Analytic Computation of the Expectation of the Linkage Disequilibrium Coefficient r^2

Yun S. Song[a] and Jun S. Song[b]

[a]Department of Computer Science, University of California, Davis, CA 95616, USA
[b]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA
E-mail addresses: yssong@cs.ucdavis.edu (Y.S. Song), jssong@jimmy.harvard.edu (J.S. Song)

Corresponding Author:
Yun S. Song
Department of Computer Science
University of California at Davis
2063 Kemper Hall
One Shields Avenue
Davis, CA 95616
U.S.A.

E-mail: yssong@cs.ucdavis.edu
Phone: +1 530 754 8577
Fax: +1 530 752 4767

To appear in Theoretical Population Biology

Abstract

The squared correlation coefficient r^2 (sometimes denoted Δ^2) is a measure of linkage disequilibrium that is widely used, but computing its expectation $\mathbb{E}[r^2]$ in the population has remained an intriguing open problem. The expectation $\mathbb{E}[r^2]$ is often approximated by the standard linkage deviation σ_d^2, which is a ratio of two expectations amenable to analytic computation. In this paper, a method of computing the population-wide $\mathbb{E}[r^2]$ is introduced for a model with recurrent mutation, genetic drift and recombination. The approach is algebraic and is based on the diffusion process approximation. In the limit as the population-scaled recombination rate ρ approaches ∞, it is shown rigorously that the asymptotic behavior of $\mathbb{E}[r^2]$ is given by $1/\rho + O(\rho^{-2})$, which, incidentally, is the same as that of σ_d^2. A computer software that computes $\mathbb{E}[r^2]$ numerically is available upon request.

Keywords: Linkage disequilibrium, squared correlation coefficient, expectation, diffusion approximation, recurrent mutation, genetic drift, recombination

1 Introduction

Linkage disequilibrium (LD), which characterizes the statistical nonindependence of alleles at different loci, depends on population history and structure, and on various evolutionary forces such as recombination, mutation and natural selection—see (Hudson, 2001a) for a general review of LD, and (Pritchard and Przeworski, 2001; International HapMap Consortium, 2005) for studies of LD in the human genome. A question that is of great interest is how one can use LD to learn about the aforementioned factors that influence its extent and patterns. Several different, albeit related, measures of LD between a pair of loci have been proposed. A commonly used measure is r^2 (defined below), which is equal to the square of the correlation coefficient between the alleles at two loci (Hill and Robertson, 1968) and is related to the power of association in gene mapping (see Pritchard and Przeworski 2001 for an explanation of this relationship).

In this paper, we consider a diallelic two-locus model for a single panmictic diploid population subject to recurrent mutation and genetic drift. This model is the same as that considered by Ohta and Kimura (1969b). The effective population size, denoted N_e, is assumed to remain constant in time. We use A and B to denote the two loci, with allele types A_0, A_1 and B_0, B_1, respectively. The marginal allele frequencies of A_0 and B_0 are respectively denoted by p and q, and the LD measure r^2 is defined as

$$r^2 := \frac{D^2}{p(1-p)q(1-q)}, \tag{1}$$

where $D = f_{00} - pq$, with f_{00} being the frequency of the gametic type A_0B_0. (Sometimes Δ^2 is used to denote the above LD measure.) We consider the following symmetric reversible mutation model:

$$A_0 \underset{u}{\overset{u}{\leftrightarrows}} A_1 \quad \text{and} \quad B_0 \underset{u}{\overset{u}{\leftrightarrows}} B_1,$$

where u is the mutation rate per generation. The recombination rate between the two loci per generation is denoted by c, and the population-scaled mutation and recombination rates are denoted by $\theta := 8N_e u$ and $\rho := 4N_e c$, respectively.

In his seminal work, Golding (1984) derived a system of recursion relations satisfied by the probabilities of sample configurations, to study the sampling distribution of LD under an infinitely-many-alleles model. This approach was further studied by Ethier and Griffiths (1990), who constructed a two-locus urn model for the distribution of sample configurations. A closed-form solution to the system of recursion relations is not known, but accurate numerical solutions can be obtained for moderate sample sizes (say, up to about 40. See Hudson 2001a). For larger sample sizes, numerically solving the recursion relations becomes intractable, but coalescent simulations can be employed to study the sampling distribution and sample estimated expectations of LD, as done by Hudson (1985, 2001b). A further study of the sampling properties of LD was carried out by Hill and Weir (1994), who used a general forward simulation approach that can easily incorporate

biologically important features such as selection. In our work, we do not consider the sampling distribution or other sampling properties of LD, but focus on the *population-wide* expectation of r^2 at equilibrium. (Roughly, the case we consider corresponds to having a very large sample size.) In what follows, p, q, and D denote population quantities.

Computing the population expectation $\mathbb{E}[r^2]$ of the squared correlation coefficient r^2 is a difficult problem, and no analytic approach has been suggested so far. In contrast, as shown by Hill and Robertson (1968) and by Ohta and Kimura (1969a,b), the ratio $\mathbb{E}[D^2]/\mathbb{E}[p(1-p)q(1-q)]$ of two expectations — called the standard linkage deviation and often denoted σ_d^2 — is much easier to tackle and closed-form formulae can be obtained for various models. Furthermore, under the neutral model, σ_d^2 admits a nice genealogical interpretation in terms of covariances in coalescence times (McVean, 2002). The expectation of ratios and the ratio of expectations are, however, of course not the same. Indeed, a previous simulation-based study, in the context of an infinitely-many-alleles model, has shown that σ_d^2 may overestimate $\mathbb{E}[r^2]$ by a substantial amount (Maruyama, 1982), although σ_d^2 is a reasonably good approximation of the sample estimated expectation of r^2 conditioned on minor allele frequencies being above 5% (Hudson, 1985). In this paper, we readdress this classic issue and develop an algebraic method of computing the expectation $\mathbb{E}[r^2]$.

As in the work of Ohta and Kimura (1969a,b), our method is formulated in the context of the diffusion process approximation, which is continuous in both time and space. Diffusion theory can be used to generate useful linear equations satisfied by certain expectations at stationarity, a state in which the effects of mutation, genetic drift, and recombination are in balance. The general idea behind our approach is to express $\mathbb{E}[r^2]$ in terms of expectations that can be computed by solving appropriate systems of linear equations arising from diffusion theory.

We have written a *C* program, called *ER2*, that solves the required systems of linear equations numerically. *ER2* is available upon request. For given values of θ and ρ, *ER2* can produce numerical estimates of $\mathbb{E}[r^2]$. We remark that the accuracy of the estimates depends on the chosen "truncation level" $\ell_{\max}$, described later in the text; $\ell_{\max} = 700$ gives very accurate answers, with the running time of a few minutes on a laptop. Our study shows that $\mathbb{E}[r^2]$ and σ_d^2 for the assumed model are quite close for large θ (say, $\theta > 4$). For small θ (say, $\theta \leq 1$), however, σ_d^2 is larger than $\mathbb{E}[r^2]$ by a substantial amount, in some cases by tens of times. If no restriction is imposed on segregation at the two loci, estimates from coalescent simulations are much closer to our theoretical computation of $\mathbb{E}[r^2]$ than they are to σ_d^2. When conditioned on segregation at both loci, however, sample estimated average r^2 increases substantially for small θ, thus resembling the behavior of σ_d^2.

Obtaining a general closed-form formula for $\mathbb{E}[r^2]$ still remains an open problem, but relevant closed-form expressions can be obtained in the limit $\rho \to \infty$. Using our approach, we rigorously show that the large ρ behavior of $\mathbb{E}[r^2]$ is given by $1/\rho + O(\rho^{-2})$. Incidentally, this asymptotic behavior of $\mathbb{E}[r^2]$ is the same as that of σ_d^2 found by Ohta and Kimura (1969a,b).

The organization of this paper is as follows. In Section 2, we describe our method of com-

puting $\mathbb{E}[r^2]$ in the context of the diffusion process approximation. In Section 3, we compare our computation of $\mathbb{E}[r^2]$ with σ_d^2 and with results from coalescent simulations. The aforementioned exact asymptotic behavior of $\mathbb{E}[r^2]$ in the large ρ limit is discussed in Section 4. In Section 5, we consider a discrete version of the assumed model and show how combinatorial techniques may be employed to compute certain expectations exactly; such computations are useful for studying the accuracy of the diffusion approximation approach. We conclude in Section 6 with an outlook on future direction.

2 Diffusion Approximation

As in (Ohta and Kimura, 1969a,b), our approach is based on diffusion approximation. Being continuous in both time and space, diffusion processes possess many nice properties not shared by discrete processes. In particular, associated to a diffusion process is a fundamental differential operator (i.e., the generator) that has a wide range of applications. As Ohta and Kimura (1969b) have shown, diffusion approximation is a powerful technique that can be used to compute certain expectations at stationarity with surprisingly little effort. In this section, we extend the work of Ohta and Kimura (1969b) to compute the expectation of r^2.

2.1 Diffusion generator

Let $\mathscr{L}$ denote the generator of a diffusion process $\boldsymbol{X}_t$ in $\mathbb{R}^n$, with t being the time parameter. Then, for f a twice continuously differentiable function with compact support, it is well known that

$$\frac{\partial}{\partial t}\mathbb{E}[f(\boldsymbol{X}_t)] = \mathbb{E}[\mathscr{L}f(\boldsymbol{X}_t)],$$

where $\mathbb{E}$ denotes the expectation with respect to the probability distribution of the diffusion process. At stationarity (i.e., $\frac{\partial}{\partial t}\mathbb{E}[f(\boldsymbol{X}_t)] = 0$), we therefore have

$$\mathbb{E}[\mathscr{L}f(\boldsymbol{X}_t)] = 0, \tag{2}$$

and this equation leads to useful algebraic relations involving various expectations. Henceforward, we refer to (2) as the *master equation*. The key idea behind our work is to choose appropriate functions f in the master equation so that expectations of interest can be computed by solving systems of linear equations.

Following Ohta and Kimura (1969a,b), we consider a diffusion process in a three-dimensional space parametrized by p, q and D (i.e., in the above notation, $\boldsymbol{X}_t = (p, q, D)_t$). The generator

corresponding to the two-locus model that we are considering in this paper is

$$\begin{aligned}\mathscr{L} &= \frac{1}{2}p(1-p)\frac{\partial^2}{\partial p^2} + \frac{1}{2}q(1-q)\frac{\partial^2}{\partial q^2} + \frac{1}{2}[p(1-p)q(1-q) + D(1-2p)(1-2q) - D^2]\frac{\partial^2}{\partial D^2} \\ &\quad + D\frac{\partial^2}{\partial p \partial q} + D(1-2p)\frac{\partial^2}{\partial p \partial D} + D(1-2q)\frac{\partial^2}{\partial q \partial D} \\ &\quad + \frac{\theta}{4}(1-2p)\frac{\partial}{\partial p} + \frac{\theta}{4}(1-2q)\frac{\partial}{\partial q} - D\left(1 + \frac{\rho}{2} + \theta\right)\frac{\partial}{\partial D}, \end{aligned} \tag{3}$$

which differs from that of Ohta and Kimura (1969b) by a factor of 2; one unit of time corresponds to $2N_e$ (rather than N_e) generations in our convention.

2.2 Reformulation of the problem: Summing over $\mathbb{E}[D^2 p^m q^n]$

The main difficulty involved in computing the expectation of r^2 comes from the fact $p(1-p)q(1-q)$ appears in the denominator. The strategy that we adopt in our work is to re-express (1) in terms of quantities whose expectations are easier to compute. First, note that

$$\frac{1}{p(1-p)q(1-q)} = \left(\frac{1}{p} + \frac{1}{1-p}\right)\left(\frac{1}{q} + \frac{1}{1-q}\right).$$

Then, we can use the convergent series expansion $1/(1-z) = \sum_{k=0}^{\infty} z^k$, where $0 \leq z < 1$, to obtain the following expressions for $0 < p < 1$:

$$\frac{1}{1-p} = \sum_{k=0}^{\infty} p^k \quad \text{and} \quad \frac{1}{p} = \frac{1}{1-(1-p)} = \sum_{k=0}^{\infty}(1-p)^k.$$

Similar results hold for $1/(1-q)$ and $1/q$. Further, since $\mathbb{E}[r^2]$ is bounded, the Lebesgue convergence theorem implies

$$\begin{aligned}\mathbb{E}[r^2] &= \mathbb{E}\left[\frac{D^2}{p(1-p)q(1-q)}\right] = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\big\{\mathbb{E}[D^2 p^m q^n] + \mathbb{E}[D^2(1-p)^m(1-q)^n] \\ &\qquad + \mathbb{E}[D^2(1-p)^m q^n] + \mathbb{E}[D^2 p^m (1-q)^n]\big\}.\end{aligned}$$

Finally, since $\mathbb{E}[D^2 p^m q^n] = \mathbb{E}[D^2(1-p)^m q^n] = \mathbb{E}[D^2 p^m (1-q)^n] = \mathbb{E}[D^2(1-p)^m(1-q)^n]$ for the assumed model of mutation,

$$\mathbb{E}[r^2] = \mathbb{E}\left[\frac{D^2}{p(1-p)q(1-q)}\right] = 4\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathbb{E}[D^2 p^m q^n]. \tag{4}$$

We have thus translated the problem of computing $\mathbb{E}[r^2]$ into a problem of computing an infinite number of expectations of form $\mathbb{E}[D^2 p^m q^n]$. As we elaborate shortly, this change in perspective is

useful for two reasons: First, $\mathbb{E}[D^2 p^m q^n]$ can be computed. Second, $4\sum_{m,n}\mathbb{E}[D^2 p^m q^n]$ converges fast as the sum $m+n$ of exponents increases, thus allowing us to truncate the summation over m and n at some appropriate level.

2.3 Warm-up exercises

Before we proceed to the computation of $\mathbb{E}[D^2 p^m q^n]$, we here demonstrate how the diffusion generator technique works by computing some simpler expectations. These expectations are used in our algorithm for computing $\mathbb{E}[D^2 p^m q^n]$.

The subsequent discussion pertains to the generator $\mathscr{L}$ given by (3). From using $f = p^n$ or $f = q^n$ in the master equation (2), it immediately follows that the expectations $\mathbb{E}[p^n]$ and $\mathbb{E}[q^n]$ satisfy

$$\mathbb{E}[p^n] = \left(\frac{\frac{\theta}{2}+n-1}{\theta+n-1}\right)\mathbb{E}[p^{n-1}] \qquad \text{and} \qquad \mathbb{E}[q^n] = \left(\frac{\frac{\theta}{2}+n-1}{\theta+n-1}\right)\mathbb{E}[q^{n-1}]. \tag{5}$$

Since $\mathbb{E}[1] = 1$, we can solve these recursions to obtain

$$\mathbb{E}[p^n] = \mathbb{E}[q^n] = \frac{\left(\frac{\theta}{2}\right)^{[n]}}{(\theta)^{[n]}}, \tag{6}$$

where $(z)^{[k]}$ denotes the *rising factorial*, defined as

$$(z)^{[k]} := z(z+1)\cdots(z+k-1). \tag{7}$$

Using $f = D$ in the master equation (2) implies $\mathbb{E}[D] = 0$, while using $f = Dp^n$, for $n \geq 1$, yields the relation $2[2+\rho+2\theta+n(3+n+\theta)]\mathbb{E}[Dp^n] = n(2+2n+\theta)\mathbb{E}[Dp^{n-1}]$. A similar relation holds for $\mathbb{E}[Dq^n]$ and $\mathbb{E}[Dq^{n-1}]$. Hence, it follows from induction that

$$\mathbb{E}[Dp^n] = \mathbb{E}[Dq^n] = 0 \text{ for all } n \geq 0. \tag{8}$$

If $f = pq$ is used, we obtain $\theta\,\mathbb{E}[pq] = \mathbb{E}[D] + \frac{\theta}{4}(\mathbb{E}[q] + \mathbb{E}[p])$. Since $\mathbb{E}[D] = 0$ and $\mathbb{E}[p] = \mathbb{E}[q] = 1/2$, we conclude that

$$\mathbb{E}[pq] = \frac{1}{4}. \tag{9}$$

Lastly, using $f = p^2 q$ yields $2(2+3\theta)\mathbb{E}[p^2 q] = 8\mathbb{E}[Dp] + \theta\,\mathbb{E}[p^2] + 2(2+\theta)\mathbb{E}[pq]$, whereas using $f = pq^2$ yields $2(2+3\theta)\mathbb{E}[pq^2] = 8\mathbb{E}[Dq] + \theta\,\mathbb{E}[q^2] + 2(2+\theta)\mathbb{E}[pq]$. Since $\mathbb{E}[Dp] = 0$, $\mathbb{E}[p^2] = \frac{\frac{\theta}{2}(\frac{\theta}{2}+1)}{\theta(\theta+1)}$, and $\mathbb{E}[pq] = \frac{1}{4}$, we conclude

$$\mathbb{E}[p^2 q] = \mathbb{E}[pq^2] = \frac{1}{4}\left(\frac{\frac{\theta}{2}+1}{\theta+1}\right). \tag{10}$$

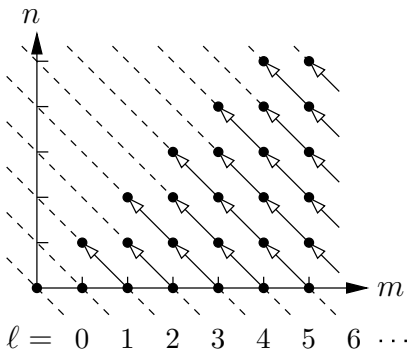

Figure 1: Pairs (m, n) corresponding to $m \geq n$, where m and n are exponents of p and q, respectively, in $\mathbb{E}[D^2 p^m q^n]$. A constant level $\ell = m + n$ is indicated by a dashed line. Within each fixed level, open arrows indicate the order in which our algorithm is carried out.

2.4 Computing $\mathbb{E}[D^2 p^m q^n]$ via solving systems of coupled linear equations

All expectations considered in the previous section were quite straightforward to compute; by using a single appropriate function f in the master equation (2), we obtained a linear equation containing the expectation that we wished to compute and other expectations that had already been computed. Computing $\mathbb{E}[D^2 p^m q^n]$ is more complicated, however, because the recursion relations now involve higher powers of D. We here present an algorithm for computing $\mathbb{E}[D^2 p^m q^n]$ systematically. It involves solving systems of coupled linear equations in a particular order.

First, note that by symmetry of the problem, p and q are exchangeable, i.e., $\mathbb{E}[D^k p^i q^j] = \mathbb{E}[D^k p^j q^i]$, for all $i, j, k \geq 0$. Without loss of generality, we may therefore assume that $m \geq n$ in $\mathbb{E}[D^2 p^m q^n]$. By a *level* we mean the sum $\ell = m + n$ of the exponents of p and q in $\mathbb{E}[D^2 p^m q^n]$. This definition is depicted in Figure 1, where the pairs (m, n) corresponding to $m \geq n$ are indicated by closed circles. Our algorithm starts from level 0 and progresses upwards in level. As illustrated in Figure 1, within each fixed level ℓ, we start from $n = 0$ and end at $n = \ell/2$ if ℓ is even or at $n = (\ell - 1)/2$ if ℓ is odd. For each pair (m, n), where $m \geq n$, we generate a system of $n + 3$ coupled linear equations by using $f = D^k p^{m+2-k} q^{n+2-k}$ in the master equation (2) for $k = 0, \ldots, n + 2$. If computations are carried out in the particular order described above, the only unknown quantities in the system of coupled linear equations thus generated will be the following $n + 3$ expectations:

$$\mathbb{E}[p^{m+2} q^{n+2}],\ \mathbb{E}[D p^{m+1} q^{n+1}],\ \mathbb{E}[D^2 p^m q^n],\ \ldots,\ \mathbb{E}[D^{2+n-1} p^{m-n+1} q],\ \mathbb{E}[D^{2+n} p^{m-n}]. \tag{11}$$

for $\ell = 0, \ldots, \ell_{\max}$ **do**
  if (ℓ is even) **then**
    set $n_{\max} = \ell/2$
  else
    set $n_{\max} = (\ell - 1)/2$
  end if
  for $n = 0, \ldots, n_{\max}$ **do**
    set $m = \ell - n$. Then,
      set up a system of $n + 3$ coupled linear equations using $f = D^k p^{m+2-k} q^{n+2-k}$
      in the master equation (2) for $0 \le k \le n + 2$, and then solve for the
      $n + 3$ expectations $\mathbb{E}[D^k p^{m+2-k} q^{n+2-k}]$, $0 \le k \le n + 2$
      if ($n \ne m$) **then**
        for $0 \le k \le n + 2$ **do**
          set $\mathbb{E}[D^k p^{n+2-k} q^{m+2-k}] = \mathbb{E}[D^k p^{m+2-k} q^{n+2-k}]$
        end for
      end if
  end for
end for

Figure 2: Algorithm for computing $\mathbb{E}[D^2 p^m q^n]$ up to some given truncation level $\ell_{\max}$.

More precisely, if $f = D^k p^i q^j$ is used in the master equation (2), then in general we obtain

$$\begin{aligned}
[k^2 + i(i-1+\theta) + j(j-1+\theta) + k(1+4i+4j+\rho+2\theta)]\,\mathbb{E}[D^k p^i q^j] \;&=\; 2ij\,\mathbb{E}[D^{k+1} p^{i-1} q^{j-1}] \\
&+\frac{1}{2}(2j^2 + j\theta + 4kj - 2j)\,\mathbb{E}[D^k p^i q^{j-1}] + \frac{1}{2}(2i^2 + i\theta + 4ki - 2i)\,\mathbb{E}[D^k p^{i-1} q^j] \\
&+k(k-1)\left(4\,\mathbb{E}[D^{k-1} p^{i+1} q^{j+1}] + \mathbb{E}[D^{k-1} p^i q^j] - 2\,\mathbb{E}[D^{k-1} p^i q^{j+1}] - 2\,\mathbb{E}[D^{k-1} p^{i+1} q^j]\right) \\
&+k(k-1)\left(\mathbb{E}[D^{k-2} p^{i+2} q^{j+2}] + \mathbb{E}[D^{k-2} p^{i+1} q^{j+1}] - \mathbb{E}[D^{k-2} p^{i+2} q^{j+1}] - \mathbb{E}[D^{k-2} p^{i+1} q^{j+2}]\right),
\end{aligned}$$

where $\mathbb{E}[D^k p^i q^j], \mathbb{E}[D^{k+1} p^{i-1} q^{j-1}], \mathbb{E}[D^{k-1} p^{i+1} q^{j+1}], \mathbb{E}[D^{k-2} p^{i+2} q^{j+2}]$ are unknown quantities. We remark that the expectations shown in (6), (9), and (10) also appear in some equations and that they are treated as known quantities. Once we solve for the $n + 3$ expectations shown in (11), we move on to the next pair (m', n') in order. A summary of the above algorithm is shown in Figure 2.

2.5 Level truncation and convergence

In (4), m and n both range from 0 to ∞. If we knew a closed-form formula for $\mathbb{E}[D^2 p^m q^n]$, then it might be possible to obtain a closed-form formula for $\mathbb{E}[r^2]$ by summing over m and n explicitly. However, it seems quite difficult to obtain a closed-form formula for $\mathbb{E}[D^2 p^m q^n]$, and therefore we have adopted a numerical approach.

We have made two independent implementations of the algorithm described in Section 2.4: one

in C and the other in *Mathematica*. Both programs are available upon request. The *Mathematica* program can compute $\mathbb{E}[D^2 p^m q^n]$ symbolically for given m and n, and can generate formulae in terms of θ and ρ. The C program is called *ER2*, and for given numerical values of θ, ρ, and $\ell_{\max}$, it computes the following level-truncated estimate of $\mathbb{E}[r^2]$:

$$\mathbb{E}[r^2]_{\ell_{\max}} := \sum_{\ell=0}^{\ell_{\max}} a_\ell, \quad \text{where} \quad a_\ell := \sum_{\substack{m,\, n \geq 0, \\ m+n=\ell}} 4\, \mathbb{E}[D^2 p^m q^n]. \tag{12}$$

Since $\mathbb{E}[r^2]$ is bounded and the sequence $\{\mathbb{E}[r^2]_{\ell_{\max}}\}_{\ell_{\max}=0}^{\infty}$ of partial sums is a monotonically increasing sequence, $\{\mathbb{E}[r^2]_{\ell_{\max}}\}_{\ell_{\max}=0}^{\infty}$ is a convergent sequence. Although an analytic expression for the rate of convergence is difficult to obtain, we have empirically observed that the sequence converges quite fast. Shown in Figure 3a and Figure 3b are plots of $\mathbb{E}[r^2]_{\ell_{\max}}$ as a function of $\ell_{\max}$ for $\rho = 1$ and $\rho = 10$, respectively. For given θ and ρ, *ER2* took a few minutes on a laptop to compute $\mathbb{E}[r^2]_{\ell_{\max}}$ up to $\ell_{\max} = 700$. The rate of convergence of $\mathbb{E}[r^2]_{\ell_{\max}}$ seems to depend on θ and ρ; $\mathbb{E}[r^2]_{\ell_{\max}}$ converges faster for smaller ρ and larger θ.

2.6 Simplification: even from odd or odd from even

The amount of computation involved in our algorithm can be reduced considerably using

$$\mathbb{E}[D^\ell p^m q^n] = (-1)^\ell\, \mathbb{E}[D^\ell p^m (1-q)^n] = (-1)^\ell\, \mathbb{E}[D^\ell (1-p)^m q^n],$$

which is valid under the assumed model of mutation. This simple observation implies the following set of relations:

$$[1-(-1)^n]\ \mathbb{E}[D^{2k} p^m q^n] = \sum_{j=0}^{n-1} \binom{n}{j} (-1)^j\, \mathbb{E}[D^{2k} p^m q^j], \tag{13}$$

$$[1-(-1)^m]\ \mathbb{E}[D^{2k} p^m q^n] = \sum_{j=0}^{m-1} \binom{m}{j} (-1)^j\, \mathbb{E}[D^{2k} p^j q^n], \tag{14}$$

$$[1+(-1)^n]\ \mathbb{E}[D^{2k+1} p^m q^n] = -\sum_{j=0}^{n-1} \binom{n}{j} (-1)^j\, \mathbb{E}[D^{2k+1} p^m q^j], \tag{15}$$

$$[1+(-1)^m]\ \mathbb{E}[D^{2k+1} p^m q^n] = -\sum_{j=0}^{m-1} \binom{m}{j} (-1)^j\, \mathbb{E}[D^{2k+1} p^j q^n]. \tag{16}$$

Relations (13) and (14) allow us to express $\mathbb{E}[D^{2k} p^m q^n]$ for odd m or odd n purely in terms of $\mathbb{E}[D^{2k} p^i q^j]$, where $i \leq m$ and $j \leq n$ are both even. In a similar vein, relations (15) and (16) allow

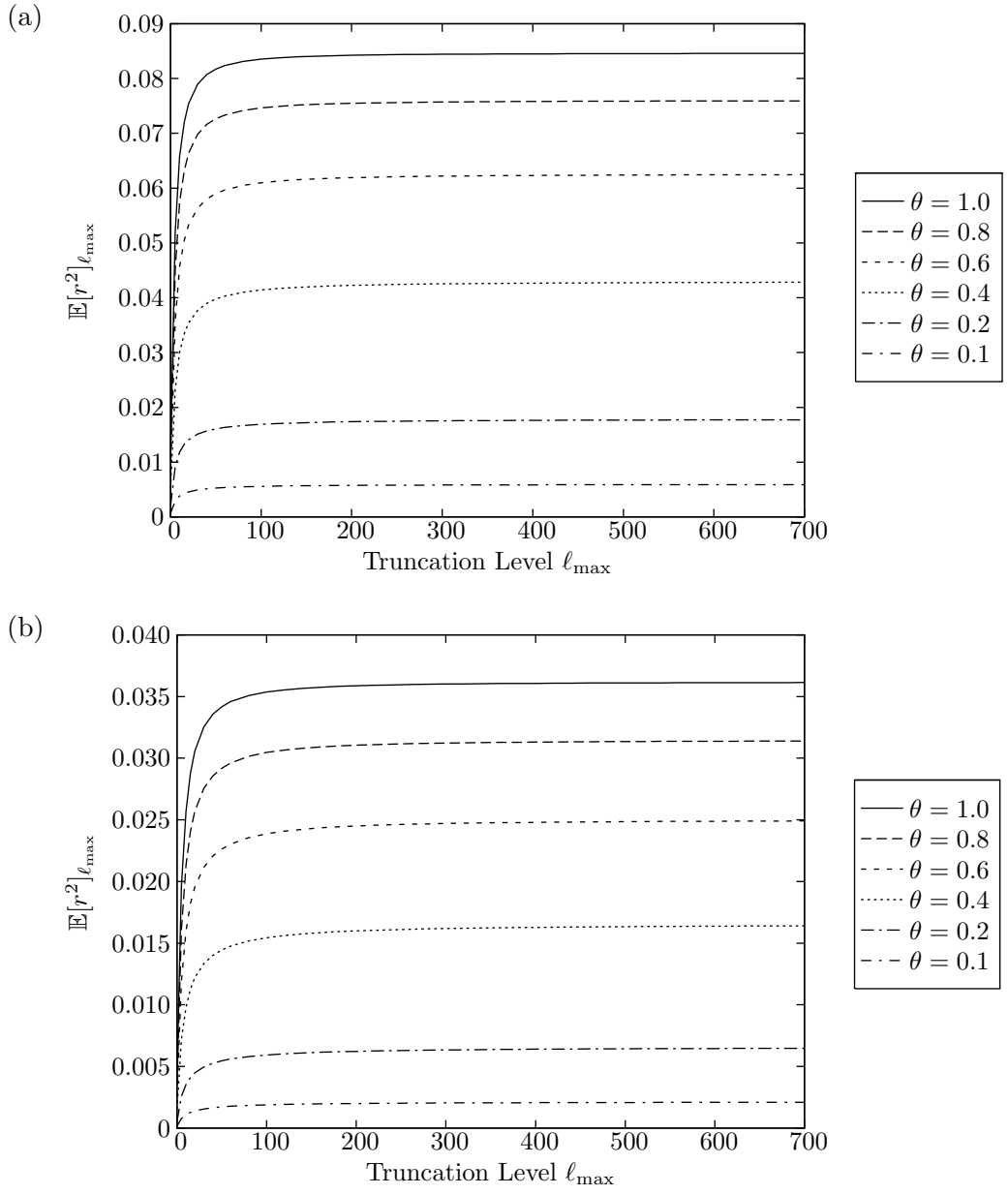

Figure 3: Plots of level truncated estimate $\mathbb{E}[r^2]_{\ell_{\max}}$, defined in (12), for various values of θ and ρ. (a) Plots for $\rho = 1$. (b) Plots for $\rho = 10$. For a given pair of θ and ρ, our software *ER2* took a few minutes on a laptop to compute $\mathbb{E}[r^2]_{\ell_{\max}}$ up to $\ell_{\max} = 700$. As these plots show, the sequence $\{\mathbb{E}[r^2]_{\ell_{\max}}\}_{\ell_{\max}=0}^{\infty}$ of partial sums converges fast in general. The precise rate of convergence of depends on θ and ρ.

us to express $\mathbb{E}[D^{2k+1}p^m q^n]$ for even m or even n purely in terms of $\mathbb{E}[D^{2k+1}p^i q^j]$, where $i \leq m$ and $j \leq n$ are both odd. Relations (13)–(16), together with the aforementioned observation that $\mathbb{E}[D^\ell p^m q^n] = \mathbb{E}[D^\ell p^n q^m]$, significantly reduce the number of expectations that need to be computed explicitly. For example, it is straightforward to show that

$$\begin{aligned}
\mathbb{E}[D^{2k}p^4q^3] = \mathbb{E}[D^{2k}p^3q^4] &= \frac{1}{4}\left(6\,\mathbb{E}[D^{2k}p^4q^2] - \mathbb{E}[D^{2k}p^4]\right), \\
\mathbb{E}[D^{2k}p^3q^3] &= \frac{1}{16}\left(\mathbb{E}[D^{2k}] - 12\,\mathbb{E}[D^{2k}p^2] + 36\,\mathbb{E}[D^{2k}p^2q^2]\right), \\
\mathbb{E}[D^{2k+1}p^4q^4] &= 4\,\mathbb{E}[D^{2k+1}p^3q^3] - 4\,\mathbb{E}[D^{2k+1}p^3q^1] + \mathbb{E}[D^{2k+1}pq], \\
\mathbb{E}[D^{2k+1}p^r] = \mathbb{E}[D^{2k+1}q^r] &= 0, \quad \text{for all } r \geq 0.
\end{aligned}$$

Furthermore, $\mathbb{E}[r^2]_{\ell_{\max}}$ in (12) can be written purely in terms of $\mathbb{E}[D^2p^mq^n]$ in which both m and n are even, thus leading to a more efficient method of computing $\mathbb{E}[r^2]_{\ell_{\max}}$.

3 Comparison of $\mathbb{E}[r^2]$ with σ_d^2 and coalescent simulations

In this section, we compare our computation of the expectation $\mathbb{E}[r^2]$ with averages of r^2 from coalescent simulations and also with the quantity

$$\sigma_d^2 = \frac{\mathbb{E}[D^2]}{\mathbb{E}[p(1-p)q(1-q)]} = \frac{10 + \rho + 4\,\theta}{22 + 13\,\rho + \rho^2 + 6\,\theta\rho + 32\,\theta + 8\,\theta^2}, \tag{17}$$

obtained by Ohta and Kimura (1969b) under the same model as in the present paper. Previous simulation-based study, in the context of an infinitely-many-alleles model, has shown that σ_d^2 can be substantially larger than $\mathbb{E}[r^2]$ (Maruyama, 1982). Hudson (1985) has shown, however, that σ_d^2 is a reasonably good approximation of the sample estimated expectation of r^2 that is conditioned on minor allele frequencies being above 5%. The discussion below pertains to our assumed recurrent mutation model, taking all frequencies into account.

Shown in Figure 4 is a plot of $\mathbb{E}[r^2]$ computed using our method, with $\ell_{\max} = 700$ as truncation level. A plot of σ_d^2 is shown in Figure 5a. As Figure 5b shows, $\mathbb{E}[r^2]$ and σ_d^2 can be considerably different for certain parameter values. The figure shows that $\mathbb{E}[r^2]$ and σ_d^2 agree well for $\theta > 4$. However, for small θ (say $\theta \leq 1$, which corresponds to a typical biologically interesting range), σ_d^2 can be larger than $\mathbb{E}[r^2]$ by a substantial amount (sometimes by tens of times), as in the aforementioned case of an infinitely-many-alleles model (Maruyama, 1982; Hudson, 1985). Numerical values of $\mathbb{E}[r^2]$ and σ_d^2 are shown in Table 1a and Table 1b, respectively. For $\theta \leq 1$, $\mathbb{E}[r^2]$ decreases as θ decreases, attaining negligibly small values for very small θ. In contrast, σ_d^2 can be very large even for very small θ. Note that σ_d^2 is a monotonically decreasing function of both θ and ρ. However, our computation indicates that, although $\mathbb{E}[r^2]$ is a monotonically decreasing function of ρ, it is not

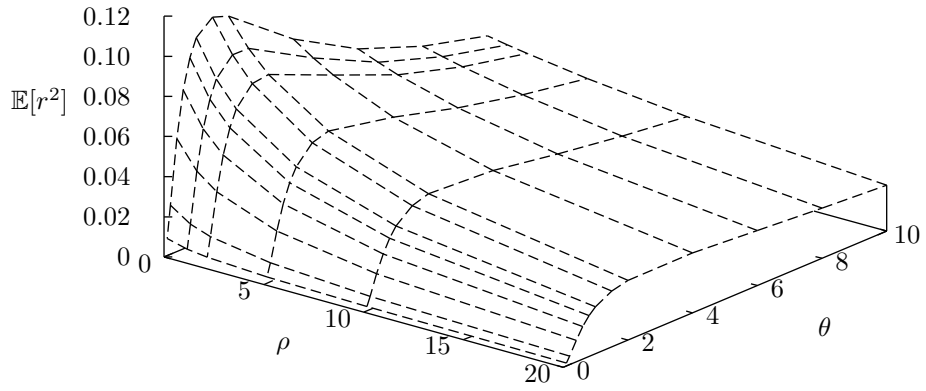

Figure 4: A plot of $\mathbb{E}[r^2]$ computed using our method, truncated at level $\ell_{\max} = 700$. Parameters θ and ρ denote the population-scaled mutation and recombination rates, respectively.

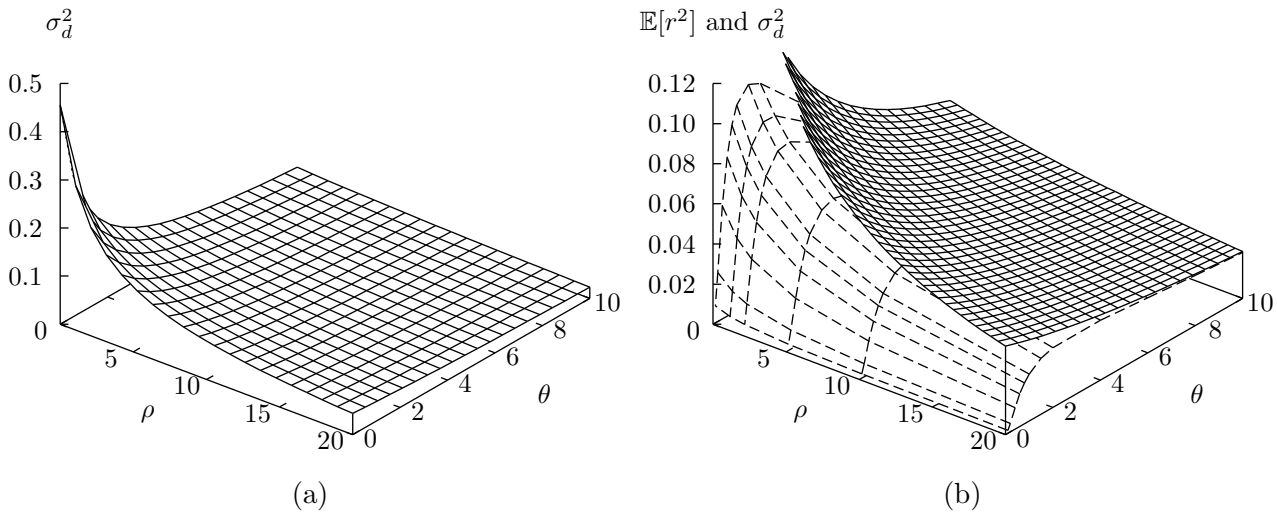

Figure 5: Comparison of σ_d^2 and our computation of $\mathbb{E}[r^2]$. (a) A plot of σ_d^2. (b) Superimposed plots of σ_d^2 and $\mathbb{E}[r^2]$, the latter (in dotted lines) being the same as that shown in Figure 4. Note that $\mathbb{E}[r^2]$ and σ_d^2 exhibit very different behavior for certain parameter ranges. In particular, for $\theta < 1$, σ_d^2 is larger than $\mathbb{E}[r^2]$ by a substantial amount.

a monotonically decreasing function of θ. For example, $\mathbb{E}[r^2]$ peaks at $\theta \approx 1.5$ for $\rho = 0$ and at $\theta \approx 2.8$ for $\rho = 20$.

We also carried out coalescent simulations to estimate average r^2. We used *Treevolve* (available from http://evolve.zoo.ox.ac.uk/software.html?id=treevolve, Grassly et al. 1999) with a constant population size of 10^4, and performed at least $10,000$ simulations for each pair of ρ and θ considered. In every simulation, we generated 200 sequences each with exactly 2 sites. Mutation parameters were chosen to simulate a symmetric recurrent mutation model with two possible alleles per site. Table 1c shows a summary of average r^2 when no restriction is imposed on segregation. For $\theta \leq 2$,

they are much closer to our theoretical computation of $\mathbb{E}[r^2]$ (see Table 1a) than they are to σ_d^2 (see Table 1b). As shown in Table 1d, however, when conditioned on segregation at both sites, sample estimated average r^2 increases substantially for small θ, resembling the behavior of σ_d^2. As Hudson (1985) noted, conditioning on minor allele frequencies being above 5% will make the sample estimated average of r^2 become even closer to σ_d^2. For large θ (say, $\theta \geq 4$), note that average r^2 estimated from simulations tends to be slightly smaller than $\mathbb{E}[r^2]$ for $\rho < 5$, while being slightly larger than $\mathbb{E}[r^2]$ for large ρ (say, $\rho > 10$).

4 Asymptotic behavior: Large ρ limit

Although a general closed-form formula for $\mathbb{E}[r^2]$ is difficult to obtain, it is possible to use our method to find relevant closed-form expressions in the limit $\rho \to \infty$. It is easy to see from (17) that the asymptotic behavior of σ_d^2 is given by

$$\sigma_d^2 = \frac{1}{\rho} + O(\rho^{-2})$$

(Ohta and Kimura, 1969b). In this section, we use our method to show rigorously that the leading term of $\mathbb{E}[r^2]$ in the limit $\rho \to \infty$ also goes like $1/\rho$, without any dependence on θ; i.e.,

$$\mathbb{E}[r^2] = \frac{1}{\rho} + O(\rho^{-2}).$$

4.1 Asymptotic behavior of $\mathbb{E}[D^2 p^m q^n]$

In our formulation (c.f., (4)), recall that computing $\mathbb{E}[r^2]$ amounts to computing $\mathbb{E}[D^2 p^m q^n]$, for $m, n \geq 0$. No closed-form formula for $\mathbb{E}[D^2 p^m q^n]$ is known for arbitrary parameter values, and therefore we have constructed an algorithm that can be used to compute $\mathbb{E}[D^2 p^m q^n]$ up to some chosen level $\ell_{\max} = m + n$. In the limit $\rho \to \infty$, however, it turns out that we can obtain a closed-form formula for the leading term in the asymptotic expansion of $\mathbb{E}[D^2 p^m q^n]$. This analysis goes as follows: To find the asymptotic behavior

$$\mathbb{E}[D^2 p^m q^n] = \frac{C(\theta)}{\rho} + O(\rho^{-2}),$$

Table 1: Numerical comparison of $\mathbb{E}[r^2]$, σ_d^2, and average r^2 from coalescent simulations. (a) Our computation of $\mathbb{E}[r^2]$ using $\ell_{\max} = 700$. (b) σ_d^2. (c) Average r^2 from coalescent simulations, with no restriction on segregation. (d) Average r^2 from coalescent simulations, conditioned on segregation at both sites. In coalescent simulations, we used $N_e = 10^4$ and generated 200 sequences each with exactly 2 sites. For each pair of ρ and θ, at least 10,000 simulated data sets were used to compute the average r^2. For $\theta \geq 4$, almost all simulated data sets had segregation at both sites, thus explaining why (c) and (d) are the same for $\theta \geq 4$.

(a)

	θ										
ρ	0.1	0.2	0.4	0.6	0.8	1.0	2.0	4.0	6.0	8.0	10.0
0.0	0.008	0.024	0.056	0.079	0.094	0.103	0.106	0.081	0.063	0.051	0.042
1.0	0.006	0.018	0.043	0.062	0.076	0.085	0.093	0.075	0.059	0.048	0.041
2.0	0.005	0.014	0.035	0.052	0.064	0.072	0.083	0.069	0.056	0.046	0.039
5.0	0.003	0.009	0.024	0.036	0.045	0.052	0.063	0.056	0.047	0.040	0.035
10.0	0.002	0.006	0.016	0.025	0.031	0.036	0.045	0.043	0.038	0.033	0.030
20.0	0.001	0.004	0.011	0.016	0.020	0.023	0.030	0.030	0.027	0.025	0.023

(b)

	θ										
ρ	0.1	0.2	0.4	0.6	0.8	1.0	2.0	4.0	6.0	8.0	10.0
0.0	0.411	0.376	0.322	0.281	0.250	0.226	0.153	0.094	0.068	0.053	0.044
1.0	0.286	0.269	0.240	0.217	0.199	0.183	0.132	0.085	0.063	0.050	0.042
2.0	0.220	0.209	0.192	0.177	0.165	0.154	0.116	0.079	0.060	0.048	0.040
5.0	0.130	0.127	0.120	0.114	0.109	0.104	0.086	0.064	0.051	0.042	0.036
10.0	0.078	0.077	0.074	0.072	0.070	0.068	0.060	0.048	0.040	0.035	0.030
20.0	0.044	0.043	0.042	0.042	0.041	0.040	0.037	0.032	0.029	0.026	0.023

(c)

	θ										
ρ	0.1	0.2	0.4	0.6	0.8	1.0	2.0	4.0	6.0	8.0	10.0
0.0	0.013	0.033	0.069	0.095	0.102	0.111	0.105	0.077	0.057	0.050	0.040
1.0	0.009	0.024	0.056	0.075	0.088	0.095	0.091	0.072	0.056	0.045	0.039
2.0	0.007	0.019	0.046	0.063	0.075	0.080	0.085	0.067	0.053	0.044	0.038
5.0	0.005	0.014	0.032	0.044	0.057	0.059	0.067	0.056	0.047	0.040	0.035
10.0	0.003	0.009	0.023	0.032	0.039	0.043	0.050	0.045	0.039	0.034	0.031
20.0	0.002	0.006	0.015	0.022	0.026	0.029	0.034	0.033	0.030	0.027	0.026

(d)

	θ										
ρ	0.1	0.2	0.4	0.6	0.8	1.0	2.0	4.0	6.0	8.0	10.0
0.0	0.131	0.128	0.126	0.125	0.121	0.119	0.106	0.077	0.057	0.050	0.040
1.0	0.093	0.093	0.097	0.099	0.102	0.103	0.095	0.072	0.056	0.045	0.039
2.0	0.076	0.078	0.081	0.082	0.088	0.091	0.084	0.067	0.053	0.044	0.038
5.0	0.051	0.052	0.057	0.059	0.062	0.066	0.067	0.056	0.047	0.040	0.035
10.0	0.037	0.038	0.041	0.042	0.046	0.048	0.051	0.045	0.039	0.034	0.031
20.0	0.024	0.026	0.028	0.029	0.031	0.032	0.035	0.033	0.030	0.027	0.026

where $C(\theta)$ is some function of θ to be determined, let $\tilde{D}^2 = \rho D^2$. Then, in the limit $\rho \to \infty$, the master equation (2) can be written in terms of p, q, and $\tilde{D}$ as

$$\begin{aligned}\mathbb{E}\left[\frac{p(1-p)}{2}\frac{\partial^2 f}{\partial p^2} + \frac{q(1-q)}{2}\frac{\partial^2 f}{\partial q^2} + \tilde{D}(1-2p)\frac{\partial^2 f}{\partial p\,\partial \tilde{D}} + \tilde{D}(1-2q)\frac{\partial^2 f}{\partial q\,\partial \tilde{D}} + \right.\\ + \frac{1}{2}\left[\rho\, pq(1-p)(1-q) + \sqrt{\rho}\tilde{D}(1-2p)(1-2q) - \tilde{D}^2\right]\frac{\partial^2 f}{\partial \tilde{D}^2} + \\ \left. + \frac{\theta}{4}(1-2p)\frac{\partial f}{\partial p} + \frac{\theta}{4}(1-2q)\frac{\partial f}{\partial q} - \frac{\rho}{2}\tilde{D}\frac{\partial f}{\partial \tilde{D}}\right] &= 0. \qquad (18)\end{aligned}$$

Substituting $p^m, q^n, p^m q^n$, $\tilde{D}p^m q^n$ and $\tilde{D}^2 p^m q^n$ for f in (18) and letting $\rho \to \infty$, we obtain the following recursion relations, respectively:

$$\begin{aligned}\mathbb{E}[p^m] &= \left(\frac{\frac{\theta}{2}+m-1}{\theta+m-1}\right)\mathbb{E}[p^{m-1}],\\ \mathbb{E}[q^n] &= \left(\frac{\frac{\theta}{2}+n-1}{\theta+n-1}\right)\mathbb{E}[q^{n-1}],\\ \mathbb{E}[p^m q^n] &= \frac{(\frac{\theta}{2}+m-1)\mathbb{E}[p^{m-1}q^n] + (\frac{\theta}{2}+n-1)\mathbb{E}[p^m q^{n-1}]}{m(\theta+m-1)+n(\theta+n-1)},\\ \mathbb{E}[\tilde{D}p^m q^n] &= 0\\ \mathbb{E}[\tilde{D}^2 p^m q^n] &= \mathbb{E}[p^{m+1}q^{n+1}] - \mathbb{E}[p^{m+1}q^{n+2}] - \mathbb{E}[p^{m+2}q^{n+1}] + \mathbb{E}[p^{m+2}q^{n+2}].\end{aligned}$$

Now, these recursions can be solved exactly; their solutions are given by

$$\begin{aligned}\mathbb{E}[p^n] &= \mathbb{E}[q^n] = \frac{\left(\frac{\theta}{2}\right)^{[n]}}{(\theta)^{[n]}},\\ \mathbb{E}[p^m q^n] &= \mathbb{E}[p^m]\,\mathbb{E}[q^n] = \frac{\left(\frac{\theta}{2}\right)^{[n]}\left(\frac{\theta}{2}\right)^{[m]}}{(\theta)^{[n]}\,(\theta)^{[m]}},\\ \mathbb{E}[\tilde{D}^2 p^m q^n] &= \left(1-\frac{\frac{\theta}{2}+n+1}{\theta+n+1}\right)\left(1-\frac{\frac{\theta}{2}+m+1}{\theta+m+1}\right)\mathbb{E}[p^{m+1}q^{n+1}] = \frac{1}{4}\,\frac{\left(\frac{\theta}{2}\right)^{[m+1]}\left(\frac{\theta}{2}\right)^{[n+1]}}{(\theta+1)^{[m+1]}\,(\theta+1)^{[n+1]}},\end{aligned}$$

where $(z)^{[k]}$ denotes the rising factorial defined in (7). Since $D^2 = \tilde{D}^2/\rho$, the last equation implies that, in the limit $\rho \to \infty$,

$$\mathbb{E}[D^2 p^m q^n] = \frac{1}{4}\,\frac{\left(\frac{\theta}{2}\right)^{[m+1]}\left(\frac{\theta}{2}\right)^{[n+1]}}{(\theta+1)^{[m+1]}\,(\theta+1)^{[n+1]}}\,\frac{1}{\rho} + O(\rho^{-2}). \qquad (19)$$

As a non-trivial check, consider $\mathbb{E}[D^2p^4q^2]$ as a rational function of ρ and θ; the numerator has degree 13 in ρ and degree 21 in θ, while the denominator has degree 14 in ρ and degree 22 in θ. In the limit $\rho \to \infty$, one can show that the asymptotic behavior of this rational function is

$$\mathbb{E}[D^2p^4q^2] = \frac{\theta^2(\theta+4)(\theta+6)(\theta+8)}{1024\,(\theta+5)(\theta^2+4\,\theta+3)^2}\;\frac{1}{\rho} + O(\rho^{-2}),$$

which agrees with (19) when $m = 4, n = 2$. We used our *Mathematica* program to make similar checks for other values of m, n.

4.2 Asymptotic behavior of $\mathbb{E}[r^2]$

Since $\mathbb{E}[r^2] = 4\sum_{m,n}\mathbb{E}[D^2p^mq^n]$, to study the behavior of $\mathbb{E}[r^2]$ in the limit $\rho \to \infty$, we need to sum over m and n in (19). For M a non-negative integer, we prove the following formula in Appendix A:

$$\sum_{m=0}^{M} \frac{\left(\frac{\theta}{2}\right)^{[m+1]}}{(\theta+1)^{[m+1]}} = 1 - \frac{\left(\frac{\theta}{2}+1\right)^{[M+1]}}{(\theta+1)^{[M+1]}}. \tag{20}$$

The right hand side of (20) can be written in terms of Γ-functions using the fact that $(x+1)^{[M+1]} = \Gamma(x+M+2)/\Gamma(x+1)$. Taking the limit $M \to \infty$ and using Stirling's formula for the asymptotic expansion of Γ-functions, we obtain, for $\theta > 0$,

$$\sum_{m=0}^{\infty} \frac{\left(\frac{\theta}{2}\right)^{[m+1]}}{(\theta+1)^{[m+1]}} = 1\,. \tag{21}$$

Thus, in the limit $\rho \to \infty$, equations (19) and (21) together imply

$$\mathbb{E}[r^2] = 4\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathbb{E}[D^2p^mq^n] = \frac{1}{\rho} + O(\rho^{-2}),$$

in which the leading term is independent of θ. As mentioned before, this behavior agrees with the asymptotic limit of $\sigma_d^2 = \mathbb{E}[D^2]/\mathbb{E}[p(1-p)q(1-q)]$ found by Ohta and Kimura (1969b).

5 Discrete process

In this section, we consider a description of the model in discrete time with a finite state space. This description was used by Hill and Robertson (1968) in their moment-generating matrix method, later extended by Ohta and Kimura (1969b) to include recurrent mutation. Although the discrete process is conceptually easy to grasp, computation in that framework is more complicated than in the diffusion approximation approach. Our goal in this section is to demonstrate that computing expectations in the discrete process can be facilitated by combinatorial techniques, and that such

computations may prove useful for studying the accuracy of the diffusion approximation approach.

5.1 Expected changes after one generation

We use p and q to denote the marginal allele frequencies, and f_{ij} to denote the frequency of the gametic type A_iB_j at the start of generation t. After mutation, the expected marginal allele frequencies are

$$p_u = (1-2u)p + u \qquad \text{and} \qquad q_u = (1-2u)q + u,$$

while the expected gametic frequencies f_{ij}^u are given by

$$\begin{aligned}
f_{00}^u &= f_{00} + u(f_{01} + f_{10}) - 2uf_{00}, \\
f_{01}^u &= f_{01} + u(f_{00} + f_{11}) - 2uf_{01}, \\
f_{10}^u &= f_{10} + u(f_{00} + f_{11}) - 2uf_{10}, \\
f_{11}^u &= f_{11} + u(f_{01} + f_{10}) - 2uf_{11}.
\end{aligned}$$

Note that, $p_u = f_{00}^u + f_{01}^u$ and $q_u = f_{00}^u + f_{10}^u$. Expected gametic frequencies after recombination are given by

$$\begin{aligned}
f_{00}^c &= (1-c)f_{00}^u + cp_uq_u, \\
f_{01}^c &= (1-c)f_{01}^u + cp_u(1-q_u), \\
f_{10}^c &= (1-c)f_{10}^u + c(1-p_u)q_u, \\
f_{11}^c &= (1-c)f_{11}^u + c(1-p_u)(1-q_u),
\end{aligned}$$

and expected frequencies after sampling can be obtained from taking expectations with respect to the following multinomial probability density:

$$\mathbb{P}_S(i,j,k \mid 2N_e) = \frac{(2N_e)!}{i!j!k!(2N_e-i-j-k)!}(f_{00}^c)^i(f_{01}^c)^j(f_{10}^c)^k(f_{11}^c)^{2N_e-i-j-k}.$$

In the remainder of this section, we use $\mathbb{E}_S$ to denote the expectation with respect to the sampling probability $\mathbb{P}_S(i,j,k \mid 2N_e)$, and $\mathbb{E}$ to denote the expectation with respect to the joint distribution of marginal allele frequencies and linkage disequilibrium.

5.2 Exact computation of $\mathbb{E}[p^l]$

In the algorithm described in Section 2.4, the expectation $\mathbb{E}[p^l]$ appears in some linear equations, and our computation of the expectation $\mathbb{E}[r^2]$ depends on $\mathbb{E}[p^l]$. The accuracy of $\mathbb{E}[p^l]$ therefore reflects

the accuracy of the diffusion process approximation. In what follows, we compute $\mathbb{E}[p^l]$ exactly in the discrete process setting and compare it with the answer from the diffusion approximation.

We first describe our exact computation of $\mathbb{E}[p^l]$. It is straightforward to show that the exponential generating function $\mathbb{E}_S[e^{x(I+J)}]$ is given by

$$\mathbb{E}_S[e^{x(I+J)}] = \sum_{i=0}^{2N_e} \sum_{j=0}^{2N_e-i} \sum_{k=0}^{2N_e-i-j} e^{x(i+j)} \mathbb{P}_S(i,j,k \mid 2N_e) = [(f_{00}^c + f_{01}^c)e^x + f_{10}^c + f_{11}^c)]^{2N_e},$$

which, using $f_{00}^c + f_{01}^c + f_{10}^c + f_{11}^c = 1$ and $f_{00}^c + f_{01}^c = p_u$, can be simplified as

$$\mathbb{E}_S[e^{x(I+J)}] = [(f_{00}^c + f_{01}^c)(e^x - 1) + 1]^{2N_e} = [p_u(e^x - 1) + 1]^{2N_e}.$$

We denote by p' the marginal allele frequency at locus 1 at the start of generation $t+1$. Then,

$$\mathbb{E}[(p')^l] = \mathbb{E}\,\mathbb{E}_S\left[\left(\frac{I+J}{2N_e}\right)^l\right] = \mathbb{E}\left[\frac{1}{(2N_e)^l}\frac{\partial^l}{\partial x^l}\mathbb{E}_S[e^{x(I+J)}]\Big|_{x=0}\right]. \tag{22}$$

We now describe how $\frac{\partial^l \mathbb{E}_S[e^{x(I+J)}]}{\partial x^l}\Big|_{x=0}$ can be computed exactly. First, define $G(x) := p_u(e^x - 1) + 1$ (i.e., $\mathbb{E}_S[e^{x(I+J)}] = [G(x)]^{2N_e}$) and

$$H(j,k) := \frac{\partial^j}{\partial x^j}[G(x)]^k\Big|_{x=0}, \tag{23}$$

for $k \geq j$. Then, since

$$\frac{\partial^j}{\partial x^j}[G(x)]^k = k\frac{\partial^{j-1}}{\partial x^{j-1}}\left\{p_u e^x [G(x)]^{k-1}\right\} = k\frac{\partial^{j-1}}{\partial x^{j-1}}\left\{[G(x)]^k - (1-p_u)[G(x)]^{k-1}\right\},$$

we see that $H(j,k)$ satisfies the following recursion relation:

$$H(j,k) = k[H(j-1,k) - (1-p_u)H(j-1,k-1)]. \tag{24}$$

The boundary condition for this recursion is $H(1,k) = kp_u$, for all $k \geq 1$, and, as shown in Appendix B, the solution is given by

$$H(j,k) = \sum_{i=1}^{j} (k)_{[i]}\, S(j,i) p_u^i, \tag{25}$$

where $(z)_{[i]}$ denotes the *falling factorial*, defined as

$$(z)_{[i]} := z(z-1)\cdots(z-i+1), \tag{26}$$

and $S(j,i)$ is the Stirling number of the second kind, defined as the number of partitions of $\{1,2,\ldots,j\}$ into i non-empty subsets.

At stationarity, $\mathbb{E}[(p')^l] = \mathbb{E}[p^l]$, and we can use (22), (23) and (25) to obtain

$$\mathbb{E}[p^l] = \frac{1}{(2N_e)^l} \sum_{i=1}^{l} (2N_e)_{[i]}\, S(l,i)\, \mathbb{E}[((1-2u)p+u)^i], \tag{27}$$

where we have substituted $p_u = (1-2u)p + u$. After some rearrangement, this equation allows us to compute $\mathbb{E}[p^l]$ recursively. We remark that $\mathbb{E}[1] = 1$ implies $\mathbb{E}[p] = 1/2$ for all u and N_e.

5.3 Approximation for small u and large N_e

We now consider a case in which $u \ll 1$ and $(1/N_e) \ll 1$, such that terms proportional to $u^i(1/N_e)^j$ where $i + j > 1$ may be ignored. For small u, note that

$$p_u^k \approx uk(p^{k-1} - 2p^k) + p^k.$$

Expanding the right hand side of (27) up to first order in either u or $1/(2N_e)$, but not both, gives

$$\mathbb{E}[p^l] \approx S(l,l)\left[\left(1 - \frac{1}{2N_e}\frac{l(l-1)}{2}\right)\mathbb{E}[p^l] + ul(\mathbb{E}[p^{l-1}] - 2\mathbb{E}[p^l])\right] + \frac{1}{2N_e} S(l,l-1)\mathbb{E}[p^{l-1}],$$

which, since $S(l,l) = 1$ and $S(l,l-1) = l(l-1)/2$, implies

$$\mathbb{E}[p^l] \approx \left(\frac{\frac{\theta}{2} + l - 1}{\theta + l - 1}\right)\mathbb{E}[p^{l-1}], \tag{28}$$

which, in turn, implies $\mathbb{E}[p^l] \approx \left(\frac{\theta}{2}\right)^{[l]} / (\theta)^{[l]}$. Similar results hold for $\mathbb{E}[q^l]$. Note that (28) is precisely the result we have obtained in (5) using diffusion theory. This example well illustrates the power of diffusion approximation; relations like (5) can be obtained very easily in that approach, whereas they may require a lot of effort to derive in the discrete analogue.

5.4 Comparison of the exact and approximate solutions

We now compare some results from the exact recursion (27) with that from the approximate recursion (28). Note that the former explicitly depends on the effective population size N_e, whereas the latter does not. Table 2 shows that, for fixed θ, (28) becomes a better approximation to the exact recursion (27) as N_e increases. Further, for small N_e, (28) becomes a significantly better approximation as the scaled mutation rate θ decreases. For $N_e \geq 10^4$, (28) is in general a good approximation for all $\theta \leq 1$.

Table 2: Numerical values of $\mathbb{E}[p^m]$ obtained from the exact recursion (27) or the approximate recursion (28). Results obtained from (28), which has no explicit dependence on the effective population size N_e, are equal to that from the diffusion process approximation.

		m							
θ	Recursion Used	50	100	150	200	250	300	350	400
0.1	Eq.(27), $N_e = 10^2$	0.404	0.392	0.386	0.383	0.380	0.378	0.377	0.376
	Eq.(27), $N_e = 10^3$	0.402	0.389	0.381	0.376	0.372	0.369	0.366	0.364
	Eq.(27), $N_e = 10^4$	0.402	0.388	0.380	0.375	0.371	0.368	0.365	0.362
	Eq.(28)	0.402	0.388	0.380	0.375	0.371	0.367	0.365	0.362
1.0	Eq.(27), $N_e = 10^2$	0.0834	0.0621	0.0533	0.0485	0.0454	0.0434	0.0419	0.0409
	Eq.(27), $N_e = 10^3$	0.0800	0.0569	0.0467	0.0407	0.0366	0.0336	0.0312	0.0294
	Eq.(27), $N_e = 10^4$	0.0796	0.0564	0.0461	0.0400	0.0358	0.0327	0.0303	0.0283
	Eq.(28)	0.0796	0.0563	0.0460	0.0399	0.0357	0.0326	0.0301	0.0282

6 Discussion

Although we have considered a symmetric recurrent mutation model for simplicity, the technique developed in this paper can be generalized to other mutation models. In particular, for a case with non-symmetric mutation rates or with unequal mutation rates at different loci, it should be straightforward to devise an algorithm similar to that in Section 2.4. Further, it should be possible to generalize our method to include natural selection, as described in (Ohta and Kimura, 1969a).

As we have shown in this paper, one can use our method to compute $\mathbb{E}[r^2]$ numerically for given θ and ρ. It would, of course, be desirable if we could obtain a closed-form formula for $\mathbb{E}[r^2]$. We believe that that is not completely out of reach. We here conclude with a description of an alternative perspective that may prove useful for future work on deriving a closed-form formula for $\mathbb{E}[r^2]$. If $f = Dp^{m+1}q^{n+1}$ is used in the master equation (2), we obtain

$$\begin{aligned} 4(1+m)(1+n)\,\mathbb{E}[D^2p^mq^n] &= 2[\rho+\theta(4+m+n)+10+m^2+n^2+5m+5n]\,\mathbb{E}[Dp^{m+1}q^{n+1}] \\ &\quad -(1+m)(4+2m+\theta)\,\mathbb{E}[Dp^mq^{1+n}]-(1+n)(4+2n+\theta)\,\mathbb{E}[Dp^{m+1}q^n]. \end{aligned} \tag{29}$$

Further, using $f = p^kq^l$, one can show that

$$\begin{aligned} \mathbb{E}[Dp^{k-1}q^{l-1}] &= \frac{1}{2kl}\Big\{[k(k-1+\theta)+l(l-1+\theta)]\,\mathbb{E}[p^kq^l] \\ &\quad -k\left(k-1+\frac{\theta}{2}\right)\mathbb{E}[p^{k-1}q^l]-l\left(l-1+\frac{\theta}{2}\right)\mathbb{E}[p^kq^{l-1}]\Big\}. \end{aligned} \tag{30}$$

Hence, if the joint expectation $\mathbb{E}[p^iq^j]$ of powers of marginal frequencies are known, then $\mathbb{E}[D^2p^mq^n]$ can easily be computed using (29) and (30), and there would be no need to solve systems of coupled

equations as described in Section 2.4. At this point, however, we do not know how to obtain a general expression for $\mathbb{E}[p^i q^j]$. Using (6), we can obtain the following generating function for $\mathbb{E}[p^k]$ and $\mathbb{E}[q^k]$:

$$\mathbb{E}[e^{\alpha p}] = \mathbb{E}[e^{\alpha q}] = 2^{\theta-1}\, e^{\frac{\alpha}{2}}\, \alpha^{\frac{1}{2}-\frac{\theta}{2}}\, I_{\frac{\theta-1}{2}}\left(\frac{\alpha}{2}\right) \Gamma\left(\frac{1}{2}+\frac{\theta}{2}\right),$$

where $I_\nu(z)$ is the modified Bessel function of the first kind. Moreover, we can show that the generating function $\mathbb{E}[e^{\alpha p+\beta q}]$ for $\mathbb{E}[p^i q^j]$ has the form

$$\mathbb{E}[e^{\alpha p+\beta q}] = \mathbb{E}[e^{\alpha p}]\,\mathbb{E}[e^{\beta q}]\,h(\alpha^2,\beta^2), \tag{31}$$

where $h(\alpha^2,\beta^2)$ is a symmetric function in α^2 and β^2. We believe that finding an explicit formula for $h(\alpha^2,\beta^2)$ is worthy of further research, since knowing the generating function $\mathbb{E}[e^{\alpha p+\beta q}]$ may lead to a closed-form formula for $\mathbb{E}[r^2]$.

Acknowledgment

We thank Charles H. Langley for helpful comments. This research is supported in part by grants CCF-0515278 and IIS-0513910 (YSS) from National Science Foundation.

References

Ethier, S. N., Griffiths, R. C., 1990. On the two-locus sampling distribution. J. Math. Biol. 29, 131–159.

Golding, G. B., 1984. The sampling distribution of linkage disequilibrium. Genetics 108, 257–274.

Grassly, N. C., Harvey, P. H., Holmes, E. C., 1999. Population dynamics of HIV-1 inferred from gene sequences. Genetics 151, 427–438, (Software webpage: http://evolve.zoo.ox.ac.uk/software.html?id=treevolve).

Hill, W. G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226–231.

Hill, W. G., Weir, B. S., 1994. Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. 54, 705–714.

Hudson, R. R., 1985. Sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics 109, 611–631.

Hudson, R. R., 2001a. Linkage disequilibrium and recombination. In: Balding, D. J., Bishop, M., Canning, C. (Eds.), Handbook of Statistical Genetics. Wiley, pp. 309–324.

Hudson, R. R., 2001b. Two-locus sampling distributions and their application. Genetics 159, 1805–1817.

International HapMap Consortium, 2005. A haplotype map of the human genome. Nature 437, 1299–1320.

Maruyama, T., 1982. Stochastic integrals and their application to population genetics. In: Kimura, M. (Ed.), Molecular Evolution, Protein Polymorphism and their Neutral Theory. Springer-Verlag, Berlin, pp. 151–166.

McVean, G. A. T., 2002. A genealogical interpretation of linkage disequilibrium. Genetics 162, 987–991.

Ohta, T., Kimura, M., 1969a. Linkage disequilibrium due to random genetic drift. Genet. Res. Camb. 13, 47–55.

Ohta, T., Kimura, M., 1969b. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63, 229–238.

Pritchard, J. K., Przeworski, M., 2001. Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69, 1–14.

Appendix

A Proof of Equation (20)

We prove the claim by induction. The lemma clearly holds for $M = 0$. Now, assume the case for M. Then, using the induction hypothesis, we have

$$\begin{aligned}
\sum_{n=0}^{M+1} \frac{\left(\frac{\theta}{2}\right)^{[n+1]}}{(\theta+1)^{[n+1]}} &= 1 - \frac{\left(\frac{\theta}{2}+1\right)^{[M+1]}}{(\theta+1)^{[M+1]}} + \frac{\left(\frac{\theta}{2}\right)^{[M+2]}}{(\theta+1)^{[M+2]}} \\
&= 1 - \frac{\left(\frac{\theta}{2}+1\right)^{[M+1]}(\theta+M+2) - \left(\frac{\theta}{2}\right)^{[M+2]}}{(\theta+1)^{[M+2]}} \\
&= 1 - \frac{\left(\frac{\theta}{2}+1\right)^{[M+1]}(\theta+M+2-\frac{\theta}{2})}{(\theta+1)^{[M+2]}} \\
&= 1 - \frac{\left(\frac{\theta}{2}+1\right)^{[M+2]}}{(\theta+1)^{[M+2]}},
\end{aligned}$$

which proves the claim.

B Proof of Equation (25)

Let X, Y and Z be finite sets of order x, y and z. Let $\Phi = \{\varphi : X \to Y\}$ be the set of all maps from X to Y and likewise for $\Psi = \{\psi : Y \to Z\}$, with order of $|\Phi| = y^x$ and $|\Psi| = z^y$. Now define $R_{\varphi,\psi} = \{(a, b, c) \in X \times Y \times Z \,|\, b = \varphi(a) \text{ and } c = \psi(b)\}$ for each pair $\varphi \in \Phi, \psi \in \Psi$. Let $R = \{R_{\varphi,\psi} | \varphi \in \Phi, \psi \in \Psi\}$ and $H_z(x, y) := |R|$. Note that $R_{\varphi,\psi}$ may be equal to $R_{\varphi',\psi'}$ for $\varphi \neq \varphi'$ and $\psi \neq \psi'$, and $H_z(x, y)$ counts the number of distinct sets $R_{\varphi,\psi}$ in R.

Lemma B.1 *$H_z(x, y)$ can be expressed as*

$$H_z(x, y) = \sum_{i=1}^{x} (y)_{[i]} S(x, i) z^i,$$

where $S(x, i)$ denotes the Stirling number of the second kind. In particular, $H_z(1, y) = y\, z$.

Proof: We first partition R as $R = R_1 \sqcup \cdots \sqcup R_x$ where $R_i = \{R_{\varphi,\psi} |\ |\varphi(X)| = i\}$. Then, there are z choices for mapping i preimages of ψ, $(y)_{[i]}$ choices for the i image of φ, and $S(x, i)$ ways of partitioning X into kernels of φ. Hence, $|R_i| = (y)_{[i]} S(x, i) z^i$. Summing over i thus gives the desired result. ■

Lemma B.2 *$H_z(x,y)$ satisfies the recursion relation*

$$H_z(x,y) = y[H_z(x-1,y) + (z-1)H_z(x-1,y-1)]$$

with initial condition $H_z(1,y) = y\,z$.

Proof: Fix an element $a \in X$ and $c \in Z$, and define $\tilde{X} = X \setminus \{a\}$ and $\tilde{Z} = Z \setminus \{c\}$. Then, we can partition R as $R = R_1 \sqcup R_2 \sqcup R_3$, where

$$\begin{aligned} R_1 &= \{R_{\varphi,\psi} \mid \varphi(a) \in \varphi(\tilde{X})\}, \\ R_2 &= \{R_{\varphi,\psi} \mid \varphi(a) \notin \varphi(\tilde{X}),\ \psi(\varphi(a)) = c\}, \\ R_3 &= \{R_{\varphi,\psi} \mid \varphi(a) \notin \varphi(\tilde{X}),\ \psi(\varphi(a)) \neq c\}. \end{aligned}$$

Now, define $\tilde{R}_{\tilde{\varphi},\psi}$ to be as before by replacing X by $\tilde{X}$. Then, to each of the $H_z(x-1,y)$ sets $\tilde{R}_{\tilde{\varphi},\psi}$, we can obtain y sets $R_{\varphi,\psi}$ by adjoining one element $(a,b,\psi(b))$ for $b \in \tilde{\varphi}(\tilde{X})$ and (a,b,c) for $b \notin \tilde{\varphi}(\tilde{X})$, and extending $\tilde{\varphi}$ to φ so that $\varphi(a) = b$. Then, there exits a one-to-one correspondence between the set of all such sets obtained and $R_1 \sqcup R_2$, and we thus see that $|R_1 \sqcup R_2| = yH_z(x-1,y)$. It is also easy to see that $|R_3| = y(z-1)H_z(x-1,y-1)$. ■

Note that even when z is not an integer, $H_z(x,y)$ still satisfies the recursion relation in Lemma B.2 as a purely algebraic relation.