On measures of association among genetic variables

Daniel Gianola^{*,†}, Eduardo Manfredi[‡] and Henner Simianer[§]

^{*}Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI, 53706, USA. [†]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432, Ås, Norway. [‡]Institut National de la Recherche Agronomique, UR631 Station d'Amélioration, Génétique des Animaux, BP 52627, 32326, Castanet-Tolosan, France. [§]Department of Animal Sciences, Georg-August-Universität, 37075, Göttingen, Germany

Summary

Systems involving many variables are important in population and quantitative genetics, for example, in multi-trait prediction of breeding values and in exploration of multi-locus associations. We studied departures of the joint distribution of sets of genetic variables from independence. New measures of association based on notions of statistical distance between distributions are presented. These are more general than correlations, which are pairwise measures, and lack a clear interpretation beyond the bivariate normal distribution. Our measures are based on logarithmic (Kullback-Leibler) and on relative 'distances' between distributions. Indexes of association are developed and illustrated for quantitative genetics settings in which the joint distribution of the variables is either multivariate normal or multivariate-*t*, and we show how the indexes can be used to study linkage disequilibrium in a two-locus system with multiple alleles and present applications to systems of correlated beta distributions. Two multivariate beta and multivariate beta-binomial processes are examined, and new distributions are introduced: the GMS-Sarmanov multivariate beta and its beta-binomial counterpart.

Keywords association indexes, entropy, gene frequencies, Kullback-Leibler distance, linkage disequilibrium, multivariate beta distributions, pleiotropy

Introduction

The analysis of systems involving several variables is of great importance in population and quantitative genetics, for example, in multi-trait prediction of breeding values or in the study of networks using multi-locus data. The problem addressed here is that of measuring departures of the joint distribution of sets of genetic variables from stochastic independence. The target variables could be, for instance, pleiotropic additive effects of many loci under the infinitesimal model of Fisher (1918), correlated allelic frequencies of loci within bins of physical distance defined by numbers of base pairs, or sets of allelic frequencies that vary at random over clusters of individuals or sub-popula-

tions. Likewise, a study may involve systems of alleles whose joint distribution departs from independence, due either to proximity within a chromosome (linkage) or to forces such as random drift or selection favoring epistatic combinations, which may also create associations among loci on different chromosomes. When the problem is that of the study of the statistical association between allele states at different loci, this usually is referred to as gametic or 'linkage disequilibrium' (LD), with pairwise correlation measures typically employed for characterizing such association. If alleles at two loci are associated due to physical linkage, equilibrium (independence) is eventually attained asymptotically as generations of random mating accrue. LD has received enormous attention in population genetics (e.g., Hill 1974; Hedrick 1987; Lewontin 1988; McVean 2007) and has become an increasingly important topic of research in the light of emerging genomic data, e.g. single nucleotide polymorphism markers and sequence data enable the study of joint evolution of genomic blocks as well as genome-wide association with complex disease or quantitative traits. This last area is one in which Professor Morris Soller has made important contributions, for example, a pioneering paper by Soller & Genizi (1978) in marker-assisted selection, and his 1990 work with

Address for correspondence

D.Gianola, Department of Animal Sciences, University of Wisconsin-Madison, WI 53706, USA.
E-mail: gianola@ansci.wisc.edu

Re-use of this article is permitted in accordance with the Terms and Conditions set out at http://wileyonlinelibrary.com/onlineopen#OnlineOpen_Terms

Accepted for publication 24 November 2011

Weller and Kashi on designs for inferring linkage between markers and quantitative trait loci in dairy cattle. The present paper is in his honor.

The objective of this work is to introduce new measures of association involving a system of variables based on notions of statistical distance between distributions. The resulting metrics are more general than correlations, which are pairwise measures only and lack a parametric interpretation beyond the bivariate normal distribution (e.g., Lewontin 1988), especially when associations are non-linear or when pairs of variables are neither independent nor identically distributed. For example, in the study of gametic disequilibrium, the distribution of correlation-based estimators is frequency dependent (Hedrick 1987), thus complicating comparisons among populations and meta-analyses. Also, there are situations such as in a multivariate *t*-distribution with a diagonal covariance matrix in which variables are uncorrelated and, yet, stochastically dependent. This illustrates that lack of correlation does not provide definitive evidence of independence: in this example, variable *X*, say, informs about *Y* even though uncorrelated. A related question is the inability of correlation to measure association between, say, 10 loci or 20 pleiotropic effects. The measures of association developed in this study are based on departure between the distributions of the variables under independence and association assumptions. Indexes measuring association between random variables, irrespective of their number or of the form of the joint distribution, are suggested and illustrated.

The article is organized as follows. Following this introduction, measures based on logarithmic and on relative ‘distances’ between distributions, as well as indexes of association, are presented. Subsequently, these indexes are developed and illustrated for quantitative genetics settings in which the joint distribution of the variables under consideration is either multivariate normal or multivariate-*t*. The fourth section shows how these indexes can be used to study LD in a two-locus system with multiple alleles; here, a correlation would not have much meaning, if at all. The fifth section presents applications to systems of correlated beta distributions; all these are generalizations of the univariate beta distribution, which has been used extensively in population genetics to study evolution of gene frequencies. In particular, two multivariate beta and multivariate beta-binomial processes are discussed, and new distributions are introduced: the GMS-Sarmanov multivariate beta and its beta-binomial counterpart. The paper concludes with a discussion of the concepts and of other procedures that have been proposed for study of multi-locus LD. Mathematical details are relegated to an appendix.

Measuring association among variables

To motivate the approach used here, consider a pair *X*, *Y* of random variables with realized values *x*, *y*; the ideas

generalize readily to distributions of higher dimension simply by replacing scalars by vectors. Let $p(x, y)$ and $p(x) \times p(y)$ be the densities of the joint distributions under association and independence, respectively; $p(x)$ and $p(y)$ are the marginal densities of *X* and *Y* respectively. Association occurs whenever $p(y|x)$ (or $p(x|y)$) differs from $p(y)$ (or $p(x)$). Measures of dependence (association) can be derived from the density ratio

$$\alpha_{xy} = \frac{p(x, y)}{p(x)p(y)} = \frac{p(y|x)}{p(y)} = \frac{p(x|y)}{p(x)}. \quad (1)$$

If $\alpha_{xy} = 1$ for all pairs of values, the ‘distance’ between distributions is 0 and independence holds. Values of α_{xy} larger or smaller than 1 suggest statistical dependence. For example, α_{xy} larger than 1 indicates ‘coupling’, i.e. the density of certain pairs of values is higher than expected under independence; values smaller than 1 would be indicative of ‘repulsion’. Likewise, strong departures of $\log \alpha_{xy}$ from 0 would be indicative of an association between *x* and *y*.

We discuss two measures of association based on either expected logarithmic distance between distributions or on expected relative distance, as conveyed by the average values of $\log \alpha_{xy}$ and of α_{xy} respectively. These measures are used subsequently to develop indexes of association, irrespective of the number of variables in a system.

Association based on Kullback-Leibler distance

The Kullback-Leibler (KL) logarithmic distance (Kullback 1968) is defined as the sum of two KL discrepancies between the distributions: one under association and the other under independence. The first discrepancy measures departure from association when the latter is true, and the other measures departure from independence when this attribute holds. The two discrepancies are positive by construction and are defined as

$$\begin{aligned} D_{AI} &= \text{Discrepancy (association} \rightarrow \text{independence)} \\ &= E_{p(x, y)}(\log \alpha_{xy}) = \int \int (\log \alpha_{xy}) p(x, y) dx dy, \end{aligned}$$

which measures discrepancy when it is true that *X* and *Y* are associated, and

$$\begin{aligned} D_{IA} &= \text{Discrepancy (independence} \rightarrow \text{association)} \\ &= E_{p(x)p(y)}(\log \alpha_{xy}^{-1}) = \int \int (\log \alpha_{xy}^{-1}) p(x)p(y) dx dy, \end{aligned}$$

measuring discrepancy when independence holds; the range of integration depends on the random processes involved. If the two distributions are identical, both discrepancies are null; on the other hand, large discrepancies reflect stochastic dependence. Above, $E_{p(x, y)}$ and $E_{p(x)p(y)}$ denote expectations taken under the assumptions that (*X*, *Y*) follow either a bivariate distribution (association) or that these variables are independently distributed.

A connection with the likelihood ratio statistic is fairly direct. Suppose that the parameters of the distributions under independence and association are estimated by maximum likelihood (ML), and that these estimates are regarded as if they were 'true' values. Then, $\log \alpha_{xy}$, with the densities evaluated at the corresponding ML estimates would be the log of a ratio between maximized likelihoods. In D_{AI} one would be taking expectations of the log-likelihood ratio over all possible realizations expected if the 'full model' (association) were true; on the other hand, in D_{IA} the expectations of the reciprocal of the log-likelihood would be taken under the assumption of independence. In this context, D_{AI} and D_{IA} are averages of log-likelihood ratios, as opposed to an evaluation at a single realization (in ML estimation parameters vary whereas observations are taken as fixed). Note that D_{AI} and D_{IA} do not involve any requirement of 'nesting' of models, contrary to likelihood ratio tests under regularity conditions, but this is a different issue.

The discrepancies are such that $D_{AI} \neq D_{IA}$, and a unique, symmetric KL distance is defined (Kullback 1968) as

$$KL = D_{AI} + D_{IA}. \quad (2)$$

Then, an index of association (taking values between 0 and 1) is obtained by normalizing D_{IA} as

$$\theta = \frac{D_{IA}}{D_{AI} + D_{IA}}, \quad (3)$$

with

$$1 - \theta = \frac{D_{AI}}{D_{AI} + D_{IA}},$$

provided that KL is not null. A large value of D_{IA} relative to D_{AI} (or the opposite) is indicative of situations in which there is more association than what would be expected under independence so that θ or $1 - \theta$ would be close to 1. Thus, values of θ or of $1 - \theta$ away from $\frac{1}{2}$ provide evidence of discrepancy between the hypotheses of random ensembles of (x, y) pairs vs. associated ensembles.

There is no closed form for D_{IA} or D_{AI} for many distributions, in which case these metrics must be calculated numerically or estimated using Monte Carlo methods, given some parameter values. If association is weak, the two logarithmic discrepancies and KL are nearly null, and this can lead to very unstable, even absurd, Monte Carlo estimates of θ . Also, the KL distance and the two discrepancies are logarithmic, and this perhaps provides a less intuitive notion of 'distance' than one based directly on α_{xy} . An alternative 'distance' measure is presented immediately below.

Association based on relative distance

Define the relative discrepancies

$$D'_{AI} = \int \int \alpha_{xy} p(x, y) dx dy = \int \int \frac{p^2(x, y)}{p(x)p(y)} dx dy, \quad (4)$$

$$D'_{IA} = \int \int \alpha_{xy}^{-1} p(x)p(y) dx dy = \int \int \frac{p^2(x)p^2(y)}{p(x, y)} dx dy \quad (5)$$

and

$$\theta' = \frac{D'_{IA}}{D'_{AI} + D'_{IA}}. \quad (6)$$

The metric D'_{AI} gives the expected value of α_{xy} under association and D'_{IA} gives the expectation of α_{xy}^{-1} under the assumption of independence. These expectations are both positive, by construction, and values of θ' away from $\frac{1}{2}$ can be construed as reflecting departure from independence. For example, $\theta' = \frac{1}{5}$ or $\frac{4}{5}$ would suggest association. The connection with likelihood ratios surfaces again; for example, D'_{IA} is the average likelihood-ratio over all possible realizations of the data under the independence assumption.

In what follows, the proposed indexes are examined for some distributions that are used often in quantitative and population genetics research.

Multivariate Gaussian and t-distributed variables

In quantitative genetics, the analysis often focuses on a set of continuous correlated variables $(X_1, X_2, \dots, X_K)$ that follow some multivariate distribution; these variates could be, for example, additive genetic effects for K traits. Suppose that a matrix of estimates of additive genetic correlations is available (and taken as a 'true' matrix), and that we seek to arrive at a single measure of association among the K genetic values, as opposed to the $\frac{K(K-1)}{2}$ pairwise measures provided by the matrix of genetic correlations. Below, examples of multivariate normal distributions with $K = 2, 3, 4$ and of a bivariate t -distribution are considered to illustrate how θ behaves.

Bivariate normal distribution

Let X_1 and X_2 be two standardized normally distributed genetic values with correlation ρ . For this distribution $D_{AI} = -\log \sqrt{1 - \rho^2}$ and $D_{IA} = \rho^2(1 - \rho^2)^{-1} + \log \sqrt{1 - \rho^2}$ (e.g. Sorensen & Gianola 2002). When the correlation is null, the two discrepancies are 0; when the correlation is perfect, the discrepancies are both ∞ . Thus, $KL = \rho^2(1 - \rho^2)^{-1}$, which is 0 when the two random variables are independent and ∞ under a perfect correlation. Then

$$\theta = \frac{D_{IA}}{D_{AI} + D_{IA}} = 1 + \frac{(1 - \rho^2) \log \sqrt{1 - \rho^2}}{\rho^2}, \quad (7)$$

measures discrepancy away from the independence situation when the latter holds, relative to the KL distance between the two distributions. It can be shown that $\lim_{\rho^2 \rightarrow 0} (\theta) = \frac{1}{2}$

and $\lim_{\rho \rightarrow 1}(\theta) = 1$, providing the lower and upper bounds of θ respectively.

The values of θ and of $1 - \theta$ are plotted against ρ in Fig. 1. The dotted line ('holds water') gives the relative discrepancy (θ) away from the independence models when this is true, and the dashed line ('spills water') depicts the trajectory of $1 - \theta$, i.e. the relative discrepancy under association; note that a given curve is a mirror image of the other. As the absolute value of the correlation increases $\theta \rightarrow 1$, as one would expect. Since an association index with values ranging between 0 and 1 may be easier to interpret, an alternative measure could be

$$\gamma = \frac{2D_{IA}}{D_{AI} + D_{IA}} - 1 = 2\theta - 1.$$

However, γ takes values in $[0,1]$ provided that $\theta \geq \frac{1}{2}$, which holds for a bivariate Gaussian distribution, but not so in general. The relationship between γ and ρ for this Gaussian model is also shown in Fig. 1 (solid line), and the association is suggested as weaker when measured by θ than when assessed with the correlation ρ . For example, if $|\rho| = \frac{1}{2}$, θ is smaller than 0.15. In multivariate systems or in other distributions, θ can be much smaller than $\frac{1}{2}$ so that γ takes negative values, as illustrated below, so it turns out that θ often provides an easier to interpret measure of association than does γ .

Multivariate normal distribution

The metrics under consideration can be applied to distributions with any number of dimensions, leading to mea-

Figure 1 Measures of association of two bivariate Gaussian variables as a function of their correlation (ρ). The straight lines give the strength of the association as measured by the absolute value of ρ . The dotted ('holds water': θ) and dashed ('spills water': $1 - \theta$) lines depict the relative contributions to the Kullback-Leibler distance due to discrepancies under independence and dependence models, respectively. Values of the association measure $\gamma = 2\theta - 1$ are represented by the dark solid line.

sures of departure from independence in more general situations, such as when alleles in a multiple-locus system are studied. To illustrate, consider the K -dimensional multivariate Gaussian distribution ($\mathbf{x} \sim N(\mathbf{0}, \mathbf{R})$, where variates are in standard deviation units so that $\mathbf{R}$ is a correlation matrix; under independence ($\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$). It can be shown (Penny & Roberts 2002) that

$$KL = \frac{1}{2} \text{tr}(\mathbf{R}^{-1} + \mathbf{R}) - K, \quad (8)$$

Where $\text{tr}(\cdot)$ denotes the sum of the diagonal elements of a squared matrix such as $\mathbf{R}$. Further,

$$D_{AI} = \frac{1}{2} [\text{tr}(\mathbf{R}) - \log|\mathbf{R}| - K], \quad (9)$$

and

$$D_{IA} = \frac{1}{2} [\log|\mathbf{R}| + \text{tr}(\mathbf{R}^{-1}) - K], \quad (10)$$

so that

$$\theta = \frac{\frac{1}{2} [\log|\mathbf{R}| + \text{tr}(\mathbf{R}^{-1}) - K]}{\frac{1}{2} \text{tr}(\mathbf{R}^{-1} + \mathbf{R}) - K}. \quad (11)$$

Examples of several multivariate normal distributions are given next.

Equi-correlated trivariate normal distribution

Let three variables be equi-correlated with correlation ρ , so that

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix},$$

with $|\mathbf{R}| = 2\rho^3 - 3\rho^2 + 1$, $\text{tr}(\mathbf{R}) = 3$ and $\text{tr}(\mathbf{R}^{-1}) = 3(\rho + 1)/(1 + \rho - 2\rho^2)$. Using (8)-(11) yields

$$D_{AI} = -\log \sqrt{(\rho - 1)^2(2\rho + 1)},$$

$$D_{IA} = \frac{3\rho^2}{1 + \rho - 2\rho^2} - D_{AI},$$

$$KL = \frac{3\rho^2}{1 + \rho - 2\rho^2},$$

and

$$\theta = 1 - \frac{(\rho - 1)(1 + 2\rho) \log \sqrt{(\rho - 1)^2(1 + 2\rho)}}{3\rho^2}, \quad (12)$$

With $0 < \theta < 1$. For example, for $\rho = 0.25, 0.50$ and 0.75 , the index of association θ takes values 0.49, 0.54 and 0.66 respectively. Figure 2 depicts θ , $1 - \theta$ and γ as a function of the coefficient of correlation. At $\rho = 0$, $\theta = 0.5$, but then it decreases slightly as the correlation increases such that γ does not attain a positive value until γ

Figure 2 Indexes of association as a function of the coefficient of correlation in an equi-correlated trivariate normal distribution. The thick solid line gives the relative Kullback-Leibler discrepancy between distributions when independence holds (θ), and the dashed line gives the relative discrepancy when association is true ($1 - \theta$). The thin line gives the trajectory of $\gamma = 2\theta - 1$.

reaches a value of about 0.3. Note that θ is not symmetric with respect to ρ (the same is true of $|\mathbf{R}|$); for instance, when $\rho = -\frac{1}{4}$, $\theta = 0.59$ so that association would be viewed as stronger than when $\rho = \frac{1}{4}$, since $\theta = 0.49$ then. This is a consequence of the values that the discrepancies D_{IA} and D_{AI} take at varying values of ρ .

Tetra-variate normal distributions

Alternative measures can be derived from eigenvalues of a correlation matrix; however, this is sensible only for distributions in which linear relationships between variables are meaningful. The index θ , derived from notions of statistical distance, does not have this limitation because KL discrepancies have a general meaning and can be calculated for any distribution, at least numerically, given values of the parameters. Naturally, any single measure of association in a multivariate distribution becomes less transparent as the system of variables increases in dimension.

Let $K = 4$ and suppose the variables are equi-correlated with coefficient ρ . Then, $|\mathbf{R}| = -(\rho - 1)^3(3\rho + 1)$, $\text{tr}(\mathbf{R}^{-1}) = -(8\rho + 4)/[(\rho - 1)(3\rho + 1)]$ and

$$\theta = 1 - (\rho - 1)(1 + 3\rho) \frac{\log \sqrt{-(\rho - 1)^3(1 + 3\rho)}}{6\rho^2}. \quad (13)$$

Figure 3 displays how θ and $1 - \theta$ vary against ρ . At $\theta = 0$ one has $\theta = 1 - \theta = \frac{1}{2}$; θ decreases somewhat subsequently and then increases, attaining 0.5 again at about $\rho = 0.45$. The values of the index of association suggest that the independence and association models are not 'too distant' unless the correlation is below -0.10 , if negative, or above $\rho = 0.50$, if positive.

Figure 3 Discrepancies from independence to association (θ , solid line), and from association to independence ($1 - \theta$, dashed line) as a fraction of the Kullback-Leibler distance between two tetravariate normal distributions. The straight lines give the absolute values of the correlation ρ .

Let now $K = 4$ and take as correlation matrix

$$\mathbf{R} = \begin{bmatrix} 1 & 0.9\tau & 0.5\tau & 0.1\tau \\ 0.9\tau & 1 & 0.2\tau & 0.1\tau \\ 0.5\tau & 0.2\tau & 1 & 0.1\tau \\ 0.1\tau & 0.1\tau & 0.1\tau & 1 \end{bmatrix},$$

where $0 \leq \tau \leq 1$, so that when $t = 0$ the four variables are independent. For $t = 1$, $\log |\mathbf{R}| = -2.5459$ and $\text{tr}(\mathbf{R}^{-1}) = 24.8980$, yielding $\theta = 0.8782$. Likewise, for $\tau = \frac{1}{2}$, $\log |\mathbf{R}| = -0.2960$ and $\text{tr}(\mathbf{R}^{-1}) = 4.6540$, so that $\theta = 0.5473$. This illustrates that the strength of the association is proportional to the strength of inter-correlation, and that θ provides a metric for measuring departure from independence.

We finish this section by posing the following question: what does θ say that is not informed by all possible pairwise correlations? Again, let $K = 4$ and take as correlation matrices

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & \frac{1}{4} & 0 \\ 0 & 1 & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix}; \quad \mathbf{R}_2 = \begin{bmatrix} 1 & 0 & 0.6 & 0 \\ 0 & 1 & 0 & 0.6 \\ 0.6 & 0 & 1 & 0.6 \\ 0 & 0.6 & 0.6 & 1 \end{bmatrix}.$$

The salient point is that the correlation structure is such that while pairs of variables (1,3), (2,4) and (3,4) are correlated, (1,2) are independent.

What is the overall strength of the association in the system? Calculations yield $\theta_1 = 0.54$ and $\theta_2 = 0.91$ for the networks of variables characterized by $\rho = \frac{1}{4}$ and $\rho = \frac{6}{10}$, respectively. While the correlations measure association between pairs of nodes, θ gives an indication of the overall degree of association.

Bivariate t -distribution

As mentioned, there are situations in which random variables are jointly dependent, but yet, their correlation is 0,

which would incorrectly suggest that X does not inform about Y . This highlights limitations of the correlation parameter as a metric for statistical dependence between sets of variables. One such case is the multivariate- t distribution, in which variables can be statistically dependent, yet uncorrelated. The univariate and multivariate t -distributions appear in robust linear regression models for quantitative genetic analysis based on pedigrees (e.g., Strandén & Gianola 1998; Sorensen & Gianola 2002; Rosa *et al.* 2003). The univariate- t distribution has also been used as a prior in Bayesian linear regression models for genome-enabled prediction of single traits (Meuwissen *et al.* 2001; Gianola *et al.* 2009).

Suppose two variables, e.g. SNP effects on a pair of traits, follow a bivariate t -distribution with a null mean vector, ν degrees of freedom and a scale matrix that is equal to a 2×2 identity matrix; here, the two effects are uncorrelated but not independent, as shown in the Appendix. For this situation, the density ratio (1) is

$$\alpha_{x_1 x_2} = \frac{\Gamma(\frac{2+\nu}{2})\Gamma(\frac{\nu}{2}) \left[1 + \frac{x_1^2 + x_2^2}{\nu}\right]^{-\frac{(2+\nu)}{2}}}{\Gamma^2(\frac{1+\nu}{2}) \left[\left(1 + \frac{x_1^2}{\nu}\right)\left(1 + \frac{x_2^2}{\nu}\right)\right]^{-\frac{(1+\nu)}{2}}},$$

so that the relative discrepancies in (4) and (5) become

$$D'_{AI} = \frac{\Gamma(\frac{2+\nu}{2})\Gamma(\frac{\nu}{2})}{\Gamma^2(\frac{1+\nu}{2})} E_{p(x_1, x_2 | \nu)} \left\{ \frac{\left[1 + \frac{x_1^2 + x_2^2}{\nu}\right]^{-\frac{(2+\nu)}{2}}}{\left[\left(1 + \frac{x_1^2}{\nu}\right)\left(1 + \frac{x_2^2}{\nu}\right)\right]^{-\frac{(1+\nu)}{2}}} \right\},$$

and

$$D'_{IA} = \frac{\Gamma^2(\frac{1+\nu}{2})}{\Gamma(\frac{2+\nu}{2})\Gamma(\frac{\nu}{2})} E_{p(x_1 | \nu)p(x_2 | \nu)} \left\{ \frac{\left[\left(1 + \frac{x_1^2}{\nu}\right)\left(1 + \frac{x_2^2}{\nu}\right)\right]^{-\frac{(1+\nu)}{2}}}{\left(1 + \frac{x_1^2 + x_2^2}{\nu}\right)^{-\frac{(2+\nu)}{2}}} \right\}.$$

The expectations above do not have closed forms but can be evaluated using Monte Carlo methods: given the value of the degrees of freedom parameter, random numbers can be drawn from a bivariate t -distribution and from two independent univariate t -distributions, to approximate D'_{AI} and D'_{IA} .

The index of association θ' in (6) was computed for t -distributions with $\mu = 1, 2, 4, 6, 8, 20$ and 100 000. The last value produces an approximation of the joint distribution of two independent normal variates, whereas the setting with $\mu = 1$ evaluates distance between a 'bivariate Cauchy' distribution and the distribution of two identically and independently distributed Cauchy random variables. The t -processes with $\mu = 1, 2, 4$ do not have a finite variance, but the distributions are proper (i.e., the density integrates to 1). As shown in Table 1, with sets of 1 million random numbers used for evaluating D'_{AI} , D'_{IA} and θ' , the 'distance' D'_{AI} from association to independence decreases as ν increases, approaching 1, whereas the distance from independence to association D'_{IA} increases, approaching 1 as well. With $\nu = 100\,000$ the random variables are essentially independent since $\theta' = 1 - \theta' = \frac{1}{2}$, leading to $\gamma' = 0$, the setting $\nu = 1$ produces the strongest possible association, with $\theta' = 0$ and both $1 - \theta'$ and θ' equal to 1.

Linkage disequilibrium

Suppose there are C and R alleles at locus A and locus B , respectively, such that data on gametic types can be arranged into an $R \times C$ contingency table. Let p_{ij} be the probability of observing a haplotype with configuration 'ij'. If n gametes are screened and the observed number having such configuration is n_{ij} , the assumption of multinomial sampling leads to

$$\begin{aligned} \alpha_{xy} &= \frac{p(x, y)}{p(x)p(y)} = \frac{\binom{n}{n_{11}, n_{21}, \dots, n_{RC}} \prod_{i=1}^R \prod_{j=1}^C p_{ij}^{n_{ij}}}{\binom{n}{n_{11}, n_{21}, \dots, n_{RC}} \prod_{i=1}^R \prod_{j=1}^C (p_i p_j)^{n_{ij}}} \\ &= \frac{\prod_{i=1}^R \prod_{j=1}^C p_{ij}^{n_{ij}}}{\prod_{i=1}^R \prod_{j=1}^C (p_i p_j)^{n_{ij}}}, \end{aligned}$$

where $p_i = \sum_{j=1}^C p_{ij}$, and $p_j = \sum_{i=1}^R p_{ij}$ are the marginal probabilities of alleles 'i' and 'j' at loci A and B respectively. Define parameters $D_{ij} = p_{ij} - p_i p_j$, measuring departure from independence (i.e. disequilibrium between alleles i and j). In a two-locus model a single D parameter is

Table 1 Departures of zero-mean bivariate t -distributions with a 2×2 identity matrix as scale parameter (so their correlation is null) from independence as measured by D'_{AI} , D'_{IA} and by the indexes of association $\theta' = D'_{IA}/(D'_{AI} + D'_{IA})$, $1 - \theta'$ and $\gamma' = 2(1 - \theta') - 1$ (when $\theta' = \frac{1}{2}$ the random variables are independent)

Degrees of freedom	D'_{AI}	D'_{IA}	$D' = D'_{AI} + D'_{IA}$	θ'	$1 - \theta'$	γ'
1	7.28×10^9	0.4376	7.28×10^9	0.0000	1.0000	1.0000
2	63438837	0.6424	6348838	0.0000	1.0000	1.0000
4	20.7443	0.7947	21.5390	0.0369	0.9631	0.9262
6	2.6299	0.8560	3.4859	0.2456	0.7544	0.5088
8	1.2715	0.8889	2.1604	0.4115	0.5885	0.1770
20	1.0584	0.9528	2.0112	0.4738	0.5262	0.0524
100 000	1.0000	1.0000	2.0000	0.5000	0.5000	0.0000

needed, and by using $p_{ij} = D + p_i p_j$, then α_{xy} would be the ratio between two likelihoods: the 'unrestricted' likelihood (indexed by parameters D, p_i, p_j) and the 'restricted' likelihood (indexed by the two allelic frequencies). The asymptotic connection between the standard log-likelihood ratio statistic and the χ^2 metrics used for multi-locus analysis of LD by Hill (1975) and more recently by Zhao *et al.* (2005) is given, for example, in Agresti (2002).

Under this sampling scheme, the discrepancies become

$$D_{AI} = E_A \sum_{i=1}^R \sum_{j=1}^C n_{ij} \log \left(\frac{p_{ij}}{p_i p_j} \right) = n \sum_{i=1}^R \sum_{j=1}^C p_{ij} \log \left(\frac{p_{ij}}{p_i p_j} \right)$$

and

$$D_{IA} = E_I \sum_{i=1}^R \sum_{j=1}^C n_{ij} \log \left(\frac{p_i p_j}{p_{ij}} \right) = n \sum_{i=1}^R \sum_{j=1}^C p_i p_j \log \left(\frac{p_i p_j}{p_{ij}} \right).$$

Hence

$$\theta = \frac{D_{IA}}{D_{IA} + D_{AI}} = \frac{\sum_{i=1}^R \sum_{j=1}^C p_i p_j \log \left(\frac{p_i p_j}{p_{ij}} \right)}{\sum_{i=1}^R \sum_{j=1}^C p_i p_j \log \left(\frac{p_i p_j}{p_{ij}} \right) - \sum_{i=1}^R \sum_{j=1}^C D_{ij} \log \left(\frac{p_i p_j}{p_{ij}} \right)}.$$

Naturally, θ is not defined when all $D_{ij} = 0$ because then the two distributions would be identical and the KL distance is 0 in that case. Note that D_{AI} is the expected value of the log-likelihood ratio, but under the assumption of association; on the other hand, when the sampling distribution of the log-likelihood statistic is considered, the reference distribution is the null model, i.e., linkage equilibrium. This illustrates an important conceptual difference, apart from the fact that D_{IA} and D_{AI} involve integration over the sampling space of allelic outcomes at the two loci in question.

To illustrate, consider a hypothetical example, patterned after data in Spiess (1989) reproduced partially by Frankham *et al.* (2010). The data pertain to three and four alleles at the major histocompatibility loci *HLA-A* and *HLA-B* respectively; the data, after a modification explained subsequently, are shown in Table 2. Frankham *et al.* (2010) presented only the data for these alleles, yet there are many more variants at these loci. Relative frequencies in Frankham *et al.* (2010) were 'normalized' such that the 12 joint frequencies would add up to one.

Taking haplotype frequencies p_{ij} as true values, the discrepancy of the independence model away from association is

$$\begin{aligned} D_{AI} &= \\ &= n \sum_{i=1}^R \sum_{j=1}^C p_{ij} \log \left(\frac{p_{ij}}{p_i p_j} \right) \\ &= 108 \left[0.0270 \log \frac{0.0270}{0.2964 \times 0.3158} \right. \\ &\quad \left. + 0.0950 \log \frac{0.0950}{0.2964 \times 0.3841} + \dots \right. \\ &\quad \left. \dots + 0.0249 \log \frac{0.0249}{0.2412 \times 0.3001} \right] \\ &= 108 \times 0.2928 = 31.6235 \end{aligned}$$

The relative contributions of each of the alleles at the *B* locus to the 0.2928 in D_{AI} above are (after rounding) 22.71%, 61.77%, 13.76% and 1.76% for *B7*, *B8*, *B35* and *B44* respectively, implying that associations with allele *B8* are responsible for most of the departure of the independence model away from association, assuming this last one is 'true'; note that the n_{ij} do not enter into the calculations, since the frequencies are viewed as true ones. For example, the relative contribution of *B8* to D_{AI} is arrived at as follows:

$$\begin{aligned} &108 \left[0.2456 \log \frac{0.2456}{0.2928 \times 0.3158} \right. \\ &\quad \left. + 0.0402 \log \frac{0.0402}{0.2928 \times 0.3841} \right. \\ &\quad \left. + 0.0070 \log \frac{0.0070}{0.2928 \times 0.3001} \right] \\ &= 108 \times 0.1809, \end{aligned}$$

and then $108 \times 0.1809 / (108 \times 0.2928) = 0.6178$. Likewise,

$$\begin{aligned} D_{IA} &= -n \sum_{i=1}^R \sum_{j=1}^C p_i p_j \log \left(\frac{p_{ij}}{p_i p_j} \right) = -108 \times (-0.4592) \\ &= 49.5987, \end{aligned}$$

and the relative contributions of the four *B*-locus alleles to D_{IA} are 16.80%, 53.93%, 12.11% and 17.16% for *B7*, *B8*, *B35* and *B44* respectively, with *B8* playing a major role again. Similar calculations can be done for the *A* locus. The KL distance is $31.6235 + 49.5987 = 81.2222$, and the KL distance per haplotype scored ($N = 108$) is $81.2222/108 = 0.7521$. The indexes of association are $\theta = \frac{49.5987}{81.2222} \approx 0.61$ and $1 - \theta \approx 0.39$.

Table 2 Hypothetical data for major histocompatibility loci *HLA-A* and *HLA-B*, with three and four alleles respectively. The p_{ij} are haplotype frequencies; $p_i = \sum_{j=1}^C p_{ij}$ and $p_j = \sum_{i=1}^R p_{ij}$ give the allelic frequencies. The hypothetical number of observed haplotypes is assumed to be 108.

Locus	A1	A2	A3	Marginal
B7	$p_{11} = 0.0270$	$p_{12} = 0.0950$	$p_{13} = 0.1743$	$p_1 = 0.2964$
B8	$p_{21} = 0.2456$	$p_{22} = 0.0402$	$p_{23} = 0.0070$	$p_2 = 0.2928$
B35	$p_{31} = 0.0106$	$p_{32} = 0.0651$	$p_{33} = 0.0939$	$p_3 = 0.1696$
B44	$p_{41} = 0.0325$	$p_{42} = 0.1838$	$p_{43} = 0.0249$	$p_4 = 0.2412$
Marginal	$p_{.1} = 0.3158$	$p_{.2} = 0.3841$	$p_{.3} = 0.3001$	1.0 ($n = 108$)

In an $R \times C$ table of alleles, a correlation has little meaning and is not invariant with respect to how alleles are scored (e.g. 0, 1, 2) at each of the intervening loci. On the other hand, the chi-square statistics of Hill (1975) and Zhao *et al.* (2005) are also invariant and have a simple interpretation in terms of deviations expected under the null distribution. Our parameter θ is also score invariant, is independent with respect to the order of rows and columns, and applies to any distribution. For instance, if the simple multinomial sampling model is violated due to, e.g. over-dispersion, an appropriate θ could be developed under such model, whereas a 'normalized' χ^2 would provide a crude, albeit probably useful, approximation.

The 'relative distance' metric takes the form

$$\alpha_{xy} = \frac{p(x, y)}{p(x)p(y)} = \prod_{i=1}^R \prod_{j=1}^C \left(\frac{p_{ij}}{p_i p_j} \right)^{n_{ij}},$$

with

$$D'_{AI} = E_A(\alpha_{xy}) = \sum_{i=1}^R \sum_{j=1}^C \left[\prod_{i=1}^R \prod_{j=1}^C \left(\frac{p_{ij}}{p_i p_j} \right)^{n_{ij}} \binom{n}{n_{11}, n_{21}, \dots, n_{RC}} \prod_{i=1}^R \prod_{j=1}^C p_{ij}^{n_{ij}} \right],$$

and

$$D'_{IA} = E_I(\alpha_{xy}^{-1}) = \sum_{i=1}^R \sum_{j=1}^C \left[\prod_{i=1}^R \prod_{j=1}^C \left(\frac{p_i p_j}{p_{ij}} \right)^{n_{ij}} \binom{n}{n_{11}, n_{21}, \dots, n_{RC}} \prod_{i=1}^R \prod_{j=1}^C (p_i p_j)^{n_{ij}} \right].$$

We evaluated D'_{AI} , D'_{IA} and $\theta' = D'_{IA}/(D'_{IA} + D'_{AI})$ using Monte Carlo sampling. In the calculation of D'_{AI} one million random trials of size 108 each, with $\{n_{ij}\}$ varying at random, were sampled from a multinomial distribution with probabilities p_{ij} ; then, $\prod_{i=1}^R \prod_{j=1}^C \left(\frac{p_{ij}}{p_i p_j} \right)^{n_{ij}}$ was evaluated for each realization and averaged over trials. Similarly, one million trials of size 108 were sampled from a multinomial distribution with probabilities $p_i p_j$ and the realizations of $\prod_{i=1}^R \prod_{j=1}^C \left(\frac{p_i p_j}{p_{ij}} \right)^{n_{ij}}$ were averaged out. The index θ' was estimated using the mean and median of the draws of θ' , yielding 0.9953 and 0.9998. The association (disequilibrium) between alleles at the A and B loci is patent, both when logarithmic (θ) and relative (θ') distances are employed to measure it.

Systems of multivariate beta processes

Bivariate beta distributions

A common finding in studies of population differentiation using Wright's F -statistics (Wright 1931; Cockerham 1969) with single nucleotide polymorphisms (SNPs) is that estimates of allelic frequencies within a bin of adjacent

SNPs are correlated (e.g. Akey *et al.* 2002; Weir *et al.* 2005; Akey 2009). This is typically due to stochastic dependence between alleles at linked loci (producing LD), but evolutionary processes causing dependence among the true frequencies themselves have been suggested; see Ohta (1982) for a population genetics model based on this concept. For example, a correlation between alleles at randomly drawn pairs of loci arises if alleles are conditionally (given the allelic frequencies) independent but with allelic frequencies varying at random according to a beta distribution.

Suppose that pairs of loci are sampled at random from some conceptual population of equi-correlated loci and that all pairs of 'true' frequencies define a probability distribution. Wright (1937) found that a beta distribution arose from a diffusion equation used to study changes in allele frequencies in finite populations, so the beta process is well grounded in population genetics. Beta distributions have been used in studies of population differentiation as mixing processes to create randomness of allelic frequencies, leading to beta-binomial likelihoods or to unrecognized posterior distributions that must be resolved using sampling procedures, e.g., Holsinger (1999), Balding (2003), Beaumont & Balding (2004) and Gianola *et al.* (2010).

Similarly, it would seem sensible to assume that allelic frequencies of pairs of loci within the hypothetical population have an association that can be modelled with bivariate beta distributions to represent random variation of true, albeit unknown, allelic frequencies. For example, suppose that there are S half-sib families that can be viewed as drawn as random from some population and that we are interested in modeling association between pairs of allelic frequencies stemming from such sampling scheme. Here, distributions proposed by Sarmanov (1966) and Olkin & Liu (2003) are discussed and generalized as candidate processes that could be used to model this type of association.

Olkin-Liu bivariate beta distribution

First, we review how a specific bivariate beta distribution arises. As in Olkin & Liu (2003), let U , V and W be random variables following independent standard Gamma distributions with parameters a , b and c respectively. By construction, $X = \frac{U}{U+W}$ and $Y = \frac{V}{V+W}$ possess the beta distributions $X: \text{Beta}(a, c)$ and $Y: \text{Beta}(b, c)$ respectively. Clearly, X and Y are positively correlated (through W), with a strength that depends on the values of a , b , c . Olkin & Liu (2003) show that X and Y have a bivariate beta distribution with density function

$$p(x, y) = \frac{\Gamma(a+b+c)x^{a-1}y^{b-1}(1-x)^{b+c-1}(1-y)^{a+c-1}}{\Gamma(a)\Gamma(b)\Gamma(c)(1-xy)^{a+b+c}}, \quad (14)$$

Where $0 < X < 1$ and $0 < Y < 1$. Moments $E(X^k Y^l)$ cannot be written in closed form but can be approximated numerically or using sampling methods. Large c and small a, b produce correlations close to 0, whereas large a, b or small c produce correlations close to 1 (Olkin & Liu 2003). For example, a correlation equal to 0.002 is obtained for $a = b = 0.01$ and $c = 5$, whereas the correlation is 0.91 for $a = 2.5$, $b = 4$ and $c = 0.1$. Figure 4 displays scatter plots of 5000 samples obtained from each of four bivariate beta distributions. Plot 1 represents a distribution in which the correlation is very low, and yet, there is considerable association between pairs of values near 0, illustrating inadequacy of correlation to reveal association. In plot 2 (resembling a 'meteorite') clustering takes place primarily at large values of X and Y . The two bottom plots depict bivariate beta distributions with similar correlations but with a completely different pattern of association. Clearly, correlation often fails as measure of statistical association.

In this bivariate beta model the density ratio takes the form

$$\alpha_{xy} = \frac{p(x, y)}{p(x)p(y)} = \frac{\Gamma(a+b+c)\Gamma(c)}{\Gamma(a+c)\Gamma(b+c)} \frac{(1-x)^b(1-y)^a}{(1-xy)^{a+b+c}}, \quad (15)$$

Where $p(x)$ and $p(y)$ are the beta densities of X and Y : $Beta(b, c)$. Then

$$\begin{aligned} D_{AI} &= \log \left[\frac{\Gamma(a+b+c)\Gamma(c)}{\Gamma(a+c)\Gamma(b+c)} \right] \\ &\quad + E_{p(x,y)} \left[\log \frac{(1-x)^b(1-y)^a}{(1-xy)^{a+b+c}} \right], \\ D_{IA} &= -\log \left[\frac{\Gamma(a+b+c)\Gamma(c)}{\Gamma(a+c)\Gamma(b+c)} \right] \\ &\quad - E_{p(x)p(y)} \left[\log \frac{(1-x)^b(1-y)^a}{(1-xy)^{a+b+c}} \right], \end{aligned}$$

so that

$$\begin{aligned} KL &= E_{p(x,y)} \{ b \log(1-x) + a \log(1-y) \\ &\quad - (a+b+c) \log(1-xy) \} \\ &- E_{p(x)p(y)} \{ b \log(1-x) + a \log(1-y) - (a+b+c) \log(1-xy) \}. \end{aligned} \quad (16)$$

The corresponding D'_{AI} , D'_{IA} and KL' can be calculated by taking expectations of density ratios as opposed to expectations of their logarithms.

Indexes of association based on D'_{AI} , D'_{IA} and KL' were calculated for the following combinations of parameters, where ρ denotes the expected correlation: 1) $a = 0.01$, $b = 0.01$, $c = 5$ ($\rho = 0.002$); 2) $a = 1$, $b = 0.5$, $c = 0.6$ ($\rho = 0.251$), 3) $a = 1$, $b = 1$, $c = 0.9$ ($\rho = 0.496$) and 4) $a = 2$, $b = 3$, $c = 0.4$ ($\rho = 0.750$). Expectations under the independence model were approximated by drawing 1 mil-

Figure 4 Scatter diagrams of 5000 samples from each of four Olkin-Liu bivariate beta distributions. Plot 1 illustrates a strong association with essentially no correlation. Plot 2 ('meteorite') depicts a limitation of the correlation as a parameter for describing association. Plot 3 suggests association clearly. Plot 4 shows a bivariate distribution that is not trivial: the true correlation (0.46) arises primarily due to weaker association in the 'middle' of the bivariate sampling space.

lion random numbers from independent $X \sim Beta(a, c)$ and $Y \sim Beta(b, c)$ distributions, and then averaging the evaluations of the appropriate expressions over the draws. In the association model, 1 million random triplets (U, V, W) were drawn from independent standard gamma variables with parameters a, b and c , to then form bivariate realizations as $x = \frac{u}{u+w}$ and $y = \frac{v}{v+w}$. Table 3 gives the results, and all quantities (save for the true ρ_{XY}) are Monte Carlo estimates. The indexes of association θ' and $1 - \theta'$ departed from 0.5 and approached 0 and 1, respectively, as the correlation in the bivariate beta distribution became stronger. Likewise, γ' and γ'' drifted away from 0, approaching $|1|$. In settings (3) and (4), θ' and $1 - \theta'$ suggest a stronger association than what would be indicated by ρ .

Olkin-Liu bivariate beta-binomial distribution

In studies of LD, focus is typically on correlation between alleles rather than between frequencies. We discuss how a correlation between alleles can arise when the association stems from frequencies, as could happen when the population consists of clusters of individuals resulting from some family structure. Consider bi-allelic locus l with allelic frequencies $\Pr(A_l) = p_l$ and $\Pr(a_l) = 1 - p_l$; A_l and a_l denote the two alleles at locus l . Suppose that a sample of N individuals is scored, and that the observed number of copies of A_l and a_l are n_{A_l} and n_{a_l} , respectively, with

Table 3 Departures of Olkin-Liu bivariate beta distributions from independence as measured by D'_{AI} , D'_{IA} and by the indexes of association θ' and $1 - \theta'$ (when $\theta' = \frac{1}{2}$ independence holds), for four combinations of parameters a , b and c . $\hat{\rho}_{XY}$ is the Monte Carlo estimate (1 million samples) of the true correlation ρ_{XY} . Means of variables X and Y are estimates from samples from association (A) or independence (I) models. The D'_{AI} and D'_{IA} parameters are estimates under the appropriate assumptions; $KL' = D'_{AI} + D'_{IA}$; $\theta' = D'_{IA}/(D'_{AI} + D'_{IA})$; $\gamma' = 2\theta' - 1$; $\gamma'' = 1 - 2\gamma'$.

Item	$a = 0.01$ $b = 0.01$ $c = 5$	$a = 0.1$ $b = 0.5$ $c = 0.6$	$a = 1$ $b = 1$ $c = 0.9$	$a = 2$ $b = 3$ $c = 0.4$
ρ_{XY}	0.002	0.2510	0.4960	0.7500
$\hat{\rho}_{XY}$	0.002	0.2499	0.4951	0.7304
ave. (X)	12.01×10^{-3} (A) 2.00×10^{-3} (I)	0.1429 (A) 0.1431 (I)	0.5262 (A) 0.5366 (I)	0.8335 (A) 0.8333 (I)
ave. (Y)	1.98×10^{-3} (A) 2.00×10^{-3} (I)	0.4548 (A) 0.4542 (I)	0.5264 (A) 0.5259 (I)	0.8825 (A) 0.8823 (I)
D'_{AI}	0.9999	0.8709	0.4292	0.0745
D'_{IA}	0.9999	0.8009	0.0779	1.944×10^{-5}
KL'	1.9999	1.5781	0.5071	0.0745
θ'	0.5000	0.4481	0.1537	0.0003
$1 - \theta'$	0.5000	0.5519	0.8463	0.9997
γ'	0.0000	-0.1037	-0.6926	-0.9995
γ''	0.0000	0.1037	0.6926	0.9995

$n_{A_i} + n_{a_i} = 2N$. Given p_i , the distribution of n_{A_i} is *Binomial* ($2N, p_i$). Now, let the allelic frequency p_i vary at random among clusters (e.g. sub-populations) according to a beta distribution with parameters (a, c) . Then, the marginal distribution of the observed number of alleles is beta-binomial (e.g., Casella & George 1992; Sorensen & Gianola 2002), that is, the probability of observing n_{A_i} copies among the $2N$ alleles scored is

$$\Pr(n_{A_i} | 2N, a, c) = \binom{2N}{n_{A_i}} \frac{\Gamma(a+c)}{\Gamma(a)\Gamma(c)} \frac{\Gamma(n_{A_i}+a)\Gamma(n_{A_i}+c)}{\Gamma(2N+a+c)}. \quad (17)$$

Consider now a pair of loci, and let the observed number of copies of the four alleles be represented as $DATA = (n_{A_1}, n_{a_1}, n_{A_2}, n_{a_2})$; haplotype frequencies are not relevant in the model that follows. If, given the allelic frequencies $\mathbf{p} = \{p_i\}$, the number of copies of A_1 and A_2 are independently distributed, i.e., the two loci are in linkage equilibrium, the distribution of the observed number of alleles is

$$I(DATA | \mathbf{p}) = \prod_{i=1}^2 \binom{2N}{n_{A_i}} p_i^{n_{A_i}} (1-p_i)^{n_{a_i}}. \quad (18)$$

Suppose now that this pair is a realization from a stochastic process where loci are sampled over a conceptual population of pairs, e.g. the two loci are drawn at random from a physical 'bin' of a certain length in kilobases. This process generates a covariance structure between allelic frequencies of pairs of loci within the 'bin'. The alleles

would be in LD, marginally, even though being in conditional (given the allelic frequencies) equilibrium. If the unobserved, 'true', allelic frequencies follow some bivariate distribution with density $g(p_1, p_2 | \theta)$, where θ denotes its parameters, the density of the marginal distribution of the observed number of allelic counts under this association model would be

$$p_A(DATA) = \left[\prod_{i=1}^2 \binom{2N}{n_{A_i}} \right] \int_0^1 \int_0^1 p_1^{n_{A_1}} (1-p_1)^{n_{a_1}} p_2^{n_{A_2}} (1-p_2)^{n_{a_2}} g(p_1, p_2 | \theta) dp_1 dp_2. \quad (19)$$

If a bivariate Olkin-Liu beta distribution with parameters $\theta = (a, b, c)$ is assumed, the joint density of allelic frequencies p_1 and p_2 as in (14). Letting

$$C(a, b, c) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)},$$

the marginal distribution (19) takes the form

$$p_A(DATA) = \left[\prod_{i=1}^2 \binom{2N}{n_{A_i}} \right] C(a, b, c) \times \int_0^1 \int_0^1 \frac{p_1^{n_{A_1}+a-1} p_2^{n_{A_2}+b-1} (1-p_1)^{n_{a_1}+b+c-1} (1-p_2)^{n_{a_2}+a+c-1}}{(1-p_1 p_2)^{a+b+c}} dp_1 dp_2. \quad (20)$$

This integral cannot be written in closed form, but it can be evaluated using Monte Carlo methods such as importance sampling (e.g. Sorensen & Gianola 2002). One possible scheme is outlined in the Appendix. Under this distribution

$$\alpha_{n_{A_1} n_{A_2}} = \frac{p_A(DATA)}{p_I(DATA)},$$

Where $p_I(DATA)$ is the distribution under independence between observed allele numbers, and given by the product of two beta-binomial distributions, that is

$$p_I(DATA) = \binom{2N}{n_{A_1}} \frac{\Gamma(a+c)}{\Gamma(a)\Gamma(c)} \frac{\Gamma(n_{A_1}+a)\Gamma(n_{A_1}+c)}{\Gamma(2N+a+c)} \times \binom{2N}{n_{A_2}} \frac{\Gamma(b+c)}{\Gamma(b)\Gamma(c)} \frac{\Gamma(n_{A_2}+b)\Gamma(n_{A_2}+c)}{\Gamma(2N+b+c)}. \quad (21)$$

Then

$$\alpha_{n_{A_1} n_{A_2}} = \frac{\Gamma(a+b+c)\Gamma(c)}{\Gamma(a+c)\Gamma(b+c)} \frac{\Gamma(2N+a+c)\Gamma(2N+b+c)}{\Gamma(2N+a+b+c)} \times \frac{\Gamma(n_{A_1}+a)\Gamma(n_{A_1}+b+c)}{\Gamma(n_{A_1}+a)\Gamma(n_{A_1}+c)\Gamma(n_{A_2}+b)\Gamma(n_{A_2}+c)} E_{B_1, B_2} [h(p_1, p_2)] \quad (22)$$

where $E_{B_1, B_2}[h(p_1, p_2)]$ is defined in the Appendix. Preceding the expectation, the first term involves the parameters of the bivariate beta distribution, the second involves sample size (N) as well, and the third one depends on observed allelic counts.

The logarithmic and relative distances between distributions are calculated as

$$D_{AI} = \sum_{n_{A_1}=0}^{2N} \sum_{n_{A_2}=0}^{2N} \log(\alpha_{n_{A_1} n_{A_2}}) p_A(\text{DATA}),$$

$$D_{IA} = - \sum_{n_{A_1}=0}^{2N} \sum_{n_{A_2}=0}^{2N} \log(\alpha_{n_{A_1} n_{A_2}}) p_I(\text{DATA}),$$

$$D'_{AI} = \sum_{n_{A_1}=0}^{2N} \sum_{n_{A_2}=0}^{2N} \alpha_{n_{A_1} n_{A_2}} p_A(\text{DATA}),$$

and

$$D'_{IA} = \sum_{n_{A_1}=0}^{2N} \sum_{n_{A_2}=0}^{2N} \alpha_{n_{A_1} n_{A_2}}^{-1} p_I(\text{DATA}),$$

With θ and θ' calculated as before. Calculations proceed on an entirely numerical basis.

A generalization of the beta process to a situation with more than two loci is in Olkin & Liu (2003). For K loci the joint density of the allelic frequencies takes the form

$$g(p_1, p_2, \dots, p_K | \theta) = \frac{\Gamma(a_1 + a_2 + \dots + a_K)}{\prod_{i=1}^K \Gamma(a_i)} \prod_{i=1}^K \frac{p_i^{a_i-1}}{(1-p_i)^{a_i+1}} \left[1 + \sum_{i=1}^K \frac{p_i}{1-p_i} \right]^{-(a_1+a_2+\dots+a_K)}.$$

The marginal distribution of allelic counts is obtained by mixing a multinomial distribution over the multivariate beta density above, but results are not available in closed form, so a numerical solution is needed again.

Lee-Sarmanov bivariate beta distribution

Lee (1996) proposed a different bivariate beta distribution following Sarmanov (1966); see Danaher & Hardie (2005). A Lee-Sarmanov bivariate beta distribution for allelic frequencies p_1 and p_2 has density

$$g(p_1, p_2 | \theta^*) = B_1(a_1, b_1) B_2(a_2, b_2) [1 + w(p_1 - \mu_1)(p_2 - \mu_2)], \quad (23)$$

Where $B_1(\cdot)$ and $B_2(\cdot)$ are beta densities as before and w is a parameter ($w=0$ produces independence); $\theta^* = (a_1, b_1, a_2, b_2, w)$. Above, $\mu_1 = \frac{a_1}{a_1+b_1}$ and $\mu_2 = \frac{a_2}{a_2+b_2}$ are the expected values of p_1 and p_2 respectively. The marginal densities of p_1 and p_2 under this model are $B_1(\cdot)$ and $B_2(\cdot)$. Further,

$$E(p_1 p_2) = \mu_1 \mu_2 + w \text{Var}(p_1) \text{Var}(p_2),$$

$$\text{Cov}(p_1, p_2) = w \text{Var}(p_1) \text{Var}(p_2),$$

so that

$$\text{Corr}(p_1, p_2) = w \sqrt{\text{Var}(p_1) \text{Var}(p_2)},$$

with

$$\text{Var}(p_i) = \frac{\mu_i(1-\mu_i)}{a_i + b_i + 1}, \quad i = 1, 2.$$

In the Lee-Sarmanov bivariate beta distribution, a null correlation implies independence. Under this model the density ratio becomes

$$\alpha_{p_1 p_2} = \frac{g(p_1, p_2 | \theta^*)}{B_1(a_1, b_1) B_2(a_2, b_2)} = [1 + w(p_1 - \mu_1)(p_2 - \mu_2)],$$

so

$$D_{AI} = E_g \{ \log[1 + w(p_1 - \mu_1)(p_2 - \mu_2)] \},$$

where E_g denotes expectation under association and

$$D_{IA} = -E_{B_1, B_2} \{ \log[1 + w(p_1 - \mu_1)(p_2 - \mu_2)] \},$$

is the corresponding expectation under independence of the distributions. Further

$$\begin{aligned} D'_{AI} &= \int_0^1 \int_0^1 [1 + w(p_1 - \mu_1)(p_2 - \mu_2)]^2 \\ &\quad B_1(a_1, b_1) B_2(a_2, b_2) dp_1 dp_2 \\ &= 1 + \text{Corr}^2(p_1, p_2) \end{aligned} \quad (24)$$

is in closed form, while

$$D'_{IA} = E_{B_1 B_2} [1 + w(p_1 - \mu_1)(p_2 - \mu_2)]^{-1}, \quad (25)$$

can be estimated using Monte Carlo sampling, assuming this expectation exists. Hence

$$\theta' = \frac{E_{B_1 B_2} [1 + w(p_1 - \mu_1)(p_2 - \mu_2)]^{-1}}{1 + \text{Corr}^2(p_1, p_2) + E_{B_1 B_2} [1 + w(p_1 - \mu_1)(p_2 - \mu_2)]^{-1}}.$$

The index θ' was evaluated for three sets of values of the a, b parameters for each of the two intervening beta distributions while setting the correlation at $0, \frac{1}{4}, \frac{3}{4}$ or $\frac{9}{10}$. Metric D'_{AI} in (24) follows directly from the correlation, while D'_{IA} was estimated using 1 million random samples from each of the two beta distributions. For each set of a, b parameters w was calibrated via

$$\begin{aligned} w &= \frac{\text{Corr}(p_1, p_2)}{\sqrt{\text{Var}(p_1) \text{Var}(p_2)}} \\ &= \text{Corr}(p_1, p_2) \sqrt{\frac{(a_1 + b_1)^2 (a_1 + b_1 + 1) (a_2 + b_2)^2 (a_2 + b_2 + 1)}{a_1 b_1 a_2 b_2}}. \end{aligned}$$

Note that D'_{AI} is solely a function of the correlation, so it was the same for the three pairs of beta distributions examined. Using (25), D'_{IA} was estimated by averaging $\alpha_{p_1 p_2}^{-1} = [1 + w(p_1 - \mu_1)(p_2 - \mu_2)]^{-1}$ over draws. Because

$\alpha_{p_1 p_2}^{-1}$ is a ratio between density functions, negative realizations were not used when forming the Monte Carlo estimate (this produces an upward bias but a more precise estimate). As shown in Table 4, the KL distance between the bivariate beta distribution and the independence process increased monotonically with the strength of the correlation. The indexes of association θ' and $1 - \theta'$ also drifted away monotonically from $\frac{1}{2}$ only in setting 2. For instance, in setting 1 θ' increased from 0.50 to 0.66 as the correlation increased from 0 to $\frac{3}{4}$, but decreased to 0.60 when the correlation was $\frac{9}{10}$. The shape of the bivariate beta distribution was examined for this setting by plotting the density at 20 000 random points drawn from each of the marginal *Beta*(2,2) distributions. This is shown in Fig. 5: the Lee-Sarmanov density 'evolves' towards a more complex topography as correlation increases. When correlation grows from 0 to $\frac{1}{4}$, the shape of the joint distribution is not too different from that under independence, as also suggested by KL and θ' . However, when the correlation increases by a factor of 3 from $\frac{1}{4}$ to $\frac{3}{4}$, the KL distance grows only about two times, while θ' and $1 - \theta'$ move away from $\frac{1}{2}$ at an even slower rate; this is expected because θ has an upper bound at 1, whereas KL takes any positive value in the real line. The bimodal shape of the density also suggests that more intensive sampling is needed to obtain reliable estimates of the indexes of association.

As shown in Danaher & Hardie (2005), the marginal distribution of the observed allelic counts DATA = $(n_{A_1}, n_{a_1}, n_{A_2}, n_{a_2})$ is obtained by mixing (18) over the Lee-Sarmanov density (23), leading to

$$p_A(\text{DATA}) = \prod_{l=1}^2 \left[\binom{2N}{n_{A_l}} \frac{\Gamma(a_l + b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(n_{A_l} + a_l)\Gamma(n_{a_l} + b_l)}{\Gamma(2N + a_l + b_l)} \right] \times \left[1 + w \frac{(n_{A_1} - 2N\mu_1)(n_{A_2} - 2N\mu_2)}{(a_1 + b_1)(a_2 + b_2)} \right], \quad (26)$$

Using the fact that under independence the joint distribution is product beta-binomial as in (21), it turns out that

$$\alpha_{n_{A_1} n_{A_2}} = 1 + w \frac{(n_{A_1} - 2N\mu_1)(n_{A_2} - 2N\mu_2)}{(a_1 + b_1)(a_2 + b_2)},$$

so that closed forms are available for

$$D'_{AI} = \sum_{n_{A_1}=0}^{2N} \sum_{n_{A_2}=0}^{2N} \left[1 + w \frac{(n_{A_1} - 2N\mu_1)(n_{A_2} - 2N\mu_2)}{(a_1 + b_1)(a_2 + b_2)} \right]^2 \times \prod_{l=1}^2 \left[\binom{2N}{n_{A_l}} \frac{\Gamma(a_l + b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(n_{A_l} + a_l)\Gamma(n_{a_l} + b_l)}{\Gamma(2N + a_l + b_l)} \right],$$

and

$$D'_{IA} = \sum_{n_{A_1}=0}^{2N} \sum_{n_{A_2}=0}^{2N} \left[1 + w \frac{(n_{A_1} - 2N\mu_1)(n_{A_2} - 2N\mu_2)}{(a_1 + b_1)(a_2 + b_2)} \right]^{-1}$$

Table 4 Departures of Lee-Sarmanov bivariate beta distributions with varying correlation from independence, as measured by D'_{AI} , D'_{IA} and by indexes of association q' and $1 - q'$ (when $\theta' = \frac{1}{2}$ independence holds). The four correlations levels depend on the value of parameter w of the bivariate distribution, and on parameters a, b of the two 'parental' beta distributions. D'_{AI} has closed form and D'_{IA} was estimated using 1 million samples from the beta distributions. $KL' = D'_{AI} + D'_{IA}$; $q' = D'_{IA} / (D'_{AI} + D'_{IA})$.

Setting/Item	D'_{AI}	D'_{IA} (% negative)	KL'	θ'	$1 - \theta'$
(1) $a_1 = b_1 = a_2 = b_2 = 2$					
Correlation (w)					
0(0)	1	1	2	0.5000	0.5000
$\frac{1}{4}$ (5)	1.0625	1.2681 (0.0)	2.3307	0.5441	0.4559
$\frac{3}{4}$ (15)	1.5625	3.0985 (8.5)	4.6610	0.6648	0.3352
$\frac{9}{10}$ (18)	1.8073	2.7369 (11.2)	4.5468	0.6019	0.3981
(2) $a_1 = 10$; $b_1 = 90$; $a_2 = 90$; $b_2 = 10$					
Correlation (w)					
0(0)	1	1	2	0.5000	0.5000
$\frac{1}{4}$ (281)	1.0625	1.1466 (0.4)	2.2091	0.5190	0.4810
$\frac{3}{4}$ (842)	1.5625	2.6824 (6.3)	4.2449	0.6319	0.3681
$\frac{9}{10}$ (1010)	1.8073	4.4937 (8.7)	6.3037	0.7128	0.2871
(3) $a_1 = .25$; $b_1 = .75$; $a_2 = .75$; $b_2 = .25$					
Correlation (w)					
0(0)	1	1	2	0.5000	0.5000
$\frac{1}{4}$ (135)	1.0625	1.3766 (0.9)	2.4391	0.5644	0.4356
$\frac{3}{4}$ (404)	1.5625	1.5094 (4.9)	3.0719	0.4914	0.5086
$\frac{9}{10}$ (485)	1.8073	1.5876 (5.6)	3.3976	0.4673	0.5327

$$\times \prod_{l=1}^2 \left[\binom{2N}{n_{A_l}} \frac{\Gamma(a_l + b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(n_{A_l} + a_l)\Gamma(n_{a_l} + b_l)}{\Gamma(2N + a_l + b_l)} \right].$$

The correlation between number of copies of the A alleles at the two loci is (Danaher & Hardie 2005)

$$\begin{aligned} \text{Corr}(n_{A_1}, n_{A_2}) &= 2Nw \sqrt{\frac{\text{Var}(p_1)\text{Var}(p_2)}{(a_1 + b_1 + 2N)(a_2 + b_2 + 2N)}} \\ &= w \sqrt{\frac{\text{Var}(p_1)\text{Var}(p_2)}{(1 + \frac{a_1 + b_1}{2N})(1 + \frac{a_2 + b_2}{2N})}}, \end{aligned}$$

so that for large N

$$\text{Corr}(n_{A_1}, n_{A_2}) \approx w \sqrt{\text{Var}(p_1)\text{Var}(p_2)} = \text{Corr}(p_1, p_2).$$

A multivariate generalization of (26) is direct but appears that it has not been reported elsewhere. The joint density of the allelic frequencies at L loci takes the form

$$\begin{aligned} g(p_1, p_2, \dots, p_L | \theta^*) &= \sum_{i=1}^L B_i(a_i, b_i) \left[1 + \sum_{j=1}^{L-1} \sum_{i=j+1}^L w_{ij} (p_1 - \mu_1)(p_2 - \mu_2) \right]. \end{aligned} \quad (27)$$

It is straightforward to verify that this density integrates to 1, and that all marginal distributions are beta. Now, let

Figure 5 Plots of the densities of four Lee-Sarmanov bivariate beta distributions having the same Beta (2,2) marginal distributions but differing in the strength of association: as the correlation increases multi-modality emerges.

$$BB_l(n_{A_l}|a_l, b_l, 2N) = \int_0^1 \binom{2N}{n_{A_l}} p_l^{n_{A_l}} (1-p_l)^{2N-n_{A_l}} B_l(a_l, b_l) dp_l$$

$$= \binom{2N}{n_{A_l}} \frac{\Gamma(a_l + b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(n_{A_l} + a_l)\Gamma(2N - n_{A_l} + b_l)}{\Gamma(2N + a_l + b_l)},$$

denote the beta-binomial distribution of the observed number of copies of allele A_l , out of $2N$ copies. The marginal distribution of allelic counts over all loci can be shown to be, after algebra

$$\Pr(n_{A_1}, n_{A_2}, \dots, n_{A_L} | \mathbf{q}^*) = \prod_{l=1}^L BB_l(n_{A_l}|a_l, b_l, 2N)$$

$$\left[1 + \sum_{i=1}^{L-1} \sum_{j=i+1}^L w_{ij} \frac{(n_{A_i} - 2N\mu_i)(n_{A_j} - 2N\mu_j)}{(a_i + b_i + 2N)(a_j + b_j + 2N)} \right]. \quad (28)$$

We propose (for obvious reasons) to name (27) and (28) as multivariate beta GMS-Sarmanov and multivariate beta-binomial GMS-Sarmanov distributions, respectively.

Discussion

The problem addressed in this study was that of measuring departures of the joint distribution of genetic variables from independence. New measures of association based on

notions of statistical distance were proposed and evaluated under several scenarios, spanning from multivariate Gaussian distributions modeling, say, pleiotropic effects, to systems of beta distributions describing association between allelic frequencies. Two hereto seemingly unreported probability distributions were derived, and termed GMS-Sarmanov multivariate beta and GMS-Sarmanov multivariate beta-binomial distributions. The standard LD problem was also dealt with to illustrate the generality of the approaches proposed.

Linkage disequilibrium analysis has been the subject of an enormous amount of research, re-energized with the availability of massive molecular marker and sequence data. For example, in the context of coalescent theory, an important issue is the decoupling of ancestries of sites at different regions of the chromosome, and this is done by studying association between alleles at different loci (Wakeley 2009). Most of the standard measures of LD employed are pairwise statistics such as correlations (e.g. Hill & Robertson 1968; Hill 1974; Hedrick 1987; Lewontin 1988; Morton *et al.* 2001; Pritchard & Przeworski 2001; McVean 2007), because of ease of calculation and, perhaps, ease of interpretation. However, as pointed out by Lewontin (1988), although a correlation is a meaningful parameter (e.g., in terms of the amount of variability of Y explained when X is observed) in a bivariate

normal distribution, it is arguably less so when applied to discrete data, a problem that is well known in quantitative genetic analysis of discrete phenotypes (Dempster & Lerner 1950; Gianola 1982). Further, pairwise measures do not characterize association well if a given genetic system involves many correlated random variables, as in multi-locus measures of LD. Actually, such measures have been reported much less often, e.g. Weir (1996) gives formulae for up to four loci, and Sabatti and Risch (2002) give a measure based on excess of heterozygosity or of homozygosity.

Nothnagel *et al.* (2002) presented an entropy-based index of LD that is related to our approach, but that differs in some respects. These authors calculate the entropy of the allelic distributions under linkage equilibrium and under disequilibrium, and express the difference in entropy as a fraction of the equilibrium entropy. This produces a normalized entropy that takes values between 0 and 1 (0 indicating no association between alleles at different loci). A formal objection is that, while entropy measures uncertainty in a distribution, its relationship to statistical distance between distributions (Kullback 1968) requires more elaboration. Their method is for bi-allelic loci only, and entropy does not always behave well in continuous distributions (Bernardo and Smith, 1994; Sorensen & Gianola 2002), whereas relative entropy measures such as the KL distance are well defined.

Closer to the spirit of our procedures, Liu & Lin (2005) suggested a measure of LD based on a relative KL discrepancy, but it differs from the ones we propose. This is mainly because they use only one of the two components (termed D_{AI} in our paper) of the invariant KL distance, and express it relative to the maximum value it can take. Using their procedures with D_{IA} would produce a different value of association, and maybe a different qualitative picture might emerge from analysis of genetic data. However, their ideas can be embedded in our approach, and expressing association relative to a maximum distance is a well taken point. It also turns out that these results can also be adapted to the continuous domain, with some care. To illustrate this, it suffices to consider two random vectors, $\mathbf{x}$ and $\mathbf{y}$, as generalization to a higher-dimensional system is straightforward. Let $H(\mathbf{x})$, $H(\mathbf{y})$, and $H(\mathbf{x}, \mathbf{y})$ be the entropies (non-negative, by construction, although pathological examples may arise in continuous distributions) of the marginal distributions of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{x}, \mathbf{y}$ respectively, with

$$H(\mathbf{x}) = - \int [\log p(\mathbf{x})] p(\mathbf{x}) d\mathbf{x},$$

$$H(\mathbf{y}) = - \int [\log p(\mathbf{y})] p(\mathbf{y}) d\mathbf{y},$$

and

$$H(\mathbf{x}, \mathbf{y}) = - \int [\log p(\mathbf{x}, \mathbf{y})] p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

The discrepancy away from the association model is

$$D_{AI} = \int \log \left[\frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right] p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}). \quad (29)$$

Note that

$$\begin{aligned} H(\mathbf{x}, \mathbf{y}) &= - \int [\log p(\mathbf{x}) + \log p(\mathbf{y}|\mathbf{x})] p(\mathbf{x}) p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} \\ &= H(\mathbf{x}) + H(\mathbf{y}|\mathbf{x}), \end{aligned}$$

Where $H(\mathbf{y}|\mathbf{x})$ is the entropy of the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$. This implies that $H(\mathbf{x}, \mathbf{y}) \geq H(\mathbf{x})$, where $\mathbf{x}$ can also be any partition of $\mathbf{x}$, e.g. its i th coordinate x_i , say, and similarly $H(\mathbf{x}, \mathbf{y}) \geq H(\mathbf{y})$. Thus, $H(\mathbf{x}, \mathbf{y}) \geq \max_i H(x_i)$. Using this in (29)

$$\begin{aligned} D_{AI} &= H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) \leq D_{AI \max} \\ &= H(\mathbf{x}) + H(\mathbf{y}) - \max_i [H(x_i), H(y_i)]. \end{aligned}$$

More generally, for a vector $\mathbf{z}$ with n elements, the KL discrepancy away from the independence model is

$$\begin{aligned} D_{AI} &= \sum_{i=1}^n H(z_i) - H(z_1, z_2, \dots, z_n) \leq D_{AI \max} \\ &= \sum_{i=1}^n H(z_i) - \max_i [H(z_i)]. \end{aligned} \quad (30)$$

Likewise

$$\begin{aligned} D_{IA} &= \int \log \left[\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right] p(\mathbf{x})p(\mathbf{y}) d\mathbf{x} d\mathbf{y} = C(\mathbf{x}, \mathbf{y}) \\ &\quad - [H(\mathbf{x}) + H(\mathbf{y})], \end{aligned}$$

where

$$C(p(\mathbf{x})p(\mathbf{y}), p(\mathbf{x}, \mathbf{y})) = - \int [\log p(\mathbf{x}, \mathbf{y})] p(\mathbf{x})p(\mathbf{y}) d\mathbf{x} d\mathbf{y}.$$

is the cross-entropy between distributions. From the expressions for D_{AI} above, $H(\mathbf{x}) + H(\mathbf{y}) \geq \max_i [H(x_i), H(y_i)]$, so that

$$\begin{aligned} D_{IA} &= C(\mathbf{x}, \mathbf{y}) - [H(\mathbf{x}) + H(\mathbf{y})] \leq D_{IA \max} \\ &= C(\mathbf{x}, \mathbf{y}) - \max_i [H(x_i), H(y_i)]. \end{aligned}$$

This leads to additional indexes of association, such as

$$\theta^* = \frac{D_{IA}}{D_{IA \max} + D_{AI \max}}, \quad (31)$$

and

$$\theta^{**} = \frac{D_{IA} + D_{AI}}{D_{IA \max} + D_{AI \max}} = \frac{KL}{KL_{\max}}, \quad (32)$$

both taking values between 0 and 1. These indexes were evaluated for a bivariate normal distribution with correlation $\rho = \frac{6}{10}$ and marginal distributions x : $N_1(0,1)$ and y : $N_2(0,1)$. One obtains

$$H(x) = H(y) = \frac{1}{2}(1 + \log 2\pi) \approx 1.4189,$$

$$D_{IA} = \rho^2(1 - \rho^2)^{-1} + \log \sqrt{1 - \rho^2} \approx 0.3394,$$

$$D_{AI} = -\log \sqrt{1 - \rho^2} \approx 0.2231,$$

$$KL = 0.3394 + 0.2231 = 0.5625,$$

$$\theta = \frac{0.3394}{0.5625} = 0.6033,$$

$$D_{AI \max} = 2 \times 1.4189 - 1.4189 = 1.4189,$$

$$C(x, y) = -\log \frac{1}{2\pi\sqrt{1 - \rho^2}} + \frac{1}{1 - \rho^2} = 3.1772,$$

$$D_{IA \max} = 3.1772 - 1.4189 = 1.7583,$$

$$\theta^* = \frac{0.3394}{1.4189 + 1.7583} = 0.1068,$$

and

$$\theta^{**} = \frac{0.5625}{1.4189 + 1.7583} = 0.1770.$$

The four indexes, ρ , θ , θ^{**} and q^{**} produce a different value of the strength of association between random variables in a distribution. This suggests that probably there is no such thing as a universal measure, although it is clear that the coefficient of correlation lacks generality.

In a nutshell, this paper presents measures of association for systems of genetic variables that go beyond standard two-dimensional statistics. The procedures apply to either continuous or discrete data, and typically require numerical implementation, because many of the expressions are not available in closed form, depending on the distribution assumed. Our procedures (like any other method) require knowledge of the parameters of the distributions under independence or association, and estimation of such parameters is not the objective of this paper. Today, computer-intensive approaches for parameter inference, such as Bayesian Markov chain Monte Carlo (Sorensen & Gianola 2002; Gelman *et al.* 2004) or approximate Bayesian computation (Beaumont *et al.* 2002), can be implemented effectively in today's high throughput systems (Wu *et al.* 2011).

Finally, since this volume is in honor of the contributions of Professor Moshe Soller, a relationship between this paper and his work should be established. As noted at the onset of this document, among many other accomplishments, he pioneered marker-assisted selection in animal and plants via exploitation of linkage and LD relationships between markers and unknown quantitative trait loci. Examples of his papers in this area include Soller & Genizi (1978) and Lipkin *et al.* (2009), and these connect with

some of the developments presented here. We look forward to many more scientific accomplishments by Moshe!

Acknowledgements

Part of this work was carried out while the senior author was a Visiting Professor at Georg-August-Universität, Göttingen (Alexander von Humboldt Foundation Senior Researcher Award) and Visiting Scientist at the Station d'Amélioration Génétique des Animaux, Centre de Recherche de Toulouse (Chaire D'Excellence Pierre de Fermat, Agence Innovation, Midi-Pyrenees). Support by the Wisconsin Agriculture Experiment as well as by the Deutsche Forschungsgemeinschaft (GRK 1644/1) is acknowledged. W.G. Hill is thanked for useful comments.

Conflicts of interest

The authors have no conflicts to declare.

References

- Agresti A. (2002) *Categorical Data Analysis*. Wiley, New York.
- Akey J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research* **19**, 711–22.
- Akey J.M., Zhang G., Zhang K., Jin L. & Shriver M.D. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* **12**, 1805–14.
- Balding D.J. (2003) Likelihood-based inference for genetic correlations. *Theoretical Population Biology* **63**, 221–30.
- Beaumont M.A. & Balding D.J. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969–80.
- Beaumont M.A., Zhang W. & Balding D.J. (2002) Approximate bayesian computation in population genetics. *Genetics* **162**, 2025–35.
- Bernardo J.M. & Smith A.F.M. (1994) *Bayram Theory*. Wiley, Chichester.
- Casella G. & George E. (1992) Explaining the Gibbs sampler. *The American Statistician* **46**, 167–74.
- Cockerham C.C. (1969) Variance of gene frequencies. *Evolution* **23**, 72–84.
- Danaher P.J. & Hardie B.G.S. (2005) Bacon with your eggs? Application of a new bivariate beta-binomial distribution. *The American Statistician* **59**, 282–6.
- Dempster E.R. & Lerner I.M. (1950) Heritability of threshold characters. *Genetics* **35**, 212–36.
- Fisher R.A. (1918) On the correlation between relatives on the supposition of Medelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433.
- Frankham R., Ballou J.D. & Briscoe D.A. (2010) *Introduction to Conservation Genetics*. Cambridge, Cambridge.
- Gelman, A., Carlin J.B., Stern H.S. & Rubin D.B. (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton.
- Gianola D. (1982) Theory and analysis of threshold characters. *Journal of Animal Science* **54**, 1079–96.

- Gianola D., de los Campos G., Hill W.G., Manfredi E. & Fernando R.L. (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* **183**, 347–63.
- Gianola D., Simianer H. & Qanbari S. (2010) A two-step method for detecting selection signatures using genetic markers. *Genetics Research* **92**, 141–55.
- Hedrick W. (1987) Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–41.
- Hill W.G. (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**, 229–39.
- Hill W.G. (1975) Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117–26.
- Hill W.G. & Robertson A. (1968) Linkage disequilibrium in finite populations. *Theoretical & Applied Genetics* **38**, 226–31.
- Holsinger K.E. (1999) Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. *Heredity* **130**, 245–55.
- Kullback S. (1968) *Information Theory and Statistics*. Dover, Mineola.
- Lee M.L.T. (1996) Properties and applications of the Sarmanon family of Bivariate distributions. *Communications in Statistics - Theory and Methods* **25**, 1207–22.
- Lewontin R.C. (1988) On measures of gametic disequilibrium. *Genetics* **120**, 849–52.
- Lipkin E., Straus K., Tal Stein R. *et al.* (2009) Extensive long-range and nonsynthetic linkage disequilibrium in livestock populations: deconstruction of a conundrum. *Genetics* **181**, 691–9.
- Liu Z. & Lin S. (2005) Multilocus LD measure and tagging SNP Selection with generalized mutual information. *Genetic Epidemiology* **29**, 353–64.
- McVean G. (2007) Linkage disequilibrium, recombination and selection. In: *Handbook of Statistical Genetics*, 3rd edn (Ed. by D.J. Balding, M. Bishop & C. Cannings), pp. 909–44. Wiley, New York.
- Meuwissen T.H., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–29.
- Morton N.E., Zhang W., Taillon-Miller P., Ennis P., Kwok P.Y. & Collins A. (2001) The optimal measure of allelic association. *Proceedings of the National Academy of Sciences* **98**, 5217–21.
- Nothnagel M., Frst R. & Rohde K. (2002) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Human Heredity* **54**, 186–98.
- Ohta T. (1982) Linkage disequilibrium with the island model. *Genetics* **101**, 139–55.
- Olkin I. & Liu R. (2003) A bivariate beta distribution. *Statistics & Probability Letters* **62**, 407–12.
- Penny W. & Roberts S. (2002) Bayesian multivariate autoregressive models with structured priors. *III Proceedings on Vision, Signal & Image Processing* **149**, 333–41.
- Pritchard J.K. & Przeworski M. (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**, 1–14.
- Rosa , G. J. M. , Padovani C.R. & Gianola D. (2003) Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* **45**, 573–90.
- Sabatti C. & Risch N. (2002) Homozygosity and linkage disequilibrium. *Genetics* **160**, 1707–19.
- Sarmanov O.V. (1966) Generalized normal correlation and two-dimensional Frechet classes. *Doklady (Soviet Mathematics)* **168**, 596–9.
- Soller M. & Genizi A. (1978) The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics* **34**, 47–55.
- Sorensen D. & Gianola D. (2002) *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer, New York.
- Spies E.B. (1989) *Genes in Populations*. Wiley, New York.
- Strandén I. & Gianola D. (1998) Attenuating effects of preferential treatment with Student-t mixed linear models: a simulation study. *Genetics, Selection, Evolution* **30**, 565–83.
- Wakeley J. (2009) *Coalescent Theory: An Introduction*. Roberts & Company, Greenwood Village.
- Weir B.S. (1996) *Genetic Data Analysis*. Sinauer, Sunderland.
- Weir B.S., Cardon L.R., Anderson A.D., Nielsen D.M. & Hill W.G. (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research* **15**, 1468–76.
- Weller J.L., Kashi Y. & Soller M. (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science* **73**, 2525–37.
- Wright S. (1931) Evolution in Mendelian populations. *Genetics* **16**, 97–159.
- Wright S. (1937) The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences* **23**, 307–20.
- Wu X.L., Beissinger T.M., Bauck S., Woodward B., Rosa G.J.M., Weigel K.A., de Leon-Gatti N. & Gianola D. (2011) A primer on high-throughput computing for genomic selection. *Frontiers in Livestock Genomics* **2**, 4. doi: 10.3389/fgene.2011.00004.
- Zhao H., Nettleton D., Soller M. & Dekkers J.C.M. (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetics Research* **86**, 77–87.

Appendix

Bivariate and univariate t-distributions

The density of a K -variate random vector following a multivariate t -distribution is

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{K+\nu}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{\frac{K}{2}}|\mathbf{S}|^{\frac{1}{2}}} \left[1 + \frac{(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\nu} \right]^{-\frac{(K+\nu)}{2}}, \quad (33)$$

where $\boldsymbol{\mu} = E(\mathbf{x})$, $\boldsymbol{\Sigma}$ is the scale matrix, $\nu > 0$ is the degrees of freedom parameter and $\Gamma(\cdot)$ is the Gamma function; the covariance matrix of this distribution is $\boldsymbol{\Sigma}\nu/(\nu - 2)$. The specification $K = 2$, $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_2$, yields a bivariate t -distribution with null mean and covariance matrix $\text{Var}(\mathbf{x}) = \mathbf{I}_2\nu/(\nu - 2)$; the two marginal distributions are univariate- t , each having a null mean, variance $\nu/(\nu - 2)$ and ν degrees of freedom. Although uncorrelated, these two random variables are statistically dependent because

$$p(x_1, x_2|v) = \frac{\Gamma(\frac{2+v}{2})}{\Gamma(\frac{v}{2})\pi v} \left[1 + \frac{\mathbf{x}'\mathbf{x}}{v}\right]^{-\frac{(2+v)}{2}}$$

≠

$$p(x_1|v)p(x_2|v) = \frac{\Gamma^2(\frac{1+v}{2})}{\Gamma^2(\frac{v}{2})\pi v} \left[\left(1 + \frac{x_1^2}{v}\right)\left(1 + \frac{x_2^2}{v}\right)\right]^{-\frac{(1+v)}{2}}.$$

The draws from the bivariate t -distribution needed for evaluating D'_{AI} and D'_{IA} are obtained as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sqrt{\frac{v}{\chi_v^2}}.$$

Under independence, the sampling from 2 univariate t -distributions is done as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} z_1 \sqrt{\frac{v}{\chi_{v,1}^2}} \\ z_2 \sqrt{\frac{v}{\chi_{v,2}^2}} \end{bmatrix}.$$

Above, the z 's are independent draws from $N(0, 1)$, χ_v^2 is a draw from a central chi-square distribution on v degrees of freedom, and $\chi_{v,1}^2$ and $\chi_{v,2}^2$ are two independent draws from the same distribution.

$$\begin{aligned} I(n_{A_1}, n_{A_2}|a, b, c) &= B_1(n_{A_1} + a, n_{A_1} + b + c) \\ &\quad B'_2[(n_{A_2} + b)p_1, (n_{A_2} + a + c)(1 - p_1)]. \end{aligned} \quad (34)$$

Distribution B'_2 depends on p_1 , and has expectation

$$\frac{(n_{A_2} + b)}{n_{A_2} + b + (n_{A_2} + a + c) \frac{(1-p_1)}{p_1}},$$

which would be equal to the mean value under independence only if $p_1 = 0.5$. The double integral in (20) can be represented as

$$\begin{aligned} &\int_0^1 \int_0^1 \frac{p_1^{n_{A_1}+a-1} p_2^{n_{A_2}+b-1} (1-p_1)^{n_{A_1}+b+c-1} (1-p_2)^{n_{A_2}+a+c-1}}{(1-p_1 p_2)^{a+b+c} B_1 B'_2} \\ &\quad B_1 B'_2 dp_1 dp_2 \\ &= \frac{\Gamma(n_{A_1} + a) \Gamma(n_{A_1} + b + c)}{\Gamma(2N + a + b + c)} E_{B_1 B_2} [h(p_1, p_2)], \end{aligned}$$

where

$$h(p_1, p_2) = \frac{\Gamma[(n_{A_2}+b)p_1] \Gamma[(n_{A_2}+a+c)(1-p_1)]}{\Gamma[(n_{A_2}+b)p_1 + (n_{A_2}+a+c)(1-p_1)]} \frac{p_2^{(n_{A_2}+b)(1-p_1)} (1-p_2)^{(n_{A_2}+a+c)p_1}}{(1-p_1 p_2)^{a+b+c}}$$

Olkin-Liu bivariate beta-binomial distribution: sampling scheme

When allelic frequencies are independent, their joint density is the product of the densities of the beta random variables $B_1(n_{A_1} + a, n_{A_1} + b + c)$ and $B_2(n_{A_2} + b, n_{A_2} + a + c)$. However, this does not provide a suitable importance sampling distribution, because dependency would not be recognized at all, with the sample space not visited appropriately. A form of introducing dependency is via an importance sampling density with two dependent beta distributions, such as

and $E_{B_1 B_2}$ denotes expectation with respect to the joint distribution with density $B_1 B'_2$. Using this in (20), the marginal distribution of the data under the association model is

$$\begin{aligned} p_A(\text{DATA}) &= \left[\prod_{l=1}^2 \binom{2N}{n_{A_l}} \right] C(a, b, c) \\ &\quad \frac{\Gamma(n_{A_1} + a) \Gamma(n_{A_1} + b + c)}{\Gamma(2N + a + b + c)} E_{B_1, B_2} [h(p_1, p_2)]. \end{aligned} \quad (35)$$

Expression (35) is a seemingly unreported discrete probability distribution which we term 'Olkin-Liu bivariate beta-binomial distribution'.