Tissue Antigens (1983) **21**, 320–328

Statistical aspects of measuring the strength of associations between HLA antigens and diseases

GLENYS THOMSON[1]*, UZI MOTRO[1]*, and STEVE SELVIN[2]

[1]Genetics Department, and [2]Department of Biomedical and Environmental Health Sciences, University of California, Berkeley, California, USA

It will be demonstrated that the δ measure (attributable risk) of Bengtsson & Thomson (1982) can be used in a wide variety of cases to determine which of a number of HLA antigens associated with a particular disease allele has the strongest genetic association. Tests of significance and confidence intervals for the δ measure will be discussed.

Received for publication 21 September, accepted 2 November 1982

A recent paper by Bengtsson & Thomson (1982) has shown that the epidemiological expression known as population attributable risk (MacMahon & Pugh 1970), or the etiological fraction (Miettinen 1974), can be given a genetic interpretation that allows it to be used to determine which of a number of human histocompatibility (HLA) antigens associated with a particular disease has the strongest genetic association with the disease. This follows from their demonstration that, in certain cases, the attributable risk, denoted δ, is a strictly increasing function of the normalized linkage disequilibrium, denoted D′. In addition, δ is a function solely of D′ and the disease allele frequency, and is not directly dependent on the antigen allele frequency. These properties imply that the antigen with the strongest genetic association with the disease, measured by the highest D′ value, will always have the highest δ value.

These results apply to the case where disease predisposition is due to a "disease" locus within the HLA region, but where this locus is distinct from the HLA–A, B, C, D, or DR loci themselves. Bengtsson & Thomson (1982) have analytically demonstrated the above results for those cases where the mode of inheritance of the "disease" allele is strictly dominant or recessive, and by simulation for more general cases of interest.

In this note an analytical demonstration will be presented to show that the results of Bengtsson & Thomson (1982) are also valid for the general "intermediate" model of disease predisposition, where it is assumed that individuals heterozygous for the "disease" allele are disease susceptible, but less so than homozygous individuals, and also for overdominant models, where heterozygotes have the highest disease susceptibility, if the disease

* Research supported by NIH grant R01 HD 12731.

allele frequency is sufficiently small. Cases where more than one antigen at an HLA locus show positive associations with the disease will also be considered.

Tests of significance for the δ measure will also be investigated, and confidence intervals specified. This will allow comparisons of δ measures to be made, and permit interpretations and inferences concerning multiple antigen associations with a disease.

Genetic properties of the attributable risk

The population attributable risk for any particular HLA antigen is denoted by δ and defined as

$$\delta = \frac{d-p}{1-p} \tag{1}$$

(Levin 1953, MacMahon & Pugh 1970, Bengtsson & Thomson 1982, Thomson & Motro 1983), where d is the frequency of the antigen among the individuals with the disease, and p is the frequency of the antigen in the general population. Bengtsson & Thomson (1982) have demonstrated that if the "disease" allele is less frequent than the antigen allele and if the mode of inheritance of the "disease" allele is recessive, then

$$\delta = D'(2 - D'), \tag{2}$$

where $D' = D/D_{max}$ (Lewontin, 1974), i.e. the linkage disequilibrium between the antigen and "disease" allele, divided by the maximum value that the disequilibrium could take with unchanged allele frequencies (clearly, $-1 \leq D' \leq 1$). If the "disease" allele is strictly dominant, the relationship becomes

$$\delta = D'[1 - p_D(1 - D')/(2 - p_D)], \tag{3}$$

where p_D is the "disease" allele frequency. If the "disease" allele is rare, then, approximately, $\delta \approx D'$ for the dominant case.

The general single locus model of disease susceptibility allows all three genotypes, *DD*, *Dd* and *dd*, at the "disease" locus to be disease susceptible, with penetrance values of f_2, f_1 and f_0, respectively (Suarez 1978, Spielman et al. 1980). For the strict recessive model, $f_2 > 0$, and $f_1 = f_0 = 0$; for the dominant model, $f_2 = f_1 > 0$, and $f_0 = 0$. An intermediate model of disease predisposition, with $f_2 > f_1 > 0$, $f_0 = 0$, has been proposed as the possible mode of inheritance of insulin dependent juvenile diabetes (Spielman et al. 1980, Svejgaard et al. 1980) and possibly other HLA associated diseases. This intermediate model will be investigated as a special case of the general disease susceptibility model with penetrance values f_2, f_1 and f_0 as given above.

In our notation, *A* denotes the antigen of interest, which is positively associated with the disease. We assume that *A* itself is not directly involved in predisposing individuals to disease; the association of *A* with the disease is assumed to be due to linkage disequilibrium between *A* and a closely linked disease susceptibility locus, with alleles *D* and *d*. It is possible to show that the attributable risk of *A*, denoted δ_A and defined in (1), satisfies the relationship

$$\delta_A = D' \left[\delta_D + \frac{1}{P(D)} (f_2 - 2f_1 + f_0)(1 - D') p_D^2 \right], \tag{4}$$

where p_D is the frequency of the disease susceptibility allele *D* ($p_d = 1 - p_D$), P(D) is the frequency of the disease in the population and, assuming Hardy-Weinberg equilibrium, $P(D) = f_2 p_D^2 + 2f_1 p_D p_d + f_0 p_d^2$. δ_D is the attributable risk of *D*, and satisfies

$$\delta_D = 1 - \frac{f_0}{P(D)}. \tag{5}$$

In addition, $D' = D/p_a p_D$, where D is the linkage disequilibrium between alleles *A* and *D*, and $p_A = 1 - p_a$ is the frequency of *A*. If we assume that $p_D < p_A$, i.e. the frequency of the disease susceptibility allele is less than the frequency of the antigen, then the maximum positive value of D is $p_a p_D$ (Thomson 1981). In such a case, $D' = D/D_{max}$. (Note that if $f_0 = 0$, then $\delta_D = 1$ for all f_2 and f_1, as expected.)

Expression (4) can also be written in the form

$$\delta_A = D' \left[\delta_D + \frac{\sqrt{V_D}\, p_D}{p_d K_p} (1 - D') \right], \quad (6)$$

where, following Suarez (1978), V_D is the dominance genetic variance

$$V_D = p_D^2 p_d^2 (f_2 - 2f_1 + f_0)^2 \quad (7)$$

and K_p is the prevalence of the disease in the population.

We note that, as for the recessive and dominant cases (Bengtsson & Thomson 1982), δ_A is not directly dependent on the antigen allele frequency p_A, but is a function of D' and p_D and in the general case also of the penetrance values f_2, f_1 and f_o.

Note that for a strict additive model where $f_2 - 2f_1 + f_0 = 0$,

$$\delta_A = D' \delta_D. \quad (8)$$

(This result also applies for non-genetic models; see Thomson & Motro 1983).

Without loss of generality we will make the assumption that

$$f_1, f_2 > f_o. \quad (9)$$

Differentiating δ_A with respect to D', we find that δ_A is a strictly increasing function of D' for $0 \leq D' \leq 1$, for all intermediate models, i.e. those with $f_2 \geq f_1$. In other words, for all possible values of the disease allele frequency and for every intermediate model of disease predisposition ($f_2 \geq f_1 > f_0$) δ is a strictly increasing function of the (positive) normalized linkage disequilibrium. Of any two antigen alleles that are in positive linkage disequilibrium with the disease allele, the one with the larger D' will also have the larger δ measure, and vice versa.

This monotony of δ_A can be extended to the negative range of D' ($-1 \leq D' < 0$) as well, provided the disease allele frequency p_D is not too large. The exact condition is

$$p_D < \frac{f_1 - f_0}{2f_1 - f_2 - f_0} \quad (10)$$

(A sufficient condition is $p_D < \frac{1}{2}$).

For overdominant models ($f_1 > f_2 > f_0$), δ_A is an increasing function of $0 \leq D' \leq 1$ if condition (10) is satisfied. The condition

$$p_D < \frac{\frac{1}{2}(f_1 - f_0)}{2f_1 - f_2 - f_0} \quad (11)$$

guarantees monotony also in the range $-1 \leq D' < 0$. (A sufficient condition in this case is $p_D < \frac{1}{4}$).

Thus it can be seen that the δ measure can be used in a wide variety of cases to determine which of a number of HLA antigens associated with a particular disease allele has the strongest genetic association.

Remark (negative antigen associations with the disease): the monotony of δ_A as a function of D' also holds when an antigen is in negative linkage disequilibrium with the disease allele, provided the disease allele frequency is not too large. However, in this case, $D' \neq D/D_{max}$, so that δ_A is in fact a function of the normalized linkage disequilibrium (D/D_{max}), the disease allele frequency and the antigen allele frequency. Thus, the genetic interpretation given above for antigens in positive linkage disequilibrium with the disease allele no longer applies for antigens showing negative

association with the disease. Similarly, the genetic interpretation cannot be given to δ_A values calculated for particular genotypes showing associations with the disease. In these cases δ_A can still be used as a measure of the association between a disease and a particular antigen genotype or an antigen showing a negative association with the disease.

Multiple antigen associations with the disease, at a single HLA locus

If two antigens at the same HLA locus, denoted A_1 and A_2, are positively associated with a disease, it is possible to calculate a δ value for each antigen separately. We denote these values by δ_{A_1} and δ_{A_2}, for antigens A_1 and A_2, respectively. In each case equation (4) holds with

$$D' = \frac{\Delta_i}{(1 - p_{A_i})\, p_D} \qquad (12)$$

where

$$\Delta_i = p_{A_iD} - p_{A_i}\, p_D, \qquad i = 1, 2. \qquad (13)$$

and p_{A_iD} denotes the frequency of the haplotype A_iD, etc.

A δ value for both antigens combined, denoted $\delta_{A_1 \cup A_2}$, can also be calculated. It also satisfies (4) with, in this case,

$$D' = \frac{\Delta_1 + \Delta_2}{(1 - p_{A_1} - p_{A_2})\, p_D} \qquad (14)$$

where Δ_1 and Δ_2 are defined in (13). However, note that the δ values are not additive, i.e. $\delta_{A_1 \cup A_2} \neq \delta_{A_1} + \delta_{A_2}$. If the mode of inheritance of the disease predisposing gene is strictly additive, then

$$\delta_{A_1 \cup A_2} > \delta_{A_1} + \delta_{A_2}. \qquad (15)$$

We recommend that the δ value be calculated separately for each antigen that shows an association with the disease.

If disease predisposition is due to two or more "disease" alleles, the same general results obtained above still apply.

Estimating δ and tests of hypotheses

In a retrospective study design, one sample (of size n_1) is drawn from the population of diseased individuals, whereas a second sample (of size n_2) is selected as the control. If the controls are sampled from the entire population, i.e. both diseased and non-diseased individuals, the results of Sheps (1959) for the relative difference can then be applied, with appropriate modifications, to the δ measure. Sheps has shown that the maximum likelihood estimate of δ is

$$\hat{\delta} = \frac{\hat{d} - \hat{p}}{1 - \hat{p}}, \qquad (16)$$

where $\hat{d}$ and $\hat{p}$ ($\hat{p} < 1$) are the sample proportions of antigen carriers among the n_1 cases and the n_2 controls, respectively. The asymptotic variance of $\hat{\delta}$ is

$$\mathrm{Var}\,(\hat{\delta}) = (1 - \delta)^2 \left[\frac{d}{n_1 (1 - d)} + \frac{p}{n_2 (1 - p)} \right] \qquad (17)$$

If the controls are sampled from the non-diseased only, as might be the case in many situations, Walter (1976) has shown that the maximum likelihood estimate of δ is

$$\hat{\delta} = \frac{(1 - \varphi)\,(\hat{d} - \hat{\pi})}{1 - \hat{\pi} - \varphi\,(\hat{d} - \hat{\pi})}, \qquad (18)$$

where $\hat{d}$ and $\hat{\pi}$ are the proportions of antigen carriers in the "disease" and the control sam-

ples, respectively, and φ is the frequency of the diseased in the population. Denoting by π the frequency of antigen carriers within the population of non-diseased, the asymptotic variance of $\hat{\delta}$ is

$$\text{Var}(\hat{\delta}) = \frac{(1-\varphi)^2(1-\delta)^4(1-\pi)^2}{(1-d)^2}\left[\frac{d}{n_1(1-d)} + \frac{\pi}{n_2(1-\pi)}\right] \quad (19)$$

If the frequency of the diseased is very small, both methods yield approximately the same results. Henceforth we assume that the controls are sampled from the entire population, and present confidence intervals and tests of hypotheses for δ.

Noether (1957) has investigated the problem of how to construct a confidence interval for the ratio of two unknown proportions. By using several methods of approximation, slightly different from each other, Noether has shown that it is possible to derive several confidence intervals for that ratio, each with approximately the desired confidence level. We present here two of these intervals, each with a $1-\alpha$ confidence level:

$$\left(1+\frac{z^2}{n_1}\right)^{-1}\left\{\hat{\delta} + \frac{z^2(1-2\hat{p})}{2n_1(1-\hat{p})} \pm (1-\hat{\delta})\, z \sqrt{\psi + \frac{z^2}{4n_1}\left[\frac{1}{n_1(1-\hat{d})^2} + \frac{4\hat{p}}{n_2(1-\hat{p})}\right]}\right\} \quad (I_1)$$

$$\hat{\delta} \pm (1-\hat{\delta})\, z \sqrt{\psi}, \quad (I_2)$$

where $\psi = \frac{\hat{d}}{n_1(1-\hat{d})} + \frac{\hat{p}}{n_2(1-\hat{p})}$

and z stands for $z_{1-\frac{1}{2}\alpha}$, defined by

$$1-\tfrac{1}{2}\alpha = \int_{-\infty}^{z_{1-\frac{1}{2}\alpha}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt,$$

and is available from tables of the standard normal distribution.

Clearly, of the two intervals, (I_2) is much simpler to compute. On the other hand, Noether points out that (I_1) has the advantage of being slightly shorter than (I_2) (in most of the cases), and the real protection level is larger than $1-\alpha$, whereas the protection level of (I_2) is smaller than $1-\alpha$. For large values of n_1, interval (I_1) becomes indistinguishable from interval (I_2).

We can also consider $\log(1-\hat{\delta})$ (see Walter 1976), whose asymptotic variance is

$$\text{Var}(\log(1-\hat{\delta})) = \frac{d}{n_1(1-d)} + \frac{p}{n_2(1-p)}.$$

This gives the confidence interval (I_3) for δ

$$(1-(1-\hat{\delta})\, e^{z\sqrt{\psi}},\ 1-(1-\hat{\delta})\, e^{-z\sqrt{\psi}}). \quad (I_3)$$

Leung & Kupper (1981) used the logit transformation $\log(\hat{\delta}/1-\hat{\delta})$ to obtain the confidence interval (I_4) for δ

$$\left\{\left[1 + \frac{1-\hat{\delta}}{\hat{\delta}} \exp(z\sqrt{\psi}/\hat{\delta})\right]^{-1},\ \left[1 + \frac{1-\hat{\delta}}{\hat{\delta}} \exp(-z\sqrt{\psi}/\hat{\delta})\right]^{-1}\right\} \quad (I_4)$$

Table 1.
The four 95% confidence intervals for δ, namely (I_1), (I_2), (I_3) and (I_4), are given for nine sets of $\hat{d}$ and $\hat{p}$ values, and for three sets of n_1 and n_2 values. See text for details.

$\hat{d}$	$\hat{p}$	n_1	n_2	$\hat{\delta}$	I_1	I_2	I_3	I_4	$\hat{R}$
0.64	0.50	50	500	0.2800	(–0.0043, 0.5243)	(0.0065, 0.5535)	(–0.0527, 0.5075)	(0.0910, 0.6016)	0.4550
0.94	0.09	50	500	0.9341	(0.8218, 0.9774)	(0.8617, 1.0064)	(0.8024, 0.9780)	(0.8140, 0.9787)	19.2063
0.42	0.07	50	500	0.3763	(0.2399, 0.5251)	(0.2285, 0.5242)	(0.2095, 0.5080)	(0.2432, 0.5312)	2.4736
0.44	0.13	50	500	0.3563	(0.2073, 0.5152)	(0.1967, 0.5160)	(0.1751, 0.4977)	(0.2163, 0.5262)	1.5516
0.96	0.27	50	500	0.9452	(0.8156, 0.9849)	(0.8707, 1.0197)	(0.7867, 0.9859)	(0.8038, 0.9864)	9.0855
0.36	0.26	50	500	0.1351	(–0.0305, 0.3277)	(–0.0502, 0.3205)	(–0.0715, 0.3020)	0.0310, 0.4328)	0.6527
0.72	0.21	50	500	0.6456	(0.4718, 0.7796)	(0.4872, 0.8039)	(0.4459, 0.7733)	(0.4769, 0.7844)	1.9321
0.01	0.20	50	500	–0.2375	(–0.2696, –0.1180)	(–0.3018, –0.1732)	(–0.3035, –0.1749)		0.1393
0.06	0.35	50	500	–0.4462	(–0.5389, –0.2568)	(–0.5837, –0.3086)	(–0.5904, –0.3150)		0.1754
0.64	0.50	100	500	0.2800	(0.0746, 0.4647)	(0.0815, 0.4785)	(0.0515, 0.4535)	(0.1269, 0.5100)	0.4550
0.94	0.09	100	500	0.9341	(0.8629, 0.9695)	(0.8829, 0.9852)	(0.8567, 0.9697)	(0.8605, 0.9702)	19.2063
0.42	0.07	100	500	0.3763	(0.2763, 0.4827)	(0.2712, 0.4814)	(0.2619, 0.4731)	(0.2783, 0.4857)	2.4736
0.44	0.13	100	500	0.3563	(0.2470, 0.4707)	(0.2424, 0.4703)	(0.2317, 0.4607)	(0.2520, 0.4764)	1.5516
0.96	0.27	100	500	0.9452	(0.8652, 0.9786)	(0.8925, 0.9979)	(0.8567, 0.9791)	(0.8618, 0.9795)	9.0855
0.36	0.26	100	500	0.1351	(0.0096, 0.2746)	(0.0003, 0.2700)	(–0.0108, 0.2600)	(0.0470, 0.3312)	0.6527
0.72	0.21	100	500	0.6456	(0.5243, 0.7462)	(0.5330, 0.7581)	(–0.5131, 0.7420)	(0.5269, 0.7487)	1.9321
0.01	0.20	100	500	–0.2375	(–0.2774, –0.2522)	(–0.2970, –0.1780)	(–0.2984, –0.1794)		0.1393
0.06	0.35	100	500	–0.4462	(–0.5390, –0.3032)	(–0.5635, –0.3288)	(–0.5684, –0.3334)		0.1754
0.64	0.50	100	100	0.2800	(0.0386, 0.5007)	(0.0448, 0.5152)	(0.0018, 0.4806)	(0.1080, 0.5553)	0.4550
0.94	0.09	100	100	0.9341	(0.8627, 0.9696)	(0.8828, 0.9854)	(0.8564, 0.9697)	(0.8603, 0.9702)	19.2063
0.42	0.07	100	100	0.3763	(0.2722, 0.4868)	(0.2671, 0.4856)	(0.2569, 0.4766)	(0.2747, 0.4901)	2.4736
0.44	0.13	100	100	0.3563	(0.2391, 0.4786)	(0.2343, 0.4783)	(0.2220, 0.4675)	(0.2454, 0.4851)	1.5516
0.96	0.27	100	100	0.9452	(0.8649, 0.9789)	(0.8922, 0.9982)	(0.8558, 0.9792)	(0.8611, 0.9796)	9.0855
0.36	0.26	100	100	0.1351	(–0.0170, 0.3013)	(–0.0269, 0.2972)	(–0.0431, 0.2829)	(0.0376, 0.3847)	0.6527
0.72	0.21	100	100	0.6456	(0.5200, 0.7506)	(0.5286, 0.7626)	(0.5069, 0.7452)	(0.5220, 0.7523)	1.9321
0.01	0.20	100	100	–0.2375	(–0.3383, –0.0914)	(–0.3612, –0.1138)	(–0.3676, –0.1198)		0.1393
0.06	0.35	100	100	–0.4462	(–0.6384, –0.2038)	(–0.6661, –0.2262)	(–0.6837, –0.2421)		0.1754

Leung & Kupper point out that, if $\hat{\delta}$ is between 0.21 and 0.79, (I_4) is shorter than (I_2). They also claim that in this range, the logit transformation method provides confidence coefficients at least as large as the stated nominal values in almost all cases. Of course, the logit transformation can be used only if we assume that the values $\hat{\delta}$ can take are always between 0 and 1.

The four confidence intervals (I_1), (I_2), (I_3) and (I_4) are given in Table 1 for various sets of $\hat{d}$ and $\hat{p}$ values and for three sets of n_1 and n_2 values. These illustrative values were chosen to cover a range of $\hat{d}$, $\hat{p}$ and $\hat{\delta}$ values, and are close to those observed in different HLA disease association studies.

As can be seen from Table 1, the right hand bound for the confidence interval of δ for (I_2) can, in some cases, exceed 1. (This phenomenon is encountered for approximate symmetric confidence limits in the neighborhood of the bounds of the estimated value.) This presents a problem, since the upper bound of δ is 1. In the case of (I_2), the right hand bound exceeds 1 if and only if

$$\left(\frac{\hat{d}}{n_1(1-\hat{d})} + \frac{\hat{p}}{n_2(1-\hat{p})}\right) z^2 > 1. \tag{21}$$

The right hand bound of (I_1) can also exceed 1, but (I_1) presents less of a problem than (I_2) since the right hand bound exceeds 1 if and only if

$$\frac{\hat{p}}{n_2(1-\hat{p})}\, z^2 > 1. \tag{22}$$

If the control sample size (n_2) is sufficiently large, there is no problem with (I_1). With (I_3) and (I_4) the right hand bound is always less than 1.

In order to test H_0: $\delta = \delta_0$, we consider the test statistic $\hat{\delta} = (\hat{d} - \hat{p})/(1 - \hat{p})$, which is the MLE of δ and is asymptotically distributed as a normal variate. The asymptotic variance of δ is given in (17) above. Hence,

$$z = \frac{\hat{\delta} - \delta_0}{\sqrt{\text{var}\,(\hat{\delta})}} \tag{23}$$

is asymptotically distributed as a standard normal variate.

Alternatively, we can consider $\log(1 - \hat{\delta})$, whose asymptotic variance is given in (20) above. Under H_0: $\delta = \delta_0$,

$$z = \frac{\log(1-\hat{\delta}) - \log(1-\delta_0)}{\sqrt{\text{var}\,(\log(1-\hat{\delta}))}} \tag{24}$$

also has asymptotically the standard normal distribution. In both cases the variances must be estimated using the estimates $\hat{\delta}$, $\hat{d}$ and $\hat{p}$, which will have little consequence for large samples.

Both variates, $\hat{\delta}$ and $\log(1 - \hat{\delta})$, approach normality as sample sizes get larger. When applying the normal approximation (where n_1 and n_2 are finite), the decision as to which of these two statistics, (23) or (24), should be used is greatly influenced by symmetry considerations. We usually prefer the statistic that has a more symmetric distribution. The Appendix presents an expression for the skewness of both test statistics. These measures of asymmetry are functions of the parameters d and p, as well as of the sample sizes n_1 and n_2.

From the two skewness coefficients it can be seen that for large values of n_1, $\log(1 - \hat{\delta})$ is more symmetric than $\hat{\delta}$, whereas the reverse is true for large values of n_2. More precisely, there exists a value R, which is a function of d and p, such that if $n_1/n_2 > R$, $\log(1 - \hat{\delta})$ is (asymptotically) more symmetric than $\hat{\delta}$, and if $n_1/n_2 < R$, $\hat{\delta}$ is more symmetric than $\log(1 - \hat{\delta})$. The precise form of R is given in the Appendix. In Table 1 an estimated value of R is given for each example, and for most of the cases considered, the dis-

tribution of $\hat{\delta}$ is (asymptotically) more symmetric than that of log $(1 - \hat{\delta})$.

Discussion

It has been demonstrated that the δ measure is a strictly increasing function of the normalized linkage disequilibrium D', $0 \leq D' \leq 1$, for all intermediate models ($f_2 \geq f_1 > f_0$) of disease predisposition, and, if the "disease" allele frequency is not too large, also for all overdominant models ($f_1 > f_2 > f_0$). This implies that δ can be used in a wide variety of cases to determine which of a number of antigens associated with a particular disease has the strongest genetic association with the "disease" allele (Bengtsson & Thomson 1982). The antigen with the highest δ value is expected to be the one most closely linked to the "disease" locus, since, other things being equal, we expect D' to decrease more rapidly with each generation the greater the recombination fraction between the antigen and "disease" loci (see, for example, Thomson 1977, Hedrick 1980).

Since the δ measure is essentially the attributable risk used in epidemiology (Bengtsson & Thomson, 1982, Thomson & Motro 1983), it can be viewed as a population parameter of association between a risk factor and a particular disease. The attributable risk is an increasing function of both the relative risk and the frequency of the risk factor in the population. It is larger for a factor with a larger relative risk, but it is also larger the more common a certain risk factor is in the population. Hence, it can be considered as a measure of the "load" the risk factor gives to the population with respect to the disease.

The δ measure can also be considered as a measure of the load contributed by the antigen to the population via its association with the "disease" allele. However, the interpretation that can be given regarding δ values of antigens at different loci and their respective recombination distances from a "disease" locus has, of course, no conceptual equivalent for the attributable risk function.

It should be borne in mind that different populations may not have equal δ values for a given antigen, since they will represent systems that have undergone different amounts of migration, selection and other such forces, all of which may alter the D' value. However, within any given population, the comparison of different δ values for a given disease is valid, since in this context it provides us with a measure of the relative strengths of the genetic associations of the antigens with the disease.

There is a problem, however, in applying statistical tests to compare two δ values, that is if there are two antigens associated with the same disease and we want to test which has a stronger association. The methodology as outlined above is strictly applicable only if the two "diseased" samples are independent (and the same for the control samples). However, in most practical instances, the same diseased individuals are used to estimate d_1 and d_2.

Appendix

Define $\lambda_1 = d/(n_1 (1 - d))$ and $\lambda_2 = p/(n_2 (1 - p))$. Employing the results of Motro et al. (1983), the asymptotic skewness ($\gamma_1 = \mu_3/\sigma^3$) of $\hat{\delta}$ is

$$- \frac{\left(\lambda_1 + 5\lambda_2\right)\left(\lambda_1 + \lambda_2\right) + \frac{\lambda_2}{n_2} - \frac{\lambda_1}{n_1}}{(\lambda_1 + \lambda_2)^{\frac{3}{2}}} \qquad \text{(A.1)}$$

and the asymptotic skewness of log $(1 - \hat{\delta})$ is

$$\frac{2\left(\lambda_2 - \lambda_1\right)\left(\lambda_1 + \lambda_2\right) + \frac{\lambda_2}{n_2} - \frac{\lambda_1}{n_1}}{(\lambda_1 + \lambda_2)^{\frac{3}{2}}} \qquad \text{(A.2)}$$

If $n_1/n_2 > R$, $\log(1-\hat{\delta})$ is more symmetric than $\hat{\delta}$, and if $n_1/n_2 < R$, $\hat{\delta}$ is more symmetric than $\log(1-\hat{\delta})$, where

$$R = \frac{-3r_1 + \sqrt{16r_1^2 + 14r_1 + \frac{2r_1^2}{r_2} + \frac{4r_1}{r_2}}}{7r_2 + 2} \qquad (A.3)$$

and $r_1 = d/(1-d)$, $r_2 = p/(1-p)$. For the special case $n_1 = n_2$, $\log(1-\hat{\delta})$ is more symmetric if

$$7r_2^2 - r_1^2 + 6r_1r_2 + 2r_2 - 2r_1 > 0, \qquad (A.4)$$

and $\hat{\delta}$ is more symmetric if the inequality sign is reversed.

Acknowledgments

We would like to thank Warren Ewens for his comments and contributions to this work.

References

Bengtsson, B. O. & Thomson, G. (1982) Measuring the strength of associations between HLA antigens and diseases. *Tissue Antigens* **18**, 356–363.

Hedrick, P. W. (1980) Hitchhiking: a comparison of linkage and partial selfing. *Genetics* **94**, 791–808.

Leung, H. M. & Kupper, L. L. (1981) Comparisons of confidence intervals for attributable risk. *Biometrics* **37**, 293–302.

Levin, M. L. (1953) The occurrence of lung cancer in man. *Acta Unionis Int Contra Cancrum* **19**, 531–541.

Lewontin, R. C. (1974) The Genetic Basis of Evolutionary Change. Columbia University Press, New York.

MacMahon, B. & Pugh, T. F. (1970) Epidemiology: Principles and Methods. Little, Brown and Co., Boston.

Miettinen, O. S. (1974) Proportion of disease caused or prevented by a given exposure, trait or invention. *Am J Epidemiol* **99**, 325–332.

Motro, U., Thomson, G. & Selvin, S. (1983) A note on the skewness of a ratio of two binomial variates. (Manuscript in preparation).

Noether, G. E. (1957) Two confidence intervals for the ratio of two probabilities and some measures of effectiveness. *J Am Stat Assoc* **52**, 36–45.

Sheps, M. C. (1959) An examination of some methods of comparing several rates or proportions. *Biometrics* **15**, 87–97.

Spielman, R. S., Baker, L. & Zmijewski, C. (1980) Gene dosage and susceptibility to insulin dependent diabetes. *Ann Hum Genet* **44**, 135–150.

Suarez, B. K. (1978) The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. *Tissue Antigens* **12**, 87–93.

Svejgaard, A., Platz, P. & Ryder, L. P. (1980) Insulin-dependent diabetes mellitus. In *Histocompatibility Testing 1980,* ed., Terasaki, P. I., pp. 638–656. UCLA Tissue Typing Laboratory, Los Angeles.

Thomson, G. (1977) The effect of a selected locus on linked neutral loci. *Genetics* **85**, 753–788.

Thomson, G. (1981) A review of theoretical aspects of HLA and disease associations. *Theor Popul Biol* **20**, 168–208.

Thomson, G. & Motro, U. (1983) The population attributable risk for non-independent risk factors. *Biometrics* (Submitted for publication).

Walter, S. D. (1976) The estimation and interpretation of attributable risk in health research. *Biometrics* **32**, 829–849.

Address:
Glenys Thomson
Genetics Department
Mulford Hall
University of California
Berkeley, CA 94720
USA