Gaussian closure scheme in the quasi-linkage equilibrium regime of evolving genome populations

E. Mauri [ID], S. Cocco and R. Monasson

Laboratoire de Physique de l'Ecole Normale Supérieure, PSL Research & CNRS UM8023, Sorbonne Université 24 rue Lhomond, Paris, France

Abstract – Describing the evolution of a population of genomes evolving in a complex fitness landscape is generally very hard. We here introduce an approximate Gaussian closure scheme to characterize analytically the statistics of a genomic population in the so-called quasi-linkage equilibrium (QLE) regime, applicable to generic values of the rates of mutation or recombination and fitness functions. The Gaussian approximation is illustrated on a short-range fitness landscape with two far away and competing maxima. It unveils the existence of a phase transition from a broad to a polarized distribution of genomes as the strength of epistatic couplings is increased, characterized by slow coarsening dynamics of competing allele domains. Results of the closure scheme are corroborated by numerical simulations.

Introduction. – Understanding how genomes stochastically change as a result of selection, mutation and sexual recombination processes is the central goal of evolutionary biology. While the case of selection acting independently on the different alleles, a regime called linkage equilibrium (LE), has been intensively studied describing the dynamics of a population of genomes evolving in a complex fitness landscape, characterized by multiple and strong couplings between alleles, it remains a formidable challenge to population geneticists and statistical physicists [1]. An important step beyond LE was done by Kimura in 1965 [2] for a 2-loci system at high recombination rate r, a regime called quasi-linkage equilibrium (QLE). Recently, Neher and Shraimann extensively studied the same regime for multi-loci systems [3]. In QLE the distribution of genomes in a population may reach a stationary regime in which the pairwise correlations between alleles are weak, proportional to the fitness coupling between these two alleles and to the inverse $1/r$ of the recombination rate. To what extent this regime remains present or valid at low recombination remains unclear.

The irruption of massive sequence data, following the rapid drop in sequencing cost, has brought additional interest to such questions. It is now legitimate to ask how relevant parameters (such as mutation or recombination rates, or the full fitness function) can be extracted from sequences. An example can be found in [4], where the effective recombination rate of HIV was inferred based on a specific model for intra-patient evolution. The large-r relation between correlations and epistatic coupling was also used to infer fitness contribution from data in [5]. However, incorporating evolutionary models into inverse statistical approaches, such as the ones developed for extracting structural and functional constraints from protein sequences [6], remains a largely open issue so far.

The daunting complexity of the relationship between fitness determinants and the statistical properties of genomic data calls for the design of accurate and practical approximation schemes. We hereafter consider a Gaussian closure scheme for the hierarchy of moments of the allele distributions. Various other proposals of general, approximate frameworks in the context of the dynamics of populations can be found in the literature, *e.g.*, [7,8] and, more recently, [9,10]. In particular Barton and Turelli introduced general recursion equations for multi-locus cumulants, which they apply to the case of interchangeable additive loci under strong selection [7]. While our approximation is less general than Barton and Turelli's as it relies on the first and second moments only, it is a practical tool to deal with generic situations in which loci contribute

differently to fitness and epistatic contributions cannot be neglected.

As we will show, our Gaussian scheme allows us to bypass the large-r expansion made in [2,3], and study the existence of QLE in a larger range of recombination and mutation rates. We apply our approximation to a specific fitness model in which two very different genomes have maximal and equal fitness values, and with short-range (along the genome) and strong epistatic contributions. As the mutation and recombination rates are varied a phase transition takes place within the QLE regime, which separates a disordered (paramagnetic) regime in which alleles are not polarized along any of the two genomes and a polarized (ferromagnetic) phase in which, after coarsening of alternating allele domains, one of the two genomes dominates in the population.

Model for the stochastic evolution of a genome population. – We consider a population of N individuals, whose genomes $\mathbf{s}^{(k)} = \{s_0^{(k)}, s_1^{(k)}, \ldots, s_{L-1}^{(k)}\}$, with $k = 1, \ldots, N$, include L loci. Each locus $i = 0, \ldots, L-1$ can carry one out of two states (alleles), $s_i = \pm 1$. Genomes stochastically evolve over time t, due to selection, mutation and sexual recombination processes. We will denote by $\langle G \rangle(t)$ the average of any function $G(\mathbf{s})$ over the genomes in the population at time t,

$$\langle G \rangle(t) = \frac{1}{N} \sum_k G(\mathbf{s}^{(k)}). \qquad (1)$$

Selection. The number of individuals carrying sequence $\mathbf{s}$ grow inside the population according to the growth rate $F(\mathbf{s})$, called *fitness*. The fitness function can be expanded in terms of the state variables,

$$F(\mathbf{s}) = F_0 + \sum_i f_i s_i + \sum_{i<j} f_{ij} s_i s_j , \qquad (2)$$

where we neglect all the interactions of order equal to, or higher than three, and consider only local (f_i) and pairwise epistatic (f_{ij}) contributions. In practice, selection can be enforced in simulations as follows. Time is discretize into very small time steps Δt. At each time step selection results in a re-sampling of the available sequences according to a Boltzmann probability distribution with the weights defined by the fitness of each sequence, *i.e.*, $p(\mathbf{s}_k) \propto e^{F(\mathbf{s}_k)}$ with $k = 1, \ldots, N$. We make sure that the population size, N, is kept fixed. Note that the constant F_0 plays no role and can be set to zero.

Mutation. Each state s_i can flip its value, $s_i \to -s_i$, with rate (probability per unit of time) μ. This rate is supposed to be uniform across loci i and individuals.

Recombination. For each infinitesimal time interval Δt, with probability $r\Delta t/N_{pairs}$, a pair of individuals is chosen uniformly at random among all $N_{pairs} = \frac{1}{2}N(N-1)$ pairs, where r is the recombination rate. These two individuals, called mother (m) and father (f), undergo outcrossing. Two new genotypes $\mathbf{s}$, $\mathbf{s}'$ are formed by inheriting some

loci from the genome of the mother, $\mathbf{s}^{(m)}$, and the complement from the father with genome $\mathbf{s}^{(f)}$. This process is described by the inheritance vector $\boldsymbol{\xi} = \{\xi_i\}$, with $\xi_i \in \{0,1\}$: the i-th allele of $\mathbf{s}$ is inherited from the mother if $\xi_i = 1$ and from the father if $\xi_i = 0$, and vice versa for $\mathbf{s}'$. Hence, $s_i = s_i^{(m)}\xi_i + s_i^{(f)}(1 - \xi_i)$. The inheritance vector $\boldsymbol{\xi}$ is itself stochastic, and we denote by $C(\boldsymbol{\xi})$ its probability. In the following, we will need to consider in particular the recombination correlation $c_{ij} \equiv \sum_{\boldsymbol{\xi}} C(\boldsymbol{\xi})[\xi_i(1-\xi_j) + (1-\xi_i)\xi_j]$, which represents the probability that loci i and j take their alleles s_i and s_j from different parents.

Equations for the moments of the s_i's and closure scheme. – The above dynamical processes entirely characterizes the stochastic evolution of the genome population. We will hereafter track the first two cumulants of the allele distribution over time,

$$\chi_i(t) = \langle s_i \rangle(t), \quad \chi_{ij}(t) = \langle s_i s_j \rangle(t) - \langle s_i \rangle(t)\langle s_j \rangle(t), \qquad (3)$$

where the averages are computed over the genomes in the population at time t. Note that $\chi_{ii}(t) = 1 - \chi_i(t)^2$.

In the infinite size limit ($N \to \infty$), these moments obey deterministic first-order differential equations derived by Neher and Shraimann [3]. Unfortunately these equations are not closed, as they involve higher-order moments in the allele variables. Our closure scheme consists in replacing these high-order moments with the values they would have if the distribution of the s_i's were Gaussian, *i.e.*, in neglecting all cumulants of order ≥ 3. Qualitatively speaking, we assume a quasi-species–like scenario, in which the population is distributed around an average sequence given by $\boldsymbol{\chi} = \{\chi_0, \ldots, \chi_{L-1}\}$, with some fluctuations encoded in the covariance matrix of components χ_{ij}. As both the average sequence and the covariance matrix depend on t, the approximation offers an analytical way to track the genome population in the course of evolution.

Within the Gaussian closure scheme we obtain the following set of $\frac{1}{2}L(L+1)$ coupled equations (with the convention $f_{ij} = f_{ji}$ if $i \neq j$ and $f_{ii} = 0$), see appendix A for derivation:

$$\dot{\chi}_i = \sum_j \chi_{ij}\left[f_j + \sum_k f_{jk}\chi_k - 2f_{ij}\chi_i\right] - 2\mu\chi_i, \qquad (4)$$

and, for $i \neq j$,

$$\begin{aligned}\dot{\chi}_{ij} = &-2\chi_{ij}[f_j\chi_j + f_i\chi_i] + \sum_{k\neq l}\chi_{ik}f_{kl}\chi_{jl} \\ &-2\chi_{ij}\sum_k[f_{ki}(\chi_{ik} + \chi_i\chi_k) + f_{kj}(\chi_{jk} + \chi_j\chi_k)] \\ &+2f_{ij}\chi_{ij}[\chi_{ij} + 2\chi_i\chi_j] - (4\mu + rc_{ij})\chi_{ij}. \qquad (5)\end{aligned}$$

Case of additive fitness. In the absence of any epistatic contribution to the fitness ($f_{ij} = 0$ for all $i \neq j$), eq. (5) for the second cumulant becomes $\dot{\chi}_{ij} = -(2f_j\chi_j + 2f_i\chi_i + 4\mu + rc_{ij})\chi_{ij}$, whose only fixed point is

$\chi_{ij} = 0$. Hence, recombination and mutations contribute to erase correlation between alleles, a scenario known as LE. The resulting value of the first moment is root of $f_i(1 - \chi_i^2) - 2\mu\chi_i = 0$; it smoothly interpolates between $\chi_i = f_i/2\mu$ at large mutation rate and $\chi_i = \text{sign}(f_i)$ at vanishingly small μ, which corresponds to monoclonality.

Case of random epistatic contributions: comparison to numerical simulations and large-r theory of QLE. As a first, empirical test of the quantitative accuracy of the Gaussian closure scheme we have simulated the stochastic evolution of a population of $N = 10000$ individuals. The fitness is defined by eq. (2), with random quenched local biases f_i and epistatic couplings f_{ij} drawn from a normal distribution with zero mean and variances equal to, respectively, Δh^2 and Δf^2. In fig. 1 we show the scatter plot of the correlations obtained from simulations once stationarity is reached, χ_{ij}^{sim}, *vs.* the solutions of (5), χ_{ij}, for one generic sample. The Gaussian Ansatz correctly estimates small correlations, but is less accurate for large ones (in absolute values). It performs substantially better that the QLE approximation in the high-r regime from [3], see red and blue dots and caption of fig. 1 for details. In the inset of fig. 1 we show the root mean square (rms) z-score,

$$\epsilon = \sqrt{\frac{2}{L(L-1)} \sum_{i<j} \left(\frac{\chi_{ij} - \chi_{ij}^{sim}}{\Delta\chi_{ij}^{sim}} \right)^2}, \qquad (6)$$

as a function of Δf in the Gaussian closure. Here $\Delta\chi_{ij}^{sim}$ stands for the statistical error (standard deviation) of the correlation computed from simulation after stationarity is reached. For these values of μ and r, we obtain decent estimates of the correlations as long as Δf is not too large, see below.

Self-consistent condition for the validity of the Gaussian closure. The computation of high-order moments, which should vanish if the Gaussian approximation were exact, is sketched in appendix B. We obtain in particular that moments of order k are, for large $k(\le L)$, bounded from above by $\sim \exp(\log k!! + \frac{k}{2}\log|\chi_{ij}| + O(\log L))$, where $|\chi_{ij}|$ denotes the order of magnitude of the 2-point correlations. Due to the asymptotic scaling of the double factorial, we conclude that the k-th moment decays exponentially with k if

$$L \times |\chi_{ij}| < 1. \qquad (7)$$

In other words, the total amount of correlations that a site, say, i, "receives" from the $(L-1)$ other sites, say, j, should be smaller than unity. This condition is sufficient and not necessary, as possible cancellation between terms of opposite signs has not been considered. It is in addition self-consistent as it must be fulfilled by the 2-point correlations computed within the Gaussian closure scheme. We will check that condition (7) is satisfied for the non-trivial model with epistatic couplings studied below.
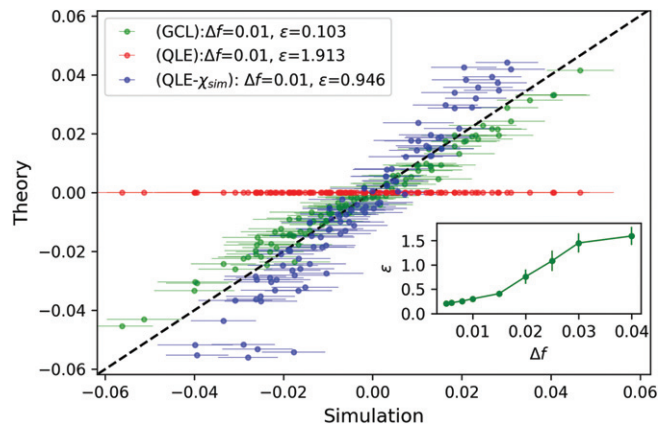

Fig. 1: Scatter plot of the correlations χ_{ij} obtained from the Gaussian closure scheme (eqs. (4), (5)) (green) or their estimates from large-r QLE theory (red: eqs. (27) of [3] for χ_i and χ_{ij}; blue: same equations for χ_{ij} only, replacing the χ_i's with their actual values χ_i^{sim} found in simulation to account for mutations) *vs.* simulations of finite populations (averaged over 200 generations once stationarity is reached) in the case of fully epistatic fitness landscape, see text and parameters in the figure. Theoretical estimates of χ_{ij} are obtained by solving the dynamical system (4), (5) within Euler approximation. Inset: rms z-score ϵ (6) as a function of the standard deviation Δf of the epistatic couplings. Parameters: $L = 15$, $\mu = 0.08$, $r = 0.85$, $\Delta h = 0.01$ and $c_{ij} = \frac{1}{2}(1 - \rho^{|i-j|})$ with $\rho = 0.6$.

Study of short-range epistatic model. – We now consider stochastic evolution under a fitness model with strong and spatially organized epistatic couplings,

$$F(\mathbf{s}) = f \sum_{i=0}^{L-1} s_i s_{i+1}, \qquad (8)$$

where $s_L \equiv s_0$ (circular chromosome); F is maximal for the two genomes $(-, -, \ldots, -)$ and $(+, +, \ldots, +)$. As for the recombination correlation, we will choose either $c_{ij} = \frac{1}{2}(1 - \rho^{|i-j|})$, which increases from 0 to $\frac{1}{2}$ with the distance between the loci, or $c_{ij} = \frac{1}{2}$ independently of the sites $i \ne j$. While the latter choice is less realistic, it does not affect the qualitative behavior of the model as we will see below, and allows for mathematical analysis in the $N \to \infty$ limit. Our goal is to understand how the distribution of alleles in the population changes as a function of μ and r. Hereafter, we define $x = \mu/f$, $y = r/(2f)$.

Paramagnetic phase. – Due to the absence of local biases ($f_i = 0$), the fitness is left unchanged by a global flip of all loci, $s_i \to -s_i$. In the absence of spontaneous symmetry breaking and in agreement with (4), we assume that $\chi_i = 0$ for all sites i. In this paramagnetic (PM) regime we look for a stationary and translation-invariant solution, $\chi_{ij} = \chi^{st}(d = |i - j|)$, to the coupled equations (5) owing to the periodic boundary conditions in (8). We find that,

for all $d \geq 1$,

$$
\chi^{st}(d+1) = [3\chi^{st}(1) + a]\chi^{st}(d) - \chi^{st}(1)^2\delta_{d,1}
$$
$$
- \frac{1}{2}\sum_{k=0}^{d-1}\sum_{s=0,1}\chi^{st}(k+s)\chi^{st}(d-k-1+s)
$$
$$
- \sum_{k=1}^{\infty}\sum_{s=0,1}\chi^{st}(k+s)\chi^{st}(d+k+1-s),
$$

$$(9)$$

where $a \equiv 2x + \frac{y}{2}$. To obtain a tractable approximation for $\chi^{st}(d)$ we further neglect all the terms with $k > d+1$ in the r.h.s. (last term) of (9). This approximation is based on the assumption that correlations decay quickly with the distance between loci, which will be checked *a posteriori*. Within this approximation the generating function $\tilde{\chi}^{st}(z) \equiv \sum_{d\geq 0}\chi^{st}(d)z^d$ is given by $G(z, u = 0)$, where

$$
G(z,u) = \frac{z(3\chi^{st}(1) + a + 4u) \pm \sqrt{\Delta(z,u)}}{1 + z^2}, \quad (10)
$$

and $\Delta(z,u) = z^2(3\chi^{st}(1) + a + 4u)^2 + (1 + z^2)[(1 - u)^2 - 2z(1-u)(2\chi^{st}(1) + 4u + a) - 2z^2\chi^{st}(1)(\chi^{st}(1) + 2u)]$; notice that $\chi^{st}(1) = d\tilde{\chi}^{st}(z)/dz|_{z=0}$.

The singularities of $\tilde{\chi}^{st}(z)$ control the asymptotic behaviour of $\chi^{st}(d)$ [11]. More precisely, if $\tilde{\chi}^{st}(z)$ behaves as $(z_c - z)^{\beta}$ close to its singularity, then

$$
\chi^{st}(d) \sim z_c^{-d}\, d^{-(\beta+1)} \quad (11)
$$

for large d; here, the square root in (10) entails that $\beta = \frac{1}{2}$. To locate the singularity z_c we identify the roots of $\Delta(z,0)$. As Δ is a polynomial of degree 4, it has two real roots and two conjugated complex roots, with modulus lower than 1. For values of $a > a_c \simeq 2.542$, there is one appropriate value for $\chi^{st}(1)$, which depends on a, such that the two complex roots merge on the real axis. This choice, accompanied by a change of sign in front of $\sqrt{\Delta}$ in (10), removes the corresponding singularity and ensures that the radius of convergence z_c of G is larger than unity, as required for the exponential decay of $\chi^{st}(d)$. The radius of convergence z_c and the corresponding value of $\chi^{st}(1)$ are shown as functions of a in fig. 2; for large a, $z_c \simeq \frac{a}{2} \gg 1$. Therefore the correlation function $\chi^{st}(d)$ (11) decays very quickly with d, as expected for large mutation or recombination rates. For $a < a_c$ removal of the singularity corresponding to the small-modulus complex roots is not possible, indicating the presence of a qualitatively different regime we will study below.

We compare in fig. 3 the average values of the correlations χ_{ij} obtained by simulating a finite population with the ones computed above using the generating function $\tilde{\chi}^{st}(z)$. The agreement is excellent for small d, and more difficult to assess at large d due to the finite-size and noise effects in the simulations.
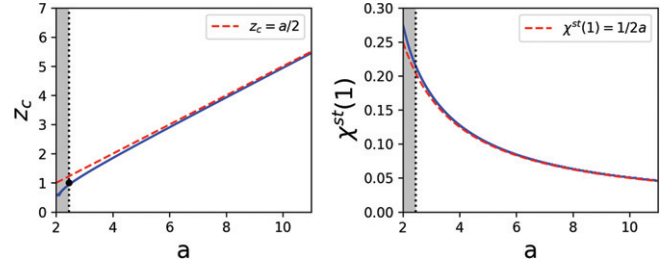

Fig. 2: Radius of convergence z_c (left) and corresponding value of $\chi^{st}(1)$ (right) *vs.* $a = 2x + \frac{y}{2}$ in the PM regime. Red dotted lines show the large-a asymptotic behaviours.
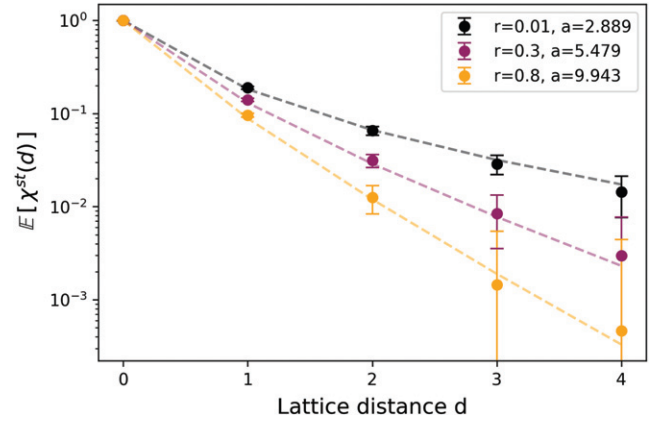

Fig. 3: Average correlations in the PM phase for a finite population (dots) against the value computed from Gaussian closure equations (4), (5) (dashed lines) through Euler approximation, for three recombination rates r. Parameters: $f = 0.028$, $L = 50$, $\mu = 0.03$, $N = 10^4$, $c_{ij} = \frac{1}{2}(1 - \rho^{|i-j|})$ with $\rho = 0.6$; results are obtained by averaging over 5000 generations.

Critical line and spontaneous symmetry breaking. – The condition $a > a_c$ is necessary for the correlations to decay with distance, but it is not sufficient for the paramagnetic phase to hold. To investigate the onset of non-zero χ_i's, we study the stability of the PM solution found above against small fluctuations of the frequencies χ_i. To do so we linearize eq. (4) at fixed correlations $\chi_{ij} = \chi^{st}(|i - j|)$, and obtain $\dot{\chi}_i = -2f\sum_j \mathcal{M}_{ij}\chi_j$, where the entries of the $\mathcal{M}$ matrix read

$$
\mathcal{M}_{ij} = [2\chi^{st}(1) + x]\delta_{ij} - \frac{1}{2}[\chi^{st}(|j-i+1|) + \chi^{st}(|j-i-1|)],
$$
$$(12)$$

Being $\mathcal{M}$ translation invariant it can be diagonalized by Fourier plane waves, and its eigenvalues λ_n are given by, for $n = 0, 1, \ldots, L-1$,

$$
\lambda_n = 2\chi^{st}(1) + x - [2\,\mathrm{Re}\,\tilde{\chi}^{st}(e^{-i\frac{2\pi}{L}n}) - 1]\cos\left(\frac{2\pi n}{L}\right),
$$
$$(13)$$

where $\tilde{\chi}$ is the generating function in (10). The PM phase is stable as long as all the eigenvalues λ_n are strictly positive. The vanishing of λ_0, the lowest eigenvalue of $\mathcal{M}$, thus defines the boundary of the paramagnetic phase. This
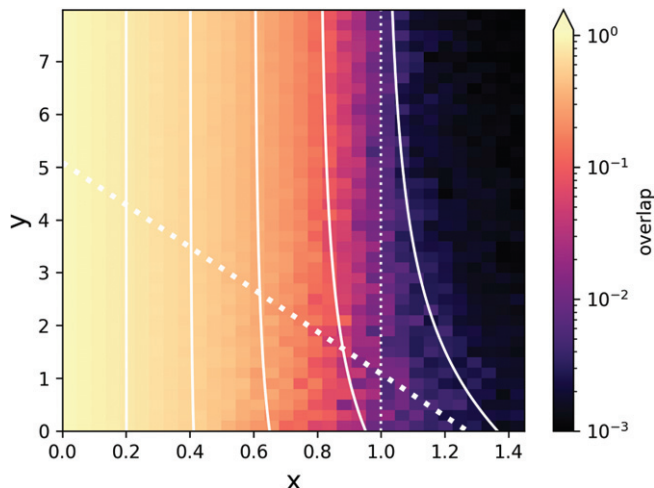
Fig. 4: Phase diagram for genomes evolving under fitness (8) in the $(x = \mu/f, y = r/(2f))$ plane. The heatmap shows the Edwards-Anderson overlap q, see text and color bar, for a simulated population of $N = 10^4$ individuals in the case of non-homogeneous recombination matrix $c_{ij} = \frac{1}{2}(1 - \rho^{|i-j|})$ with $\rho = 0.6$. The theoretical white lines were derived for long-range recombination correlation $c_{ij} = \frac{1}{2}$, see main text. The critical line $x = 2\tilde{\chi}^{st}(1) - 2\chi^{st}(1) - 1$ (rightmost solid curve) separates PM (right) and FM (left) phases, and approaches the $x = 1$ line (thin dashed) when $y \gg 1$. Theoretical contour lines corresponding to $\chi^2_{FM} = 0, .2, .4, .6, .8$ (from right to left) are shown in the FM phase (full lines), in fair agreement with numerical findings for the short-range model, see color code. The thick dashed line stands for $a = a_c \approx 2.542$, showing that $a > a_c$ throughout the PM phase.

critical line is shown in the (x, y) plane in fig. 4, and corresponds to the condition $x = 2\tilde{\chi}^{st}(1) - 2\chi^{st}(1) - 1$. We compared our analytical results, valid for $N \to \infty$, with simulations for a large but finite population size. We report the values of the Edward-Anderson overlap $q = \frac{1}{L}\sum_i \chi_i^2$ in fig. 4. We observe that q takes very low values to the right of the critical line where the paramagnetic phase is predicted to be stable, which justifies *a posteriori* our approximation above. To the left of the critical line, $q > 0$, which signals the onset of a ferromagnetic phase; the mutation rate is too weak to ensure that the average values of the s_i's are zero.

Ferromagnetic phase. – In the broken symmetry phase, hereafter referred to as ferromagnetic (FM), we expect the allele frequencies not to vanish any longer, and be equal to $\chi_i = \chi_{FM}$, up to a global reversal symmetry. Repeating the procedure followed in the PM phase we again consider the generating function $\tilde{\chi}^{st}(z)$ for the correlations $\chi^{st}(d)$ at distances $d \geq 1$. According to eq. (4) the average allele frequency reads

$$
\begin{aligned}
\chi^2_{FM} &= 1 - x + 2\sum_{d \geq 2}\chi^{st}(d) \\
&= x - [2\tilde{\chi}^{st}(1) - 2\chi^{st}(1) - 1],
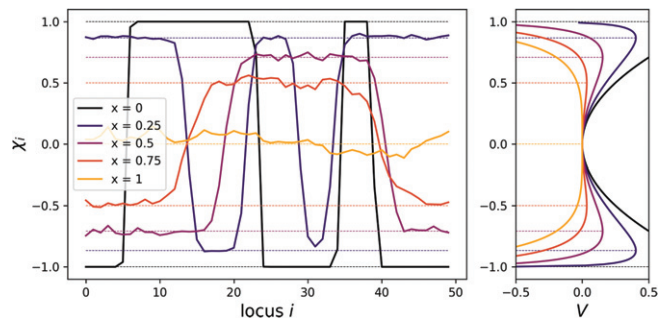\end{aligned} \tag{14}
$$

Fig. 5: Left: allele frequencies χ_i for one population evolved over 500 generations in the FM phase and for different values of the ratio x. Parameters: $f = 0.028$, $N = 10^4$, $L = 50$, $r = 0.85$ and $c_{ij} = \frac{1}{2}(1 - \rho^{|i-j|})$ with $\rho = 0.6$. Dashed horizontal lines show $\pm\sqrt{1-x}$. Right: effective potential $V(\chi)$ for the respective value of x shown in the legend and horizontal lines locating the maxima of the potential in $\chi_{FM} = \pm\sqrt{1-x}$.

where the generating function of the correlations is now given by $\tilde{\chi}^{st}(z) = G(z, \chi^2_{FM})$, see (10). As in the PM case, $\chi^{st}(1)$ is appropriately chosen to remove the spurious singularities inside the unit circle in the complex z-plane. By solving numerically the self-consistent equations (10) and (14) we were able to compute χ^2_{MF} throughout the (x, y) plane. As expected, χ_{MF} is non-zero to the left of the border line of the region of stability of the PM phase; contour lines for χ^2_{FM} in the FM phase are shown in fig. 4. Notice that, for a large range of parameter values, correlations decay very fast with distance as expected in the QLE phase, and $\chi_{FM} \simeq \sqrt{1-x}$ according to eq. (14). In addition, we have numerically checked that condition (7) is met everywhere in the FM and PM phases, *i.e.*, $2\sum_{d \geq 1}\chi^{st}(d) < 1$.

Short-term dynamics: allele domains. Simulations show that, after a transient, the χ_i's reach apparently stationary values, with up and down domains of height $\pm\chi_{FM}$, see fig. 5. Domains are of variable lengths, and are separated by characteristic domain walls.

The shape of walls separating contiguous domains can be analytically characterized in the $r \to \infty$ limit. For x close to 1 and $L \gg 1$, a smooth, continuum limit for the index $i \to z$ can be taken. In this approximation, we write $\chi_{i+1} + \chi_{i-1} - 2\chi_i \simeq \chi''(z)$ in the stationary regime, and find, according to (4),

$$
\chi''(z) = \frac{2x\chi(z)}{1 - \chi^2(z)} - 2\chi(z). \tag{15}
$$

After integration, we obtain $\frac{1}{2}\chi'(z)^2 + V(\chi(z)) = E$, where E is a constant and the potential is $V(\chi) = x\ln(1 - \chi^2) + \chi^2$. Hence, finding the stationary profile of the spin average values is equivalent to a classical mechanics problem for a point-like particle of coordinate χ and unit mass, moving in the potential energy V over fictitious time z. The potential $V(\chi)$ is symmetric in 0, has two maxima in $\chi = \pm\chi_{FM}(x)$ and a local minimum
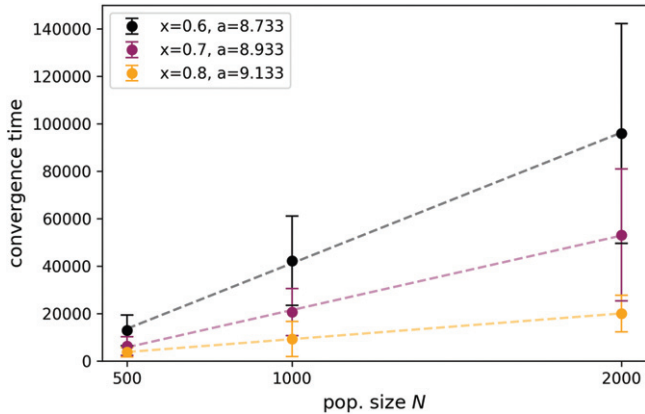
Fig. 6: FM phase: number of generations required for convergence towards the uniform frequency state $\chi = \chi_{FM}$ vs. population size N for different values of x. The dashed lines stand for the linear regression of the data. Each point is the average over 10 simulations. Parameters: $r = 0.85$, $f = 0.028$, $L = 50$, $c_{i \neq j} = \frac{1}{2}$.

in $\chi = 0$ see fig. 5 (right). The trajectories $\chi(z)$ with conserved energy $E = V(\chi_{FM})$ define the domain wall profiles. These profiles depend on the value of x, and are in good agreement with fig. 5 (left).

Long-term dynamics: domain coarsening. On longer time scales, the domain walls described in fig. 5 undergo a diffusive process, due to the noise resulting from finite (population) size sampling. Domain walls merge when they encounter, and the system will eventually settle into a final state with uniform frequency $\pm\chi_{FM}$. The characteristic time the system needs to converge into the uniformly magnetized state is shown in fig. 6. It depends on the population size N and on the genome length L, and is expected to scale as

$$\tau \sim \tau_0 \times L^2 \times N, \tag{16}$$

where τ_0 depends on the intensive parameters, such as the mutation/fitness ratio x. The scaling in (16) stems from the following simple argument. The typical size of domains grow with time as $\sim t^{1/2}$ according to the Allen-Cahn theory of coarsening for non-conserved fields [12]. Hence a single domain will occupy the whole genome after a time given by L^2/D, where D is the diffusion coefficient. Here, D is expected to be inversely proportional to the population size N. This scaling is corroborated by the numerical results reported in fig. 6.

Conclusions. – We have here introduced a Gaussian closure scheme for the dynamics of multi-allele correlations across an infinite-size population of genomes evolving under recombination, mutation and selection. Informally speaking, the distribution of genomes at time t is described as a cloud of covariance matrix $\chi_{ij}(t)$ around an "average" genome (set of allele frequencies) $\chi_i(t)$. These 1- and 2-point correlations obey a set of $\frac{1}{2}L(L+1)$ coupled

deterministic, non-linear first-order differential equations, where L is the genome length. For the sake of simplicity, we have assumed that alleles can be in one out of two possible states. While this assumption may already lead to quantitative behaviors similar to real systems [13] it can be easily extended to multi-categorical (Potts) variables [5].

Our scheme can be formally seen as a practical way to investigate the QLE regime beyond the large-r perturbative approach initiated by Kimura [2]. While our equations (5) allow us to find back the QLE expressions for χ_{ij} at high recombination [14], they include contributions to all orders in powers of $1/r$ and can be exploited for generic recombination and mutation rates, as well as for any fitness function, and unveil non-trivial behaviors taking place at intermediate values of their defining parameters. An illustration is provided by the model analyzed in detail in this paper, in which two competing genomes, far away from each other in terms of allele composition, have maximal and equal fitness. We find, at low enough mutation rate, a phase transition from a paramagnetic (PM) phase in which the distribution is broad and encompasses the two fittest genomes to a ferromagnetic (FM) regime, in which one of the genomes eventually dominates. We stress that, both in the PM and FM phases, correlations decay exponentially with the distance between the alleles along the genome, which is also their distance on the epistatic interaction graph. Hence, both phases formally belong to the QLE regime, even when r is not large and μ is not small. However, contrary to a usual, informal definition of QLE, the genome distribution is not parametrized by the allele frequencies χ_i only, but also by the correlations χ_{ij}, as a consequence of the finiteness of r and μ.

Some aspects of the present work would deserve further investigation. First, while determining the range of validity of closure schemes is generally a complicated issue, *e.g.*, in the case of small chemical systems [15], we have here derived some sufficient condition, see eq. (7). This condition will break down if the genome distribution is multi-modal and correlations are not small, as is the case in the Clonal Competition (CC) regime [16]. However, even in CC, we may expect that our scheme remains adequate to describe locally the distribution of genomes attached to a clone. Therefore, it would be interesting to consider mixtures of Gaussian Ansätze, each describing a clone, an approximation which could be numerically tractable for sufficiently small numbers of clones and alleles.

Second, our approach offers, in principle, a practical way to infer the fitness determinants (f_i, f_{ij}) from the statistics of sequence data. Approaches to interpret correlations are needed as epistasis is notoriously difficult to infer from observation of mean fitness alone [17]. Recently, the present Gaussian scheme was successfully used to reconstruct epistatic contributions from synthetic data [14]. We hope the present approach will help put the inverse statistical approaches to protein sequence data [6] onto firm population genetics grounds.

* * *

Appendix A: derivation of dynamical equations for the first and second moments. − Under the selection, mutation and recombination steps described in the main text the distribution of genomes at time t, $P(\mathbf{s},t)$, obeys the Fokker-Planck–like equation:

$$\dot{P}(\mathbf{s},t) = (F(\mathbf{s}) - \langle F \rangle)P(\mathbf{s},t) + \mu \sum_{i=1}^{L}(P(\mathbf{s}^{(i)},t) - P(\mathbf{s},t))$$
$$+ r \sum_{\boldsymbol{\xi},\mathbf{s}'} C(\boldsymbol{\xi})(P(\mathbf{s}^m,t)P(\mathbf{s}^f,t) - P(\mathbf{s},t)P(\mathbf{s}',t)),$$

$$(A.1)$$

which is identical to eq. (7) in [3]. In the above equation, $\mathbf{s}^{(i)}$ denotes genome $\mathbf{s}$ in which allele i has been flipped, $s_i \to -s_i$. By multiplying the above equation by s_i or $s_i s_j$ and summing over all genomes $\mathbf{s}$ we obtain the evolution equations for the moments χ_i or χ_{ij}, that is, $\dot{\chi}_i = \langle s_i(F(\mathbf{s}) - \langle F \rangle) \rangle - 2\mu \chi_i$ and $\dot{\chi}_{ij} = \langle (s_i - \chi_i)(s_j - \chi_j)(F(\mathbf{s}) - \langle F \rangle) \rangle - (4\mu + rc_{ij})\chi_{ij}$, where the fitness function is defined in (2). The above equations coincide with eqs. (8) and (10) in [3]. We are left with the evaluation of the averages $\langle \cdot \rangle$ in the above equations. After replacement of quadratic terms s_k^2 with 1, the polynomial to be averaged is of degree zero or one in each variable $a_k = s_k - \chi_k$. We then forget about the discrete nature of the a_k's, and take them as multivariate Gaussian variables with zero means and covariance matrix $\chi_{k\ell}$. Hence all odd moments vanish, and even moments can be computed from Wick's theorem, e.g., $\langle a_k a_\ell a_m a_n \rangle = \chi_{k\ell}\chi_{mn} + 2$ other pairings. A systematic evaluation of all terms lead to (4) and (5).

Appendix B: contributions to the k-th moment. − The approximation scheme described in appendix A can be used to compute moments of any order, such as $\chi_{i_1,i_2,\ldots,i_k} = \langle \prod_{\alpha=1}^{k}(s_{i_\alpha} - \chi_{i_\alpha}) \rangle$ with $k \geq 3$. In the stationary regime, we obtain

$$\chi_{i_1,i_2,\ldots,i_k} = \frac{1}{G_k}\left\langle \left\{ \prod_{\alpha=1}^{k} a_{i_\alpha} - \sum_{\alpha=1}^{k}(a_{i_\alpha} + \chi_{i_\alpha})\left\langle \prod_{\beta(\neq\alpha)} a_{i_\beta} \right\rangle \right\} \right.$$
$$\left. \times \left(\sum_i \hat{f}_i a_i + \sum_{i<j} f_{ij}(a_i a_j - \chi_{ij}) \right) \right\rangle \qquad (B.1)$$

where $\hat{f}_i = \sum_{j(\neq i)} f_{ji}\chi_j$, $G_k = 2k\mu + rc_{i_1,i_2,\ldots,i_k}$, and $c_{i_1,i_2,\ldots,i_k}$ is the probability (computed with $C(\xi)$) that all k alleles do not come from the same parent (father

or mother). Careful inspection of the polynomial in the a_k variables (after replacements $a_i^2 \to 1 - \chi_i^2 - 2\chi_i a_i$ to enforce $s_i^2 = 1$) and of the effects of Wick's contraction shows that terms appearing in the average $\langle \cdot \rangle$ above are of the form $M \times C \times (\chi_{ij})^E \times I$, where χ_{ij} stands for a generic 2-point moment, and 1) the exponent E is equal to $\frac{\ell}{2}$, where ℓ is an integer comprised between $k-2$ and $k+2$, depending on the parity of k and the term considered; 2) the combinatorial factor C is equal to $(E-1)!!$; 3) the multiplicity M is bounded by L^3; 4) the interaction term I is a polynomial of degree ≤ 2 in two of the χ_{i_α}'s ($\alpha = 1,\ldots,k$), and is proportional to one local fitness term $\hat{f}_i$ or epistatic coupling f_{ij}.

Note added in proofs: In the absence of recombination ($y = 0$) our fitness model is equivalent to a quantum Ising chain in transverse field and the transition takes place at $x = \mu/f = 1$ [18].

REFERENCES

[1] BARTON N. H. and DE VLADAR H. P., *Genetics*, **181** (2009) 997.
[2] KIMURA M., *Genetics*, **52** (1965) 875.
[3] NEHER R. A. and SHRAIMAN B. I., *Rev. Mod. Phys.*, **83** (2011) 1283.
[4] NEHER R. A. and LEITNER T., *PLoS Comput. Biol.*, **6** (2010) 1.
[5] GAO C.-Y., CECCONI F., VULPIANI A., ZHOU H.-J. and AURELL E., *Phys. Biol.*, **16** (2019) 026002.
[6] COCCO S., FEINAUER C., FIGLIUZZI M., MONASSON R. and WEIGT M., *Rep. Prog. Phys.*, **81** (2018) 032601.
[7] TURELLI M. and BARTON N. H., *Genetics*, **138** (1994) 913.
[8] BUERGER R., *The Mathematical Theory of Selection, Recombination, and Mutation* (John Wiley & Sons) 2000.
[9] JAIN K. and STEPHAN W., *Mol. Biol. Evol.*, **34** (2017) 3169.
[10] BARGHI N., HERMISSON J. and SCHLOTTERER C., to be published in *Nat. Rev. Genet.* (2020), https://doi.org/10.1038/s41576-020-0250-z.
[11] FLAJOLET P. and SEDGEWICK R., *Analytic Combinatorics* (Cambridge University Press, Cambridge) 2009.
[12] BRAY A. J., *Adv. Phys.*, **43** (1994) 357.
[13] SHEKHAR K., RUBERMAN C. F., FERGUSON A. L., BARTON J. P., KARDAR M. and CHAKRABORTY A. K., *Phys. Rev. E*, **88** (2013) 062705.
[14] ZENG H.-L., MAURI E., DICHIO V., COCCO S., MONASSON R. and AURELL E., arXiv:2006.16735 (2020).
[15] SCHNOERR D., SANGUINETTI G. and GRIMA R., *J. Chem. Phys.*, **143** (2015) 185101.
[16] NEHER R. A., VUCELJA M., MEZARD M. and SHRAIMAN B. I., *J. Stat. Mech.: Theory Exp.*, **2013** (2013) P01008.
[17] McCANDLISH D. M., OTWINOWSKI J. and PLOTKIN J. B., *Evolution*, **69** (2015) 2359.
[18] BAAKE E., BAAKE M. and WAGNER H., *Phys. Rev. Lett.*, **78** (1997) 559.