Statistical genetics and evolution of quantitative traits

Richard A. Neher

*Kavli Institute for Theoretical Physics, University of California,
Santa Barbara, California 93106, USA,
and Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany*

Boris I. Shraiman

*Kavli Institute for Theoretical Physics, and Department of Physics, University of California,
Santa Barbara, California 93106, USA*

The distribution and heritability of many traits depends on numerous loci in the genome. In general, the astronomical number of possible genotypes makes the system with large numbers of loci difficult to describe. Multilocus evolution, however, greatly simplifies in the limit of weak selection and frequent recombination. In this limit, populations rapidly reach quasilinkage equilibrium (QLE) in which the dynamics of the full genotype distribution, including correlations between alleles at different loci, can be parametrized by the allele frequencies. This review provides a simplified exposition of the concept and mathematics of QLE which is central to the statistical description of genotypes in sexual populations. Key results of quantitative genetics such as the generalized Fisher's "fundamental theorem," along with Wright's adaptive landscape, are shown to emerge within QLE from the dynamics of the genotype distribution. This is followed by a discussion under what circumstances QLE is applicable, and what the breakdown of QLE implies for the population structure and the dynamics of selection. Understanding the fundamental aspects of multilocus evolution obtained through simplified models may be helpful in providing conceptual and computational tools to address the challenges arising in the studies of complex quantitative phenotypes of practical interest.

PACS numbers: 87.23.Kg, 87.10.Vg

CONTENTS

I. INTRODUCTION

R. A. Fisher's celebrated "fundamental theorem of natural selection," relating the rate of change in the average fitness to the genetic variance in fitness, occupies a place in population genetics similar to Newton's $F = ma$ in physics. Yet conceptually Fisher's law and the whole subject of "quantitative genetics" (QG) (Falconer and Mackay, 1996; Lynch and Walsh, 1998), which studies the response of quantitative traits to selection, is closer to thermodynamics. Thermodynamics is a phenomenological description of readily measurable physical properties (e.g., average energy or pressure) of a large ensemble of molecules. Quantitative genetics is a phenomenological description of readily observable phenotypic traits of a population. Thermodynamics takes macroscopic averages over the random motion of individual molecules in thermal equilibrium. Quantitative genetics similarly focuses on the behavior of population-wide averages (and variances) over many genetically diverse individuals. The genetic composition of the population is governed by natural selection and random drift along with recombination and mutation, all acting on individuals. The phenotype distribution is related to the genotype distribution by the largely unknown genotype-to-phenotype map, which is further obscured by environmental effects which can cause phenotypic variation even between genetically identical individuals. Yet deterministic laws of thermodynamics emerge despite the complexity and chaos of molecular motion. In fact, they emerge thanks to the microscopic complexity and chaos and are made possible by

the extensive self-averaging that dominates macroscopic behavior of physical matter. Similarly, simple laws of quantitative population genetics emerge when phenotypic traits depend on large numbers of polymorphic genetic loci.

While the analogy between QG and thermodynamics is most appealing and has been noted by many including R. A. Fisher himself (Fisher, 1930), see Iwasa (1988), Sella and Hirsh (2005), and Barton and Vladar (2009) for recent work, fundamental issues such as the lack of an energylike conserved quantity in population genetics impede direct transcription of thermodynamic laws to QG. Instead, the analogy must be pursued as an approach to the construction of a coarse-grained phenomenological theory bridging the gap between ensemble averaged (read population averaged) observables and the hidden microscale (read individual genotype) dynamics. One must be careful to define an averaging ensemble that equilibrates on the time scale of the observation, e.g., the response to selection in QG. In particular, as we illustrate in Fig. 1, dynamics in sexually reproducing populations are characterized by two widely different time scales: (1) mating and recombination reshuffle the polymorphic loci,
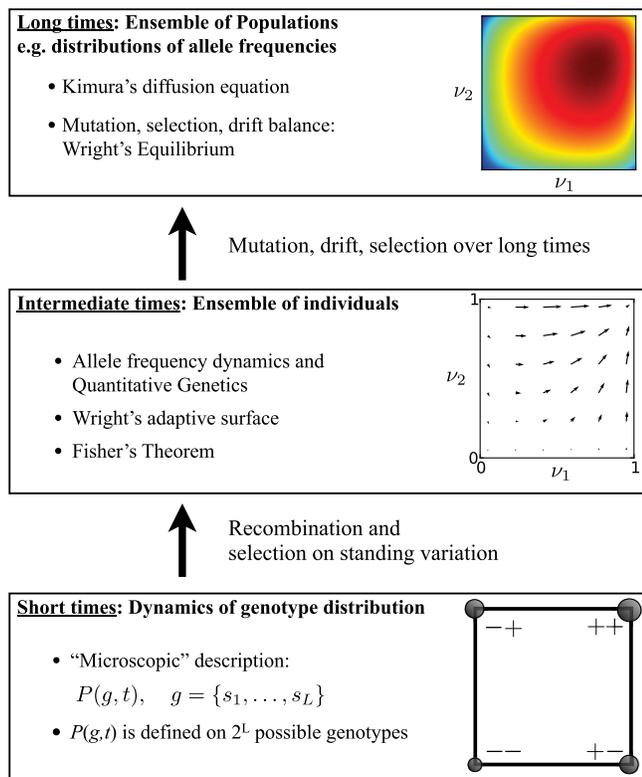

FIG. 1 (color online). Time scales in sexual populations. A population is described by the distribution of 2^L genotypes, on which selection, recombination, and mutation acts. In sexual populations, mating and recombination is the fastest process, so that different loci are only weakly correlated (close to linkage equilibrium) and its dynamics can be approximately described via L allele frequencies, a number much smaller than the 2^L possible genotypes that would have to be tracked otherwise. Allele frequencies change slowly and means of quantitative traits follow the laws of quantitative genetics. Over the much longer time scale of μ^{-1}, allele frequencies themselves tend to an equilibrium between selection, mutation, and genetic drift (assuming a constant environment).

allowing exploration of the space of genotypes on a short time scale; and (2) mutation and population drift control genetic variation on much longer time scales, often long enough to render the ensemble meaningless.

The bridge between the dynamics of the genotype distribution and the coarse-grained, QG-type description is built on understanding multilocus evolution. Our review focuses on the intermediate time scale in the above-mentioned hierarchy. We show how the genotype distribution $P(g, t)$ can be parametrized by slowly varying allele frequencies, while mating and recombination lead to rapid equilibration of $P(g, t)$ given a set of allele frequencies. In this ensemble trait distributions are determined by allele frequencies and the dynamics of trait averages can be expressed in terms of the dynamics of allele frequencies. This in turn gives rise to the familiar laws of quantitative genetics in terms of additive variances and covariances. In this sense, a statistical multilocus theory plays the role of statistical mechanics, which explains how the deterministic laws of thermodynamics emerge from the erratic motion of many microscopic particles. Hence the subject of this review should be thought of as "statistical genetics," a term introduced in a closely related context by Wright (1942).

Classical quantitative genetics (Falconer and Mackay, 1996) is based on the assumption that genotypes are random reassortments of alleles, each occurring with a certain frequency. This absence of correlations between alleles at different loci is termed "linkage equilibrium," (LE) implying that recombination (breaking linkage) has relaxed correlations between loci. This drastic simplification has earned QG a derogatory epithet of "beanbag genetics" from the pen of Ernst Mayr (1963) [see, however, Haldane (1964) in defense of beanbag genetics]. Yet in this review we see that the key phenomenological laws of QG extend beyond the assumption of linkage equilibrium. This understanding emerges from the studies of multilocus selection which began with two alleles and/or two loci systems (Kimura, 1956; Lewontin and Kojima, 1960; Karlin and Feldman, 1969). Kimura (1965) showed that a two-locus system tends toward a state where allele frequencies change slowly and correlations are small and steady. He termed this state quasilinkage equilibrium (QLE), which is the subject of this review. Subsequently, several comprehensive treatments of multilocus evolution were developed (Christiansen, 1990; Barton and Turelli, 1991; Bürger, 1991; Nagylaki, 1993; Prügel-Bennett and Shapiro, 1994; Baake, 2001) [for a monograph see Bürger (2000)] with Barton and Turelli (1991) and Nagylaki (1993), in particular, generalizing and justifying the QLE approximation in multilocus systems.

In addition to studying the generic behavior of systems with a very large number of loci, explicit multilocus modeling of smaller systems has been used to study the evolution of recombination (Barton, 1995a; Roze and Barton, 2006) and patterns of genetic variation produced by positive selection (Stephan, Song, and Langley, 2006). Recent work produced interesting examples (Weinreich *et al.*, 2006; de Visser, Park, and Krug, 2009) of empirically determined fitness landscapes with five or more loci. The dynamics of populations on these landscapes can be studied in laboratory experiments and comparison to theoretical models is possible (de Visser, Park, and Krug, 2009). Quantitative understanding of multilocus

evolution is also essential when studying the emergence of drug resistance in HIV, which often depends on several interacting loci in a recombining population (Kellam *et al.*, 1994; Bretscher *et al.*, 2004; Nora *et al.*, 2007).

Our discussion of the multilocus selection problem will follow the Barton-Turelli course of making and keeping it simple (Barton and Turelli, 1991; Turelli and Barton, 1994), by attempting to make it simpler still. We define a streamlined conceptual and analytic framework which will not only reproduce classic results, but also readily generate some new extensions. To accomplish this we formulate and analyze a "minimal model" of multilocus evolution: continuous time selection in a haploid model with a general (epistatic) fitness function of L loci. We show how the generalized Fisher theorem and other results of quantitative genetics follow from a straightforward cumulant perturbation theory similar to that used extensively (for high-temperature expansions) in statistical physics (McQuarrie, 1973). The perturbative regime corresponds to Kimura's QLE. Using this formulation of QLE, we present systematic generalizations of QG results and of Kimura's diffusion theory, typically formulated in complete linkage equilibrium, to include weak correlations between loci. We also discuss how QLE breaks down when the ratio of characteristic strength of selection to the rate of recombination exceeds a critical value that depends on the strength of epistasis. While the QLE regime corresponds to the selection of individual alleles based on their effect on fitness averaged over genetic backgrounds, the breakdown of QLE follows the appearance of strong correlations between alleles at different loci and represents a transition to the effective selection of genotypes. In Sec. VIII we connect the transition from the "allele" selection to the "genotype" selection to the closely related spin-glass transition (modeling the behavior of disordered magnets) studied in statistical physics (Mezard, Parisi, and Virasoro, 1987). We also discuss its implications for quantitative genetics.

II. RELATING QUANTITATIVE TRAITS AND GENOTYPES

We focus on the fitness which is the most important "quantitative trait," although everything we say about "fitness landscapes" in this section applies directly to any quantitative phenotype. A fitness landscape is a metaphor for a map from the high dimensional space of genotypes to expected reproductive success. While the map itself is unambiguous, several different ways of parametrizing fitness landscapes with alleles and groups of alleles have been proposed (Barton and Turelli, 1991; Weinberger, 1991; Hansen and Wagner, 2001; Hansen, 2006).

Consider a haploid genome of L loci with two alleles each, such that a genotype is uniquely characterized by L binary variables $g = \{s_1, \ldots, s_L\}$. We choose $s_i \in \{-1, 1\}$, $i = 1, \ldots, L$ instead of $s_i \in \{0, 1\}$ (more commonly used in population genetics literature), since the symmetric choice simplifies the algebra below. (The relation between representations can be found in Appendix A.) Functions of the genotype, e.g., as the population distribution, fitness, or any other quantitative trait, live therefore on an L-dimensional hypercube. Any such function on the hypercube can be decomposed into a sum of monomials in s_i:

$$F(g) = \bar{F} + \sum_i f_i s_i + \sum_{i<j} f_{ij} s_i s_j + \sum_{i<j<k} f_{ijk} s_i s_j s_k + \cdots,$$

(1)

where the first sum represents independent contributions of L single loci, the second sum which runs over all $L(L-1)/2$ pairs of loci represents contributions of pairs, and the higher order terms account for the effect of each and every possible subgroup of loci. The first order contribution f_i defines the additive effect of locus i which is independent of all other loci considered. Higher order terms which include locus i define the genetic background dependence of the effect of the s_i allele. Collectively terms of order higher than 1 represent genetic interactions also known as "epistasis." The contribution of each locus or subgroup of loci is determined by unbiased (i.e., each genotype enters with weight 2^{-L}) averaging over the remainder of the genome: thus the coefficients are given by

$$\bar{F} = 2^{-L} \sum_g F(g), \qquad f_i = 2^{-L} \sum_g s_i F(g),$$
$$f_{ij} = 2^{-L} \sum_g s_i s_j F(g).$$

(2)

One easily convinces oneself that plugging Eq. (1) into Eq. (2) reduces to the desired coefficients. In total, there are 2^L coefficients $f^{(k)}_{i_1,\ldots,i_k}$, as it has to be for an exact representation of a function on a hypercube. In fact, the coefficient of the expansion of $F(g)$ into monomials is nothing but the Fourier transform of the original function on the hypercube, which was used in the context of genotype-fitness maps by Hordijk *et al.* (1998), Stadler and Wagner (1998), and Weinberger (1991). In addition to this genetic contribution to the trait, the trait value of a given individual will also depend on environmental (and epigenetic) factors which are not modeled here.

It proves useful to define a "density of states" $\rho(F) = 2^{-L} \sum_g \delta(F - F(g))$, where $\delta(F)$ is a Dirac delta function. The fraction of genotypes with fitness in the interval $[F, F + \delta F]$ is then given by $\int_F^{F+\delta F} dF' \rho(F')$. Provided $F(g)$ receives contributions of many terms of similar magnitude in Eq. (1), the central limit theorem will apply making the density of states approximately Gaussian in shape. The width of this Gaussian is given by the (square root of the) variance of $F(g)$ over the hypercube:

$$\bar{\sigma}^2 = 2^{-L} \sum_g (F(g) - \bar{F})^2$$
$$= \sum_i f_i^2 + \sum_{i<j} f_{ij}^2 + \sum_{i<j<k} f_{ijk}^2 + \cdots.$$

(3)

This simple decomposition of variance is the equivalent of Parseval's theorem for the Fourier transform. Note that this variance is an intrinsic property of the fitness landscape completely independent of any population that may be evolving on it. It should not be confused with the population variance that we discuss later. We use $\bar{\sigma}$ as a measure of selection strength.

The sums in Eq. (3) for $\bar{\sigma}$ can be interpreted as the power spectrum of the $F(g)$. A falling or rising power spectrum gives rise to qualitatively different landscapes: If most of the variation of the fitness function were captured by the first

order terms, the landscape would be smooth and simple. If higher order terms dominate the fitness variance, the landscape is multipeaked and rugged. The properties of smooth versus rugged landscapes [parametrized in the manner of Eq. (1)] has been a subject of extensive study in statistical physics as it relates to the theory of spin glasses (Mezard, Parisi, and Virasoro, 1987). It is known that the key consequences of complexity of the general landscape appear already in the class of functions involving only pairwise interactions [also known as the Sherrington-Kirkpatrick model (Sherrington and Kirkpatrick, 1975)] (Mezard, Parisi, and Virasoro, 1987). Here for simplicity we consider only pairwise interactions. [An alternative instructive simplification would be to consider $F(g)$ defined by a fixed random function on the hypercube, known in population genetics as the "house-of-cards" model (Kingman, 1978) or NK models (Kauffman and Weinberger, 1989) and in physics as a "random energy" model (Derrida, 1981).]

Before moving on to population dynamics, it is instructive to discuss the implication of the combinatorial explosion of higher order interactions: In principle there are $\binom{L}{k}$ interactions of order k, a number which increases with L as L^k. Hence increasing the number of loci without changing the statistics of the coefficients would shift the power spectrum toward higher order, making the function more rugged. It seems more likely that the interactions are sparse with the number of "partners" of a typical locus not growing in proportion to the total number of loci: In particular, one may posit that each locus interacts with a finite number of other loci, independent of L and set all other coefficients to zero. Unfortunately, despite some recent progress (Brem and Kruglyak, 2005; Ehrenreich *et al.*, 2010), we still know little about the generic structure of genotype-phenotype maps. One must also be aware of the fact that, because of selection, statistics of genetic interactions observed among cosegregating polymorphisms within a breeding population may be quite different from that for a random set of loci or for polymorphisms created by crossing two isolated populations (Jinks *et al.*, 1966). Indeed, most immediate evidence for epistasis is provided by the "outcrossing depression": suppression in the fitness of progeny issuing from a cross of diverged strains (Jinks *et al.*, 1966; Seidel, Rockman, and Kruglyak, 2008).

III. DYNAMICS OF THE GENOTYPE DISTRIBUTION

Selection, mutation, and recombination operate on individuals and change the distribution of genotypes $P(g, t)$ in the population. The fitness $F(g)$ of a genotype g is defined as the expected reproductive success, i.e., the rate at which the proportion of a genotype increases or shrinks in the population due to (natural or artificial) selection. During the time interval Δt, *selection* changes the distribution of genotypes according to

$$P(g, t + \Delta t) = \frac{e^{\Delta t F(g)}}{\langle e^{\Delta t F} \rangle} P(g, t), \qquad (4)$$

where $\langle e^{\Delta t F} \rangle = \sum_g e^{\Delta t F(g)} P(g, t)$ denotes the population average. The genetic diversity that selection acts upon is due to *mutations*, which change the genotype distribution as follows:

$$P(g, t + \Delta t) = P(g, t) + \Delta t \mu \sum_{i=1}^{L} [P(M_i g, t) - P(g, t)]. \quad (5)$$

Here $M_i g$ is shorthand for genotype g with s_i replaced by $-s_i$. Despite the importance of mutations for generating polymorphisms and maintaining genetic diversity in the long run, the effect of mutation on the dynamics of significantly polymorphic sites can be neglected if mutation rates are much smaller than selection coefficients.

In addition to selection and mutation, the dynamics of the genotype distribution in sexual populations are driven by *mating* and *recombination*. Gametes are formed during meiosis crossing over homologous parental chromosomes. Assuming random pairing of gametes and outcrossing with rate r, the genotype distribution changes during recombination as follows:

$$P(g, t + \Delta t) = (1 - \Delta t r) P(g, t)$$
$$+ \Delta t r \sum_{\{\xi_i\}\{s_i'\}} C(\{\xi\}) P(g^{(m)}, t) P(g^{(f)}, t). \qquad (6)$$

The first term accounts for those individuals that did not outcross during the Δt time interval. In the event of outcrossing, a new genotype is formed from genetic material of the mother with genotype $g^{(m)}$ and a father with genotype $g^{(f)}$. The novel recombinant genotype g inherits a subset of his loci from the mother and the complement from the father, which in Eq. (6) is described by the set of random variables $\{\xi_i\}$. If $\xi_i = 1$, gene i is inherited from the mother, if $\xi_i = 0$ from the father. Using this notation, the maternal genotype is $s_i^{(m)} = \xi_i s_i + (1 - \xi_i) s_i'$ and equivalently the paternal genotype is $s_i^{(f)} = (1 - \xi_i) s_i + \xi_i s_i'$. The part of the maternal and paternal genome which is not passed on to the offspring $\{s_i'\}$ is summed over. Each particular realization of $\{\xi_i\}$, i.e., a pattern of crossovers, has probability $C(\{\xi\})$, which depends on the crossover rates between different loci. In addition to the summation over all $\{s_i'\}$, we have to sum over possible crossover patterns $\{\xi_i\}$. A similar notation was used by Christiansen (1990). While our presentation so far is completely general, dealing with diploid genomes inflates the required bookkeeping as we proceed with the analysis. Since our goal is to present the key effects and ideas in the simplest possible form, we from here on restrict to considering only haploids, two of which recombine upon mating producing a haploid offspring. Although this model is chosen for simplicity, it is sufficient to describe diploids in the absence of dominance. It also describes haploid yeast going through a mating-sporulation-germination cycle or the population genetics of many RNA viruses such as HIV and influenza.

Provided selection is weak [$\Delta t F(g) \ll 1$], we can use a continuous time description of the dynamics,

$$\frac{d}{dt} P(g, t) = [F(g) - \langle F \rangle] P(g, t) + \mu \sum_{i=1}^{L} [P(M_i g) - P(g)]$$
$$+ r \sum_{\{\xi_i\}\{s_i'\}} C(\{\xi\}) [P(g^{(m)}, t) P(g^{(f)}, t)$$
$$- P(g', t) P(g, t)]. \qquad (7)$$

This equation describes the dynamics of the genotype distribution in the limit $N \to \infty$, where each genotype is sampled

TABLE I. Population genetic quantities in our notation.

Quantity	Our notation
Allele at locus i, $\{a_i, A_i\}$	$s_i \in \{-1, 1\}$
Allele frequency ν_i	$\nu_i = (1 + \chi_i)/2$, where $\chi_i = \langle s_i \rangle$
Linkage disequilibrium D_{ij} ($i \neq j$)	$4D_{ij} = \chi_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle$
Heterozygocity $H_i = 2\nu_i(1 - \nu_i)$	$2H_i = \chi_{ii} = 1 - \chi_i^2$

by enough individuals to neglect sampling noise which arises during reproduction. This stochastic component to the dynamics of the genotype distribution is known as "random genetic drift." We discuss random drift in Sec. VI. Our focus here will be on the interplay between selection and recombination, which dominates the behavior of Eq. (7).

Instead of specifying $P(g, t)$ for every g, $P(g, t)$ can be parametrized by its cumulants. The cumulants of first and second order are defined as $\chi_i = \langle s_i \rangle$ and $\chi_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle$, which are related to allele frequencies and pairwise linkage disequilibria (LD) (see Table I). In total there are $2^L - 1$ cumulants, with higher order ones $\chi_{ij\cdots k}$ more easily defined via the cumulant generating function (McQuarrie, 1973). However, only the first and second order cumulants will be needed in the present context.

To obtain dynamical equations for χ_i, we multiply Eq. (7) by s_i and sum over all possible genotypes. One finds

$$\dot{\chi}_i = \langle s_i[F(g) - \langle F \rangle] \rangle - 2\mu \langle s_i \rangle, \tag{8}$$

where we used $M_i s_i = -s_i$ and the notation $\dot{\chi}_i$ for total derivative with respect to time. The dynamics of χ_i do not depend explicitly on the recombination rate, which is intuitive since recombination does not create or destroy alleles. In order to evaluate $\langle s_i F(g) \rangle$ in Eq. (8) we need to know higher order cumulants, i.e., we are faced with a hierarchy of cumulant equations.

In contrast to first order cumulants, the dynamics of higher order cumulants depend explicitly on recombination, which has the tendency to destroy associations between alleles and drives higher order cumulants to zero. To write down an equation for the dynamics of the second order cumulants, $\chi_{ij} = \langle s_i s_j \rangle - \chi_i \chi_j$, we have to evaluate $d\langle s_i s_j \rangle / dt$, which explicitly depends on recombination. Evaluating the recombination term only, we find

$$r \sum_{\{\xi_i\}} C(\{\xi\}) \sum_{g, g'} s_i s_j [P(g^{(m)}, t) P(g^{(f)}, t)$$
$$- P(g', t) P(g, t)] = -rc_{ij} \chi_{ij}, \tag{9}$$

where c_{ij} is the probability that loci i and j derive from different parents: $c_{ij} = \sum_{\{\xi\}} C(\{\xi\})[\xi_i(1 - \xi_j) + (1 - \xi_i)\xi_j]$. To arrive at this result, we substituted $s_i = \xi_i s_i^{(m)} + (1 - \xi_i) s_i^{(f)}$ (analogously for s_j), and averaged over the maternal and paternal genomes. The second term evaluates simply to $r\langle s_i s_j \rangle$. This result holds more generally for central moments of the genotype distribution (Barton and Turelli, 1991). Together with selection and mutation, we find (for $i \neq j$)

$$\dot{\chi}_{ij} = \langle (s_i - \chi_i)(s_j - \chi_j)[F(g) - \langle F \rangle] \rangle$$
$$- 4\mu \chi_{ij} - rc_{ij} \chi_{ij}. \tag{10}$$

We see that selection drives χ_{ij} away from zero, while χ_{ij} relaxes through mutation and recombination. In the absence of selection $P(g, t)$ tends to a steady state of LE with vanishing cumulants χ_{ij} (for $i \neq j$) implying complete decorrelation of alleles at different loci corresponding to factorization of the genotype distribution: $P_0(g) = \prod_{i=1}^{L} p_i(s_i)$. It is easy to see that the recombination term in Eq. (7) vanishes whenever $P(g) = P_0(g)$. In Sec. V, we, starting from $P_0(g)$, develop the QLE approximation by systematically accounting for small linkage disequilibria (χ_{ij}).

IV. TRAIT DISTRIBUTIONS AND THE DYNAMICS OF POPULATION AVERAGES

In most cases, $P(g, t)$ cannot be observed directly. Instead, the subject of quantitative genetics are distributions of traits in the population. Trait distributions can be obtained from genotype distributions by projection. The probability of finding in the population an individual with fitness (or any other trait) in the interval $[F, F + \Delta F]$ defines the density

$$p(F, t) = \sum_g \delta(F - F(g)) P(g, t), \tag{11}$$

where $\delta(F)$ is the Dirac delta function $[\int dF \delta(F) = 1]$. Applying this projection to Eq. (7) yields an equation for the dynamics of the trait distribution. Before addressing the dynamics of traits in sexual populations, it is instructive to consider the dynamics of the fitness distribution $p(F, t)$ in the absence of mutation and recombination, in which case one obtains simply

$$\frac{d}{dt} p(F, t) = [F - \langle F \rangle] p(F, t), \tag{12}$$

where $\langle F \rangle = \int dF F p(F, t)$. Multiplying this equation by F and integrating over F (i.e., the first moment of this equation) yields Fisher's fundamental theorem in the asexual case,

$$\frac{d}{dt} \langle F \rangle = \langle [F - \langle F \rangle]^2 \rangle = \sigma^2. \tag{13}$$

Evidently this is just the first in the hierarchy of infinitely many moment equations that characterize the dynamics of $p(F, t)$ given explicitly by Eq. (12). The second moment expresses the dynamics of σ^2 in terms of the third moment, etc. This hierarchy of equations is not closed, yet under certain conditions higher moments may be suppressed, making the σ^2 a slowly varying function of time. One notes that Eq. (12) has a Gaussian traveling wave solution $p(F, t) = C \exp[-(F - \nu t)^2 / 2\nu]$ with an arbitrary constant variance $\sigma^2 = \nu$ setting the rate of fitness growth $d\langle F \rangle / dt = \nu$ in agreement with Eq. (13). A traveling wave with constant speed requires that genotypes with arbitrarily high fitness are populated with at least one individual, which requires an infinitely large population with infinitely many polymorphic loci with limits taken in this order. Otherwise genetic diversity disappears and adaptation stalls. The evolution of the shape of the fitness distribution in finite populations has been studied in the context of genetic algorithms by

Prügel-Bennett and Shapiro (1994). Prügel-Bennett and Shapiro studied the effect of selection and recombination on the cumulants of the fitness distribution and observed how the fitness variation vanishes as the population condenses into a local fitness maximum. To prevent this condensation, new variation has to be constantly supplied by mutation. Quite generally σ is determined by the balance generation of genetic variation through mutations or recombination and its removal by selection and drift, which requires careful stochastic treatment (Tsimring, Levine, and Kessler, 1996; Rouzine, Wakeley, and Coffin, 2003; Rouzine and Coffin, 2005; Desai and Fisher, 2007; Neher, Shraiman, and Fisher, 2010; Hallatschek, 2011).

One can also consider the dynamics of an arbitrary trait $G(g)$ different from fitness. In analogy to Eq. (11), we can study the joint distribution $p(F, G, t)$ of this trait with fitness. The population average of the trait obeys

$$\frac{d}{dt}\langle G\rangle = \langle GF\rangle - \langle G\rangle\langle F\rangle = \mathrm{Cov}(F, G)_{P(g)}, \qquad (14)$$

i.e., its rate of change is given by its covariance with fitness (Price, 1970). This statement is also known as the "secondary theorem" of natural selection (Robertson, 1966).

With mutation and recombination, the dynamics of trait means are no longer that simple. To evaluate the mutation and recombination terms, we utilize the orthogonal expansion of the fitness function in Eq. (1). Restricting ourselves to pairwise interactions, we use Eqs. (8) and (9) to obtain

$$\frac{d}{dt}\langle F\rangle = \sigma^2 - \mu\Delta_\mu - r\sum_{i<j} c_{ij}f_{ij}\chi_{ij}, \qquad (15)$$

where Δ_μ is the average loss in fitness due to mutation. The latter can be calculated by observing that each moment decays through mutation with rate $2\mu k$, where k is the order of the moment,

$$\Delta_\mu = \mu\left[2\sum_i f_i\chi_i + 4\sum_{i<j} f_{ij}(\chi_i\chi_j + \chi_{ij}) + \cdots\right]. \qquad (16)$$

Higher moments decay faster because they have a greater mutation target. The second term in Eq. (15) is the loss in fitness through recombination, which reflects the tendency of recombination to factorize the genotype distribution such that contributions similar to $f_{ij}\chi_{ij}$ to $\langle F\rangle$ decay with rate rc_{ij}. Later we will see that the previous form of Fisher's theorem can be recovered by a suitable definition of an *additive* fitness variance. To do so, however, we have to understand how the genotype distribution evolves under selection and recombination.

V. BEYOND LINKAGE EQUILIBRIUM: QUASILINKAGE EQUILIBRIUM

We have already seen that without selection or epistasis, $P(g)$ will relax to a product of independent distributions at different loci: the linkage equilibrium state. Next we account for the correlations between loci induced by selection. For simplicity we omit the mutational contribution, which we will restore once we understand the basis of QLE.

A. QLE: A perturbation expansion at high recombination rates

If selection on the time scale of recombination is weak, i.e., $\bar{\sigma} \ll r$, the induced correlation is also weak and can be calculated using perturbation theory (Kimura, 1965; Barton and Turelli, 1991). To this end, we parametrize the genotype distribution as follows:

$$\log P(g, t) = \Phi(t) + \sum_i \phi_i(t)s_i + \sum_{i<j} \phi_{ij}(t)s_is_j, \qquad (17)$$

which is the already familiar Fourier representation of functions on the genotype space. The factorized distribution $P_0(g)$ corresponds to the coefficients ϕ_{ij} of all multilocus contributions being zero. The second order terms capture (to leading order) the correlations induced by selection and (in the limit under consideration) are assumed to be small. The genotype independent term $\Phi(t)$ is fixed by the normalization of the probability distribution,

$$e^{-\Phi(\{\phi\})} = \sum_g \exp\left[\sum_i \phi_is_i + \sum_{i<j} \phi_{ij}s_is_j + \cdots\right] \qquad (18)$$

and acts as the generator of the cumulants via

$$\chi_i = -\frac{\partial\Phi}{\partial\phi_i}, \qquad \chi_{ij} = -\frac{\partial^2\Phi}{\partial\phi_i\partial\phi_j}. \qquad (19)$$

The generating function Φ is evaluated perturbatively for small ϕ_{ij} in Appendix B yielding

$$\chi_i \approx \tanh(\phi_i) + \sum_{j\neq i} \phi_{ij}[1 - \tanh^2(\phi_i)]\tanh(\phi_j), \qquad (20)$$

$$\chi_{ij} \approx (1 - \chi_i^2)(1 - \chi_j^2)\phi_{ij} \quad \text{for } i \neq j, \qquad (21)$$

$$\chi_{ii} = 1 - \chi_i^2, \qquad (22)$$

which is correct to the leading order in $|\phi_{ij}|$. The distribution given by Eq. (17) may be thought of as a maximum entropy distribution constrained to have certain first and second order cumulants: Parameters ϕ_i and ϕ_{ij} are the Lagrange multipliers that impose the constraints.

We rewrite Eq. (7) as an equation for the dynamics of $\log P(g)$ which yields

$$\dot{\Phi} + \sum_i \dot{\phi}_is_i + \sum_{i<j} \dot{\phi}_{ij}s_is_j$$

$$= F(g) - \langle F\rangle + r\sum_{\{\xi_i\}\{s_i'\}} C(\{\xi\})P(g')\left[\frac{P(g^{(m)})P(g^{(f)})}{P(g)P(g')} - 1\right]$$

$$\approx \bar{F} - \langle F\rangle + \sum_i f_is_i + \sum_{i<j} f_{ij}s_is_j$$

$$+ r\sum_{i<j} c_{ij}\phi_{ij}[(s_i\langle s_j\rangle + \langle s_i\rangle s_j) - (s_is_j + \langle s_is_j\rangle)], \qquad (23)$$

where the recombination part has been evaluated approximately by expanding the exponential that defines $P(g)$ (see Appendix B). We now collect terms with the same monomials in s_i to obtain the equations governing the time evolution of ϕ_i and ϕ_{ij}:

$$\dot{\phi}_i = f_i + r\sum_{j\neq i} c_{ij}\phi_{ij}\langle s_j\rangle, \qquad (24)$$

$$\dot{\phi}_{ij} = f_{ij} - rc_{ij}\phi_{ij}. \tag{25}$$

At large crossover rates rc_{ij}, the ϕ_{ij} rapidly approach a steady state $\phi_{ij} = f_{ij}/rc_{ij}$. This has to be contrasted with the behavior in the absence of recombination, in which case ϕ_{ij} grows linearly as $f_{ij}t$. Recombination prevents effective selection on interactions. Instead, the higher order contributions to fitness affect the dynamics of ϕ_i after averaging over possible genetic backgrounds: Substituting the steady-state relation into the equation for ϕ_i yields

$$\dot{\phi}_i = f_i + \sum_{j\neq i} f_{ij}\chi_j = \hat{f}_i, \tag{26}$$

where we defined $\hat{f}_i = f_i + \sum_j f_{ij}\chi_j$, which is the effective strength of selection acting on locus i in linkage equilibrium. It is obtained from the general expression for $F(g)$ in Eq. (1) by replacing $s_j \rightarrow \chi_j$ and differentiating with respect to χ_i (and is truncated here at second order because we assumed, for simplicity, that genetic interactions are limited to that order).

Converting ϕ's to χ's using Eq. (20), we find $\dot{\chi}_i = (1 - \chi_i^2)\dot{\phi}_i$, correct to the leading order. For the discussion below, it is useful to derive equations for χ_i and χ_{ij} also to the subleading order

$$\dot{\chi}_i = \sum_j \chi_{ij}[\hat{f}_j - \chi_i f_{ij}] + \bar{\sigma}\mathcal{O}(\bar{\sigma}^2/r^2),$$

$$\chi_{ij} = \frac{(1 - \chi_i^2)(1 - \chi_j^2)f_{ij}}{2\hat{f}_i\chi_i + 2\hat{f}_j\chi_j + rc_{ij}} + \mathcal{O}(\bar{\sigma}^2/r^2) \text{ for } i \neq j. \tag{27}$$

In QLE, correlations χ_{ij} between loci ($i \neq j$) are determined by the balance between epistatic selection and recombination. (Note, in contrast, the diagonal elements $\chi_{ii} = \langle s_i^2 \rangle - \langle s_i \rangle^2 = 1 - \chi_i^2$ are determined by the allele frequencies.)

Wright (1931) showed that in linkage equilibrium the dynamics of allele frequencies are driven by the gradient in mean fitness. The result can be generalized to include correlations between loci arising in QLE. Starting with the exact equation for the allele frequency dynamics and using our parametrization of $P(g)$ via the "fields" ϕ_i given in Eq. (17), we find

$$\dot{\chi}_i = \langle s_i F \rangle - \chi_i \langle F \rangle = \partial_{\phi_i} \langle F \rangle \approx \sum_j \partial_{\phi_i} \chi_j \partial_{\chi_j} \langle F \rangle$$

$$= \sum_j \chi_{ij} \partial_{\chi_j} \langle F \rangle, \tag{28}$$

where we used the chain rule of differentiation and the fact that $\partial \chi_j / \partial \phi_i = \chi_{ij}$ following directly from Eq. (19). The correlation matrix χ_{ij} acts as a mobility matrix for allele frequencies. The nondiagonal entries of order $\bar{\sigma}/r$ imply that selection on locus j, via the correlation with locus i, affects the rate of change of χ_i. Equation (28) describes the dynamics of allele frequencies as the population ascends Wright's "adaptive landscape." While allele frequencies still evolve to maximize $\langle F \rangle$, their dynamics now are coupled by correlations captured in the off-diagonal terms of χ_{ij}.

The key point emerging from the analysis of the weak selection and rapid recombination limit is the remarkable simplicity of multilocus dynamics: The 2^L ordinary differential

equations for all cumulants or equivalently for all genotypes are reduced to L differential equations describing the dynamics of allele frequencies. Higher order cumulants are slaved to allele frequencies and can be obtained by solving algebraic equations defining the L dimensional quasi-linkage-equilibrium manifold. The distribution of genotypes in the population can therefore be parametrized by time-dependent allele frequencies, with all other features of the distribution constrained by the QLE equations. In mathematical terms, the dynamics of genotype distribution are approximately reducible to the dynamics on the "center manifold" formed by the set of allele frequencies (Guckenheimer and Holmes, 1997). Within the QLE approximation, population averages of any trait $G(g)$ can be parametrized by $\{\chi_1(t), \ldots, \chi_L(t)\}$ and the time derivative of the trait mean is therefore given by

$$\frac{d}{dt}\langle G \rangle \approx \sum_i \partial_{\chi_i} \langle G \rangle \partial_t \chi_i(t)$$

$$= \sum_{ij} \chi_{ij} \partial_{\chi_i} \langle G \rangle \partial_{\chi_j} \langle F \rangle - 2\mu \sum_i \chi_i \partial_{\chi_i} \langle G \rangle, \tag{29}$$

where we have restored the contribution of mutations through its effect on allele frequencies as it appeared in Eq. (8). This result has a simple interpretation: The rate of change of the trait mean is the product of the rate of change of allele frequencies through selection and the susceptibility of the trait mean to the allele frequency. The second term accounts for the effect of mutation on the trait mean. Since the first term is the additive covariance between fitness and the trait G, this equation is the analog of Eq. (14) in a recombining population. The QLE approximation breaks down when recombination is not sufficiently rapid to confine the genotype distribution to the L dimensional manifold defined by quasisteady correlations between loci. This breakdown will be discussed in more detail below.

B. Additive genetic variance and Fisher's theorem in QLE

Fisher's theorem in sexual populations posits that the rate of mean fitness increase is equal to the additive variance. We now discuss how Fisher's theorem emerges from Eq. (15) and how it compares with Eq. (29) which obviously can be used to calculate $d\langle F \rangle/dt$. Additive variance is typically defined as the variance captured by a linear model of the form

$$F_A(g) = a_0 + \sum_i a_i s_i, \tag{30}$$

where the coefficients are determined by minimizing

$$\sigma_I^2 = \sum_g [F_A(g) - F(g)]^2 P(g, t). \tag{31}$$

The remaining variance σ_I^2 is commonly called *epistatic* or *interaction* variance. Minimization yields $a_0 = \langle F \rangle - \sum_i a_i \chi_i$ with a_i determined by the linear equation

$$\sum_j \chi_{ij} a_j = \langle s_i F \rangle - \chi_i \langle F \rangle = \partial_{\phi_i} \langle F \rangle. \tag{32}$$

We have seen the right-hand side (rhs) of this equation already in Eq. (28): It is the contribution of selection to $\dot{\chi}_i$. In the high recombination limit, $\partial_{\phi_i} \langle F \rangle \approx \sum_j \chi_{ij} \partial_{\chi_j} \langle F \rangle$. Hence

Richard A. Neher and Boris I. Shraiman: Statistical genetics and evolution of …

the additive fitness coefficients (defined by linear regression) are $a_i = \partial_{\chi_i}\langle F \rangle$, which is accurate to order $\bar{\sigma}/r$. The additive variance therefore is

$$\sigma_A^2 = \sum_{ij} a_i \chi_{ij} a_j \approx \sum_{ij} \chi_{ij} \partial_{\chi_i}\langle F \rangle \partial_{\chi_j}\langle F \rangle + \bar{\sigma}^2 \mathcal{O}(\bar{\sigma}^2/r^2). \tag{33}$$

Recalling the QLE equation for mean trait dynamics, Eq. (29), and using fitness as a trait, we have

$$\frac{d}{dt}\langle F \rangle \approx \sum_{ij} \chi_{ij} \partial_{\chi_i}\langle F \rangle \partial_{\chi_j}\langle F \rangle - 2\mu \sum_i \chi_i \partial_{\chi_i}\langle F \rangle, \tag{34}$$

and comparing to the definition of σ_A^2 we arrive at the generalized Fisher's fundamental theorem

$$\frac{d}{dt}\langle F \rangle = \sigma_A^2 - \mu \Delta_\mu + \mathcal{O}(\bar{\sigma}^4/r^2), \tag{35}$$

which limits growth of fitness to the additive variance. Comparing to the general expression for mean fitness given before in Eq. (15) we see that the loss in fitness due to disruption of favorable combinations of alleles through recombination exactly cancels the epistatic $\sigma_I^2 = \sigma^2 - \sigma_A^2$ part of total variance. In other words, in a sexually reproducing species the uncertainty in the phenotype of the offspring in relation to that of its parents limits the effect of selection to the additive component of variance. The latter is that genetic component of the trait that "survives" reshuffling of genes by reassortment and recombination which depends on the genetic distance to the mate. Hence, this decomposition of genetic variation in additive and nonadditive components is explicitly population dependent.

One must, of course, remember that the generalized Fisher's law as stated only holds in this rapid recombination and weak selection limit and only after correlations have relaxed to their steady QLE values. During the initial transient toward QLE or at low recombination rates the mean fitness can exhibit very different dynamics. The meaning of Fisher's theorem has been subject to extensive discussion in the literature (Feldman and Crow, 1970; Price, 1972; Ewens, 1989; Frank and Slatkin, 1992; Edwards, 1994) caused by Fisher's insistence that his statement was exact. Price (1972), in particular, suggested that Fisher's intention was to describe not the total rate of change of mean fitness, but only the "partial rate" due to a change in allele frequencies: i.e., just the first term on the rhs of Eq. (29). The "theorem" would in that case become an exact statement, but not a very useful one. Following Kimura (1958) and Nagylaki (1993) our Eq. (35) sticks to $d\langle F \rangle/dt$ so that the generalized Fisher's theorem is an unambiguous, but approximate statement. The above analysis assumed that the population is subject to a constant fitness function and the mean fitness provides a useful measure of adaptation. If the fitness function itself depends on time, the increase in mean fitness due to adaptation of the population is superimposed with the dynamics of the fitness function. In the latter case, an unambiguous measure of adaptation, the fitness flux, can be defined in analogy to fluctuation theorems of nonequilibrium statistical mechanics (Mustonen and Lässig, 2010).

The off-diagonal terms in the additive variance $a_i \chi_{ij} a_j$ have interesting implications for the evolution of recombination: If two alleles that are selected with the same sign ($a_i a_j > 0$) are anticorrelated ($\chi_{ij} < 0$), the rate of adaptation is smaller than it would be in linkage equilibrium. This is the basis for the often made statement that recombination accelerates adaptation by reducing negative linkage disequilibria and thereby increasing the additive variance (Barton and Otto, 2005). There is, however, an additional effect of recombination on adaptation that is not captured by deterministic multilocus dynamics and is likely to be more important: Recombination greatly increases the likelihood that a novel beneficial mutation establishes and ultimately fixates in the population (Fisher, 1930; Muller, 1932; Barton, 1995b; Neher, Shraiman, and Fisher, 2010). Thereby the number of simultaneously polymorphic loci is increased, which in turn increases the fitness variance and speeds up adaptation. The reason for this is again that recombination breaks down negative linkage disequilibria (a tendency of beneficial alleles to be anticorrelated), which are generated by chance and amplified by selection (Barton and Otto, 2005). Analysis of this phenomenon requires going beyond QLE (see below).

VI. FINITE POPULATION DRIFT AND WRIGHT'S MUTATION, SELECTION, AND DRIFT EQUILIBRIUM

So far our formulation of the genotype dynamics Eq. (7) and the dynamics of allele frequencies Eq. (27) was deterministic, i.e., we neglected random drift. Random drift is a consequence of the stochastic nature of birth and death in a finite population of size N. In the simplest models of stochastic population genetics, called Fisher-Wright models, stochasticity is introduced by resampling the population from a multinomial distribution parametrized with the current genotype (or gamete) frequencies of each generation.

We have seen that the genotype frequency distribution can be parametrized by allele frequencies when recombination is rapid, and we now discuss how resampling of genotypes leads to stochastic contributions to the dynamics of allele frequencies and cumulants. For alleles that are present in large numbers, the relative sizes of fluctuations due to resampling are small and random drift can be accurately described by a diffusion approximation (Kimura, 1964). To derive a diffusion equation for allele frequencies, we generalize the ordinary differential equations (27) to stochastic differential equations [Langevin equations (Gardiner, 2004)]. For a finite time step Δt, one has

$$\chi_i(t + \Delta t) = \chi_i(t) + \Delta t \left[\sum_j \chi_{ij} \partial_{\chi_j}\langle F \rangle - 2\mu \chi_i \right] + \sqrt{\Delta t}\, \zeta_i(t), \tag{36}$$

$$\chi_{ij}(t + \Delta t) = \chi_{ij}(t) + \Delta t[(1 - \chi_i^2)(1 - \chi_j^2)f_{ij} - rc_{ij}]\chi_{ij} + \sqrt{\Delta t}\, \zeta_{ij}(t), \tag{37}$$

where we neglected terms much smaller than rc_{ij} in the relaxation rate of χ_{ij}. $\zeta_i(t)$ and $\zeta_{ij}(t)$ are white noise terms with zero mean and a covariance matrix determined by the multinomial sampling of the genotypes. One finds

Rev. Mod. Phys., Vol. 83, No. 4, October–December 2011

$$\langle \zeta_i(t)\zeta_j(t') \rangle = \frac{\chi_{ij}}{N} \delta(t - t'), \tag{38}$$

$$\langle \zeta_{ij}(t)\zeta_{ij}(t') \rangle \approx \frac{(1 - \chi_i^2)(1 - \chi_j^2)}{N} \delta(t - t'), \tag{39}$$

while other covariances are of order $\bar{\sigma}/Nr$ or smaller; see Appendix C. The leading order presented in Eq. (39) will be sufficient for the analysis below. The joint stochastic dynamics of allele frequencies and the correlation between loci has been studied by Ohta and Kimura (1969) using a two-locus model. Here we study a multilocus model making the simplifying assumption that the recombination is faster than all other processes.

In this case, the second order cumulants relax much faster than allele frequencies change and we can solve the equation for χ_{ij} assuming fixed χ_i. The solution can be decomposed into a deterministic component due to the competition between epistatic selection and recombination and a stochastic component:

$$\chi_{ij}(t) = \frac{f_{ij}(1 - \chi_i^2)(1 - \chi_j^2)}{rc_{ij}} + \delta\chi_{ij}. \tag{40}$$

The deterministic component is the familiar QLE value from Eq. (27), while the stochastic component $\delta\chi_{ij}$ has an autocorrelation

$$\langle \delta\chi_{ij}(t)\delta\chi_{ij}(t + \Delta t) \rangle = \frac{(1 - \chi_i^2)(1 - \chi_j^2)}{2Nrc_{ij}} e^{-rc_{ij}\Delta t};$$

see Appendix C. We now use this result to study the Langevin equation for χ_i. We have to distinguish the case where the deterministic component to χ_{ij} dominates over the stochastic term or vice versa. In order to compare the stochastic to the deterministic term, we have to average the former over the time scale of the dynamics of χ_i given by the inverse of $\partial\langle F\rangle/\partial\chi_i \approx \hat{f}_i$. Recalling that in equilibrium $\chi_i \approx 1 - \mu/\hat{f}_i$, we find that the deterministic contribution to χ_{ij} dominates if $N\mu \gg 1$ and $f_{ij} \gg \mu$. In the opposite limit, the stochastic contribution $\delta\chi_{ij}$ affects the dynamics of χ_i more strongly than the deterministic one. We now show how the equilibrium distribution of allele frequencies is affected by correlation between loci in these two cases.

A. Wright's equilibrium in the QLE approximation

Assuming we can neglect the stochastic contribution to χ_{ij}, the Langevin equation for the χ_i (interpreted in the Îto sense) corresponds to the following forward Kolmogorov equation for the dynamics of the probability distribution of allele frequencies denoted by $Q(\{\chi_i\}, t)$ (Gardiner, 2004):

$$\partial_t Q(\{\chi_i\}, t) = \sum_i \partial_{\chi_i} \left[\frac{1}{2N} \sum_j \partial_{\chi_j} (\chi_{ij} Q(\{\chi_i\}, t)) \right.$$
$$\left. + Q(\{\chi_i\}, t) \left(2\mu\chi_i - \sum_j \chi_{ij} \partial_{\chi_j} \langle F \rangle \right) \right]. \tag{41}$$

This multilocus version of the diffusion equation for allele frequencies in linkage equilibrium (no correlations) appears already in Kimura (1955). It has a steady solution where all probability flux vanishes, i.e., where the term in brackets is zero for each i. In complete linkage equilibrium, the matrix χ_{ij} is diagonal and different allele frequencies decouple. One obtains the equilibrium distribution

$$Q(\{\chi_i\}) = Ce^{2NF(\{\chi_i\})} \prod_i (1 - \chi_i^2)^{2N\mu-1}, \tag{42}$$

where $F(\{\chi_i\})$ is the mean fitness evaluated in linkage equilibrium obtained by replacing each s_i by its moment χ_i in Eq. (1). The term $e^{2NF(\{\chi_i\})}$ is analogous to the contribution of energy to a Gibbs measure, while $\prod_i (1 - \chi_i^2)^{2N\mu-1}$ plays the role of an entropy. Note that for $2N\mu < 1$, the distribution is singular at $|\chi_i| = 1$. In the opposite case $2N\mu > 1$, $Q(\{\chi_i\})$ vanishes if any of the $|\chi_i| = 1$. Instead $Q(\{\chi_i\})$ has a maximum in the interior of the hypercube defined by $|\chi_i| < 1$.

The corresponding solution for QLE, where χ_{ij} has small but steady off-diagonal entries, is derived in Appendix C with the result,

$$Q(\{\chi_i\}) = Ce^{2N\langle F\rangle + 4N\mu \sum_{i<j} f_{ij}\chi_i\chi_j/rc_{ij}} \prod_{i=1}^L (1 - \chi_i^2)^{2N\mu-1}. \tag{43}$$

The genotype distribution assumes this exponential (Boltzmann) form $\sim e^{N\langle F\rangle}$ since the mobility matrix χ_{ij} is proportional to the autocorrelation of the genetic drift. Equation (43) provides a systematic extension of Wright's equilibrium to QLE, which appears to be a new result.

B. Wright's equilibrium with stochastic linkage disequilibrium

In the absence of epistasis or in cases where selection is weak or comparable to the strength of genetic drift (diffusion constant), the deterministic expectation for χ_{ij} is small compared to its fluctuations. The coupling between different allele frequencies in Eq. (36) has, therefore, a fluctuating sign and acts as an additional noise source with autocorrelation time $(rc_{ij})^{-1}$ (for derivation see Appendix C.2). Such an increased noise level increases the diffusion constant in the Fokker-Planck equation for each of the χ_i by a factor

$$\frac{N}{N_e} = 1 + \frac{1}{2} \sum_{i \neq j} (1 - \chi_j^2) \left(\frac{1}{rc_{ij}} \frac{\partial\langle F\rangle}{\partial\chi_j} \right)^2. \tag{44}$$

This increase in diffusion constant is often phrased as a reduction in effective population size N_e and is known as a manifestation of the Hill-Robertson effect (Hill and Robertson, 1966). Note that the correction has the structure of the additive variance in fitness where each term is compared to the square of the recombination rate between the loci i and j. This result was derived in the context of fixation probabilities of novel mutations by Barton (1995b). It has been shown that this effective increase in the diffusion constant through stochastic correlations of loci can select for increased recombination rates (Barton and Otto, 2005).

C. Equilibration toward a steady state

The approach to the equilibrium distribution is governed by the smallest nonzero eigenvalue of Eq. (41). For $2N\mu > 1$ and smooth fitness landscapes, this relaxation rate is governed by the larger of μ and the scale of selection on individual loci.

The corresponding time scales can be very long. Furthermore, if different parts of sequence space are separated by fitness valleys ("energy barriers"), relaxation to the steady state can take exponentially long (Weissman, Feldman, and Fisher, 2010).

Similar equations for the distribution of allele frequencies apply in the context of spatially structured populations, in which case the role of mutation is played by migration of individuals. The latter problem was the subject of work by Wright (1932). Migration rates and the associated influx of foreign alleles are often much larger than mutation rates and rapid equilibration is plausible.

VII. BREAKDOWN OF QLE

QLE greatly simplifies the dynamics of the genotype distribution, but its perturbation theory nature leaves one with the question about the range of its validity. In particular, we know from statistical physics that Gibbs measures of the form of Eq. (17) can lead to a so-called glass transition where the structure of the distribution changes qualitatively. Below the glass transition, different realizations of the system have a nonvanishing probability to be (largely) identical, which is quantified by the overlap distribution (Parisi order parameter) (Mézard and Montanari, 2009). A related transition in which the population condenses into a small number of genotypes is driven by the competition between epistasis and recombination. It occurs already in the deterministic mean-field setting and is discussed in Sec. VII.A. QLE can also become unstable at low recombination rates even in the absence of epistasis because of the discreteness of contributions of individual loci in a finite genome. The instability in that case is driven by fluctuations due to finite population size and is discussed in Sec. VII.B.

A. Infinite N and L limit: Alleles versus genotypes

To gain some heuristic insight into the range of validity of the perturbation expansion in $\bar{\sigma}/r$, it is useful to study the following coarse-grained quantitative genetic version of QLE which yields an explicit criterion for the validity of QLE (Neher and Shraiman, 2009). Instead of following the entire genotype distribution, consider the joint distribution $P(A, E, t)$ of additive A and epistatic E contributions to fitness defined via $A = F_A(g)$ [cf. Eq. (30)] and $E = F - A$. Hence additive and epistatic contributions are defined with reference to the current distribution of genotypes. The joint distribution of A and E evolves according to

$$\partial_t P(A, E, t) = (A + E - \langle A \rangle - \langle E \rangle) P(A, E, t)$$
$$+ r \left(\rho(E) \int dE' P(A, E', t) - P(A, E, t) \right),$$

$$\text{(45)}$$

where $\langle A \rangle$ and $\langle E \rangle$ are the mean additive and epistatic fitness in the population. Here we have assumed that the epistatic fitness of novel recombinants is independent of its parents and given by a random sample from the density of possible epistatic fitness values $\rho(E)$ [the house-of-cards model (Kingman, 1978)]. We assume $\rho(E)$ to be a Gaussian with

the variance equal to σ_I^2—the epistatic component of fitness variance defined in Eq. (31). Additive fitness of recombinants is a random sample from the current distribution of additive fitness in the population, i.e., the marginal $\int dE' P(A, E', t)$. The model does not include finite population size effects and assumes that both A and E are from a continuous distribution. The latter implies that the number of loci L that contribute to fitness is very large, while the individual contributions of loci are small (cf. Sec. II). In this sense, it is a deterministic mean-field model.

Equation (45) has a factorized solution $P(A, E, t) = \theta(A, t) \omega(E)$ with

$$\theta(A, t) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-(A - \langle A \rangle)^2 / 2\sigma_A^2}, \quad \text{where} \quad \frac{d}{dt}\langle A \rangle = \sigma_A^2,$$

$$\text{(46)}$$

$$\omega(E) = \frac{r\rho(E)}{r + \langle E \rangle - E}, \quad \text{where} \quad \langle E \rangle = \int dEE\omega(E).$$

The mean epistatic fitness is determined by enforcing the normalization of $\omega(E)$, i.e., $\int dE\omega(E) = 1$. Note that this solution is a QLE solution: Fitness increases with a rate given by the additive variance, while the epistatic contribution to fitness is steady with a magnitude controlled by recombination. Unlike Eq. (27), Eq. (46) implies a condition on r and the density of states: $\rho(E)$ has to vanish for $E \geq r + \langle E \rangle$. Otherwise, ω is not normalizable. The density of states $\rho(E)$ is typically of Gaussian form, and given 2^L states has its maximum at $E_{\max} \approx \sigma_I^2 \sqrt{2L \log 2}$, however, for $N \ll 2^L$, as will be generically the case, $E_{\max} \approx \sigma_I^2 \sqrt{2 \log N}$. Hence QLE is expected to break down at $r_c \approx E_{\max} - \langle E \rangle \approx E_{\max}$.

The dynamics of the distribution $P(A, E)$ change dramatically as r falls below r_c. For $r > r_c$, no genotypes with $E \geq r + \langle E \rangle$ exist. Hence, all genotypes are destroyed by recombination and are short lived. At $r < r_c$, however, many genotypes with $E \geq r + \langle E \rangle$ exist which can outrun recombination and grow exponentially. The genotype distribution is no longer a product of additive and epistatic parts, but contains clones which are populated by many individuals. Selection now operates on the entire genotype over many generations and the relevant dynamical quantities are now clone sizes rather than allele frequencies, which are slaved to the performance of the clones. The alleles that make up the most successful genotype will fixate, not necessarily those with the most favorable additive effect. The transition line between the two regimes is sketched in Fig. 2 with the ratio of recombination to selection on the x axis and the "heritability," the ratio of the additive variance to the total variance $h^2 = \sigma_A^2 / (\sigma_A^2 + \sigma_I^2)$, on the y axis. [Note that heritability also measures the correlation between fitness of a recombinant offspring and parental mean (Lynch and Walsh, 1998).] At low recombination rates and strong epistatic interactions, selection operates on genotypes while at high recombination rates, or in absence of epistasis, selection operates on the additive effects of alleles. The distinction between genotype and allele selection regimes goes back to Franklin and Lewontin (1970) and Slatkin (1972), who showed that a related transition occurs in models with a strong heterozygote advantage. The regimes of allele and genotype selection are summarized in Fig. 2. In absence of epistatic interactions or heterozygote advantage, a similar condensation phenomenon
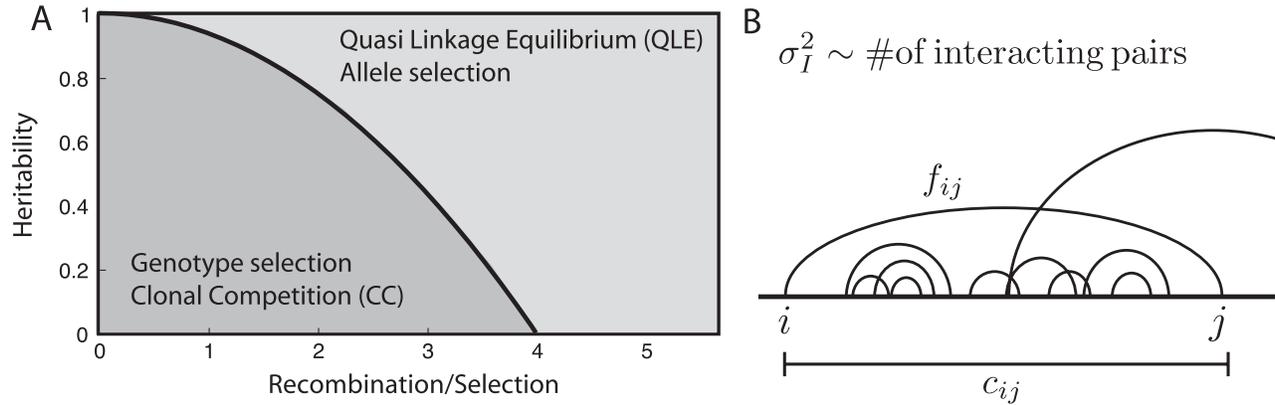
FIG. 2. Genetic interactions and the breakdown of QLE. (A) The range of validity of QLE as a function of $\bar{\sigma}/r$ and the heritability, i.e., the ratio of additive variance to the total fitness variance. Below the transition line, strong linkage equilibrium is expected and selection operates on genotypes rather than alleles. (B) An illustration of a possible "interaction graph" of polymorphisms on a block of chromosome. Loci interact with nearby loci, as well as with distant loci outside the block. The epistatic fitness variance solely within the block, i.e., averaged over the rest of chromosome, is proportional to the number of interaction terms (arcs in the figure).

occurs only at very low outcrossing rates $r = \mathcal{O}(N^{-1})$ (Rouzine and Coffin, 2005).

The condensation of genotypes goes along with a dramatic speedup of the allele frequency dynamics: In QLE (allele selection) each allele frequency is driven by an effective additive coefficient $a_i \sim \bar{\sigma}/\sqrt{L}$ (each a_i accounts for $\sim L^{-1}$ of the additive variance σ_A^2). When selection operates on genotypes, the time scale of selection is driven by fitness differences between individuals, which are of order $\bar{\sigma}$. The rate of change of allele frequencies is therefore greater by a factor $\sqrt{L}$ which could be a large effect (Neher and Shraiman, 2009). We emphasize that the stationarity of the distribution of the epistatic component of fitness $\omega(E)$ holds only on time scales short compared to that of the allele frequency dynamics.

The simple picture of the transition in a facultatively outcrossing species can also apply to blocks of chromosome in obligate sexuals. Consider a block that harbors l loci spread over a map distance c (on average c recombination events within the block per generation). If epistatic fitness within the block exceeds c, QLE will break down since individual haplotypes will be amplified by selection above less fit recombinants. Whether such local breakdown of QLE will occur depends on how fitness variance and recombination rate depend on the block size. The recombination rate is proportional to the block size, and, assuming constant density of polymorphism, will be proportional to the number l of polymorphic loci. Similarly, the additive variance is proportional to l and the rms is therefore $\sim\sqrt{l}$. The epistatic variance within the block scales with the number of interactions between loci within the block, as illustrated in Fig. 2(b). Any given locus will interact only with a fraction of all other loci, i.e., f_{ij} is sparse, and the number of interactions between loci within a block depends on whether these sparse interactions tend to be local or not. If any two loci are equally likely to interact, the number of interactions within the block is $\sim l^2$, so that rms epistatic fitness is $\sim l \sim c$. Hence the ratio of recombination within the block and the epistatic fitness are independent of the block size and QLE is either globally stable or unstable. A different conclusion is reached if

interactions are local and each locus interacts with k nearby other loci. As before additive fitness $\sim\sqrt{l}$, but the number of interactions within the block is $\sim lk$. Hence the typical epistatic fitness is $\sim\sqrt{lk}$, which decreases less fast than c as the block length is decreased. We therefore expect that QLE is unstable on scales below a critical block size l_c, where local epistasis overwhelms rare recombination. This local selection on coadapted haplotypes can coexist with the establishment of QLE on longer genomic scales (Neher and Shraiman, 2009). We return to the recombination and selection on different chromosomal scales in Sec. VIII.

B. Validity of QLE for finite N and L

The above discussion of the breakdown of QLE focused on the competition between genetic interactions driving and recombination destroying correlations in the limit where fluctuations are negligible and contributions of individual loci are small. We now discuss how the discrete contributions of individual loci and the number fluctuations in finite populations can drive populations off the QLE manifold. This problem has a long history in population genetics and was mainly discussed for scenarios without genetic interactions, i.e., on the line where the heritability equals 1 in Fig. 2(a). In this case the only source of correlations are the initial condition or fluctuations. Maynard Smith (1968) showed that without genetic interactions, and with no correlations in the initial condition, correlations do not develop in an infinite population at any recombination rate, in accordance with Fig. 2(a). However, novel mutations arise in single copies on random genomes, giving rise to correlations: QLE has to be stable with respect to these perturbations.

The hallmark of the QLE approximation are slowly changing allele frequencies and steady and perturbative correlations between loci. The latter will be true only if the correlations relax, i.e., are governed by an equation of the form $\dot{\chi}_{ij} = \beta - \alpha\chi_{ij}$ with $\alpha = 2(\hat{f}_i\chi_i + \hat{f}_j\chi_j) + rc_{ij} > 0$ (ignoring mutations). Hence the QLE state is unstable if $-2(\hat{f}_i\chi_i + \hat{f}_j\chi_j) > rc_{ij}$. In that case any small deviation from $\chi_{ij} = 0$,

which could be due to stochastic fluctuations, will grow. This effect has important implications for the evolution of recombination: Consider two closely linked loci at which beneficial mutations happen. Both novel mutations exists initially as a single copy ($\chi_i \approx -1$) and will most likely reside in different individuals, i.e., are anticorrelated or in negative linkage disequilibrium. Selection will now amplify the initial χ_{ij} if $2\hat{f}_i + 2\hat{f}_j > rc_{ij}$, generating predominantly negative linkage disequilibrium. This growth of correlations due to selection on individual loci slows down adaptation and can result in the loss of beneficial alleles. This phenomenon is known as the Hill-Robertson interference and it contributes to potential benefits of sexual reproduction (Hill and Robertson, 1966; Barton, 1995b; Barton and Otto, 2005). While this cumulant based approach to interference between sweeping loci is tractable for a few loci, it becomes intractable in populations in which many sweeping loci are tightly linked (Cohen, Kessler, and Levine, 2005; Rouzine and Coffin, 2005; Neher, Shraiman, and Fisher, 2010).

C. Cumulant analysis beyond QLE

Even though the QLE approximation breaks down when correlations are no longer slaved variables, the cumulant expansion can be useful to study the short term dynamics of systems with a small number of loci, in particular, if the initial conditions are such that higher order cumulants are small. Furthermore, if only a few isolated pairs of tightly linked loci are present, cumulants between these pairs can be treated as dynamical variables, while all other pairs for which the χ_{ij} are stable are treated in QLE. Such an analysis has, for example, been performed by Stephan, Song, and Langley (2006) to study linkage disequilibrium patterns between neutral markers following a selective sweep.

Explicit modeling of stochastic multilocus systems typically requires computer simulations, which are computationally expensive when the number of loci or the population size is large. However, making use of the fast-Fourier transformation (FFT) on the 2^L dimensional genotype space, one can speed up such simulation from a run time that scales as 8^L to 3^L. The FFT allows one to calculate and reuse the frequency of subsets of loci from which the distribution of recombinant genomes can be assembled. An efficient implementation of multilocus evolution for arbitrary fitness functions and genetic maps is available from the author's Web site. Cumulant equations to higher order involve "bookkeeping" of many terms and is best done with computer algebra systems. A package for MATHEMATICA has been developed by Kirkpatrick, Johnson, and Barton (2002). A implementation for MAPLE is available from the authors.

VIII. DISCUSSION

We have presented a review of the dynamics of multilocus genotype distributions and the resulting dynamics of quantitative traits. We focused, in particular, on how the distribution of genotypes can be parametrized by allele frequencies in the weak selection and fast recombination limit. This description extends beyond beanbag genetics allowing also for weak correlation (i.e., linkage disequilibrium) between loci. The

central element is the quasi-linkage-equilibrium approximation pioneered by Kimura. QLE emerges as a perturbation expansion in the weak selection and rapid recombination limit similar to high-temperature expansion in statistical physics. In a suitably defined system, the population genetics can be classified by the ratio of the strength of selection and the rate of recombination and the degree to which the fitness variation is additive or epistatic; see Fig. 2. At high recombination and additivity, QLE is an accurate approximation. This regime is separated from a regime at low recombination and strong epistasis, where QLE breaks down and the population condenses into a few fit genotypes.

Our exposition assumes a panmictic, random mating and haploid population. While the former are common assumptions, assuming haploidy in recombining population might raise objections. Our aim was to discuss the interplay between selection, genetic interactions, and recombination in multilocus systems. Dominance is a special kind of genetic interaction, where a locus interacts with itself, giving rise to additional nonlinearities. These nonlinearities can stabilize loci at intermediate allele frequencies, a process not possible in haploid populations. The effects of dominance, however, are well understood at the single locus level, as well as when many loci with a heterozygote advantage are close to each other (Franklin and Lewontin, 1970). Within QLE, the dynamics of allele frequencies in diploid populations is still relaxational and maximizes the mean diploid fitness. A full parametrization of the diploid populations and diploid fitness requires a straightforward, if somewhat tedious, generalization: To represent diploids one should (i) double the number of loci, (ii) define a genetic "transfer function" $C(\{\xi\})$ that represents meiotic crossover of the parental genomes, and (iii) extend the fitness function $F(g)$ to $2L$ hypercube to parametrize the 3^L states (homozygocity and heterozygocity at L loci). Another simplification of our exposition was the use of the continuous time description, in contrast to the more common discrete generation formulations of population genetics. Continuous time formulations assume that the population changes little in one generation. If this is the case, the results are completely equivalent and one can make use of calculus instead of recursions and difference equations.

The QLE approximation will often be appropriate for panmictic populations where genetic variation is replenished by de novo mutations. In this scenario, novel mutations establish if they blend in well with the genetic makeup of the population. This is manifest in the QLE equation (27) where alleles are selected on the basis of their additive effect, i.e., their effect on fitness marginalized over the distribution alleles at other loci in the population. The fitness of individual genotypes is not relevant to the evolutionary dynamics, since genotype frequencies are determined by allele frequencies (and sampling noise in finite populations). This issue was recently discussed by Livnat *et al.* (2008).

A very different evolutionary dynamics follows a hybridization event (Orr, 1995; Barton, 2001; Nolte and Tautz, 2010), i.e., a situation when two strains of one species that have been evolving in isolation for some time come in contact again. The two strains differ at many loci and these differences have never been tested for compatibility. Crossing two

such diverged strains can result in a phenotypically diverse population from which novel hybrid species can emerge (Nolte and Tautz, 2010). Such speciation after hybridization is similar to the clonal population structure observed in theoretical models of the selection dynamics after hybridization (Neher and Shraiman, 2009). In this regime of clonal competition, the allele frequencies are slaved to the dynamics of the clones and the average effect of an individual mutation affects the fate of a clone only very mildly. The lucky accident that produced through recombination a very fit genotype that contains the allele determines whether the allele can fixate or not. The possibility of a sharp transition between mixing and not mixing of two populations in a hybrid zone has been described by Barton (1983), who used a model of hybrid inferiority. In the limit of a large number of contributing loci, there exists a critical ratio of recombination rate and selection against hybrids, which separates the regimes of mixing and nonmixing. Similar transitions are expected if the reason for outbreeding depression is epistasis rather than dominance.

The qualitative differences between the genotype and allele selection regimes also sheds light on the importance of stochasticity (genetic drift). Allele frequencies are well sampled by $\mathcal{O}(N)$ copies, unless the allele is very young (or about to become extinct). Stochasticity therefore matters only during the establishment phase of the allele. As soon as the frequency exceeds $(N\hat{f}_i)^{-1} \approx \sqrt{L}/N\bar{\sigma}$, selection dominates. In the genotype selection phase, however, the founding of each genotype can, if it is exceptionally fit, change the fate of the population dramatically.

The transition to genotype selection driven by epistatic interaction is related to spin-glass transition in models for disordered physical systems and magnets. Within these models, the probability of finding the system in a particular state $\{s_i\}$ is given by

$$P(\{s_i\}) \sim e^{-\mathcal{H}(\{s_i\})/kT} = e^{-(1/kT)[\sum_i h_i s_i + \sum_{ij} J_{ij} s_i s_j + \cdots]}, \quad (47)$$

and hence completely analogous to Eq. (17). Such a system generically resides in one of three states: *paramagnetic*, *ferromagnetic*, or *glassy*. In the paramagnetic state at high temperature different parts of the system are uncorrelated, which is analogous to QLE. The perturbation expansion in $\bar{\sigma}/r$ is very similar to a high-temperature expansion in statistical physics. At low temperature, the behavior depends on the structure of the Hamiltonian $\mathcal{H}(\{s_i\})$. If most of the J_{ij} have the same sign, the system will go to an energetically favored ordered state where spins are aligned, giving rise to a ferromagnet. In this case $\mathcal{H}(\{s_i\})$ has one heavily preferred energy minimum, corresponding to a fit genotype.

A different low-temperature behavior is found when the J_{ij} have erratic sign. In that case, not all interactions can be in their favorable state simultaneously and the resulting landscape has many minima and maxima. At low temperature, the system condenses into one of the minima. This spin-glass phase is characterized by a nontrivial overlap distribution: Different realizations of the system, drawn from the ensemble defined by Eq. (47), will fall into clusters of different degrees of similarity (measured by the Hamming distance). The clusters themselves have subclusters, giving rise to a

hierarchical ultrametric structure (Mezard, Parisi, and Virasoro, 1987). This is in contrast to the high-temperature phase, where the overlap distribution is Gaussian. These qualitatively different overlap distributions above and below the spin-glass transition have a direct analogy to population structure and heterozygosity: In the high recombination limit, genotypes in the population are assembled from the available alleles more or less at random such that any two individuals differ at about $2\sum_{i=1}^L \nu_i(1 - \nu_i)$ sites (ν_i being the allele frequency at locus i). At low recombination (or substantial inbreeding), the population will condense into fit genotypes (or inbred groups) that are much more similar to each other than to members of the general population.

Figure 2(b) illustrates pairwise interaction between polymorphic loci along the chromosome. In general, we expect a complex and possibly hierarchical pattern of interactions: A given pair of distant genes will have only a low probability to interact substantially, while polymorphisms within one gene and its regulatory elements are much more likely to strongly interact. Nearby polymorphisms in a protein (Callahan *et al.*, 2011) will still be more likely to interact. In obligate sexuals, the sparse long range interactions rarely suffice to produce appreciable correlations between loci. Within small stretches of chromosomes, however, recombination rates are low and if the strength of interactions within this stretch is sufficiently high, QLE will locally break down. Consider, for example, a 1 centimorgan long region, which in humans corresponds to about 1 mega base and harbors around 1000 polymorphisms. If the typical epistatic contribution fitness of this stretch of chromosome were on the order of 1%, we expect a runaway selection on coadapted haplotypes and strong correlations. Since distant parts of the genome are in QLE, one expects a "module" selection regime, where loosely linked and weakly interacting modules are in QLE, but strong interactions and infrequent recombination has led locally to condensation into coadapted haplotypes (Neher and Shraiman, 2009). [An excellent early discussion of such epistasis driven "coagulation" in the "soup" of genes is found in Turner (1967).] Stated another way, we can view such a system as consisting of weakly interacting mesoscopic loci, at which several superalleles segregate. These superalleles are "destructible," in the sense that recombination within leads to reduced fitness and purging by selection. However, since recombination within these alleles is rare, quantitative traits would be highly heritable on short time scales and quantitative genetics would work as usual.

Our discussion of the multilocus theory and QLE was guided by ideas of statistical physics. The explicit form of the (approximate) genotype distribution function $P(g)$ parametrized by instantaneous allele frequencies is the central pillar connecting the dynamics of population average traits, the subject of QG, to the individual-based evolutionary process. It is essential that the QLE distribution is reached on a relatively fast time scale of mating and recombination. Allele frequencies are well defined and vary slowly on this time scale. The QLE ensemble should not be confused with a very different mutation, selection, or drift ensemble, which could be rightfully termed the "Wright equilibrium" [Eq. (42)], which is often invoked as a link between evolutionary dynamics and statistical physics. The Wright equilibrium gives

a stationary distribution of allele frequencies which would be established in a finite population (N playing the role of inverse temperature) on a time scale longer than the inverse mutation rate μ, provided stationary selection pressures. Ruggedness of the fitness landscape could further increase this equilibration time scale exponentially (Weissman, Feldman, and Fisher, 2010). Clearly, this type of equilibrium applies on a very different time scale than the phenomena addressed in the present work. Related ideas were developed in the context of quasispecies theory to study the conditions under which hereditary information can be maintained over long times (Eigen, 1971; Franz and Peliti, 1997). The focus of these studies was prebiotic evolution, where fidelity of replication was most likely low and stability of genomic information can be sensibly studied using a simple equilibrium model. Equilibrium arguments were also applied to the evolution of the codon bias (Iwasa, 1988) and the evolution of transcription factor binding sites (Mustonen and Lassig, 2005). In the latter two cases, an ensemble can be constructed by combining many instances of the same sequence motive which was under constant selection pressure for a very long time (conserved transcription factor binding motive or conserved preference of certain codons over others). In many cases, however, the equilibrium state is of little relevance.

In conclusion, in this review we provided a derivation of the genotype distribution in the QLE approximation, providing a systematic generalization of Fisher's theorem, Kimura's diffusion theory, and Wright's equilibrium from LE to QLE, which includes the effect of (weak) correlations between loci. We also discussed the limitation of the QLE approximation and the structure of the genotype distribution at low recombination rates.

It is our hope that better understanding of the QLE approximation will promote progress in understanding the effects associated with its breakdown, whether due to strong epistasis or strong physical linkage, such as, for example, the Hill-Roberson effects (hitchhiking and background selection) which still await comprehensive treatment.

GLOSSARY

Allele	State of a locus, for example, the base A, C, G, or T at a certain position
Crossover rate	In meiosis, parental chromosomes are paired up and crossed over. The density of crossovers on the chromosome is called the crossover rate.
Dominance	Interaction of the two alleles at the same locus in diploid organisms
Epistasis	Genetic interactions between alleles at different loci, i.e., a dependence of the effect of an allele at one locus on the remainder of the genome.
Fitness	Expected reproductive success of an organism. For modeling purposes, this is often equated with the growth rate (Malthusian or log fitness) or the average number of offspring in the subsequent generation (absolute fitness).
Gametes	Egg and sperm.
Genetic drift	Sampling fluctuations of genotype or allele frequencies. Genetic drift enters as the diffusion term in the Fokker-Planck equation for the dynamics of the distribution of allele frequencies.
Genetic map	The cumulative crossover rate along the chromosome. The average number of crossover events per chromosome is the map length.
Genotype	State of the genome, i.e., the set of alleles an individual carries.
Haplotype	Alleles inherited from one parent. In diploids, two haplotypes make one genotype.
Heritability	Broad sense heritability is the genetic component of traits, i.e., the concordance of traits between monozygotic twins. Narrow sense heritability refers to the genetic component of traits that is inherited in sexual reproduction, i.e., the correlation between trait values of parents and children.
Heterozygosity	Fraction individuals in a diploid population that carry distinct alleles at a locus.
Homozygosity	The complement of heterozygosity.
Linkage	Loci on the same chromosome are linked and share history until crossover events separate them.
Linkage (dis)equilibrium	Absence (presence) of correlations between loci, often abbreviated LE and LD.
Locus	Location on the chromosome, e.g., a gene.
Mean fitness	To preserve overall population size, fitness is often measured with respect to the mean fitness of the population.
Meiosis	Division of a diploid cell to produce haploid gametes.
Panmictic	A population is panmictic if each individual is equally likely to compete and interact with any other individual. In practice, this requires that dispersal is fast compared to population genetic time scales.
Polymorphism	A locus with variation, i.e., the population contains several alleles at this locus.
Random mating	Simplifying assumption that mating is independent of genotype, phenotype, and environment.
Recombination	Process of reshuffling of the genetic material in sexual reproduction.
Outcrossing	Fertilization with sperm or pollen from a different individual.
Selfing	Many plants and other organisms have female and male sexual organs and can self-fertilize or self-pollinate.

LIST OF SYMBOLS AND ABBREVIATIONS

g	Haploid genotype: $g = \{s_1, \ldots, s_L\}$.
$P(g, t)$	Genotype distribution in the population.

$\langle \cdots \rangle$	Population average.
$F(g)$	Fitness (growth rate) of genotype g.
$f_{i_1 \cdots i_k}$	Contribution to fitness of the $i_1 \cdots i_k$ set of loci.
a_i	Additive effect of locus i.
σ^2, σ_A^2, σ_I^2	Total, additive, and epistatic variance in fitness.
$\xi_i \in \{0, 1\}$	Origin of locus i, i.e., maternal or paternal.
$C(\{\xi_i\})$	Probability of the recombination pattern $\{\xi_i\}$.
c_{ij}	Probability that loci i and j derive from different parents.
μ, r	Mutation and outcrossing rate.

ACKNOWLEDGMENTS

APPENDIX A: NOTATION

The specification of genotypes and parametrization of genotype-phenotype maps is not unique and our notation differs from the traditional population genetics choice. Conventionally, one chooses one "wild-type" reference genome $(0, 0, \ldots, 0)$ and enumerates deviations from this reference. This is useful when a well-defined wild-type genotype exists. In diverse populations, for example, the progeny of cross between diverged strains, the reference free parametrization we are using here is more natural. The allelic state at each locus is denoted symmetrically by ± 1, e.g., whether an allele comes from one or the other strain. The two different parametrizations are completely equivalent and related to each other by a simple linear transformation (see Table I). In the present context the reference free notation simplifies the algebra since the $s_i = \pm 1$ basis is orthogonal when averaging over the genotype space. The relation to the Fourier transform allows an unambiguous decomposition of the fitness function into additive parts and epistatic components of different order (Parceval's theorem), while in the reference based parametrization of fitness functions, more akin to a Taylor expansion, coefficients depend explicitly on the choice of reference. We also deviated from the traditional D_{ij} notation for linkage disequilibrium because we want to use the diagonal $\chi_{ii} = 1 - \chi_i^2$ components of the cumulant matrix (2 times the heterozygosity at locus i) on the same footing as the off-diagonal ones.

APPENDIX B: QLE IN TERMS OF EFFECTIVE FIELDS

In this Appendix, we discuss how the fields ϕ_i and ϕ_{ij} introduced to parametrize the genotype distribution $P(g, t)$ in Eq. (17) are related to the cumulants of $P(g, t)$. We also detail how the recombination term in Eq. (7) can be evaluated explicitly within the QLE perturbation theory. We parametrized the genotype distribution via

$$\log P(g, t) = \Phi(t) + \sum_i \phi_i(t) s_i + \sum_{i<j} \phi_{ij}(t) s_i s_j. \quad \text{(B1)}$$

The constant term is determined by the normalization of the distribution, the coefficients $\phi_i(t)$ are related to frequencies, and the second order coefficients $\phi_{ij}(t)$ to the connected correlation between loci. In the limit under consideration, the second order contributions are small and we evaluate the coefficients to leading order in $\phi_{ij}(t)$,

$$e^{-\Phi} = \sum_g e^{\sum_i \phi_i s_i + \sum_{i<j} \phi_{ij} s_i s_j} \approx \sum_g e^{\sum_i \phi_i s_i} \left(1 + \sum_{i<j} \phi_{ij} s_i s_j\right)$$

$$= 2^L \left(1 + \sum_{k<j} \phi_{kj} \tanh(\phi_k) \tanh(\phi_j)\right) \prod_{i=1}^L \cosh(\phi_i). \quad \text{(B2)}$$

The relations between χ_i, χ_{ij} and ϕ_i, ϕ_{ij} given in Eq. (20) follow by differentiation.

To arrive at the equations for the time evolution of the fields ϕ_i and ϕ_{ij} [Eq. (24)], we have to evaluate the recombination term in Eq. (23). This is done below. The terms proportional to $\phi_i(t)$ cancel exactly between numerator and denominator and we are left with

$$\sum_{\{\xi_i\}\{s_i'\}} C(\{\xi\}) P(g') \left[\frac{P(g^{(m)}) P(g^{(f)})}{P(g) P(g')} - 1\right] = \sum_{\{\xi_i\}\{s_i'\}} C(\{\xi\}) P(g') \left[e^{\sum_{i<j} \phi_{ij}[(\xi_i s_i + \bar{\xi}_i s_i')(\xi_j s_j + \bar{\xi}_j s_j') + (\bar{\xi}_i s_i + \xi_i s_i')(\bar{\xi}_i s_i + \xi_j s_j') - s_i s_j - s_i' s_j']} - 1\right]$$

$$= \sum_{\{\xi_i\}\{s_i'\}} C(\{\xi\}) P(g') \left[e^{\sum_{i<j} \phi_{ij}[(\xi_i \xi_j + \bar{\xi}_i \bar{\xi}_j - 1)(s_i s_j + s_i' s_j') + (\xi_i \bar{\xi}_j + \bar{\xi}_i \xi_j)(s_i s_j' + s_i' s_j)]} - 1\right]. \quad \text{(B3)}$$

In the limit under consideration, the second order contributions have to be small enough that the entire exponent is small. In this case, the exponential can be expanded and the different terms averaged individually,

$$\sum_{\{\xi_i\}\{s_i'\}} C(\{\xi\}) P(g') \left[\frac{P(g^{(m)}) P(g^{(f)})}{P(g) P(g')} - 1\right] \approx \sum_{\{\xi_i\}} C(\{\xi\}) \sum_{i<j} \phi_{ij}[(\xi_i \xi_j + \bar{\xi}_i \bar{\xi}_j - 1)(s_i s_j + \langle s_i s_j \rangle) + (\xi_i \bar{\xi}_j + \bar{\xi}_i \xi_j)(s_i \langle s_j \rangle + \langle s_i \rangle s_j)]$$

$$= \sum_{i<j} c_{ij} \phi_{ij}[(s_i \langle s_j \rangle + \langle s_i \rangle s_j) - (s_i s_j + \langle s_i s_j \rangle)], \quad \text{(B4)}$$

where c_{ij} is the probability that an odd number of crossovers happened between loci i and j.

APPENDIX C: DIFFUSION THEORY AND WRIGHT'S EQUILIBRIUM

In this Appendix, we detail intermediate steps to arrive at the diffusion equation for the allele frequencies in QLE and the generalized Wright equilibrium. The noise terms in the Langevin equation (36) stem from the multinomial sampling of the genotypes or gametes. From the covariance of the multinomial distribution, we can therefore determine the covariance of the noise terms ζ_i for χ_i and ζ_{ij} for χ_{ij}. The covariance of the changes in χ_i and χ_j, for example, can be calculated as follows:

$$
\begin{aligned}
\langle \Delta\chi_i \Delta\chi_j \rangle &= \left\langle \sum_g s_i \Delta P(g) \sum_g s'_j \Delta P(g') \right\rangle \\
&= \sum_{gg'} s_i s'_j \langle \Delta P(g) \Delta P(g') \rangle \\
&= \frac{1}{N}\left[\sum_g s_i s_j P(g)[1 - P(g)] \right. \\
&\qquad \left. - \sum_{g \neq g'} s_i s'_j P(g) P(g') \right] \\
&= \frac{\chi_{ij}}{N}.
\end{aligned}
\tag{C1}
$$

The other covariance terms can be calculated analogously. In particular, one finds $\langle \Delta\chi_{ij}^2 \rangle \approx N^{-1}\chi_{ii}\chi_{jj} = N^{-1}(1 - \chi_i^2) \times (1 - \chi_j^2)$.

1. The effect of deterministic correlations

If deterministic correlations dominate over the fluctuations in χ_{ij} the forward Kolmogorov equation for the distribution of the χ_i is given by

$$
\partial_t Q(\{\chi_i\}, t) = \sum_i \partial_{\chi_i}\left[\frac{1}{2N}\sum_j \partial_{\chi_j}(\chi_{ij} Q(\{\chi_i\}, t)) + Q(\{\chi_i\}, t)\left(2\mu\chi_i - \sum_j \chi_{ij}\partial_{\chi_j}\langle F\rangle \right) \right].
\tag{C2}
$$

In the steady state, all probability fluxes vanish. The i component of the probability flux is precisely the expression in brackets above and hence has to be equal to zero. Multiplying the bracket with $2N\chi_{ki}^{-1}$ and summing over i (χ_{ki}^{-1} is the ki element of the matrix inverse of χ_{ij}), we have

$$
\partial_{\chi_k} Q(\{\chi_i\}) = Q(\{\chi_i\})\left(-\sum_{ij}\chi_{ki}^{-1}\partial_{\chi_j}\chi_{ij} \right. \\
\left. - 4N\mu\sum_i \chi_{ki}^{-1}\chi_i + 2N\partial_{\chi_k}\langle F\rangle \right).
\tag{C3}
$$

Next, we use the fact that the off-diagonal elements of χ_{ij} are small and $\chi_{ij} = \gamma_{ij}(1 - \chi_i^2)(1 - \chi_j^2)$, while $\chi_{ii} = 1 - \chi_i^2$. To first order in the off-diagonal elements, the inverse is given by $\chi_{ii}^{-1} = (1 - \chi_i^2)^{-1}$ and off-diagonal elements $\chi_{ij}^{-1} = -\chi_{ij}(1 - \chi_j^2)^{-1}(1 - \chi_i^2)^{-1} = -\gamma_{ij}$. Going over the terms in Eq. (C3) one by one, we have

$$
\begin{aligned}
\sum_{ij}\chi_{ki}^{-1}\partial_{\chi_j}\chi_{ij} &= \sum_i \chi_{ki}^{-1}\sum_{j\neq i}\partial_{\chi_j}\chi_{ij} + \sum_i \chi_{ki}^{-1}\partial_{\chi_i}\chi_{ii} \\
&= \sum_i \chi_{ki}^{-1}\chi_{ij}\sum_{j\neq i}\partial_{\chi_j}\log(1 - \chi_j^2) \\
&\quad + \sum_i \chi_{ki}^{-1}\chi_{ii}\partial_{\chi_i}\log(1 - \chi_i^2) \\
&= \partial_{\chi_k}\log(1 - \chi_k^2).
\end{aligned}
\tag{C4}
$$

The mutation term can be evaluated as follows:

$$
\begin{aligned}
4N\mu\sum_i \chi_{ki}^{-1}\chi_i &= -2N\mu\partial_{\chi_k}\log(1 - \chi_k^2) - 4N\mu\sum_{i\neq k}\gamma_{ik}\chi_i \\
&= -2N\mu\partial_{\chi_k}\log(1 - \chi_k^2) \\
&\quad - 4N\mu\partial_{\chi_k}\sum_{i\neq k}\gamma_{ik}\chi_i\chi_k.
\end{aligned}
\tag{C5}
$$

Substituting these terms into Eq. (C3) and $\gamma_{ij} = f_{ij}/rc_{ij}$, we have

$$
\begin{aligned}
\partial_k Q(\{\chi_i\}) = Q(\{\chi_i\})\partial_{\chi_k}\Big(&(2N\mu - 1)\log(1 - \chi_k^2) \\
&+ 2N\Big[\langle F\rangle + 2\mu\sum_{i\neq k}\frac{f_{ik}\chi_i\chi_k}{rc_{ik}}\Big]\Big),
\end{aligned}
\tag{C6}
$$

which is straightforwardly integrated to

$$
Q(\{\chi_i\}) = Ce^{2N\langle F\rangle + 4N\mu\sum_{i<k}f_{ik}\chi_i\chi_k/rc_{ik}}\prod_{i=1}^{L}(1 - \chi_i^2)^{2N\mu - 1}.
\tag{C7}
$$

2. The effect of fluctuating correlations between loci

Even when associations between loci are zero on average, fluctuations of χ_{ij} can affect the allele frequency dynamics. The coupling between different loci acts as an additional noise source on the dynamics of allele frequencies. Grouping deterministic and stochastic forces, the corresponding Langevin equation for χ_i is given by

$$
\begin{aligned}
\chi_i(t + \Delta t) - \chi_i(t) &= \Delta t[\chi_{ii}\partial_{\chi_i}\langle F\rangle - 2\mu\chi_i] \\
&\quad + \int_t^{t+\Delta t}dt'\Big[\sum_{j\neq i}\chi_{ij}(t')\partial_{\chi_j}\langle F\rangle + \zeta_i\Big],
\end{aligned}
\tag{C8}
$$

where the integral constitutes the fluctuating noise term. Solving the Langevin equation for $\chi_{ij}(t)$ assuming constant χ_i and χ_j, one finds

$$
\langle \chi_{ij}(t)\chi_{ij}(t + \Delta t)\rangle = \frac{(1 - \chi_i^2)(1 - \chi_j^2)e^{-rc_{ij}\Delta t}}{2Nrc_{ij}}.
\tag{C9}
$$

Averaging the square of the noise term in Eq. (C8), we find

$$\left\langle \int_t^{t+\Delta t} dt' \int_t^{t+\Delta t} dt'' \left[\sum_{j\neq i}\chi_{ij}\partial_{\chi_j}\langle F\rangle + \zeta_i\right]\right.$$
$$\left.\times \left[\sum_{k\neq i}\chi_{ik}\partial_{\chi_j}\langle F\rangle + \zeta_i\right]\right\rangle$$
$$\approx \frac{\chi_{ii}\Delta T}{N}\left[1 + \frac{1}{2}\sum_{j\neq i}\chi_{jj}\left(\frac{1}{rc_{ij}}\frac{\partial\langle F\rangle}{\partial\chi_j}\right)^2\right]. \quad (C10)$$

The cross term is of order $N^{-3/2}$ and can be neglected.

REFERENCES

Baake, E., 2001, J. Math. Biol. **42**, 455.
Barton, N., 1983, Evolution **37**, 454.
Barton, N., and H. D. Vladar, 2009, Genetics **181**, 997.
Barton, N. H., 1995a, Genet. Res. **65**, 123.
Barton, N. H., 1995b, Genetics **140**, 821.
Barton, N. H., 2001, Molecular Ecology **10**, 551.
Barton, N. H., and S. P. Otto, 2005, Genetics **169**, 2353.
Barton, N. H., and M. Turelli, 1991, Genetics **127**, 229.
Brem, R. B., and L. Kruglyak, 2005, Proc. Natl. Acad. Sci. U.S.A. **102**, 1572.
Bretscher, M. T., C. L. Althaus, V. Müller, and S. Bonhoeffer, 2004, BioEssays **26**, 180.
Bürger, R., 1991, J. Math. Biol. **30**, 199.
Bürger, R., 2000, *The Mathematical Theory of Selection, Recombination, and Mutation* (Wiley, New York).
Callahan, B., R. Neher, D. Bachtrog, P. Andolfatto, and B. Shraiman, 2011, PLoS Genetics **7**, e1001315.
Christiansen, F. B., 1990, Theor. Popul. Biol. **37**, 39.
Cohen, E., D. A. Kessler, and H. Levine, 2005, Phys. Rev. Lett. **94**, 098102.
Derrida, B., 1981, Phys. Rev. B **24**, 2613.
Desai, M. M., and D. S. Fisher, 2007, Genetics **176**, 1759.
de Visser, J. A. G. M., S.-C. Park, and J. Krug, 2009, Am. Nat. **174**, S15.
Edwards, A. W., 1994, Biol. Rev. Camb. Philos. Soc. **69**, 443.
Ehrenreich, I. M., N. Torabi, Y. Jia, J. Kent, S. Martis, J. A. Shapiro, D. Gresham, A. A. Caudy, and L. Kruglyak, 2010, Nature (London) **464**, 1039.
Eigen, M., 1971, Naturwissenschaften **58**, 465.
Ewens, W. J., 1989, Theor. Popul. Biol. **36**, 167.
Falconer, D. S., and T. F. C. Mackay, 1996, *Introduction to Quantitative Genetics* (Pearson, Harlow, UK).
Feldman, M. W., and J. F. Crow, 1970, Theor. Popul. Biol. **1**, 371.
Fisher, R. A., 1930, *The Genetical Theory of Natural Selection* (Clarendon, Oxford).
Frank, S. A., and M. Slatkin, 1992, Trends Ecol. Evol. **7**, 92.
Franklin, I., and R. C. Lewontin, 1970, Genetics **65**, 707.
Franz, S., and L. Peliti, 1997, J. Phys. A **30**, 4481.
Gardiner, C. W., 2004, *Handbook of Stochastic Methods for Physics, Chemistry and The Natural Sciences* (Springer, New York).
Guckenheimer, J., and P. Holmes, 1997, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer, New York).
Haldane, J., 1964, Perspect. Biol. Med. **7**, 343.
Hallatschek, O., 2011, Proc. Natl. Acad. Sci. U.S.A. **108**, 1783.
Hansen, T. F., 2006, Annu. Rev. Ecol. Evol. Syst. **37**, 123.
Hansen, T. F., and G. P. Wagner, 2001, Theor. Popul. Biol. **59**, 61.
Hill, W. G., and A. Robertson, 1966, Genet. Res. **8**, 269.

Hordijk, W., and P. F. Stadler, 1998, Adv. Compl. Syst. **1**, 39.
Iwasa, Y., 1988, J. Theor. Biol. **135**, 265.
Jinks, J. L., C. E. Caten, G. Simchen, and J. H. Croft, 1966, Heredity **21**, 227.
Karlin, S., and M. W. Feldman, 1969, Proc. Natl. Acad. Sci. U.S.A. **62**, 70.
Kauffman, S. A., and E. D. Weinberger, 1989, J. Theor. Biol. **141**, 211.
Kellam, P., C. A. Boucher, J. M. Tijnagel, and B. A. Larder, 1994, J. Gen. Virol. **75**, 341.
Kimura, M., 1955, Cold Spring Harbor Symposia on Quantitative Biology **20**, 33.
Kimura, M., 1956, Evolution **10**, 278.
Kimura, M., 1958, Heredity **12**, 145.
Kimura, M., 1964, J. Appl. Probab. **1**, 177.
Kimura, M., 1965, Genetics **52**, 875.
Kingman, J., 1978, J. Appl. Probab. **15**, 1.
Kirkpatrick, M., T. Johnson, and N. H. Barton, 2002, Genetics **161**, 1727.
Lewontin, R., and K. Kojima, 1960, Evolution **14**, 458.
Livnat, A., C. Papadimitriou, J. Dushoff, and M. W. Feldman, 2008, Proc. Natl. Acad. Sci. U.S.A. **105**, 19803.
Lynch, M., and B. Walsh, 1998, *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, MA).
Maynard Smith, J., 1968, Am. Nat. **102**, 469.
Mayr, E., 1963, *Animal Species and Evolution* (Belknap Press, Cambridge, MA).
McQuarrie, D., 1973, *Statistical Mechanics* (University Science Books, Sausalito, CA).
Mezard, M., G. Parisi, and M. A. Virasoro, 1987, *Spin Glass Theory and Beyond*, Lecture Notes in Physics (World Scientific, Singapore).
Mézard, M., and A. Montanari, 2009, *Information, Physics, and Computation* (Oxford University Press, New York).
Muller, H. J., 1932, Am. Nat. **66**, 118.
Mustonen, V., and M. Lassig, 2005, Proc. Natl. Acad. Sci. U.S.A. **102**, 15 936.
Mustonen, V., and M. Lässig, 2010, Proc. Natl. Acad. Sci. U.S.A. **107**, 4248.
Nagylaki, T., 1993, Genetics **134**, 627.
Neher, R. A., and B. I. Shraiman, 2009, Proc. Natl. Acad. Sci. U.S.A. **106**, 6866.
Neher, R. A., B. I. Shraiman, and D. S. Fisher, 2010, Genetics **184**, 467.
Nolte, A. W., and D. Tautz, 2010, Trends Genet. **26**, 54.
Nora, T., C. Charpentier, O. Tenaillon, C. Hoede, F. Clavel, and A. J. Hance, 2007, J. Virol. **81**, 7620.
Ohta, T., and M. Kimura, 1969, Genetics **63**, 229.
Orr, H. A., 1995, Genetics **139**, 1805.
Price, G. R., 1970, Nature (London) **227**, 520.
Price, G. R., 1972, Ann. Hum. Genet. **36**, 129.
Prügel-Bennett, A., and J. Shapiro, 1994, Phys. Rev. Lett. **72**, 1305.
Robertson, A., 1966, Animal Production **8**, 95.
Rouzine, I. M., and J. M. Coffin, 2005, Genetics **170**, 7.
Rouzine, I. M., J. Wakeley, and J. M. Coffin, 2003, Proc. Natl. Acad. Sci. U.S.A. **100**, 587.
Roze, D., and N. H. Barton, 2006, Genetics **173**, 1793.
Seidel, H. S., M. V. Rockman, and L. Kruglyak, 2008, Science **319**, 589.
Sella, G., and A. E. Hirsh, 2005, Proc. Natl. Acad. Sci. U.S.A. **102**, 9541.
Sherrington, D., and S. Kirkpatrick, 1975, Phys. Rev. Lett. **35**, 1792.
Slatkin, M., 1972, Genetics **72**, 157.

Stadler, P. F., and G. P. Wagner, 1998, IEEE Trans. Evol. Comput. **5**, 241.

Stephan, W., Y. S. Song, and C. H. Langley, 2006, Genetics **172**, 2647.

Tsimring, L., H. Levine, and D. Kessler, 1996, Phys. Rev. Lett. **76**, 4440.

Turelli, M., and N. H. Barton, 1994, Genetics **138**, 913.

Turner, J., 1967, Evolution **21**, 645.

Weinberger, E., 1991, Biol. Cybern. **65**, 321.

Weinreich, D. M., N. F. Delaney, M. A. Depristo, and D. L. Hartl, 2006, Science **312**, 111.

Weissman, D., M. Feldman, and D. Fisher, 2010, Genetics **186**, 1389.

Wright, S., 1931, Genetics **16**, 97.

Wright, S., 1932, in *Proceedings of the Sixth International Congress of Genetics* (Brooklyn Botanic Garden, New York), pp. 356–366.

Wright, S., 1942, Bull. Am. Math. Soc. **48**, 223.