Volume Measures for Linkage Disequilibrium

Yuguo Chen[1], Chia-Ho Lin[2], and Chiara Sabatti[2,3]

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign IL 61820[1].

Department of Statistics[2], and Department of Human Genetics[3], UCLA, Los Angeles CA 90095-7088.

UCLA Statistics Department Preprint # 465

April 2006

Running head Volume Measures for LD

Keywords Linkage disequilibrium measures; volume tests; D'; Homozygosity; Mahalanobis distance.

Corresponding author Chiara Sabatti

Department of Human Genetics
UCLA School of Medicine
695 Charles E. Young Drive South
Los Angeles, California 90095-7088 (USA)
FAX: (310) 794-5446
Phone: (310) 794-9567
e-mail: csabatti@mednet.ucla.edu

Abstract

We discuss the value of volume measures for linkage disequilibrium, showing how they are robust to small sample variation and easily generalized to multi-allelic markers. In particular we introduce $Dvol$, a volume analogue to D' and show that it performs substantially better when the sample size is small to moderate. $Mvol$ is proposed as a generalization of this measure to multi-allelic markers. Finally a measure based on homozygosity $Hvol$ is suggested as a generalization of R^2. To evaluate these measures, we introduce a sequential importance sampling algorithm. We illustrate their performance on simulated and real data.

Linkage disequilibrium (LD), that is the association between alleles at different markers, is influenced by the recombination fraction between the markers, the population history, the marker mutation rates, etc. The relation between linkage disequilibrium and the amount of recombination between markers can be described with the following simplified population model. Suppose each individual has one chromosome and that the chromosomes of the current generation are obtained by either sampling one of the chromosomes of the previous generation and not recombining it or sampling two and recombining them. We are interested in considering two loci, and let π_{ij}^t be the joint frequency of allele i at locus 1 and allele j at locus 2 at generation t. Let θ be the probability of recombination between the two loci. Then, $\pi_{ij}^{t+1} = \pi_{ij}^t(1-\theta) + p_{i\cdot}p_{\cdot j}\theta$; if we call disequilibrium at generation t in the cell ij the difference $D_{ij}^t = \pi_{ij}^t - p_{i\cdot}p_{\cdot j}$, we have $D^t = (1-\theta)^t D^0$, which illustrates the connection between the amount of disequilibrium and frequency of recombination. This relation can be interpreted as to say that, within one population, one might expect higher levels of disequilibrium between markers that are closer together (in terms of lower recombination fraction) rather than far apart. This assumption is at the base of the idea of association mapping. However, due to differential heterozygosity and mutation rates between markers, it is impossible to see such relation hold between all marker pairs, while it appears to hold as an average behavior.

A necessary step towards the study of variation of linkage disequilibrium is the definition of an appropriate measure of LD. Generally speaking, one seeks to standardize the amount of disequilibrium observed in a table in order to be able to compare it across different marker pairs and populations. With this goal, a measure whose value are ranging between 0 (in correspondence of perfect equilibrium) and 1 (maximum disequilibrium) is desired (for a general review of measures of LD, see Devlin and Risch [1995]). For the case of biallelic markers, two such measures are currently used by practitioners: D' and R^2. The case of biallelic markers is particularly simple as the absolute values of the entries D_{ij} is the same for all i and j. To simplify notation, we will use the following parameterization for the haplotype distribution at two biallelic markers:

$$\begin{array}{c|c|c|c} & 1 & 0 & \\ \hline 1 & x & p-x & p \\ \hline 0 & q-x & 1-p-q+x & 1-p \\ \hline & q & 1-q & 1 \end{array} \tag{1}$$

The $|R|$ measure is derived by treating the allele values $0, 1$ as quantitative and calculating the correlation coefficient between the two random variables corresponding to the two alleles:

$$|R| = \frac{|x - pq|}{\sqrt{pq(1-p)(1-q)}}.$$

It is clear that $|R| \leq 1$, however this bound is quite crude in that the value of 1 can be achieved only when $p = q$ or $p = 1 - q$. That is, $|R|$ is going to equal 1 only when knowing the allele at one of the markers always allows to predict perfectly the allele at the other marker (irrespectively of which marker is known and which allele value). For this reason, $|R|$ (or R^2) is preferred when the goal of linkage disequilibrium measure is too assess how predictable the alleles at a polymorphic site are, given the alleles at neighboring markers [Pritchard and Przeworski, 2001]

The maximum value of D conditional on p, q can also be easily calculated in a 2x2 table. This was done, for example, by Lewontin, also taking into account the direction of disequilibrium. The resulting D' measure is defined as:

$$D' = \begin{cases} \dfrac{|D|}{\min(p(1-q), q(1-p))} & \text{if} \quad D \geq 0 \\ \\ \dfrac{|D|}{\min(pq, (1-p)(1-q))} & \text{if} \quad D < 0 \end{cases},$$

where we have taken the absolute value of D, since its sign carries no genetically relevant information. Clearly one could define another measure that standardize D conditional on p, q, but irrespectively of its directions (the denominator would be the maximum of the two different denominators appearing in the D' definition). It is relevant to note that $D' = 1$ as soon as one of the entries in the contingency table is equal to zero (while $|R| = 0$ only if two are): this correspond to the situation of no recombination having ever occurred between the two markers since the arising of one of the polymorphisms and hence is favored by geneticists that try to measure the amount of recombination in terms of LD.

While by far the most used in practice, the two described measures of LD have some substantial limitations. First and foremost, they are only defined on 2x2 tables and it is not possible to extend them directly to tables of general sizes $r \times c$. Moreover, they are defined on population haplotype frequencies and their properties when evaluated on the corresponding sample frequencies are not entirely clear or satisfactory. For example, it is easy to see that the fact that one zero entry suffices to generate a $D' = 1$ may lead to inflated values of D' when one of the haplotypes has low frequency and the size of the studied sample is not sufficiently large. Detailed analysis of this phenomena are available in Teare *et al.* [2002] and Tenesa *et al.* [2004]. To avoid spurious results, researchers often calculate empirical confidence intervals for D', using resampling schemes (see, for example, Gabriel *et al.* [2002]). While this certainly takes care of the variability of D', it does not result in a “corrected” measure of linkage disequilibrium.

We here suggest the use of volume measures both to take effectively into account variability due to sample size and to successfully define measures that are applicable to multi-allelic markers.

We have argued before about their relevance as measure of association [Sabatti, 2002]. We here precise their role in generalizing linkage disequilibrium measures and present a novel and fast algorithm to evaluate them on finite samples.

1 Volume measures of linkage disequilibrium

The generic notion of a volume measure [Sabatti, 2002; Diaconis and Efron, 1985; Hotelling, 1939] is based on a different approach to the standardization of the distance of a given table from equilibrium. Given a discrepancy measure, rather than using its maximum value to standardize it, one calculates the proportion of probability distributions that lead to lower values of such discrepancy among all the possible probability distributions for the problem. This immediately leads to some of the appealing features of volume measures: 1) they are scale independent, in that only the ordering of probability distributions induced by the discrepancy matter; 2) they are easily interpretable in terms of probability that a distribution selected uniformly among all the possible ones results in a lower discrepancy from equilibrium than the observed one.

In the case of 2x2 tables, it is easy to evaluate volume measures, and it is perhaps useful to compare them in some detail with the two measures we have previously described. In terms of the parameterization (1), $\max(0, p+q-1) \leq x \leq \min(p, q)$ represent all the possible haplotypes distributions on 2 biallelic markers with marginal allele frequencies p and q. In Figure 1, the range of x is represented on the horizontal coordinate, a specific value of z corresponding to one distribution is put in evidence, and the values of the curve $|x - pq|$ are drawn. The measures of disequilibrium D' (red) and $|R|$ (green) can be described as ratios of values read on the y coordinate: the numerators are indicated with broken lines and the denominators with solid ones. Note that the denominator in $|R|$ typically does not correspond to the achievable maximum for $|x - pq|$. Values of volume measures are, instead, ratios of quantities identifiable on the x axis. Two measures are described: $Dvol$ (red) and $Mvol$ (blue). $Dvol$ is defined as the ratio of the

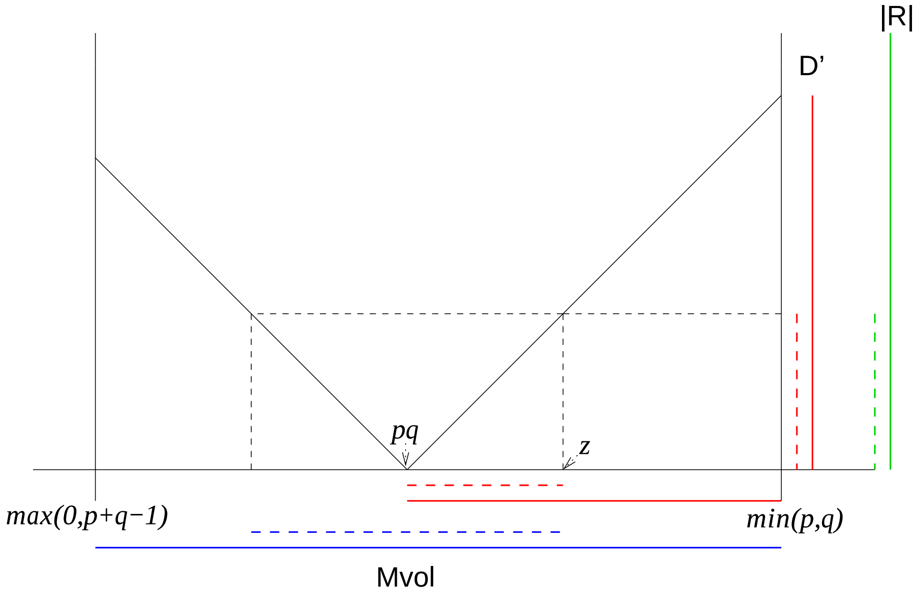

Figure 1: Measures of association on 2x2 tables. The value of the x entry of table (1) is displayed on the horizontal coordinate. The point z corresponds to a specific table under consideration. The highlighted point pq corresponds to linkage equilibrium (independence). The values of the measures D', $|R|$, and $Mvol$ for the table identified by z are presented as ratios between broken and solid lines of the following, respective, colors: red, green, and blue.

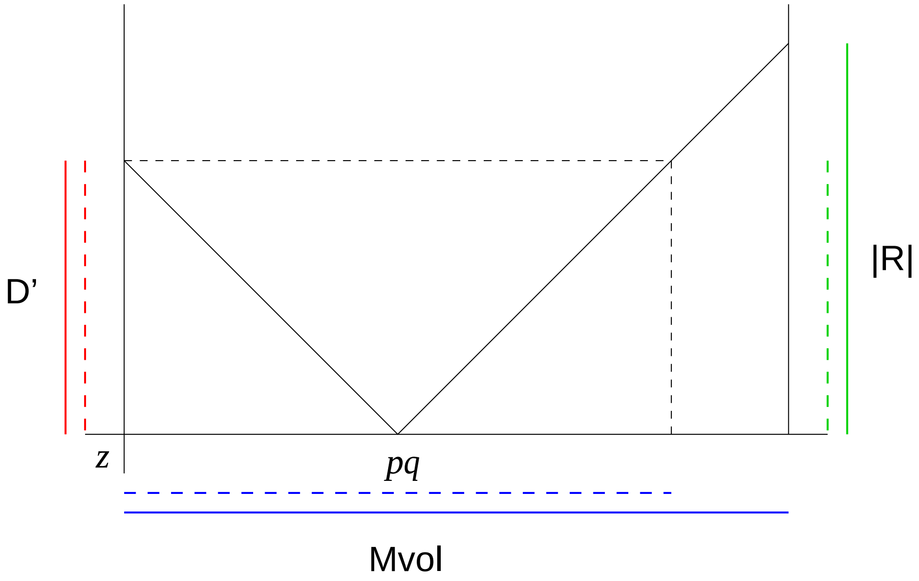

Figure 2: Measures of association on 2x2 tables. The value of the x entry of table (1) is displayed on the horizontal coordinate. The point z corresponds to a specific table under consideration. The highlighted point pq corresponds to linkage equilibrium (independence). We assumed here that $p = q$. The values of the measures D', $|R|$, and $Mvol$ for the table identified by z are presented as ratios between broken and solid lines of the following, respective, colors: red, green, and blue.

volume of the space of distributions for which $|x - pq| < |z - pq|$ and $(x - pq)(z - pq) \geq 0$ and the volume of the space of distributions for which $(x - pq)(z - pq) \geq 0$: a simple geometric argument shows that $Dvol = D'$. The $Mvol$ measure is based on Mahalanobis distance between the specific distribution represented by z and the independence one. As in the 2x2 case this is equal to $\frac{(x-pq)^2}{pq(1-p)(1-q)}$, the ratio of the blue lines in Figure 1 identify $Mvol$. To further clarify the differences between the considered measures, we illustrated in Figure 2 a case where $p = (1 - q)$ and $z = 0$.

The definition of $Mvol$ can be easily extended to the case of multiallelic markers, even if in this case the evaluation of the volumes is not easy. The definition of $Dvol$, with its consideration of the sign of $(x - pq)$ is tied to the case of 2x2 tables. Later we will introduce another volume measure based on excess of heterozygosity and show how this can be considered an extension of $|R|$ to multiallelic case, in the sense that it focuses on the predicting power of alleles at one marker for alleles at the other. At this point, however, we would like to introduce one fundamental characteristic of volume measures that makes them particularly attractive in our eyes and allows also to overcome the difficulties presented by their analytical evaluation. So far, we have described measures as defined on the population distribution of haplotypes. This is, however, rarely known and it is typically approximated with a sample distribution. Measures as D' and $|R|$ are then evaluated on the sampling distribution as if this was the true one. This clearly results in variability of the estimated measures, even this has often been overlooked. To take into account of this variability, researchers have sometimes calculated confidence intervals for D' based on resampling procedures. This is certainly appreciable, but it adds to the computational burden and, the results (a measure, and its confidence bounds) do not easily lead themselves to the visual displays that have become customary in the field.

Volume measures can be defined directly on the observed contingency tables of sample haplotypes. Instead of calculating the volumes of spaced of probability distributions, one counts the number of contingency tables that have the same marginal totals and satisfy other characteristics of interest. This definition of volume measures corresponds more directly to the one of volume tests introduced by Hotelling (1939) and analyzed by Diaconis and Efron (1985). Indeed defined in this way, volume measures have a test-like property in that they evaluate how strongly the observed table differs from what one would expect under independence, considering the number of observations available from the population distribution. We will devote the following sections to the explicit definition of these volume measures and the illustration of a sequential importance sampling algorithm that makes it possible to evaluate them efficiently.

2 Finite sample evaluation of $Dvol$ and $Mvol$

Consider the table of observed haplotypes counts T:

$$T = \begin{array}{c|c|c|c|c|c}
 & B_1 & B_2 & \cdots & B_c & \\ \hline
A_1 & n_{11} & n_{12} & \cdots & n_{1c} & n_{1\cdot} \\ \hline
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline
A_r & n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r\cdot} \\ \hline
 & n_{\cdot 1} & n_{\cdot 2} & \cdots & n_{\cdot c} & n
\end{array}$$

where B_i represent alleles at marker B and A_i alleles at marker A.

When $c = r = 2$, let Ω_1 denote the set of all contingency tables with the same row and column sums as the observed table and the same sign of $(n_{11} - n_{1\cdot}n_{\cdot 1}/n)$ as in the observed table. We then define $Dvol$ as

$$Dvol = \frac{1}{|\Omega_1|} \sum_{T' \in \Omega_1} 1_{\{M(T') < M(T)\}}, \tag{2}$$

where $M(T)$ is defined as

$$\sum_{i,j} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}.$$

For general $r \times c$ tables, let Ω_2 denote the set of all contingency tables with the same row and column sums as the observed table. We define

$$Mvol = \frac{1}{|\Omega_2|} \sum_{T' \in \Omega_2} 1_{\{M(T') < M(T)\}}. \tag{3}$$

Note that the above definitions use the strict inequality sign. The choice of this over $\leq$ is irrelevant for large n, but it makes a difference in the case of small n, where strict inequality allows us to better discriminate against apparent association due to small sample.

3 A volume measure based on excess of homozygosity

Given the framework we presented in the previous section to evaluate volume measures, it is clear that one can apply the same procedure for a variety of definitions of discrepancies, substituting the Mahalanobis distance M with other that may be of interest. In particular we would like to point the attention to the excess of homozygosity H:

$$H(T) = \sum_{i,j} n_{ij}^2 - \sum_i n_{i\cdot}^2 \sum_j n_{\cdot j}^2 / n^2,$$

that leads to the definition of $Hvol$:

$$Hvol = \text{sign}(H(T)) \frac{\sum_{T' \in \Omega_2} 1_{\{|H(T')| < |H(T)|\}} 1_{\{H(T)H(T') \geq 0\}}}{\sum_{T' \in \Omega_2} 1_{\{H(T)H(T') \geq 0\}}}. \tag{4}$$

The relation between excess of homozygosity and LD has been explored extensively in Sabatti and Risch [2002]. In particular, it is worth recalling that joint homozygosity at two markers is a measure of agreement between the two markers and hence it provides a better sense of how the alleles at one can be used to predict alleles at the other than a measure as $Mvol$ might. Indeed, despite all its limitations as a measure of disequilibrium, $Hvol$ carries some of the information in $|R|$ that is absent from D'. Since $Hvol$ is not limited to 2x2 tables, as R, it can be effectively used to explore the relation between markers that are multiallelic from this viewpoint. Figures 3 and 4 illustrate this point using approximately 2500 haplotype distributions obtained by looking at adjacent SNPs typed on chr 22, in 200 individuals from Costa Rica (data from the study by Service et al. [2006]). In panels (a) the relation between $|R|$ and D' (Figure 3) and $Mvol$ (Figure 4) is illustrated: clearly there are a number of tables with high D' or $Mvol$ values, that have low $|R|$. In this display red circles correspond to tables that had an excess of heterozygosity and green circles to tables with excess homozygosity: clearly excess heterozygosity correlates with low values of $|R|$. The positive correlation between $|R|$ and $Hvol$ is explicitely illustrated in panels (b), while panels (c) indicate that the information in $Hvol$ is largely orthogonal to the one contained in D' or $Mvol$.

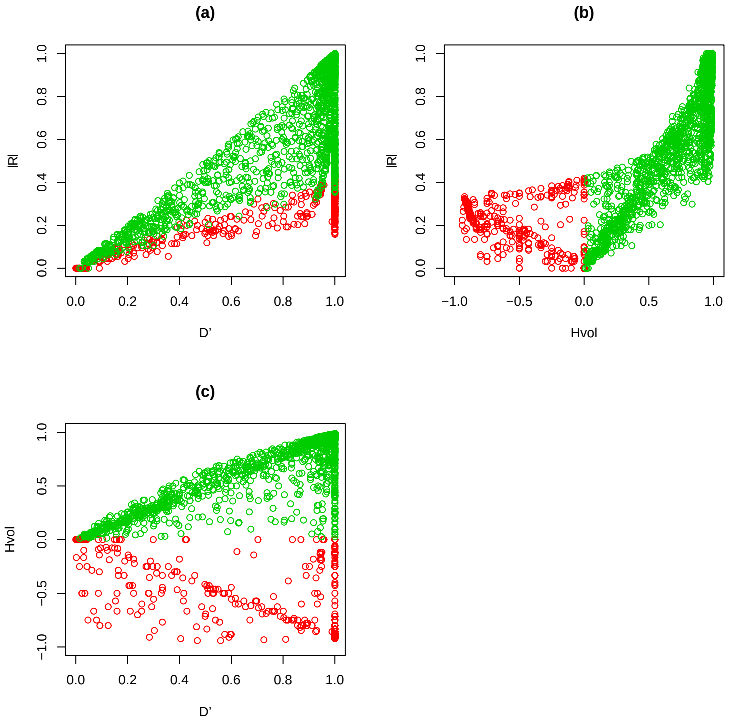

Figure 3: Illustration of relationships among D', $|R|$, and $Hvol$. We are plotting $|R|$ rather than R^2, as its scale is directly comparable to the one of D'. Panel (a), (b), (c) respectively show the relationships between D' and $|R|$, $Hvol$ and $|R|$, D' and $Hvol$. Each point corresponds to a SNP pair. The red circles represent tables with excess heterozygosity, i.e. $Hvol \leq 0$, and green circles represent tables with excess homozygosity, i.e. $Hvol > 0$.

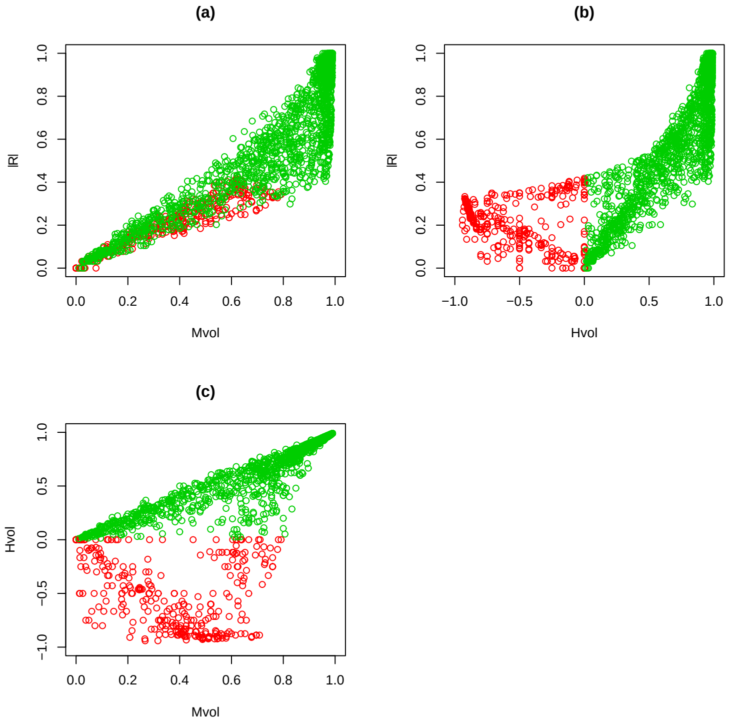

Figure 4: Illustration of relationships among $Mvol$, $|R|$, and $Hvol$. Panels (a), (b), (c) respectively show the relationships between $Mvol$ and $|R|$, $Hvol$ and $|R|$, $Mvol$ and $Hvol$. Each point corresponds to a SNPs pair. The red circles represent tables with excess heterozygosity, i.e. $Hvol \leq 0$, and green circles represent tables with excess homozygosity, i.e. $Hvol > 0$.

Because $Hvol$ captures part of the information on predictability contained in $|R|$ and can be defined on tables of any size, we suggest calculating it as complementary to $Mvol$.

4 Algorithms for the evaluation of volume measures

To evaluate these measures defined in the previous section, we need to explore the space of all tables with the the same margins. In the case of 2x2 tables, as those involved in Figures 3 and 4, this can be done by simple enumeration. With particular reference to the case of $Dvol$, that is defined only on 2x2 tables, to enumerate all tables in Ω_1, it is useful to notice that n_{11} must satisfy

$$\max(0, n_{\cdot 1} - n_{2\cdot}) \leq n_{11} \leq \min(n_{1\cdot}, n_{\cdot 1}), \tag{5}$$

and after n_{11} is chosen, we can fill in other entries of the 2×2 table by the marginal sum constraints. Therefore we can enumerate tables in Ω_1 by assigning all possible integers satisfying (5) to n_{11}, and keeping those tables such that $(n_{11} - n_{1\cdot} n_{\cdot 1}/n)$ has the same sign as the observed table.

When we consider the case of generic $r \times c$ tables, exhaustive enumeration is going to be too time consuming. For this purpose, we have successfully implemented a sequential importance sampling (SIS) algorithm originally introduced in Chen *et al.* [2005]. Following is a brief description of the algorithms implemented in our C code, which is available at the following url: `http://www.stat.uiuc.edu/˜yuguo/software/volume/`.

Let $u(T')$ be the uniform distribution over all tables in Ω_2. Then $Mvol$ can be treated as the expectation of the indicator function $1_{\{M(T')<M(T)\}}$ with respect to $u(T')$. It is hard to sample directly from $u(T')$. The idea of importance sampling is to sample tables from another proposal distribution $g(T')$, and then estimate $Mvol$ by

$$\frac{\sum_{i=1}^{N} 1_{\{M(T'_i)<M(T)\}} \frac{u(T'_i)}{g(T'_i)}}{\sum_{i=1}^{N} \frac{u(T'_i)}{g(T'_i)}} = \frac{\sum_{i=1}^{N} 1_{\{M(T'_i)<M(T)\}} \frac{1}{g(T'_i)}}{\sum_{i=1}^{N} \frac{1}{g(T'_i)}}, \tag{6}$$

where $T'_1, \ldots, T'_N$ are N independent and identically distributed (i.i.d.) samples from $g(T')$. SIS generates a table cell by cell by decomposing the proposal distribution $g(T')$ as

$$g(T') = g(n_{11})g(n_{21}|n_{11}) \cdots g(n_{rc}|n_{r-1,c}, \ldots, n_{11}). \tag{7}$$

Notice that the support for the first entry n_{11} is

$$\max(0, n_{\cdot 1} + n_{1\cdot} - n) \leq n_{11} \leq \min(n_{1\cdot}, n_{\cdot 1}).$$

We sample an integer uniformly from the above range for n_{11}, i.e., $g(n_{11})$ is the uniform distribution on the support of n_{11}. Recursively, suppose we have chosen $n_{i1} = n_{i1}^*$ for $i = 1, \ldots, k-1$. Then the support for n_{k1} is $\max\Big(0, \Big(n_{\cdot 1} - \sum_{i=1}^{k-1} n_{i1}^*\Big) - \sum_{i=k+1}^{r} n_{i\cdot}\Big) \leq n_{k1} \leq \min\Big(n_{k\cdot}, n_{\cdot 1} - \sum_{i=1}^{k-1} n_{i1}^*\Big)$. We sample an integer uniformly from the above range for n_{k1}. The procedure is continued until all the entries in the first column have been considered. Then we update the row sums by subtracting the realization of the first column from the original row sum, and sample the second column of the table in the same way.

5 Examples

5.1 The effect of small sample size on D' and $Dvol$

It has been noticed that D' tends to be biased upwards in small samples [Teare *et al.*, 2002; Tenesa *et al.*, 2004]. We conducted a simulation study to illustrate how this problem is less severe when using $Dvol$. We generated 100 two-markers haplotype tables with 200 observations, each under the hypothesis of linkage equilibrium between the markers. The distribution of the frequency of the minor alleles of the simulated SNPs matched a random sample of markers on chromosome 22, that were used in Service et al. [2006]. In a situation where the true population value of D' is equal to zero, any sample based estimator is going to be upward biased, since 0 is the minimum value that D' can achieve. The point of our investigation was to compare the severity

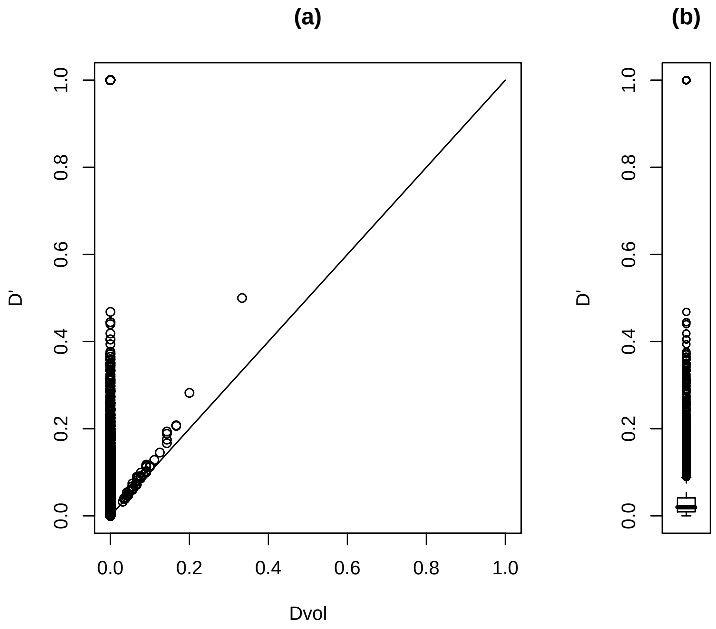

Figure 5: Comparison of D' and $Dvol$ on tables generated under linkage equilibrium. (a) scatterplot of the values of D' and $Dvol$. (b) Boxplot of the values of D'.

of this bias. Figure 5 illustrates the results: D' is always larger than $Dvol$, and it is occasionally equal to 1; $Dvol$ is actually equal to zero in the majority of cases.

5.2 Patterns of LD between multiallelic markers

Our next example focuses on the application of volume measures to multiallelic markers. The data consists in 157 phase-known non-transmitted chromosomes 2 of parents of BP-I persons from the Central Valley of Costa Rica. The chromosomes were typed with 85 markers in the course of the study by Ophoff *et al.* [2002]. Using volume measures $Mvol$ and $Hvol$ we were able to evaluate the level of disequilibrium between all the possible marker pairs in this sample. Figure 6 gives a graphical representation of the values of $Mvol$ and $Hvol$ in this data set as well as the

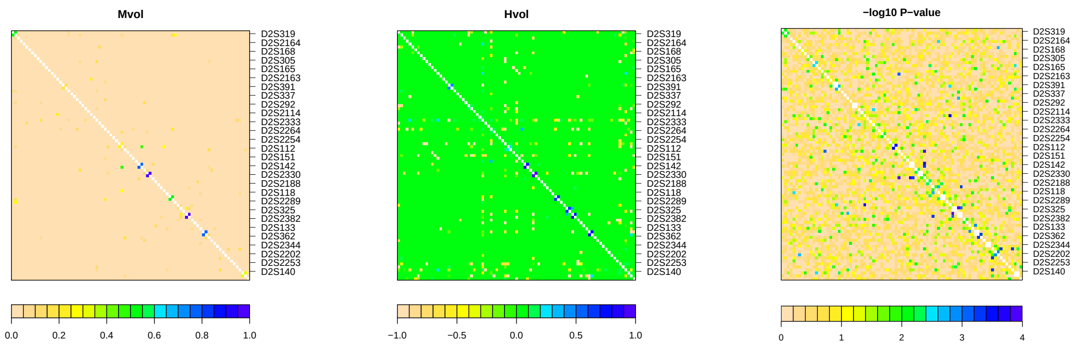

Figure 6: Measure of disequilibrium between microsatellites. Each square in this symmetric picture corresponds to a marker pair (the same markers are reported on both rows and columns). The three panels report, from left to right, $Mvol$, $Hvol$, and the negative of the log10 of the p-value for a Fisher exact test of independence.

negative of log10 p-value for a Fisher exact test of independence. This last one is reported for comparison purposes, as it is often used as a measure of dependence, despite the fact that it is rather inappropriate with this goal [Diaconis and Efron, 1985].

5.3 Consistency of LD patterns on chr 22 in 12 populations

We have used the measures D', $|R|$, $Dvol$, $Mvol$ and $Hvol$ to assess the distribution and extent of linkage disequilibrium on chromosome 22 in samples of 200 persons from each of eleven population isolates and in an out-bred Caucasian sample, using 2486 SNP markers spaced at a density of approximately one marker every 13.8 kb. [Service et al., 2006; Wang *et al.*, 2006]. To conduct a complete analysis of the linkage disequilibrium patterns in the 12 population samples, we restricted our attention to the SNPs with sample minor allele frequencies larger then 0.1. We did so for uniformity with previous studies (for example, Hinds et al. [2005]) and to make sure that our results were not strongly influenced by the rare markers with exceptionally high homozygosity. This leads us to work with 1920 SNPs. Five measures, D, $Dvol$, $Mvol$, R^2 and $Hvol$ are calculated for each of the 1,842,240 pairs of SNPs. The results were summarized averaging the measured disequilibrium within windows of 1.7 Mb sliding along chromosome 22. Figure 7 reports the values of the five measures in the Costa Rica population. The observed relation between the measures is consistent across populations. In particular it can be noted that the average values of $Dvol$ are lower than the ones of D', while clearly exhibiting very similar patterns. This testifies that even if the sample size is moderately large (200 individuals) and only markers with MAF>0.1 are considered, D' is inflated. Turning now our attention to $Hvol$, it can be seen that its values are much closer to the R^2 ones.

Aknowledgements

We thank Nelson Freimer and his lab for making their data available to us. We also thank Hui Wang for help in drawing figures. Y. Chen is partially supported by NSF grant DMS-0503981. C. Sabatti acknowledges support from NSF grant DMS0239427, NIH grants R01NS037484 and R01MH049499, and USPHS grant GM53275.

References

Chen, Y., Diaconis, P., Holmes, S., and Liu, J.S. (2005) Sequential Monte Carlo Methods for Statistical Analysis of Tables. *Journal of the American Statistical Association*, **100**, 109–120.

Devlin, B. and Risch, N. (1995) A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics*, **29**, 311–322.

Diaconis, P. and Efron, B. (1985)Testing for Independence in a Two-Way Table: New Interpretations of the Chi-Square Statistics. *The Annals of Statistics*, **13**, 845–874.

Diaconis, P., and Sturmfels, B. (1997) Algebraic Algorithms for Sampling from Conditional Distributions. *The Annals of Statistics*, **26**, 363–397.

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M., and Altshuler, D. (2002) The Structure of Haplotype Blocks in the Human Genome. *Science*, **21**, 2225–2229.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: Whole-genome patterns of common DNA variation in three human populations. Science 2005; 307:1072-1079.

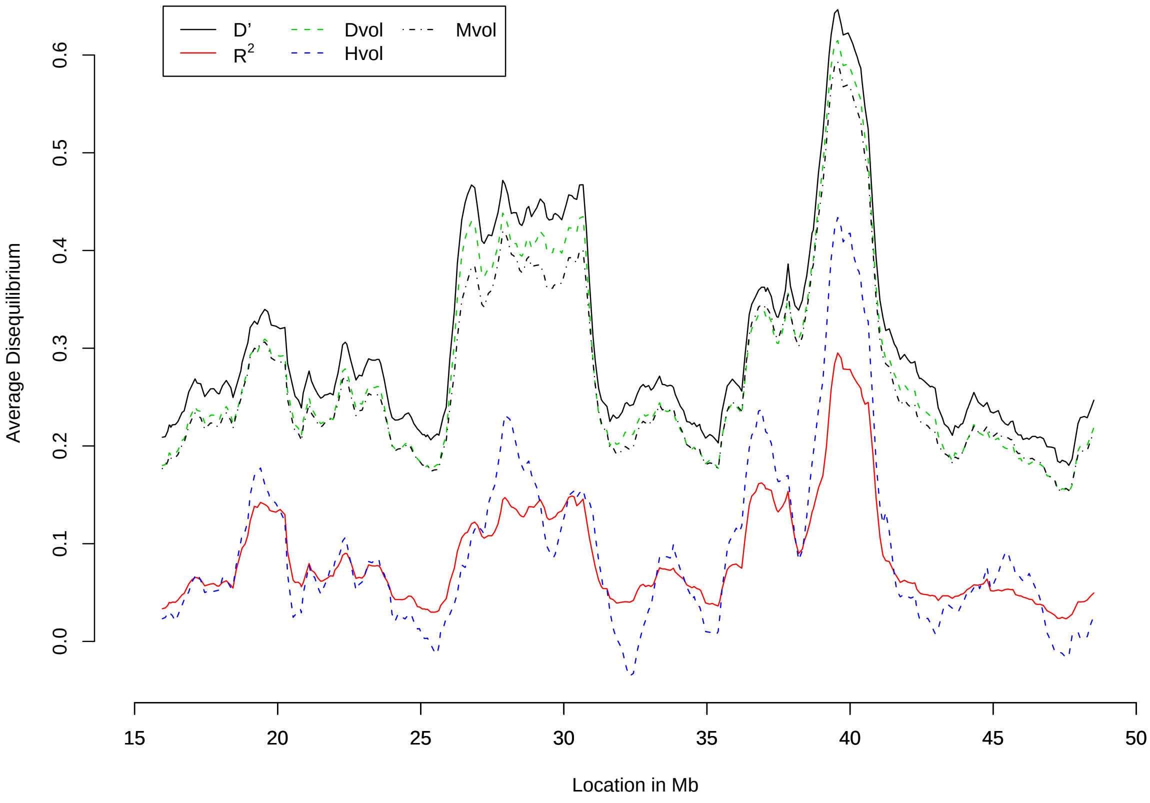

Figure 7: Linkage disequilibrium of chromosome 22 in Costa Rica according to five different measures. D', R^2,$Dvol$, $Mvol$ and $Hvol$ are represented, respectively, with a solid black, as solid red, a broken red, broken red, and a broken blue line. The average value of the measures, between markers that are within a 1.7 Mb window, is plotted against the middle point of the window, with the x axis representing the length of chromosome 22.

Hotelling, H. (1939) Tubes and Spheres in n-spaces, and a Class of Statistical Problems. *American Journal of Mathematics*, **61**, 440–460.

Ophoff, R., Escamilla, M., Service, S., Spesny, M., Meshi, D., Poon,W., Molina, J., Fournier, E., Gallegos, A., Mathews, C., Neylan, T., Batki, S., Roche, E., Ramirez, M., Silva, S., De Mille, M., Dong, P., Leon, P., Reus, V., Sandkuijl, L., and Freimer, N. (2002) Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate. *Am J Hum Genet* **71**, 565–74.

Pritchard, J. K., and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Gen.*, **68**, 1–14.

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

Sabatti, C. (2002) Measuring dependence with volume tests, *The American Statistician* **50**, 191–195.

Service, S., DeYoung, J., Karayiorgou, M., Louw Roos, J., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J., Heutink, P., Aulchenko, Y., Oostra, B., van Duijn, C., Jarvelin, M., Varilo, T., Peddle, L., Rahman, P., Piras, G., Monne, M., Murray, S., Galver, L., Peltonen, L., Sabatti, C., Collins, A., Freimer, N. (2006) Extent, distribution and magnitude of linkage disequilibrium in eleven population isolates. *Nature Genetics* published on line April 2, 2006.

Teare, M., Dunning, A., Durocher, F., Rennart, G., and Easton, D. F. (2002) Sampling distribution of summary linkage disequilibrium measures. *Ann. Hum. Genet.* **66**, 223–233.

Tenesa, A. *et al.* (2004) Extent of linkage disequilbirum in a Sardinian sub-isolate: sampling and methodological considerations. *Hum. Mole. Genet.* **13**, 25–33.

Wang, H., Lin, C., Service, S., Chen,Y., Freimer, N., Sabatti, C. (2006) Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density, *Submitted*.