Bull Math Biol (2011) 73: 459–494
DOI 10.1007/s11538-010-9534-3

ORIGINAL ARTICLE

The Site-Frequency Spectrum of Linked Sites

Xiaohui Xie

Received: 11 December 2009 / Accepted: 8 March 2010 / Published online: 27 March 2010

Abstract The site-frequency spectrum, representing the distribution of allele frequencies at a set of polymorphic sites, is a commonly used summary statistic in population genetics. Explicit forms of the spectrum are known for both models with and without selection if independence among sites is assumed. The availability of these explicit forms has allowed for maximum likelihood estimation of selection, developed first in the Poisson random field model of Sawyer and Hartl, which is now the primary method for estimating selection directly from DNA sequence data. The independence assumption, which amounts to assume free recombination between sites, is, however, a limiting case for many population genetics models. Here, we extend the site-frequency spectrum theory to consider the case where the sites are completely linked. We use diffusion approximation to calculate the joint distribution of the allele frequencies of linked sites for models without selection and for models with equal coefficient selection. The joint distribution is derived by first constructing Green's functions corresponding to multiallele diffusion equations. We show that the site-frequency spectrum is highly correlated between frequencies that are complementary (i.e., sum to 1), and the correlation is significantly elevated by positive selection. The results presented here can be used to extend the Poisson random field to allow for estimating selection for correlated sites. More generally, the Green's function construction should be able to aid in studying the genetic drift of multiple alleles in other cases.

Keywords Site-frequency spectrum · Wright–Fisher model · Diffusion approximation · Green's function

X. Xie (✉)
Department of Computer Science, Center for Complex Biological Systems, Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697, USA
e-mail: xhx@ics.uci.edu

1 Introduction

The site-frequency spectrum is the distribution of allele frequencies at a set of polymorphic sites. The statistic was originally developed to study models that assume irreversible mutations. Fisher and Wright first demonstrated that although no equilibrium can be reached at each individual site in these models, the distribution of allele frequencies across polymorphic loci does reach an equilibrium if population size and selection intensities are both kept constant (Fisher 1930; Wright 1938). Kimura later extended the theory to consider mutations at individual bases of DNA sequences by introducing the concept of "infinitely-many-sites" (Kimura 1969).

Under the infinitely-many-sites model, the distribution of the frequency (x) of a mutant base at a polymorphic site is proportional to $1/x$ if the locus is evolving without selection, the population size is constant, and the mating of individuals is random (Wright 1942; Kimura 1964; Durrett 2008). A deviation of the frequency spectrum away from the $1/x$ distribution would suggest a violation of one or more of the three hypotheses. Test statistics derived from the site-frequency spectrum are widely used in population genetics studies, including tests of neutrality (Nielsen 2005; Bustamante et al. 2001; Przeworski 2002; Braverman et al. 1995; Drake et al. 2006), studies of population structure (De and Durrett 2007), investigations of demographic histories (Nei et al. 1975; Tajima 1989; Marth et al. 2004; Wakeley et al. 2001; Adams and Hudson 2004), and so on.

The distribution of the allele frequencies of polymorphic sites under selection was first derived by Fisher and Wright (Fisher 1930; Wright 1938). The theory of the site-frequency spectrum under selection was later reviewed and extended by Griffiths (2003). Using diffusion approximation (Kimura 1964; Karlin and Taylor 1981), Griffiths showed that the spectrum in a finite sample can be derived from a solution to the backward diffusion equation by assuming sampling with replacement. The theory of the site-frequency spectrum has also been extended along several other directions to consider other factors, such as varying population size (Griffiths 2003; Griffiths and Tavaré 1998; Polanski and Kimmel 2003; Evans et al. 2007), background selection, or genetic hitchhiking (Braverman et al. 1995; Fay and Wu 2000; Kim and Stephan 2002), etc.

The intensity of selection can be inferred from the observed site-frequency spectrum in a finite sample of chromosomes using the Poisson random field (PRF) model of Sawyer and Hartl (1992). The PRF model is currently the most widely used method for estimating selection directly from DNA sequence data in population genetics.

The PRF model assumes independence among sites, which greatly limits its utility to most realistic population genetics datasets. The assumption amounts to assume free recombination between polymorphic sites. However, for typical DNA sequences, polymorphic sites at the same genetic locus often segregate simultaneously because of the lack of recombination, or are completely linked in the case of haploid genomes. An analysis done by Bustamante et al. found that the selection estimated by the PRF model can be quite misleading for linked sites, and recommended to use the PRF model only for truly independent genetic variation (Bustamante et al. 2001).

Here, we extend the theory of the site-frequency spectrum to the case where the alleles are completely linked. We use diffusion approximation to derive a formula

on the joint distribution of the allele frequencies at two linked segregating sites. The technique that we use is based on constructing Green's functions (Karlin and Taylor 1981; Roach 1982) of multiallele diffusion equations with appropriate boundary conditions. Although not explored here, we believe the results presented here can be used to extend the PRF model to allow for selection estimation for dependent sites. The rest of the paper is organized as follows: In Sect. 2, we provide some basic definitions of the models. In Sect. 3, we derive the Green's functions corresponding to multiallele diffusion equations with or without selection. In Sect. 4, we calculate the mean occupation time at the diffusion boundaries. In Sect. 5, we use the results from Sects. 3 and 4 to calculate the joint distribution of the allele frequencies of linked segregating sites.

2 Basic Definitions

2.1 Wright–Fisher Model of Random Genetic Drift

Consider a genetic locus with K different alleles in a haploid population of constant size N (or a diploid population of size $N/2$) with nonoverlapping generations that undergoes random mating. The Wright–Fisher model describes the stochastic process of the genetic drift at the locus in the population as random sampling with replacement. More specifically, suppose the allele frequencies at generation t are $X(t) = (x_1(t), x_2(t), \ldots, x_K(t))$, and the relative fitness of each allele is $1 + s_k$ for the kth allele (assuming additive selection for diploids). Then the allele frequencies in the next generation will follow the multinomial distribution with parameters $p = (p_1, \ldots, p_K)$ where $p_i = x_i(t)(1 + s_i)/\sum_i x_i(t)(1 + s_i)$.

We assume that the mutation process is described by the infinite-many-sites model of Kimura in which mutations always occur at distinct sites of a DNA, and each new mutation introduces a new allele into the population. As time goes, most mutations will become either extinct or fixed in the population. Our focus is on the polymorphic sites that are neither fixed nor extinct in the present population. One specific objective is to derive the joint distribution of the allele frequencies at two or more sites conditioned on the fact that they are polymorphic.

2.2 K-allele Diffusion Approximation

Although the Wright–Fisher model provides a straightforward way for simulating genetic drift, it is not amenable to mathematical analysis. Instead, we will study the model through diffusion approximation which has a long tradition in population genetics, pioneered by Kolmogorov, Wright, Fisher, Kimura, and others (Fisher 1930; Wright 1942; Kimura 1964; Ewens 1979). In particular, when the population size is large, selection intensity is relatively weak, and the model is running at a N-generation time scale, the Wright–Fisher model can be well approximated by a multidimensional diffusion process (Ewens 1979; Durrett 2008) with infinitesimal generator

$$L = \frac{1}{2} \sum_{i,j=1}^{K} a_{ij}(x) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{K} b_i(x) \frac{\partial}{\partial x_i} \tag{1}$$

where $b(x)$ is the infinitesimal drift vector and $a(x)$ is the infinitesimal covariance matrix, and they take the following form:

$$b_i(x) = x_i \sum_{j=1}^{K} (\gamma_i - \gamma_j) x_j, \qquad a_{ij}(x) = x_i(\delta_{ij} - x_j) \tag{2}$$

where $\gamma_i = Ns_i$ is the scaled selection intensity. The diffusion model assumes values in the $(K-1)$-dimensional simplex

$$\Delta_K = \left\{ x = (x_1, \ldots, x_K) : x_1 \geq 0, \ldots, x_K \geq 0, x_1 + \cdots + x_K = 1 \right\} \tag{3}$$

so effectively (1) describes a $M = K - 1$ dimensional diffusion.

The diffusion model describes the evolution of the allele frequencies under random drift and selection, but without mutations. Watterson (1977) and Li (1977) discussed the stationary solution of the model under selection as well as mutation. Kimura (1956) and Littler and Fackerell (1975) first provided a solution on the transition density function of the diffusion process without selection and mutation. Shimakura (1977) and Griffiths (1979) later provided a solution of the transition density function in the neural case with parent-independent mutation by using explicit eigenfunction expansion. Simpler solutions for the neutral case with parent-independent mutation were later found by Griffiths and Li (1983) and Tavaré (1984) by considering a genealogical process associated with the model, and by Baxter et al. (2007) by using Jacobi polynomial expansion. Different from these previous results, our focus here is to derive a Green's function (Roach 1982; Karlin and Taylor 1981) associated with the K-allele diffusion. Although the Green's function associated with two-allele diffusion is well studied (Karlin and Taylor 1981; Durrett 2008), a general solution associated with K-allele diffusion with $K > 2$ has not been described before.

3 Green's Function of the K-allele Diffusion

Definition 1 (Green's function of the K-allele diffusion) The Green's function is the solution to

$$LG(x; x') = -\delta(x - x') \tag{4}$$

subject to the boundary condition of $G(x; x') = 0$ for all $x \in \Delta_K^b$, where $\Delta_K^b := \{x = (x_1, \ldots, x_K) : \exists k$ such that $x_k = 0, x \in \Delta_K\}$ is the boundary of the simplex Δ_K, x' is an interior point of Δ_K, and $\delta(x - x')$ is the Dirac delta function.

Before we proceed to derive a formula for the Green's function, we first provide an alternative and more intuitive interpretation of it. The following is an extension of the one-dimensional result described in the book by Durrett (2008).

Theorem 1 *Consider the random process X_t with K-dimensional infinitesimal generator L. Suppose V is a subset of $\mathbb{R}^K$, which is compact has a piecewise smooth*

boundary ∂V. *Suppose it is possible to reach* ∂V *from any interior point of* V. *Let* $\tau = \inf\{t : X_t \in \partial V\}$ *be the time of the first visit to* ∂V *when* $X(0) = x$. *Then*

$$g(x) = E_x\left[\int_0^\tau f(X_t)\,dt\right] \tag{5}$$

is the unique solution of $Lg = -f$ *for all* $x \in V$ *with the boundary condition*: $g(y) = 0$ *for all* $y \in \partial V$.

Proof Since ∂V is reachable from any interior point of V, we have $\sup_{x\in V} E_x[\tau] < \infty$, and thus $g(x)$ is well defined. Note that

$$\frac{d}{dt}E_x\left[g(X_t) + \int_0^t f(X_s)\,ds\right] = E_x\big[Lg(X_t) + f(X_t)\big] = 0 \tag{6}$$

Thus, $E_x[g(X_t) + \int_0^t f(X_s)\,ds] = C$ is a constant. Consider two cases: (a) when $t \to \infty$, $C = E_x[\int_0^\tau f(X_s)\,ds]$, and (b) when $t = 0$, $C = g(x)$. So, we must have $g(x) = E_x[\int_0^\tau f(X_t)\,dt]$. □

As a consequence, we have the following interpretation on the Green's function.

Corollary 1 *Suppose* $G(x; y)$ *is the solution of*

$$LG(x; y) = -\delta(x - y) \tag{7}$$

for all $x \in V$ *with the boundary condition of* $G(z; y) = 0$ *for all* $z \in \partial V$, *where* y *is an interior point of* V. *Then*

$$G(x; y) = \int_0^\infty p\big(X(t) = y | X(0) = x\big)\,dt \tag{8}$$

where $p(X(t) = y|X(0) = x)$ *is the transitional probability density.*

In another words, for sufficiently small δy, $G(x; y)\delta y$ is the mean occupation time of $[y, y + \delta y]$ before hitting the boundary ∂V. Given G, the solution to $Lg = -f$ can be simply written as $g(x) = \int G(x, y) f(y)\,dy$.

3.1 Change of Variables

In its present form, (4) is not easy to solve because the variables are not separable. Next, we describe a change of variables, which was first proposed by Kimura (1956) and later extended by Baxter et al. (2007), to eliminate the cross-covariance terms. We consider separately the diffusion models with or without selection.

3.1.1 Without Selection

In this case, the infinitesimal generator is

$$L = \frac{1}{2}\sum_{i,j=1}^{K} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} \tag{9}$$

The following change of variables will be able to eliminate the cross-covariance terms in L.

Lemma 1 (Change of variables in K allele diffusion without selection) *Consider the infinitesimal generator for K-allele diffusion without selection in* (9). *Change variables from $x = (x_1, \ldots, x_K)$ to $z = (z_1, \ldots, z_K)$ according to*

$$z_1 = x_1, \qquad z_2 = \frac{x_2}{1 - x_1}, \ldots, \qquad z_i = \frac{x_i}{1 - \sum_{j=1}^{i-1} x_j} \tag{10}$$

for $i = 2, \ldots, K$. Then the infinitesimal generator of the diffusion process in terms of z is

$$L = \frac{1}{2} z_1 (1 - z_1) \frac{\partial^2}{\partial z_1^2} + \frac{1}{2} \sum_{i=2}^{K-1} \frac{z_i (1 - z_i)}{\prod_{j=1}^{i-1} (1 - z_j)} \frac{\partial^2}{\partial z_i^2} \tag{11}$$

Proof Denote $\bar{b}(z)$ the new infinitesimal drift vector and $\bar{a}(z)$ the new infinitesimal covariance matrix after the change of variables. Let $D_i = \frac{\partial}{\partial x_i}$ and $D_{ij} = \frac{\partial^2}{\partial x_i \partial x_j}$. For a general transformation of variables, $\bar{b}(z)$ and $\bar{a}(z)$ are related to the original drift vector $b(x)$ and covariance matrix $a(x)$ through

$$\begin{aligned} \bar{b}_i(z) &= \sum_m D_m z_i(x) b_m(x) + \frac{1}{2} \sum_{m,n} D_{mn} z_i(x) a_{mn}(x) \\ \bar{a}_{ij}(z) &= \sum_{m,n} D_m z_i(x) D_n z_j(x) a_{mn}(x) \end{aligned} \tag{12}$$

Given the particular form $z_i(x) = x_i / (1 - \sum_{j<i} x_j)$, we have

$$\begin{aligned} D_m z_i &= \mathrm{I}(m = i) \frac{1}{1 - \sum_{j<i} x_j} + \mathrm{I}(m < i) \frac{x_i}{(1 - \sum_{j<i} x_j)^2} \\ D_{mn} z_i &= \mathrm{I}(n < m = i \text{ or } m < n = i) \frac{1}{(1 - \sum_{j<i} x_j)^2} + \mathrm{I}(m, n < i) \frac{2x_i}{(1 - \sum_{j<i} x_j)^3} \end{aligned}$$

Substituting the above to (12), we have $\bar{b}_i = 0$ for all i, that is, the drift vector stays zero. For the covariance matrix, let $u_i = 1 - \sum_{k<i} x_k$. Then we have

$$\begin{aligned} \bar{a}_{ij} = \frac{1}{u_i u_j} \big\{ & a_{ij} + z_i z_j u_i \big[u_j - \mathrm{I}(j \le i) \big] + z_i z_j u_j \big[u_i - \mathrm{I}(i \le j) \big] \\ & + z_i z_j \big[u_j (1 - u_i) \mathrm{I}(i < j) + u_i (1 - u_j) \mathrm{I}(j \le i) \big] \big\} \end{aligned}$$

So, we have $\bar{a}_{ij} = z_i (1 - z_i) / u_i$ for $i = j$ and 0 otherwise. Note that $u_i = u_{i-1}(1 - z_{i-1})$, so $u_i = \prod_{j<i} (1 - z_j)$. Thus, this completes the proof. □

3.1.2 With Selection

When the selection intensity is nonzero, the infinitesimal generator L in (1) contains interaction terms between variables in both the drift term and the covariance term. For general forms of selection intensity, the above change of variables scheme will not be able to completely separate variables in L, and consequently an explicit solution cannot be derived using this method. Note that in general the diffusion with selection is more difficult to analyze, which is reflected by the fact that no explicit solution of its transition density function is currently known. Recently, Barbour et al. (2000) and Etheridge and Griffiths (2009) derived a transition density expansion in terms of the transition functions of a dual birth-death process, which however did not yield a closed-form solution.

We consider a special case relevant to the study of the site-frequency spectrum in which the selection intensity is chosen such that $\gamma_1 = 0$ and $\gamma_i = \gamma$ for all $i > 1$. This corresponds to the scenario where the first allele is a wild type allele, and all the others are mutant types derived from the wild type, each of which has the same fitness. Alternatively, this can also represent the case where one of the alleles is under selection with intensity $-\gamma$ while all others are neutral.

With this choice of selection intensity, the infinitesimal generator for the diffusion becomes

$$L = \frac{1}{2}\sum_{i,j=1}^{K} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} - \gamma x_1(1-x_1)\frac{\partial}{\partial x_1} + \gamma x_1 \sum_{i=2}^{K} x_i \frac{\partial}{\partial x_i} \tag{13}$$

That is, the infinitesimal drift $b_1(x) = -\gamma x_1(1-x_1)$ and $b_i(x) = \gamma x_1 x_i$ for all $i > 1$.

Lemma 2 (Change of variables in K-allele diffusion with selection) *Consider the infinitesimal generator for K-allele diffusion with selection in* (13). *Change variables from $x = (x_1, \ldots, x_K)$ to $z = (z_1, \ldots, z_K)$ according to*

$$z_1 = x_1, \qquad z_2 = \frac{x_2}{1-x_1}, \ldots, \qquad z_i = \frac{x_i}{1 - \sum_{j=1}^{i-1} x_j} \tag{14}$$

Then the infinitesimal generator of the diffusion process in terms of z is

$$L = \frac{1}{2} z_1(1-z_1)\frac{\partial^2}{\partial z_1^2} + \frac{1}{2}\sum_{i=2}^{K-1} \frac{z_i(1-z_i)}{\prod_{j=1}^{i-1}(1-z_j)}\frac{\partial^2}{\partial z_i^2} - \gamma z_1(1-z_1)\frac{\partial}{\partial z_1} \tag{15}$$

Proof The proof is similar to the one presented in Lemma 1. The covariance terms stay the same. We only need to check the drift term.

$$\bar{b}_i = \sum_m D_m z_i(x) b_m(x) + \frac{1}{2}\sum_{m,n} D_{mn} z_i(x) a_{mn}(x) = \sum_m D_m z_i(x) b_m(x)$$

The case for $i = 1$ is straightforward since $\bar{b}_1 = b_1(x) = -\gamma z_1(1-z_1)$. For $i > 1$,

$$\tilde{b}_i = \sum_m \left[\frac{\mathrm{I}(m=i)}{u_i} + \frac{\mathrm{I}(m<i)x_i}{u_i^2}\right] b_m(x)$$

$$= \frac{b_i(x)}{u_i} + \frac{x_i}{u_i^2} \sum_{m=1}^{i-1} b_m(x)$$
$$= 0$$

Thus, this completes the proof. □

3.2 Green's Function of Three-allele Diffusion

For the clarity of discussion, we consider first a simple case when the total number of alleles is $K = 3$, and generalize the result to an arbitrary K in a later section. When $K = 3$, the corresponding diffusion model will be a two-dimensional diffusion, involving two free variables.

3.2.1 Without Selection

Our goal here is to find a solution to

$$\left[\frac{x_1(1-x_1)}{2}\frac{\partial^2}{\partial x_1^2} - x_1x_2\frac{\partial^2}{\partial x_1 \partial x_2} + \frac{x_2(1-x_2)}{2}\frac{\partial^2}{\partial x_2^2}\right] G(x_1, x_2; x_1', x_2')$$
$$= -\delta(x - x') \tag{16}$$

with both (x_1, x_2) and $(x_1', x_2') \in \Delta_3$ where $\Delta_3 = \{(y_1, y_2) : y_1, y_2 \in [0, 1], y_1 + y_2 \leq 1\}$, and with the boundary condition of $G(x_1, x_2; x_1', x_2') = 0$ for all x_1 and x_2 that satisfy $x_1 = 0$, $x_2 = 0$, or $x_1 + x_2 = 1$. Our approach is to expand the Green's function using orthogonal polynomials, more specifically, the Jacobi polynomials (Abramowitz and Stegun 1965) in this case.

Theorem 2 (Green's function of $K = 3$ diffusion without selection) *The Green's function corresponding to three-allele diffusion without selection, that is, the solution of the* (16), *is*

$$G(x_1, x_2; x_1', x_2') = \sum_{n=0}^{\infty} \frac{2(n+2)n!(n+2)!}{(2n+2)!} \frac{\Phi_n(x_1, x_1')}{x_1'(1-x_1')^2}$$
$$\times P_n^{(1,1)}(1-2z_2')P_n^{(1,1)}(1-2z_2)z_2(1-z_2) \tag{17}$$

where $z_2 = x_2/(1-x_1)$, $z_2' = x_2'/(1-x_1')$, $P_n^{(1,1)}$ *is the Jacobi polynomial, and the function* Φ_n *is defined as*

$$\Phi_n(x_1, x_1') = (1-x_1)^r(1-x_1')^r$$
$$\times \begin{cases} x_1 {}_2F_1(r, r+1; 2; x_1) {}_2F_1(r, r-1; 2r; 1-x_1') & \text{if } x_1 < x_1' \\ x_1' {}_2F_1(r, r+1; 2; x_1') {}_2F_1(r, r-1; 2r; 1-x_1) & \text{o.w.} \end{cases} \tag{18}$$

where $r = n + 2$ *and* ${}_2F_1$ *is the Gauss hypergeometric function.*

Proof The proof consists of the following three steps:
Step 1. Change of variables
Let $z_1 = x_1$, $z_2 = x_2/(1-x_1)$, $z_1' = x_1'$, and $z_2' = x_2'/(1-x_1')$. Then according to Lemma 1, (16) can be rewritten as

$$\left[\frac{z_1(1-z_1)}{2}\frac{\partial^2}{\partial z_1^2} + \frac{z_2(1-z_2)}{2(1-z_1)}\frac{\partial^2}{\partial z_2^2}\right]G(z_1, z_2; z_1', z_2') = -\frac{1}{1-z_1}\delta(z-z') \tag{19}$$

subject to the boundary condition of $G(z_1, z_2; z_1', z_2') = 0$ for all $z_1, z_2 = 0$ or 1. Since both z_1' and z_2' are viewed as parameters of the differential equation, in the following, we will also use the notation of $G(z_1, z_2)$ to represent G.

Step 2. Expansion using orthogonal polynomials
Next, we propose to expand the dependency of G on z_2 using orthogonal polynomials

$$G(z_1, z_2) = \sum_{n=0}^{\infty} A_n(z_1) z_2(1-z_2) P_n^{(1,1)}(1-2z_2) \tag{20}$$

where $P_n^{(1,1)}$ is the n-th order Jacobi polynomial, which has the general form of $P_n^{(\alpha,\beta)}$ with $P_n^{(\alpha,\beta)}(1-2x)$ being the solution of the hypergeometric function

$$x(1-x)y'' + \big[(\alpha+1) - (\alpha+\beta+2)x\big]y' + n(n+\alpha+\beta+1)y = 0 \tag{21}$$

for $\alpha, \beta > -1$ and $x \in [0, 1]$.

Let $B_n(u) \equiv u(1-u)P_n^{(1,1)}(1-2u)$. It can be shown that B_n satisfies

$$\frac{u(1-u)}{2}B_n''(u) = -\lambda_n B_n(u) \tag{22}$$

for all $u \in [0, 1]$, where $\lambda_n = (n+1)(n+2)/2$ with $n = 0, 1, \ldots$. According to the Sturm–Liouville theory, $\{B_n(u) : n = 0, 1, \ldots\}$ forms a complete set of orthogonal basis functions for any function $f(u)$ on $u \in [0, 1]$, with f and f' being piece-wise continuous and satisfying the boundary condition $f(0) = f(1) = 0$. Thus, the expansion of $G(z_1, z_2)$ in terms of (20) is always possible.

Substituting (20) to (19), we have

$$\sum_{m=0}^{\infty}\left[\frac{z_1(1-z_1)}{2}A_m''(z_1) - \lambda_m\frac{A_m(z_1)}{1-z_1}\right]B_m(z_2) = -\frac{1}{1-z_1}\delta(z_1-z_1')\delta(z_2-z_2') \tag{23}$$

Multiply both sides by $P_n^{(1,1)}(1-2z_2)$, take integral over z_2, and use the orthogonal property of Jacobi polynomials

$$\int_{-1}^{1}\big(1-u^2\big)P_n^{(1,1)}(u)P_m^{(1,1)}(u)\,du = \frac{8(n+1)}{(n+2)(2n+3)}\delta_{nm} \tag{24}$$

We find that $A_n(z_1)$ has to satisfy

$$\begin{aligned} &\frac{z_1(1-z_1)^2}{2}A_n''(z_1) - \lambda_n A_n(z_1) = -C_n\delta(z_1 - z_1') \\ &\text{s.t.} \quad A_n(0) = A_n(1) = 0 \end{aligned} \tag{25}$$

where $C_n = \frac{(n+2)(2n+3)}{n+1}P_n^{(1,1)}(1-2z_2')$.

Step 3. Derivation of the one-dimensional Green's function
Our next step is to find a solution to (25), which is the Green's function for an one-dimensional second-order ODE.

Let $A_n(z_1) = (1-z_1)^r\phi(z_1)$ and substitute it to (25). We find that the left side of (25) becomes

$$\begin{aligned} \text{LHS} = \frac{1}{2}(1-z_1)^{r+1}\Big[&z_1(1-z_1)\phi'' - 2rz_1\phi' - r(r-1)\phi \\ &- \frac{r(r-1)-2\lambda_n}{1-z_1}\phi\Big] \end{aligned} \tag{26}$$

$$= \frac{1}{2}(1-z_1)^{r+1}\big[z_1(1-z_1)\phi'' - 2rz_1\phi' - r(r-1)\phi\big] \tag{27}$$

where the second equation holds if we choose r satisfying $r(r-1) = 2\lambda_n$, i.e., $r = n+2$. With this choice of r, function $\phi(x)$ should satisfy

$$\begin{aligned} &x(1-x)\phi'' - 2rx\phi' - r(r-1)\phi = -\frac{2C_n}{(1-x)^{r+1}}\delta(x-x') \\ &\text{s.t.} \quad \phi(0) = 0 \text{ and } \phi(1) = \text{finite} \end{aligned} \tag{28}$$

It can be shown that the two homogenous solutions of the above equation are $\phi^{(1)}(x) = x\,{}_2F_1(r, r+1; 2; x)$ and $\phi^{(2)}(x) = {}_2F_1(r, r-1; 2r; 1-x)$, where $\phi^{(1)}$ satisfies the boundary condition at $x=0$ and $\phi^{(2)}$ satisfies the boundary condition at $x=1$. Consequently, the two homogenous solutions of (25) are

$$A_1(x) = (1-x)^r x\,{}_2F_1(r, r+1; 2; x) \tag{29}$$

$$A_2(x) = (1-x)^r\,{}_2F_1(r, r-1; 2r; 1-x) \tag{30}$$

where $A_1(x)$ satisfies boundary condition at $x=0$ and $A_2(x)$ satisfies boundary condition at $x=1$. Thus, the solution of (25) is $A_n(z_1, z_1') = d\Phi_n(z_1, z_1')$, where

$$\Phi_n(z_1, z_1') = \begin{cases} A_1(z_1)A_2(z_1') & \text{if } z_1 < z_1' \\ A_1(z_1')A_2(z_1) & \text{if } z_1 \geq z_1' \end{cases} \tag{31}$$

and $d = 2C_n/[W(A_1, A_2)(z_1')z_1'(1-z_1')^2]$, where

$$W(A_1, A_2)(x) \equiv A_1'(x)A_2(x) - A_1(x)A_2'(x) \tag{32}$$

is the Wronskian function, and should be a constant since (25) does not contain the first derivative of $A_n(z_1)$. So, $W(A_1, A_2)(x) = W(A_1, A_2)(0) = A_2(0)$ for all $x \in [0, 1]$.

Note that

$$ {}_2F_1(a, b; c; 1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \tag{33} $$

holds if $c - a - b > 0$ and $c \neq 0, -1, -2, \ldots$. Thus

$$ W(A_1, A_2)(x) = {}_2F_1(r, r-1; 2r; 1) = \frac{\Gamma(2r)}{\Gamma(r)\Gamma(r+1)} = \frac{(2n+3)!}{(n+1)!(n+2)!} \tag{34} $$

for all $x \in [0, 1]$. So,

$$ A_n(z_1, z_1') = \frac{2C_n(n+1)!(n+2)!}{(2n+3)!} \frac{\Phi_n(z_1, z_1')}{z_1'(1-z_1')^2} \tag{35} $$

Substituting the expression of $A_n(z_1, z_1')$ back to (20) gives the final formula of the Green's function, and thus completes the proof. □

3.2.2 With Selection

Our goal here is to find a solution to

$$ \left[\frac{x_1(1-x_1)}{2}\frac{\partial^2}{\partial x_1^2} - x_1x_2\frac{\partial^2}{\partial x_1 \partial x_2} + \frac{x_2(1-x_2)}{2}\frac{\partial^2}{\partial x_2^2} - \gamma x_1(1-x_1)\frac{\partial}{\partial x_1} + \gamma x_1 x_2 \frac{\partial}{\partial x_2}\right] G(x_1, x_2; x_1', x_2') = -\delta(x - x') \tag{36} $$

with both (x_1, x_2) and $(x_1', x_2') \in \Delta_3$, and with the boundary condition of $G(x_1, x_2; x_1', x_2') = 0$ for all x_1 and x_2 that satisfy $x_1 = 0$, $x_2 = 0$, or $x_1 + x_2 = 1$.

Theorem 3 (Green's function of $K = 3$ diffusion with selection) *The Green's function corresponding to three-allele diffusion with selection, i.e., the solution of* (36) *is*

$$ G(x_1, x_2; x_1', x_2') = \sum_{n=0}^{\infty} \frac{2(n+2)(2n+3)}{(n+1)\bar{F}(n+2, n+1; 2(n+2); 1; \gamma^2)} \frac{e^{\gamma(x_1 - x_1')}\bar{\Phi}_n(x_1, x_1')}{x_1'(1-x_1')^2} \times P_n^{(1,1)}(1-2z_2')P_n^{(1,1)}(1-2z_2)z_2(1-z_2) \tag{37} $$

where $z_2 = x_2/(1-x_1)$, $z_2' = x_2'/(1-x_1')$, $P_n^{(1,1)}$ *is the Jacobi polynomial, and the function* $\bar{\Phi}_n$ *is defined as*

$$ \bar{\Phi}_n(x_1, x_1') = (1-x_1)^r(1-x_1')^r \times \begin{cases} x_1\bar{F}\big(r, r+1; 2; x_1; \gamma^2\big)\bar{F}\big(r, r-1; 2r; 1-x_1'; \gamma^2\big) & \text{if } x_1 \le x_1' \\ x_1'\bar{F}\big(r, r+1; 2; x_1'; \gamma^2\big)\bar{F}\big(r, r-1; 2r; 1-x_1; \gamma^2\big) & \text{o.w.} \end{cases} \tag{38} $$

with $r = n + 2$. $\bar{F}$ *is an extension of the Gauss hypergeometric function defined as*

$$\bar{F}(a, b; c; x; d) = \sum_{n=0}^{\infty} a_n x^n, \quad \textit{where}$$
$$a_{n+1} = \frac{(n+a)(n+b)a_n + d(a_{n-1} - a_{n-2})}{(n+1)(n+c)} \tag{39}$$

with $a_0 = 1$, *and* $a_n = 0$ *when* $n < 0$.

Proof The proof consists of the following three steps:
Step 1. Change of variables
Let $z_1 = x_1$, $z_2 = x_2/(1 - x_1)$, $z_1' = x_1'$, and $z_2' = x_2'/(1 - x_1')$. Then according to Lemma 2, (36) can be rewritten as

$$\left[\frac{z_1(1-z_1)}{2}\frac{\partial^2}{\partial z_1^2} + \frac{z_2(1-z_2)}{2(1-z_1)}\frac{\partial^2}{\partial z_2^2} - \gamma z_1(1-z_1)\frac{\partial}{\partial z_1}\right] G(z_1, z_2; z_1', z_2')$$
$$= -\frac{1}{1-z_1}\delta(z - z') \tag{40}$$

with the boundary condition of $G(z_1, z_2; z_1', z_2') = 0$ for all $z_1, z_2 = 0$ or 1.

Step 2. Expansion using orthogonal polynomials
Similar to the argument presented in the proof of Theorem 2, G can be expanded using orthogonal polynomials

$$G(z_1, z_2) = \sum_{n=0}^{\infty} A_n(z_1) z_2(1-z_2) P_n^{(1,1)}(1 - 2z_2) \tag{41}$$

Substituting the expanded G to (40), we have

$$\sum_{m=0}^{\infty}\left[\frac{z_1(1-z_1)}{2}A_m''(z_1) - \gamma z_1(1-z_1)A_m'(z_1) - \lambda_m\frac{A_m(z_1)}{1-z_1}\right]B_m(z_2)$$
$$= -\frac{\delta(z_1 - z_1')\delta(z_2 - z_2')}{1-z_1}$$

Multiplying both sides by $P_n^{(1,1)}(1 - 2z_2)$, taking integral over z_2, and using the orthogonal property of Jacobi polynomials, we find that $A_n(z_1)$ has to satisfy

$$\frac{z_1(1-z_1)}{2}A_n''(z_1) - \gamma z_1(1-z_1)A_m'(z_1) - \frac{\lambda_n}{1-z_1}A_n(z_1) = -\frac{C_n\delta(z_1 - z_1')}{1-z_1} \tag{42}$$
$$\text{s.t.} \quad A_n(0) = A_n(1) = 0$$

where $C_n = \frac{(n+2)(2n+3)}{n+1}P_n^{(1,1)}(1 - 2z_2')$.

Step 3. Derivation of the one-dimensional Green's function
Our next step is to find a solution to (42). Letting $A_n(z_1) = e^{\gamma z_1}\psi(z_1)$ and substituting it to (42), we have

$$\frac{z_1(1-z_1)}{2}\psi''(z_1) - \left[\frac{\gamma^2}{2}z_1(1-z_1) + \frac{\lambda_n}{1-z_1}\right]\psi = -C_n\frac{e^{-\gamma z_1}}{1-z_1}\delta(z_1 - z_1') \quad \text{s.t.} \quad \psi(0) = \psi(1) = 0 \tag{43}$$

Now, we further let $\psi(x) = (1-x)^r\phi(x)$. Then the left side of (43) becomes

$$\begin{aligned}&\frac{1}{2}(1-z_1)^r\Big[z_1(1-z_1)\phi'' - 2rz_1\phi' - \big[r(r-1) + \gamma^2 z_1(1-z_1)\big]\phi \\ &\qquad - \frac{r(r-1) - 2\lambda_n}{1-z_1}\phi\Big] \\ &= \frac{1}{2}(1-z_1)^r\big[z_1(1-z_1)\phi'' - 2rz_1\phi' - \big[r(r-1) + \gamma^2 z_1(1-z_1)\big]\phi\big]\end{aligned}$$

where the second equation holds if we choose r satisfying $r(r-1) = 2\lambda_n$, i.e., $r = n+2$.

With this choice of r, function $\phi(x)$ should satisfy

$$x(1-x)\phi'' - 2rx\phi' - \big[r(r-1) + \gamma^2 x(1-x)\big]\phi = -\frac{2C_n e^{-\gamma x}}{(1-x)^{r+1}}\delta(x - x') \quad \text{s.t.} \quad \phi(0) = 0 \text{ and } \phi(1) = \text{finite} \tag{44}$$

Our next step is to find two homogenous solutions of the above equation that satisfy the boundary conditions. For this purpose, we consider a general form of the secondary order ODE

$$x(1-x)y'' + \big[c - (a+b+1)x\big]y' - \big[ab + dx(1-x)\big]y = 0 \tag{45}$$

where a, b, c, and d are constants. Here, $x = 0$ is a regular singular point. We consider a series solution around $x = 0$. Let $y(x) = x^r\sum_{n=0}^{\infty} a_n x^n$. The indicial equation is $r(r-1+c) = 0$, so $r = 0$ or $r = 1-c$. If $c > 1$ (which is the case we will be considering), the solution corresponding to $r = 1-c$ diverges at $x = 0$, so we only consider the solution corresponding to $r = 0$. Substituting the series solution to the ODE, we find $y(x) = \bar{F}(a, b; c; x; d)$ defined in Theorem 3. The series converge for all $|x| < 1$, and reduce to the Gauss hypergeometric function when $d = 0$.

In terms of $\bar{F}$, the two homogenous solutions of (44) can be written as $\phi^{(1)}(x) = x\bar{F}(r, r+1; 2; x; \gamma^2)$ and $\phi^{(2)}(x) = \bar{F}(r, r-1; 2r; 1-x; \gamma^2)$ where $\phi^{(1)}$ satisfies the boundary condition at $x = 0$ and $\phi^{(2)}$ satisfies the boundary condition at $x = 1$. Consequently, the two homogenous solutions of (43) are

$$A_1(x) = (1-x)^r x\bar{F}\big(r, r+1; 2; x; \gamma^2\big) \tag{46}$$

$$A_2(x) = (1-x)^r\bar{F}\big(r, r-1; 2r; 1-x; \gamma^2\big) \tag{47}$$

where $A_1(x)$ satisfies boundary condition at $x = 0$ and $A_2(x)$ satisfies boundary condition at $x = 1$. Thus, the solution of (42) is $A_n(z_1, z_1') = de^{\gamma z_1}\bar{\Phi}_n(z_1, z_1')$, where

$$\bar{\Phi}_n(z_1, z_1') = \begin{cases} A_1(z_1)A_2(z_1') & \text{if } z_1 < z_1' \\ A_1(z_1')A_2(z_1) & \text{if } z_1 \geq z_1' \end{cases} \tag{48}$$

and $d = 2C_n e^{-\gamma z_1'}/[W(A_1, A_2)(z_1')z_1'(1 - z_1')^2]$, where

$$W(A_1, A_2)(x) \equiv A_1'(x)A_2(x) - A_1(x)A_2'(x) \tag{49}$$

is the Wronskian function, and should be a constant since (43) does not contain the first derivative of $\psi(z_1)$. So, $W(A_1, A_2)(x) = W(A_1, A_2)(0) = A_2(0)$ for all $x \in [0, 1]$. Thus,

$$d = \frac{2C_n}{\bar{F}(r, r-1; 2r; 1; \gamma^2)} \frac{e^{-\gamma z_1'}}{z_1'(1 - z_1')^2} \tag{50}$$

This completes the proof. □

3.3 Green's Function for General K

Next, we consider the general case of any $K \geq 3$.

Theorem 4 (Green's function of K-allele diffusion without selection) *The Green's function corresponding to the K-allele diffusion (i.e. $M = K - 1$ dimensional diffusion) without selection is*

$$G(x; x') = \sum_{l \in \mathbb{N}^{M-1}} a_l \frac{\Phi_{n_2}(z_1, z_1')}{z_1'(1 - z_1')^M} \times \prod_{j=2}^{M} \frac{z_j(1 - z_j)^{r_j} P_{l_j}^{(1, 2r_j - 1)}(1 - 2z_j) P_{l_j}^{(1, 2r_j - 1)}(1 - 2z_j')}{(1 - z_j')^{M - j - r_j + 1}} \tag{51}$$

where $z_1 = x_1$, $z_i = x_i/(1 - \sum_{j<i} x_j)$ for all $i > 2$, and similarly for $(z_1', \ldots, z_M')$. a_l is a coefficient indexed by $l = (l_2, \ldots, l_M)$ with $l_i = 0, 1, \ldots$, and is defined to be

$$a_l = 2\frac{(n_2 + 1)!(n_2 + 2)!}{(2n_2 + 3)!} \prod_{j=2}^{M} \frac{(2l_j + 2r_j + 1)(l_j + 2r_j)}{l_j + 1} \tag{52}$$

where $n_i = \sum_{j=i}^{M} l_j + M - i$ for all $i = 2, \ldots, M$, and $r_i = n_{i+1} + 2$ when $i = 1, \ldots, M - 1$ and equal to 1 when $i = M$. $P_n^{(\alpha,\beta)}$ is the Jacobi polynomial, and the function Φ_n is defined the same as in Theorem 2.

A proof of Theorem 4 can be found in the Appendix.

Theorem 5 (Green's function of K-allele diffusion with selection) *The Green's function corresponding to the K-allele diffusion (i.e., $M = K - 1$ dimensional diffusion) with selection defined in the form of (13) is*

$$G(x; x') = \sum_{l \in \mathbb{N}^{M-1}} \bar{a}_l \frac{e^{\gamma(z_1 - z_1')} \bar{\Phi}_{n_2}(z_1, z_1')}{z_1'(1 - z_1')^M} \times \prod_{j=2}^{M} \frac{z_j(1 - z_j)^{r_j} P_{l_j}^{(1,2r_j-1)}(1 - 2z_j) P_{l_j}^{(1,2r_j-1)}(1 - 2z_j')}{(1 - z_j')^{M-j-r_j+1}} \tag{53}$$

where $z_1 = x_1$, $z_i = x_i/(1 - \sum_{j<i} x_j)$ for all $i > 2$, and similarly for $(z_1', \ldots, z_M')$. $\bar{a}_l$ is a coefficient indexed by $l = (l_2, \ldots, l_M)$ with $l_i = 0, 1, \ldots$, and is defined to be

$$\bar{a}_l = \frac{2}{\bar{F}(r_1, r_1 - 1; 2r_1; 1; \gamma^2)} \prod_{j=2}^{M} \frac{(2l_j + 2r_j + 1)(l_j + 2r_j)}{l_j + 1} \tag{54}$$

where $n_i = \sum_{j=i}^{M} l_j + M - i$ for all $i = 2, \ldots, M$, and $r_i = n_{i+1} + 2$ when $i = 1, \ldots, M - 1$ and equal to 1 when $i = M$. $P_n^{(\alpha,\beta)}$ is the Jacobi polynomial, and the functions $\bar{F}$ and $\bar{\Phi}_n$ are defined as in Theorem 3.

Theorem 5 can be proved using a combination of the proofs shown for Theorems 4 and 3, and is not shown here.

4 Occupation Time at Diffusion Boundaries

For the K-allele diffusion in (1), we have considered so far only the behavior within the diffusion boundaries. However, the diffusion will eventually reach one of the boundaries, which corresponds to the extinction of one of the allele types. With the Wright–Fisher model, the genetic drift afterward will continue to be modeled with random mating and sampling with replacement, and the corresponding diffusion approximation will be a $(K - 1)$-allele diffusion.

In this section, we derive the mean occupation time spent at different points of the diffusion boundaries before any of the remaining allele types becomes further extinct. Since the Green's function corresponding to the $(K - 1)$-allele diffusion can calculate the mean occupation time at these points conditioned on a particular initial condition, the key step here is to derive the probability of hitting each entry point of the K-allele diffusion boundaries. Although the probability of the fixation of an allele or the probability of a particular sequence of extinction are well studied (Kimura 1955; Littler 1975; Ewens 1979; Durrett 2008), the problem on the probability of hitting a particular boundary point has not been thoroughly investigated before. For the simplicity of discussion, we consider only the case of $K = 3$ in the following, although the results can be generalized to any K in a straightforward manner.

4.1 Without Selection

Theorem 6 (Probability of hitting and the time of occupying different boundary points of three-allele diffusion without selection) *Consider the diffusion model describing the frequencies $X(t) = (X_1, X_2, 1 - X_1 - X_2)$ of three-alleles without selection. Suppose the initial state is $X_1(0) = x_1$ and $X_2(0) = x_2$, and let $z_2 = x_2/(1 - x_1)$ and $r = n + 2$. Then the probability density of first hitting the $X_1 = 0$ boundary at $X_2 = y$ is*

$$p_b(y; x_1, x_2) = \sum_{n=0}^{\infty} \frac{(n+2)n!(n+2)!}{(2n+2)!}(1-x_1)^r {}_2F_1(r, r-1; 2r; 1-x_1) \times P_n^{(1,1)}(1-2y) P_n^{(1,1)}(1-2z_2) z_2(1-z_2) \tag{55}$$

And the mean time occupying $X_2 \in [y, y + \delta y]$ at the $X_1 = 0$ boundary is $T(y; x_1, x_2)\delta y$, where

$$T(y; x_1, x_2) = \sum_{n=0}^{\infty} \frac{2(n+2)n!n!}{(2n+2)!}(1-x_1)^r {}_2F_1(r, r-1; 2r; 1-x_1) \times P_n^{(1,1)}(1-2z_2) z_2(1-z_2) P_n^{(1,1)}(1-2y) \tag{56}$$

Proof Apply the change of variables described in Lemma 1 to the diffusion operator of the $K = 3$ diffusion without selection. Then in terms of the z variables, the probability density function of z should satisfy the continuity equation

$$\frac{\partial p(z,t)}{\partial t} = -\sum_{i=1}^{2} \frac{\partial J_i(z,t)}{\partial z_i} \tag{57}$$

where two currents are

$$J_1(z,t) = -\frac{1}{2}\frac{\partial}{\partial z_1}\big[z_1(1-z_1)p(z,t)\big] \tag{58}$$

$$J_2(z,t) = -\frac{1}{2(1-z_1)}\frac{\partial}{\partial z_2}\big[z_2(1-z_2)p(z,t)\big] \tag{59}$$

The probability density of hitting the $z_1 = 0$ boundary at z_2 is then

$$P(z_1 = 0, z_2) = -\int_0^{\infty} J_1(z_1 = 0, z_2, t)\, dt = \frac{1}{2}\int_0^{\infty} p(z_1 = 0, z_2, t)\, dt = \frac{G(x_1, x_2; x_1' = 0, x_2' = z_2)}{2} \tag{60}$$

Using the expression of the Green's function described in Theorem 2 leads to the first part of the theorem.

To calculate the mean occupation time spend at the boundary, we note that the mean time spent at $X_2 = y$ is

$$T(y; x_1, x_2) = \int_0^1 p_b(u; x_1, x_2) G(u, y)\, du \tag{61}$$

where $G(u, y) = 2u/y I(u < y) + 2(1-u)/(1-y) I(u \geq y)$ is the one-dimensional Green's function without selection. Plug in the expression of $p_b(u; x_1, x_2)$ and note that the integral can be evaluated by using

$$\begin{aligned}
&\int_0^1 G(u, y) P_n^{(1,1)}(1-2u)\, du \\
&= \frac{1}{n+2} \frac{2}{y(1-y)} \int_0^y P_{n+1}(1-2u)\, du \\
&= \frac{1}{(n+2)(2n+3)} \frac{1}{y(1-y)} \big[P_n(1-2y) - P_{n+2}(1-2y)\big] \\
&= \frac{2}{(n+1)(n+2)} P_n^{(1,1)}(1-2y)
\end{aligned}$$

where P_n is the Legendre polynomial. This leads to the formula of the mean occupation time. □

The formula for the probability density of hitting other boundaries and the corresponding mean occupation time can be derived using a symmetry argument. In particular, according to the theorem, the probability density of first hitting the $X_1 + X_2 = 1$ boundary at $X_2 = y$ is $p_b(y; 1 - x_1 - x_2, x_2)$, and the mean times occupying $X_2 \in [y, y + \delta y]$ at the $X_1 + X_2 = 1$ boundary is $T(y; 1 - x_1 - x_2, x_2)\delta y$.

4.2 With Selection

Theorem 7 (Probability of hitting different boundary points of three-allele diffusion with selection) *Consider the diffusion model describing the frequencies $X(t) = (X_1, X_2, 1 - X_1 - X_2)$ of three-alleles with selection described in (13). Suppose the initial state is $X_1(0) = x_1$ and $X_2(0) = x_2$, and let $z_2 = x_2/(1 - x_1)$ and $r = n + 2$. Then*

(*a*) *The probability density of first hitting the $X_2 = 0$ boundary at $X_1 = y_1$ is*

$$\begin{aligned}
p_b^1(y_1; x_1, x_2) = &\sum_{n=0}^{\infty} \frac{(n+2)(2n+3)}{\bar{F}(n+2, n+1; 2(n+2); 1; \gamma^2)} \frac{e^{\gamma(x_1 - y_1)} \bar{\Phi}_n(x_1, y_1)}{y_1(1-y_1)^2} \\
&\times P_n^{(1,1)}(1-2z_2) z_2 (1-z_2)
\end{aligned} \tag{62}$$

(*b*) *The probability density of first hitting $X_1 = 0$ boundary at $X_2 = y_2$ is*

$$p_b^2(y_2; x_1, x_2) = \sum_{n=0}^{\infty} \frac{(n+2)(2n+3)}{(n+1)\bar{F}(n+2, n+1; 2(n+2); 1; \gamma^2)} e^{\gamma x_1} (1-x_1)^r$$

$$\times \bar{F}\big(r, r-1; 2r; 1-x_1; \gamma^2\big)$$
$$\times P_n^{(1,1)}(1-2y_2) P_n^{(1,1)}(1-2z_2) z_2(1-z_2) \tag{63}$$

Proof Apply the change of variables described in Lemma 2 to the diffusion operator of the $K=3$ diffusion with selection. Then in terms of the z variables, the probability density function of z should satisfy the continuity equation

$$\frac{\partial p(z,t)}{\partial t} = -\sum_{i=1}^{2} \frac{\partial J_i(z,t)}{\partial z_i} \tag{64}$$

with the currents

$$J_1(z,t) = -\frac{1}{2}\frac{\partial}{\partial z_1}\big[z_1(1-z_1)p(z,t) + \gamma z_1(1-z_1)p(z,t)\big] \tag{65}$$

$$J_2(z,t) = -\frac{1}{2(1-z_1)}\frac{\partial}{\partial z_2}\big[z_2(1-z_2)p(z,t)\big] \tag{66}$$

Thus, the probability of hitting the $X_2=0$ boundary is the total flux into $z_2=0$ boundary, which is

$$P(z_1, z_2=0) = -\int_0^\infty J_2(z_1, z_2=0, t)\,dt$$
$$= \frac{1}{2(1-z_1)}\int_0^\infty p(z_1, z_2=0, t)\,dt \tag{67}$$
$$= \frac{\bar{G}(z_1, z_2=0)}{2(1-z_1)} = \frac{1}{2}G(x_1, x_2; x_1'=z_1, x_2'=0) \tag{68}$$

where $\bar{G}(z_1, z_2)$ is the Green's function of z variables, and G is the green function in terms of x variables. And similarly, the probability of hitting the $X_1=0$ boundary is the total flux

$$P(z_1=0, z_2) = -\int_0^\infty J_1(z_1=0, z_2, t)\,dt = \frac{1}{2}\int_0^\infty p(z_1=0, z_2, t)\,dt$$
$$= \frac{1}{2}G(x_1, x_2; x_1'=0, x_2'=z_2) \tag{69}$$

Substituting the expression of the Green's function described in Theorem 3 leads to the theorem. □

Theorem 8 (Mean occupation time at different boundaries of $K=3$ diffusion with selection) *Consider the diffusion model describing the frequencies* $X(t) = (X_1, X_2, 1-X_1-X_2)$ *of three-alleles with selection described in* (13). *Suppose the initial state is* $X_1(0)=x_1$ *and* $X_2(0)=x_2$, *and let* $z_2 = x_2/(1-x_1)$ *and* $r=n+2$.

Then

(*a*) *The mean time of occupying* $X_1 \in [y_1, y_1 + \delta y]$ *at the* $X_2 = 0$ *boundary is* $T^1(y_1; x_1, x_2)\delta y$, *where*

$$T^1(y_1; x_1, x_2) = \sum_{n=0}^{\infty} \frac{(n+2)(2n+3)}{\bar{F}(n+2, n+1; 2(n+2); 1; \gamma^2)} \Psi_n(x_1, y_1; \gamma) \times P_n^{(1,1)}(1-2z_2)z_2(1-z_2) \tag{70}$$

where function $\Psi_n(x, y; \gamma) = \int_0^1 e^{\gamma(x-u)} \bar{\Phi}_n(x, u) G(u; y, -\gamma)/[u(1-u)^2]du$ *with*

$$G(u; y, -\gamma) = I(u \le y) \frac{1 - e^{2\gamma u}}{1 - e^{2\gamma}} \frac{1 - e^{2\gamma(1-y)}}{-\gamma y(1-y)} + I(u > y) \frac{e^{2\gamma u} - e^{2\gamma}}{1 - e^{2\gamma}} \frac{e^{-2\gamma y - 1}}{-\gamma y(1-y)} \tag{71}$$

(*b*) *The mean time of occupying* $X_2 \in [y_2, y_2 + \delta y]$ *at the* $X_1 = 0$ *boundary is* $T^2(y_2; x_1, x_2)\delta y$, *where*

$$T^2(y_2; x_1, x_2) = \sum_{n=0}^{\infty} \frac{2(2n+3)}{(n+1)^2 \bar{F}(n+2, n+1; 2(n+2); 1; \gamma^2)} e^{\gamma x_1}(1-x_1)^r \times \bar{F}\big(r, r-1; 2r; 1-x_1; \gamma^2\big) \times P_n^{(1,1)}(1-2z_2)z_2(1-z_2)P_n^{(1,1)}(1-2y_2) \tag{72}$$

Proof To prove (a), note that after hitting the $X_2 = 0$ boundary, the random drift of X_1 follows the one-dimensional diffusion with selection, for which the Green's function is $G(u; y, -\gamma)$ if the starting state is $X_1 = u$.

So, the overall mean time spent in $X_1 = y$ after taking into the account the probability density of hitting different points of the boundary is

$$T^1(y_1; x_1, x_2) = \int_0^1 p_b^1(u, ; x_1, x_2) G(u; y, -\gamma)\, du \tag{73}$$

Substituting into it the definition of Ψ_n leads to the result in part (a).

To prove (b), note that because the first allele is the wild-type allele, conditioned on $X_1 = 0$, the evolution of X_2 follows a one-dimensional diffusion without selection. Thus, the mean time spent in $X_2 = y_2$ is

$$T^2(y_2; x_1, x_2) = \int_0^1 p_b^2(u; x_1, x_2) G(u, y)\, du \tag{74}$$

where $G(u, y) = 2u/y I(x < y) + 2(1-u)/(1-y) I(x \ge y)$ is the one-dimensional Green's function without selection. After evaluating the integral, we derive the result in part (b). □

4.3 Simulation Results

To confirm the accuracy of the above derivations of the Green's functions and the boundary occupation times, we performed a simulation study, and compared the results obtained from the theoretical calculations to the one obtained from computer simulations.

The simulations were carried out using the Wright–Fisher model of the genetic drift of $K = 3$ alleles, with one wild type allele and two mutant alleles. Throughout the examples, the population size is chosen to be $N = 500$, and the initial frequencies of two mutant alleles are chosen to be $X_1(0) = 0.15$ and $X_2(0) = 0.5$, respectively. Because no additional mutations were introduced to the model, the population eventually converged to one of three allele types. For each Wright–Fisher run, we recorded the total time spent at each of the states when all three allele types are present in the population and the total time spent at each of the states when only two allele types are present. To obtain the mean values of the occupation times, each run was repeated 500,000 times, which took about 6 hours in a Matlab implementation.

Figure 1 shows the mean occupation time spent at each state when all three allele types are present in the population in two cases: (1) without selection (Fig. 1A, B), and (2) with selection intensity of $\gamma = 10$ for the mutation alleles (Fig. 1C, D). Plotted on top of the simulation results are calculations based on Theorems 2 and 3. The results demonstrate a good consistency between the simulation results and the theoretical calculations in both cases.

Figure 2 shows the boundary behavior, plotting the mean occupation time spent at each of the states when one of the allele becomes extinct from the population. Also, two cases are shown, without selection (Fig. 2A) or with selection (Fig. 2B). The results obtained from the calculations described in Theorems 6 and 8 are also plotted, and show a good consistency with the simulation results.

Selection intensity has a significant impact on the distribution of mean occupation time at different states. Figure 3 shows the mean occupation time of different states before reaching boundaries for different selection intensities ($\gamma = 0$, 10, 20, or 50). With the increasing of γ, the center of mass of the diffusion is clearly shifted toward the $x_1 + x_2$ boundary, corresponding to a much higher chance of the wild-type allele becoming extinct first. This effect will have notable implications on the distribution of the site-frequency spectrum for alleles under selection - namely it will lead to much higher correlations between two allele frequencies that are complementary to each other (i.e., sum to 1).

5 Site-frequency Spectrum of Linked Alleles

Next, we use the results presented in the previous sections to study the site-frequency spectrum of linked alleles. We will consider a population of N chromosomes (or N segments of DNA sequences), and assume an infinitely-many-sites model without recombination.

Consider two mutations that have occurred within the chromosomes in the past (Fig. 4). Suppose the first mutation a occurred at time t_1, and the second mutation b

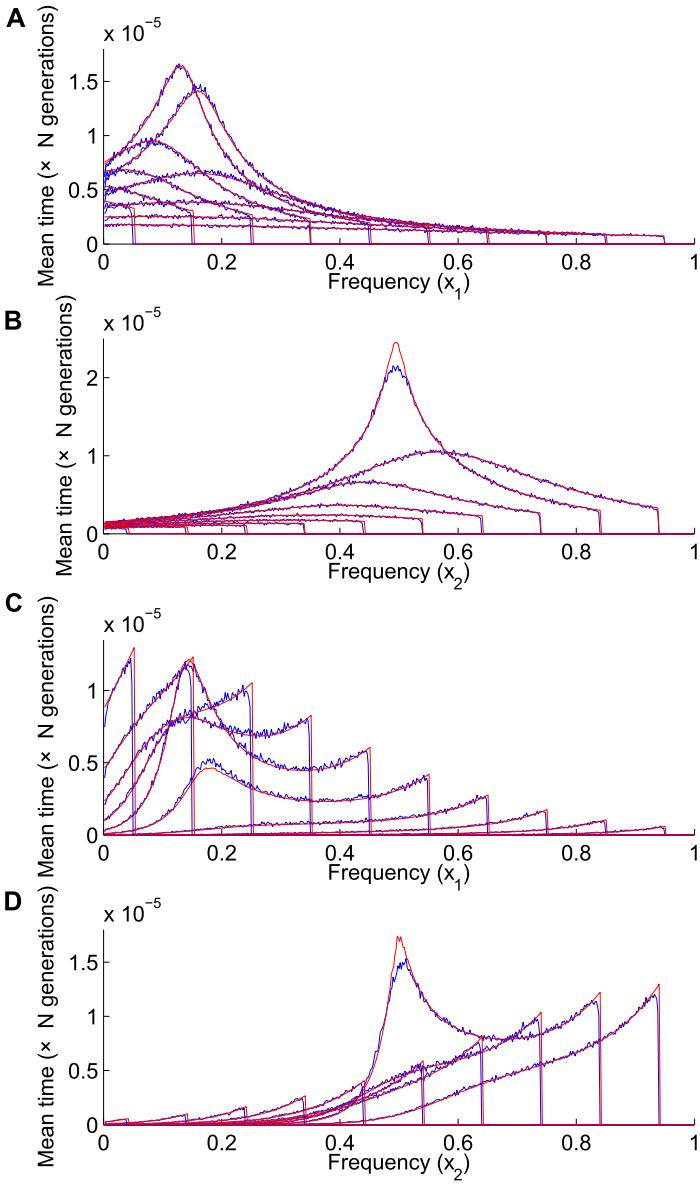

Fig. 1 Comparison of the mean occupation time obtained by simulating the Wright–Fisher model (*blue*) and the one calculated using the Green's function formula (*red*). Panels (**A**) and (**B**) show the results without selection, while panels (**C**) and (**D**) plot the results with selection ($\gamma = 10$). Panels (**A**) and (**C**) plot the mean occupation time as a function of the frequency of the first allele (x_1), while the frequency of the second allele (x_2) is fixed from the *left-to-right* direction at ten evenly distributed values between 0.05 and 0.95. Similarly, panels (**B**) and (**D**) plot the mean occupation time as a function of x_2 while x_1 is fixed. Number of different alleles $K = 3$, with the first two representing the mutant types and the third representing the wild type. Population size $N = 500$, and the initial states are $x_1 = 0.15$ and $x_2 = 0.5$. (Color figure online)

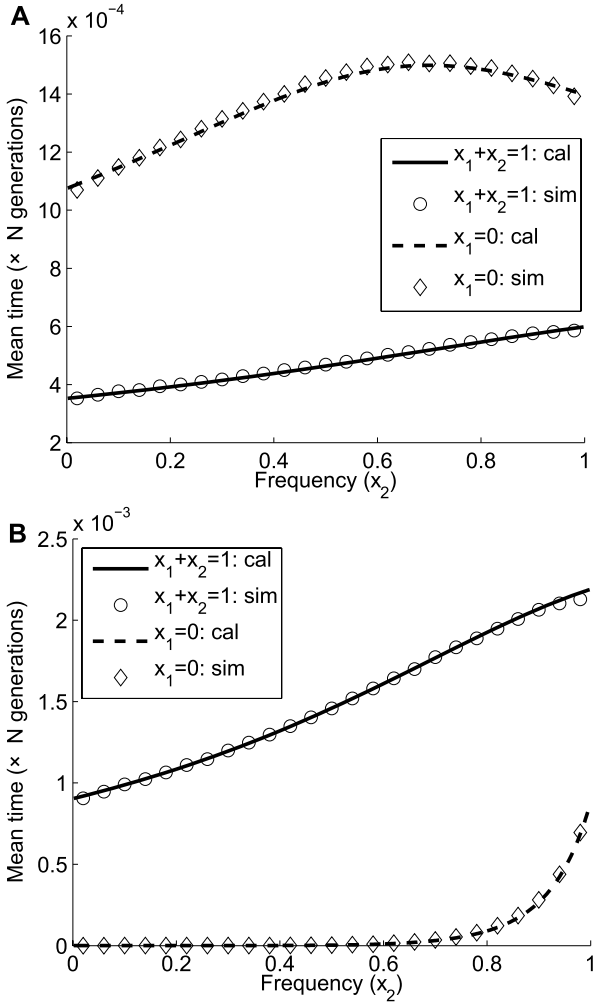

Fig. 2 Comparison of the mean occupation time at the boundaries by simulating the Wright–Fisher model (*circles* and *diamonds*) and the one calculated using the Green's function formula (*lines*). Panel (**A**) shows the results without selection, while panel (**B**) plots the results with selection ($\gamma = 10$). Number of different alleles $K = 3$, with the first two representing the mutant types and the third representing the wild type. Population size $N = 500$, and the initial states are $x_1 = 0.15$ and $x_2 = 0.5$. Two boundaries are shown including the $x_1 + x_2 = 1$ boundary (*solid lines*) and the $x_1 = 0$ boundary (*dashed lines*)

occurred at time t_2 with $t_1 < t_2 < 0$ (measured in the unit of N generations). In terms of these two mutations, the chromosomes in the present population can be classified into four allele types shown in Table 1. We use $X_1(t)$, $X_2(t)$, $X_3(t)$, and $X_4(t)$ to denote the population frequencies of the four alleles.

According to the infinitely-many-sites model, the two mutations must have occurred at different sites of the chromosomes. Suppose both sites are polymorphic in the present population with frequencies p_1 and p_2 at the sites corresponding to mutation a and b respectively. Our first goal is to calculate the joint distribution of p_1 and p_2 conditioned on the fact that $0 < p_1, p_2 < 1$.

Next, we describe how to calculate the joint distribution of p_1 and p_2 in two separate cases depending on in which allele type the second mutation occurred. In each case, we further consider four evolutionary scenarios according to the selection intensity associated with the mutant alleles.

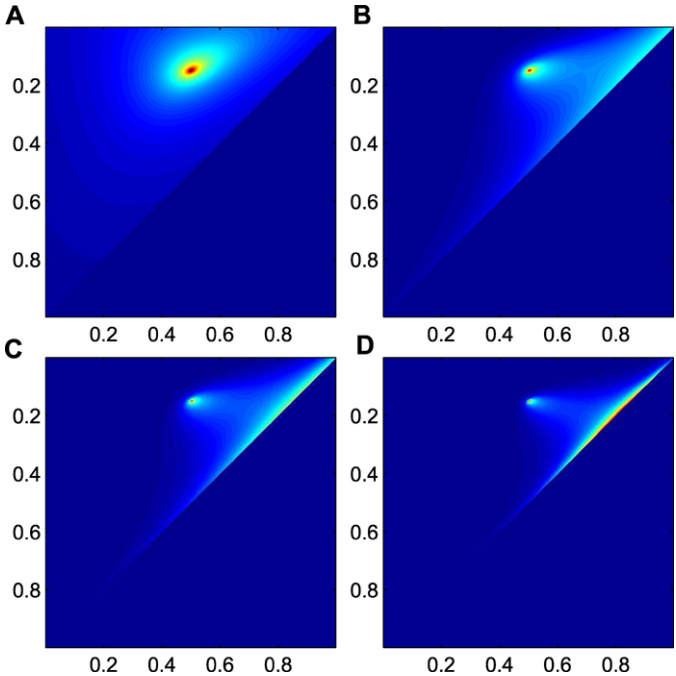

Fig. 3 Comparison of the Green's functions with different selection intensities. Number of different alleles $K = 3$, with the first two representing the mutant types and the third representing the wild type. Population size $N = 500$, and the initial states are $x_1 = 0.15$ and $x_2 = 0.5$. Shown here are the image representations of the Green's functions with values from high-to-low encoded by pseudo-colors from *red*-to-*blue*. The selection intensity γ is equal to 0 (**A**), 10 (**B**), 20 (**C**), and 50 (**D**). (Color figure online)

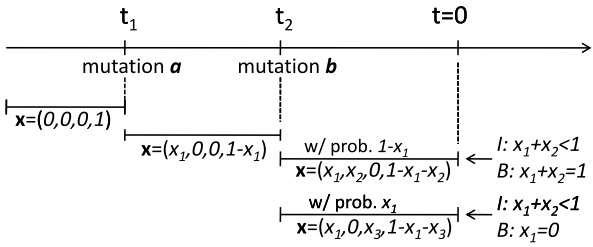

Fig. 4 A diagram of two mutation events and the corresponding allele types. Mutation a and b occurred at time t_1 and t_2, respectively. The mutation b can create a new allele type A_2 if the mutation occurred within the wild type allele, or A_3 if the mutation occurred within the A_1 allele. The two cases correspond to two different boundary requirements, with the boundary of $x_1 + x_2 = 1$ considered for the first case and the boundary of $x_1 = 0$ considered for the second case

5.1 Joint Distribution of the Allele Frequencies of Two Polymorphic Sites

At the time when the second mutation occurred, the population consists of two allele types (A_1 or A_4). Depending upon which allele type the second mutation landed on, the newly derived allele can be either (1) an allele carrying mutation b only (A_2), or (2) an allele carrying both mutation a and b (A_3). The chance of each case depends

Table 1 List of four possible allele types after two mutations. Four selection scenarios are considered according to the selection intensity of the mutant alleles: (1) $\gamma_1 = \gamma_2 = \gamma_3 = 0$; (2) $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$; (3) $\gamma_1 = \gamma_3 = \gamma$, $\gamma_2 = 0$; and (4) $\gamma_2 = \gamma_3 = \gamma$, $\gamma_1 = 0$

Allele	Frequency	Allele type	Selection
A_1	X_1	Carry mutation *a* only	γ_1
A_2	X_2	Carry mutation *b* only	γ_2
A_3	X_3	Carry both mutation *a* and *b*	γ_3
A_4	X_4	Wild type, no mutation	0

on the frequency of the A_1 allele at the time t_2, with the probability of seeing the first case being $1 - X_1(t_2)$ and the second case being $X_1(t_2)$. Because the two cases involve two different boundary requirements, next we consider them separately.

5.1.1 Case 1: The Second Mutation Occurred Within the Wild Type Allele

In this case, the alleles present in the population after mutation *b* has occurred can only be A_1, A_2, or A_4 (Table 1). Conditioned on both of the mutation sites being polymorphic, the alleles present at time $t = 0$ must include both A_1 and A_2 alleles. But there is no constraint on the frequency of the wild type allele A_4, which can either be present, or become extinct in the present population. This means that we will need to consider both the occupation time of different states within the boundary of the $K = 3$ diffusion, and the occupation time at the boundary of $X_1 + X_2 = 1$.

Let $\delta = 1/N$ denote the frequency of a mutant when it first appeared. Denote $P(X_1(t_2) = u|X_1(t_1) = \delta)$ the transition probability of X_1 from δ at t_1 to u at t_2, and $P(X_1(0) = x_1, X_2(0) = x_2|X_1(t_2) = u, X_2(t_2) = \delta)$ the transition probability of X_1 and X_2 from u and δ at t_2 to x_1 and x_2 at time 0. Then the probability of $X_1(0) = x_1$ and $X_2(0) = x_2$ is proportional to

$$f_1(x_1, x_2) = \int_{-\infty}^{0} dt_2 \int_{-\infty}^{t_2} dt_1 \int_0^1 du(1-u)P\big(X_1(t_2) = u|X_1(t_1) = \delta\big) \times P\big(X_1(0) = x_1, X_2(0) = x_2|X_1(t_2) = u, X_2(t_2) = \delta\big) \tag{75}$$

after integrating over all possible mutation times, and the intermediate value of $X_1(t_2)$. Using Theorem 1, f_1 can be rewritten as

$$f_1(x_1, x_2) = \int_0^1 (1-u)G(\delta; u)T_1(u, \delta; x_1, x_2)\, du \tag{76}$$

where $G(\delta; u)$ is the Green's function of the one-dimensional diffusion, and $T_1(u, \delta; x_1, x_2)$ represents the mean time spent in $X_1 = x_1$ and $X_2 = x_2$ when the starting frequencies are u and δ respectively. $T_1(u, \delta; x_1, x_2)$ consists of two components: one corresponds to the case of $x_1 + x_2 < 1$ (i.e., the Green's function calculated inside the boundary of the diffusion), and the other one corresponds to the case of $x_1 + x_2 = 1$ (i.e., diffusion along the boundary of $x_1 + x_2 = 1$).

Both G and T_1 can be calculated explicitly as described in the previous sections. We consider the following four scenarios according to the selection intensity associated with the two mutant alleles, denoted by γ_1 and γ_2 for allele A_1 and A_2, respectively. Let T_I^N and T_B^N denote the functions defined respectively in (17) and (56), and let T_γ^I, T_γ^B, and $T_\gamma^{B'}$ denote the functions defined in (37), (72), and (70), respectively.

1. $\gamma_1 = \gamma_2 = 0$ (Both mutant alleles are neutral.). The one-dimensional Green's function $G(\delta; u) = 2\delta/u$, and function T_1 can be expressed as

$$T_1(u, \delta; x_1, x_2) = T_I^N(\delta, u; x_2, x_1) + T_B^N(x_2; 1 - u - \delta, \delta)\delta(x_1 + x_2 - 1) \quad (77)$$

   The $\delta(\cdot)$ function is used to constrain x_1 and x_2 to be sum 1 along the boundary.
2. $\gamma_1 = \gamma_2 = \gamma$ (Both alleles are under selection.). The one-dimensional Green's function $G(\delta; u) = 2\delta(1 - e^{-2\gamma(1-y)})/[(1 - e^{-2\gamma})y(1 - y)]$, and T_1 can be expressed as

$$\begin{aligned} T_1(u, \delta; x_1, x_2) = {} & T_\gamma^I(1 - u - \delta, \delta; 1 - x_1 - x_2, x_2) \\ & + T_\gamma^B(x_2; 1 - u - \delta, \delta)\delta(x_1 + x_2 - 1) \end{aligned} \quad (78)$$

3. $\gamma_1 = \gamma$ and $\gamma_2 = 0$ (Allele A_1 is under selection and A_2 is neutral.). The one-dimensional Green's function corresponding to A_1 is $G(\delta; u) = 2\delta(1 - e^{-2\gamma(1-y)})/[(1 - e^{-2\gamma})y(1 - y)]$. To use the Green's function formulas derived above, notice that the selection intensity vector $(\gamma, 0, 0)$ associated with the three alleles can be equivalently represented as $(0, -\gamma, -\gamma)$. Consequently, function T_1 can be expressed as

$$T_1(u, \delta; x_1, x_2) = T_{-\gamma}^I(u, \delta; x_1, x_2) + T_{-\gamma}^{B'}(x_1; u, 1 - u - \delta)\delta(x_1 + x_2 - 1) \quad (79)$$

4. $\gamma_1 = 0$ and $\gamma_2 = \gamma$ (Allele A_1 is neutral and A_2 is under selection.). The one-dimensional Green's function corresponding to A_1 is $G(\delta; u) = 2\delta/u$. Function T_1 can be expressed as

$$T_1(u, \delta; x_1, x_2) = T_{-\gamma}^I(\delta, u; x_2, x_1) + T_{-\gamma}^{B'}(x_2; \delta, 1 - u - \delta)\delta(x_1 + x_2 - 1) \quad (80)$$

5.1.2 Case 2: The Second Mutation Occurred Within the A_1 Allele

In this case, the alleles present in the population after mutation b can only be A_1, A_3, or A_4. The population frequencies of the two mutation sites at $t = 0$ are $X_1(0) + X_2(0)$ and $X_2(0)$, respectively. Conditioned on the fact that both mutation sites are polymorphic, the alleles present at $t = 0$ must include both A_3 and A_4, that is, $X_3(0) > 0$ and $X_4(0) > 0$. The A_4 allele must be present because otherwise the first mutation site would appear fixed in the population. However, $X_1(0)$ can be zero because the first mutation site will be polymorphic as long as A_3 is present.

Similar to the argument presented in Case 1, the probability of $X_1(0) = x_1$ and $X_3(0) = x_3$ is proportional to

$$f_2(x_1, x_3) = \int_0^1 uG(\delta; u)T_2(u, \delta; x_1, x_3)\, du \quad (81)$$

where $T_2(u, \delta; x_1, x_3)$ represents the mean time spent in $X_1 = x_1$ and $X_3 = x_3$ when the starting frequencies are u and δ, respectively. $T_2(u, \delta; x_1, x_2)$ also consists of two components: one corresponds to the case of $x_1 + x_2 < 1$, as in Case 1, and the other one corresponds to the case of $x_1 = 0$, corresponding to the extinction of the A_1 allele.

Again, T_2 can be calculated explicitly using the results from the previous sections. Both the one-dimensional Green's functions and the Green's function corresponding to diffusion inside the boundaries are the same as those described in Case 1. The only difference is the contribution resulting from the different boundary. We consider the following four scenarios according to the selection intensity associated with the two mutant alleles, denoted by γ_1 and γ_3 for allele A_1 and A_3, respectively.

1. $\gamma_1 = \gamma_3 = 0$ (Both alleles are neutral.). $G(\delta; u) = 2\delta/u$, and function T_2 is

$$T_2(u, \delta; x_1, x_3) = T_I^N(\delta, u - \delta; x_3, x_1) + T_B^N(x_3; u - \delta, \delta)\delta(x_1) \tag{82}$$

2. $\gamma_1 = \gamma_3 = \gamma$ (Both alleles are under selection.). $G(\delta; u) = 2\delta(1 - e^{-2\gamma(1-y)})/[(1 - e^{-2\gamma})$, and function T_2 is

$$\begin{aligned} T_2(u, \delta; x_1, x_3) = {} & T_\gamma^I(1 - u, \delta; 1 - x_1 - x_3, x_3) \\ & + T_\gamma^{B'}(1 - x_3; 1 - u, u - \delta)\delta(x_1) \end{aligned} \tag{83}$$

   Note that this scenario can arise in two situations. One corresponds to the choice of $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$, for which we assume that the epistatic interactions between the two mutation sites are antagonistic. One example of such a scenario is the two loss-of-function mutations occurred at two nearby sites within the same gene. The second situation corresponds to the choice of $\gamma_1 = \gamma_3 = \gamma$ and $\gamma_2 = 0$, that is, only the alleles carrying the first mutation are under selection.
3. $\gamma_1 = \gamma$ and $\gamma_3 = 0$ (Allele A_1 is under selection and A_3 is neutral.). $G(\delta; u) = 2\delta(1 - e^{-2\gamma(1-y)})/[(1 - e^{-2\gamma})$, and function T_2 is

$$T_2(u, \delta; x_1, x_3) = T_{-\gamma}^I(u - \delta, \delta; x_1, x_3) + T_{-\gamma}^B(x_3; u - \delta, \delta)\delta(x_1) \tag{84}$$

4. $\gamma_1 = 0$ and $\gamma_3 = \gamma$ (Allele A_1 is neutral and A_3 is under selection.). $G(\delta; u) = 2\delta/u$, and function T_2 is

$$T_2(u, \delta; x_1, x_3) = T_{-\gamma}^I(\delta, u - \delta; x_3, x_1) + T_{-\gamma}^{B'}(x_3; \delta, u - \delta)\delta(x_1) \tag{85}$$

5.1.3 Joint Frequency Distribution

Combining the two cases described above, we conclude that the probability of the frequencies of the two polymorphic sites being p_1 and p_2 is equal to

$$g(p_1, p_2) = f_1(p_1, p_2) + f_2(p_1 - p_2, p_2)I(p_1 > p_2) \tag{86}$$

up to a difference in normalization constant.

In many cases, we are interested in the allele frequency distribution of two segregating sites within a sample of n chromosomes. Suppose the population size is large,

then the chromosomes can be approximated as sampling from the population with replacement (Griffiths 2003). So, for the Case 1 considered above, the frequency of observing b_1 A_1 alleles and b_2 A_2 alleles should be proportional to

$$q_1(b_1, b_2) = \int_0^1 dy_1 \int_0^{1-y_1} dy_2 \frac{n!}{b_1! b_2! (n - b_1 - b_2)!} \\ y_1^{b_1} y_2^{b_2} (1 - y_1 - y_2)^{n-b_1-b_2} f_1(y_1, y_2) \tag{87}$$

And similarly, we can find the sample frequency distribution of the A_1 and A_3 alleles considered in Case 2. Combining them, we can then derive the joint frequency distribution $q(b_1, b_2)$ of two segregating sites within a given sample, and symmetrize q when the order of mutations is unknown.

Table 2 shows the joint distribution on the allele frequencies of two segregating sites in a sample of size $n = 8$, calculated for both Cases 1 and 2, when all mutant alleles are neutral or under selection with equal intensity. Note the high probabilities associated with the antidiagonal entries in Case 1, which is even more prominent for models with selection (Table 2c). This reflects a significant contribution from the occupation time spent at the boundary of the diffusion, corresponding to the extinction of the wild type allele in the population. The joint distribution after combining Cases 1 and 2 is shown in Table 3. Note that the combined distribution is mostly dominated by the contribution from Case 1, with a ratio of 2.94 between Case 1 and Case 2 for the neutral case and 3.78 for the selection case.

We used Matlab to calculate the Jacobi polynomials and hypergeometric functions. We found the series in the Green's functions converge quickly, and used only the first 100 terms to evaluate the functions. The integrations were done numerically using the trapezoidal method. Overall, the computation is fast and the results reported here can be found within a few seconds using a modest laptop.

5.2 Site-frequency Spectrum Covariance

Suppose there are S segregating sites in a sample of n chromosomes. Let $(u_1, u_2, \ldots, u_S)$ denote the frequency of mutant allele at each of these sites. Denote η_k the number of sites where the mutant allele has frequency k, i.e., $\eta_k = \sum_{i=1}^{S} I(u_i = k)$. The above calculations can also be used to calculate the summary statistics of η_k. In particular, the mean of η_k is

$$E[\eta_k] = \sum_{i=1}^{S} E\big[I(u_i) = k\big] = SE\big[I(u_i) = k\big] = Sq_k \tag{88}$$

where q_k is the marginal distribution of $q(b_1, b_2)$, the joint frequency distribution of two sites. And the covariance is

$$\mathrm{Var}[\eta_k \eta_l] = E[\eta_k \eta_l] - S^2 q_k q_l = \big(S^2 - S\big) q_{kl} + Sq_k(\delta_{kl} - Sq_l) \tag{89}$$

Table 4 shows the covariance matrix of the site frequency spectrum in a sample of size $n = 8$ that contains $S = 10$ segregating sites. Note that cross-covariance terms are all negative except those entries at the antidiagonal. And the positive correlations at the antidiagonal entries increase significantly when the selection is introduced.

Table 2 Joint distribution of the allele frequencies of two segregating sites in a sample of size $n = 8$ for two selection models: (a, b) Without selection, and (c, d) with equal selection intensity ($\gamma_1 = \gamma_2 = \gamma_3 = \gamma$) in both Cases 1 and 2. b_1 and b_2 denote the number of samples carrying the first and the second mutation, respectively

$b_1 \setminus b_2$	1	2	3	4	5	6	7
1	0.1701	0.0549	0.0262	0.0151	0.0097	0.0071	0.0295
2	0.1030	0.0375	0.0192	0.0116	0.0082	0.0312	0
3	0.0687	0.0270	0.0146	0.0096	0.0331	0	0
4	0.0480	0.0200	0.0118	0.0356	0	0	0
5	0.0342	0.0156	0.0388	0	0	0	0
6	0.0249	0.0434	0	0	0	0	0
7	0.0517	0	0	0	0	0	0

(a) Without selection: Case 1

$b_1 \setminus b_2$	1	2	3	4	5	6	7
1	0.1418	0	0	0	0	0	0
2	0.0492	0.0934	0	0	0	0	0
3	0.0492	0.0210	0.0727	0	0	0	0
4	0.0492	0.0210	0.0123	0.0606	0	0	0
5	0.0492	0.0210	0.0123	0.0083	0.0523	0	0
6	0.0492	0.0210	0.0123	0.0083	0.0061	0.0463	0
7	0.0492	0.0210	0.0123	0.0083	0.0061	0.0047	0.0417

(b) Without selection: Case 2

$b_1 \setminus b_2$	1	2	3	4	5	6	7
1	0.0255	0.0111	0.0070	0.0056	0.0054	0.0101	0.0843
2	0.0195	0.0099	0.0071	0.0065	0.0113	0.0889	0
3	0.0170	0.0099	0.0081	0.0129	0.0942	0	0
4	0.0163	0.0110	0.0151	0.1006	0	0	0
5	0.0172	0.0187	0.1085	0	0	0	0
6	0.0256	0.1188	0	0	0	0	0
7	0.1340	0	0	0	0	0	0

(c) With selection: Case 1 ($\gamma = 10$)

$b_1 \setminus b_2$	1	2	3	4	5	6	7
1	0.0459	0	0	0	0	0	0
2	0.0320	0.0254	0	0	0	0	0
3	0.0404	0.0132	0.0198	0	0	0	0
4	0.0534	0.0178	0.0083	0.0184	0	0	0
5	0.0752	0.0257	0.0120	0.0067	0.0198	0	0
6	0.1148	0.0404	0.0193	0.0108	0.0068	0.0243	0
7	0.1916	0.0702	0.0342	0.0194	0.0121	0.0082	0.0339

(d) With selection: Case 2 ($\gamma = 10$)

Table 3 Joint distribution of the allele frequencies of two segregating sites in a sample of size $n = 8$ after combining Cases 1 and 2 for two selection models: (a) without selection, and (b) with equal selection intensity. The distribution is symmetrized to account for the situation where the order in which the two mutations occurred is unknown

$b_1 \setminus b_2$	1	2	3	4	5	6	7
1	0.1629	0.0651	0.0416	0.0298	0.0226	0.0182	0.0365
2	0.0651	0.0517	0.0199	0.0145	0.0115	0.0305	0.0027
3	0.0416	0.0199	0.0293	0.0095	0.0284	0.0016	0.0016
4	0.0298	0.0145	0.0095	0.0419	0.0011	0.0011	0.0011
5	0.0226	0.0115	0.0284	0.0011	0.0133	0.0008	0.0008
6	0.0182	0.0305	0.0016	0.0011	0.0008	0.0118	0.0006
7	0.0365	0.0027	0.0016	0.0011	0.0008	0.0006	0.0106

(a) Without selection

$b_1 \setminus b_2$	1	2	3	4	5	6	7
1	0.0298	0.0154	0.0137	0.0142	0.0168	0.0261	0.1064
2	0.0154	0.0132	0.0081	0.0088	0.0145	0.0864	0.0073
3	0.0137	0.0081	0.0106	0.0119	0.0814	0.0020	0.0036
4	0.0142	0.0088	0.0119	0.0834	0.0007	0.0011	0.0020
5	0.0168	0.0145	0.0814	0.0007	0.0041	0.0007	0.0013
6	0.0261	0.0864	0.0020	0.0011	0.0007	0.0051	0.0009
7	0.1064	0.0073	0.0036	0.0020	0.0013	0.0009	0.0071

(b) With selection ($\gamma = 10$)

6 Conclusion

In conclusion, we have used diffusion approximation to derive an analytical formula on the distribution of allele frequencies of linked polymorphic sites for models with or without selection. The main technique we use is based on constructing the Green's functions of the multiallele diffusion equations by expanding the Green's functions with orthogonal polynomials. We used numerical simulations to confirm the theoretical calculations. We found that the allele frequencies of linked sites are highly correlated, more prominently between the frequencies that are complementary to each other, and the correlation between complementary frequencies can be significantly affected by the selection intensities associated with mutant alleles.

In this paper, we have focused our analysis on the joint distribution of the allele frequencies of two linked sites. However, the Green's function results obtained here can be further used to derive the joint distribution on the frequencies of more than two sites, and thus allow us to consider even higher order statistics.

The site-frequency spectrum covariance can also be derived using the coalescence approximation when there is no selection (Fu 1995). However, for models with selection, the coalescence method is not applicable, and thus cannot be used to extend the Poisson random field model for estimating selection intensities. Our methods based

Table 4 The covariance matrix of the site-frequency spectrum in a sample of size $n = 8$ that contains $S = 10$ segregating sites

C	1	2	3	4	5	6	7
1	3.0500	−1.0761	−0.8490	−0.7155	−0.6241	−0.5340	0.7487
2	−1.0761	2.0077	−0.5385	−0.4293	−0.3377	0.8984	−0.5245
3	−0.8490	−0.5385	1.5849	−0.3013	0.9269	−0.4564	−0.3665
4	−0.7155	−0.4293	−0.3013	2.5129	−0.4370	−0.3485	−0.2811
5	−0.6241	−0.3377	0.9269	−0.4370	0.9787	−0.2802	−0.2267
6	−0.5340	0.8984	−0.4564	−0.3485	−0.2802	0.9092	−0.1885
7	0.7487	−0.5245	−0.3665	−0.2811	−0.2267	−0.1885	0.8387

(a) Without selection

C	1	2	3	4	5	6	7
1	0.2803	−1.3240	−1.1017	−0.9428	−0.7607	−0.2781	4.1270
2	−1.3240	0.4552	−0.8378	−0.7109	−0.3628	3.6336	−0.8533
3	−1.1017	−0.8378	0.5391	−0.3589	3.5537	−0.9145	−0.8798
4	−0.9428	−0.7109	−0.3589	4.6919	−0.8954	−0.8926	−0.8913
5	−0.7607	−0.3628	3.5537	−0.8954	0.2735	−0.8959	−0.9124
6	−0.2781	3.6336	−0.9145	−0.8926	−0.8959	0.3055	−0.9580
7	4.1270	−0.8533	−0.8798	−0.8913	−0.9124	−0.9580	0.3678

(b) With selection ($\gamma = 10$)

on diffusion approximation offer several additional advantages. First, we are able to derive the full joint distribution of allele frequencies rather than just correlations. Second, it can be easily generalized to derive higher order correlations.

Although not explored here, the formula we have derived here should be able to be utilized to extend the Poisson random field model for estimating selection intensities. A straightforward extension would be to consider a Markov model that accounts for the correlation between neighboring sites. It would be interesting to investigate how the extended model can improve the selection estimation.

A factor we have not considered in this paper is the recombination between polymorphic sites, which can have a significant effect on the site-frequency spectrum (Hill and Robertson 2009; Przeworski et al. 2001). Recombination can be incorporated into the diffusion model by adding the effect of recombination into the drift term as follows (Ohta and Kimura 1969; Durrett 2008). Consider two segregating loci, each having two alleles separated by recombination with a probability r per generation. Let $X = (X_1, X_2, X_3, X_4)$ denote the frequencies of the four haplotypes (aB, Ab, ab, AB) (Table 1). Assume a Wright–Fisher model with random union of gametes. If the model is run at a rate of N generations, then X can be approximated by a $K = 4$ diffusion in (1) with the infinitesimal drift vector $b_i(x) = x_i \sum_{j=1}^{K} (\gamma_i - \gamma_j) x_j + RD(x)\eta_i$ and the infinitesimal covariance vector $a_{ij}(x) = x_i(\delta_{ij} - x_j)$, where $\eta = (-1, -1, 1, 1)$, $R = Nr$, and $D(x) = x_3x_4 - x_1x_2$, representing linkage disequilibrium. A future direction would be to extend the tech-

niques developed here to calculate the Green's function of the diffusion equation with recombination. There are, however, significant obstacles, since the diffusion equation is not easily amenable to mathematical analysis after recombination is introduced. More likely, numerical methods would be needed to obtain a solution.

Acknowledgement This work was partially supported by National Science Foundation grant DBI-0846218.

Appendix: Proof of Theorem 4

To prove the theorem, we first introduce two lemmas.

Lemma A.1 *Consider the following differential equation*:

$$x(1-x)y'' - \frac{\mu}{1-x}y' = \nu y \tag{90}$$

where $x \in [0,1]$, $\mu \geq 0$, *and the boundary condition is* $y(0) = y(1) = 0$. *A solution exists only when* ν *takes one of the following eigenvalues*

$$\nu = -(n+r)(n+r+1) \tag{91}$$

for $n = 0, 1, \ldots,$ *where* $r = (1+\sqrt{1+4\mu})/2$. *And the corresponding eigenfunction is* $x(1-x)^r P_n^{(1,2r-1)}(1-2x)$.

Proof Let $y(x) = (1-x)^r g(x)$. Then

$$\text{LHS} = (1-x)^r \left(x(1-x)g'' - 2rxg' + \left[\frac{r(r-1)-\mu}{1-x} - \big[r(r-1)+\nu\big] \right] g \right) \tag{92}$$

If $r = \frac{1+\sqrt{1+4\mu}}{2}$, then $r(r-1) = \mu$. Thus

$$x(1-x)g'' - 2rxg' - \big[r(r-1)+\nu\big]g = 0 \tag{93}$$

subject to boundary condition $g(0) = 0$ if $r > 0$. Let $g(x) = xf(x)$, then

$$x(1-x)f'' + \big[2-(2+2r)x\big]f' - \big[r(r+1)+\nu\big]f = 0 \tag{94}$$

subject to boundary condition $f(0) =$ finite and $f(1) =$ finite. The solution is the Gauss hypergeometric function $f(x) = {}_2F_1(a,b;c;x)$, with $c = 2; a+b = 2r+1; ab = r(r+1)+\nu$. Note that because $c-a-b = 1-2r \leq -1$, ${}_2F_1(a,b;c;x)$ diverges at $x = 1$ unless the series is finite. In this case, the Gauss series reduces to a polynomial of degree n in x when a or b is equal to $-n$, $(n = 0, 1, 2, \ldots)$. In another word, $\nu = -n(n+2r+1) - r(r+1) = -(n+r)(n+r+1)$. In this case, ${}_2F_1(a,b;c;x) = \frac{1}{n+1}P_n^{(1,2r-1)}(1-2x)$. □

A special case of the Lemma is when $\mu = (m+1)(m+2)$ with $m = 0, 1, \ldots$. Then $r = m+2$ and $\nu = -(n+m+2)(n+m+3)$ for all $n = 0, 1, \ldots$.

Lemma A.2 *Let*

$$L_M = \sum_{i=1}^{M} \frac{z_i(1-z_i)}{2\prod_{j<i}(1-z_j)} \frac{\partial^2}{\partial z_i^2} \tag{95}$$

with $M \geq 1$. Suppose f is an eigenfunction of L_M, that is, $L_M f = \lambda f(z)$, with boundary condition $f(z) = 0$ for all z with $z_i = 0$ or 1 for all $i = 0, \ldots, M$. Let $l_j = 0, 1, \ldots$ for all $j = 1, \ldots, M$, and $n_i = \sum_{j=i}^{M} l_j + M - i$ for all $i = 1, \ldots, M$. Then the eigenvalues, indexed by $(l_1, \ldots, l_M)$, are

$$\lambda_{l_1,\ldots,l_M} = -\frac{(n_1+1)(n_1+2)}{2} \tag{96}$$

And the corresponding eigenfunction is

$$\phi_{(l_1,\ldots,l_M)}(z_1, \ldots, z_M) = \prod_{j=1}^{M} z_j(1-z_j)^{r_j} P_{l_j}^{(1,2r_j-1)}(1-2z_j) \tag{97}$$

where $r_j = n_{j+1} + 2$ with $n_{M+1} \equiv -1$.

Proof We prove the lemma by induction.

When $M = 1$, the eigenvalues are $\lambda_{l_1} = -(l_1+1)(l_1+2)/2$ and the corresponding eigenfunction is $z_1(1-z_1)P_{l_1}^{(1,1)}(1-2z_1)$. So, the result holds.

Now suppose the lemma is true for $M = m$. Then $L_{m+1} f$ becomes

$$L_{m+1} f = \frac{z_1(1-z_1)}{2} \frac{\partial^2}{\partial z_1^2} f + \frac{1}{1-z_1} L_m(z_2, \ldots, z_M) f \tag{98}$$

Let $f = g(z_1)\phi_{(l_2,\ldots,l_M)}(z_2, \ldots, z_M)$. Then

$$\begin{aligned} L_{m+1}[f] &= \left[\frac{z_1(1-z_1)}{2} g''(z_1) + \frac{\lambda_{l_2,\ldots,l_M}}{1-z_1} g(z_1)\right] \phi_{(l_2,\ldots,l_M)}(z_2, \ldots, z_M) \\ &= -\frac{(l_1+r)(l_1+r+1)}{2} \\ &\quad \times z_1(1-z_1)^r P_{l_1}^{(1,2r-1)}(1-2z_1)\phi_{(l_2,\ldots,l_M)}(z_2, \ldots, z_M) \end{aligned}$$

where $r = \frac{1+\sqrt{1-8\lambda_{l_2,\ldots,l_M}}}{2} = n_2 + 2$. In addition, $n_1 = l_1 + r - 1 = l_1 + n_2 + 1$. Thus, the result also holds true for $M = m+1$. By induction, the result has to be true for all $M > 1$. □

Proof of Theorem 4

Proof To prove the theorem, we go through three steps.

Step 1. Change of variables
Let $z_1 = x_1$ and $z_i = x_i/(1-\sum_{j<i} x_j)$ for all $i > 1$. Then in terms of z_i, (4) can be rewritten as

$$\left[\frac{z_1(1-z_1)}{2}\frac{\partial^2}{\partial z_1^2} + \sum_{i=2}^{M} \frac{z_i(1-z_i)}{2\prod_{j<i}(1-z_j)}\frac{\partial^2}{\partial z_i^2}\right] G(z;z') = -\frac{1}{\prod_{i=1}^{M}(1-z_i)^{M-i}}\delta(z-z') \tag{99}$$

subject to boundary condition

$$G(z;z') = 0 \quad \text{for all } z \text{ with } z_i = 0 \text{ or } 1 \text{ for any } i \tag{100}$$

Step 2. Expansion using orthogonal polynomials
We expand the Green's function in variable $z_2, \ldots, z_M$ in terms of orthogonal polynomials

$$G(z_1, z_2, \ldots, z_M) = \sum_{l\in\mathbb{N}^{M-1}} A_l(z_1)\phi_l(z_2, \ldots, z_M) \tag{101}$$

where $l = (l_2, \ldots, l_M)$ is a $M-1$ dimensional index with each $l_i = 0, 1, \ldots$, that is, $l_i \in \mathbb{N}$. $\phi_l(z_2, \ldots, z_M) = \prod_{j=2}^{M} z_j(1-z_j)^{r_j} P_{l_j}^{(1,2r_j-1)}(1-2z_j)$.

Substituting it to (99), we have

$$\sum_{l\in\mathbb{N}^{M-1}} \left[\frac{z_1(1-z_1)}{2}A_l''(z_1) + \frac{\lambda_l}{1-z_1}A_l\right]\phi_l(z_2, \ldots, z_M) = -\frac{1}{\prod_{i=1}^{M}(1-z_i)^{M-i}}\delta(z-z') \tag{102}$$

where $\lambda_l = -(n_2+1)(n_2+2)/2$ with $n_2 = \sum_{j=2}^{M} l_j + M - 2$. Multiply both sides by $\prod_{j=2}^{M}(1-z_j)^{r_j-1}P_{l_j'}^{(1,2r_j-1)}(1-2z_j)$, integrate over $z_2, \ldots, z_M$, and use the orthogonal property of Jacobi polynomials

$$\int_0^1 x^\alpha(1-x)^\beta P_m^{(\alpha,\beta)}(1-2x)P_n^{(\alpha,\beta)}(1-2x)\,dx = \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{(2n+\alpha+\beta+1)\Gamma(n+\alpha+\beta+1)n!}\delta_{nm}$$

We have

$$\begin{aligned} &\frac{z_1(1-z_1)}{2}A_l''(z_1) + \frac{\lambda_l}{1-z_1}A_l = -C_l\delta(z_1 - z_1') \\ &\text{s.t} \quad A_l(0) = A_l(1) = 0 \end{aligned} \tag{103}$$

where

$$C_l = \prod_{j=2}^{M} \frac{(2l_j + 2r_j + 1)(l_j + 2r_j)}{l_j + 1} \frac{P_{l_j}^{(1,2r_j-1)}(1 - 2z_j')}{(1 - z_1')^{M-1} \prod_{j=2}^{M} (1 - z_j')^{M-j-r_j+1}} \tag{104}$$

Step 3. Derivation of the one-dimensional Green's function
Our next step is to find a solution to (103), which is the Green's function for a one-dimensional second-order ODE.

Let $A_l(z_1) = (1 - z_1)^r \phi(z_1)$ and substitute it to (103). We find that the left side of (103) becomes

$$\begin{aligned} \text{LHS} &= \frac{1}{2}(1 - z_1)^r \left[z_1(1 - z_1)\phi'' - 2rz_1\phi' - r(r-1)\phi - \frac{r(r-1) - 2\lambda_l}{1 - z_1}\phi \right] \\ &= \frac{1}{2}(1 - z_1)^r \left[z_1(1 - z_1)\phi'' - 2rz_1\phi' - r(r-1)\phi \right] \end{aligned}$$

where the second equation holds if we choose r satisfying $r(r-1) = 2\lambda_l$, i.e., $r = n_2 + 2$.

With the choice of r, function $\phi(x)$ is the solution of

$$\begin{aligned} &x(1 - x)\phi'' - 2rx\phi' - r(r-1)\phi = -\frac{2C_l}{(1 - x)^r}\delta(x - x') \\ &\text{s.t.} \quad \phi(0) = 0 \text{ and } \phi(1) = \text{finite} \end{aligned} \tag{105}$$

It can be shown that the two homogenous solutions of the above equation are $\phi^{(1)}(x) = x\,{}_2F_1(r, r+1; 2; x)$ and $\phi^{(2)}(x) = {}_2F_1(r, r-1; 2r; 1 - x)$, where $\phi^{(1)}$ satisfies the boundary condition at $x = 0$ and $\phi^{(2)}$ satisfies the boundary condition at $x = 1$. Consequently, the two homogenous solutions of (103) are

$$A_1(x) = (1 - x)^r x\,{}_2F_1(r, r+1; 2; x) \tag{106}$$

$$A_2(x) = (1 - x)^r\,{}_2F_1(r, r-1; 2r; 1 - x) \tag{107}$$

where $A_1(x)$ satisfies boundary condition at $x = 0$ and $A_2(x)$ satisfies boundary condition at $x = 1$. Thus, the solution of (103) is $A_l(z_1, z_1') = d\phi_l(z_1, z_1')$, where

$$\phi_l(z_1, z_1') = \begin{cases} A_1(z_1)A_2(z_1') & \text{if } z_1 < z_1' \\ A_1(z_1')A_2(z_1) & \text{if } z_1 > z_1' \end{cases} \tag{108}$$

and $d = 2C_l/[W(A_1, A_2)(z_1')z_1'(1 - z_1')]$, where

$$W(A_1, A_2)(x) \equiv A_1'(x)A_2(x) - A_1(x)A_2'(x) \tag{109}$$

is the Wronskian function, and should be a constant since (103) does not contain the first derivative of $A_n(z_1)$. So, $W(A_1, A_2)(x) = W(A_1, A_2)(0) = A_2(0)$ for all $x \in [0, 1]$, that is,

$$W(A_1, A_2)(x) = {}_2F_1(r, r-1; 2r; 1) = \frac{\Gamma(2r)}{\Gamma(r)\Gamma(r+1)}$$

$$= \frac{(2n_2+3)!}{(n_2+1)!(n_2+2)!} \tag{110}$$

for all $x \in [0, 1]$. So,

$$A_l(z_1) = \frac{2C_l(n_2+1)!(n_2+2)!}{(2n_2+3)!} \frac{\phi_l(z_1, z_1')}{z_1'(1-z_1')} \tag{111}$$

Substituting the expression of $A_l(z_1)$ back to (101) gives the final formula of the Green's function, and thus completes the proof. □

References

Abramowitz, M., Stegun, I., 1965. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Courier Dover, New York.

Adams, A., Hudson, R., 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics 168*(3), 1699.

Barbour, A., Ethier, S., Griffiths, R., 2000. A transition function expansion for a diffusion model with selection. Ann. Appl. Probab., 123–162.

Baxter, G., Blythe, R., McKane, A., 2007. Exact solution of the multi-allelic diffusion model. *Math. Biosci. 209*(1), 124–170.

Braverman, J., Hudson, R., Kaplan, N., Langley, C., Stephan, W., 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics 140*(2), 783.

Bustamante, C., Wakeley, J., Sawyer, S., Hartl, D., 2001. Directional selection and the site-frequency spectrum. *Genetics 159*(4), 1779.

De, A., Durrett, R., 2007. Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics 176*(2), 969.

Drake, J., Bird, C., Nemesh, J., Thomas, D., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S., Dermitzakis, E., et al., 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet. 38*(2), 223–227.

Durrett, R., 2008. *Probability Models for DNA Sequence Evolution*. Springer, Berlin.

Etheridge, A., Griffiths, R., 2009. A coalescent dual process in a Moran model with genic selection. Theor. Popul. Biol.

Evans, S., Shvets, Y., Slatkin, M., 2007. Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol. 71*(1), 109–119.

Ewens, W., 1979. *Mathematical Population Genetics*. Springer, New York.

Fay, J., Wu, C., 2000. Hitchhiking under positive Darwinian selection. *Genetics 155*(3), 1405.

Fisher, R., 1930. The distribution of gene ratios for rare mutations. In: Proc. R. Soc. Edinb., vol. 50, pp. 204–219.

Fu, Y., 1995. Statistical properties of segregating sites. *Theor. Popul. Biol. 48*(2), 172–197.

Griffiths, R., 1979. A transition density expansion for a multi-allele diffusion model. *Adv. Appl. Probab. 11*(2), 310–325.

Griffiths, R., 2003. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor. Popul. Biol. 64*(2), 241–251.

Griffiths, R., Li, W., 1983. Simulating allele frequencies in a population and the genetic differentiation of populations under mutation pressure. *Theor. Popul. Biol. 23*(1), 19.

Griffiths, R., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. *Stoch. Models 14*(1), 273–295.

Hill, W., Robertson, A., 2009. The effect of linkage on limits to artificial selection. *Genet. Res. 8*(03), 269–294.

Karlin, S., Taylor, H., 1981. *A Second Course in Stochastic Processes*. Academic Press, New York.

Kim, Y., Stephan, W., 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics 160*(2), 765.

Kimura, M., 1955. Random genetic drift in multi-allelic locus. *Evolution 9*(4), 419–435.

Kimura, M., 1956. Random genetic drift in a tri-allelic locus; exact solution with a continuous model. *Biometrics 12*(1), 57–66.

Kimura, M., 1964. Diffusion models in population genetics. *J. Appl. Probab. 1*(2), 177–232.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics 61*(4), 893.

Li, W., 1977. Maintenance of genetic variability under mutation and selection pressures in population. *Proc. Natl. Acad. Sci. USA 74*(6), 2509–2513.

Littler, R., 1975. Loss of variability at one locus in a finite population. *Math. Biosci. 25*(1–2), 151–163.

Littler, R., Fackerell, E., 1975. Transition densities for neutral multi-allele diffusion models. *Biometrics 31*(1), 117–123.

Marth, G., Czabarka, E., Murvai, J., Sherry, S., 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics 166*(1), 351.

Nei, M., Maruyama, T., Chakraborty, R., 1975. The bottleneck effect and genetic variability in populations. *Evolution 29*(1), 1–10.

Nielsen, R., 2005. Molecular signatures of natural selection. *Annu. Rev. Genet. 39*, 197–218.

Ohta, T., Kimura, M., 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics 63*(1), 229.

Polanski, A., Kimmel, M., 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics 165*(1), 427.

Przeworski, M., 2002. The signature of positive selection at randomly chosen loci. *Genetics 160*(3), 1179.

Przeworski, M., Wall, J., Andolfatto, P., 2001. Recombination and the frequency spectrum in Drosophila melanogaster and Drosophila simulans. *Mol. Biol. Evol. 18*(3), 291.

Roach, G., 1982. *Green's Functions*. Cambridge University Press, Cambridge.

Sawyer, S., Hartl, D., 1992. Population genetics of polymorphism and divergence. *Genetics 132*(4), 1161.

Shimakura, N., 1977. Equations differentielles provenant de la genetique des populations. *Tohoku Math. J. 29*, 287.

Tajima, F., 1989. The effect of change in population size on DNA polymorphism. *Genetics 123*(3), 597.

Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol. 26*(2), 119–164.

Wakeley, J., Nielsen, R., Liu-Cordero, S., Ardlie, K., 2001. The discovery of single-nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet. 69*(6), 1332–1347.

Watterson, G., 1977. Heterosis or neutrality? *Genetics 85*(4), 789.

Wright, S., 1938. The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. USA 24*(7), 253.

Wright, S., 1942. Statistical genetics and evolution. *Bull. Am. Math. Soc. 48*(4), 223–246.