

SUPPLEMENTAL MATERIAL: FILE S2

DIFFUSION APPROXIMATIONS FOR WRIGHT-FISHER MODELS WITH SELECTION

We present the formulation of the diffusion approximation for the Wright-Fisher model with selection here and refer to Durrett (2008) for a rigorous proof, especially for one- and two-locus population dynamics. For simplicity of notation, we introduce a short-hand notation for the conditional expectation of a random variable given the population of the haplotype frequencies \mathbf{x} in generation k

$$E^{(k,\mathbf{x})}(\cdot) = E\left(\cdot \mid \mathbf{X}^{(N)}(k) = \mathbf{x}\right),$$

and a function ρ_{uv}^i of three variables u, v and i , defined as

$$\rho_{uv}^i = \frac{1}{2}(\delta_{ui} + \delta_{vi}),$$

where δ_{ui} and δ_{vi} are the Kronecker delta functions.

Let $\Delta X_i^{(N)}(k)$ denote the change in the frequency of haplotype i from generation k to the next, and we have

$$E^{(k,\mathbf{x})}\left(\Delta X_i^{(N)}(k)\right) = \sum_{u,v} \rho_{uv}^i E^{(k,\mathbf{y})}\left(\Delta Y_{uv}^{(N)}(k)\right) \quad (1)$$

$$E^{(k,\mathbf{x})}\left(\Delta X_i^{(N)}(k)\Delta X_j^{(N)}(k)\right) = \sum_{u,v} \sum_{m,n} \rho_{uv}^i \rho_{mn}^j E^{(k,\mathbf{y})}\left(\Delta Y_{uv}^{(N)}(k)\Delta Y_{mn}^{(N)}(k)\right). \quad (2)$$

Based on the construction of the Wright-Fisher model with selection, we have

$$E^{(k,\mathbf{y})}\left(\Delta Y_{uv}^{(N)}(k)\right) = q_{uv} - y_{uv} \quad (3)$$

$$E^{(k,\mathbf{y})}\left(\Delta Y_{uv}^{(N)}(k)\Delta Y_{mn}^{(N)}(k)\right) = \frac{1}{N}q_{uv}(\delta_{um}\delta_{vn} - q_{mn}) + (q_{uv} - y_{uv})(q_{mn} - y_{mn}). \quad (4)$$

where q_{uv} is the frequency of genotype made up of haplotype u and v of an effectively infinite population after the possible mechanisms of evolutionary change (except population regulation) at intermediate stages of the life cycle such as natural selection, meiosis and random mating within generation k . Substituting Eqs. (3) and (4) into Eqs. (1) and (2), respectively, we have

$$E^{(k,\mathbf{x})}\left(\Delta X_i^{(N)}(k)\right) = p_i - x_i$$

$$E^{(k,\mathbf{x})}\left(\Delta X_i^{(N)}(k)\Delta X_j^{(N)}(k)\right) = \frac{1}{2N}p_i(\delta_{ij} - p_j) + (p_i - x_i)(p_j - x_j),$$

where

$$p_i = \sum_{u,v} \rho_{uv}^i q_{uv} \quad (5)$$

is the frequency of haplotype i of an effectively infinite population after the possible mechanisms of evolutionary change (except population regulation) at intermediate stages of the life cycle such as natural selection, meiosis and random mating within generation k .

Considering the limits as the population size N goes to infinity, we can formulate the infinitesimal mean vector $\boldsymbol{\mu}(t, \mathbf{x})$ as

$$\begin{aligned} \mu_i(t, \mathbf{x}) &= \lim_{N \rightarrow \infty} 2NE^{([2Nt], \mathbf{x})} \left(\Delta X_i^{(N)}([2Nt]) \right) \\ &= \lim_{N \rightarrow \infty} 2N(p_i - x_i) \end{aligned}$$

and the infinitesimal covariance matrix $\boldsymbol{\Sigma}(t, \mathbf{x})$ as

$$\begin{aligned} \Sigma_{ij}(t, \mathbf{x}) &= \lim_{N \rightarrow \infty} 2NE^{([2Nt], \mathbf{x})} \left(\Delta X_i^{(N)}([2Nt]) \Delta X_j^{(N)}([2Nt]) \right) \\ &= \lim_{N \rightarrow \infty} p_i(\delta_{ij} - p_j) + 2N(p_i - x_i)(p_j - x_j), \end{aligned}$$

where $[\cdot]$ is used to denote the integer part of the value in the brackets, according to standard techniques of diffusion theory (see, for example, Karlin and Taylor, 1981).

Therefore, the process $\mathbf{X}^{(N)}$ converges to a diffusion process, denoted by $\mathbf{X} = \{\mathbf{X}(t), t \geq 0\}$ and referred to as the Wright-Fisher diffusion with selection, satisfying the stochastic differential equation of the form

$$d\mathbf{X}(t) = \boldsymbol{\mu}(t, \mathbf{X}(t)) dt + \boldsymbol{\sigma}(t, \mathbf{X}(t)) d\mathbf{W}(t),$$

where the diffusion coefficient matrix $\boldsymbol{\sigma}(\mathbf{x})$ satisfies the relation that

$$\boldsymbol{\sigma}(t, \mathbf{x}) \boldsymbol{\sigma}^T(t, \mathbf{x}) = \boldsymbol{\Sigma}(t, \mathbf{x})$$

and $\mathbf{W}(t)$ is a multi-dimensional standard Brownian motion.

One-locus population dynamics. For the population evolving at a single locus, we use a common category of fitness values for a diploid population at a single locus (see Hamilton, 2011, for other categories of fitness values presented in terms of selection coefficients), which can be presented as follows: genotypes $\mathcal{A}_1\mathcal{A}_1$, $\mathcal{A}_1\mathcal{A}_2$ and $\mathcal{A}_2\mathcal{A}_2$ at a given locus \mathcal{A} have fitness values 1, $1 - h_{\mathcal{A}}s_{\mathcal{A}}$ and $1 - s_{\mathcal{A}}$, respectively, where $s_{\mathcal{A}}$ is the selection coefficient and $h_{\mathcal{A}}$ is the dominance parameter. Substituting the sampling probabilities $q_{uv}^{(v)}$ and $q_{uv}^{(f)}$ for $u, v = 1, 2$ into Eq. (5), respectively, we have

$$\begin{aligned} p_1^{(v)} &= \frac{x_1^2 + (1 - h_{\mathcal{A}}s_{\mathcal{A}})x_1x_2}{1 - (2h_{\mathcal{A}}x_1x_2 - x_2^2)s_{\mathcal{A}}} \\ p_2^{(v)} &= \frac{(1 - h_{\mathcal{A}}s_{\mathcal{A}})x_1x_2 + (1 - s_{\mathcal{A}})x_2^2}{1 - (2h_{\mathcal{A}}x_1x_2 - x_2^2)s_{\mathcal{A}}} \end{aligned}$$

and

$$p_1^{(f)} = \frac{x_1^2 + (1 - h_{\mathcal{A}}s_{\mathcal{A}})x_1x_2}{1 - (2h_{\mathcal{A}}x_1x_2 - x_2^2)s_{\mathcal{A}}}$$

$$p_2^{(f)} = \frac{(1 - h_{\mathcal{A}}s_{\mathcal{A}})x_1x_2 + (1 - s_{\mathcal{A}})x_2^2}{1 - (2h_{\mathcal{A}}x_1x_2 - x_2^2)s_{\mathcal{A}}},$$

That is, we have

$$\left| p_i^{(v)} - p_i^{(f)} \right| = 0$$

for $i = 1, 2$, which implies that we have the same representation of the infinitesimal mean vectors and covariance matrices for the cases that natural selection takes the form of viability or fecundity selection,

$$\begin{aligned} \mu_1(\mathbf{x}) &= \alpha_{\mathcal{A}}x_1x_2[x_1h_{\mathcal{A}} + x_2(1 - h_{\mathcal{A}})] \\ \mu_2(\mathbf{x}) &= -\alpha_{\mathcal{A}}x_1x_2[x_1h_{\mathcal{A}} + x_2(1 - h_{\mathcal{A}})] \end{aligned} \quad (6)$$

and

$$\Sigma_{ij}(\mathbf{x}) = x_i(\delta_{ij} - x_j) \quad (7)$$

for $i, j = 1, 2$. Eqs. (6) and (7) ensure the same SDE-representation of the one-locus Wright-Fisher diffusion with selection for the two types of natural selection, viability and fecundity selection.

Two-locus population dynamics. For the population evolving at two linked loci, we use a same category of fitness values for a diploid population at a single locus and assume that the fitness values of two-locus genotypes are determined multiplicatively from fitness values at individual loci, *e.g.*, the fitness value of the $\mathcal{A}_1\mathcal{B}_2/\mathcal{A}_2\mathcal{B}_2$ genotype is $(1 - h_{\mathcal{A}}s_{\mathcal{A}})(1 - s_{\mathcal{B}})$. We let r denote the recombination rate between the two loci (*i.e.*, the rate that a recombinant gamete is produced at meiosis). Substituting the sampling probabilities $q_{uv}^{(v)}$ and $q_{uv}^{(f)}$ for $u, v = 1, 2, 3, 4$ into Eq. (5), respectively, and sing Taylor expansions with respect to the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ and the recombination rate r , we have

$$\begin{aligned} p_1^{(v)} &= x_1 - (x_1x_4 - x_2x_3)r + d_{\mathcal{A}}(x_1(x_3 + x_4) - (x_3 + x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\ &\quad + d_{\mathcal{B}}(x_1(x_2 + x_4) - (x_2 + x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2 + r^2) \\ p_2^{(v)} &= x_2 + (x_1x_4 - x_2x_3)r + d_{\mathcal{A}}(x_2(x_3 + x_4) + (x_3 + x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\ &\quad - d_{\mathcal{B}}(x_2(x_1 + x_3) + (x_1 + x_3)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2 + r^2) \\ p_3^{(v)} &= x_3 + (x_1x_4 - x_2x_3)r - d_{\mathcal{A}}(x_3(x_1 + x_2) + (x_1 + x_2)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\ &\quad + d_{\mathcal{B}}(x_3(x_2 + x_4) + (x_2 + x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2 + r^2) \\ p_4^{(v)} &= x_4 - (x_1x_4 - x_2x_3)r - d_{\mathcal{A}}(x_4(x_1 + x_2) - (x_1 + x_2)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\ &\quad - d_{\mathcal{B}}(x_4(x_1 + x_3) - (x_1 + x_3)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2 + r^2) \end{aligned}$$

and

$$\begin{aligned}
p_1^{(f)} &= x_1 - (x_1x_4 - x_2x_3)r + d_{\mathcal{A}}(x_1(x_3 + x_4) + (x_1 + x_2 - x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\
&\quad + d_{\mathcal{B}}(x_1(x_2 + x_4) + (x_1 - x_2 + x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2) \\
p_2^{(f)} &= x_2 + (x_1x_4 - x_2x_3)r + d_{\mathcal{A}}(x_2(x_3 + x_4) - (x_1 + x_2 - x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\
&\quad - d_{\mathcal{B}}(x_2(x_1 + x_3) + (x_1 - x_2 + x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2) \\
p_3^{(f)} &= x_3 + (x_1x_4 - x_2x_3)r - d_{\mathcal{A}}(x_3(x_1 + x_2) + (x_1 + x_2 - x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\
&\quad + d_{\mathcal{B}}(x_3(x_2 + x_4) - (x_1 - x_2 + x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2) \\
p_4^{(f)} &= x_4 - (x_1x_4 - x_2x_3)r - d_{\mathcal{A}}(x_4(x_1 + x_2) - (x_1 + x_2 - x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{A}} \\
&\quad - d_{\mathcal{B}}(x_4(x_1 + x_3) - (x_1 - x_2 + x_3 - x_4)(x_1x_4 - x_2x_3)r)s_{\mathcal{B}} + \mathcal{O}(s_{\mathcal{A}}^2 + s_{\mathcal{A}}s_{\mathcal{B}} + s_{\mathcal{B}}^2),
\end{aligned}$$

where

$$\begin{aligned}
d_{\mathcal{A}} &= (x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}}) \\
d_{\mathcal{B}} &= (x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}}).
\end{aligned}$$

That is, we have

$$\left| p_i^{(v)} - p_i^{(f)} \right| = \mathcal{O}\left(\frac{1}{N^2}\right)$$

for $i = 1, 2, 3, 4$ under the assumption that the selection coefficients $s_{\mathcal{A}}$ and $s_{\mathcal{B}}$ and the recombination rate r are all of order $\mathcal{O}(1/N)$, which implies that we have the same representation of the infinitesimal mean vectors and covariance matrices for the cases that natural selection takes the form of viability or fecundity selection,

$$\begin{aligned}
\mu_1(\mathbf{x}) &= \alpha_{\mathcal{A}}x_1(x_3 + x_4)[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})] \\
&\quad + \alpha_{\mathcal{B}}x_1(x_2 + x_4)[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})] - R(x_1x_4 - x_2x_3) \\
\mu_2(\mathbf{x}) &= \alpha_{\mathcal{A}}x_2(x_3 + x_4)[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})] \\
&\quad - \alpha_{\mathcal{B}}x_2(x_1 + x_3)[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})] + R(x_1x_4 - x_2x_3) \\
\mu_3(\mathbf{x}) &= -\alpha_{\mathcal{A}}x_3(x_1 + x_2)[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})] \\
&\quad + \alpha_{\mathcal{B}}x_3(x_2 + x_4)[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})] + R(x_1x_4 - x_2x_3) \\
\mu_4(\mathbf{x}) &= -\alpha_{\mathcal{A}}x_4(x_1 + x_2)[(x_1 + x_2)h_{\mathcal{A}} + (x_3 + x_4)(1 - h_{\mathcal{A}})] \\
&\quad - \alpha_{\mathcal{B}}x_4(x_1 + x_3)[(x_1 + x_3)h_{\mathcal{B}} + (x_2 + x_4)(1 - h_{\mathcal{B}})] - R(x_1x_4 - x_2x_3),
\end{aligned} \tag{8}$$

and

$$\Sigma_{ij}(\mathbf{x}) = x_i(\delta_{ij} - x_j) \tag{9}$$

for $i, j = 1, 2, 3, 4$, where $\alpha_{\mathcal{A}} = 2Ns_{\mathcal{A}}$, $\alpha_{\mathcal{B}} = 2Ns_{\mathcal{B}}$ and $R = 2Nr$ are the scaled selection coefficients and recombination rate, respectively. Eqs. (8) and (9) guarantee the same SDE-representation of

the two-locus Wright-Fisher diffusion with selection for the two types of natural selection, viability and fecundity selection.

Therefore, we can conclude that we have the same diffusion approximations for the Wright-Fisher models with selection for the population dynamics at a single locus or two linked loci no matter whether viability or fecundity selection is occurring.

REFERENCES

- Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer-Verlag, New York.
- Hamilton, M. (2011). *Population Genetics*. Wiley-Blackwell, Chichester.
- Karlin, S. and Taylor, H. E. (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.