Conditional Asymmetric Linkage Disequilibrium (ALD): Extending the Biallelic r^2 Measure

Glenys Thomson* and Richard M. Single†,1

*Department of Integrative Biology, University of California, Berkeley, California 94720 and †Department of Mathematics and Statistics, University of Vermont, Burlington, Vermont 05405

ORCID ID: 0000-0001-6054-6505 (R.M.S.)

ABSTRACT For multiallelic loci, standard measures of linkage disequilibrium provide an incomplete description of the correlation of variation at two loci, especially when there are *different* numbers of alleles at the two loci. We have developed a complementary pair of *conditional asymmetric* linkage disequilibrium (ALD) measures. Since these measures do not assume symmetry, they more accurately describe the correlation between two loci and can identify heterogeneity in genetic variation not captured by other symmetric measures. For biallelic loci the ALD are symmetric and equivalent to the correlation coefficient *r*. The ALD measures are particularly relevant for disease-association studies to identify cases in which an analysis can be stratified by one of more loci. A stratified analysis can aid in detecting primary disease-predisposing genes and additional disease genes in a genetic region. The ALD measures are also informative for detecting selection acting independently on loci in high linkage disequilibrium or on specific amino acids within genes. For SNP data, the ALD statistics provide a measure of linkage disequilibrium on the same scale for comparisons among SNPs, among SNPs and more polymorphic loci, among haplotype blocks of SNPs, and for fine mapping of disease genes. The ALD measures, combined with haplotype-specific homozygosity, will be increasingly useful as next-generation sequencing methods identify additional allelic variation throughout the genome.

THE definition of the linkage disequilibrium (LD) parameter D_{ij} of nonrandom association between a pair of alleles A_i and B_j at two loci (*A* and *B*) is straightforward and unequivocal. It is the difference between the observed (or estimated) haplotype (chromosomal or gametic) frequency (f_{ij}) and that expected under random association of the two allele frequencies (p_{A_i} and p_{B_j}) : $D_{ij} = f_{ij} - p_{A_i}p_{B_j}$. While this is the base of all other measures of LD, defining the *strength* of any observed nonrandom association is complicated by the fact that the maximum value D_{ij} can take is a function of the observed allele frequencies. A number of normalized measures to reflect the strength of LD have been proposed; both for *bi-* and *multi*allelic data (Hedrick 1987; Lewontin 1988). However, since these are all a single summary of multidimensional data, no proposed measure of the strength of LD can be perfect; although each may have strengths and weaknesses with respect to the question being addressed.

The two most common measures of the strength of LD are: (1) the normalized measure of the individual LD values (Lewontin 1964), $D_{ij}' = D_{ij}/D_{\max}$ (see Supporting Information, File S1 for details) and (2) the correlation coefficient *r* for *biallelic* data, which is most often reported as $r^2 = {D_{ij}}^2 / (p_{A1}\, p_{A2}\, p_{B1}\, p_{B2})$. Hedrick (1987) extended the D' measure for multiallelic data as a weighted average over all alleles at each locus of the individual normalized LD values: $D' = \Sigma_i \Sigma_j\, p_{A_i}\, p_{B_j} |D_{ij}'|$. The multiallelic extension of the r^2 measure is

$$W_n^2 = \left[\sum_i \sum_j D_{ij}^2 \Big/ \left(p_{A_i} p_{B_j}\right)\right] \Big/ \min(k_A - 1,\ k_B - 1),$$

where k_A and k_B indicate the number of alleles at each locus. It is also known as Cramer's *V* statistic (Cramer 1946), defined on the contingency table relating two categorical variables and is a reexpression of the χ^2 statistic, normalized to be between zero and one (Hill 1975; Hedrick 1987; Single *et al.* 2007, 2011). With *N* individuals ($2N$ alleles/haplotypes), $(2N)(W_n^2)\min(k_A - 1, k_B - 1)$ has a χ^2 distribution

doi: 10.1534/genetics.114.165266
Manuscript received April 15, 2014; accepted for publication July 8, 2014; published Early Online July 14, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165266/-/DC1.
[1]Corresponding author: Department of Mathematics and Statistics, University of Vermont, 16 Colchester Ave., Burlington, VT 05405. E-mail: richard.single@uvm.edu

Table 1 Linkage disequilibrium and genetic diversity measures

Description	Definition of measures[a]
1. Single locus homozygosity (F) and heterozygosity (H)[b]	$F_A = \sum_i p_{Ai}^2$, $H_A = 1 - F_A$
2. Haplotype-specific homozygosity (HSF)[c]	$F_{A/Bj} = \sum_i (f_{ij} / p_{Bj})^2$, $F_{B/Ai} = \sum_j (f_{ij} / p_{Ai})^2$
3. Overall weighted HSF values $F_{A/B}$ and $F_{B/A}$	$F_{A/B} = \sum_j (F_{A/Bj})\,(p_{Bj})$, $F_{B/A} = \sum_i (F_{B/Ai})\,(p_{Ai})$
4. Multiallelic ALD squared (overall asymmetric LD squared)	$W_{A/B}^2 = (F_{A/B} - F_A) / (1 - F_A) = [\sum_i \sum_j (D_{ij}^2 / p_{Bj})] / (1 - F_A)$ $W_{B/A}^2 = (F_{B/A} - F_B) / (1 - F_B) = [\sum_i \sum_j (D_{ij}^2 / p_{Ai})] / (1 - F_B)$

[a] In all cases, $\sum_i$ indicates summation over all $i = 1, 2, \ldots, k_A$, and similarly $\sum_j$ over all $j = 1, 2, \ldots, k_B$, where k_A and k_B are the number of alleles at the A and B loci, respectively, with $\sum_i p_{Ai} = 1$, and $\sum_j p_{Bj} = 1$. $\sum_i \sum_j D_{ij} = 0$, $\sum_i D_{ij} = 0$, $\sum_j D_{ij} = 0$, $\sum_i \sum_j f_{ij} = 1$, $\sum_j f_{ij} = p_{Ai}$, $\sum_i f_{ij} = p_{Bj}$. For biallelic data: $D_{11} = -D_{12} = -D_{21} = D_{22} = D$. Alternate expressions are given in the *Appendix* along with biallelic results.

[b] The values of F_A and H_A ($F_A + H_A = 1$) are those expected under Hardy–Weinberg proportions.

[c] The $HSF_{A/Bj}$ values are the extension of the single-locus F_A values but now restricted to haplotypes containing the allele B_j (similarly for $HSF_{B/Ai}$). Haplotype-specific heterozygosity values are $HSH_{A/Bj} = 1 - HSF_{A/Bj}$, and $HSH_{B/Ai} = 1 - HSF_{B/Ai}$.

with $(k_A - 1)(k_B - 1)$ degrees of freedom and can be used to test for significant LD between two loci.

For *biallelic* data, $D' = 1$ whenever one or more of the four possible haplotypes are *not* observed, irrespective of the expected frequencies. In contrast, r directly measures the correlation coefficient of the biallelic variation at two loci. Specifically, $r = 1$ only when the allelic variations at the two loci show 100% correlation, *i.e.*, when both loci have equal allele frequencies and only two complementary haplotypes are observed. This correlation property is of interest to many research questions. For example, if two loci show associations with a disease but r is close or equal to one (*i.e.*, nearly complete allelic association), then there is little or no variation that can be assessed by a stratified analysis for risk heterogeneity between two potentially disease-predisposing genetic variants.

Due to these inherent differences between the properties of the D' measure and the correlation measure r, we focus on the correlation measure and its multiallelic extension W_n. We developed the pair of *conditional asymmetric* LD (ALD) measures, $W_{A/B}$ and $W_{B/A}$, to complement the W_n measure especially when there are *different* numbers of alleles at the two loci. This leads to cases where W_n is equal or close to one while one of the two ALD measures is substantially less than one.

Other conditional LD measures have been proposed (Nei and Li 1980; Chakravarti *et al.* 1984; Hudson 1985; Guo 1997). Nei and Li (1980) developed a statistic that quantifies the association between alleles at a marker locus and a disease locus for studies where individuals are not randomly sampled from a single population, but sampling intensity varies within (disease) categories (Kaplan and Weir 1992; Maiste and Weir 1992). See File S1 for additional detail. In contrast to the above, the ALD measures introduced below are defined for a randomly ascertained sample from a demographically defined population or control group.

When there are *different numbers of alleles* at the two loci, the direct correlation property discussed above for the r measure *is not* retained by its multiallelic extension W_n. Consider *example 1* with *two* and *three* alleles at the first and second loci, with $f_{11} = 0.3$, $f_{22} = 0.5$, $f_{23} = 0.2$ for A_1B_1, A_2B_2, and A_2B_3 haplotypes. $W_n = 1$; however, there is variation at the B locus on haplotypes containing the A_2 allele. Thus, there is not 100% correlation, and there *never* can be with differing numbers of alleles at the two loci. In this example the two ALD measures (defined below) reflect that while there is no variation of A locus alleles on any of the haplotypes conditioned on the B locus alleles ($W_{A/B} = 1$), there is variation in the B_2 and B_3 alleles on haplotypes carrying A_2 ($W_{B/A} = 0.73$). The ALD measures directly indicate that with appropriate sample size, stratification analyses could be carried out for certain comparisons. In contrast, a naive interpretation of the fact that $W_n = 1$ could result in passing over these data for conditional or stratified haplotype analyses of risk heterogeneity (Thomson *et al.* 2008).

The definition of the ALD measures begins with the homozygosity (F) and heterozygosity (H) values expected under Hardy–Weinberg proportions (HWP) at a single locus (see Table 1). While there are other measures of association and LD that are based on allelic diversity statistics (see File S1 for details), these measures are all symmetric (Ohta 1980; Maruyama 1982; Hedrick and Thomson 1986; Hedrick 1987). The composite LD measure of Wu *et al.* (2008) is designed to test interaction between two unlinked loci.

The conditional two-locus extensions of F and H, called haplotype-specific homozygosity (*HSF*) and haplotype-specific heterozygosity (*HSH*), measure the level of genetic variation at locus A on haplotypes with a specific allele on the B locus (and vice versa), *i.e.*, $F_{A/Bj}$, and $F_{B/Ai}$ (see Table 1). We developed the *HSF* and *HSH* measures (Malkki *et al.* 2005) to ascertain informative microsatellites (MSATs) in HLA transplantation and disease studies. The complementary pair of conditional ALD measures are defined by normalizing an extension of the HSF measure across all haplotypes.

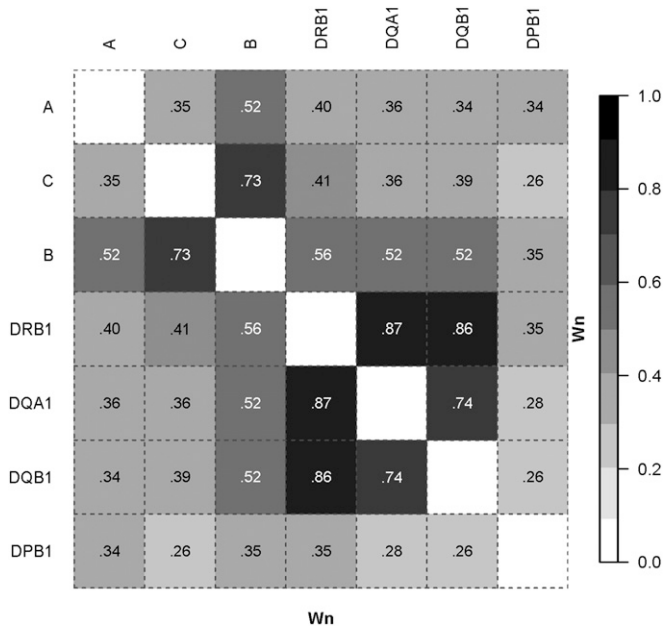

Figure 1 W_n Measure for classical HLA genes. LD plot based on the W_n measure (the multiallelic extension of the r correlation measure). Data source: ImmPort Study#SDY26 with N = 300 controls typed for HLA (Wilson 2010). The number of alleles (k) are as follows (given in parentheses after each locus): *A* (33), *C* (29), *B* (61), *DRB1* (40), *DQA1* (9), *DQB1* (14), *DPB1* (24).

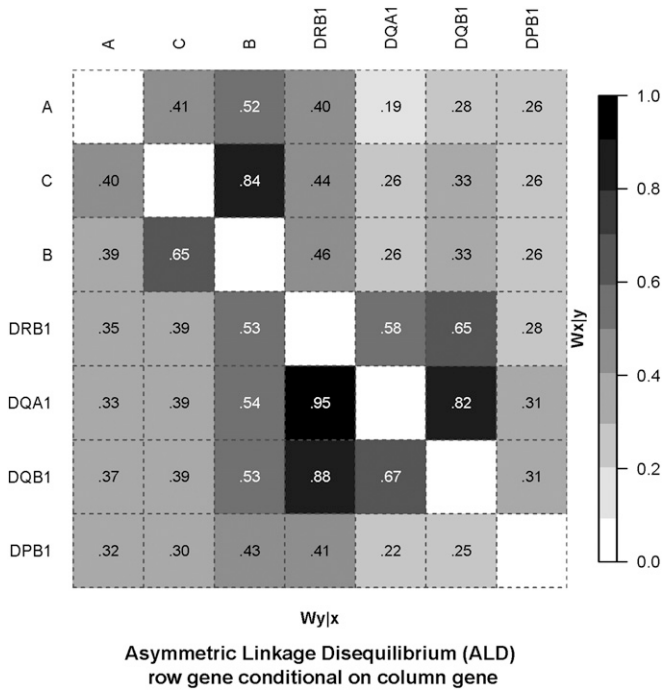

Figure 2 ALD measures for classical HLA genes. LD plot based on the ALD measures. Our usual generic notation for the ALD measures of $W_{A/B}$ and $W_{B/A}$ is replaced by $W_{x/y}$ and $W_{y/x}$ to avoid confusion with the HLA-*A* and -*B* loci used in this example. Data source as for Figure 1.

Materials and Methods

Definition of the asymmetric LD measures

There are two conditional ALD measures, depending on which locus is conditioned upon. For simplicity, we often describe the measure in detail conditioning on the *B* locus. The derivation of the complementary measure, conditioning on the *A* locus, is given by swapping the roles of loci *A* and *B*.

The *individual HSF* values (Table 1) are combined as a weighted average over all alleles at the conditioned locus to obtain the two *overall* haplotype specific homozygosity measures: $F_{A/B}$ and $F_{B/A}$ (Table 1 and see *Appendix* for alternate expressions). The maximum value $F_{A/B}$ can take is 1.0, when each *A* allele occurs with only one *B* allele.

${W_{A/B}}^2$ (the square of the ALD measure) is obtained by normalizing the *overall weighted HSF* value based on the range of possible values that it can achieve (Table 1):

$$W_{A/B}^2 = \left(F_{A/B} - F_A\right) \Big/ (1 - F_A)$$
$$= \left[\sum_i \sum_j \left(D_{ij}^2 \Big/ p_{Bj}\right)\right] \Big/ (1 - F_A).$$

For biallelic data at both loci ${W_{A/B}}^2 = {W_{B/A}}^2 = D^2/(p_{A_1}p_{A_2}p_{B_1}p_{B_2}) = r^2$ (see *Appendix*).

Once we deviate from having two alleles at both loci, the two ALD measures are only equal in certain specific cases (see below). For biallelic data the correlation coefficient is given by r; for multiallelic data W_n and the ALD measures, $W_{A/B}$ and $W_{B/A}$, give the appropriate correlation coefficients.

Other factors being equal, the ALD increases with stronger LD between the two loci. The ALD values are also influenced by the number of alleles at each locus. Specifically, for multiallelic loci with *unequal* numbers of alleles, *e.g.*, $k_A < k_B$ (with $k_A \geq 2$), in the extreme case each B_j allele will occur with only one A_i allele and $W_{A/B} = 1$ (indicating no variation at the A locus on any haplotype containing a specific B_j allele) and also $W_n = 1$ (mirroring this effect). However, $W_{B/A} < 1$ reflects the required variation, given the inequality of allele numbers, at the B locus on some or all haplotypes containing a specific A_i allele (see special case e, below).

Special cases

a. *Biallelic* loci with *two* haplotypes of the four possible, *e.g.*, A_1B_1 and A_2B_2, (hence $p_{A_1} = p_{B_1}$ and $p_{A_2} = p_{B_2}$). LD is maximal with $D = p_{A_1}p_{B_1}$, and there is symmetry in all measures: $D' = 1$ and $r = W_n = W_{A/B} = W_{B/A} = 1$.

b. *Biallelic* loci with *three* haplotypes of the four possible, *e.g.*, A_1B_1, A_1B_2, and A_2B_2. With the following allele frequencies $p_{A_2} = p_{B_2}, p_{A_1} < p_{B_1}$, and $p_{A_1} < p_{B_2}$, LD is maximal ($D = p_{A_2}p_{B_1}$): $D' = 1$, but r ($= W_n = W_{A/B} = W_{B/A}$) < 1. This reflects that the allele frequencies at the two loci are not 100% correlated.

c. Multiallelic loci with *equal* number of alleles (*i.e.*, $k_A = k_B = k$) and only symmetric haplotypes (*i.e.*, $f_{ii} > 0$, for all $i = 1, 2, \ldots, k$, and $f_{ij} = 0$ otherwise). As above for the biallelic case a, there is complete symmetry and 100% correlation of allele frequencies at the two loci: $D' = 1$,

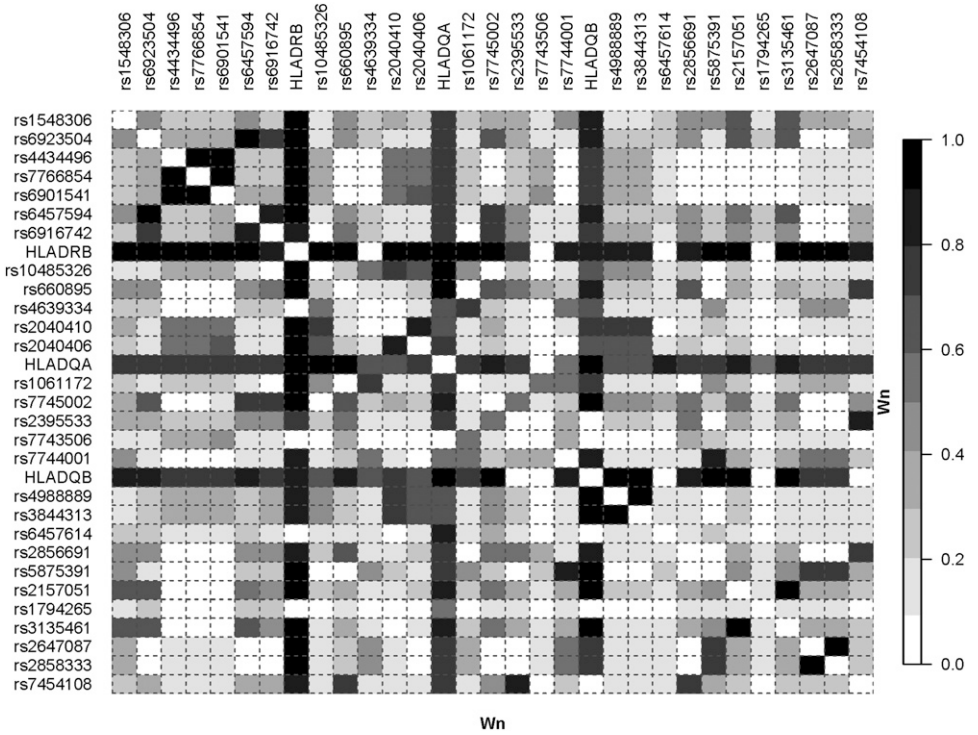

Figure 3 The W_n measure applied to SNP and HLA–*DRB1*, *DQA1*, and *DQB1* data. LD plot based on the W_n measure. The data are for 90 unrelated individuals from de Bakker *et al.* (2006) with European ancestry from the CEU obtained from the Tagger/MHC webpage.

and $W_n = W_{A/B} = W_{B/A} = 1$. An example with three alleles at both loci is $f_{11} = 0.5$, $f_{22} = 0.3$, $f_{33} = 0.2$, with all other $f_{ij} = 0$. There is no variation of A locus alleles on any of the haplotypes conditioned on the B locus alleles, and vice versa.

d. The same as c above, except that one or more of $f_{ij} > 0$ for $i \neq j$: $W_n < 1$, $W_{A/B} < 1$, $W_{B/A} < 1$.

e. Multiallelic loci with *unequal* number of alleles (*e.g.*, $k_A < k_B$), with each B_j allele occurring with only one A_i allele (see example 1 in the Introduction). While $W_n = W_{A/B} = 1$, $W_{B/A} < 1$.

f. One locus biallelic and the other multiallelic (*e.g.*, $k_A = 2$, $k_B > 2$): $W_n = W_{A/B} \neq W_{B/A}$. In a variety of cases examined, $W_{B/A} < W_{A/B}$, but we have no proof that this is always the case.

See File S1 for proofs of special cases c–f.

Results

HLA classical loci

We applied the ALD measures to data for the polymorphic HLA classical genes (Wilson 2010): class I (*A*, *C*, and *B*) and class II *(DRB1*, *DQA1*, *DQB1*, and *DPB1*). Figure 1 and Figure 2 respectively show the standard overall LD measure W_n and the ALD measures $W_{A/B}$ and $W_{B/A}$. The W_n measure assumes/forces symmetry (as does the overall D' measure, not shown) even though with more than two alleles per locus, differing numbers of alleles at each locus, and different levels of LD between loci this is not the case.

The ALD values show considerable heterogeneity. For example (with numbers of alleles for each locus given in parentheses), the *ALD* for *DRB1*(40) conditioning on *DQA1* (9) is 0.58 = $W_{DRB1/DQA1}$; *i.e.*, the overall variation for *DRB1* is relatively high given specific *DQA1* alleles. In contrast, the *ALD* for *DQA1* conditioning on *DRB1* is 0.95 = $W_{DQA1/DRB1}$; *i.e.*, the overall variation for *DQA1* is relatively low given specific *DRB1* alleles. This reflects both the smaller number of alleles at *DQA1* compared to *DRB1* and the high LD between the two loci (most *DRB1* alleles occur with only one *DQA1* allele, but not vice versa). Similarly with the *B*(61) and *C*(29) loci, $W_{B/C} = 0.65$, and $W_{C/B} = 0.84$. In both these examples the standard (symmetric) overall pairwise LD values are intermediate to the ALD values: $W_n = 0.87$ and 0.73 for the *DRB1*:*DQA1* and *C*:*B* locus pairs, respectively. In almost all comparisons, if the number of alleles $k_X > k_Y$ then $W_{X/Y} < W_{Y/X}$. An exception is with the *A*(33) and *C*(29) loci, *i.e.*, $k_A > k_C$, but $W_{A/C}$ (0.41) > $W_{C/A}$ (0.40).

SNP and HLA data

HLA and SNP data from de Bakker *et al.* (2006) characterized patterns of LD among highly polymorphic HLA genes and a large number of SNP sites. The extensive LD across the extended HLA region (~8 Mb) makes the identification of additional non-HLA genomic effects on disease difficult to assess. The SNP sites used here were selected on the basis of their ability to identify or tag specific alleles at each of the HLA classical loci (*i.e.*, tag-SNPs for HLA alleles). We chose this example, with a subset of the HLA and SNP data in the class II region, to highlight the properties of the ALD measures and what distinguishes them from the symmetric *r* and W_n measures.

Figure 3 and Figure 4 show plots of the W_n and ALD measures for 90 unrelated individuals with European ancestry

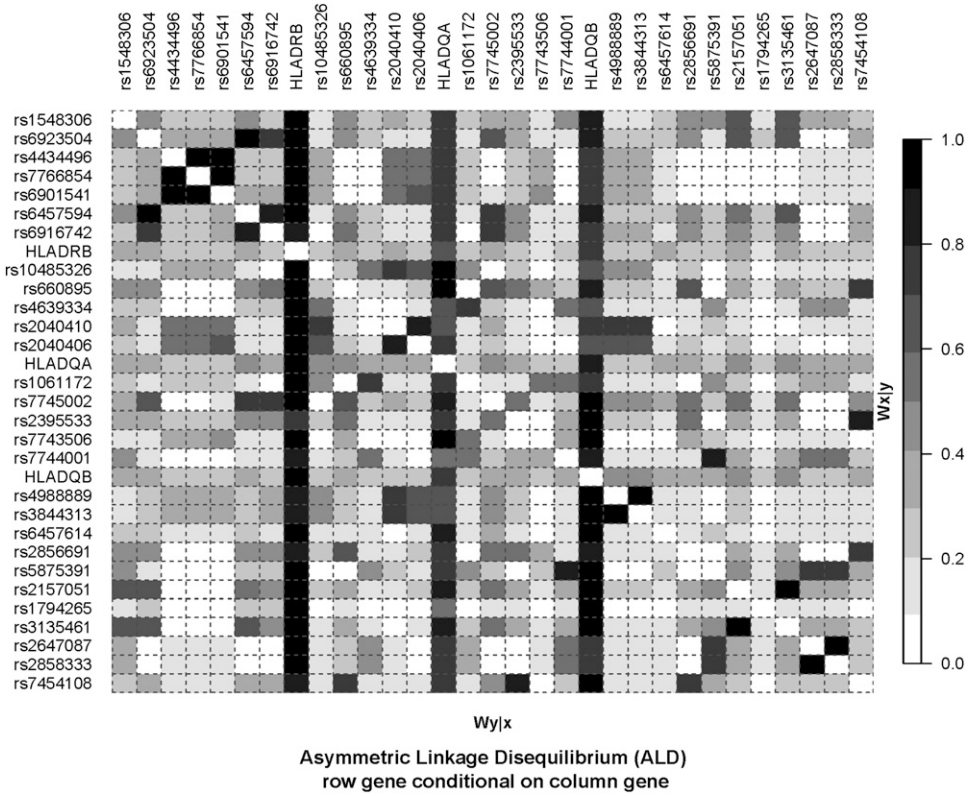

Figure 4 The ALD measures applied to SNP and HLA–*DRB1*, *DQA1* ,and *DQB1* data. LD plot based on the ALD measures. Data source as for Figure 3.

from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (CEU) obtained from the Tagger/MHC webpage. The ALD measures (Figure 4) provide a visualization of the tag-SNP properties that is not captured by the symmetric W_n measure. Looking down the column for any one of the HLA loci (*i.e.*, conditioning on an HLA locus), one can see the particular SNPs that tag specific HLA alleles. These show up as a dark column in the figure. However, conditioning on any given SNP does not show this pattern of high LD. In contrast with the figure for W_n, there are no dark *rows* of high LD for the ALD measures, indicating that the ALD measures capture the different degree of overall association for each individual SNP.

Note that the information displayed in Figure 3 and Figure 4 captures different aspects of LD from the results reported in the de Bakker *et al.* (2006) article, as we present *overall* LD between each pair of loci. The r^2 values reported in their article represent the squared correlation between a given SNP and presence/absence of each particular HLA allele (*e.g.*, A*0101 *vs.* other). The tag-SNPs were chosen such that this r^2 value is 1.0 (or nearly so) for a specific HLA allele, not for the overall locus. The values in Figure 3 and Figure 4 represent *overall* LD combining over all alleles at both loci.

For example, the SNP rs4988889 is listed as a tag-SNP in the CEU population for the HLA-DQB1*02:01 *allele* in Table S3 of de Bakker *et al.* (2006), with an r^2 (symmetric) value of 0.958. It does not show up as a tag-SNP for any other HLA allele in their Table S3. In Table 2 below, one can see that the values for $W_{\mathrm{HLA}|\mathrm{SNP}}$ and $W_{\mathrm{SNP}|\mathrm{HLA}}$ are quite different (0.4083 *vs.* 0.9788). The rs7743506 SNP is listed in de Bakker *et al.* (2006) as a tag-SNP for three class II alleles, each with an r^2 value of 1.0: HLA-*DQA1**04:01, HLA-*DQB1**04:02, and HLA-*DRB1**08:01. Thus, allele 2 for this SNP is completely correlated with the presence of each of these three class II HLA alleles. This 100% correlation is captured by the ALD measure ($W_{\mathrm{SNP}|\mathrm{HLA}} = 1.0$), while the low values for $W_{\mathrm{HLA}|\mathrm{SNP}}$ for each of the three class II loci indicates that there is a large amount of variability remaining at the HLA loci after conditioning on this SNP. Note that for the examples in Table 2, $W_{\mathrm{SNP}|\mathrm{HLA}}$ is equal to W_n. This is an example of special case f above.

HLA disease association data

The HLA class II *DRB1* gene is strongly associated with juvenile idiopathic arthritis (oligoarticular-persistent) (JIA-OP), with a hierarchy of predisposing through intermediate ("neutral") to protective effects (Hollenbach *et al.* 2010; Thomson *et al.* 2010). Amino-acid position 13 (AA13) of *DRB1* shows the strongest single AA association with JIA-OP. This association is also stronger than other potentially biologically relevant combinations of AAs defined under sequence feature variant-type (SFVT) analysis (Karp *et al.* 2010; Thomson *et al.* 2010). AA13 is also identified as potentially causative in disease using an extension of Salamon's unique combinations algorithm (Salamon *et al.* 1996; Thomson *et al.* 2010). The overall AA LD (W_n) patterns are quite complex for each of the classical HLA loci, with *DRB1* control data for

Table 2 Overall LD measures applied to data from de Bakker *et al.* (2006)

$W_{HLAISNP}$	$W_{SNPIHLA}$	D'	W_n	Locus 1[a]	Locus 2
0.2611	0.8214	0.9150	0.8214	*DRB1*	rs4988889
0.2824	0.6256	0.8255	0.6256	*DQA1*	rs4988889
0.4083	0.9788	1.0000	0.9788	*DQB1*	rs4988889
0.1980	1.0000	1.0000	1.0000	*DRB1*	rs7743506
0.2164	1.0000	1.0000	1.0000	*DQA1*	rs7743506
0.2056	1.0000	1.0000	1.0000	*DQB1*	rs7743506

[a] The loci listed under locus 1 are the three classical class II loci HLA–*DRB1*, –*DQA1*, and –*DQB1*

JIA-OP shown in Figure 5. AA13 shows high LD via the W_n measure with quite a few other AAs (note only AAs 9–38 within exon 2 are shown). However, ALD analyses show additional variation that can be tested via conditional analyses (Figure 6).

For illustration, we consider the block of high LD AAs 11(6), 12(2), and 13(6) (the number of "alleles," or different AA residues segregating, at each AA site are given in parentheses). AA 10(2) and AA 12 are 100% correlated apart from a very rare allele, and hence AA 10 is not considered here. The ALD values indicate which pairs of AAs may allow for stratification and conditional analyses. For example (see Figure 5 and Figure 6), with AAs 11 and 12, $W_n = 1$, and while $W_{12/11} = 1$, $W_{11/12} = 0.64$, and hence some stratification analyses can be carried out (this is also an illustration of special case f above). Table 3 shows the results of specific tests of risk heterogeneity: variation at AA 13 is significantly associated with disease on haplotypes with AA 11 and AA 12. In contrast, AA 11 does not show heterogeneity on haplotypes with AA 13. This does not exclude a role for AA 11, nor AA 12, in disease predisposition, but the conditional analyses do show a potential role for AA 13 in being directly involved in disease risk.

Selection on HLA–DRB1 amino acids

A role for balancing selection maintaining much of the extensive variation at the HLA classical loci is well established (Meyer and Thomson 2001; Meyer *et al.* 2006). In particular, application of the Ewens–Watterson (EW) neutrality test of allele-frequency distributions at the classical HLA loci has revealed the action of balancing selection in maintaining diversity at the HLA-*A*, *-C*, *-B*, *DRB1*, *DQA1*, and

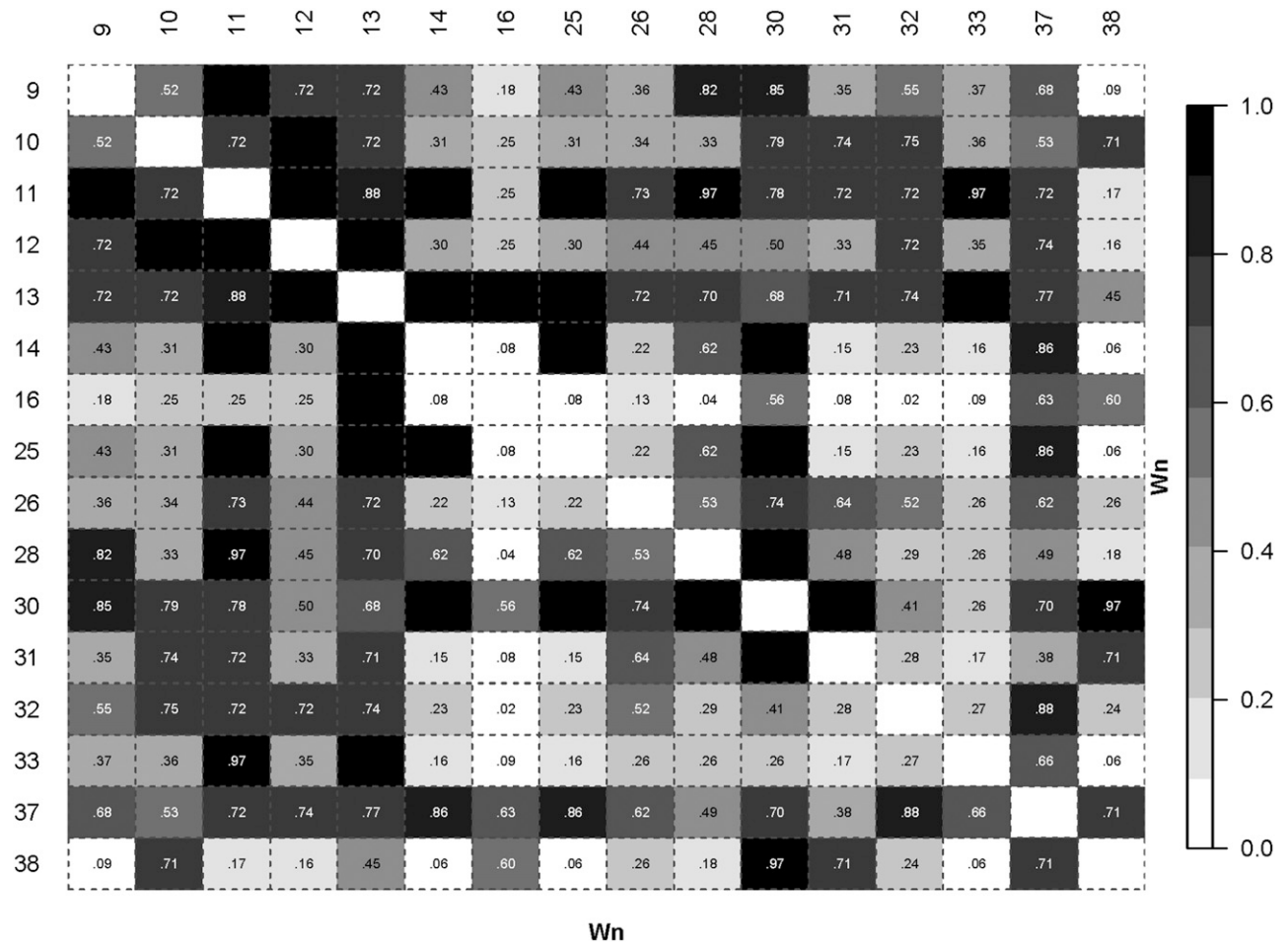

Figure 5 The W_n measure applied to HLA-*DRB1* AMINO Acids 9-38. LD plot based on the W_n measure. Black colored cells have a numeric value of 1. The data are for the control sample in a study of JIA.(Hollenbach *et al.* 2010; Thomson *et al.* 2010) For clarity, only amino acids 9-38 are shown.

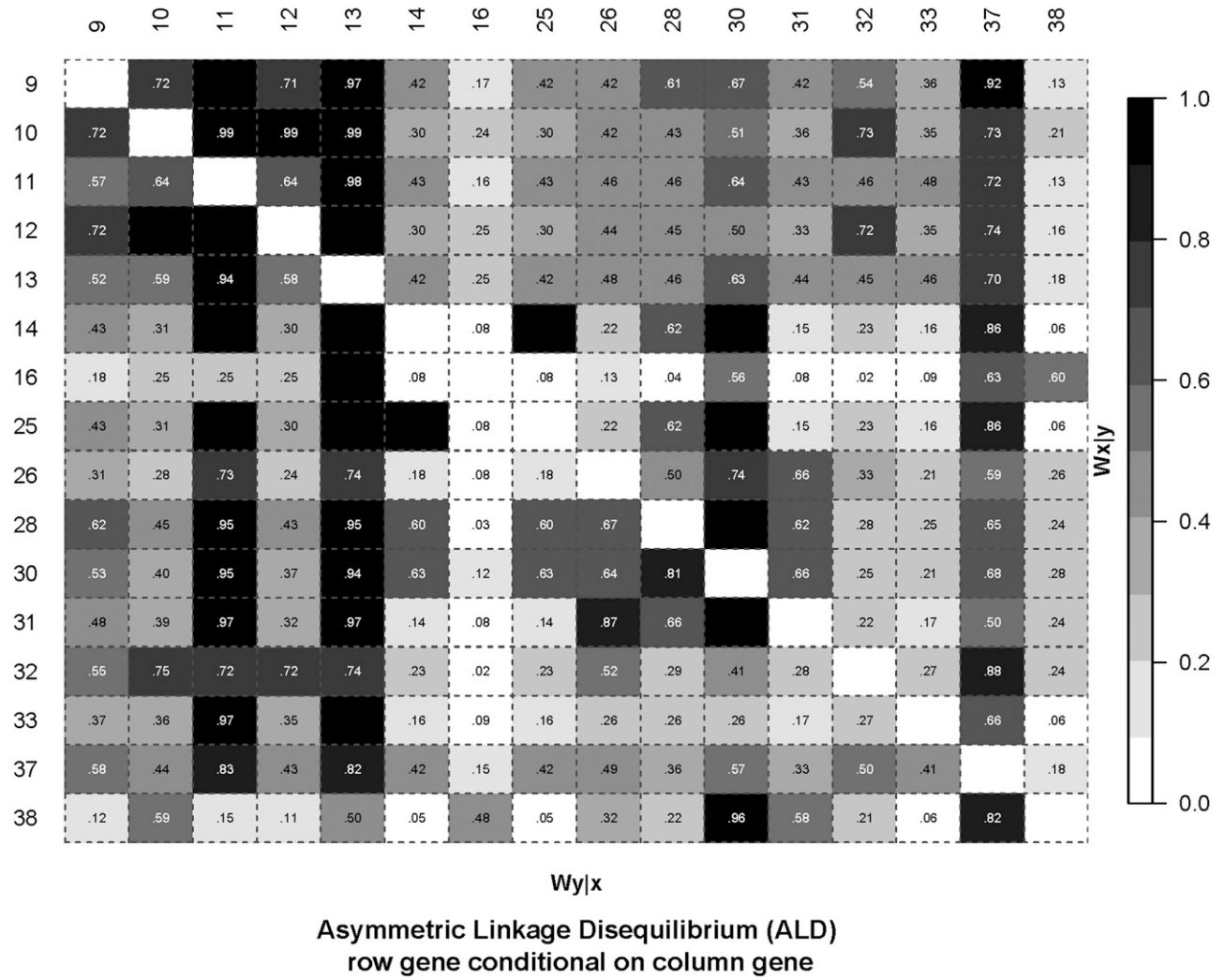

Figure 6 The *ALD* measures applied to HLA-*DRB1* AMINO Acids 9-38. LD plot based on the ALD measures. The data are as for Figure 5.

DQB1 loci (Salamon *et al.* 1999; Lancaster 2006; Solberg *et al.* 2008). Allele frequency distributions at these loci are generally more even than expected under neutral conditions. The distributions of *DPB1* alleles do not show evidence of balancing selection (Salamon *et al.* 1999; Begovich *et al.* 2001; Lancaster 2006; Tsai and Thomson 2007; Solberg *et al.* 2008). However, extension of the EW test to the AA level has shown evidence for balancing selection acting on some AAs for all the classical HLA loci, including *DPB1* (Salamon *et al.* 1999; Valdes *et al.* 1999; Lancaster 2006).

At both the allele and AA levels, the statistic used for the above analyses is the mean across populations of the normalized deviate F_{nd} of the homozygosity statistic F (Salamon *et al.* 1999). Balancing selection results in significantly negative F_{nd} values compared to neutral expectations, whereas directional selection, along with certain demographic events, leads to significant positive values. An observation of interest from previous studies is that pairs of AAs that show high LD may nonetheless show quite different F_{nd} values (Salamon *et al.* 1999; Lancaster 2006). To illustrate this point in the context of ALD measures applied to the JIA-OP *DRB1* control data, consider AA positions 37 and 38, which have a moderately high W_n value of 0.71 (Figure 5). However, the ALD values are quite disparate ($W_{37/38} = 0.18$ and $W_{38/37} = 0.82$) (Figure 6), and explain how the observed F_{nd} values can show different evolutionary histories with significant evidence for balancing selection for AA 37 and possible directional selection for AA 38 (Figure 7). This pattern is not unique to this particular population. Similar patterns of this differential selection can be seen in meta-analyses across several populations (see Figure S1 for F_{nd} values across 57 populations for DRB1 data (Lancaster 2006)). For these data, *P*-values for deviation from neutral expectations in the direction of balancing selection are $2.5E{-}24$ and 0.11 for AAs 37 and 38, respectively (Lancaster 2006).

Discussion

From analyses of allele and haplotype data in disease-association studies, HLA researchers have long recognized

Table 3 Conditional Analyses of JIA-OP Data for HLA *DRB1* Amino Acids 11-13[a]

11_13	Cases	Controls	χ^2	*P*-value	OR	
S-G	121	22	46.12	1.1E-11	4.89	*P*<3.6E-06
S-S	363	200	14.72	0.0001	1.81	
D-F	9	6	0.08	0.7821	1.15	ns
L-F	87	66	0.01	0.9198	1.01	
V-H	46	84	23.49	1.3E-06	0.38	
P-R	50	99	31.79	1.7E-08	0.34	
G-Y	30	65	23.92	1.0E-06	0.33	
Total	708	546	140.12	9.4E-28		

12_13	Cases	Controls	χ^2	*P*-value	OR	
T-G	121	22	46.12	1.1E-11	4.91	*P*<3.6E-06
T-S	363	200	14.72	0.0001	1.82	
K-F	98	76	0.002	0.9708	0.994	*P*<1.2E-05
K-H	46	84	23.49	1.3E-06	0.382	
K-R	50	99	31.79	1.7E-08	0.343	
K-Y	30	65	23.92	1.0E-06	0.327	
Total	708	546	140.04	1.8E-28		

[a] Standard ^{2} tests of heterogeneity (overall and individual contributions to the overall test) are carried out for the overall data (ranked by odds ratio (OR) values) for the amino acid haplotypes 11 and 13, and 12 and 13. Note that the test of each individual row contribution to the overall significance are based on a ^{2} with 1 df. These individual p-values are conservative due to the assumption of a single df ^{2}, however the p-values can be used for a relative ranking of the allelic effects.

that high pairwise LD (W_n) between two loci has limited our ability in some cases to distinguish the primary disease gene or genes. It is also well known that there are instances, particularly with differing numbers of alleles at two loci, where the W_n value does not accurately reflect our ability to perform stratified or conditional analyses to identify disease-risk heterogeneity. With multiallelic data, the ALD measures presented here are more appropriate and informative than the W_n measure. For example, with type 1 diabetes (T1D), *DRB1–DQB1* haplotypes carrying the *DRB1**04:01 allele can be subdivided by the *DQB1**03:02 (predisposing) and *03:01 (protective) alleles. This approach, termed for HLA studies "within serogroup comparisons" (based on a specific variant in the first field, or serotype, of the *DQB1* allele name, and comparing AA variation related to disease risk in the second field) focuses on a smaller number of AAs to compare. In this case the analysis of *DRB1–DQB1* haplotypes is stratified on *DQB1* based on the presence of *DRB1**04:01. This led to identification of AA 57 of *DQB1* in T1D risk. In fact, for T1D both *DRB1* and *DQB1* are directly involved in disease risk, with confirmation coming from cross-ethnic studies (Thomson *et al.* 2007, 2008, 2011; Erlich *et al.* 2008).

Another example of stratification on a particular site aiding in the identification of additional effects comes from a SNP in the *PTPN22* gene. In a study of rheumatoid arthritis, Begovich *et al.* (2004) demonstrated an association with the minor allele of the R620W missense SNP (*rs2476601*) in *PTPN22*. In a follow-up study, similar to the above HLA study on T1D, Carlton *et al.* (2005) used AA analyses of closely related haplotypes of SNPs to show a direct role of R620W in risk heterogeneity. With stratification of the data by R620W, the role in disease risk of at least one additional SNP in *PTPN22* was identified.

The ALD measures were initially developed to aid two separate lines of research for AA variation at classical HLA genes: to determine the actual disease-predisposing AAs in disease-association studies and to identify which AA sites are independently subject to selection in population studies. The major problem encountered in both research areas is the high level and complex patterns of LD between many AA sites, combined with more than two (and up to six) distinct AAs ("alleles"), seen at many sites. When evidence of strong balancing selection is seen at a number of AA sites (Salamon *et al.* 1999; Valdes *et al.* 1999; Lancaster 2006), how does one determine which AA sites could potentially show independent evolution *vs.* correlation due to high LD? Similarly with disease-association studies of individual AAs and biologically relevant sequence features (SFs) and their variant

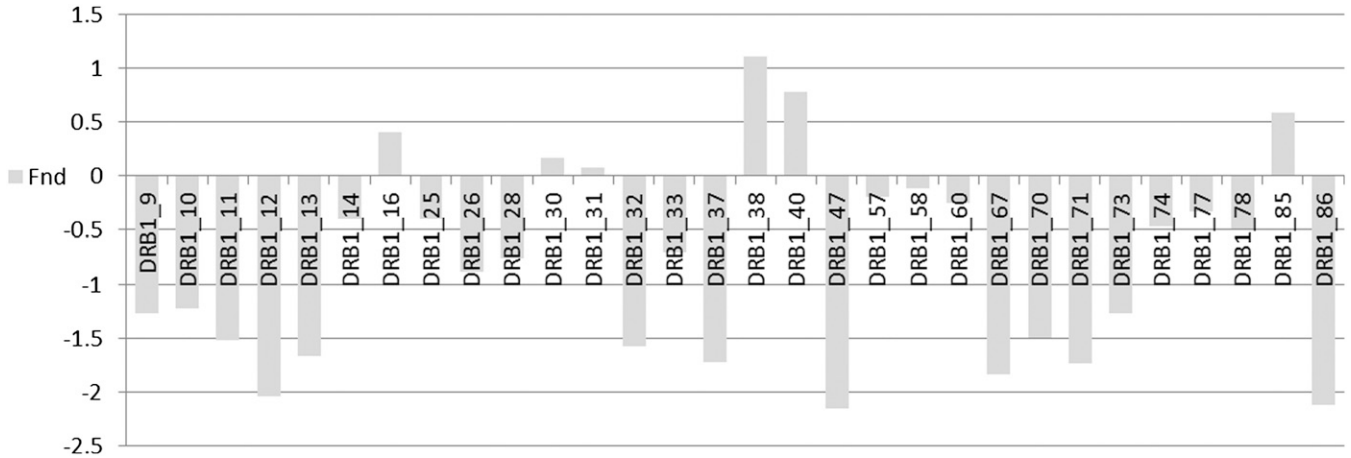

Figure 7 Selection at *DRB1* AA sites measured by F_{nd} values. The F_{nd} measure of selection for polymorphic amino acid sites in HLA–*DRB1* exon 2. The data are as for Figure 5 and Figure 6.

types (VTs) (Karp *et al.* 2010; Thomson *et al.* 2010), how can one distinguish between potentially causal effects *vs.* those due to LD? These AA-level analyses showed that there are cases with different numbers of "alleles" (AAs or SFVTs) at two loci where $W_n = 1$; nonetheless a stratified analysis could be applied to potentially distinguish disease predisposing variants. Also in population studies there are cases of two AA sites with $W_n \approx 1$, which show variation that appears to be under different selection pressures (Salamon *et al.* 1999; Lancaster 2006). The ALD measures can help provide additional insight in these situations.

The ALD measures are applicable to the study of any genetic variation, and the fact that they are measured on the same scale as the well-documented correlation measure r enhances their comparability and interpretation. They will be increasingly useful as next-generation sequencing methods identify more allelic variation, including nonbiallelic SNPs, insertion/deletion polymorphisms, and copy-number variants. Currently, these nonbiallelic SNP sites are often excluded from analyses. Linkage disequilibrium analyses among SNPs and among polymorphic genes are typically handled separately and polymorphic genes are often recoded as a set of dichotomous indicator variables (presence/absence of each allele) to simplify analyses at the expense of interpretation. The ALD statistics provide a measure of linkage disequilibrium that is on the same scale for comparisons among SNPs, among SNPs and more polymorphic loci, among haplotype blocks of SNPs, and for fine mapping of disease genes. The ALD measures are especially useful when there is asymmetry in the number of alleles at each locus, and it is suspected that even with very high W_n values, some haplotypes will allow for a stratified analysis. The ALD values, combined with the HSF values (Table 1), give us a numeric evaluation of the variation available for stratification analyses. It can be challenging to conduct several analyses, synthesizing results from various combinations and types of genetic variants as risk factors. The ALD measures form a base for such studies, along with consideration of other complementary summary measures of the strength and structure of LD in multiallelic data.

Acknowledgments

We thank Diogo Meyer, Montgomery Slatkin, and two anonymous reviewers for their helpful comments. We also thank Alex Lancaster for the use of his thesis data. This work was supported in part by National Institutes of Health (NIH) Contract HHSN272201200028C (G.T. and R.M.S.), NIH grant MH096262 (G.T.), and a 2013-14 REACH grant from the University of Vermont (R.M.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Data used in this paper can be found at the tagger/MHC webpage (http://www.broadinstitute.org/mpg/tagger/mhc.html) and at the Immunology Database and Analysis Portal (ImmPort - immport.niaid.nih.gov - SDY26 and SDY313).

Literature Cited

Begovich, A. B., P. V. Moonsamy, S. J. Mack, L. F. Barcellos, L. L. Steiner *et al.*, 2001 Genetic variability and linkage disequilibrium within the HLA-DP region: analysis of 15 different populations. Tissue Antigens 57: 424–439.

Begovich, A. B., V. E. Carlton, L. A. Honigberg, S. J. Schrodi, A. P. Chokkalingam *et al.*, 2004 A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am. J. Hum. Genet. 75: 330–337.

Carlton, V. E., X. Hu, A. P. Chokkalingam, S. J. Schrodi, R. Brandon *et al.*, 2005 PTPN22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. Am. J. Hum. Genet. 77: 567–581.

Chakravarti, A., C. C. Li, and K. H. Buetow, 1984 Estimation of the marker gene frequency and linkage disequilibrium from conditional marker data. Am. J. Hum. Genet. 36: 177–186.

Cramer, H., 1946 *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

de Bakker, P. I., G. McVean, P. C. Sabeti, M. M. Miretti, T. Green *et al.*, 2006 A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. 38: 1166–1172.

Erlich, H., A. M. Valdes, J. Noble, J. A. Carlson, M. Varney *et al.*, 2008 HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. Diabetes 57: 1084–1092.

Guo, S. W., 1997 Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum. Hered. 47: 301–314.

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

Hedrick, P. W., and G. Thomson, 1986 A two-locus neutrality test: applications to humans, *E. coli* and lodgepole pine. Genetics 112: 135–156.

Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Popul. Biol. 8: 117–126.

Hollenbach, J. A., S. D. Thompson, T. L. Bugawan, M. Ryan, M. Sudman *et al.*, 2010 Juvenile idiopathic arthritis and HLA class I and class II interactions and age-at-onset effects. Arthritis Rheum. 62: 1781–1791.

Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics 109: 611–631.

Kaplan, N., and B. S. Weir, 1992 Expected behavior of conditional linkage disequilibrium. Am. J. Hum. Genet. 51: 333–343.

Karp, D. R., N. Marthandan, S. G. Marsh, C. Ahn, F. C. Arnett *et al.*, 2010 Novel sequence feature variant type analysis of the HLA genetic association in systemic sclerosis. Hum. Mol. Genet. 19: 707–719.

Lancaster, A., 2006 Interplay of selection and molecular function in HLA genes. Ph.D. Thesis, University of California, Berkeley, CA.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49–67.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics 120: 849–852.

Maiste, P. J., and B. S. Weir, 1992 Estimating linkage disequilibrium from conditional data. Am. J. Hum. Genet. 50: 1139–1140.

Malkki, M., R. Single, M. Carrington, G. Thomson, and E. Petersdorf, 2005 MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic transplantation and disease association studies. Tissue Antigens 66: 114–124.

Maruyama, T., 1982 Stochastic integrals and their application to population genetics, pp. 151–166 in *Molecular Evolution*, *Protein*

Polymorphism and the Neutral Theory, edited by M. Kimura. Japan Scientific Societies Press, Tokyo.

Meyer, D., and G. Thomson, 2001 How selection shapes variation of the human major histocompatibility complex: a review. Ann. Hum. Genet. 65: 1–26.

Meyer, D., R. M. Single, S. J. Mack, H. A. Erlich, and G. Thomson, 2006 Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. Genetics 173: 2121–2142.

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

Ohta, T., 1980 Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. Genet. Res. 36: 181–197.

Salamon, H., J. Tarhio, K. Ronningen, and G. Thomson, 1996 On distinguishing unique combinations in biological sequences. J. Comput. Biol. 3: 407–423.

Salamon, H., W. Klitz, S. Easteal, X. Gao, H. A. Erlich *et al.*, 1999 Evolution of HLA class II molecules: allelic and amino acid site variability across populations. Genetics 152: 393–400.

Single, R., D. Meyer, and G. Thomson, 2007 Statistical methods for analysis of population genetic data, pp. 518–522 in *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress*, edited by J. Hansen. IHWG Press, Seattle, WA.

Single, R., P.-A. Gourraud, H. Maldonado-Torres, A. Lancaster, F. Briggs *et al.*, 2011 Estimating haplotype frequencies and linkage disequilibrium parameters in the HLA and KIR Regions. NIAID/NIH's ImmPort. https://immport.niaid.nih.gov/docs/standards/MethodsManual_HaplotypeFreqs+LD_v8.pdf.

Solberg, O. D., S. J. Mack, A. K. Lancaster, R. M. Single, Y. Tsai *et al.*, 2008 Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum. Immunol. 69: 443–464.

Thomson, G., A. M. Valdes, J. A. Noble, I. Kockum, M. N. Grote *et al.*, 2007 Relative predispositional effects of HLA class II DRB1–DQB1 haplotypes and genotypes on type 1 diabetes: a meta-analysis. Tissue Antigens 70: 110–127.

Thomson, G., L. F. Barcellos, and A. M. Valdes, 2008 Searching for additional disease loci in a genomic region. Adv. Genet. 60: 253–292.

Thomson, G., N. Marthandan, J. A. Hollenbach, S. J. Mack, H. A. Erlich *et al.*, 2010 Sequence feature variant type (SFVT) analysis of the HLA genetic association in juvenile idiopathic arthritis. Pac. Symp. Biocomput., 359–370.

Thomson, G., S. Mack, A. Valdes, L. Barcellos, J. Hollenbach *et al.*, 2011 HLA disease associations: detecting primary and secondary disease predisposing genes. NIAID/NIH's ImmPort. https://immport.niaid.nih.gov/docs/standards/MM-HL-Adisease-version010.pdf.

Tsai, Y., and G. Thomson, 2007 Selection intensity differences in seven HLA loci in many populations, pp. 705–746 in *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Congress*, edited by J. Hansen. IHWG Press, Seattle, WA.

Valdes, A. M., S. K. McWeeney, D. Meyer, M. P. Nelson, and G. Thomson, 1999 Locus and population specific evolution in HLA class II genes. Ann. Hum. Genet. 63: 27–43.

Wilson, C., 2010 Identifying polymorphisms associated with risk for the development of myopericarditis following smallpox vaccine. Study #26. NIAID/NIH's ImmPort. https://immport.niaid.nih.gov/immportWeb/clinical/study/displayStudyDetails.do?itemList=SDY26.

Wu, X., L. Jin, and M. Xiong, 2008 Composite measure of linkage disequilibrium for testing interaction between unlinked loci. Eur. J. Hum. Genet. 16: 644–651.

Communicating editor: J. Wall

Appendix: Alternate Expressions for *ALD* Statistics

Alternate expressions for $F_{A/B}$ and $F_{B/A}$ for multiallelic data

The two overall HSF measures can also be expressed as haplotype and allele frequencies (line 1 below), or as a deviation from the single-locus homozygosity (second line below) using individual LD (D_{ij}) values and allele frequencies.

$$F_{A/B} = \sum_j \left(F_{A/Bj}\right)\left(p_{Bj}\right) = \sum_j \sum_i \left(f_{ij}/p_{Bj}\right)^2 \left(p_{Bj}\right) = \sum_i \sum_j \left[f_{ij}^2/p_{Bj}\right] = \sum_i \sum_j \left[p_{Ai}p_{Bj} + D_{ij}\right]^2/p_{Bj}$$
$$= \sum_i \sum_j \left[p_{Ai}^2 p_{Bj} + 2p_{Ai}D_{ij} + D_{ij}^2/p_{Bj}\right] = F_A + \sum_i \sum_j D_{ij}^2/p_{Bj}$$

(since $\sum_j D_{ij} = 0$; see Table 1 footnote). Similarly, $F_{B/A} = F_B + \sum_i \sum_j D_{ij}^2/p_{Ai}$. It follows that $F_{A/B} \geq F_A$ with equality only when all $D_{ij} = 0$ (a "Wahlund" effect).

Alternate expressions for $F_{A/B}$ and $F_{B/A}$ for biallelic data

If both loci are biallelic:

$$F_{A/B} = F_A + \sum_i \sum_j D_{ij}^2/p_{Bj} = F_A + [D_{11}^2/p_{B_1}] + [D_{12}^2/p_{B_2}] + [D_{21}^2/p_{B_1}] + [D_{22}^2/p_{B_2}]$$
$$= F_A + 2[D^2/p_{B_1}] + 2[D^2/p_{B_2}], \quad \text{since } D_{11} = -D_{12} = -D_{21} = D_{22} = D,$$
$$= F_A + 2\ D^2/(p_{B_1}\ p_{B_2}).$$

Similarly, $F_{B/A} = F_B + 2\ D^2/(p_{A_1}\ p_{A_2})$.

Alternate expressions for $W_{A/B}^2$ and $W_{B/A}^2$ for multiallelic data

$W_{A/B}^2$ and $W_{B/A}^2$ (Table 1) can also be expressed using haplotype and allele frequencies or using individual LD (D_{ij}) values and allele frequencies:

$$W_{A/B}^2 = \left(F_{A/B} - F_A\right)/(1 - F_A) = \left(\sum_i \sum_j \left[f_{ij}^2/p_{Bj}\right] - F_A\right)/(1 - F_A) = \left[\sum_i \sum_j \left(D_{ij}^2/p_{Bj}\right)\right]/(1 - F_A),$$

$$W_{B/A}^2 = \left(F_{B/A} - F_B\right)/(1 - F_B) = \left(\sum_i \sum_j \left[f_{ij}^2/p_{Ai}\right] - F_B\right)/(1 - F_B) = \left[\sum_i \sum_j \left(D_{ij}^2/p_{Ai}\right)\right]/(1 - F_B).$$

Alternate expressions for $W_{A/B}^2$ and $W_{B/A}^2$ for biallelic data

If both loci are biallelic:

$$W_{A/B}^2 = \left(F_{A/B} - F_A\right)/(1 - F_A) = \left(\sum_i \sum_j \left[f_{ij}^2/p_{Bj}\right] - F_A\right)/(1 - F_A) = \left[\sum_i \sum_j \left(D_{ij}^2/p_{Bj}\right)\right]/(1 - F_A)$$
$$= D^2/(p_{A_1}\ p_{A_2}\ p_{B_1}\ p_{B_2}) = r^2.$$

Similarly, $W_{B/A}^2 = r^2$.

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.165266/-/DC1

Conditional Asymmetric Linkage Disequilibrium (ALD): Extending the Biallelic r^2 Measure

Glenys Thomson and Richard M. Single

DOI: 10.1534/genetics.114.165266

File S1

A. Constraints on individual LD values

The normalized individual LD values are $D_{ij}' = D_{ij} / D_{max}$, where $D_{max} = \min[p_{Ai}(1-p_{Bj}), (1-p_{Ai})\,p_{Bj}]$ if $D_{ij}>0$, and $D_{max} = \min[p_{Ai}\,p_{Bj,}\,(1-p_{Ai})(1-p_{Bj})]$ when $D_{ij}<0$. The range of D_{ij}' is (−1, 1).

B. Other measures of conditional LD

The conditional association between the alleles at *marker* locus B and *disease* locus A was defined by Nei and Li (1980) as $d_* = (f_{A2B1} / p_{A2} - f_{B1A1} / p_{A1})$ for two bi-allelic loci. It was developed to account for study designs for rare diseases with known modes of inheritance and full, or close to full, penetrance (e.g., sickle cell anemia and Duchenne muscular dystrophy) where individuals are not randomly sampled from a single population. This measure is equivalent to Somer's *D* statistic, *D(C|R)*, conditioning on the rows of the contingency table relating two categorical variables, as shown below.

$$d_* = f_{A2B1} / p_{A2} - f_{A1B1} / p_{A1} = \left(p_{A2}\,p_{B1} + D\right) / p_{A2} - \left(p_{A1}\,p_{B1} - D\right) / p_{A1}$$
$$= p_{B1} + \left(D / p_{A2}\right) - p_{B1} + \left(D / p_{A1}\right) = D / \left(p_{A1}p_{A2}\right)$$

Somer's *D* statistic is defined as twice the difference between the number of concordant and discordant entries in the contingency table, divided by $n^2 - \Sigma_i\, n_{i\cdot}^2$, where n is the table total and $n_{i\cdot}$ is the i^{th} row total.

$$D(C\,|\,R) = 2\left(n_{11}\,n_{22} - n_{21}\,n_{12}\right) / \left(n^2 - \sum\nolimits_i n_{i\cdot}^2\right) = 2\left(f_{A1B1}f_{A2B2} - f_{A2B1}f_{A1B2}\right) / \left(1^2 - \sum\nolimits_i p_{Ai}^2\right)$$
$$= 2D / \left(1 - F_A\right) = 2D / \left(2p_{A1}p_{A2}\right) = D / \left(p_{A1}p_{A2}\right)$$

C. Other measures of LD based on diversity statistics

Other measures of association and LD that are based on allelic diversity statistics (homozygosity and heterozygosity) have been defined. However, these measures are all symmetric. Ohta (1980) suggested a measure, *F'*, that divides the difference between the two-locus haplotypic homozygosity (F_{AB}) and the product of the two single locus homozygosity values by the product of the single locus heterozygosity values: $F' = (F_{AB} - F_A\,F_B) / [(1 - F_A)\,(1 - F_B)]$ (Ohta 1980). The *D** measure (Maruyama 1982; Hedrick and Thomson 1986) standardizes the disequilibrium between all alleles at two loci by the product of the single locus heterozygosity values: $D^* = \Sigma_i\,\Sigma_j\, D_{ij}^2 / [(1 - F_A)\,(1 - F_B)]$. Note that in the bi-allelic case *D** is equivalent to the square of the correlation measure (r^2) (Hedrick 1987).

D: Proofs of Special Cases (c) – (f)

D.1. Multi-Allelic Case (c): $k_A = k_B = k$, $f(AiBi) > 0$, $i = 1, 2, \ldots k$, and $f(AiBj) = 0$ for all $i \neq j$

Summary: $W_n = W_{A/B} = W_{B/A} = 1$. There is complete symmetry and 100% correlation of alleles at the two loci.

The *A* and *B* locus allele frequencies are equal and the notation is simplified as follows: $p_{Ai} = p_{Bi} = p_i$, $i = 1, 2, \ldots k$,

$f(A_iB_i) = p_i = p_i^2 + D_{ii}$, i.e., $D_{ii} = p_i\,(1 - p_i)$, $f(A_iB_j) = 0 = p_i\,p_j + D_{ij}$, i.e., $D_{ij} = -\,p_i\,p_j$, $i \neq j$.

$W_{A/B}^2 = \{\Sigma_i\,\Sigma_j\,[f_{ij}^2 / p_{Bj}] - F_A\} / (1 - F_A) = [\{\Sigma_i\, f_{ii}^2 / p_i\} - F_A]\} / (1 - F_A)$

$= [\{\Sigma_i\, p_i^2/p_i\} - F_A] / (1 - F_A) = [\{\Sigma_i\, p_i\} - F_A] / (1 - F_A) = (1 - F_A) / (1 - F_A) = 1.$

Similarly, $W_{B/A}^2 = 1$.

$W_n^2 = \{\Sigma_i \Sigma_j\, D_{ij}^2 / (p_{Ai}\, p_{Bj})\} / (k-1) = [\{\Sigma_i\, D_{ii}^2 / (p_i^2)\} + \{\Sigma_i \Sigma_{j \neq i}\, D_{ij}^2 / (p_i\, p_j)\}] / (k-1)$

$= [\{\Sigma_i\, p_i^2\, (1 - p_i)^2 / (p_i^2)\} + \{\Sigma_i \Sigma_{j \neq i}\, (p_i\, p_j)^2 / (p_i\, p_j)\}] / (k-1) = [\Sigma_i\, (1 - p_i)^2 + \{\Sigma_i \Sigma_{j \neq i}\, (p_i\, p_j)] / (k-1)$

$= [\Sigma_i\, (1 - p_i)^2 + p_i\, (1 - p_i)] / (k-1) = [\Sigma_i\, (1 - p_i)] / (k-1) = (k-1) / (k-1) = 1.$

D.2. Multi-Allelic Case (d): $k_A = k_B = k$, $f(AiBi) > 0$, $i = 1, 2, \ldots k$, and $f(A1B2) \neq 0$

In this case $W_{A/B} < 1$, $W_{B/A} < 1$, and $W_n < 1$ (Proof given below for $k = 3$, with the same result holding for any value of k).

Summary: $W_{A/B}^2 = \{1 - 2\,\delta\,[p_{A2} / p_{B2}] - F_A\} / (1 - F_A)$, $W_{B/A}^2 = \{1 - 2\,\delta\,[p_{B1} / p_{A1}] - F_B\} / (1 - F_B)$,

$W_n^2 = 1 - \{\delta\,(p_{A1} + p_{A2}) / [2\, p_{A1}\, p_{B2}]\}$

Allele frequencies are (A locus): p_{A1}, p_{A2}, and p_{A3}; (B locus): $p_{B1} = p_{A1} - \delta$, $p_{B2} = p_{A2} + \delta$, and $p_{B3} = p_{A3}$, with $\delta > 0$ and $\delta < p_{A1}$.

Haplotype	Frequency*	Frequency*	LD
A1B1	$p_{A1} - \delta$	$p_{A1}\,(p_{A1} - \delta) + D_{11}$	$D_{11} = (p_{A1} - \delta)\,(1 - p_{A1})$
A1B2	δ	$p_{A1}\,(p_{A2} + \delta) + D_{12}$	$D_{12} = -\,p_{A1}\,(p_{A2} + \delta) + \delta$
A1B3	0	$p_{A1}\, p_{A3} + D_{13}$	$D_{13} = -\,p_{A1}\, p_{A3}$
A2B1	0	$p_{A2}\,(p_{A1} - \delta) + D_{21}$	$D_{21} = -\,p_{A2}\,(p_{A1} - \delta)$
A2B2	p_{A2}	$p_{A2}\,(p_{A2} + \delta) + D_{22}$	$D_{22} = p_{A2}\,(1 - p_{A2} - \delta)$
A2B3	0	$p_{A2}\, p_{A3} + D_{23}$	$D_{23} = -\,p_{A2}\, p_{A3}$
A3B1	0	$p_{A3}\,(p_{A1} - \delta) + D_{31}$	$D_{31} = -\,p_{A3}\,(p_{A1} - \delta)$
A3B2	0	$p_{A3}\,(p_{A2} + \delta) + D_{32}$	$D_{32} = -\,p_{A3}\,(p_{A2} + \delta)$
A3B3	p_{A3}	$p_{A3}\, p_{A3} + D_{33}$	$D_{33} = p_{A3}\,(1 - p_{A3})$

* These two columns are different, but equivalent, formats for the same haplotype frequency.

$W_{A/B}^2 = \{\Sigma_i \Sigma_j\, [f_{ij}^2 / p_{Bj}] - F_A\} / (1 - F_A)$

$= \{[(p_{B1}^2) / (p_{B1}) + (\delta^2) / (p_{B2}) + (p_{A2}^2) / (p_{B2}) + (p_{A3}^2) / (p_{A3})] - F_A\} / (1 - F_A)$

$= \{[p_{B1} + (\delta^2) / (p_{B2}) + (p_{A2}^2) / (p_{B2}) + p_{A3}] - F_A\} / (1 - F_A)$

$= \{[(p_{A1} - \delta) + (\delta^2) / (p_{A2} + \delta) + (p_{A2})^2 / (p_{A2} + \delta) + p_{A3} + p_{A2} - p_{A2}] - F_A\} / (1 - F_A)$

$= \{[\, p_{A1} + p_{A2} + p_{A3} - [(\delta(p_{A2} + \delta) - (\delta^2) - (p_{A2})^2 + p_{A2}\,(p_{A2} + \delta)] / (p_{A2} + \delta)] - F_A\} / (1 - F_A)$

$= \{1 - 2\,\delta\,[p_{A2} / p_{B2}] - F_A\} / (1 - F_A) < 1$ always since $\delta > 0$

$W_{B/A}^2 = \{\Sigma_i \Sigma_j\, [f_{ij}^2 / p_{Ai}] - F_B\} / (1 - F_B)$

$= \{[(p_{A1} - \delta)]^{\,2}) / (p_{A1}) + (\delta^2) / (p_{A1}) + p_{A2} + p_{A3}] - F_B\} / (1 - F_B)$

$= \{[\, p_{A1} + p_{A2} + p_{A3} - [(p_{A1})^2 - (\delta^2) - (p_{A1} - \delta)^2] / (p_{A1})] - F_B\} / (1 - F_B)$

$= \{1 - 2\,\delta\,[(p_{A1} - \delta) / p_{A1}] - F_B\} / (1 - F_B) = \{1 - 2\,\delta\,[p_{B1} / p_{A1}] - F_B\} / (1 - F_B) < 1$ always since $\delta > 0$.

$W_n^2 = \{\Sigma_i \Sigma_j\, D_{ij}^2 / (p_{Ai}\, p_{Bj})\} / (k-1)$

$= \{[(p_{A1} - \delta)\,(1 - p_{A1})^{\,2} / p_{A1}] + p_{A1}\,(p_{A2} + \delta) - 2\delta + \delta^2 / [\, p_{A1}\,(p_{A2} + \delta)] + p_{A1}\, p_{A3} + p_{A2}\,(p_{A1} - \delta) + [p_{A2} / (p_{A2} + \delta)] - 2\, p_{A2}$

$+ p_{A2}(p_{A2}+\delta) + p_{A2}p_{A3} + p_{A3}(p_{A1}-\delta) + p_{A3}(p_{A2}+\delta) + (1-p_{A3})^2\} / 2$

$= \{(1-p_{A1})^2 - [\delta(1-p_{A1})^2 / p_{A1}] + p_{A1}p_{A2} + \delta p_{A1} - 2\delta + \delta^2 / [p_{A1}(p_{A2}+\delta)] + p_{A1}p_{A3} + p_{A2}p_{A1} - \delta p_{A2} + [p_{A2}/(p_{A2}+\delta)]$

$+ [1-(p_{A2}+\delta)/(p_{A2}+\delta)] - 2p_{A2} + p_{A2}^2 + \delta p_{A2} + p_{A2}p_{A3} + p_{A3}p_{A1} - \delta p_{A3} + p_{A3}p_{A2} + \delta p_{A3} + (1-p_{A3})^2\} / 2$

$= \{(1-p_{A1})^2 + 2p_{A1}p_{A2} + 2p_{A1}p_{A3} + 2p_{A2}p_{A3} + (1-p_{A2})^2 + (1-p_{A3})^2 - [\delta(1-p_{A1})^2] / [p_{A1}] + \delta p_{A1} - 2\delta + \delta^2 / [p_{A1}(p_{A2}+\delta)]$

$- [\delta / (p_{A2}+\delta)]\} / 2$

$= \{3 - 2(p_{A1}+p_{A2}+p_{A3}) + (p_{A1}+p_{A2}+p_{A3})^2 - [\delta / p_{A1}] + 2\delta - \delta p_{A1} + \delta p_{A1} - 2\delta + \delta^2 / [p_{A1}(p_{A2}+\delta)] - [\delta / (p_{A2}+\delta)]\} / 2$

$= 1 - \{\delta(p_{A1}+p_{A2}) / [2p_{A1}p_{B2}]\} < 1$ always since $\delta > 0$.

D.3. Multi-Allelic Case (e): $k_A < k_B$, with each B_j allele occurring with only one A_i allele

In this case $W_n = W_{A/B} = 1$, and $W_{B/A} < 1$ (Proof given below for $k_A = 3$ and $k_B = 4$, and the equivalent result holds if $k_A > k_B$.)

Summary: $W_{A/B}^2 = 1$, $W_{B/A}^2 = \{1 - [2p_{B3}p_{B4} / (p_{B3}+p_{B4})] - F_B\} / (1 - F_B)$, $W_n^2 = 1$

Haplotype	Frequency*	Frequency*	LD
A1B1	p_{B1}	$p_{B1}^2 + D_{11}$	$D_{11} = p_{B1}(1-p_{B1})$
A1B2	0	$p_{B1}p_{B2} + D_{12}$	$D_{12} = -p_{B1}p_{B2}$
A1B3	0	$p_{B1}p_{B3} + D_{13}$	$D_{13} = -p_{B1}p_{B3}$
A1B4	0	$p_{B1}p_{B4} + D_{14}$	$D_{14} = -p_{B1}p_{B4}$
A2B1	0	$p_{B2}p_{B1} + D_{21}$	$D_{21} = -p_{B2}p_{B1}$
A2B2	p_{B2}	$p_{B2}^2 + D_{22}$	$D_{22} = p_{B2}(1-p_{B2})$
A2B3	0	$p_{B2}p_{B3} + D_{23}$	$D_{23} = -p_{B2}p_{B3}$
A2B4	0	$p_{B2}p_{B4} + D_{24}$	$D_{24} = -p_{B2}p_{B4}$
A3B1	0	$p_{B1}(p_{B3}+p_{B4}) + D_{31}$	$D_{31} = -p_{B1}(p_{B3}+p_{B4})$
A3B2	0	$p_{B2}(p_{B3}+p_{B4}) + D_{32}$	$D_{32} = -p_{B2}(p_{B3}+p_{B4})$
A3B3	p_{B3}	$p_{B3}(p_{B3}+p_{B4}) + D_{33}$	$D_{33} = p_{B3}(1-p_{B3}-p_{B4})$
A3B4	p_{B4}	$p_{B4}(p_{B3}+p_{B4}) + D_{34}$	$D_{34} = p_{B4}(1-p_{B3}-p_{B4})$

Allele frequencies at the *A* locus are: p_{A1}, p_{A2}, and $p_{A3} = (p_{B3}+p_{B4})$, and at the *B* locus they are: p_{B1} (= p_{A1}), p_{B2} (= p_{A2}), p_{B3} and p_{B4}.

$W_{A/B}^2 = \{\Sigma_i \Sigma_j [f_{ij}^2 / p_{Bj}] - F_A\} / (1 - F_A) = \{p_{B1}+p_{B2}+p_{B3}+p_{B4} - F_A\} / (1 - F_A) = 1$

$W_{B/A}^2 = \{\Sigma_i \Sigma_j [f_{ij}^2 / p_{Ai}] - F_B\} / (1 - F_B)$

$= \{p_{B1} + p_{B2} + [p_{B3}^2 / (p_{B3}+p_{B4})] + [p_{B4}^2 / (p_{B3}+p_{B4})] - F_B\} / (1 - F_B)$

$= \{p_{B1} + p_{B2} + [(p_{B3}^2 + p_{B4}^2) / (p_{B3}+p_{B4})] - F_B\} / (1 - F_B)$

$= \{p_{B1} + p_{B2} + [(p_{B3}+p_{B4})^2 / (p_{B3}+p_{B4})] - [2p_{B3}p_{B4} / (p_{B3}+p_{B4})] - F_B\} / (1 - F_B)$

$= \{1 - [2p_{B3}p_{B4} / (p_{B3}+p_{B4})] - F_B\} / (1 - F_B) < 1$ always

$W_n^2 = \{\Sigma_i \Sigma_j D_{ij}^2 / (p_{Ai}p_{Bj})\} / (k_A - 1)$

$= \{(1-p_{B1})^2 + p_{B1}p_{B2} + p_{B1}p_{B3} + p_{B1}p_{B4} + p_{B2}p_{B1} + (1-p_{B2})^2 + p_{B2}p_{B3} + p_{B2}p_{B4} + p_{B1}(p_{B3}+p_{B4}) + p_{B2}(p_{B3}+p_{B4}) + 1 - 2p_{A3} + p_{A3}^2\} / 2$

$= \{3 - 2(p_{B1}+p_{B2}+p_{A3}) + (p_{B1}+p_{B2}+p_{A3})^2\} / 2 = 1$

D.4. One Locus Bi-Allelic and the Other Multi-Allelic, Case (f): $k_A = 2$, $k_B > 2$.

In this case $W_n = W_{A/B} \neq W_{B/A}$ (Proof given below for $k_A = 2$ and $k_B = 3$, with the obvious extension for any value of k_B)

Summary: $W_{A/B}^2 = [\{D_{11}^2 / (p_{B1})\} + \{D_{12}^2 / (p_{B2})\} + \{D_{11} + D_{12})^2 / (p_{B3})\}] / (p_{A1}\, p_{A2})$, $W_{B/A}^2 = [\{D_{11}^2 + D_{12}^2 + (D_{11} + D_{12})^2] / \{(p_{A1}\, p_{A2})\,(1 - F_B)\}$,

$W_n^2 = W_{A/B}^2$

Haplotype	Frequency	LD*
A1B1	$p_{A1}\, p_{B1} + D_{11}$	D_{11}
A1B2	$p_{A1}\, p_{B2} + D_{12}$	D_{12}
A1B3	$p_{A1}\, p_{B3} + D_{13}$	$- D_{11} - D_{12}$
A2B1	$p_{A2}\, p_{B1} + D_{21}$	$- D_{11}$
A2B2	$p_{A2}\, p_{B2} + D_{22}$	$- D_{12}$
A2B3	$p_{A2}\, p_{B3} + D_{23}$	$D_{11} + D_{12}$

* See Table 1 footnote for constraints on LD parameters, such that two LD variables, D_{11} and D_{12}, along with three allele frequencies (p_{A1}, p_{B1}, and p_{B2}) are sufficient to describe the five independent haplotype frequencies in this case.

$W_{A/B}^2 = \{\Sigma_i \Sigma_j\, D_{ij}^2 / (p_{Bj})\} / (1 - F_A)$

$= [\{D_{11}^2 / (p_{B1})\} + \{D_{12}^2 / (p_{B2})\} + \{D_{11} + D_{12})^2 / (p_{B3})\} + \{D_{11}^2 / (p_{B1})\} + \{D_{12}^2 / (p_{B2})\} + \{D_{11} + D_{12})^2 / (p_{B3})\}] / (2\, p_{A1}\, p_{A2})$

$= [\{D_{11}^2 / (p_{B1})\} + \{D_{12}^2 / (p_{B2})\} + \{D_{11} + D_{12})^2 / (p_{B3})\}] / (p_{A1}\, p_{A2})$

$W_{B/A}^2 = \{\Sigma_i \Sigma_j\, D_{ij}^2 / (p_{Ai})\} / (1 - F_B)$

$= [\{D_{11}^2 / (p_{A1})\} + \{D_{12}^2 / (p_{A1})\} + \{D_{11} + D_{12})^2 / (p_{A1})\} + \{D_{11}^2 / (p_{A2})\} + \{D_{12}^2 / (p_{A2})\} + \{D_{11} + D_{12})^2 / (p_{A2})\}] / (1 - F_B)$

$= [\{D_{11}^2 + D_{12}^2 + (D_{11} + D_{12})^2] / \{(p_{A1}\, p_{A2})\,(1 - F_B)\}$

$W_n^2 = \{\Sigma_i \Sigma_j\, D_{ij}^2 / (p_{Ai}\, p_{Bj})\} / (2-1)$

$= [\{D_{11}^2 / (p_{A1}\, p_{B1})\} + \{D_{12}^2 / (p_{A1}\, p_{B2})\} + \{D_{11} + D_{12})^2 / (p_{A1}\, p_{B3})\} + \{D_{11}^2 / (p_{A2}\, p_{B1})\} + \{D_{12}^2 / (p_{A2}\, p_{B2})\} + \{D_{11} + D_{12})^2 / (p_{A2}\, p_{B3})\}]$

$= [\{D_{11}^2 / (p_{B1})\} + \{D_{12}^2 / (p_{B2})\} + \{D_{11} + D_{12})^2 / (p_{B3})\}] / (p_{A1}\, p_{A2}) = W_{A/B}^2$

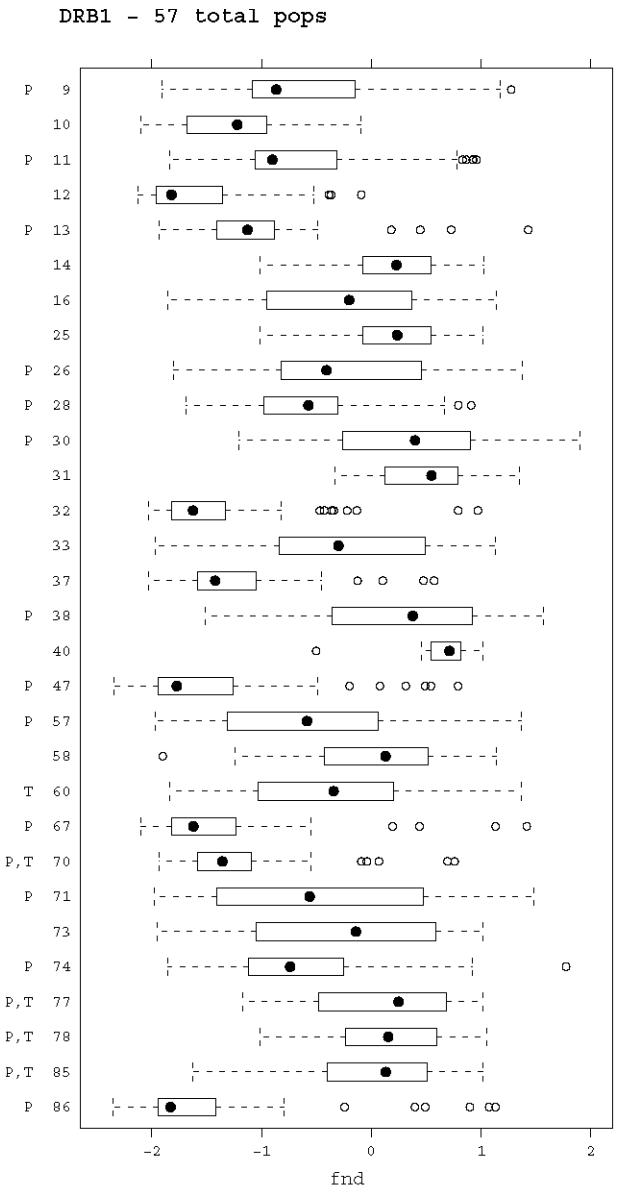

Figure S1 Boxplots of F_{nd} values for *DRB1* Amino Acids 9 to 86 from 57 Populations (Lancaster 2006)*

*AA position is indicated on the vertical axis with P indicating a peptide interacting site and T indicating a T-cell interacting site. The boxes indicate the middle 50% of the Fnd values across the 57 populations with a line drawn at the median value. The spread of the central half of the data is called the interquartile range. Extreme observations that are more than 1.5 times the interquartile range away from the central box are identified with circles.