File S2

Models for allele frequencies

The vector of allele counts at locus $\boldsymbol{j}$ in an isolated population $\boldsymbol{A}$ for generation $\boldsymbol{t+1}$ follows the multinomial distribution

$$n_{Aj}(t+1)|n_{Aj}(t)\sim\mathrm{Mult}\big(2n_A, p_{Aj}(t)\big) \quad (\mathrm{Eq.\,S8})$$

where $\boldsymbol{p}_{Aj}(t) = \frac{\boldsymbol{n}_{Aj}(t)}{2n_A}$ is the allele frequency for the generation t. While $\boldsymbol{n}_{Aj}$ follows a multinomial random walk, $\boldsymbol{p}_{Aj}$ follows a corresponding process on an $n_j - 1$ dimensional simplex. This discrete process is often approximated by a continuous-valued random process, the so-called Wright-Fisher diffusion (see e.g. Nicholson *et al.* 2002). Kimura (1955) first derived the exact solution for the distribution of allele frequencies of a biallelic locus under Wright-Fisher diffusion. This solution is not Gaussian, because the diffusion is non-isotropic. Solutions have also been obtained for multiallelic loci (Tavaré 1984; Xie 2011). However, implementing these solutions in the AFM framework would pose considerable computational challenges because of the need to iterate infinite, high-dimensional sums. In case of biallelic loci, the solution of Wright-Fisher diffusion is often approximated by a truncated normal distribution (e.g. Balding 2003; Coop *et al.* 2010; Nicholson *et al.* 2002). However, this approximation cannot be applied on multiallelic loci as such, because the distribution of $\boldsymbol{p}_{Aj}$ needs to be restricted on the simplex Δ^{n_j-1}. The alternative that we apply here is to use the Dirichlet distribution as a phenomenological, i.e. non-mechanistic, model for allele frequencies.

Application of the Dirichlet distribution as a model of pure drift may be considered questionable for two reasons. Firstly, the Dirichlet distribution is known to arise as an equilibrium distribution from the balance of random drift and mutation or migration (e.g. Nicholson *et al.* 2002; Rannala 1996), but not as a result of pure random drift in an isolated population. Secondly, the Dirichlet distribution is a continuous distribution such that each component is restricted on the open interval $]0,1[$, which gives a zero probability for the fixation of any one allele. However, with a small value of the parameter a_A, the Dirichlet distribution can have much of its probability mass very close to the boundaries. Thus, when supplemented with a sampling model for a finite population,

$$\boldsymbol{n}_{Aj}\sim\mathrm{Mult}\big(2\boldsymbol{n}_A, \boldsymbol{p}'_{Aj}\big),$$

$$\boldsymbol{p}'_{Aj}\sim\mathrm{Dirichlet}\big(a_A\boldsymbol{q}_j\big),$$

the Dirichlet model is able to predict a high probability of fixation. In the AFM, we use Dirichlet-distributed allele frequencies z_{Aj} to model the evolutionary history of the independent lineages, and the multinomial step naturally follows

from the fact that the sample of genotypes is finite, even if the whole subpopulation is sampled. Below, we investigate this model by a comparison with the truncated normal distribution in a biallelic case where both distributions are easily tractable.

We consider a closed population of N individuals that mates randomly for T generations, and assume that the initial frequency of allele 1 has been q_{j1}. In this Supplement, we focus on four representative cases: symmetric allele frequencies with moderate drift (Scenario 1, Fig. S1), symmetric allele frequencies with a high amount of drift (Scenario 2, Fig. S1), uneven allele frequencies with moderate drift (Scenario 3, Fig. S1) and uneven allele frequencies with a high amount of drift (Scenario 4, Fig. S1). To sample from this model, we first generated a sample of size 10^5 from the last generation by using the true model (repeated application of Eq. S8). Then, we derived a corresponding sample from the Dirichlet approximation by randomizing $\boldsymbol{p}'_{Aj}$ for 10^5 times and sampling the allele counts for each realization from $\mathrm{Mult}\left(2n_A, \boldsymbol{p}'_{Aj}\right)$. Finally, we considered the model of allele frequencies under the truncated normal approximation. As suggested by Nicholson et al. (2002), we specified the allele frequency as

$$p_{Aj1} \sim \mathrm{N}\left(q_{j1}, cq_{j1}(1-q_{j1})\right) \coloneqq \Phi$$

so that the extinction probability of allele 1 was calculated as $\Phi(0)$ and the fixation probability as $1-\Phi(1)$. We calculated the pointwise probabilities of the discrete classes as $\Phi'(p_{Aj1})/2n_A$, i.e. by dividing the Gaussian density function by the number of discrete values in]0,1]. While theoretical values exist for the drift parameters c and a_A given the demographic model, we optimized the values of these parameters in each scenario by minimizing the square distance (denoted D^2) with the true (empirical) distribution.

The results show that both the Dirichlet and truncated normal are imperfect approximations. In scenario 1, where drift is moderate and fixations do not occur, both approximations are qualitatively good, while the truncated normal distribution has a better goodness of fit (Multinomial-Dirichlet $D^2 = 9.2\times10^{-8}$; truncated normal $D^2 = 5.9\times10^{-8}$). In scenario 2, the truncated normal approximation has a better goodness of fit ($D^2 = 5.9\times10^{-7}$), than the Dirichlet approximation ($D^2 = 1.5\times10^{-5}$) which has an inconveniently convex shape in this case. In scenario 3, the truncated normal approximation has a better goodness of fit ($D^2 = 1.1\times10^{-5}$), but it is qualitatively different from the data by having a clear mode in the interior of $[0,1]$ which the Dirichlet approximation ($D^2 = 2.5\times10^{-5}$) does not have. In scenario 4, the Dirichlet approximation is better ($D^2 = 3.8\times10^{-4}$ as opposed to truncated normal $D^2 = 6.1\times10^{-4}$). In general, both distributions have problems in coping with the data when the amount of drift is high, which shows in the increase of the square distances. Finally, we note that the expectation of the truncated normal distribution is not strictly q_{j1} which would be expected under pure random drift. On the other hand, this is likely to be unimportant when the amount of drift is low.

References

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol **63:** 221-230.

Coop, G., D. Witonsky, A. Di Rienzo and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics **185:** 1411-1423.

Kimura, M., 1955 Solution of a Process of Random Genetic Drift with a Continuous Model. Proceedings of the National Academy of Sciences of the United States of America **41:** 144-150.

Nicholson, G., A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson *et al.*, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society Series B-Statistical Methodology **64:** 695-715.

Rannala, B., 1996 The Sampling Theory of Neutral Alleles in an Island Population of Fluctuating Size. Theor Popul Biol **50:** 91-104.

Tavare, S., 1984 Line-of-Descent and Genealogical Processes, and Their Applications in Population-Genetics Models. Theoretical Population Biology **26:** 119-164.

Xie, X. H., 2011 The Site-Frequency Spectrum of Linked Sites. Bulletin of Mathematical Biology **73:** 459-494.