# The nonlinear structure of linkage disequilibrium

Reginald D. Smith

*Supreme Vinegar LLC, 3430 Progress Dr. Suite D, Bensalem, PA 19020*

**Abstract**

The allele frequency dependence of the ranges of all measures of linkage disequilibrium is well-known. The maximum values of commonly used parameters such $r^2$ and $D$ are absolutely dependent on the minor allele frequencies at each loci. However, though this phenomenon is recognized and accounted for in many studies, the comprehensive mathematical framework underlying the limits of linkage disequilibrium measures at various frequency combinations is often heuristic or empirical. Here, it is demonstrated that underlying this behavior is the fundamental shift between linear and nonlinear dependence in the linkage disequilibrium structure between loci. The proportion of linear and nonlinear dependence can be estimated and it demonstrates how even the same values of $r^2$ can have different implications for the nature of the overall dependence.

*Keywords:* linkage disequilibrium, allele frequency, nonlinearity

## 1. Definitions of linkage disequilibrium

Soon after the rediscovery of Mendel's work and the dawn of modern genetics, one of his laws, his Second Law of the independent assortment of genes, was found to be at variance with experimental results in several organisms (Bateson Punnett & Saunders (1905, 1906); Morgan (1911)). The concept of linkage, where certain pairs of genes on chromosomes segregated together during recombination much more often than other pairs, became a fundamental concept in theoretical and applied genetics. The mathematical theory of linkage between two loci was first expounded soon afterwards (Weinberg (1908); Jennings (1917); Robbins (1918)). The more general case

*Email address:* rsmith@supremevinegar.com (Reginald D. Smith)

for multiple loci was derived in detail by Geiringer (1944). The modern formation of the problem, and the definition of the term linkage disequilibrium, was made by Lewontin & Kojima (1960) where $D$, one of the most common measures of linkage disequilibrium, is first defined. Many other measures of linkage disequilibrium would be defined such as $D'$ (Lewontin (1964)), $r^2$, the square of the correlation coefficient (Hill & Robertson (1968)), and $d$ (Nei & Li (1980)).

Across two bi-allelic loci which have possible alleles $A/a$ and $B/b$, the two most well-known measures of linkage equilibrium, are $D$ and $r^2$.

$$D = P_{AB} - P_A P_B = P_{AB} P_{ab} - P_{aB} P_{Ab} \tag{1}$$

$$r^2 = \frac{D^2}{P_A(1 - P_A) P_B (1 - P_B)} \tag{2}$$

Given the allele frequencies at each locus are Bernoulli distributed, $D$ is equal to the covariance and $P_A(1 - P_A)$ and $P_B(1 - P_B)$ are the variances at the respective loci. The range of $D$ varies based on loci minor allele frequencies with a maximum of $[-0.25, 0.25]$ where $P_A = P_B = 0.5$. Given its difficult interpretation, the use of $D$ is often modified and represented by the normalized indicator, $D'$ (Lewontin (1964)) which normalizes the value of $D$ to a range of $[0, 1]$ based on its maximum given the underlying allele frequencies at each locus. For $D > 0$

$$D' = \frac{D}{\min\left[P_A(1 - P_B), P_B(1 - P_A)\right]} \tag{3}$$

and for $D < 0$

$$D' = \frac{D}{\max\left[-P_A P_B, -(1 - P_B)(1 - P_A)\right]} \tag{4}$$

The measures of $r^2$ and $D'$ are often used separately or jointly due to the fact they have a common, intuitive, and restricted range of values (Hedrick (1987); Lewontin (1988)). However, there can still be issues of interpretation with both variables. First, though the full range of values can be $[0, 1]$ for each, only $D'$ can take any value in the range for all values of minor allele frequencies. The range of $r^2$, similar to $D$, is restricted based on the loci minor allele frequencies. Second, sometimes the two measures can have wildly

different, seemingly contradictory, values for the same measurement of linkage disequilibrium. For example, in some cases $r^2$ can take small values near zero while $D' = 1$.

Regarding the first exception, an in-depth analysis on the maximum values of $r^2$ across different pairs of minor allele frequencies was given by VanLiere & Rosenberg (2008). In the paper, equations 2-5 illustrate the derived maximum values of $r^2$, dubbed $r^2_{\text{max}}$, for various minor allele frequency combinations across loci. Depending on the relative values of the major and minor allele frequencies between loci, $r^2_{\text{max}}$ would take the value of either $P_A(1-P_B)/P_B(1-P_A)$, $P_B(1-P_A)/P_A(1-P_B)$, $P_AP_B/(1-P_A)(1-P_B)$, or $(1-P_A)(1-P_B)/P_AP_B$. One of the key findings of the paper was that $r^2_{\text{max}}$ can only be equal to 1, allowing $r^2$ to take any value, under the two conditions of either $P_A = P_B$ or $P_A = 1 - P_B$. The first condition, $P_A = P_B$, corresponds to positive values of $r$, while $P_A = 1-P_B$ corresponds to negative values of $r$. The description of $r^2_{\text{max}}$ was important for several reasons. For example, it helped give theoretical validation to the long espoused practice of trying to match minor allele frequencies in association studies when inferring a trait locus from a marker locus it is in linkage disequilibrium with. It was known empirically that the limited values of $r^2$ when minor allele frequencies are mismatched can cause problems with inference and increase the necessary population sizes in association studies (Nei & Li (1980); Kaplan & Weir (1992); Kruglyak (1999); Wray (2005)).

However, the nature of $r^2_{\text{max}}$ also raises questions about the nature of linkage disequilibrium. Sometimes $r^2_{\text{max}}$ can be extremely limited and much less than one. Does this imply that $r^2_{\text{max}}$ is the upper bound of any dependence between the two loci or is there additional dependence not encapsulated by $r^2$? Particularly, should $r^2_{\text{max}}$ values correspond with complete linkage disequilibrium as defined by $D' = 1$ or is the relationship more complicated? In VanLiere & Rosenberg (2008), there is a clue about a relationship where they demonstrate a proportion of minor allele frequency combinations where $D' = r^2/r^2_{\text{max}}$ showing that there can often be such a relationship but it is not consistent over all allele frequency values. Where this relation holds and $r^2 = r^2_{\text{max}}$ we have complete linkage disequilibrium. However, for those allele frequency combinations where this relationship does not hold, does $D'$ provide any additional information on dependence beyond that of $r^2$?

One of the key reasons for the limitations of $r^2$ and the divergences between $r^2$ and $D'$ is that $r^2$, like all correlation coefficient derived measures, is strictly a measure of the linear dependence between alleles at different loci.

On the other hand, $D$ and $D'$ both account for both linear and nonlinear dependence. The issues with $r^2$ across different allele frequencies is in fact, not just an issue of the definition of $r^2$ but rather a larger one of nonlinear dependence between loci. This may seem like a subtle or even unimportant detail, however, in this paper it will be demonstrated that the effects of this dichotomy are far from trivial and can demonstrate the error of relying on $r^2$ alone in certain situations.

## 2. Defining linear linkage disequilibrium

While all values of linkage disequilibrium have a linear component as evidenced by a nonzero value of $r^2$, how linear the overall relationship is between the two loci should be quantified in order to understand the importance of $r^2$ in the context of total dependence. As a first step, we will analyze under what conditions the linkage disequilibrium is completely linear and thus has all dependence described by $r^2$ alone. While the joint probabilities of two alleles at each locus is symmetric, $P_{AB} = P_{BA}$ the conditional probabilities are distinct and defined as $P_{B|A} = P_{AB}/P_A$ and $P_{A|B} = P_{AB}/P_B$. Given the conditional probabilities describe the dependence of one allele on the other, we can use them to define the conditions of completely linear dependence.

The linkage disequilibrium between two loci is fully linear when the dual conditional probabilities of any two alleles can be described by a system of linear equations

$$\begin{aligned} P_{B|A} &= c_1 P_A + c_2 P_B + c_3 \\ P_{A|B} &= c_4 P_A + c_5 P_B + c_6 \end{aligned} \tag{5}$$

In equation 5 $c_1$ to $c_6$ are constants. From the definitions of conditional probabilities and equations 1 and 2 we can derive the two equations

$$\begin{aligned} P_{B|A} &= r\sqrt{\frac{P_B(1-P_A)(1-P_B)}{P_A}} + P_B \\ P_{A|B} &= r\sqrt{\frac{P_A(1-P_A)(1-P_B)}{P_B}} + P_A \end{aligned} \tag{6}$$

This system of equations is not linear and would seem to imply that there is no case of complete linear dependence, however, there are two conditions

where equation 6 reduces to a system of linear equations, in particular the cases of $P_A = P_B$ or $P_A = 1 - P_B$, the identical conditions where $r^2_{\text{max}} = 1$. Under both conditions the equations reduce to

$$\begin{aligned} P_{B|A} &= P_B + r(1 - P_A) \\ P_{A|B} &= P_A + r(1 - P_B) \end{aligned} \tag{7}$$

Therefore, only under the conditions of $r^2_{\text{max}} = 1$ can the linkage disequilibrium be described as completely linear and $r^2$ captures all dependence. The primacy of $r^2$ under these conditions of perfectly linear linkage disequilibrium is emphasized by the fact that for these matchings of allele frequencies $D'^2 = r^2$. Further, by Bayes' Theorem, $P_{B|A}P_A = P_{A|B}P_B$, we can show that under perfectly linear linkage disequilibrium

$$\begin{aligned} P_{B|A} &= P_{A|B} \text{ for } r > 0 \\ P_{b|A} &= P_{A|b} \text{ for } r < 0 \end{aligned} \tag{8}$$

Since the conditions of perfectly linear linkage disequilibrium are extremely narrow, it is now a question of defining the relative strengths of linear and nonlinear dependence in linkage disequilibrium for all other possible combinations of allele frequencies across loci.

## 3. Defining nonlinear linkage disequilibrium

Under the conditions of perfectly linear linkage disequilibrium, $D'^2 = r^2$. However, under all other conditions, $D'$ can have a range of values for any fixed value of linear linkage disequilibrium as measured by $r^2$. To understand the ranges of $D'$ for a given linear contribution from $r^2$, we can define the minimum and maximum values of $D'$. For any $r$, the maximum value of $D'$ is always one where linkage disequilibrium measured by $D$ is at its maximum value. However, the minimum of $D'$ can vary. For fixed $r > 0$,

$$\begin{aligned} D' &= r\sqrt{\frac{P_B(1 - P_A)}{P_A(1 - P_B)}} \text{ where } P_A(1 - P_B) \text{ is smaller} \\ D' &= r\sqrt{\frac{P_A(1 - P_B)}{P_B(1 - P_A)}} \text{ where } P_B(1 - P_A) \text{ is smaller} \end{aligned} \tag{9}$$

or in the case $r < 0$

$$
\begin{aligned}
D' &= r\sqrt{\frac{(1-P_B)(1-P_A)}{P_A P_B}} \text{where } P_A P_B \text{ is smaller} \\
D' &= r\sqrt{\frac{P_A P_B}{(1-P_B)(1-P_A)}} \text{where } (1-P_A)(1-P_B) \text{ is smaller}
\end{aligned}
\tag{10}
$$

Both possibilities for positive or negative $r$ depend on the maximum values of $D$ per equations 3 and 4. However, since the values in the denominator are smaller than the numerator for each case, the terms in the square roots have a minimum value of one. The maximum of $D'$ is 1 which means that at complete linkage disequilibrium those square root terms equal $1/r$ and are at a maximum. Therefore, a value of 1 for the square root terms designates the minimum value of $D'$ giving $D' = |r|$. Since it is already established previously that $D'^2 = r^2$ designates perfectly linear linkage disequilibrium the following two points can be determined.

First, for a given value of $r$, the range of $D'$ is limited to $[|r|,1]$ where the condition of $D' = |r|$ or $D'^2 = r^2$ indicates perfectly linear dependence between the alleles across both loci where $P_{B|A} = P_{A|B}$ or $P_{b|A} = P_{A|b}$ for positive or negative values of $D$ and $r$ respectively. Second, for a given value of $r^2$, any $D'^2 > r^2$ indicates the presence of nonlinear dependence in addition to the linear dependence between alleles across both loci.

Therefore, $D'^2$ more accurately measures full dependence, linear and nonlinear. However, the linear and nonlinear proportions of dependence cannot be simply separated just relying on the value of $D'^2$. In order to measure the linear and nonlinear components of $D'^2$ we will rely on the most common and comprehensive measure of total dependence, mutual information. First derived by Claude Shannon when creating information theory (Shannon (1948)) mutual information is a measure of total dependence between two variables and how much reduction in uncertainty we can expect to know about one variable given other variable. For two bi-allelic loci the mutual information between them is calculated as

$$
I_{ij} = \sum_{i \in A,a} \sum_{j \in B,b} P_{ij} \log \frac{P_{ij}}{P_i P_j} \tag{11}
$$

This equation sums the components over all possible allele pairs between the loci. In particular, we can use the mutual information to calculate a

ratio of the linear dependence to total dependence, hereafter termed $\Lambda$ that will allow us to measure what proportion of total dependence in $D'$ is linear under any given scenario. The ratio represented by $\Lambda$ was first worked out in terms of determining the relative linear dependence in data via the residuals of linear regression (Smith (2015)). In that procedure the ratio of the mutual information of transformed residuals and the independent variable and the mutual information of the original function and independent variable is used to determine the approximate proportion of linear dependence in the original function or data.

For linkage disequilibrium, however, a real valued linear regression is not possible but the mutual information of the base linear case for $D'^2$ can be calculated by calculating the mutual information for $D'^2 = r^2$. This can then be compared with the total mutual information for any other given value of $D'^2 > r^2$ for the same $r^2$.

$$\Lambda = \frac{I(i \in A, a, j \in B, b; D'^2 = r^2)}{I(i \in A, a, j \in B, b; D'^2)} \tag{12}$$

The calculation of both the values of mutual information was done through numerical simulation where $D'^2$ and $r^2$ were simulated across all possible values for each for every compatible pair of minor allele frequencies at each locus.

Based on this procedure we can create heat plots of $\Lambda$ for different values of $P_A$ and $P_B$. In fact, what we find is regardless the value of $D'^2$, the heat plots are the same with two different plots: one for positive values of $r$ and one for negative values of $r$. These are shown in Figure 1.

As shown in the plots, for positive $r$ the maximum values of $\Lambda$, indicating strict linear dependence, only occur in the case of $P_A = P_B$. Outside this region, the linear dependence as a proportion of all dependence rapidly falls off and becomes lower as the ratio of the allele frequencies at each locus become increasingly larger or smaller. Where $r$ is negative, we find that the strict linear dependence range changes to where $P_A$ and $(1 - P_B)$ are equal. However, like the other situation, the proportion of linear dependence rapidly drops off outside this region.

These plots closely reflect the regions of $r^2_{\text{max}}$ described by VanLiere & Rosenberg (2008) who plotted the maximum possible value of $r^2$ by allele frequency pairs for positive or negative correlation. Therefore, we find in a completely different perspective that the proportion of linear dependence to

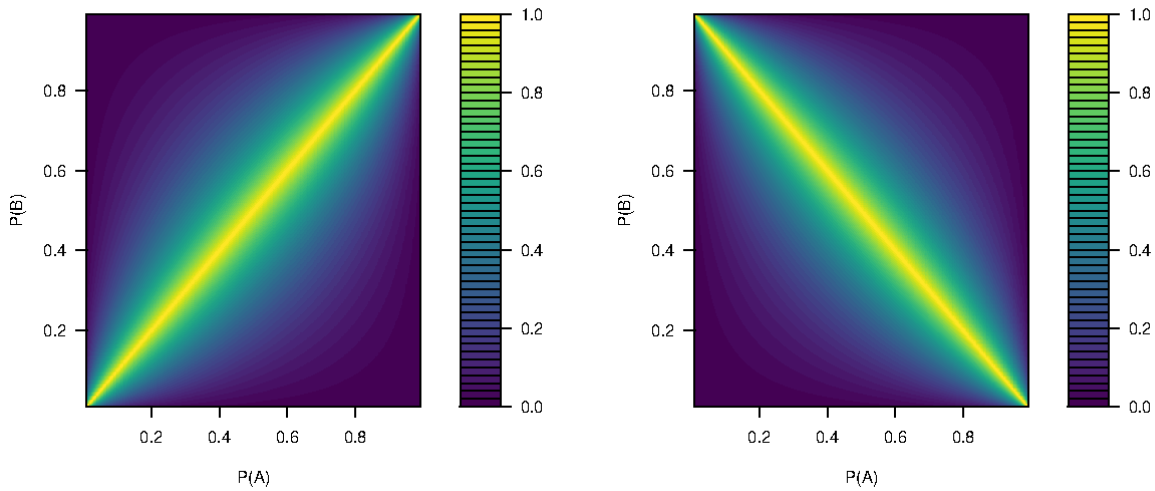


Figure 1: Heat plots of the value of $\Lambda$ for positive and negative $r$ respectively. Yellow indicates a $\Lambda \approx 1$ while violet indicates lower values.

total dependence closely matches the values of $r^2_{\text{max}}$ for all allele frequency pairs. Note that $\Lambda$ keeps the same value for allele frequencies regardless of the actual value of $r^2$, even if this value is less than $r^2_{\text{max}}$.

## 4. Estimating $\Lambda$ from allele frequencies

Given the consistent relationships shown in Figure 1 and the fact that $D'^2 = r^2$ for perfect linearity, it is likely $\Lambda$ is somehow related to $D'^2$ and $r^2$. What is found is that there is a nearly linear fit across all allele frequencies between $\Lambda$ and $r^2/D'^2$. This is shown in Figure 2.

Given the definitions of $r^2$ and $D'^2$, we can show that

$$\Lambda \approx \frac{r^2}{D'^2} \approx \min\left[\frac{P_A(1-P_B)}{P_B(1-P_A)}, \frac{P_B(1-P_A)}{P_A(1-P_B)}\right] \tag{13}$$

These again match the equations for $r^2_{\text{max}}$ in VanLiere & Rosenberg (2008) across most of the values of the allele frequencies. In addition, while in VanLiere & Rosenberg (2008) it was shown that $D'^2 = r^2/r^2_{\text{max}}$ over part of the allele pair frequency pair space, this analysis shows that the value of $D'^2 = r^2/\Lambda$ fits across the entire space. Thus, an important finding is

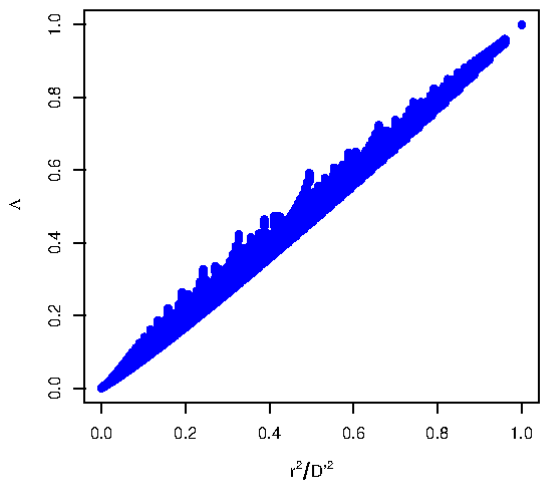


Figure 2: Scatterplot of $\Lambda$ against $r^2/D'^2$.

that the relative nature of the linear and nonlinear dependence between two associated loci is dictated only by the frequencies of the minor alleles at each locus and not the strength of their linkage disequilibrium. For given allele frequencies, the relative linear and nonlinear proportions of that dependence are identical across all values of linkage disequilibrium. Equation 13 fits both positive or negative values of $r$.

Having an approximate equation for $\Lambda$ it is now possible to generalize the relationships between the conditional probabilities $P_{A|B}$ and $P_{B|A}$ for an arbitrary value of $\Lambda$. Since $P_{A|B} = P_{B|A}$ only in perfect linear dependence ($\Lambda = 1$), it follows that the introduction of nonlinear dependence fundamentally skews this so that one allele is more dependent on the other for co-occurrence than the reverse. How nonlinear dependence coherently fits into the picture can be shown with $\Lambda$. Given equation 13 and Bayes' Theorem

$$P_{B|A} \approx \Lambda P_{A|B} \frac{1 - P_B}{1 - P_A} \tag{14}$$

Multiplying out we can then include the joint probabilities as well

$$P_{B|A} - P_{AB} \approx \Lambda (P_{A|B} - P_{AB}) \tag{15}$$

This results in a final consolidated expression decomposing the conditional probability into its linear and nonlinear dependencies

$$P_{B|A} \approx \Lambda P_{A|B} + (1-\Lambda)P_{AB} \tag{16}$$

The second term indicates the nonlinear contribution to dependence which gradually reduces the symmetry between co-occurrences at different loci as $\Lambda$ decreases. At $\Lambda = 0$, the dependence is absolute: the conditional probability of finding $B$ given $A$ is equal to their joint probability implying allele $A$ has become fixed. Equation 16 can also be stated in terms of product instead of a sum

$$P_{B|A} \approx P_{A|B}\big[1-(1-\Lambda)(1-P_B)\big] \tag{17}$$

Equations 16 and 17 help illustrate how increasingly nonlinear dependence between alleles at different loci makes using $r^2$ in calculations problematic where it is explicitly assumed the allele co-occurrences exhibit a symmetric dependence from correlation alone. To illustrate the nature of dependence between alleles at different loci, the diagram in Figure 3 shows the regions of mostly linear, mostly nonlinear, and perfectly linear linkage disequilibrium. Of note is that in the allowed regions of linkage disequilibrium, both the mostly linear and mostly nonlinear regions cover the same area.

The conditional variance between loci can be calculated similar to the marginal variances as $\sigma^2_{B|A} = P_{B|A}(1-P_{B|A})$. This allows one to calculate variance from either the traditional definition of $P_{B|A}$ calculated from $P_{AB} = D + P_A P_B$ as

$$P_{B|A} = r\frac{\sigma_B}{\sigma_A}(1-P_A) + P_B \tag{18}$$

or the definition using $\Lambda$ in equation 16. These yield conditional variances of

$$\sigma^2_{B|A} = \sigma^2_B - r\frac{\sigma_B}{\sigma_A}(1-P_A)\Big[P_{B|A} + P_B - 1\Big] \tag{19}$$

$$\sigma^2_{B|A} = \Lambda P_{A|B}\Big[1-\Lambda P_{A|B}\Big] + (1-\Lambda)P_{AB}\Big[1-(1-\Lambda)P_{AB}\Big] - 2\Lambda(1-\Lambda)P_{A|B}P_{AB} \tag{20}$$

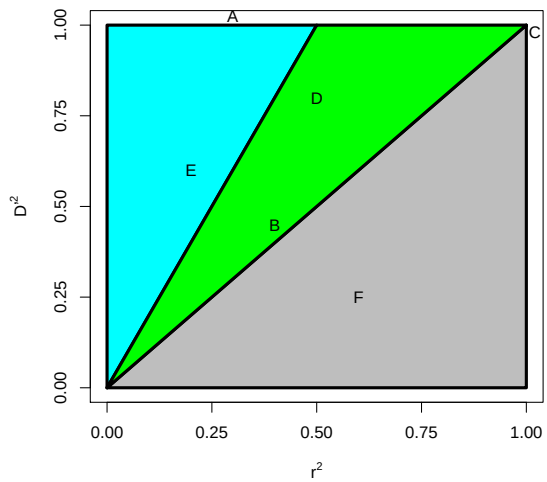


Figure 3: A plot of $D'^2$ vs. $r^2$ demonstrating the different types of linkage disequilibrium for various values of each measure. Labeled regions indicate (A) complete linkage disequilibrium $D' = 1$, (B) perfectly linear linkage disequilibrium $D'^2 = r^2$, (C) perfect linkage disequilibrium $D'^2 = r^2 = 1$, (D) mostly linear linkage disequilibrium (green) $\Lambda \geq 0.5$, (E) mostly nonlinear linkage disequilibrium (light blue) $\Lambda < 0.5$, and (F) disallowed combinations of each measure (gray).

Equation 19 is the more intuitive and easier to calculate. In addition, we can change this equation to one where the relationship is linear and $P_A = P_B = P$ to obtain

$$\sigma^2_{B|A} = \sigma^2(1 - r^2) - r(1 - p)(1 - r)(2p - 1) \tag{21}$$

In equation 21 we see the familiar relationship of the conditional variance and $r^2$. While the equation does not exactly equal $\sigma^2(1 - r^2)$ except when $p = 0.5$, the product in the second term is typically very small for most correlations across many allele frequencies so the traditional $r^2$ estimation is still widely valid when the minor allele frequencies at both loci are equal and of moderate frequency. The exceptions will be described in the next section.

## 5. Insights of nonlinear linkage disequilibrium

When there is any nonlinearity in the dependence between two loci, the frequencies and thus variances differ. Therefore, the conditional probabil-

ities of the alleles are not symmetric and one allele will give much more information about the presence of the other than vice versa. In extreme cases of weakly linear relationships, this causes a situation where one lower frequency allele at one locus almost always co-occurs with another higher frequency allele at a different locus. This has long been recognized for de novo mutations or rare alleles where $r^2 \approx 0$ and $D' = 1$ but it can still be applicable at even moderate values of $r^2$.

In Tables 1, 2, and 3 this is shown for a variety of situations where $p_B$ and $r^2$ are fixed but $\Lambda$ is allowed to vary. It is clear that for an increasingly nonlinear linkage disequilibrium, one allele becomes a less reliable predictor of the other despite all the situations having the same correlation. An additional complication is added when trying to predict how the variance at one locus can be predicted based on the variance at another locus it is in linkage with.

It is standard process in many analyses such as association studies to measure the variance of groups of single nucleotide polymorphisms (SNPs) with respect to a trait being studied. The SNP density along and across chromosomes is often chosen with the understanding that even if the SNPs themselves are not causal variants, the causal variants, often termed quantitative trait loci (QTL), are likely in linkage disequilibrium with multiple SNPs and thus SNP variance with respect to trait variance is related to the causal variant variance with respect to the trait. The typical expectation is that given a casual variant's variance, $2p(1-p)$, the SNP variance with respect to the variance of the trait is $2p(1-p)r^2$.

Assuming the SNP is represented by the alleles at locus $A$ and the causal variant by those at locus $B$ and using the first term of equation 21 we can determine that $p_B(1-p_B)r^2$ is approximately equivalent to

$$\sigma_B^2 r^2 = \sigma_B^2 - \sigma_{B|A}^2 \tag{22}$$

So the expected variance at the SNP that is explained by the trait variance is the causal variant variance minus the conditional variance of the causal allele given the SNP allele. When there is statistical independence between the loci the conditional variance equals the marginal variance and thus there is zero variance due to the trait at the independent locus. Under complete dependence the conditional variance is zero and the trait allele variance explains all the variance at the SNP. Under linear conditions equation 22 is approximately $r^2\sigma_B^2$ as expected for moderate minor allele frequencies. Under nonlinear conditions though this relationship is altered by the same

asymmetry that affects the conditional probabilities.

For consistency, we will define $P_A \geq P_B$ and $P_A \geq 0.5$. The proportion of the trait variance that can be inferred at a SNP in linkage, thus becomes

$$\frac{\sigma_B^2 - \sigma_{B|A}^2}{\sigma_B^2} \tag{23}$$

In Tables 1, 2, and 3 this is shown with respect to both alleles at different values of $\Lambda$. For the lower frequency allele $P_B$, the proportion of variance in the SNP accounted for with respect to trait variance steadily declines from approximately $r^2$ at $\Lambda = 1$ to lower values for $\Lambda = r^2$ and $D' = 1$. The exact amount can be calculated based on equation 19.

$$\frac{\sigma_B^2 - \sigma_{B|A}^2}{\sigma_B^2} = \frac{r\sqrt{\Lambda}}{P_A}\left[P_{A|B} + P_A - 1\right] \tag{24}$$

The other locus shows the opposite trend with respect to the variance of the causal allele explained by the variance in the SNP. It increases from approximately $r^2$ at $\Lambda = 1$ to 1 at $\Lambda = r^2$. This is due to $D' = 1$ where the variant is nearly completely associated with the SNP but not the opposite.

$$\frac{\sigma_A^2 - \sigma_{A|B}^2}{\sigma_A^2} = \frac{r}{\sqrt{\Lambda}P_B}\left[P_{B|A} + P_B - 1\right] \tag{25}$$

Figure 4 shows the deviations of $(\sigma_A^2 - \sigma_{A|B}^2)/\sigma_A^2$ from $r^2$ where negative values indicate a proportion of variance explained lower than $r^2$ would predict while positive values predict a higher proportion of variance than $r^2$ would predict. Most values, especially where $P \approx 0.5$ are close to $r^2$ and large deviations typically appear for very low minor allele frequencies.

These results may appear to contradict the relationships of linear dependence where $r^2$ may underestimate total dependence when nonlinearity is present but $r^2$ cannot overestimate dependence. However, the overall dependence between the two loci remains relatively balanced as the dependence in one direction increases while that in the other direction decreases. When looking at the corresponding relationship between the minor alleles at these same loci, the proportion variation explained does differ, however the direction of change with $\Lambda$, reverses between $\sigma_{a|b}^2$ and $\sigma_{b|a}^2$ compared to its trend for the major alleles at the same loci so that the higher frequency minor allele still has a lower proportion of its variance explained by the lower frequency minor allele as the relationship becomes increasingly nonlinear.

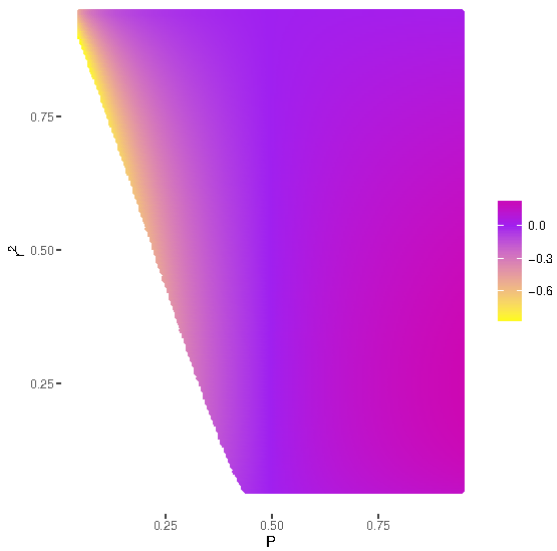


Figure 4: Positive or negative deviation of $\frac{\sigma_A^2 - \sigma_{A|B}^2}{\sigma_A^2}$ from $r^2$ across values of $P_A = P_B = P$ and $r^2$.

This can have significant implications for models or experiments who are trying to relate the variance of one locus to another using only $r^2$ as a dependence measure. For differential allele frequencies where $\Lambda$ is less than one, especially when $r^2$ is small or modest, the actual variance measured based on observed SNP variance can substantially underestimate or overestimate the variance at an unknown site that is in linkage. Therefore, allele frequency matching becomes even more crucial at lower values of $r^2$ across loci since it can skew results in an unexpected manner.

## 6. Nonlinearity of linkage disequilibrium under evolution

Besides an additional descriptor of the dependence structure between a pair of loci, the relative linearity of linkage equilibrium can show distinct patterns under some conditions of evolution. In addition, because it is based only on allele frequencies, $\Lambda$ is more resistant to the decay effects of recombination compared to linkage disequilibrium measures. $D'^2$ and $r^2$ decay due to recombination at the same rates while the effects of recombination on $\Lambda$ generally cancel out except for selective or stochastic effects on allele frequencies.

| $P_A$ | $P_B$ | $r^2$ | $\Lambda$ | $P_{A\|B}$ | $P_{B\|A}$ | $\frac{\sigma_A^2-\sigma_{A\|B}^2}{\sigma_A^2}$ | $\frac{\sigma_B^2-\sigma_{B\|A}^2}{\sigma_B^2}$ |
|---|---|---|---|---|---|---|---|
| 0.665 | 0.5 | 0.5 | 0.5 | 1.0 | 0.75 | 1.0 | 0.25 |
| 0.625 | 0.5 | 0.5 | 0.6 | 0.97 | 0.77 | 0.87 | 0.3 |
| 0.588 | 0.5 | 0.5 | 0.7 | 0.94 | 0.8 | 0.75 | 0.35 |
| 0.555 | 0.5 | 0.5 | 0.8 | 0.91 | 0.82 | 0.66 | 0.4 |
| 0.525 | 0.5 | 0.5 | 0.9 | 0.88 | 0.84 | 0.57 | 0.45 |
| 0.5 | 0.5 | 0.5 | 1 | 0.85 | 0.85 | 0.5 | 0.5 |

Table 1: The structure of linkage disequilibrium where $r^2 = 0.5$ and $p_B = 0.5$ for different amounts of nonlinearity.

| $P_A$ | $P_B$ | $r^2$ | $\Lambda$ | $P_{A\|B}$ | $P_{B\|A}$ | $\frac{\sigma_A^2-\sigma_{A\|B}^2}{\sigma_A^2}$ | $\frac{\sigma_B^2-\sigma_{B\|A}^2}{\sigma_B^2}$ |
|---|---|---|---|---|---|---|---|
| 0.769 | 0.5 | 0.3 | 0.3 | 1.0 | 0.65 | 1.0 | 0.09 |
| 0.665 | 0.5 | 0.3 | 0.4 | 0.96 | 0.67 | 0.82 | 0.12 |
| 0.665 | 0.5 | 0.3 | 0.5 | 0.93 | 0.69 | 0.69 | 0.15 |
| 0.625 | 0.5 | 0.3 | 0.6 | 0.89 | 0.71 | 0.58 | 0.18 |
| 0.588 | 0.5 | 0.3 | 0.7 | 0.86 | 0.73 | 0.5 | 0.21 |
| 0.555 | 0.5 | 0.3 | 0.8 | 0.83 | 0.75 | 0.42 | 0.24 |
| 0.525 | 0.5 | 0.3 | 0.9 | 0.8 | 0.76 | 0.36 | 0.27 |
| 0.5 | 0.5 | 0.3 | 1 | 0.77 | 0.77 | 0.3 | 0.3 |

Table 2: The structure of linkage disequilibrium where $r^2 = 0.3$ and $p_B = 0.5$ for different amounts of nonlinearity.

| $P_A$ | $P_B$ | $r^2$ | $\Lambda$ | $P_{A\|B}$ | $P_{B\|A}$ | $\frac{\sigma_A^2-\sigma_{A\|B}^2}{\sigma_A^2}$ | $\frac{\sigma_B^2-\sigma_{B\|A}^2}{\sigma_B^2}$ |
|---|---|---|---|---|---|---|---|
| 0.886 | 0.7 | 0.3 | 0.3 | 1.0 | 0.79 | 1.0 | 0.21 |
| 0.854 | 0.7 | 0.3 | 0.4 | 0.98 | 0.80 | 0.85 | 0.25 |
| 0.824 | 0.7 | 0.3 | 0.5 | 0.96 | 0.82 | 0.74 | 0.29 |
| 0.796 | 0.7 | 0.3 | 0.6 | 0.94 | 0.83 | 0.66 | 0.32 |
| 0.769 | 0.7 | 0.3 | 0.7 | 0.92 | 0.84 | 0.59 | 0.35 |
| 0.744 | 0.7 | 0.3 | 0.8 | 0.90 | 0.85 | 0.53 | 0.38 |
| 0.722 | 0.7 | 0.3 | 0.9 | 0.88 | 0.86 | 0.48 | 0.41 |
| 0.7 | 0.7 | 0.3 | 1 | 0.86 | 0.86 | 0.44 | 0.44 |

Table 3: The structure of linkage disequilibrium where $r^2 = 0.3$ and $p_B = 0.7$ for different amounts of nonlinearity.

*6.1. Mutation and drift*

For a de novo mutation on a locus with a fixed ancestral allele, the frequency of the new allele is $1/2N_e$ where $N_e$ is the effective population size. Given the existence of only three haplotypes with any other locus at this first generation, $D' = 1$ and $r \approx 0$. There the linkage disequilibrium induced by a new mutation is always extremely nonlinear with $\Lambda \approx 0$. Considering the case of a selectively neutral allele, under conditions of mutation-drift balance, the long-term expected linkage disequilibrium with another locus is $r^2 \approx 1/4N_e c$ (Hill & Robertson (1968)) where $c$ is the rate of recombination per generation between the two loci.

If mutation of ancestral alleles to derived alleles is allowed, but not the reverse, any given locus will not see a stable equilibrium for allele frequencies. However, across an infinite number of sites, a dynamic steady state across all sites does have a stable allele frequency spectrum (Ewens (2004)) where the number of derived alleles with frequency $p$ , $\Phi(p)\delta p$, is given by

$$\Phi(p)\delta p = \frac{4N_e\mu}{p}\delta p \tag{26}$$

where $\mu$ is the mutation rate per site per generation. Under these conditions of mutation-drift balance without selection or migration, the typical allele frequency spectrum is one of many rare alleles, who often do not have the chance to establish themselves due to random loss, and very few alleles of moderate frequency. Where two loci can be considered independent for their respective minor allele frequency probabilities and have a recombination rate, $c$ the expected linkage disequilibrium $r^2$ can be calculated as well as $\Lambda$ based on the statistical probability of minor allele frequencies at each locus. Defining the joint probability of one locus having a minor allele frequency $P_A$ and the other $P_B$ as

$$\frac{\Phi(P_A)\Phi(P_B)}{(\sum_i \Phi(p))^2} = \frac{4N_e\mu}{P_A \sum_i \Phi(p)} \frac{4N_e\mu}{P_B \sum_i \Phi(p)} \tag{27}$$

Following on this, we can then use these probabilities and Equation 13 to calculate the probability of a range of $\Lambda$ between two loci in linkage disequilibrium. Across a range of values of $4N_e\mu$ and $4N_e c$ the distribution is consistent and is shown in Figure 5 from numerical calculations based on the previous equations as well as from a forward time simulation in Python module SimuPop (Peng & Kimmel (2005)) in Figure 6. Due to the plethora

of loci with rare alleles and approximately similar minor allele frequencies, much of the linkage disequilibrium generated by mutation and drift is of linear character, though very weak as typical $r^2$ induced by drift-mutation balance.

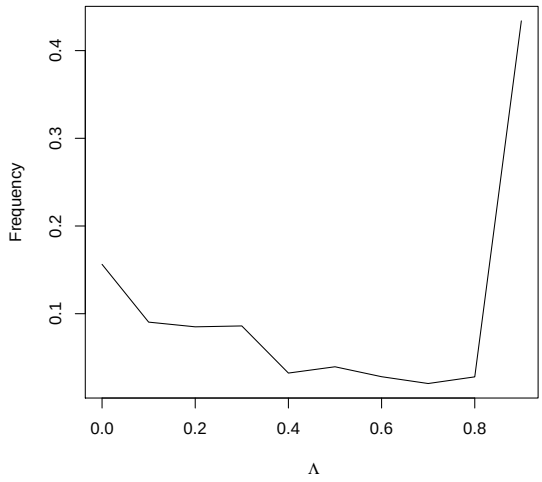


Figure 5: Frequency distribution of the $\Lambda$ between two alleles in an infinite site model of mutation-drift balance based on numerical calculations.

*6.2. Selective sweeps and hitchhiked alleles*

As discussed previously, $\Lambda$ decays much more slowly under the forces of recombination than other linkage disequilibrium measures due to its allele frequency basis. The relative resilience of $\Lambda$ is shown in Figure 7 where the decay of $r$ and $D'$ are superimposed with that of $\Lambda$ for two loci in a simple population with high initial linkage disequilibrium but with no subsequent selection.

While favorable mutations can be relatively rare, those that are favored by selection can see their frequency rapidly increased throughout the population in a selective sweep. Selective sweeps starting as de novo mutations are associated with specific neutral alleles at nearby loci on the chromosome and these alleles can also see their frequencies increase dramatically due to the effect known as hitchhiking. Selective sweeps, particularly hard sweeps, also generate distinct patterns of linkage disequilibrium where loci on either side

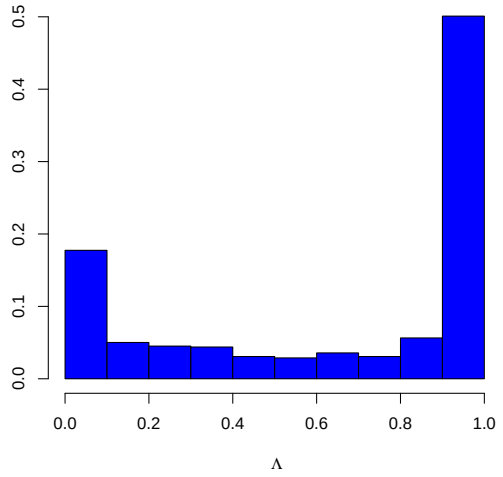


Figure 6: Frequency distribution of the $\Lambda$ between two loci in an infinite site model of mutation-drift balance based off of a forward time simulation of three $N = 200$ populations in SimuPop over 1,000 generations.

of the selected site will develop high linkage disequilibrium with each other even though the linkage disequilibrium with the selected allele can plummet as the latter reaches fixation. It can be demonstrated as well that despite the various values of the linkage disequilibrium in neighboring sites generated by the sweep, the linearity increases, often from nearly nonlinear to nearly linear, and that this pattern of linearity is preserved in the linkage disequilibrium even if selection is relaxed and recombination begins to degrade $D'$ and $r^2$.

During a hard sweep, if neighboring linked neutral alleles obtain a relatively high frequency with minor alleles being relatively rare, they will both be in linkage disequilibrium with each other as well as likely have a linear relationship due to similar allele frequencies at each locus. Not only this, in the event that selection is relaxed, especially before the selected allele reaches fixation, recombination will reassert itself decreasing the linkage disequilibrium but not affecting $\Lambda$ as long as allele frequencies at the loci are not substantially altered by selection, migration, or drift.

A selective sweep with an adjacent locus was simulated in SimuPop. In Figure 8 the linkage disequilibrium between two neighboring loci, the selected site and a neighboring locus, is shown for 400 generations. Selection was

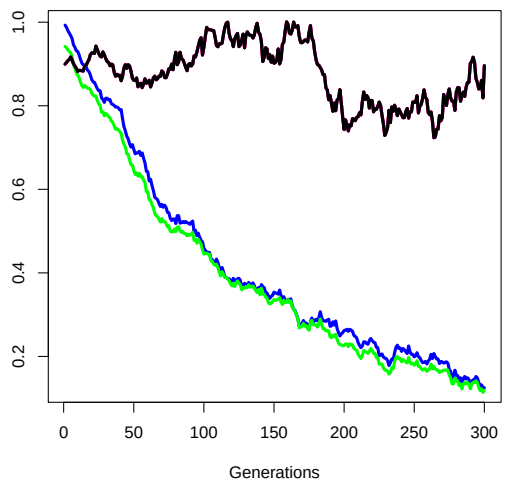


Figure 7: $D'$ (blue), $r$ (green), and $\Lambda$ (black) in a $N = 10,000$ population with initial linkage disequilibrium at $t = 0$ but without selection where $4N_e c = 280$.

relaxed at generation 50, near the point where the allele is roughly halfway to fixation, and the two measures of $r$ and $D'$ are shown along with $\Lambda$ for each time step. The linkage disequilibrium rapidly increases with the sweep as expected and then declines after selection is relaxed and no longer counters recombination. For many generations, until linkage disequilibrium almost disappears completely, the high degree of linearity induced by the sweep is a relatively stable marker of the former evolutionary dynamics.

In a region where a selective sweep is expected to have occurred in the past, if allele frequencies can be assumed stable, a marker of the former association could be the level of linearity in linkage disequilibrium in surrounding regions. Assuming the association was once strong, where a $D' = 1$ is usually more likely than a comparatively high $r^2$ due to the initial linkage disequilibrium of the favorable mutation, a rough measure of the number of generations since selection was relaxed could be given by

$$t = \frac{logD'}{log(1-c)} \tag{28}$$

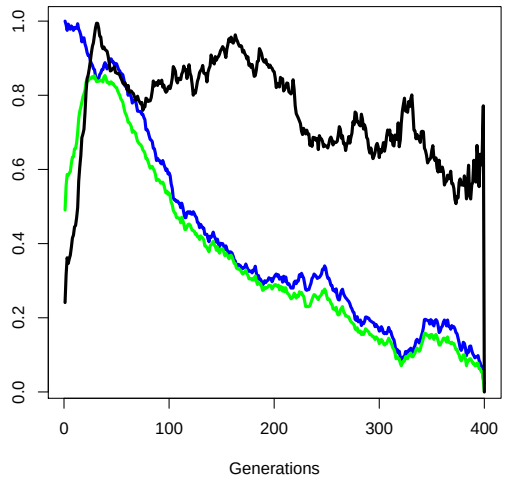


Figure 8: $D'$ (blue), $r$ (green), and $\Lambda$ (black) in a $N = 10,000$ population with two loci, one of which undergoes a selective sweep from generation 1 to 50, after which selection is relaxed. $s = 0.2$, $h = 0.5$, and $4N_e c = 280$

## 7. Conclusion

Linkage disequilibrium is a useful and ubiquitous aspect of all genomes. While its measurement given frequency data may seem relatively straightforward, it would behoove all researchers to understand that the measurement of a value of linkage disequilibrium with any preferred measure may mask underlying complexities in the relationship between the loci in question. One measure, be it $r^2$ or $D'$, cannot reconcile these issues. In particular, the realization that $r^2$ measures only linear dependence and is not the same for different allele frequency pairs is essential. Theoretical and experimental analyses based on assumptions of only linear dependence may inadvertently skew estimates of frequency or variance dependence between loci pairs.

Individual measures may be safely used as summary statistics and for comparisons between loci pairs where $\Lambda$ is roughly equivalent. Though the variable $\Lambda$ in and of itself may or may not have any direct evolutionary relevance, it is a useful summary to supplement linkage disequilibrium measures in order to fully elucidate the frequency dependence between alleles at separate loci. Hopefully in this role it can make studies that rely on linkage

disequilibrium measurements more accurate and insightful.

Declarations of interest: none. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## References

Bateson, W., Saunders E.R., & Punnett, R.C. (1905). Reports to the Evolution Committee of the Royal Society II. Experimental Studies in the Physiology of Heredity. London.

Bateson, W., Saunders E.R., & Punnett, R.C. (1906). Reports to the Evolution Committee of the Royal Society III. Experimental Studies in the Physiology of Heredity. London.

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(9), e142. doi:10.1371/journal.pgen.0020142

Ewens, W.J. (2004) Mathematical population genetics 1: theoretical introduction (Vol. 27). New York: Springer.

Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. The Annals of Mathematical Statistics, 15(1), 25-57.

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. Genetics, 117(2), 331-341.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. Theoretical and Applied Genetics, 38(6), 226-231. doi:10.1007/BF01245622

Jennings, H.S., (1917). The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. Genetics, 2(2), 97-154.

Kaplan, N., & Weir, B. S. (1992). Expected behavior of conditional linkage disequilibrium. American journal of human genetics, 51(2), 333.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature genetics, 22(2), 139. doi:10.1038/9642

Lewontin, R. C. & Kojima, K. I. (1960). The evolutionary dynamics of complex polymorphisms. Evolution, 14(4), 458-472. doi:10.1111/j.1558-5646.1960.tb03113.x

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. Genetics, 49(1), 49-67.

Lewontin, R. C. (1988). On measures of gametic disequilibrium. Genetics, 120(3), 849-852.

Morgan, T. H. (1911). Random segregation versus coupling in Mendelian inheritance. Science, 34(873), 384. doi:10.1126/science.34.873.384

Nei, M. & Li, W. H. (1980). Non-random association between electromorphs and inversion chromosomes in finite populations. Genetics Research, 35(1), 65-83. doi:10.1017/S001667230001394X

Peng, B. & Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. Bioinformatics, 21(18), 3686-3687. doi:10.1093/bioinformatics/bti584

Robbins, R.B., (1918). Applications of mathematics to breeding problems II. Genetics, 3(1), 73-92.

Shannon, C. E. (1948). (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423. doi:10.1002/j.1538-7305.1948.tb00917.x

Smith, R. (2015). A mutual information approach to calculating nonlinearity. Stat, 4(1), 291-303. doi:10.1002/sta4.96

VanLiere, J. M. & Rosenberg, N. A. (2008). Mathematical properties of the $r^2$ measure of linkage disequilibrium. Theoretical Population Biology, 74(1), 130-137. doi:10.1016/j.tpb.2008.05.006

Weinberg, W., (1908). Über vererbungsgesetze beim menschen. Zeitschrift für induktive Abstammungs-und Vererbungslehre, 1(1), 440-460.

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(2), 87-94. doi:10.1375/twin.8.2.87