Theoretical Population Biology 74 (2008) 130–137

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Mathematical properties of the r^2 measure of linkage disequilibrium

Jenna M. VanLiere [a,*], Noah A. Rosenberg [a,b,c]

[a] Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI 48109, United States
[b] Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, United States
[c] Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, United States

ARTICLE INFO

Article history:
Received 22 February 2008
Available online 1 June 2008

Keywords:
Allele frequency
Linkage disequilibrium
Recombination

ABSTRACT

Statistics for linkage disequilibrium (LD), the non-random association of alleles at two loci, depend on the frequencies of the alleles at the loci under consideration. Here, we examine the r^2 measure of LD and its mathematical relationship to allele frequencies, quantifying the constraints on its maximum value. Assuming independent uniform distributions for the allele frequencies of two biallelic loci, we find that the mean maximum value of r^2 is ∼0.43051, and that r^2 can exceed a threshold of 4/5 in only ∼14.232% of the allele frequency space. If one locus is assumed to have known allele frequencies – the situation in an association study in which LD between a known marker locus and an unknown trait locus is of interest – we find that the mean maximum value of r^2 is greatest when the known locus has a minor allele frequency of ∼0.30131. We find that in 1/4 of the space of allowed values of minor allele frequencies and haplotype frequencies at a pair of loci, the unconstrained maximum r^2 allowing for the possibility of recombination between the loci exceeds the constrained maximum assuming that no recombination has occurred. Finally, we use $r^2_{\max}$ to examine the connection between r^2 and the D' measure of linkage disequilibrium, finding that $r^2/r^2_{\max} = D'^2$ for ∼72.683% of the space of allowed values of (p_a, p_b, p_{ab}). Our results concerning the properties of r^2 have the potential to inform the interpretation of unusual LD behavior and to assist in the design of LD-based association-mapping studies.

1. Introduction

Linkage disequilibrium (LD) refers to a non-random association in the occurrence of alleles at two loci (Hudson, 2001; Pritchard and Przeworski, 2001; Slatkin, 2008). LD finds applications in diverse contexts, including the inference of demographic events in human evolutionary history (Tishkoff et al., 1996; Plagnol and Wall, 2006), the fine-mapping of disease genes after localization via linkage analysis (Devlin and Risch, 1995), and the modeling, selection, and evaluation of sets of informative single-nucleotide polymorphisms for use in detecting disease-susceptibility alleles in genome-wide association studies (Kruglyak, 1999; Carlson et al., 2004; Eberle et al., 2007). Measurements of LD are typically based on comparisons of the observed frequencies of haplotypes to the frequencies expected based on the frequencies of the alleles that comprise the various haplotypes. Statistically estimated haplotype frequencies are used in place of observed frequencies when observed frequencies are unavailable.

One of the challenges inherent in measuring LD is that the ranges of LD measures can depend on the frequencies of alleles at the loci under consideration. Hedrick (1987) showed that for several LD statistics, holding the allele frequencies of one of the two loci in a pair constant, the maximal values of the statistics could occur only when the allele frequencies of the second locus were equal to those of the first locus; further, in some cases, the maximum itself was frequency dependent. The only statistic considered by Hedrick (1987) whose range was frequency independent was D' (Lewontin, 1964), which ranges from −1 to 1 for any set of allele frequencies for a pair of polymorphic biallelic loci. However, even D' is not independent of allele frequencies in most senses of the concept of "independence" (Lewontin, 1988).

For biallelic loci, one of the most commonly used measures for LD is r^2 (Hill and Robertson, 1968), the square of the correlation coefficient between two indicator variables — one representing the presence or absence of a particular allele at the first locus and the other representing the presence or absence of a particular allele at the second locus. In a disease association context, the r^2 statistic is often used in calculations of power to detect disease-susceptibility loci. Under some conditions, the power to detect disease association with a marker locus when using a case-control sample of size N is approximately equal to the power to detect disease association with the true causal locus when using a sample of size Nr^2, where r^2 here denotes the value of the r^2 statistic for the marker locus and the causal locus

* Corresponding author.
E-mail address: jennav@umich.edu (J.M. VanLiere).

0040-5809/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2008.05.006

(Pritchard and Przeworski, 2001; Jorgenson and Witte, 2006; Terwilliger and Hiekkalinna, 2006). The r^2 statistic also underlies popular methods for identifying informative markers for use in LD-based association studies (Carlson et al., 2004; de Bakker et al., 2005).

Like most LD statistics, r^2 has a frequency-dependent range. The maximum value of r^2 as a function of the allele frequencies of two loci under consideration drops sharply with the extent of the minor allele frequency difference between the loci (Wray, 2005; Eberle et al., 2006; Amos, 2007). Thus, in some settings, matching loci by allele frequencies prior to measurement of LD can provide a way to circumvent the frequency dependence of r^2. Using genotypes from 71 unrelated individuals of European, African-American, and Chinese descent, Eberle et al. (2006) found that by restricting their calculations to matched loci with similar allele frequencies, their ability to identify high LD values using r^2 increased considerably, revealing excess LD in genic regions.

Although the frequency dependence of r^2 has often been noted (e.g. Devlin and Risch (1995) and Zondervan and Cardon (2004)), relatively little is known about the mathematical properties of this dependence. Wray (2005) showed that if two loci have a value of r^2 above a specified cutoff and one of the loci has known allele frequencies, then the frequencies at the other locus must lie in a narrow range. Eberle et al. (2006) studied the properties of r^2 in a genealogical context, examining the predictions made by a coalescent model about the expected value of r^2 conditional on the allele frequencies at a pair of loci in the absence of recombination. In this paper, we consider the mathematical relationship between r^2 and allele frequencies in detail. We investigate the maximum possible value of r^2 for a given set of allele frequencies, compute the mean value of $r^2_{\max}$ when frequencies at one of the loci are assumed to be known, and determine the range of possible allele frequencies for one locus when r^2 and the frequencies for the other locus are known. We also use two possible genealogical histories (a scenario similar to that of Eberle et al. (2006)) to investigate the effect of recombination on the value of r^2. Finally, we determine the relationship between r^2 and D' using a connection to $r^2_{\max}$, the maximum value of r^2 possible given the allele frequencies at a pair of loci.

2. Theory

Consider two biallelic loci, locus 1 with alleles a and A and locus 2 with alleles b and B. Suppose the frequencies for alleles a and A are respectively p_a and $1-p_a$, and the frequencies for alleles b and B are p_b and $1-p_b$. Since p_a and p_b range from 0 to 1, the pair (p_a, p_b) ranges over the (open) unit square. The set of combinations of allele frequencies (p_a, p_b) can be split into eight components, which we label $S_1, S_2, \ldots, S_8$ for convenience (Fig. 1). Each of the other seven components, $S_2, \ldots, S_8$, corresponds to a transformation of S_1 in which alleles are swapped at locus 1, alleles are swapped at locus 2, loci 1 and 2 are swapped, or two or more of these exchanges are performed. We will use this symmetry to simplify some of our calculations.

The r^2 measure of linkage disequilibrium is defined as

$$r^2(p_a, p_b, p_{ab}) = \frac{(p_{ab} - p_a p_b)^2}{p_a(1-p_a)p_b(1-p_b)}, \tag{1}$$

where p_{ab} is the frequency of haplotypes having allele a at locus 1 and allele b at locus 2 (Hill and Robertson, 1968). As the square of a correlation coefficient, $r^2(p_a, p_b, p_{ab})$ can range from 0 to 1 as p_a, p_b and p_{ab} vary.

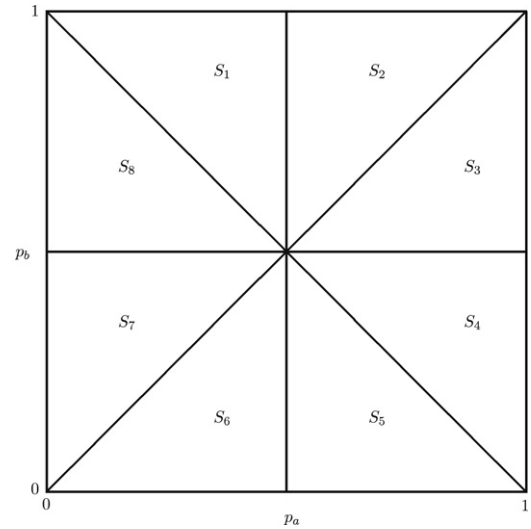

Fig. 1. The unit square of possible combinations of allele frequencies (p_a, p_b), divided into eight components. The other seven components can all be obtained from a transformation of S_1 by switching alleles at locus 1 (reflection over the line $p_a = 1/2$), switching alleles at locus 2 (reflection over the line $p_b = 1/2$), switching loci (reflection over the line $p_b = p_a$), or some combination of these three exchanges.

2.1. $r^2_{\max}(p_a, p_b)$

Our first computation is of $r^2_{\max}(p_a, p_b)$, the maximum value of r^2 for given values of p_a and p_b, considering all possible values of p_{ab}. Given (p_a, p_b), the denominator of r^2 is fixed. Therefore, to maximize r^2, it suffices to choose the value for p_{ab} that maximizes the numerator. The possible values of p_{ab} are constrained by the fact that the frequency of a haplotype can be no more than the frequency of the least frequent allele that it contains and no less than 0 or the minimum overlap that can occur between two alleles based on their frequencies. It is at one of these extremes – the highest or lowest possible haplotype frequency – that the numerator is maximized. Thus, the maximum value of r^2 occurs either at $p_{ab} = \min(p_a, p_b)$ or at $p_{ab} = \max(0, p_a + p_b - 1)$, depending on the component, S_i, in which the given (p_a, p_b) is located. For S_1 and S_4, the maximum occurs at $p_{ab} = p_a + p_b - 1$, so

$$r^2_{\max}(p_a, p_b) = \frac{(1-p_a)(1-p_b)}{p_a p_b}. \tag{2}$$

For S_2 and S_7, the maximum occurs at $p_{ab} = p_a$:

$$r^2_{\max}(p_a, p_b) = \frac{p_a(1-p_b)}{(1-p_a)p_b}. \tag{3}$$

For S_3 and S_6, the maximum occurs at $p_{ab} = p_b$:

$$r^2_{\max}(p_a, p_b) = \frac{(1-p_a)p_b}{p_a(1-p_b)}. \tag{4}$$

Finally, for S_5 and S_8, the maximum occurs at $p_{ab} = 0$:

$$r^2_{\max}(p_a, p_b) = \frac{p_a p_b}{(1-p_a)(1-p_b)}. \tag{5}$$

Table 1 summarizes these results. Note that $r^2_{\max}$ is continuous on the boundaries between components. An important consequence of Eqs. (2)–(5) is that $r^2_{\max}(p_a, p_b) = 1$ if and only if $p_b = p_a$ or $p_b = 1 - p_a$.

Combining Eqs. (2)–(5), Fig. 2 shows a three-dimensional plot of $r^2_{\max}$ for all combinations (p_a, p_b). The X-shape of the figure illustrates the symmetries of r^2 as a function of p_a and p_b, as well as the property that r^2 can only equal 1 if the two loci have the

Table 1
$r^2_{\max}$, $D^2_{\max}$, and $D^2_{r^2_{\max}}$ in the eight components of the space of possible allele frequencies.

Component	$p_a < \frac{1}{2}$	$p_b < \frac{1}{2}$	$p_a < p_b$	$p_a + p_b < 1$	p_{ab} that produces $r^2_{\max}$	$r^2_{\max}$	$D^2_{\max}$		$D^2_{r^2_{\max}}$
							$D < 0$	$D > 0$	
S_1	Yes	No	Yes	No	$p_a + p_b - 1$	$\frac{(1-p_a)(1-p_b)}{p_a p_b}$	$[(1-p_a)(1-p_b)]^2$	$[p_a(1-p_b)]^2$	$[(1-p_a)(1-p_b)]^2$
S_2	No	No	Yes	No	p_a	$\frac{p_a(1-p_b)}{(1-p_a)p_b}$	$[(1-p_a)(1-p_b)]^2$	$[p_a(1-p_b)]^2$	$[p_a(1-p_b)]^2$
S_3	No	No	No	No	p_b	$\frac{(1-p_a)p_b}{p_a(1-p_b)}$	$[(1-p_a)(1-p_b)]^2$	$[(1-p_a)p_b]^2$	$[(1-p_a)p_b]^2$
S_4	No	Yes	No	No	$p_a + p_b - 1$	$\frac{(1-p_a)(1-p_b)}{p_a p_b}$	$[(1-p_a)(1-p_b)]^2$	$[(1-p_a)p_b]^2$	$[(1-p_a)(1-p_b)]^2$
S_5	No	Yes	No	Yes	0	$\frac{p_a p_b}{(1-p_a)(1-p_b)}$	$(p_a p_b)^2$	$[(1-p_a)p_b]^2$	$(p_a p_b)^2$
S_6	Yes	Yes	No	Yes	p_b	$\frac{(1-p_a)p_b}{p_a(1-p_b)}$	$(p_a p_b)^2$	$[(1-p_a)p_b]^2$	$[(1-p_a)p_b]^2$
S_7	Yes	Yes	Yes	Yes	p_a	$\frac{p_a(1-p_b)}{(1-p_a)p_b}$	$(p_a p_b)^2$	$[p_a(1-p_b)]^2$	$[p_a(1-p_b)]^2$
S_8	Yes	No	Yes	Yes	0	$\frac{p_a p_b}{(1-p_a)(1-p_b)}$	$(p_a p_b)^2$	$[p_a(1-p_b)]^2$	$(p_a p_b)^2$

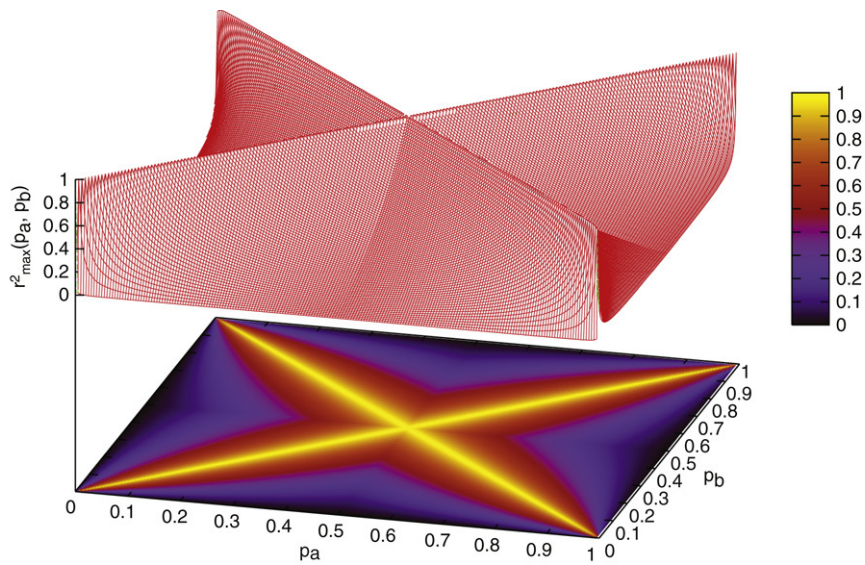

Fig. 2. $r^2_{\max}(p_a, p_b)$ as a three-dimensional plot, with a contour plot shown below.

same minor allele frequency. Additionally, the graph shows a very steep decay of $r^2_{\max}$ moving away from the diagonals, indicating that even small differences in allele frequency between the two loci, especially if the frequencies are not near 1/2, can reduce the range of possible values for r^2 considerably.

We can quantify the effect of differences in minor allele frequency observed in Fig. 2 by calculating the average $r^2_{\max}$ value assuming independent Uniform(0,1) distributions for p_a and p_b. This computation amounts to evaluating the volume below $r^2_{\max}$ over the unit square. Using symmetry, the total volume can be calculated by finding the volume over one of the eight components in Fig. 1 and multiplying by eight. Denoting the volume of $r^2_{\max}$ over component S_1 by V_1, we have

$$\begin{aligned} V_1 &= \int_0^{\frac{1}{2}} \int_{1-p_a}^{1} \frac{(1-p_a)(1-p_b)}{p_a p_b} dp_b dp_a \\ &= -\int_0^{\frac{1}{2}} \frac{(1-p_a)}{p_a}[\ln(1-p_a) + p_a] dp_a \\ &= \frac{1}{12}\pi^2 - \frac{1}{2}(\ln 2)^2 + \frac{1}{2}\ln 2 - \frac{7}{8} \\ &\approx 0.05381. \end{aligned}$$

The last step uses the dilogarithm function $\int_z^0 \ln(1-t)/t\, dt = \mathrm{Li}_2(z)$, where $\mathrm{Li}_2(0) = 0$ and $\mathrm{Li}_2(1/2) = \pi^2/12 - (\ln 2)^2/2$ (Weisstein, 2003). Consequently the mean $r^2_{\max}$ given $p_a \sim$ Uniform(0, 1), $p_b \sim$ Uniform(0, 1), and assuming p_a and p_b are independent is $8V_1 = 2\pi^2/3 - 4(\ln 2)^2 + 4(\ln 2) - 7 \approx 0.43051$.

This result and the shape of Fig. 2 suggest that it is only possible to achieve high values of r^2 over relatively small portions of the space of possible values of p_a and p_b. For a constant c, $0 \le c \le 1$, we can calculate the proportion of the allele frequency space where it is possible for r^2 to exceed c, $p(c)$. Again using symmetry, we can restrict our attention to S_6. Using Eq. (4), the portion of S_6 in which $r^2_{\max}(p_a, p_b) \ge c$, whose area we denote by A_6, satisfies

$$p_b \ge \frac{cp_a}{1 - p_a + cp_a}.$$

Considering the complement of the area of interest in S_6, we have

$$\begin{aligned} \frac{1}{8} - A_6 &= \int_0^{\frac{1}{2}} \int_0^{\frac{cp_a}{1-p_a+cp_a}} 1 dp_b dp_a \\ &= -\frac{c}{2(1-c)} - \frac{c\ln\left(\frac{1}{2} + \frac{1}{2}c\right)}{(1-c)^2}. \end{aligned}$$

Thus, the proportion of the allele frequency space where it is possible for r^2 to exceed c is $8A_6$, or

$$p(c) = 1 + \frac{4c}{1-c} + \frac{8c\ln\left(\frac{1}{2} + \frac{1}{2}c\right)}{(1-c)^2}. \tag{6}$$

Fig. 3 shows that the proportion of the allele frequency space where it is possible for r^2 to exceed c declines faster than linearly. For example, only over ∼0.39709 of the allele frequency space is it possible for r^2 to exceed 1/2 and only over ∼0.14232 of the space is it possible for r^2 to exceed 4/5.

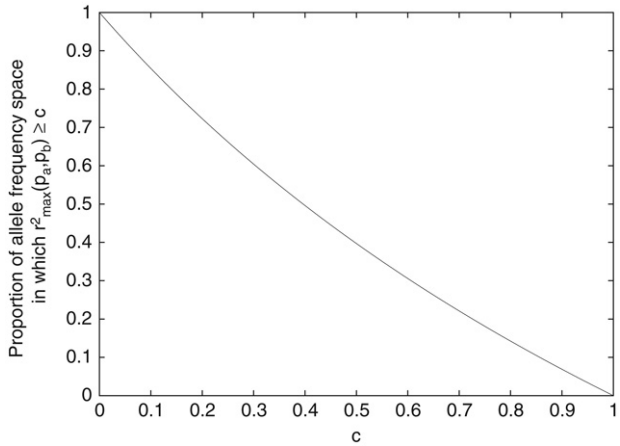

Fig. 3. The proportion of the allele frequency space where it is possible for r^2 to be greater than a constant c. By "allele frequency space" we mean the unit square in which p_a and p_b range from 0 to 1.

2.2. $r^2_{\max}(p_a, p_b)$ with p_a fixed

In contrast to the previous computations, in which we performed integrations over possible values of p_a and p_b, we now consider the case in which p_a is fixed. This computation enables us to identify the allele frequencies for a locus that is able to have high r^2 values across the broadest range of allele frequencies for a second locus. Assuming p_b has a Uniform(0,1) distribution, we can calculate $m(p_a) = E[r^2_{\max}(p_a, p_b)|p_b \sim \text{Uniform}(0, 1)]$, the mean maximum r^2 value as a function of p_a. We can assume that $p_a \leq 1/2$ and then consider $p_a > 1/2$ by observing that $m(p_a) = m(1 - p_a)$. For $p_a \leq 1/2$ we perform piecewise integration across components S_6, S_7, S_8, and S_1 using Eqs. (2)–(5):

$$\begin{aligned} m(p_a) &= \int_0^{p_a} \frac{(1-p_a)p_b}{p_a(1-p_b)} dp_b + \int_{p_a}^{\frac{1}{2}} \frac{p_a(1-p_b)}{(1-p_a)p_b} dp_b \\ &\quad + \int_{\frac{1}{2}}^{1-p_a} \frac{p_a p_b}{(1-p_a)(1-p_b)} dp_b \\ &\quad + \int_{1-p_a}^{1} \frac{(1-p_a)(1-p_b)}{p_a p_b} dp_b \\ &= -\frac{2(1-p_a)}{p_a}[p_a + \ln(1-p_a)] \\ &\quad + \frac{2p_a}{(1-p_a)}\left[\ln\left(\frac{1}{2}\right) - \frac{1}{2} - \ln p_a + p_a\right]. \end{aligned}$$

Fig. 4 shows that the mean of $r^2_{\max}(p_a, p_b)$, averaging over values of p_b, has an m-shape as a function of p_a. The maximum of this mean occurs at $p_a \approx 0.30131$ and $p_a \approx 0.69869$ and equals ~0.53091. Notice that the largest values of $m(p_a)$ occur for intermediate minor allele frequencies rather than for minor allele frequencies close to 1/2. This finding can be explained by examining the contour plot of Fig. 2, which suggests that slices through the graph made at intermediate frequencies for p_a contain more space with higher values of $r^2_{\max}$ than do other slices.

2.3. $r^2_{\max}(p_a, p_b)$ with p_a and $p_a - p_b$ fixed

We now consider the situation in which p_a and the difference between allele frequencies $|p_a - p_b|$ are known. This situation is similar to the scenario considered by Wray (2005) in which r^2 was assumed to exceed some known threshold, p_a was assumed to be known, and $p_b = p_a + v$ was investigated.

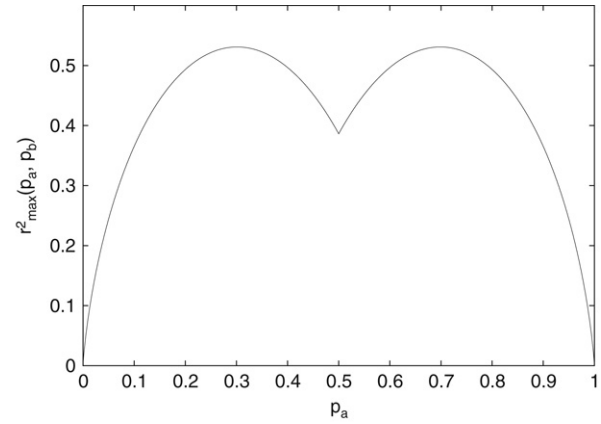

Fig. 4. The mean of $r^2_{\max}(p_a, p_b)$ assuming p_a is constant and p_b is distributed uniformly on the unit interval, as a function of p_a.

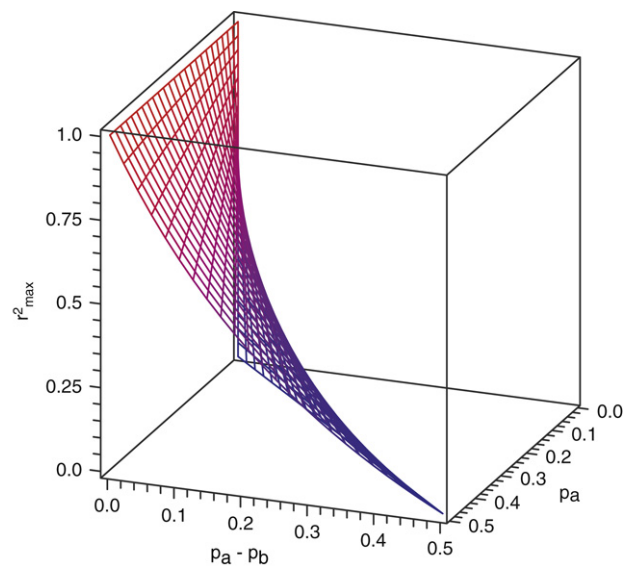

Fig. 5. $r^2_{\max}$ as a function of p_a and $p_a - p_b$ assuming $p_b \leq p_a \leq 1/2$. The domain of the graph corresponds to the component S_6 in Fig. 1.

Let p_a and p_b be minor allele frequencies ($\leq 1/2$) with $p_a \geq p_b$, so that we are considering component S_6. Define $d = p_a - p_b \geq 0$. Treating $r^2_{\max}$ as a function of p_a and d, we can rewrite Eq. (4):

$$r^2_{\max}(p_a, d) = 1 - \frac{d}{p_a(1 + d - p_a)}. \tag{7}$$

Fig. 5 shows a three-dimensional plot of $r^2_{\max}$ as a function of the larger minor allele frequency (p_a) and the difference between minor allele frequencies ($p_a - p_b$). The twisted surface illustrates that for p_a fixed, $r^2_{\max}$ decreases faster with the difference in minor allele frequency when p_a has smaller values. This observation corresponds to the steeper decline from the diagonals further from the center in Fig. 2.

Holding d constant and non-negative in Eq. (7), the maximum of $r^2_{\max}$ for $p_a \leq 1/2$ occurs at $p_a = 1/2$:

$$r^2 \leq 1 - \frac{d}{\frac{1}{2}(1 + d - \frac{1}{2})}. \tag{8}$$

By rearranging this equation, we can calculate the maximum value of $|p_a - p_b|$ possible given a known value of r^2,

$$d \leq \frac{1 - r^2}{2(1 + r^2)}. \tag{9}$$

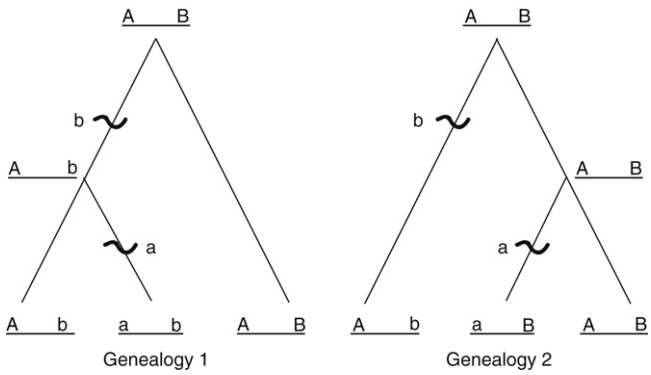

Fig. 6. Two possible non-recombinant genealogies. In Genealogy 1, allele *a* arises on a haplotype that contains allele *b*. In Genealogy 2, allele *a* arises on a haplotype that does not contain allele *b*.

As Eq. (9) is based on the maximum of $r^2_{\max}$ over all possible minor allele frequency values for p_a, it represents the broadest range possible for the difference in allele frequencies. For d to achieve the maximum value of $(1-r^2)/[2(1+r^2)]$, p_a must equal 1/2.

The computation above assumes a known r^2 and determines the maximum for d. However, if we know p_a in addition to r^2, then we can solve exactly for the set of allowable values of p_b. Assuming again that p_a and p_b are minor allele frequencies (that is, at most 1/2), we must consider two cases: $p_a \geq p_b$ and $p_a \leq p_b$. For the first case, (p_a, p_b) is in S_6. Rearranging Eq. (4),

$$p_b \geq \frac{r^2 p_a}{1 + r^2 p_a - p_a}, \tag{10}$$

so that $r^2 p_a/(1+r^2 p_a - p_a) \leq p_b \leq p_a$. In the second case, (p_a, p_b) is in S_7 so we can rearrange Eq. (3) to obtain

$$p_b \leq \frac{p_a}{r^2 - r^2 p_a + p_a}. \tag{11}$$

Recalling our assumption that $p_b \leq 1/2$ and combining Eqs. (10) and (11), we find

$$\frac{r^2 p_a}{1 + r^2 p_a - p_a} \leq p_b \leq \min\left(\frac{1}{2}, \frac{p_a}{r^2 - r^2 p_a + p_a}\right).$$

This result accords with the values that appear in Table 2 of Wray (2005).

2.4. $r^2_{\max}(p_a, p_b)$ and recombination

We have previously been examining r^2 with the assumption that it is possible for p_{ab} to take any value within its allowable range. This amounts to an assumption that we are not constraining the recombination history of the two loci under consideration. In this section, we consider a different situation: how does recombination affect r^2 for two loci that have not previously experienced recombination? This depends on the genealogical history of the loci.

Consider two possible genealogies (Fig. 6). In Genealogy 1, a mutation at locus 1 arises later, but on the same side of the tree, as a mutation at locus 2. In Genealogy 2, the mutations at loci 1 and 2 arise on different sides of the tree so that no haplotypes carry both mutations. Assuming $p_a \leq p_b \leq 1/2$ so that the minor alleles are derived rather than ancestral, then without recombination, $p_{ab} = p_a$ for Genealogy 1, and $r^2(p_a, p_b) = p_a(1-p_b)/[(1-p_a)p_b]$ (Eberle et al., 2006). For Genealogy 2, without recombination $p_{ab} = 0$, so $r^2(p_a, p_b) = p_a p_b/[(1-p_a)(1-p_b)]$ (Eberle et al., 2006).

In typical settings, recombination reduces linkage disequilibrium, as recombination separates new alleles from the haplotypic background on which they arose. For Genealogy 1 in Fig. 6, the unconstrained maximum r^2 allowing p_{ab} to take on any possible value is precisely $r^2_{\max}(p_a, p_b)$ (Eq. (3)), the value taken when $p_{ab} = p_a$ and no recombination has occurred. Thus, any recombination events that reduce p_{ab} below p_a will lead to a decrease in r^2. However, with Genealogy 2 we can see that situations do exist in which recombination can lead to an increase in LD. Consider Genealogy 2 and suppose recombination occurs such that the frequency of the recombinant haplotype (ab) becomes $p_{ab} > 0$. This haplotype can arise through recombination events between *Ab* haplotypes and *aB* haplotypes. Is it possible for r^2, with recombination events allowed, to be greater than r^2 in the absence of recombination? Solving the inequality

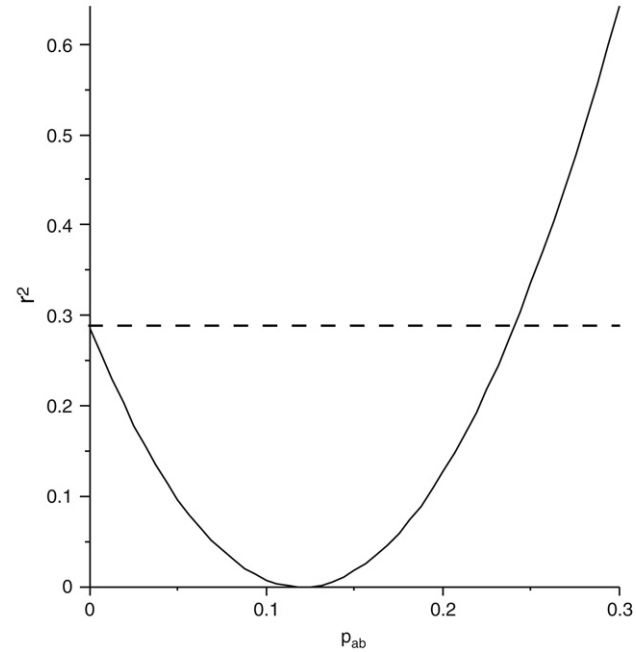

Fig. 7. Example of a situation in which recombination can produce an increase in r^2, for $p_a = 0.3$ and $p_b = 0.4$. Here, p_{ab} is the frequency of the recombinant haplotype in the setting of Genealogy 2 in Fig. 6. The dashed line is the value of r^2 assuming no recombination.

$$\frac{(p_{ab} - p_a p_b)^2}{p_a(1-p_a)p_b(1-p_b)} > \frac{p_a p_b}{(1-p_a)(1-p_b)},$$

we obtain

$$p_{ab} > 2p_a p_b.$$

For each $p_a \leq 1/2$ and $p_b \leq 1/2$ it is possible to choose a value of p_{ab} that satisfies $p_{ab} \geq 2p_a p_b$. Recall our assumption that $p_a \leq p_b$, which restricts $p_{ab} \leq p_a$. Thus, if the fraction of recombinant haplotypes satisfies

$$2p_a p_b < p_{ab} \leq p_a, \tag{12}$$

then the occurrence of recombination produces an increase in r^2 compared to the maximum possible value had no recombination occurred on the genealogy. Fig. 7 shows an example of the variation in r^2 as a function of p_{ab} for $p_a = 0.3$ and $p_b = 0.4$. Once p_{ab} exceeds $2p_a p_b = 0.24$, the value of r^2 between loci increases above the initial value in the absence of recombination.

Using inequality (12), we can determine the fraction of the space of allowed values for (p_a, p_b, p_{ab}) in which the unconstrained maximum r^2 permitting recombination (p_{ab} not necessarily equal to 0) exceeds the maximum under the assumption that no recombination occurs ($p_{ab} = 0$). The volume of the region where

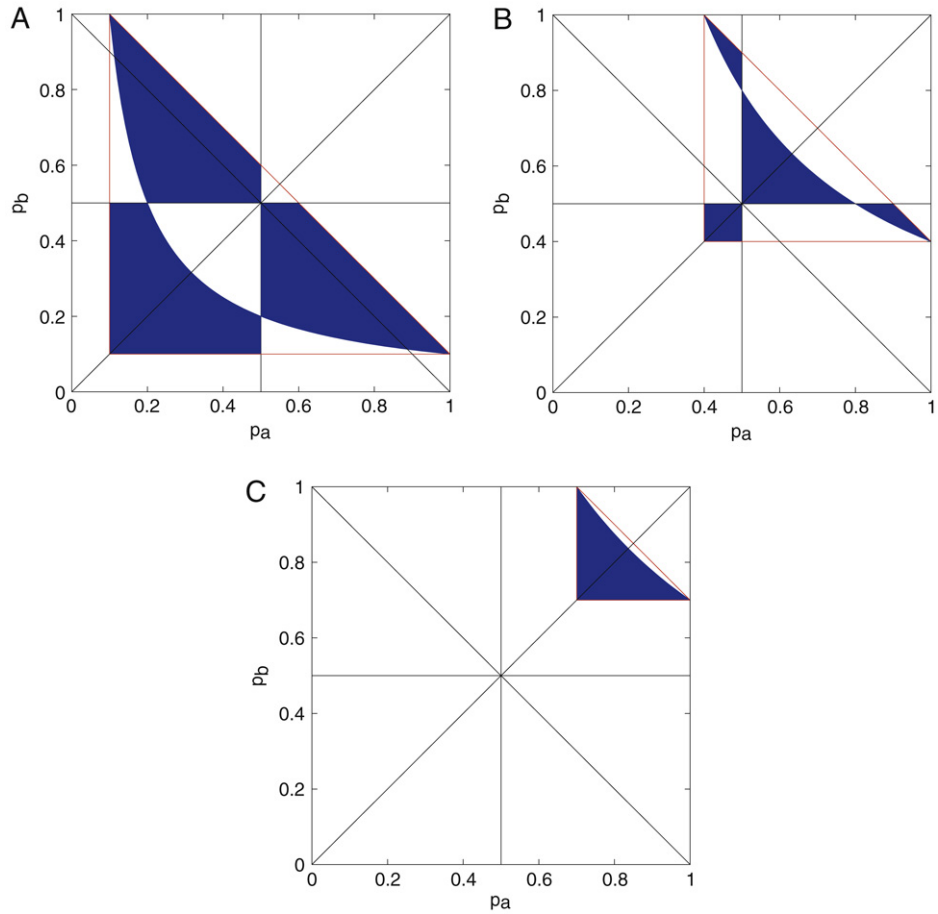

Fig. 8. The portion of the space of values of (p_a, p_b) in which $D'^2 = r^2/r^2_{\max}$. The red triangle indicates the boundary of the set of values of (p_a, p_b) possible for a given p_{ab}. The blue shaded space indicates regions where $D'^2 = r^2/r^2_{\max}$. Each of the three plots represents a "slice" of the three-dimensional space for (p_a, p_b, p_{ab}), holding p_{ab} constant. (A) $p_{ab} = 0.1$. (B) $p_{ab} = 0.4$. (C) $p_{ab} = 0.7$.

recombination inflates r^2 is

$$\int_0^{\frac{1}{2}} \int_0^{p_b} \int_{2p_ap_b}^{p_a} 1 dp_{ab} dp_a dp_b = \frac{1}{192}.$$

The volume of the region of allowed values for (p_a, p_b, p_{ab}), assuming $p_{ab} \le p_a \le p_b \le 1/2$, is

$$\int_0^{\frac{1}{2}} \int_0^{p_b} \int_0^{p_a} 1 dp_{ab} dp_a dp_b = \frac{1}{48}.$$

Taking the quotient of these results, the fraction of the space of possible values in which recombination inflates r^2 is 1/4. Thus, averaging over possible values for (p_a, p_b) with $p_a \le p_b \le 1/2$, on average 1/4 of possible values for p_{ab} lead to $r^2(p_a, p_b, p_{ab}) > r^2(p_a, p_b, 0)$.

2.5. The relationship between $r^2(p_a, p_b, p_{ab})$ and $D'(p_a, p_b, p_{ab})$

So far, we have focused on the r^2 measure of LD and on various properties of its maximum value. A second LD statistic, namely D', is defined based on maxima and minima. Our computations with $r^2_{\max}$ provide a basis for examining the connection between r^2 and D'.

D' is defined as

$$D' = \frac{D}{D_{\max}}, \tag{13}$$

where $D = p_{ab} - p_ap_b$, $D_{\max} = \min(p_ap_b, (1-p_a)(1-p_b))$ if $D < 0$, and $D_{\max} = \min(p_a(1-p_b), (1-p_a)p_b)$ if $D > 0$ (Lewontin, 1964). Given any values for p_a and p_b, D' can take on any value from -1 to 1, thus differing from r^2 in that its range is not frequency dependent (Hedrick, 1987; Lewontin, 1988).

D' is equal to D normalized by its maximum given the allele frequencies; r^2 can similarly be normalized by its maximum to obtain $r^2/r^2_{\max}$. This quotient is the squared correlation coefficient between allelic indicator variables at two loci, standardized by the maximum squared correlation possible given the frequencies of the alleles at the two loci. As D'^2 and $r^2/r^2_{\max}$ both have numerator D^2, it is natural to compare their different normalization procedures to determine if they represent the same quantity. We can rewrite $r^2/r^2_{\max}$ as

$$\frac{r^2}{r^2_{\max}} = \frac{D^2}{D^2_{r^2_{\max}}}. \tag{14}$$

Here, $D_{r^2_{\max}}$ is defined as $p_{ab} - p_ap_b$ evaluated at the value of p_{ab} that produces the maximum of D^2 as a function of p_a and p_b. This quantity differs across components of the allele frequency space, as described in Section 2.1. Comparing Eq. (14) to Eq. (13), we find that $D'^2 = r^2/r^2_{\max}$ if $D^2_{\max} = D^2_{r^2_{\max}}$.

The sign of D determines how $D_{\max}$ is computed. Thus, whether $D_{\max} = D_{r^2_{\max}}$, and consequently $D'^2 = r^2/r^2_{\max}$, depends on the sign of D. For example, consider S_1, in which $p_a < 1/2$, $p_b > 1/2$, and $p_a + p_b > 1$. In this component, $D_{\max}$ equals $D_{r^2_{\max}} = (1-p_a)(1-p_b)$ only when D is less than 0 ($p_{ab} < p_ap_b$). In each of the eight components, $D^2_{\max} = D^2_{r^2_{\max}}$ either when $D < 0$ or when $D > 0$, but not in both cases (Table 1). Thus, the region in which $D'^2 = r^2/r^2_{\max}$

includes some but not all of the space of possible values of p_a, p_b, and p_{ab}. When $D'^2 \neq r^2/r^2_{\max}$, D'^2 is always greater than $r^2/r^2_{\max}$.

As D'^2 and $r^2/r^2_{\max}$ are functions of p_a, p_b, and p_{ab}, we can fix one of these three variables and examine the relationship between D'^2 and $r^2/r^2_{\max}$ as a function of the other two variables. If we fix p_{ab}, then the domain for (p_a, p_b) is a triangle, as $p_a \geq p_{ab}$, $p_b \geq p_{ab}$, and $p_a + p_b - 1 \leq p_{ab}$. Inside this triangle, Fig. 8 shows the values of (p_a, p_b) where $D'^2 = r^2/r^2_{\max}$ for $p_{ab} = 0.1$, 0.4, and 0.7. The three graphs represent the three qualitatively different patterns observed for such graphs as p_{ab} varies from 0 to 1. For $p_{ab} = 0.1$, the domain spans all eight components, S_1 to S_8. For $p_{ab} = 0.4$, the domain spans all eight components, but in two of these components there is no region in which $D'^2 = r^2/r^2_{\max}$ and in two other components there is no region in which $D'^2 \neq r^2/r^2_{\max}$. Finally, for $p_{ab} = 0.7$, the domain spans only two components, S_2 and S_3. The transition points between the three cases occur at $p_{ab} = 1/4$, where the boundary line $p_{ab} = p_a p_b$ crosses the point $(1/2, 1/2)$, and at $p_{ab} = 1/2$ where the space of allowable (p_a, p_b) becomes restricted to the upper right quadrant.

As a function of p_{ab}, we can calculate the fraction of the space of possibilities where $D'^2 = r^2/r^2_{\max}$. For a given p_{ab}, the space of possible values of (p_a, p_b) is bounded by $p_a = p_{ab}$, $p_b = p_{ab}$, and $p_{ab} = p_a + p_b - 1$, producing a triangle of area $(1 - p_{ab})^2/2$. For $0 \leq p_{ab} \leq 1/4$, we calculate the area where $D'^2 = r^2/r^2_{\max}$ by subtracting the area where the two quantities are not equal from the total area possible, yielding

$$\frac{1}{2}(1-p_{ab})^2 - \left[2\int_{\frac{1}{2}}^{1}\int_{p_{ab}}^{\frac{p_{ab}}{p_a}} 1 dp_b dp_a + \frac{1}{2}p_{ab}^2 + \int_{2p_{ab}}^{\frac{1}{2}}\int_{\frac{p_{ab}}{p_a}}^{\frac{1}{2}} 1 dp_b dp_a\right]. \tag{15}$$

For $1/4 \leq p_{ab} \leq 1/2$, we calculate the area where $D'^2 = r^2/r^2_{\max}$ by summing areas in each quadrant and noting that the upper left and lower right quadrants have the same area. This area is

$$2\int_{p_{ab}}^{\frac{1}{2}}\int_{\frac{p_{ab}}{p_a}}^{1+p_{ab}-p_a} 1 dp_b dp_a + \int_{\frac{1}{2}}^{2p_{ab}}\int_{\frac{1}{2}}^{\frac{p_{ab}}{p_a}} 1 dp_b dp_a + \left(\frac{1}{2} - p_{ab}\right)^2. \tag{16}$$

For $1/2 \leq p_{ab} \leq 1$, the calculation of the area where $D'^2 = r^2/r^2_{\max}$ is simplified due to the restriction of the space of possible values of (p_a, p_b) to the upper right quadrant. This area is

$$\int_{p_{ab}}^{1}\int_{p_{ab}}^{\frac{p_{ab}}{p_a}} 1 dp_b dp_a. \tag{17}$$

Computing the integrals in Eqs. (15)–(17) and then dividing by the area of the space of possible values of (p_a, p_b), we find that the fraction of the space where $D'^2 = r^2/r^2_{\max}$ is

$$\begin{aligned} &\frac{\frac{1}{4} + p_{ab}(1 - 4\ln 2) - p_{ab}\ln p_{ab}}{\frac{1}{2}(1-p_{ab})^2}, \quad 0 < p_{ab} \leq \frac{1}{4} \\ &\frac{\frac{5}{4} + p_{ab}(4\ln 2 - 3) + 3p_{ab}\ln p_{ab}}{\frac{1}{2}(1-p_{ab})^2}, \quad \frac{1}{4} < p_{ab} \leq \frac{1}{2} \\ &\frac{p_{ab}(p_{ab} - 1 - \ln p_{ab})}{\frac{1}{2}(1-p_{ab})^2}, \quad \frac{1}{2} < p_{ab} < 1. \end{aligned} \tag{18}$$

Fig. 9 shows a plot of this function. The minimum fraction of the space where $D'^2 = r^2/r^2_{\max}$ is ∼0.31357, which occurs at $p_{ab} \approx 0.37162$. The fraction is generally large for large p_{ab}; when p_{ab} is large the probability is quite high that $D > 0$. In S_2 and S_3, positive D leads to $D'^2 = r^2/r^2_{\max}$.

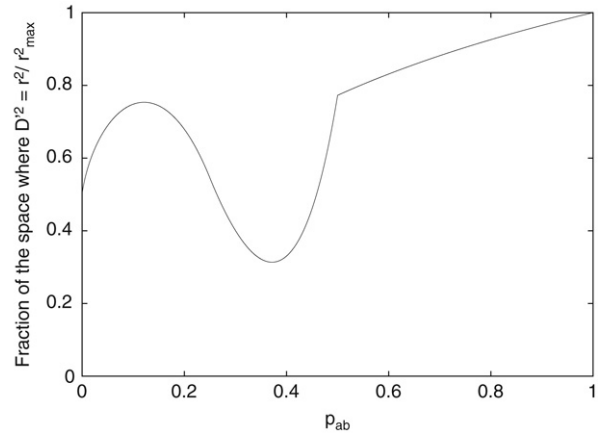

Fig. 9. The fraction of the space of possible values of (p_a, p_b, p_{ab}) for which $D'^2 = r^2/r^2_{\max}$, as a function of p_{ab}.

By integrating the function in Eq. (18) from 0 to 1, we can obtain the fraction of the space of all three variables – p_a, p_b, and p_{ab} – in which $D'^2 = r^2/r^2_{\max}$. Again using the dilogarithm, we obtain

$$\frac{1}{3}\pi^2 - 4(\ln 2)^2 + \frac{3}{2} - 8\mathrm{Li}_2\left(\frac{1}{4}\right)$$

for the probability that a set of values of p_a, p_b and p_{ab} chosen from the space of possible values leads to $r^2/r^2_{\max} = D'^2$. Numerically, this probability is ∼0.72683.

3. Discussion

In this paper, we have examined the mathematical relationship between r^2 and allele frequencies, producing a variety of results concerning the frequency dependence of r^2. By evaluating the volume below $r^2_{\max}(p_a, p_b)$, we found that the mean $r^2_{\max}$ over the space of possible allele frequencies is only ∼0.43051. This number is rather low, implying that for much of the allele frequency space, the value of r^2 is severely restricted. We also calculated the formula for the proportion of the allele frequency space where it is possible for r^2 to exceed some constant c (Eq. (6)). Using the cutoff of $c = 4/5$ commonly employed for examining the genomic coverage of a set of "tag SNPs" in association studies (e.g. Jorgenson and Witte, 2006), we find that it is possible for r^2 to be greater than or equal to this value in only ∼0.14232 of the allele frequency space.

An additional scenario that we considered is the case in which one of the allele frequencies was set to a fixed known value. This is the situation, for example, in an association study in which a marker locus with fixed known allele frequencies is used to detect a trait locus of unknown allele frequencies. By assuming a uniform distribution for the frequency of an allele at the other locus, we found that the marker minor allele frequency able to detect high LD with the largest range of values for the minor allele frequency of the trait locus was ∼0.30131, not 1/2 as might have been expected from an assumption that the most polymorphic markers have the greatest potential for LD detection. Although the specific location of the optimum may change with the distribution of allele frequencies in an actual population, this result has the implication that algorithms that choose informative markers for detecting LD might produce improved performance if they ensure that a considerable fraction of markers near the optimum frequency are selected. The sharp allele frequency dependence of r^2 may also mean that it is desirable to choose a range of allele frequencies among "tag SNP" markers in order to increase the probability of capturing LD with unknown trait loci.

Another perhaps surprising result, obtained by considering the effect of recombination on the value of r^2 for different genealogical histories, is that in certain contexts recombination can increase rather than decrease the value of r^2. This is somewhat counterintuitive; a typical scenario of loss of LD with recombination involves a decoupling of derived mutations that have occurred sequentially on the same lineage, such as in recombination events between haplotypes *ab* and *AB* of Genealogy 1 in Fig. 6. In our scenarios where recombination can increase LD, in Genealogy 2 of Fig. 6, the LD is produced by recombination that produces sufficient coupling between derived mutations that have occurred in parallel on separate lineages. This type of scenario is likely to be a rather unusual outcome under common assumptions about evolutionary processes; however, we did observe that such scenarios accounted for a non-trivial proportion of the space of possible values for (p_a, p_b, p_{ab}).

Finally, we considered the relationship between r^2 and another commonly used measure of LD, D'. We found that a close connection exists between r^2 and D', in that D'^2 is often equal to $r^2/r^2_{\max}$. For any haplotype frequency p_{ab}, this equality occurs over at least ∼31.357% of the space of possible allele frequencies (p_a, p_b), and when $r^2/r^2_{\max}$ and D'^2 are not equal, $r^2/r^2_{\max}$ is always less than D'^2. Due to its connections to both r^2 and D', there may exist some potential for $r^2/r^2_{\max}$, which we term $r^{2\prime}$, to serve as a useful LD measure. Although many measures of LD have situations in which they are particularly applicable (Hedrick, 1987; Devlin and Risch, 1995; Hudson, 2001; Morton et al., 2001), $r^{2\prime}$ – the squared correlation coefficient between allelic indicator variables at two loci standardized by the maximum squared correlation possible given the frequencies of the alleles at the two loci – is one of relatively few that can be used when a measure with allele-frequency-independent range is desired.

Note that in various computations we have considered the entire unit square as the domain for p_a and p_b. Some treatments of LD reorient alleles and loci so that only S_6 or S_7 is examined (e.g. Amos, 2007), or otherwise use a reorientation that spans more than one of the eight components in Fig. 1 (e.g. Morton et al., 2001). Consideration of only a single component in some cases will yield results that are identical on the allowed domain to those presented (e.g. Fig. 2). Particularly in the comparison between $r^{2\prime}$ and D'^2, however, restriction of the space of allele frequencies may lead to somewhat different results. Within a component, $r^2_{\max}$ is achieved when the haplotype with the major alleles at both loci has as high a frequency as possible, so that the normalization in the computation of $r^{2\prime}$ depends only on the allele frequencies p_a and p_b. However, the normalization in the computation of D' additionally takes into account which alleles are coupled, so that it depends on whether or not p_{ab} exceeds $p_a p_b$. Thus, reorienting alleles so that $p_a \leq p_b$, $p_a \leq 1/2$, and $D > 0$, as is done by Morton et al. (2001), leads to a domain for p_a and p_b that cannot be obtained by dividing the plots in Fig. 8 along one of their lines of symmetry. Consequently, given p_{ab}, the reorientation of Morton et al. (2001) will produce a different result for the probability that $r^{2\prime}$ is equal to D'^2 over the allowed domain.

We have additionally assumed Uniform(0,1) distributions of allele frequencies in many computations. This assumption can be viewed as a basis for assessing functions of allele frequencies across their entire ranges, rather than as an assumption that these distributions apply in any particular population. Our primary interest has been to provide details on the theoretical properties of r^2; future work may have the potential to exploit the properties that we have uncovered, such as in interpreting unusual LD behavior, or in improving the design of disease-mapping studies that rely on patterns of LD.

Acknowledgments

We thank the two reviewers for their comments. This work was supported by NIH grants R01 GM081441 and T32 HG00040 and by grants from the Alfred P. Sloan Foundation and the Burroughs Wellcome Fund.

References

Amos, C.I., 2007. Successful design and conduct of genome-wide association studies. Human Molecular Genetics 16, R220–R225.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., Nickerson, D.A., 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. American Journal of Human Genetics 74, 106–120.

de Bakker, P.I.W., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., Altshuler, D., 2005. Efficiency and power in genetic association studies. Nature Genetics 37, 1217–1223.

Devlin, B., Risch, N., 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29, 311–322.

Eberle, M.A., Ng, P.C., Kuhn, K., Zhou, L., Peiffer, D.A., Galver, L., Viaud-Martinez, K.A., Taylor Lawley, C., Gunderson, K.L., Shen, R., Murray, S.S., 2007. Power to detect risk alleles using genome-wide tag SNP panels. PLoS Genetics 3, 1827–1837.

Eberle, M.A., Rieder, M.J., Kruglyak, L., Nickerson, D.A., 2006. Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genetics 2, 1319–1327.

Hedrick, P.W., 1987. Gametic disequilibrium measures: Proceed with caution. Genetics 117, 331–341.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38, 226–231.

Hudson, R.R., 2001. Linkage disequilibrium and recombination. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 309–324 (Chapter 11).

Jorgenson, E., Witte, J.S., 2006. Coverage and power in genomewide association studies. American Journal of Human Genetics 78, 884–888.

Kruglyak, L., 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22, 139–144.

Lewontin, R.C., 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49, 49–67.

Lewontin, R.C., 1988. On measures of gametic disequilibrium. Genetics 120, 849–852.

Morton, N.E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y., Collins, A., 2001. The optimal measure of allelic association. Proceedings of the National Academy of Sciences USA 98, 5217–5221.

Plagnol, V., Wall, J.D., 2006. Possible ancestral structure in human populations. PLoS Genetics 2, 972–979.

Pritchard, J.K., Przeworski, M., 2001. Linkage disequilibrium in humans: Models and data. American Journal of Human Genetics 69, 1–14.

Slatkin, M., 2008. Linkage disequilibrium – Understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9, 477–485.

Terwilliger, J.D., Hiekkalinna, T., 2006. An utter refutation of the 'Fundamental Theorem of the HapMap'. European Journal of Human Genetics 14, 426–437.

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., Kidd, K.K., 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271, 1380–1387.

Weisstein, E.W., 2003. CRC Concise Encyclopedia of Mathematics, 2nd ed. Chapman & Hall/CRC, Boca Raton.

Wray, N.R., 2005. Allele frequencies and the r^2 measure of linkage disequilibrium: Impact on design and interpretation of association studies. Twin Research and Human Genetics 8, 87–94.

Zondervan, K.T., Cardon, L.R., 2004. The complex interplay among factors that influence allelic association. Nature Reviews Genetics 5, 89–100.