Published in final edited form as:

Hum Hered. 2019 ; 84(3): 127–143. doi:10.1159/000504171.

Mathematical properties of linkage disequilibrium statistics defined by normalization of the coefficient $D = p_{AB} - p_A p_B$

Jonathan T. L. Kang[*,1], **Noah A. Rosenberg**[1]

[1]Department of Biology, Stanford University, Stanford, CA 94305 USA

Abstract

Background: Many statistics for measuring linkage disequilibrium (LD) take the form of a normalization of the linkage disequilibrium coefficient *D*. Different normalizations produce statistics with different ranges, interpretations, and arguments favoring their use.

Methods: Here, to compare the mathematical properties of these normalizations, we consider five of these normalized statistics, describing their upper bounds, the mean values of their maxima over the set of possible allele frequency pairs, and the size of the allele frequency regions accessible given specified values of the statistics.

Results: We produce detailed characterizations of these properties for the statistics d and ρ, analogous to computations previously performed for r^2. We examine the relationships among the statistics, uncovering conditions under which some of them have close connections.

Conclusion: The results contribute insight into LD measurement, particularly the understanding of differences in the features of different LD measures when computed on the same data.

Keywords

allele frequencies; linkage disequilibrium; population genetics; statistical genetics; statistics

1 Introduction

Linkage disequilibrium (LD) refers to the non-random association of the alleles at a pair of genetic loci. It manifests as a deviation of observed haplotype frequencies from the frequencies expected under the assumption that alleles at the two loci associate independently. As a fundamental concept in population genetics, LD appears in a wide variety of contexts, such as association mapping and detection of natural selection [1–5].

[*]**Corresponding author:** Jonathan T. L. Kang, Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305, (650) 724-5122, jonathan.tl.kang@gmail.com.

AUTHOR CONTRIBUTIONS

J.T.L.K. and N.A.R. conceived of the study, performed the mathematical computations, analyzed the results, and wrote the manuscript. J.T.L.K. performed the data analysis.

STATEMENT OF ETHICS

The authors have no ethical conflicts to disclose.

DISCLOSURE STATEMENT

The authors have no conflicts of interest to declare.

The original measure of LD for a pair of biallelic loci, one with alleles A and a and the other with alleles B and b, was $D = p_{AB} - p_A p_B$, where p_A and p_B represent the frequencies of alleles A and B, respectively, and p_{AB} is the frequency of the two-locus haplotype containing alleles A and B [6]. The frequencies p_A and p_B can be measured for the two loci separately, each in the absence of information on the other locus, whereas evaluation of the frequency p_{AB} uses information on co-occurrence within individuals of alleles at the two loci.

Because LD is a property of a relationship between a pair of loci, values of the allele frequencies at the two loci under consideration can affect the potential strength of that relationship. This dependence is a recognized feature of LD measurement: soon after the initial development of the measure D, the quantity $|D'|$ was introduced as a normalization of D that has the same maximal value irrespective of the allele frequencies at the constituent loci [7].

Many measures of LD have been proposed, each with different arguments favoring its use [1, 3, 8–12]. For example, the popular measure r^2 [13] has the property that it can be interpreted as a squared correlation coefficient between indicator variables for the presence of allele A at the first locus and allele B at the second locus. Each allelic indicator variable is a Bernoulli trial, so that the squared covariance in the numerator of r^2, D^2, is obtained by examining the probability that both indicator variables simultaneously equal 1. Features of r^2 in a population evolving according to a standard neutral model are closely related to the population recombination rate $4N_e c$, where N_e is the effective population size and c is the recombination rate between two loci [14–15]. In addition, a calculation of the sample size necessary to detect disease association at a marker locus in linkage disequilibrium with a disease locus relies specifically on a measurement of r^2 between the marker and disease loci [3, 16].

The measure d [17] contains an asymmetry between the pair of loci that can be useful if ascertainment of haplotypes forces specific frequencies for one of the loci. This asymmetry is potentially of use in association mapping in the context of a case-control study, where the B/b locus is taken to contain the disease allele, with A/a being a marker locus [9, 18]. In this context, d can also be interpreted as the difference in the proportions of disease and normal alleles found on the same haplotype with a particular marker allele [9].

The measure ρ has been argued to be informative in a model-based perspective on LD, in which it is treated as the probability that a haplotype chosen at random descends without recombination from a population of haplotypes that excludes the aB haplotype [19]. Specifically, given a set of allele and haplotype frequencies, ρ satisfies

$$\rho \begin{bmatrix} p_B & p_A - p_B \\ 0 & 1 - p_A \end{bmatrix} + (1-\rho) \begin{bmatrix} p_A p_B & p_A(1-p_B) \\ (1-p_A)p_B & (1-p_A)(1-p_B) \end{bmatrix} = \begin{bmatrix} p_{AB} & p_{Ab} \\ p_{aB} & p_{ab} \end{bmatrix}. \quad (1)$$

A fifth measure, a normalization of r^2 termed $r^2/r^2_{\max}$ [20], has the same property as $|D'|$ that its maximum is invariant with respect to the values of p_A and p_B.

All of these normalized measures — $|D'|, r^2, d, \rho,$ and $r^2/r^2_{\max}$ — have numerators that are functions of D and denominators that are functions of the single-locus quantities p_A and p_B. The normalizations introduce different consequences for the maximal values of the statistics as functions of p_A and p_B [8, 20–21]. They also affect the symmetries of the statistics both with respect to exchanges of the two loci and with respect to exchanges of the alleles at one or both loci.

In applying LD statistics, many uses implement numerical cutoffs to assess if a desired degree of association has been met by a pair of loci, with only those locus pairs whose LD value exceeds the threshold regarded as having done so. For example, pairwise LD thresholds have been used in defining the boundaries of haplotype blocks [22–23]. They have also been applied to select tag SNP sets to assay in association studies, choosing tags by the number of non-tags with which they achieve a minimum LD cutoff and evaluating the fraction of non-tags that achieve an LD cutoff with at least one tag [24–25]. LD thresholds have also been employed for such purposes as visualizing tiered LD levels [26], pruning correlated markers in polygenic risk score calculations [27], and generating networks whose vertices represent loci and whose edges connect locus pairs with LD values exceeding a cutoff [28].

The frequent use of pairwise LD thresholds motivates studies of the implicit properties of allele frequencies forced by the thresholds, and more generally, of the way in which numerical values and interpretations of the various statistics depend on allele frequencies. This paper examines such properties and other mathematical features of the various D-based statistics. Although the statistics all range from 0 to 1, owing to their different normalizations and constraints, the meaning of a numerical value of one statistic can differ from the meaning of the same value of another statistic. Our goal is to characterize properties of the range, dependencies, and typical magnitudes of the measures, in order to assist in giving insight about values observed in empirical and theoretical studies of LD.

VanLiere and Rosenberg [20] studied the maximal value of r^2 as a function of p_A and p_B (see also [29]), in addition to considering such quantities as the mean maximal value of r^2 over the unit square for the pair of allele frequencies, the mean maximal value for r^2 over values of p_B for fixed values of p_A, and the set of permissible values of p_B given r^2 and p_A (see also [30]). With the current emphasis on rare variants in human genetics [31–32], a salient observation concerning r^2 is that if rare mutations occur at two loci on the same common haplotype in different individuals, then r^2 for the pair of loci is likely to have an extremely low value [33–34], complicating the use of r^2 in comparing LD across locus pairs. Here, we examine aspects of the mathematical properties of LD measures for each of the five normalized measures. We also consider the relationships between pairs of measures, finding that some pairs of measures are equal in particular scenarios.

2 Theory

2.1 Setting

We consider two biallelic loci, locus 1 with alleles A and a, and locus 2 with alleles B and b. The population frequencies of these alleles are then given by

$p_A, p_a = 1 - p_A, p_B,$ and $p_b = 1 - p_B$, respectively. Because both p_A and p_B lie in [0, 1], a set of frequencies can be characterized as a point in the unit square with axes p_A and p_B. For ease of notation, following VanLiere and Rosenberg [20], we split this square into octants $S_1, S_2,...,S_8$, as illustrated in Figure 1. The conditions on p_A and p_B that characterize these octants appear in Table 1. We henceforth assume that the loci are both polymorphic, so that p_A, p_a, p_B, and p_b all lie in (0, 1). The two pairs of alleles associate into four distinct haplotypes: AB, Ab, aB, and ab, with frequencies p_{AB}, p_{Ab}, p_{aB}, and p_{ab}, respectively (Table 2).

We consider parametric values for the allele frequencies, so that our interest is in LD statistics computed as functions of quantities $p_A, p_a, p_B, p_b, p_{AB}, p_{Ab}, p_{aB},$ and p_{ab}. This setting amounts to considering the statistics in an idealized setting of an infinite population.

2.2 The five normalized LD measures

As mentioned earlier, the most basic measure of LD is $D = p_{AB} - p_A p_B$, the difference between the observed frequency of the AB haplotype and its expected frequency under independence of loci 1 and 2. Expressions for D can also be formulated using each of the three other possible combinations of alleles at the two loci (Ab, aB, and ab). The four formulations all give an identical value, up to a change in sign.

If no association exists between the two loci, then we expect $p_{AB} = p_A p_B$, and hence $D = 0$. We consider several LD measures, each of which is a normalization of D. For instance, D' is obtained by normalizing D by its maximal magnitude, given the sign of D:

$$D' = \frac{D}{D_{\max}}, \quad \text{where } D_{\max} = \begin{cases} \min[p_A(1-p_B), (1-p_A)p_B] & \text{if } D > 0 \\ \min[p_A p_B, (1-p_A)(1-p_B)] & \text{if } D < 0 \end{cases}. \tag{2}$$

The r^2 measure is defined as D^2 normalized by the product of all four allele frequencies:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}. \tag{3}$$

The next two LD measures represent the two ways in which D can be normalized by the product of two of the four allele frequencies. If the two frequencies represent alleles from the same locus, then we have d, given by Nei and Li [17]:

$$d = \frac{D}{p_B(1-p_B)}. \tag{4}$$

By convention, in an association mapping setting, locus 2 is designated as a disease locus, and hence B or b is regarded as a potential disease-causing allele.

If the two frequencies instead represent alleles from different loci, then we have ρ, given by Collins and Morton [35]:

$$\rho = \frac{D}{(1 - p_A)p_B}. \quad (5)$$

Unlike D' and r^2, both d and ρ introduce an asymmetry in the pair of loci by virtue of the choice of alleles assigned to their denominators.

Lastly, as was noted by VanLiere and Rosenberg [20], the maximal value of r^2 is constrained by the values of the allele frequencies p_A and p_B. Let $r^2_{\max}$ be the maximal value of r^2 possible given p_A and p_B. The measure $r^2/r^2_{\max}$, introduced by VanLiere and Rosenberg [20], is then simply equal to r^2 normalized by $r^2_{\max}$.

2.3 Prescribed domains

The measures D' and r^2 can be applied for all values of p_A and p_B in (0,1), that is, in all octants in Figure 1. $r^2/r^2_{\max}$, being derived from r^2, also has all octants available. However, d and ρ are defined only on part of the domain (0, 1) × (0, 1). For d, because locus 2 is usually taken to be a disease locus—with one relatively rare allele—and locus 1 is the marker locus, it is assumed that $\min(p_B, p_b) \leqslant \min(p_A, p_a)$. This assumption restricts d to S_1, S_2, S_5, and S_6. For ρ, the allele frequencies are assigned labels such that $D \geqslant 0$, $p_{aB} \leqslant p_{Ab}$, $p_B \leqslant p_A$, and $p_B \leqslant 1 - p_B$ [19]. Note that $p_{aB} \leqslant p_{Ab}$ is equivalent to $p_B \leqslant p_A$, as $p_{aB} = p_B - p_{AB}$ and $p_{Ab} = p_A - p_{AB}$. Together, these conditions restrict the available octants to S_4, S_5, and S_6. Domain restrictions are summarized in Table 3, and for d and ρ, we restrict our subsequent analysis to octants in which these measures apply.

2.4 Upper bounds, mean maximum values, and accessible regions

We are interested in analyzing mathematical properties of the five LD measures. Because the magnitude of these measures is the quantity of interest, we work with the absolute values $|D'|$ and $|d|$. r^2 and $r^2/r^2_{\max}$ are always non-negative owing to the fact that D^2 is used in their expressions, and ρ is always non-negative because its definition requires $D \geqslant 0$.

We seek to determine the upper bound, mean maximal value, and accessible region for each of the five measures, given the values of p_A and p_B. The mean maximum $\mathbb{E}[m_{\max}]$ of a measure m is defined as its average maximum value, assuming (p_A, p_B) follows a bivariate uniform distribution on the permissible domain over which m applies. Its accessible region for a constant $c \in [0, 1]$, $p_m(c)$, is defined as the proportion of the domain in which the upper bound for the measure is greater than or equal to c.

To determine these mathematical properties, we first must choose a value of p_{AB} that maximizes $|D|$, because all other variables in the expressions for the five statistics are fixed given p_A and p_B. The values of p_{AB} that achieve this maximum are the same as those given by VanLiere and Rosenberg [20] for finding the upper bound on r^2 (Figure 2A), because $|D|$ is maximized if and only if D^2 is maximized. Hence, on S_1 and S_4, the maximum $|D|$ occurs if $p_{AB} = p_A + p_B - 1$; on S_2 and S_7, if $p_{AB} = p_A$; on S_3 and S_6, if $p_{AB} = p_B$; and on S_5 and S_8,

if $p_{AB} = 0$. These values appear in TABLE 1. For all five measures, the values of $\mathbb{E}[m_{\max}]$ and $p_m(c)$ appear in TABLE 4.

$|D'|$: Because $|D'|$ is simply $|D|$ normalized by its maximum value $D_{\max}$, both its upper bound and the mean maximum $\mathbb{E}[|D'|_{\max}]$ are equal to 1. Furthermore, its accessible region $p_{|D'|}(c)$ is also 1, irrespective of the value of c.

r^2: The upper bound of r^2 as a function of p_A and p_B, for each octant $S_1,...,S_8$, was calculated in eqs. 2–5 of VANLIERE AND ROSENBERG [20]. These results appear in TABLE 1, and a contour plot of $r^2_{\max}$ is reproduced in FIGURE 2A. In addition, VANLIERE AND ROSENBERG [20] derived the mean maximum of r^2, obtaining $\mathbb{E}[r^2_{\max}] = 2\pi^2/3 - 4(\ln 2)^2 + 4\ln 2 - 7 \approx 0.43051$, as well as its accessible region, which is

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

A plot of $p_{r^2}(c)$ appears in FIGURE 3.

$|d|$: By substituting the appropriate value of p_{AB} into the expression for $|d|$, we obtain $|d|_{\max}$ as a function of p_A and p_B in octants S_1, S_2, S_5, and S_6:

$$S_1: |d|_{\max}(p_A, p_B) = \frac{|p_A + p_B - 1 - p_A p_B|}{p_B(1-p_B)} = \frac{1-p_A}{p_B} \tag{7}$$

$$S_2: |d|_{\max}(p_A, p_B) = \frac{p_A - p_A p_B}{p_B(1-p_B)} = \frac{p_A}{p_B} \tag{8}$$

$$S_5: |d|_{\max}(p_A, p_B) = \frac{|0 - p_A p_B|}{p_B(1-p_B)} = \frac{p_A}{1-p_B} \tag{9}$$

$$S_6: |d|_{\max}(p_A, p_B) = \frac{p_B - p_A p_B}{p_B(1-p_B)} = \frac{1-p_A}{1-p_B}. \tag{10}$$

These results are summarized in TABLE 1. FIGURE 2B shows a contour plot of $|d|_{\max}$ in S_1, S_2, S_5, and S_6, combining eqs. 7–10. We note some similarities, as well as some differences, with the plot of $r^2_{\max}$ in FIGURE 2A. Examining the characteristic X-shape of the figure, we see that $|d|_{\max}$ can equal 1 if and only if the allele frequencies are identical at the two loci, $p_A = p_B$ or $p_A = p_b$, as is the case with $r^2_{\max}$. However, instead of having a symmetric shape over all octants, $|d|_{\max}$ is symmetric with respect to an exchange of p_A and p_a or p_B and p_b, but not with respect to an exchange of p_A and p_B (and thus also p_a and p_b) or p_A and p_b (and thus also p_a and p_B). Its shape is symmetric over S_1, S_2, S_5, and S_6, the four octants on

which it can be calculated. Unlike $r^2_{\max}$, $|d|_{\max}$ does not approach 0 as p_B approaches either 0 or 1. This feature enables $|d|$ to maintain a considerable range of allowable values, even if the minor allele frequency (MAF) at locus 2 is low, as is likely the case in a mapping study in which locus 2 is regarded as causal for a rare disease.

We can quantify the difference in range for $|d|$ and r^2 by comparing the mean maximum value of $|d|$ to that of r^2. First, we compute the volume V_2, which we define to be the volume of $|d|_{\max}$ over the octant S_2:

$$\begin{aligned} V_2 &= \int_{\frac{1}{2}}^{1} \int_{p_A}^{1} \frac{p_A}{p_B} dp_B dp_A \\ &= \frac{3}{16} - \frac{1}{8}\ln 2 \approx 0.10086. \end{aligned} \tag{11}$$

Owing to symmetry, V_2 is equal to corresponding values V_1, V_5, and V_6. Assuming a uniform joint distribution of p_A and p_B over the octants S_1, S_2, S_5, and S_6, and noting that these octants have a total area $\frac{1}{2}$, the mean maximum value of $|d|_{\max}$ is

$$\begin{aligned} \mathbb{E}[|d|_{\max}] &= 4V_2 / \frac{1}{2} \\ &= \frac{3}{2} - \ln 2 \approx 0.80685. \end{aligned} \tag{12}$$

This value exceeds $\mathbb{E}[r^2_{\max}] \approx 0.43051$ derived by VanLiere and Rosenberg [20] under the same assumption of a uniform distribution on the domain, suggesting that $|d|$ can achieve a high magnitude over a considerably larger portion of the allele frequency space than is seen for r^2. To quantify this difference, we calculate the accessible region $p_{|d|}(c)$. We first focus on S_6, and extend the result to the remaining octants using symmetry.

Let A_6 denote the area of the portion of S_6 in which $|d|_{\max} \geq c$. Using eq. 10, the portion of S_6 in which $|d|_{\max} \geq c$ satisfies $p_B \geq (p_A + c - 1)/c$. We now set up an integral to calculate the complement of the desired area, the area of the portion of S_6 in which $|d|_{\max} \geq c$. Observe from Figure 2B that in S_6, for $c \geq \frac{1}{2}$, the horizontal plane $|d|_{\max} = c$ intersects the upper bound at $p_A = 1 - c$ if $p_B = 0$. Therefore, we have

$$\begin{aligned} \frac{1}{8} - A_6 &= \int_{1-c}^{\frac{1}{2}} \int_{0}^{\frac{p_A + c - 1}{c}} 1 dp_B dp_A = \frac{(2c-1)^2}{8c} \\ A_6 &= \frac{(4c-1)(1-c)}{8c}. \end{aligned} \tag{13}$$

For $c \leq \frac{1}{2}$, all points in octants S_1, S_2, S_5, and S_6 have $|d|_{\max} \geq c$ (Figure 2B). Hence, in this situation, $p_{|d|}(c)$ is simply 1. For $c \geq \frac{1}{2}$, applying eq. 13,

$$p_{|d|}(c) = 4A_6/\frac{1}{2} = \frac{(4c-1)(1-c)}{c}. \quad (14)$$

The piecewise function $p_{|d|}(c)$ appears in FIGURE 3, alongside a plot of $p_{r^2}(c)$ from eq. 6. The permissible fraction of the frequency space for $|d|$ decreases more slowly as a function of c than does the corresponding function for r^2.

ρ: For the upper bound on ρ, the following conditions all must be satisfied when assigning labels to the alleles: $D \geqslant 0, p_B \leqslant p_A,$ and $p_B \leqslant 1 - p_B$ [19, 36]. The latter two conditions imply that ρ applies only in S_4, S_5, and S_6. The condition $p_B \leqslant p_A$ implies $p_B - p_A p_B \leqslant p_A - p_A p_B$, which in turn implies $(1 - p_A)p_B \leqslant p_A(1 - p_B)$. This result, in addition to the requirement that $D \geqslant 0$, indicates that ρ is exactly equal to $|D'|$ under the conditions in which ρ applies. Consequently, the upper bound of ρ, its mean maximum $\mathbb{E}[\rho_{\max}]$, and its accessible region $p_\rho(c)$ all equal 1.

$r^2/r^2_{\max}$: The upper bound, the mean maximum $\mathbb{E}[\{r^2/r^2_{\max}\}_{\max}]$. and the accessible region $p_{r^2/r^2_{\max}}(c)$ all equal 1, by the definition of the statistic as r^2 normalized by $r^2_{\max}$.

2.5 Mean maximum of r^2 and $|d|$ under a beta distribution

In SECTION 2.4, we examined the mean maximum of the five measures, assuming (p_A, p_B) follows a bivariate uniform distribution. For r^2 and $|d|$, the two measures that do not have a mean maximum of 1, we can also calculate their mean maximum value under less restrictive assumptions. We now assume p_A and p_B follow independent beta distributions. To preserve symmetry between loci and exchangeability of the alleles at a locus, we consider $p_A, p_B \sim$ Beta-(α, α), and compute $\mathbb{E}[|d|_{\max}]$ and $\mathbb{E}[r^2_{\max}]$ as functions of α.

By analogy with eq. 12, again using octant S_2, we can set up an integral for $\mathbb{E}[|d|_{\max}]$:

$$\begin{aligned}\mathbb{E}[|d|_{\max}] &= 8\int_{\frac{1}{2}}^{1}\int_{p_A}^{1}\frac{p_A}{p_B}\frac{p_A^{\alpha-1}(1-p_A)^{\alpha-1}}{\mathrm{B}(\alpha,\alpha)}\frac{p_B^{\alpha-1}(1-p_B)^{\alpha-1}}{\mathrm{B}(\alpha,\alpha)}dp_B dp_A \\ &= \frac{8}{[\mathrm{B}(\alpha,\alpha)]^2}\int_{\frac{1}{2}}^{1}\int_{p_A}^{1}\frac{p_A}{p_B}\left(p_A - p_A^2\right)^{\alpha-1}\left(p_B - p_B^2\right)^{\alpha-1}dp_B dp_A.\end{aligned} \quad (15)$$

Here, $\mathrm{B}(\alpha, \alpha) = [\Gamma(\alpha)]^2/\Gamma(2\alpha)$. To compute $\mathbb{E}\left[r^2_{\max}\right]$, we use $r^2_{\max}$ on S_2 (TABLE 1):

$$\begin{aligned}\mathbb{E}\left[r^2_{\max}\right] &= 8\int_{\frac{1}{2}}^{1}\int_{p_A}^{1}\frac{p_A(1-p_B)}{(1-p_A)p_B}\frac{p_A^{\alpha-1}(1-p_A)^{\alpha-1}}{\mathrm{B}(\alpha,\alpha)}\frac{p_B^{\alpha-1}(1-p_B)^{\alpha-1}}{\mathrm{B}(\alpha,\alpha)}dp_B dp_A \\ &= \frac{8}{[\mathrm{B}(\alpha,\alpha)]^2}\int_{\frac{1}{2}}^{1}\int_{p_A}^{1}\frac{p_A(1-p_B)}{(1-p_A)p_B}\left(p_A - p_A^2\right)^{\alpha-1}\left(p_B - p_B^2\right)^{\alpha-1}dp_B dp_A.\end{aligned} \quad (16)$$

We evaluate eqs. 15 and 16 numerically, using values of α ranging from 0.2 to 5. The results appear in FIGURE 4. Low values of α imply that the distribution of allele frequencies is skewed toward loci with a low MAF, whereas high α values correspond to greater density in loci with a high MAF. Note that setting $\alpha = 1$ gives a uniform distribution and recovers the values derived in SECTION 2.4 for the case of $p_A, p_B \sim$ Uniform-(0,1).

From FIGURE 4, we observe that $\mathbb{E}[r^2_{\max}]$ varies considerably as a function of the allele frequency distribution, whereas $\mathbb{E}[|d|_{\max}]$ is more stable as α changes.

2.6 The five measures as functions of p_{AB} for fixed p_A, p_B

For most of our subsequent calculations, to facilitate comparison, we restrict our analysis to values of p_A and p_B in S_6, as all five measures have S_6 in their prescribed domains. Any point not in S_6 can be mapped to a point in S_6 by performing one or more of a set of transformations: (i) reflection over $p_B = \frac{1}{2}$ (corresponding to exchanging the p_A and p_a labels), (ii) reflection over $p_A = \frac{1}{2}$ (exchanging the p_B and p_b labels), and (iii) reflection over the $p_A = p_B$ line (exchanging the p_A and p_B labels, and thus also the p_a and p_b labels).

We first compare how each measure varies with the haplotype frequency p_{AB}. FIGURE 5 illustrates $|D'|, r^2, |d|,$ and $r^2/r^2_{\max}$ as functions of the haplotype frequency p_{AB}, for fixed values of p_A and p_B. In this analysis, ρ is omitted because under the conditions in which it applies, it is exactly equal to $|D'|$. Each of the measures has a value of 0 in the case of linkage equilibrium, at which $p_{AB} = p_A p_B$. Using $p_{AB} = p_A p_B$ as a reference point, we can split the plots for each of the measures into two portions: the right arm, where $p_{AB} \geqslant p_A p_B$ (or $D \geqslant 0$, corresponding to the case with an excess of haplotypes containing both minor alleles), and the left arm, where $p_{AB} \leqslant p_A p_B$ (or $D \leqslant 0$, corresponding to the case with a deficit of haplotypes containing both minor alleles).

$|D'|$: $|D'|$ varies linearly with p_{AB}. However, its left and right arms are in general not symmetric about the line $p_{AB} = p_A p_B$; the absolute value of the derivative of $|D'|$ as a function of p_{AB} differs in the two arms. This phenomenon results from the different normalizations applied in obtaining D', depending on whether D is positive or negative. The left and right arms of $|D'|$ are symmetric only if $p_A = \frac{1}{2}, p_B = \frac{1}{2}$, or both. The value of $|D'|$ can always reach 1 irrespective of whether the haplotype containing both minor alleles is in excess or in deficit; in general, no such result holds for the other three measures.

r^2: r^2 varies quadratically as a function of p_{AB}, with the measure increasing at a faster rate the further p_{AB} is from $p_A p_B$. As reported by VANLIERE AND ROSENBERG [20], in S_6, r^2 can only reach 1 if $p_A = p_B$, and even then only if an excess of haplotypes containing both minor alleles occurs. Finally, ignoring the truncation imposed by the lower limit of $p_{AB} = 0$, the arms of r^2 are symmetric with respect to the line $p_{AB} = p_A p_B$.

$|d|$: $|d|$, like $|D'|$, varies linearly as a function of p_{AB}. However, unlike for $|D'|$, the left and right arms of $|d|$ are symmetric, and they have the same absolute value of the derivative as a function of p_{AB}. This pattern occurs because $|d|$ does not necessarily have to reach 1 at the

points where p_{AB} lies at its maximum or minimum values, given p_A and p_B. Like r^2, $|d|$ can only reach 1 on its right arm if $p_A = p_B$. As a result, $|D'| = |d|$ if $p_A = p_B$ and $D \geqslant 0$, as can be observed from FIGURE 5.

$r^2/r^2_{\max}$: $r^2/r^2_{\max}$ varies quadratically as a function of p_{AB}, but increases more quickly compared to r^2 as p_{AB} moves away from $p_A p_B$. It can always reach 1 irrespective of the values of p_A and p_B but does so only if the haplotype containing both minor alleles is in excess. If $p_A = p_B$, then $r^2/r^2_{\max} = r^2$, as the maximum of r^2 is 1 in this case.

Comparison: In general, for all values of p_A, p_B, and p_{AB}, $|D'| \geqslant |d| \geqslant r^2$. To demonstrate this, we first show that $|D'| \geqslant |d|$. Consider $D > 0$, where D is normalized by $\min[p_A(1 - p_B), (1 - p_A)p_B]$ in the calculation of $|D'|$. If $p_A \geqslant p_B$, then $p_A(1 - p_B) \geqslant p_B(1 - p_B) \geqslant (1 - p_A)p_B$, but if $p_A \leqslant p_B$, then $p_A(1 - p_B) \leqslant p_B(1 - p_B) \leqslant (1 - p_A)p_B$. In either case, $p_B(1 - p_B) \geqslant \min[p_A(1 - p_B), (1 - p_A)p_B]$, and therefore $|D'| \geqslant |d|$.

similar calculation shows also that $|D'| \geqslant |d|$ if $D < 0$.

To see that $|d| \geqslant r^2$ where $|d|$ applies (S_1, S_2, S_5, S_6), we show $r^2/|d| \leqslant 1$. We have

$$\frac{r^2}{|d|} = \frac{|D|}{p_A(1 - p_A)} = \frac{|p_{AB} - p_A p_B|}{p_A(1 - p_A)}. \tag{17}$$

Consider S_6, where $p_A \geqslant p_B$ and $p_{AB} = p_B$ maximizes $|D|$ (TABLE 1):

$$\frac{|p_{AB} - p_A p_B|}{p_A(1 - p_A)} \leqslant \frac{p_B - p_A p_B}{p_A(1 - p_A)} = \frac{p_B}{p_A}. \tag{18}$$

Because $p_A \geqslant p_B$, $r^2/|d| \leqslant 1$. Eq. 18 and other similar calculations for S_1, S_2, and S_5 show that $|d| \geqslant r^2$ for all possible values of p_A and p_B.

2.7 Mean LD values given fixed p_A and p_B

In SECTION 2.4, we have described upper bounds of the five measures given values of p_A and p_B. We have also computed the mean of the maximum value. Next, we specify a distribution on p_{AB} and calculate the mean values of the measures over the possible domain.

For this section, we again use S_6; analogous results for other octants follow the same framework. Because $p_B \leqslant p_A$ in S_6, p_{AB} lies in $[0, p_B]$. If we further assume $p_{AB} \sim$ Uniform-$(0, p_B)$, then we can compute the mean value of each LD measure as a function of p_A and p_B. For completeness, associated variances are derived in the APPENDIX.

$|D'|$: $D > 0$ if $p_{AB} > p_A p_B$, and $D < 0$ if $p_{AB} < p_A p_B$. We split the integral for the mean to account for both cases. If p_{AB} is distributed uniformly on $(0, p_B)$, then $\mathbb{E}[|D'|]$ has a constant value of $\frac{1}{2}$, and does not depend on p_A or p_B.

$$\mathbb{E}[|D'|] = \frac{1}{p_B}\left[\int_0^{p_A p_B} \frac{p_A p_B - p_{AB}}{p_A p_B} dp_{AB} + \int_{p_A p_B}^{p_B} \frac{p_{AB} - p_A p_B}{(1-p_A)p_B} dp_{AB}\right] = \frac{1}{2}. \tag{19}$$

r^2: For r^2, it is not necessary to split the integral.

$$\mathbb{E}\left[r^2\right] = \frac{1}{p_B}\int_0^{p_B} \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)} dp_{AB} = \frac{p_B\left(1 - 3p_A + 3p_A^2\right)}{3p_A(1-p_A)(1-p_B)}. \tag{20}$$

The result is plotted in Figure 6A.

$|d|$: For d, we again split the integral as we did for $|D'|$.

$$\mathbb{E}[|d|] = \frac{1}{p_B}\left[\int_0^{p_A p_B} \frac{p_A p_B - p_{AB}}{p_B(1-p_B)} dp_{AB} + \int_{p_A p_B}^{p_B} \frac{p_{AB} - p_A p_B}{p_B(1-p_B)} dp_{AB}\right] = \frac{1 - 2p_A + 2p_A^2}{2(1-p_B)}. \tag{21}$$

The result is plotted in Figure 6B.

$r^2/r^2_{\max}$: The result appears in Figure 6C. Note that in octant S_6, $\mathbb{E}\left[r^2/r^2_{\max}\right]$ is a function of only p_A (or in general, allele frequencies at the locus with the higher MAF).

$$\mathbb{E}\left[r^2/r^2_{\max}\right] = \frac{1}{p_B}\int_0^{p_B} \frac{\frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}}{\frac{(1-p_A)p_B}{p_A(1-p_B)}} dp_{AB} = \frac{1 - 3p_A + 3p_A^2}{3(1-p_A)^2}. \tag{22}$$

$\mathbb{E}\left[r^2/r^2_{\max}\right]$ varies less across the domain for p_A than do $\mathbb{E}[r^2]$ and $\mathbb{E}[|d|]$.

2.8 Constraints on one allele frequency given an LD value and the allele frequency at the other locus

In this section, we examine for each of the five LD measures the allowable values of one allele frequency (either p_A or p_B) while fixing the allele frequency at the other locus and specifying the value of the LD measure. Owing to symmetry in the loci, for $|D'|, r^2,$ and $r^2/r^2_{\max}$, fixing p_A is equivalent to fixing p_B, and we need only examine one case. For $|d|$ and ρ, owing to asymmetries in the formulation of the measures, two cases must be considered. For the calculations in this section, we assume for convenience that p_A and p_B

are the minor allele frequencies (octants S_6 and S_7), and that the constraints on the major allele frequencies will follow accordingly. The results of this section are summarized in TABLE 5.

Allowable values of p_B, given $|D'|$ and p_A: Having a value of $|D'|$ and a value of p_A does not constrain p_B, as all values of $|D'|$ between 0 and 1 are accessible given a pair of allele frequencies p_A and p_B. This result can be shown from the fact that given p_A and p_B, $|D'|$ is a continuous rational function of p_{AB}. Because 0 and 1 are the extreme values of $|D'|$, by the intermediate value theorem, $|D'|$ can take on any value in [0, 1] (also see FIGURE 5).

Allowable values of p_B, given r^2 and p_A: Assuming that $p_A, p_B \leqslant \frac{1}{2}$, given p_A and r^2, the constraint on p_B is

$$\frac{r^2 p_A}{1 + r^2 p_A - p_A} \leqslant p_B \leqslant \min\left(\frac{1}{2}, \frac{p_A}{r^2 - r^2 p_A + p_A}\right). \tag{23}$$

This result has been previously reported in eqs. 10 and 11 of VANLIERE AND ROSENBERG [20] and TABLE 2 of WRAY [30].

Allowable values of p_B, given $|d|$ and p_A: Because $|d|$ is not symmetric in the two loci, we first assume $|d|$ and p_A are specified, and solve for the range of p_B. Recalling that d applies only in S_1, S_2, S_5, and S_6, and assuming $p_A, p_B \leqslant \frac{1}{2}$, we consider S_6. From eq. 10,

$$\begin{aligned} |d| &\leqslant \frac{1 - p_A}{1 - p_B} \\ p_B &\geqslant \frac{p_A + |d| - 1}{|d|}. \end{aligned} \tag{24}$$

Taking into account $0 \leqslant p_B \leqslant p_A \leqslant \frac{1}{2}$, we have

$$\max\left(0, \frac{p_A + |d| - 1}{|d|}\right) \leqslant p_B \leqslant p_A. \tag{25}$$

Allowable values of p_A, given $|d|$ and p_B: Next, we assume $|d|$ and p_B are specified, and solve for the range of p_A. In S_6, from eq. 10,

$$\begin{aligned} |d| &\leqslant \frac{1 - p_A}{1 - p_B} \\ p_A &\leqslant 1 - |d| + |d| p_B. \end{aligned} \tag{26}$$

Taking into account $0 \leqslant p_B \leqslant p_A \leqslant \frac{1}{2}$, we have

$$p_B \leqslant p_A \leqslant \min\left(\frac{1}{2}, 1 - |d| + |d| p_B\right). \quad (27)$$

The upper bound here corresponds to the upper bound for the "frequency difference" measure of LD reported in TABLE 2 of WRAY [30], noting that labels p_A and p_B are reversed in that study. However, a difference exists between the reported lower bounds, which can be attributed to the fact that $p_A \leqslant \frac{1}{2}$ is not mandated by WRAY [30].

Allowable values of p_B, given ρ and p_A: Because all values of ρ in [0, 1] can be reached with any given set of allele frequencies in the permissible domain, no additional constraint exists on p_B given ρ and p_A.

Allowable values of p_A, given ρ and p_B: For the same reason as in the case in which p_A is instead specified, no additional constraint exists on p_A given ρ and p_B.

Allowable values of p_B, given $r^2/r^2_{\max}$ and p_A: Being given a value of $r^2/r^2_{\max}$ and p_A does not constrain the values p_B can take, as all values of $r^2/r^2_{\max}$ in [0, 1] are accessible for a given set of allele frequencies p_A and p_B.

3 DATA ILLUSTRATION

We now examine how LD distributions from data, as given by the various measures, relate to our bounds. We use the 1000 Genomes Project (data at http://csg.sph.umich.edu/abecasis/MACH/download/1000G-PhaseI-Interim.html), considering LD values on chromosome 22 of its pooled European population, consisting of 381 individuals: 87 Utah residents of Northern and Western European ancestry, 93 Finnish from Finland, 89 British from England and Scotland, 14 Iberians from Spain, and 98 Toscani from Italy. To ensure inclusion of locus pairs with substantial LD, calculations are restricted to pairs of loci that lie at most 1,000 base pairs apart. Once again, for ease of comparison, all pairs of allele frequency values outside S_6 are mapped to corresponding points within S_6.

3.1 Pairs of loci for which $|D'| = 1$

From 225,159 loci biallelic in the European data (of 494,975 loci in total), we obtain 1,742,020 pairs of loci separated by at most 1,000 base pairs. Of these, 1,465,140 pairs have $|D'| = 1$, indicating the presence of only two or three of the four possible haplotypes. Recalling that $|D'|$ can reach 1 either if an excess or a deficit of haplotypes containing both minor alleles occurs, 349,837 locus pairs belong to the former case, and 1,115,303 to the latter.

r^2: We now examine on S_6 how r^2 is distributed in relation to the upper bound. If $|D'| = 1$ for a pair of loci, then the r^2 value lies on one of two surfaces. If $|D'| = 1$ as the result of an excess of haplotypes containing both minor alleles, corresponding to $p_{AB} = p_B$, then r^2 lies on the surface that defines the upper bound on S_6, or

$$r^2 = \frac{(1 - p_A)p_B}{p_A(1 - p_B)}, \tag{28}$$

as given by eq. 4 in VanLiere and Rosenberg [20]. In S_6, with a deficit of haplotypes containing both minor alleles, $|D'| = 1$ is achieved if $p_{AB} = 0$, which results in the r^2 surface

$$r^2 = \frac{(0 - p_A p_B)^2}{p_A(1 - p_A)p_B(1 - p_B)} = \frac{p_A p_B}{(1 - p_A)(1 - p_B)}. \tag{29}$$

The surfaces and data points for these two cases appear in Figures 7A and 7B.

$|d|$: As was seen with r^2, $|d|$ for a locus pair lies on one of two surfaces if $|D'| = 1$ (Figures 7C and 7D). Once again, if $p_{AB} = p_B$, then the associated surface is the upper bound of $|d|$ on S_6, as in eq. 10. If $p_{AB} = 0$, then the points lie on

$$|d| = \frac{|0 - p_A p_B|}{p_B(1 - p_B)} = \frac{p_A}{1 - p_B}. \tag{30}$$

$r^2/r^2_{\max}$: Finally, we repeat the analysis for $r^2/r^2_{\max}$ values. If $p_{AB} = p_B$, then $r^2 = r^2_{\max}$, and therefore $r^2/r^2_{\max} = 1$. If instead $p_{AB} = 0$, then using eqs. 28 and 29, we obtain

$$\frac{r^2}{r^2_{\max}} = \frac{\frac{p_A p_B}{(1 - p_A)(1 - p_B)}}{\frac{(1 - p_A)p_B}{p_A(1 - p_B)}} = \left(\frac{p_A}{1 - p_A}\right)^2 \tag{31}$$

The surfaces and points corresponding to these two cases appear in Figures 7E and 7F.

3.2 Pairs of loci for which $|D'| < 1$

In the special case in which $|D'| = 1$ for a pair of loci, we have seen that corresponding values for r^2, $|d|$, and $r^2/r^2_{\max}$ lie on well-defined surfaces. Although most locus pairs from our data fall within this category, for 276,880 of 1,742,020 pairs, $|D'| < 1$. For these pairs, we examine how the values of $|D'|$, r^2, $|d|$, and $r^2/r^2_{\max}$ are distributed within their ranges.

Recognizing that the four measures can exhibit different distribution patterns at different allele frequencies, we sample pairs of loci for which p_A and p_B have MAF values within four specified ranges, representing very low, low, intermediate, and high MAF: (0, 0.02], [0.04, 0.06], [0.24, 0.26], and [0.44, 0.46]. Distributions of values for the measures appear as a series of histograms in Figure 8, with each panel representing one pair of allele frequency ranges for p_A and p_B; because $p_B \leqslant p_A$ in S_6, only 10 of the 16 possible combinations of joint allele frequency ranges are possible.

From Figure 8, we can observe a few properties of the distributions. First, in accordance with our theoretical results, the range of values for r^2 and $|d|$ does not extend to 1 in the

panels that are off the diagonal, where $p_A = p_B$ is not possible. In particular, the limitation on the range of r^2 is more pronounced than that of $|d|$, when comparing within similar allele frequency ranges. This constraint also results in a large number of r^2 values being close to 0, especially if p_B is small (bottom row).

In addition, although we selected only locus pairs at most 1,000 bp apart, if p_B is small, then relatively few pairs have a high LD value. This result holds especially for r^2, but also for measures such as $|D'|$ and $r^2/r^2_{\max}$ that always have upper bound 1. If one of the loci has a low MAF, then small changes in the haplotype frequency can have large effects on LD measures, especially those normalized to potentially reach 1 irrespective of the marginal allele frequencies (see also the $p_B = 0.1$ panels of FIGURE 5).

Comparing scenarios on the diagonal, where it is possible in principle for all four measures to achieve high LD values, we see that high LD values are more frequently observed for high and intermediate MAF than for low and very low MAF. For pairs of loci with low MAF, it is unusual for haplotypes to contain the rare allele at both loci, as the rare alleles likely result from relatively recent mutations that have taken place on the same common haplotype, but in different individuals. Thus, because the rare alleles are unlikely to co-occur, the nature of evolutionary descent makes it improbable that the LD-maximizing scenario that couples the rare variants will obtain. These considerations support a cautious perspective when interpreting LD measures in the case that one or both loci have a low MAF.

4 DISCUSSION

In this paper, we have described the domains of five LD measures that are defined by normalizations of D or its square, with a function of p_A and p_B in the denominator. Based on these domains, we have calculated the upper bound, mean maximum, and accessible region for each of the five measures. Three of the measures ($|D'|$, ρ, and $r^2/r^2_{\max}$) can be considered "unrestricted," in that their upper bound, mean maximum, and accessible region are all equal to 1. However, for the remaining two measures (r^2 and $|d|$), these values depend on the allele frequencies of the pair of loci under consideration.

For each of the five measures, its description, proposed usages, and mathematical properties are summarized in TABLE 6. The table provides examples illustrating how a measure's mathematical properties can inform its use. For instance, $|d|$ allows for a theoretically wider range of values compared to r^2, with a mean maximum of 0.80685 compared to 0.43051 for r^2. The increased range of $|d|$ is evident in the analysis of genetic data, which suggests that empirical $|d|$ values are more differentiated than corresponding r^2 values (FIGURE 8).

In a sense, $|d|$ can be considered a measure that is "intermediate" between $|D'|$ and r^2. First, its value always lies between r^2 and $|D'|$. It also has properties in common each with r^2 and $|D'|$. Like $|D'|$, given p_A and p_B, $|d|$ varies linearly as a function of p_{AB}, possessing a property that r^2 does not share. However, like r^2 but unlike $|D'|$, $|d|$ is symmetric in p_{AB} around the linkage equilibrium value $p_{AB} = p_A p_B$ (FIGURE 5).

We have also identified situations in which some of these measures are equal to one another. Among the measures, ρ uniquely requires $D \geqslant 0$. This requirement, along with the conditions $p_B \leqslant p_A$ and $p_B \leqslant 1 - p_B$, can be satisfied by (i) reflection over $p_A = \frac{1}{2}$ (exchanging the p_A and p_a labels), (ii) reflection over $p_B = \frac{1}{2}$ (exchanging p_B and p_b), or both. Under these prescribed conditions for the use of ρ, it exactly equals $|D'|$, a fact that had also been noted by Shete [36] and Mangin *et al.* [37]. Furthermore, $|d| = |D'|$ if $p_A = p_B$ and the haplotype containing both minor alleles is in excess. This result can be seen by observing the right arms of $|D'|$ and $|d|$ in plots along the diagonal of FIGURE 5, and also from noting that in the right arm, where $D \geqslant 0$, the normalizations $\min[p_A(1 - p_B), (1 - p_A)p_B]$ and $p_B(1 - p_B)$ for $|D'|$ and $|d|$, respectively, agree if $p_A = p_B$.

By quantifying the degree to which values for the different LD statistics change in response to shifts in allele and haplotype frequencies, the results provide context to the use of LD thresholds in various statistical genetics applications. Some uses impose thresholds in LD measures alongside minimal-MAF cutoffs [24–25, 27], and our results can be used to understand the behavior of the statistics in permissible ranges specified by simultaneous LD and MAF thresholds. Additionally, the results are useful for low-MAF loci, for which the rare alleles are unlikely to occur on the same haplotype. In particular, they illustrate that r^2 is the most tightly constrained measure (FIGURE 5, eq. 18), so that other measures might provide a broader range of values when computing LD statistics for loci with rare variants.

We note that we have focused on parametric aspects of LD measures rather than LD estimated from samples. Sampling properties can be examined, both in models that view alleles as draws from a parametric allele frequency distribution and in coalescent perspectives whose allele frequencies represent outcomes of a generative model (e.g. [38–40]). The functional forms of estimators can then potentially be combined with bounds on parametric LD measures to produce corresponding bounds on the estimators (e.g. [41, p. 1590]).

Many other measures of LD exist that are not included in our analysis [1, 3, 8–11]. Other measures are sometimes normalized by a quantity that includes a haplotype frequency, rather than a function of allele frequencies only, and thus do not lend themselves well to the framework in this paper. Similarly, LD measures used specifically in cases pertaining to multiallelic loci, such as the multiallelic $|D'|$ [8, 42–43], require additional parameters. Of the LD measures that are described by a ratio of a function of D to a product of allele frequencies for biallelic loci, we have taken a comprehensive look at the most natural statistics with that form.

We initially assumed Uniform-(0,1) distributions on the allele frequencies to perform computations for the mean maximum of the various measures. This choice, as in VANLIERE AND ROSENBERG [20], permits us to obtain mathematical insight into those measures across their prescribed ranges. In some applications, weighted distributions, such as the Beta-(α, α) distribution we subsequently used, can be applied in place of the uniform distribution.

Acknowledgments

We thank M. Edge for discussions.

FUNDING SOURCES

We acknowledge NIH grant R01 HG005855 for support.

Appendix

In this appendix, we provide the variances of $|D'|, r^2, |d|,$ and $r^2/r^2_{\max}$, under the assumptions in SECTION 2.7. For computing the variances, we use eqs. 19, 20, 21, and 22 to supply the means $\mathbb{E}[|D'|]$, $\mathbb{E}[r^2]$, $\mathbb{E}[|d|]$, and $\mathbb{E}[r^2/r^2_{\max}]$, respectively.

$$\begin{aligned}\mathbb{E}\left[|D'|^2\right] &= \frac{1}{p_B}\left[\int_0^{p_A p_B}\frac{(p_A p_B - p_{AB})^2}{p_A^2 p_B^2}dp_{AB} + \int_{p_A p_B}^{p_B}\frac{(p_{AB} - p_A p_B)^2}{(1-p_A)^2 p_B^2}dp_{AB}\right] \\ &= \frac{1}{3}.\end{aligned} \tag{32}$$

$$\mathrm{Var}\left[|D'|\right] = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}. \tag{33}$$

$$\begin{aligned}\mathbb{E}\left[r^4\right] &= \frac{1}{p_B}\int_0^{p_B}\frac{(p_{AB} - p_A p_B)^4}{p_A^2(1-p_A)^2 p_B^2(1-p_B)^2}dp_{AB} \\ &= \frac{\left[p_A^5 + (1-p_A)^5\right]p_B^2}{5p_A^2(1-p_A)^2(1-p_B)^2}.\end{aligned} \tag{34}$$

$$\begin{aligned}\mathrm{Var}\left[r^2\right] &= \frac{\left[p_A^5 + (1-p_A)^5\right]p_B^2}{5p_A^2(1-p_A)^2(1-p_B)^2} - \left[\frac{\left(1 - 3p_A + 3p_A^2\right)p_B}{3p_A(1-p_A)(1-p_B)}\right]^2 \\ &= \frac{\left(4 - 15p_A + 15p_A^2\right)p_B^2}{45p_A^2(1-p_A)^2(1-p_B)^2}.\end{aligned} \tag{35}$$

$$\begin{aligned}\mathbb{E}\left[|d|^2\right] &= \frac{1}{p_B}\int_0^{p_B}\frac{(p_{AB} - p_A p_B)^2}{p_B^2(1-p_B)^2}dp_{AB} \\ &= \frac{1 - 3p_A + 3p_A^2}{3(1-p_B)^2}.\end{aligned} \tag{36}$$

$$\begin{aligned}\mathrm{Var}[|d|] &= \frac{1-3p_A+3p_A^2}{3(1-p_B)^2} - \left[\frac{1-2p_A+2p_A^2}{2(1-p_B)}\right]^2 \\ &= \frac{1-12p_A^2+24p_A^3-12p_A^4}{12(1-p_B)^2}.\end{aligned} \tag{37}$$

$$\begin{aligned}\mathbb{E}\left[\left(r^2/r^2_{\max}\right)^2\right] &= \frac{1}{p_B}\int_0^{p_B}\left[\frac{\frac{(p_{AB}-p_Ap_B)^2}{p_A(1-p_A)p_B(1-p_B)}}{\frac{(1-p_A)p_B}{p_A(1-p_B)}}\right]^2 dp_{AB} \\ &= \frac{p_A^5+(1-p_A)^5}{5(1-p_A)^4}.\end{aligned} \tag{38}$$

$$\begin{aligned}\mathrm{Var}\left[r^2/r^2_{\max}\right] &= \frac{p_A^5+(1-p_A)^5}{5(1-p_A)^4} - \left[\frac{1-3p_A+3p_A^2}{3(1-p_A)^2}\right]^2 \\ &= \frac{4-15p_A+15p_A^2}{45(1-p_A)^4}.\end{aligned} \tag{39}$$

References

[1]. Hudson RR: Linkage disequilibrium and recombination; in Balding DJ, Bishop M, Cannings C (eds): Handbook of Statistical Genetics. Chichester, Wiley, 2001, pp 309–324.

[2]. Nordborg M, Tavaré S: Linkage disequilibrium: what history has to tell us. Trends Genet 2002; 18: 83–90. [PubMed: 11818140]

[3]. McVean G: Linkage disequilibrium, recombination and selection; in Balding DJ, Bishop M, Cannings C (eds): Handbook of Statistical Genetics, ed 3 Chichester, Wiley, 2007, pp 909–944.

[4]. Slatkin M: Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 2008; 9: 477–485. [PubMed: 18427557]

[5]. Weir BS: Linkage disequilibrium and association mapping. Annu Rev Genomics Hum Genet 2008; 9: 129–142. [PubMed: 18505378]

[6]. Lewontin RC, Kojima K: The evolutionary dynamics of complex polymorphisms. Evolution 1960; 14: 458–472.

[7]. Lewontin RC: The interaction of selection and linkage.I. General considerations; heterotic models. Genetics 1964; 49: 49–67. [PubMed: 17248194]

[8]. Hedrick PW: Genetic disequilibrium measures: proceed with caution. Genetics 1987; 117: 331–341. [PubMed: 3666445]

[9]. Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 1995; 29: 311–322. [PubMed: 8666377]

[10]. Sabatti C, Risch N: Homozygosity and linkage disequilibrium. Genetics 2002; 160: 1707–1719. [PubMed: 11973323]

[11]. Mueller JC: Linkage disequilibrium for different scales and applications. Brief Bioinform 2004; 5: 355–364. [PubMed: 15606972]

[12]. Zapata C: On the uses and applications of the most commonly used measures of linkage disequilibrium from the comparative analysis of their statistical properties. Hum Hered 2011; 71: 186–195. [PubMed: 21778738]

[13]. Hill WG, Robertson A: Linkage disequilibrium in finite populations. Theor Appl Genet 1968; 38: 226–231. [PubMed: 24442307]

[14]. Hudson RR: The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics 1985; 109: 611–631. [PubMed: 3979817]

[15]. McVean G: A genealogical interpretation of linkage disequilibrium. Genetics 2002; 162: 987–991. [PubMed: 12399406]

[16]. Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. Am J Hum Genet 2001; 69: 1–14. [PubMed: 11410837]

[17]. Nei M, Li W-H: Non-random association between electromorphs and inversion chromosomes in finite populations. Genet Res 1980; 35: 65–83. [PubMed: 7450499]

[18]. Kaplan N, Weir BS: Expected behavior of conditional linkage disequilibrium. Am J Hum Genet 1992; 51: 333–343. [PubMed: 1353663]

[19]. Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok P-Y, Collins A: The optimal measure of allelic association. Proc Natl Acad Sci USA 2001; 98: 5217–5221. [PubMed: 11309498]

[20]. VanLiere JM, Rosenberg NA: Mathematical properties of the r^2 measure of linkage disequilibrium. Theor Popul Biol 2008; 74: 130–137. [PubMed: 18572214]

[21]. Lewontin RC: On measures of gametic disequilibrium. Genetics 1988; 120: 849–852. [PubMed: 3224810]

[22]. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al.: The structure of haplotype blocks in the human genome. Science 2002; 296: 2225–2229. [PubMed: 12029063]

[23]. Wall JD, Pritchard JK: Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 2003; 4: 587–601. [PubMed: 12897771]

[24]. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004; 74: 106–120. [PubMed: 14681826]

[25]. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. Nat Genet 2005; 37: 1217–1223. [PubMed: 16244653]

[26]. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005; 21: 263–265. [PubMed: 15297300]

[27]. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al.: Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet 2015; 97: 576–592. [PubMed: 26430803]

[28]. Kemppainen P, Knight CG, Sarma DK, Hlaing T, Prakash A, Maung Maung YN, et al.: Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. Mol Ecol Resour 2015; 15: 1031–1045. [PubMed: 25573196]

[29]. Eberle MA, Rieder MJ, Kruglyak L, Nickerson DA: Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. PLoS Genet 2006; 2: e142. [PubMed: 16965180]

[30]. Wray NR: Allele frequencies and the r^2 measure of linkage disequilibrium: impact on design and interpretation of association studies. Twin Res Hum Genet 2005; 8: 87–94. [PubMed: 15901470]

[31]. Auer PL, Lettre G: Rare variant association studies: considerations, challenges and opportunities. Genome Med 2015; 7: 16. [PubMed: 25709717]

[32]. Bomba L, Walter K, Soranzo N: The impact of rare and low-frequency genetic variants in common disease. Genome Biol 2017; 18: 77. [PubMed: 28449691]

[33]. Li B, Liu DJ, Leal SM: Identifying rare variants associated with complex traits via sequencing. Curr Protoc Hum Genet 2013; 78: 1.26.1–1.26.22.

[34]. Turkmen A, Lin S: Are rare variants really independent?. Genet Epidemiol 2016; 41: 363–371.

[35]. Collins A, Morton NE: Mapping a disease locus by allelic association. Proc Natl Acad Sci USA 1998; 95: 1741–1745. [PubMed: 9465087]

[36]. Shete S: A note on the optimal measure of allelic association. Ann Hum Genet 2003; 67: 189–191. [PubMed: 12675694]

[37]. Mangin B, Garnier-Géré P, Cierco-Ayrolles C: The estimator of the optimal measure of allelic association: mean, variance and probability distribution when the sample size tends to infinity. Stat Appl Genet Mol Biol 2008; 7: 20.

[38]. Weir BS: Genetic Data Analysis II. Sunderland, Sinauer, 1996.

[39]. Rosenberg NA, Blum MGB: Sampling properties of homozygosity-based statistics for linkage disequilibrium. Math Biosci 2007; 208: 33–47. [PubMed: 17157882]

[40]. Song YS, Song JS: Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. Theor Popul Biol 2007; 71: 49–60. [PubMed: 17069867]

[41]. Alcala N, Rosenberg NA: Mathematical constraints on F_{ST}: biallelic markers in arbitrarily many populations. Genetics 2017; 206: 1581–1600. [PubMed: 28476869]

[42]. Zapata C: The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution 2000; 54: 1809–1812. [PubMed: 11108607]

[43]. Payseur BA, Place M, Weber JL: Linkage disequilibrium between STRPs and SNPs across the human genome. Am J Hum Genet 2008; 82: 1039–1050. [PubMed: 18423524]

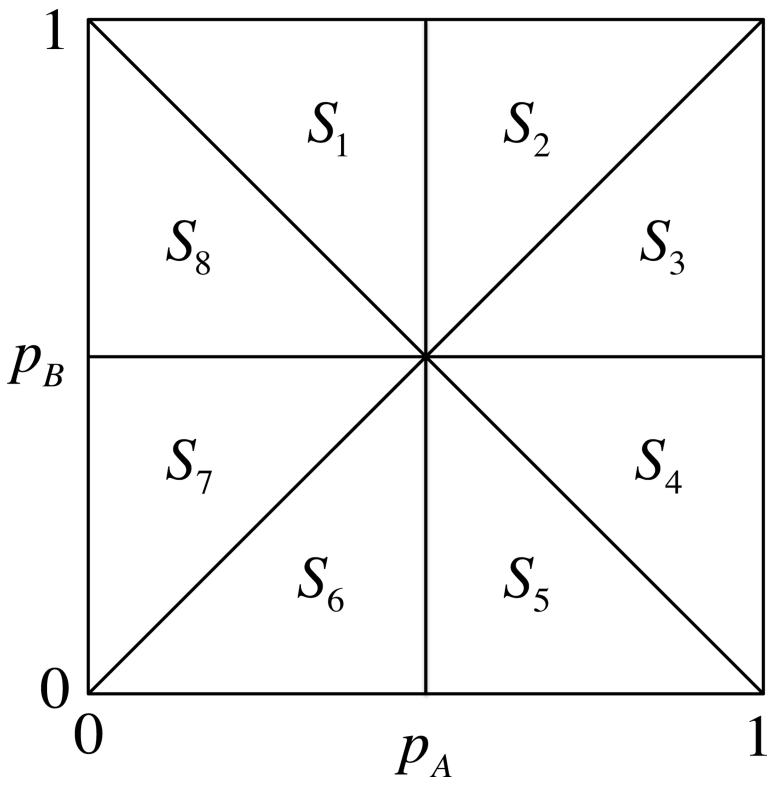

Figure 1:
A unit square showing all possible combinations of the frequencies p_A and p_B. The region is subdivided into eight octants $S_1,...,S_8$.

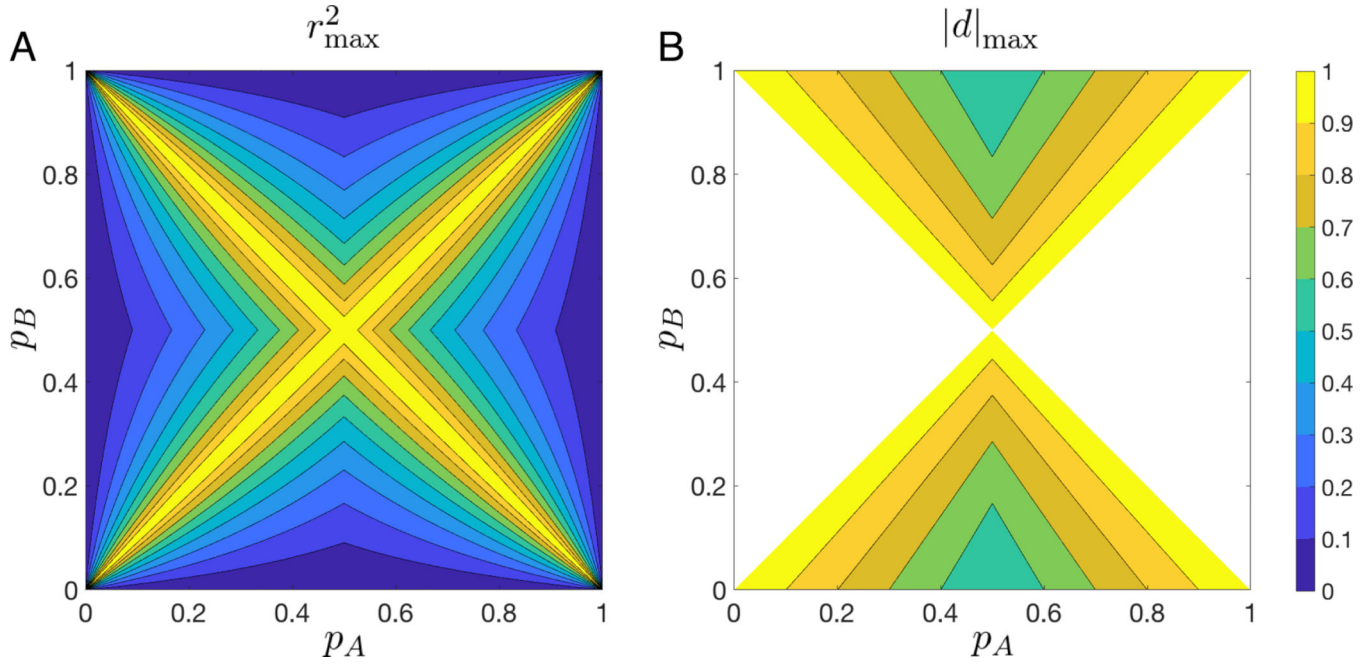

Figure 2:

$r^2_{\max}$ and $|d|_{\max}$ as functions of p_A and p_B. (A) Contour plot of $r^2_{\max}$. (B) Contour plot of $|d|_{\max}$. The plots consider the maximum over all possible values of p_{AB}. The functions plotted appear in TABLE 1. $|d|$ is defined only in octants S_1, S_2, S_5, and S_6.

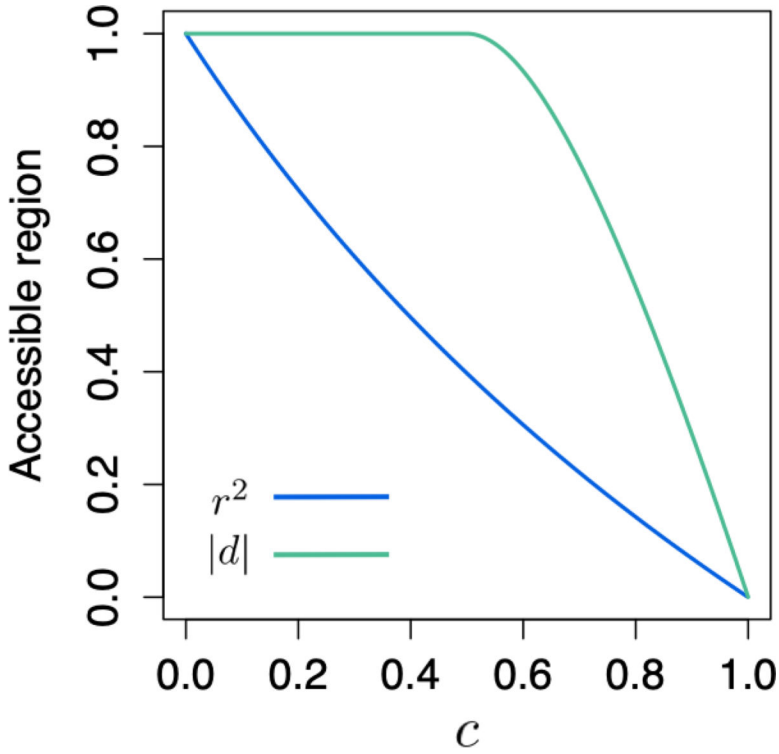

Figure 3:
The portion of the permissible allele frequency space where r^2 and $|d|$ can exceed a specific value of c, as a function of c. $p_{r^2}(c)$ is taken from eq. 6, $p_{|d|}(c)$ is taken from eq. 14, and both functions appear in TABLE 4.

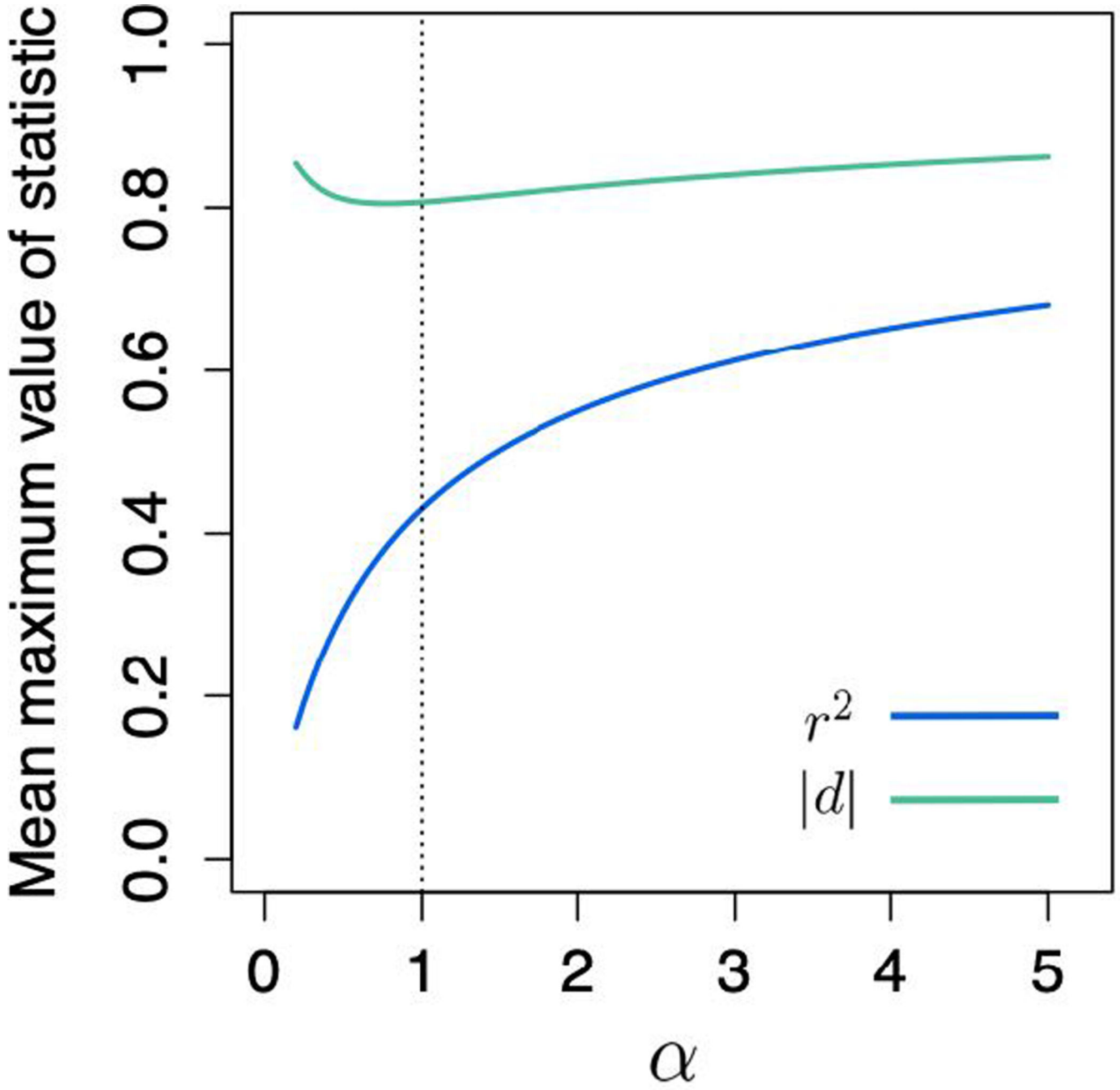

Figure 4:

Mean maximum value of r^2 and $|d|$, if p_A and p_B are drawn from independent Beta-(a, a) distributions. The dotted line indicates a = 1, which gives values that are identical to the case in which p_A and p_B are drawn from independent Uniform-(0, 1) distributions.

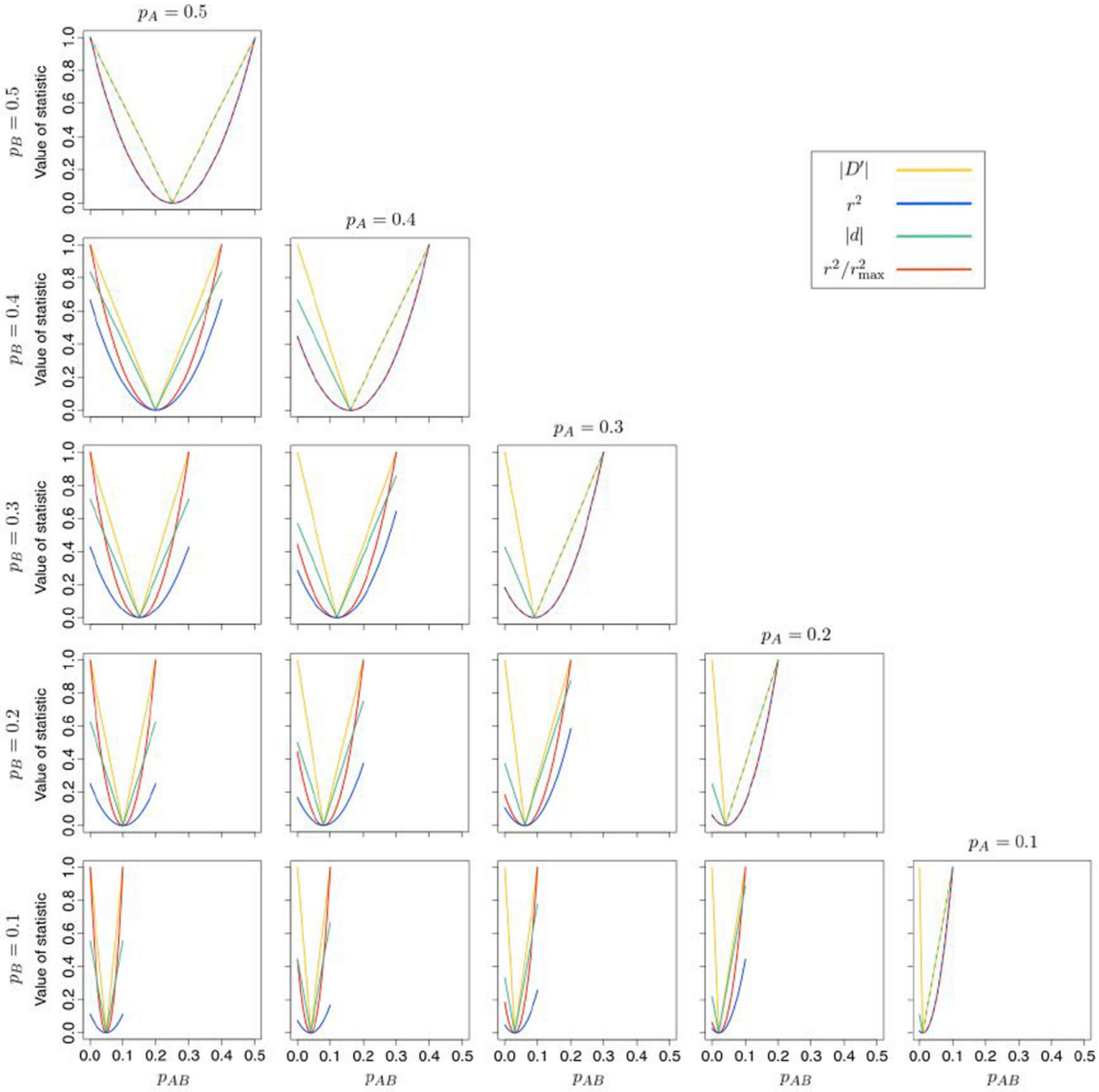

Figure 5:
Values of linkage disequilibrium statistics as functions of the haplotype frequency p_{AB}, for fixed values of p_A and p_B in S_6.

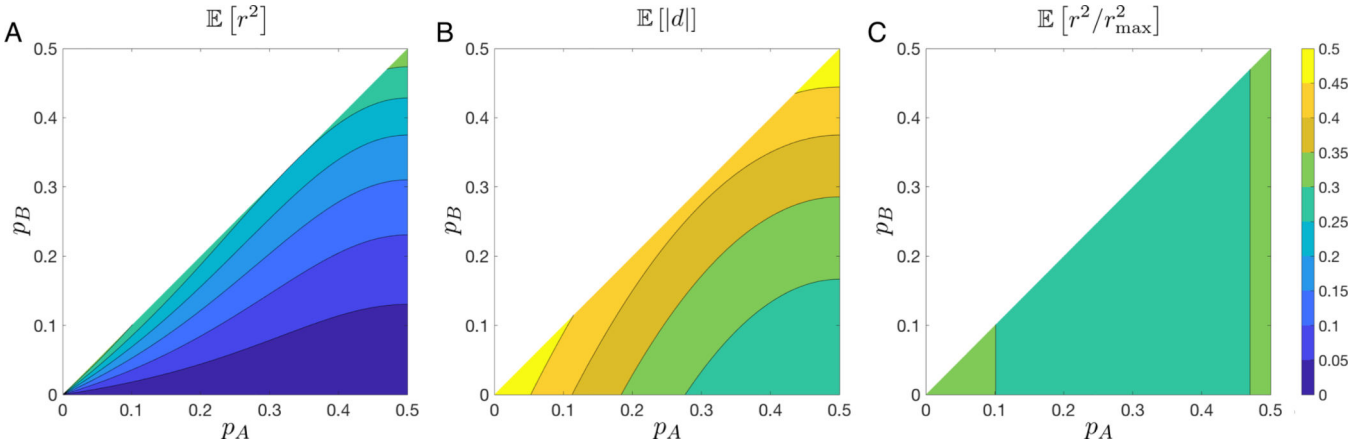

Figure 6:

Mean value of three linkage disequilibrium statistics in S_6 as functions of p_A and p_B, assuming $p_{AB} \sim$ Uniform-$(0, p_B)$. (A)r^2. (B)$|d|$. (C)$r^2/r^2_{\max}$.

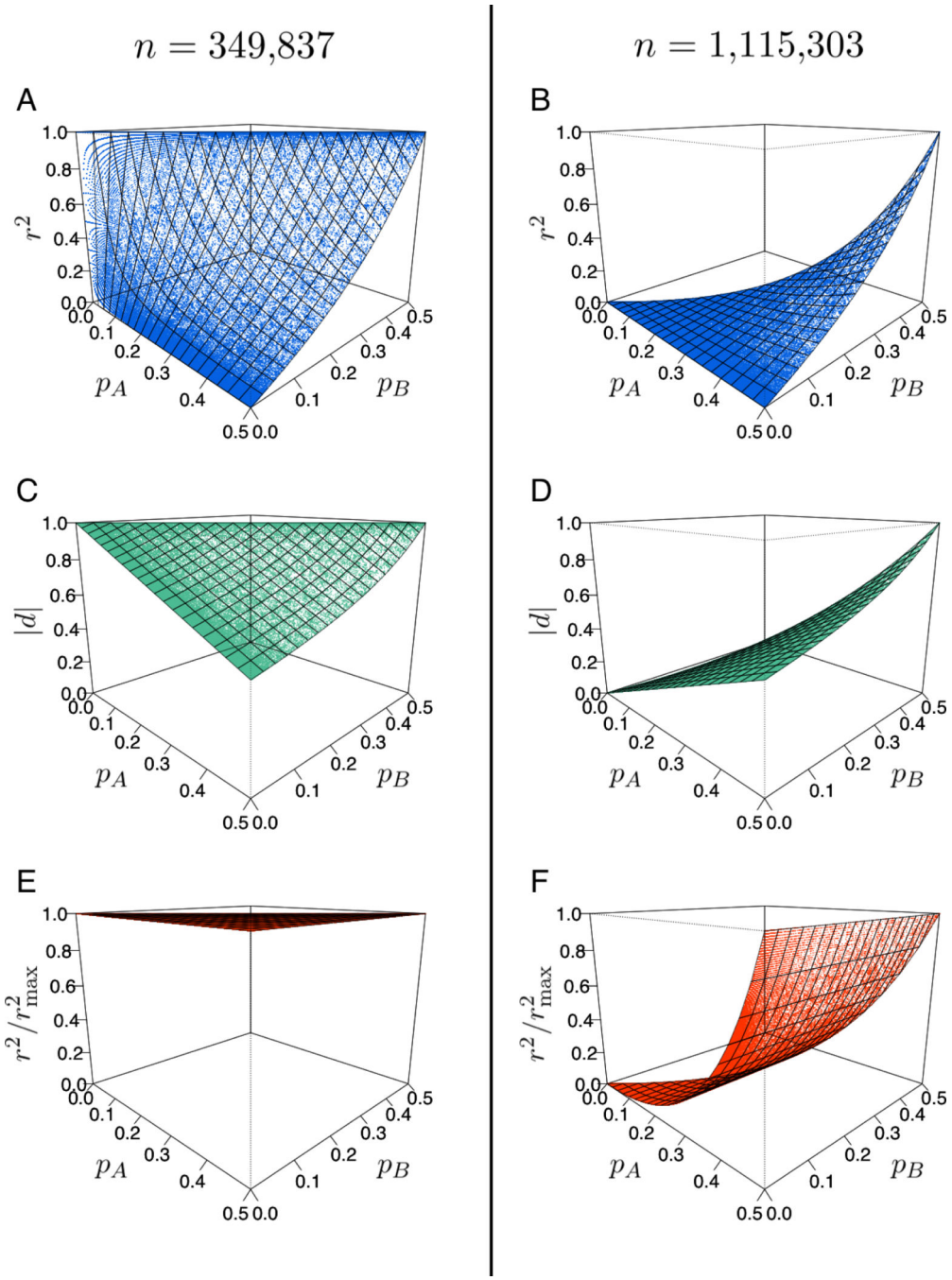

Figure 7:

The distributions of r^2, $|d|$, and $r^2/r^2_{\max}$ values calculated from data in S_6, when $|D'| = 1$. (A) r^2 values if $p_{AB} = p_B$, lying on the surface $r^2 = (1 - p_A)p_B/[p_A(1 - p_B)]$. (B) r^2 values if $p_{AB} = 0$, lying on the surface $r^2 = p_A p_B/[(1 - p_A)(1 - p_B)]$. (C) $|d|$ values if $p_{AB} = p_B$, lying on the surface $|d| = (1 - p_A)/(1 - p_B)$. (D) $|d|$ values if $p_{AB} = 0$, lying on the surface $|d| = (1 - p_A)/(1 - p_B)$. (E) $r^2/r^2_{\max}$ values if $p_{AB} = p_B$, lying on the surface $r^2/r^2_{\max} = 1$. (F) $r^2/r^2_{\max}$ values if $p_{AB} = 0$, lying on the surface $r^2/r^2_{\max} = [p_A/(1 - p_A)]^2$.

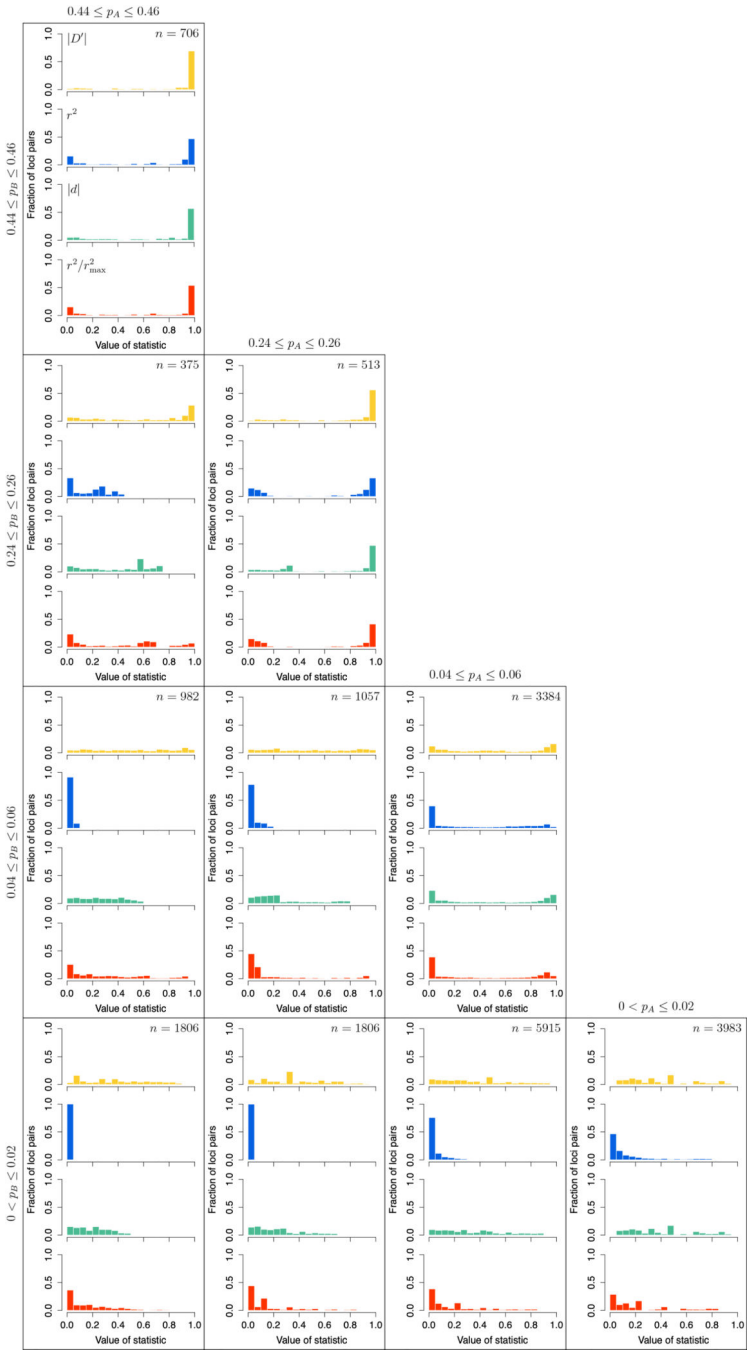

Figure 8:

The distributions of values of linkage disequilibrium statistics calculated from data in S_6, given specific ranges of values for p_A and p_B. For each of four windows for p_A and four windows for p_B, we divide points into bins based on their values of each of four statistics $(|D'|, r^2, |d|, r^2/r^2_{\max})$. The number of locus pairs falling into a pair of bins appears in the top right corner of the group of four histograms associated with the bin pair.

Table 1:

The eight octants in the space of possible allele frequencies, along with their associated $r^2_{\max}$, $|d|_{\max}$, and $\rho_{\max}$ values.

Octant	Condition				p_{AB} achieving maximal $\lvert D\rvert$	$r^2_{\max}$	$\lvert d\rvert_{\max}$	$\rho_{\max}$
	$p_A < \frac{1}{2}$	$p_B < \frac{1}{2}$	$p_A < p_B$	$p_A + p_B < 1$				
S_1	Yes	No	Yes	No	$p_A + p_B - 1$	$\frac{(1-p_A)(1-p_B)}{p_A p_B}$	$\frac{1-p_A}{p_B}$	-
S_2	No	No	Yes	No	p_A	$\frac{p_A(1-p_B)}{(1-p_A)p_B}$	$\frac{p_A}{p_B}$	-
S_3	No	No	No	No	p_B	$\frac{(1-p_A)p_B}{p_A(1-p_B)}$	-	-
S_4	No	Yes	No	No	$p_A + p_B - 1$	$\frac{(1-p_A)(1-p_B)}{p_A p_B}$	-	1
S_5	No	Yes	No	Yes	0	$\frac{p_A p_B}{(1-p_A)(1-p_B)}$	$\frac{p_A}{1-p_B}$	1
S_6	Yes	Yes	No	Yes	p_B	$\frac{(1-p_A)p_B}{p_A(1-p_B)}$	$\frac{1-p_A}{1-p_B}$	1
S_7	Yes	Yes	Yes	Yes	p_A	$\frac{p_A(1-p_B)}{(1-p_A)p_B}$	-	-
S_8	Yes	Yes	Yes	Yes	0	$\frac{p_A p_B}{(1-p_A)(1-p_B)}$	-	-

Table 2:

Notation for the allele and haplotype frequencies for a pair of biallelic loci.

		Locus 1		
		A	a	Total
Locus 2	B	p_{AB}	p_{aB}	p_B
	b	p_{Ab}	p_{ab}	$1-p_B$
	Total	p_A	$1-p_A$	1

Table 3:

Octants of the allele frequency space in which the different linkage disequilibrium measures can be applied.

Octant	D'	r^2	d	ρ	$r^2/r^2_{\max}$
S_1	Yes	Yes	Yes	No	Yes
S_2	Yes	Yes	Yes	No	Yes
S_3	Yes	Yes	No	No	Yes
S_4	Yes	Yes	No	Yes	Yes
S_5	Yes	Yes	Yes	Yes	Yes
S_6	Yes	Yes	Yes	Yes	Yes
S_7	Yes	Yes	No	No	Yes
S_8	Yes	Yes	No	No	Yes

Table 4:

Mean maximum values and accessible regions for the five measures. The mean maximum value of a measure is its average maximum value over its prescribed domain, assuming p_A and p_B are independent and uniformly distributed over the domain. The accessible region of a measure for a constant $c \in [0, 1]$ is defined as the proportion of the applicable domain in which the upper bound for the measure is greater than or equal to c.

	Mean maximum value	Accessible region
$\lvert D' \rvert$	1	1
r^2	$2\pi^2/3 - 4(\ln 2)^2 + 4 \ln 2 - 7 \approx 0.43051$	$1 + \frac{4c}{1-c} + \frac{8c \ln(\frac{1}{2} + \frac{1}{2}c)}{(1-c)^2}$
$\lvert d \rvert$	$\frac{3}{2} - \ln 2 \approx 0.80685$	1, if $c \leqslant 0.5$; $\frac{(4c-1)(1-c)}{c}$, if $c > 0.5$
ρ	1	1
$r^2/r^2_{\max}$	1	1

Table 5:

Constraints on one allele frequency given an LD value and allele frequency at the other locus, for the five measures. Here, we assume $p_A, p_B \leqslant \frac{1}{2}$. Owing to symmetry in the loci, for $|D'|, r^2,$ and $r^2/r^2_{\max}$, fixing p_A is equivalent to fixing p_B.

	Fixed allele frequency	
	p_A	p_B
$\lvert D'\rvert$	$0 \leqslant p_B \leqslant 1$	$0 \leqslant p_A \leqslant 1$
r^2	$\frac{r^2 p_A}{1 + r^2 p_A - p_A} \leqslant p_B \leqslant \min\left(\frac{1}{2}, \frac{p_A}{r^2 - r^2 p_A + p_A}\right)$	$\frac{r^2 p_B}{1 + r^2 p_B - p_B} \leqslant p_A \leqslant \min\left(\frac{1}{2}, \frac{p_B}{r^2 - r^2 p_B + p_B}\right)$
$\lvert d\rvert$	$\max\left(0, \frac{p_A + \lvert d\rvert - 1}{\lvert d\rvert}\right) \leqslant p_B \leqslant p_A$	$p_B \leqslant p_A \leqslant \min\left(\frac{1}{2}, 1 - \lvert d\rvert + \lvert d\rvert p_B\right)$
ρ	$0 \leqslant p_B \leqslant 1$	$0 \leqslant p_A \leqslant 1$
$r^2/r^2_{\max}$	$0 \leqslant p_B \leqslant 1$	$0 \leqslant p_A \leqslant 1$

Table 6:

Description, usages, and mathematical properties of five LD measures for biallelic loci.

Statistic	Description	Noted usages in the literature	Mathematical properties
$\|D'\|$	Normalization of $\|D\|$ by its theoretical maximum value for a given set of allele frequencies	Detecting "complete" LD (where one of the four haplotypes is absent), an indication of whether recombination has occurred between the two loci [11]	$\|D'\|$ varies linearly as a function of p_{AB} (FIGURE 5) Upper bound of 1 for all allele frequencies Assuming a uniform distribution of p_{AB} over the range of values it can take, the mean and variance of $\|D'\|$ are both constant values (eqs. 19 and 33)
r^2	Squared correlation coefficient measure between allelic indicator variables	Testing for independence between a pair of loci by a χ^2 test [16] Association studies, where a mathematical relationship exists between r^2 and the sample size needed to detect association between a marker and disease phenotype [16]	r^2 varies quadratically as a function of p_{AB} (FIGURE 5) Low upper bound and small range of values if MAF is low (FIGURE 2A) Mean maximum value varies considerably as a function of the allele frequency distribution (FIGURE 4)
$\|d\|$	Difference in the proportions of disease and normal alleles found on the same haplotype with a particular marker allele	Association mapping for rare diseases in which case-control sampling is employed [9]	$\|d\|$ varies linearly as a function of p_{AB} (FIGURE 5) Upper bound has an intermediate value; measure still has a considerable range even at low MAF (FIGURE 2B) Mean maximum value relatively stable as a function of the allele frequency distribution (FIGURE 4)
ρ	Probability that a haplotype chosen at random descends without recombination from a population of haplotypes that excludes one of the four possible haplotypes	Mapping of marker association and localization of disease loci [19]	Identical to $\|D'\|$ in the octants in which it can be applied
$r^2/r^2_{\max}$	Normalization of r^2 by its theoretical maximum value for a given set of allele frequencies	If a range that is independent of allele frequency is desired, but the measure still maintains some connection to r^2 [20]	$r^2/r^2_{\max}$ varies quadratically as a function of p_{AB} (FIGURE 5) Upper bound of 1 for all allele frequencies Assuming a uniform distribution of p_{AB} over the range of values it can take, the mean and variance of $r^2/r^2_{\max}$ both depend only on the locus with the larger MAF (eqs. 22 and 39)