Letter to the Editor

Gametic and Zygotic Associations

Rong-Cai Yang[1]

Alberta Agriculture, Food and Rural Development, Edmonton, Alberta T6H 5T6, Canada and Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, Alberta T6G 2P5, Canada

Manuscript received July 3, 2002
Accepted for publication February 5, 2003

NONRANDOM associations between genes at different loci are often assessed in population genetic and evolution studies because such associations provide the basis for inferring about demographic and genetic events in the past, such as population history and evolutionary forces governing the loci. Current intensive interest in the association studies largely stems from the prospect of exploiting the relation between the extent of association and the recombination fraction for fine-scale mapping of quantitative trait loci (QTL) controlling complex diseases in humans (Ardlie *et al.* 2002) or quantitative traits of economical or adaptive importance in animals and plants (Farnir *et al.* 2002). In either case, the focus has been on the use of gametic association or commonly called linkage disequilibrium (LD). Several statistical measures have been proposed to characterize LD (see Hedrick 1987 for review), but the use of these measures is often limited to a pair of alleles at two loci. With increasing availability of multiallelic systems such as microsatellites, pairwise LD measures may be too numerous to be readily manageable and interpretable in initial genome-wide studies. More importantly, unless a stringent significance level is imposed, the large number of required pairwise tests under commonly used significance levels 5 and 1% may produce spurious association realizations (Karlin and Piazza 1981).

Recently, Sabatti and Risch (2002) suggested the use of haplotype homozygosity as a possible measure of LD to circumvent the problem of measuring multilocus associations relating to multiple alleles and loci. When zygotes result from the random union of gametes (*i.e.*, Hardy-Weinberg equilibrium) as assumed in Sabatti and Risch (2002), LD can be estimated from observed homozygosities and heterozygosities. The advantage of this approach is that the homozygosities and heterozygosities are defined independently of the number of alleles per locus, thereby allowing one to measure LD between highly polymorphic markers. In the presence of Hardy-Weinberg disequilibrium as often in natural populations, however, LD is only one of several genic disequilibria that are required for a complete characterization of nonrandom associations at different loci (Cockerham and Weir 1973). In a similar but independent development, Yang (2000, 2002) advocated a direct characterization and test of zygotic associations at multiple loci regardless of whether or not the population is in Hardy-Weinberg equilibrium. The purposes of this letter are (i) to elucidate the relationship between the two approaches by Sabatti and Risch (2002) and by Yang (2000, 2002) and (ii) to point out possible bias in calculating LD if other nonzero genic disequilibria are ignored.

For simplicity, the consideration is given only to the case of two loci (say j and l), each with multiple alleles ($j_1, j_2, \ldots, j_r$; $l_1, l_2, \ldots, l_s$). Frequencies of zygotes at loci j and l from the union of gametes, $j_u\ l_y$ and $j_v\ l_z$ ($u, v = 1, 2, \ldots, r$ and $y, z = 1, 2, \ldots, s$), are written as $P_{vz}^{uy} = P_{uy}^{vz}$. At each locus, a zygote can be either homozygous (denoted as 0) or heterozygous (denoted as 1). Thus, there are four classes of zygotic frequencies at the two loci: (i) double homozygotes [$f(00)$], (ii) homozygotes at locus j and heterozygotes at locus l [$f(01)$], (iii) heterozygotes at locus j and homozygotes at locus l [$f(10)$], and (iv) double heterozygotes [$f(11)$]:

$$f(00) = \sum_{u=1}^{r}\sum_{y=1}^{s} P_{uy}^{uy}; \qquad f(01) = \sum_{u=1}^{r}\sum_{y\neq z}\sum P_{uz}^{uy};$$
$$f(10) = \sum_{u\neq v}\sum\sum_{y=1}^{s} P_{vy}^{uy}; \quad f(11) = \sum_{u\neq v}\sum\sum_{y\neq z}\sum P_{vz}^{uy}.$$

The marginal zygotic frequencies at the two individual loci are: $f(0\cdot) = f(00) + f(01) = 1 - H_j$, $f(1\cdot) = f(10) + f(11) = H_j$, $f(\cdot 0) = f(00) + f(10) = 1 - H_l$, and $f(\cdot 1) = f(01) + f(11) = H_l$, where H_j and H_l are the population heterozygosities at loci j and l, respectively. Sabatti and Risch (2002) used these relations to set up the two-way contingency table for homozygosities and heterozygosities at the two loci, but all in terms of double and single homozygosities. Using the above notation, the four

[1]*Address for correspondence:* Department of Agricultural, Food and Nutritional Science, 410 Agriculture/Forestry Centre, University of Alberta, Edmonton, AB T6G 2P5, Canada.
E-mail: rong-cai.yang@ualberta.ca

classes of zygotic frequencies given in Equation 9 of Sabatti and Risch (2002) are: $f(00)$, $f(\cdot 0) - f(00)$, $f(0\cdot) - f(00)$, and $1 - f(\cdot 0) - f(0\cdot) + f(00)$. Yang (2000, 2002) explicitly described nonrandom association between zygotes at loci j and l called zygotic association (ω) in terms of the following relations:

$$\begin{aligned}\omega &= f(00)f(11) - f(10)f(01)\\ &= f(00) - f(0\cdot)f(\cdot 0)\\ &= -[f(01) - f(0\cdot)f(\cdot 1)]\\ &= -[f(10) - f(1\cdot)f(\cdot 0)]\\ &= f(11) - f(1\cdot)f(\cdot 1).\end{aligned}$$

Furthermore, Yang (2002) showed that this zygotic association could be expressed as a function of various genic disequilibria including LD using Cockerham and Weir's (1973) disequilibrium functions,

$$\omega = \sum_{u=1}^{r}\sum_{y=1}^{s}[2p_uD_{\cdot y}^{uy} + 2q_yD_{u\cdot}^{uy} + 2p_uq_yD_{\cdot\cdot}^{uy} + 2p_uq_yD_{\cdot y}^{u} + (D_{\cdot\cdot}^{uy})^2 + (D_{\cdot y}^{u})^2 + D_{uy}^{uy}], \quad (1)$$

where p_u and q_y are the frequencies of allele u at locus j and allele y at locus l, respectively, and each genic disequilibrium is the deviation of a frequency from that based on random association of genes and accounting for any lower-order disequilibria. For example, LD ($D_{\cdot\cdot}^{uy}$) is the deviation of frequency of gamete $j_u\ l_y$ from the product of frequencies of allele u at locus j and allele y at locus l, $D_{\cdot\cdot}^{uy} = P_{\cdot\cdot}^{uy} - p_uq_y$ with $P_{\cdot\cdot}^{uy} = \Sigma_{v=1}^{r}\Sigma_{z=1}^{s}P_{vz}^{uy}$.

When zygotes result from random union of gametes, all nongametic disequilibria including Hardy-Weinberg disequilibrium disappear (*e.g.*, $D_{u\cdot}^{u\cdot} = D_{\cdot y}^{u\cdot} = D_{\cdot y}^{uy} = D_{uy}^{uy} = 0$). Thus, Equation 1 reduces to

$$\omega = \sum_{u=1}^{r}\sum_{y=1}^{s}[2p_uq_yD_{\cdot\cdot}^{uy} + (D_{\cdot\cdot}^{uy})^2]. \quad (2)$$

Furthermore, if there are only two alleles per locus (u, $y = 1, 2$), the zygotic association becomes

$$\omega = 2(p_1 - p_2)(q_1 - q_2)D + 4D^2, \quad (3)$$

where $D\,(= D_{\cdot\cdot}^{11} = -D_{\cdot\cdot}^{12} = -D_{\cdot\cdot}^{21} = D_{\cdot\cdot}^{22})$ is LD. Equation 3 is essentially the same as Equation 3 of Sabatti and Risch (2002) and is the basis for the homozygosity measure of gametic disequilibrium.

Evidently, since the zygotic association is a composite measure, the direct one-to-one relationship between zygotic and gametic associations is possible only when there are two alleles at each of the two loci with the absence of all nongametic disequilibria (*i.e.*, Equation 3). Thus, with knowledge of ω and allelic frequencies (p's and q's), LD can be calculated by solving the equation $4D^2 + 2(p_1 - p_2)(q_1 - q_2)D - \omega = 0$. In the special case of $p_1 = p_2 = 0.5$ or $q_1 = q_2 = 0.5$,

$$D = \pm\frac{\sqrt{\omega}}{2}. \quad (4a)$$

Unfortunately, ω must be nonnegative to obtain a solution for D. In all other cases,

$$D = \frac{-(p_1 - p_2)(q_1 - q_2)}{4}\left[1 \pm \sqrt{1 + \frac{4\omega}{(p_1 - p_2)^2(q_1 - q_2)^2}}\right] \quad (4b)$$

with the condition of $\omega \geq -[(p_1 - p_2)(q_1 - q_2)/2]^2$. It remains unclear which of the two solutions at a given ω is the right solution for D.

Numerical calculation is carried out to examine patterns of the solutions for D. Consider first the case where all genic disequilibria except for LD are zero. For a given set of gene frequencies, LD falls in the range of $D_{\max}^{-} \leq D \leq D_{\max}^{+}$, where $D_{\max}^{-} = \max(-p_1q_1, -p_2q_2)$ and $D_{\max}^{+} = \min(p_1q_2, p_2q_1)$ with $D_{\min}^{-} = D_{\min}^{+} = 0$. Using disequilibrium functions of Cockerham and Weir (1973), I construct frequencies of 10 genotypes at loci j and l (with double heterozygotes being distinguished) and then group them into the four frequency classes of homozygotes and heterozygotes [$f(00)$, $f(01)$, $f(10)$, and $f(11)$]. The zygotic association is calculated as $\omega = f(00)f(11) - f(10)f(01)$. Given gene frequencies and ω, the two solutions for D as obtained from (4b) are D_1 (for taking the negative at "$\pm$" sign) and D_2 (for taking the positive at $\pm$ sign). Table 1 presents the solutions for six gene frequencies that are equal at the two loci (*i.e.*, $p_1 = q_1 = 0.0$, 0.1, 0.2, 0.3, 0.4, and 0.5), at five levels of D ($D_{\max}^{-}$, 0.5 $D_{\max}^{-}$, 0, 0.5 $D_{\max}^{+}$, and $D_{\max}^{+}$). The equality of two zygotic frequencies, $f(10)$ and $f(01)$, is expected for equal gene frequencies at the two loci. It is evident from Table 1 that when gene frequencies are low, only D_1 is the correct solution, but when gene frequencies are increased toward intermediate ($p_1 = q_1 = 0.5$), D_1 is the correct solution for $D \geq 0$ and D_2 is the correct solution for $D < 0$.

Because ω is a summary statistic at the zygote level, it may represent a loss of haplotype information such as gametic disequilibrium. In other words, zero zygotic association ($\omega = 0$) does not preclude the existence of certain nonzero gametic disequilibria ($D \neq 0$) as evident from Equation 3. Thus, with $\omega = 0$, the nontrivial solution as derived from Equation 4b for LD, $D = -(p_1 - p_2)(q_1 - q_2)/2$, is not necessarily zero unless there are symmetric allele frequencies ($p_1 = p_2 = 0.5$ or $q_1 = q_2 = 0.5$). For example, if $p_1 = q_1 = 0.3$, the nontrivial solution for LD is $D = -0.08$, but zygotic frequencies are $f(00) = 0.3364$, $f(01) = f(10) = 0.2436$, and $f(11) = 0.1764$, leading to $\omega = (0.3364)(0.1764) - (0.2436)^2 = 0$.

In the presence of all genic disequilibria, the relationship between zygotic and gametic associations becomes far less clear (*cf.* Equation 1). Table 2 presents five selected examples of solutions for LD (D_1 and D_2) from zygotic associations (ω). For each of five gene frequencies that are equal at the two loci (*i.e.*, $p_1 = q_1 = 0.1$,

TABLE 1

Solutions for linkage disequilibrium (D_1 and D_2) from zygotic associations (ω) in Hardy-Weinberg equilibrium populations

		Zygotic frequency						
$p_1 = q_1$	D	$f(00)$	$f(01)$	$f(10)$	$f(11)$	ω	D_1	D_2
0.0	0.000	1.0000	0.0000	0.0000	0.0000	0.0000	0.000*	−0.500
0.1	−0.010	0.6600	0.1600	0.1600	0.0200	−0.0124	−0.010*	−0.310
0.1	−0.005	0.6661	0.1539	0.1539	0.0261	−0.0063	−0.005*	−0.315
0.1	0.000	0.6724	0.1476	0.1476	0.0324	0.0000	0.000*	−0.320
0.1	0.045	0.7381	0.0819	0.0819	0.0981	0.0657	0.045*	−0.365
0.1	0.090	0.8200	0.0000	0.0000	0.1800	0.1476	0.090*	−0.410
0.2	−0.040	0.4400	0.2400	0.2400	0.0800	−0.0224	−0.040*	−0.140
0.2	−0.020	0.4496	0.2304	0.2304	0.0896	−0.0128	−0.020*	−0.160
0.2	0.000	0.4624	0.2176	0.2176	0.1024	0.0000	0.000*	−0.180
0.2	0.080	0.5456	0.1344	0.1344	0.1856	0.0832	0.080*	−0.260
0.2	0.160	0.6800	0.0000	0.0000	0.3200	0.2176	0.160*	−0.340
0.3	−0.090	0.3400	0.2400	0.2400	0.1800	0.0036	0.010	−0.090*
0.3	−0.045	0.3301	0.2499	0.2499	0.1701	−0.0063	−0.035	−0.045*
0.3	0.000	0.3364	0.2436	0.2436	0.1764	0.0000	0.000*	−0.080
0.3	0.105	0.4141	0.1659	0.1659	0.2541	0.0777	0.105*	−0.185
0.3	0.210	0.5800	0.0000	0.0000	0.4200	0.2436	0.210*	−0.290
0.4	−0.160	0.3600	0.1600	0.1600	0.3200	0.0896	0.140	−0.160*
0.4	−0.080	0.2896	0.2304	0.2304	0.2496	0.0192	0.060	−0.080*
0.4	0.000	0.2704	0.2496	0.2496	0.2304	0.0000	0.000*	−0.020
0.4	0.120	0.3376	0.1824	0.1824	0.2976	0.0672	0.120*	−0.140
0.4	0.240	0.5200	0.0000	0.0000	0.4800	0.2496	0.240*	−0.260
0.5	−0.250	0.5000	0.0000	0.0000	0.5000	0.2500	0.250	−0.250*
0.5	−0.125	0.3125	0.1875	0.1875	0.3125	0.0625	0.125	−0.125*
0.5	0.000	0.2500	0.2500	0.2500	0.2500	0.0000	0.000*	0.000
0.5	0.125	0.3125	0.1875	0.1875	0.3125	0.0625	0.125*	−0.125
0.5	0.250	0.5000	0.0000	0.0000	0.5000	0.2500	0.250*	−0.250

Zygotic frequencies are constructed using disequilibrium functions involving linkage disequilibrium (D) with all other genic disequilibria being zero at different gene frequencies ($p_1 = q_1$). The correct solution is accompanied with an asterisk (*).

0.2, 0.3, 0.4, and 0.5), minimum and maximum values of Hardy-Weinberg disequilibria (HWD), nonallelic digenic disequilibria including both gametic (D) and nongametic disequilibria (D'), trigenic disequilibria (TRID), and quadrigenic disequilibria (QD) are determined just as LD is determined for Table 1. As with LD, the strength of each genic disequilibrium is represented by the five levels (maximum negative, half-maximum negative, zero, half-maximum positive, and maximum positive). Thus, a total of 3125 ($5 \times 5 \times 5 \times 5 \times 5$) combinations are examined. Frequencies of 10 genotypes are calculated using Cockerham and Weir's (1973) disequilib-

TABLE 2

Selected examples of solutions for linkage disequilibrium (D_1 and D_2) from zygotic associations (ω) in Hardy-Weinberg disequilibrium populations

					Zygotic frequency						
$p_1 = q_1$	HWD	$D = D'$	TRID	QD	$f(00)$	$f(01)$	$f(10)$	$f(11)$	ω	D_1	D_2
0.1	−0.0050	0.0000	0.0000	0.0000	0.6561	0.1539	0.1539	0.0361	0.0000	0.0000	−0.3200
0.2	−0.0200	−0.0200	0.0000	0.0008	0.4000	0.2400	0.2400	0.1200	−0.0096	−0.0145	−0.1655
0.3	0.1050	0.0000	−0.0135	0.0081	0.6997	0.0903	0.0903	0.1197	0.0756	0.1032	−0.1832
0.4	0.1200	−0.0800	−0.0320	0.0128	0.6992	0.0608	0.0608	0.1792	0.1216	0.1646	−0.1846
0.5	−0.1250	0.0000	0.0000	0.0313	0.1875	0.0625	0.0625	0.6875	0.1250	0.1768	−0.1768

Zygotic frequencies are constructed using disequilibrium functions involving Hardy-Weinberg disequilibria (HWD), nonallelic digenic disequilibria including both gametic (D) and nongametic disequilibria (D'), trigenic disequilibria (TRID), and quadrigenic disequilibria (QD) at different gene frequencies ($p_1 = q_1$).

rium functions involving these genic disequilibria and 4 zygotic frequencies are simply appropriate sums of the 10 genotypic frequencies. In the first example, all nonallelic genic disequilibria ($D = D'$, TRID, and QD) are zeros, zygotic association is zero ($\omega = 0$) as expected, and the first solution ($D_1 = 0$) corresponds to the absence of gametic disequilibrium ($D = 0$). However, because one or more nonallelic genic disequilibria are present in each of the remaining four examples, there is no correspondence between either of the two solutions (D_1 or D_2) and D. In the third and fifth examples, there is no LD ($D = 0$), but because of nonzero TRID and/or QD, neither solution is zero. In particular, the fifth example represents a well-known scenario where nonzero quadrigenic disequilibrium between two unlinked loci is present in a population undergoing mixed selfing and random mating with s being the proportion of selfing (*e.g.*, Weir and Cockerham 1973). For the case of two alleles at each of the two loci, the zygotic association is $\omega = \Sigma_{u=1}^{2}\Sigma_{y=1}^{2} D_{uy}^{uy}$, where

$$D_{uy}^{uy} = \frac{4s(1-s)}{(4-s)(2-s)^2} p_u(1-p_u)\, q_y(1-q_y).$$

Clearly, $D_{uy}^{uy} \neq 0$ unless $s = 0$ or $s = 1$. Thus, because of nonzero zygotic association, neither D_1 nor D_2 is even close to zero for a gametic equilibrium ($D = 0$) population.

While the selected examples in Table 2 are somewhat arbitrary, the point is clear: there is little correspondence between gametic and zygotic associations when other types of genic disequilibria are present. Sabatti and Risch (2002, p. 1718) also noted that "unfortunately, the relation between homozygosity and recombination fraction is not always direct . . ." although they considered only the haplotype homozygosity and heterozygosity in a Hardy-Weinberg equilibrium population. The important values of zygote-based measures may lie in (i) their ability to quickly detect suspected "hot spots" of associations in genome-wide scans (Sabatti and Risch 2002) and (ii) the comparative assessment of gametic *vs.* zygotic associations to infer about adaptive significance of genotypes at different loci (Yang 2002). For the genome scanning, the primary purpose of the zygotic association analysis, just like that of the LD analysis, is to detect markers that are tightly linked to QTL. In such detection, spurious associations (false positives) between markers and QTL may occur in two ways. First, strong associations between unlinked loci may arise from many evolutionary factors (see below for a discussion). Genetic designs and statistical tests are now available to avoid these kinds of false-positive findings (Gibson and Muse 2002). Second, the huge number of comparisons that are required to scan the genome for association will inevitably produce abundant false positives unless a significance level that is much more stringent than 5% or 1% is imposed (Karlin and Piazza 1981).

Most current LD studies, whether on evolution or on QTL mapping, focus on patterns of LD as predicted by simple demographic models of population expansions or contractions, but do often acknowledge the impact of other factors such as natural selection, random drift, admixture, or gene flow and inbreeding (*e.g.*, Pritchard and Przeworski 2001; Ardlie *et al.* 2002). In essence, these factors cause the departure from Hardy-Weinberg equilibrium, thereby producing the zygotic association even in a gametic equilibrium population (*cf.* Yang 2000, Table 2, case 4). Thus, if these factors are present but ignored, LD will be definitely over- or underemphasized in evolution or QTL-mapping studies.

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada grant OGP0183983.

LITERATURE CITED

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

Cockerham, C. C., and B. S. Weir, 1973 Descent measures for two loci with some applications. Theor. Popul. Biol. **4:** 300–330.

Farnir, F., B. Grisart, W. Coppieters, J. Riquet, P. Berzi *et al.*, 2002 Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics **161:** 275–287.

Gibson, G., and S. V. Muse, 2002 *A Primer of Genome Science.* Sinauer Associates, Sunderland, MA.

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

Karlin, S., and A. Piazza, 1981 Statistical methods for assessing linkage disequilibrium at the HLA-A, B, C loci. Ann. Hum. Genet. **45:** 79–94.

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

Sabatti, C., and N. Risch, 2002 Homozygosity and linkage disequilibrium. Genetics **160:** 1707–1719.

Weir, B. S., and C. C. Cockerham, 1973 Mixed self and random mating at two loci. Genet. Res. **21:** 247–262.

Yang, R.-C., 2000 Zygotic associations and multilocus statistics in a nonequilibrium diploid population. Genetics **155:** 1449–1458.

Yang, R.-C., 2002 Analysis of multilocus zygotic associations. Genetics **161:** 435–445.

Communicating editor: M. A. Asmussen