Supplement
Competition between recombination and epistasis can cause a transition from allele to genotype selection

Richard A. Neher* and Boris I. Shraiman*†
*Kavli Institute for Theoretical Physics and †Department of Physics
University of California, Santa Barbara, CA 93106, USA

March 17, 2009

Linkage disequilibrium in QLE

The limit of rapid recombination can be understood in perturbation theory in $\sigma \ll r$ pioneered by Kimura[1] and further developed by a number of authors[2;3]. Kimura showed (for $L = 2$) that in this limit, alleles at different loci develop correlations proportional to the epistatic interaction between these loci and the dynamics of the joint probability distribution function comes to a quasi-steady state which he termed quasi linkage equilibrium (QLE). In the QLE state linkage disequilibrium tends to a fixpoint where the rate of build-up of correlations due to selection equals the rate of their break-up by recombination. In the $\sigma \ll r$ limit the QLE state evolves slowly, following the dynamics of L individual allele frequencies ν_i. In QLE, the linkage disequilibrium between two loci i and j is given by

$$D_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle \approx \frac{f_{ij}\nu_i(1-\nu_i)\nu_j(1-\nu_j)}{\rho|j-i|}, \tag{1}$$

where ρ is the recombination rate per locus and $|j-i|$ is the distance between the two loci. Since the allele frequencies are changing over time, it is more convenient to study the quantity

$$\psi_{ij} = \frac{D_{ij}}{\nu_i(1-\nu_i)\nu_j(1-\nu_j)} \approx \frac{f_{ij}}{\rho|j-i|}. \tag{2}$$

Since we are not interested in the disequilibrium between a particular pair of loci but in general properties of the population, we consider the sum of all ψ_{ij}^2. For f_{ij} drawn from a Gaussian distribution with variance $\frac{2\sigma^2}{L(L-1)}$, we have

$$\sum_{ij} \psi_{ij}^2 \approx \frac{2\sigma^2}{\rho^2 L(L-1)} \sum_{i<j} \frac{1}{(j-i)^2} = \frac{2\sigma^2}{\rho^2 L(L-1)} \sum_{k=1}^{L} \frac{L-k}{k^2} \approx \frac{\pi^2\sigma^2}{3\rho^2 L} \tag{3}$$

For a circular chromosome, where each recombination event results in two crossovers, the result is $\frac{\pi^2\sigma^2}{12\rho^2 L}$.

In addition to the deterministic contribution to linkage disequilibrium, there is also a contribution due to sampling fluctuations, i.e. random drift. The contribution of drift to ψ_{ij}^2 is proportional to $(N\rho|j-i|)^{-1}$ for linked loci and obligate mating or proportional to $(Nr)^{-1}$ for unlinked loci with outcrossing rate r. The N dependence in the latter case is well confirmed by the data shown in the Fig. S1.

The critical recombination rate

The self-consistency condition for the QLE state presented in the main text builds on an equation for the evolution of the joint probability distribution $P(A,E;t)$ of the additive and epistatic fitness contributions A and E.

$$\partial_t P(A,E;t) = (F - \bar{F} - r)P(A,E;t) + r\rho(E)\vartheta(A;t) \tag{4}$$

As explained in the manuscript, the last term describing the generation of genotype by recombination term is the product of the distribution of possible epistatic fitness values $\rho(E)$ (which depends on the model and is Gaussian in our case) and the marginal distribution of additive fitness $\vartheta(A) = \int dE P(A,E)$.

Equation 4 is solved by a factorized ansatz $P(A,E;t) = \vartheta(A;t)\omega(E)$. The additive part $\vartheta(A;t)$ evolves according to $\partial_t\vartheta(A;t) = (A - \bar{A}(t))\vartheta(A;t)$, which has the solution

$$\vartheta(A;t) = \frac{1}{\sqrt{2\pi V_A}} e^{-\frac{(A-\bar{A}(t))^2}{2V_A}} \tag{5}$$

with $\bar{A}(t) = A_0 + V_A t$. The distribution $\omega(E)$ of the epistatic fitness E is time-independent and given by

$$\omega(E) = \frac{r\rho(E)}{r + \bar{E} - E}, \tag{6}$$

where $\bar{E}$ is determined by the condition that $\omega(E)$ has to be normalized. Here, we discuss when this solution ceases to exist as the recombination rate decreases. Obviously, no such solution exists if $\rho(E) > 0$ for arbitrarily large E since the denominator vanishes at $E = \bar{E} + r$. However, any finite genotype space will have some maximal $E = E_{max}$. Furthermore, the population size is often much smaller than the number of available genotypes, in which case the typical maximal E encountered by the population will play the role of E_{max}, see below.

For now, assume the population is infinite and completely samples $\rho(E)$. As r decreases, the distribution $\omega(E)$ of E in the population shifts to larger values and $\bar{E}$ increases. The susceptibility of $\bar{E}$ to changes in r depends on the upper tail of the distribution $\rho(E)$ and is smaller for more rapidly decaying $\rho(E)$. The reason for this behavior is that changing r does not affect $\omega(E)$ much in the bulk where the $r + \bar{E} - E$ is large, but has strong effect in the upper tail where the $r + \bar{E} - E$ is small. Thus, changes in r affect $\bar{E}$ only as much as the tail of

$\omega(E)$ contributes to the mean. If $\rho(E)$ decreases linearly or faster at E_{max}, a critical r_c exists below which no self-consistent solution for $\omega(E)$ can be found. Below r_c, genotypes with $E > \bar{E} + r$ form clones that grow exponentially and the distribution of epistatic fitness E is no longer time independent.

A finite population undersamples the genotype space and the range of values of E the population encounters determines r_c. In our case, the available states are 2^L random samples from a Gaussian distribution with variance V_I. For this choice of $\rho(E)$, it turns out that $\bar{E}$ responds very weakly to changes in r and remains much smaller than $\sqrt{V_I}$ for $r > r_c$, such that $r_c \approx E_{max}$. The initial population sampled N genotypes and new genotypes are generated through mating with rate Nr. However, even a sufficiently fit genotype with $E > \bar{E} + r$ will only establish with probability $E - (\bar{E} + r)$, which is at most the spacing between the extremal samples of E given by $p_{fix} \sim \sqrt{V_I}/\sqrt{2 \ln Nr\tau}$[1]. This leads to an approximate equation for r_c

$$\bar{E} + r_c \approx \sqrt{2V_I \ln(Nr_c\tau p_{fix})}, \tag{7}$$

where the time to fixation τ is determined by the the additive fitness component. Using $N = 10^5$, $\sigma^2 = 0.005$, $\tau = 500$ and correcting for the discrete recombination scheme yields $r_c \approx 0.3$, in very good agreement with the simulation result $r_c \approx 4\sigma \approx 0.28$.

The non-monotonicity of the final fitness

In the main text, we presented data showing that the fitness of the fixated genotype F_{final} has a non-monotonous dependence on the outcrossing rate, exhibiting a peak just below r_c. Underlying reason for the peak is the different population dynamics in the two phases. In the CC phase, the population explores genotypes and fixates the fittest found. The number of genotypes sampled grows in time as Nrt, which, assuming independent samples from a Gaussian distribution, typically yields a maximal fitness of $\approx \sqrt{2V_I \ln Nrt}$, i.e. the maximum grows with N and r. The time t is limited by the fixation time, which grows as $\ln N$ in the CC phase. The prefactor is determined by the strength of selection and the number of times a new fitter genotype is created and sweeps to frequencies of order one before fixation. The fitness of the fittest genotype created during CC evolution increases therefore even faster with N since the genotype space is explored for longer times for larger N. In the QLE phase, the population dynamics is determined by the dynamics of the allele frequencies, resulting in a more or less deterministic path to fixation independent of N (assuming N^{-1} is smaller than single locus effects, see Fig. S3). In the PE model, the QLE dynamics leads typically to a $F_{\text{final}} \sim \sqrt{V_I L}$([4]). For sufficiently large N, the fittest genotype found in CC will be fitter than the QLE fixate, resulting in a peak of F_{final} just below r_c. In practice, limited population sizes and the in-

[1] Logarithmic factors in p_{fix} influence the results only very weakly.

crease of F_{final} with L in the QLE dynamics of the PE model implies that a peak will only be observed for sufficiently small L.

References

[1] Kimura M (1965) Attainment of Quasi Linkage Equilibrium When Gene Frequencies Are Changing by Natural Selection. *Genetics* **52**:875–890.

[2] Barton N. H, and Turelli M (1991) Natural and sexual selection on many loci. *Genetics* **127**:229–255.

[3] Nagylaki T (1993) The evolution of multilocus systems under weak selection. *Genetics* **134**:627–647.

[4] Parisi G (1995) On the Statistical Properties of the Large Time Zero Temperature Dynamics of the SK Model. *arXiv* cond-mat/9501045v1.