Genetic Epidemiology 22:186–195 (2002)

# Haplotype Frequency Estimation in Patient Populations: The Effect of Departures From Hardy-Weinberg Proportions and Collapsing Over a Locus in the HLA Region

**Richard M. Single,[1,2]* Diogo Meyer,[1] Jill A. Hollenbach,[1] Mark P. Nelson,[1] Janelle A. Noble,[3] Henry A. Erlich,[3] and Glenys Thomson[1]**

*[1]Department of Integrative Biology, University of California, Berkeley*
*[2]Department of Medical Biostatistics, University of Vermont, Burlington*
*[3]Children's Hospital Oakland Research Institute, Oakland; and Department of Human Genetics, Roche Molecular Systems, Alameda, California*

Haplotype analyses are an important area in the study of the genetic components of human disease. Associations between markers and disease loci that are not evident with a single marker locus may be identified in multi-locus marker analyses using estimated haplotype frequencies (HFs). Procedures that make use of the expectation-maximization (EM) algorithm to estimate HFs from unphased genotype data are in common use in genetic studies. The EM algorithm uses these unphased genotype frequencies along with the assumption of Hardy-Weinberg proportions (HWP) to converge on HF estimates. In this paper, we assess the accuracy of EM estimates of HFs in patients with type I diabetes for whom the true haplotypes are known, but the data are analyzed ignoring family information to allow comparison between estimated and true frequencies. The data consist of six HLA loci with high levels of polymorphism and a range of departures from HWP and linkage equilibrium. While the overall accuracy of the EM estimates is good, there can be large over- and underestimates of particular HFs, even for common haplotypes, especially when the loci involved deviate significantly from HWP. Estimating HFs for three or more loci and then collapsing over loci so as to generate two locus haplotypes can improve the accuracy of the estimation. The collapsing procedure is most beneficial when one of the loci in the two-locus haplotype of interest deviates significantly from HWP and the locus collapsed over is in linkage disequilibrium with the other loci. Genet. Epidemiol. 22:186–195, 2002. 

**Key words: expectation-maximization algorithm; estimation accuracy; Hardy-Weinberg proportions; linkage disequilibrium**

Contract grant sponsor: National Institutes of Health; Contract grants: GM35326, CA84497 (R.M.S., DM., J.A.H., G.T.), and DK46626 (J.AN., H.A.E.); Contract grant sponsor: American Diabetes Association, Career Development Award (J.A.N.).
*Corresponding to: Richard M. Single, Medical Biostatistics, University of Vermont, Hills Science Building, Burlington, VT 05405-0082. E-mail: single@allele5.biol.berkeley.edu
Received for publication 25 June 2001; revision accepted 1 October 2001


**DOI 10.1002/gepi.0163**

## INTRODUCTION

Genetic data, in the form of multi-locus genotypes, are often compatible with many haplotypic arrangements when linkage phase is unknown. The accuracy of haplotype frequency (HF) estimates is of increasing interest with regard to population and disease association studies. Marker-disease associations that are not detected with a single marker locus may be detected in a multi-locus marker analysis [Barcellos et al., 1997; Valdes et al., 1999a]. Even when single-locus marker associations with disease are found, haplotype analyses may greatly aid localization of the genetic region involved in disease [Johnson and Todd, 2000]. Linkage disequilibrium (LD) patterns also influence our ability to identify the actual disease predisposing/protective variants and determine whether additional genetic factors in a region are involved in disease [Valdes and Thomson, 1997].

The standard programs for HF estimation when family data are not available, such as HAUPT [Baur and Danilovs, 1980], ARLEQUIN [Schneider et al., 2000], HAPLO [Hawley and Kidd, 1995], and 3LOCUS.PAS [Long et al., 1995], employ an iterative expectation-maximization (EM) algorithm and the assumption of Hardy-Weinberg proportions (HWP) to estimate HFs [Dempster et al., 1977]. Application of an EM algorithm with the assumption of HWP could lead to inaccuracies because for all disease models except recessive, the genotype frequencies at a marker locus in LD with the disease locus are not expected to fit HWP [Thomson, 1995a].

In this study, we use a family-based data set in which the true HFs are known. We analyze the data ignoring the family information so estimated HFs can be directly compared with the real values. Our work extends that of others. For instance, Schipper et al. [1998] compared three methods for estimating A-B HLA HFs using data in which neither locus deviated significantly from HWP. Tishkoff et al. [2000] studied frequency estimation for two-locus haplotypes at the CD4 locus. Both polymorphisms showed non-significant deviation from HWP. Fallin and Schork [2000] studied haplotype estimation for biallelic variation in a population context. The authors simulated data with varying numbers of loci, allele frequencies, and degrees of LD and departure from HWP. Our study involves multi-allelic variation at six HLA loci for patient data. As well as considering the effects of LD and HWP on haplotype estimation, we consider effects of sample size by analyzing subsets of the data.

When more than two marker loci are available, two-locus haplotypes can be estimated directly or by summing over the appropriate HFs from a multi-locus estimation. We refer to the latter procedure as collapsing over a third locus. Several estimation programs estimate two-locus HFs by collapsing over estimates of three-locus HFs (e.g., HAUPT and 3LOCUS.PAS). Other programs utilize two-locus HFs obtained by collapsing over a third locus to compute statistics such as LD values (e.g., HAPLO). Still other programs do not utilize higher order HF estimates when estimating two-locus HFs, even when these frequency estimates are available (e.g., ARLEQUIN). In this paper, we assess the circumstances under which it is beneficial to estimate two-locus HFs by collapsing over a third locus.

## SUBJECTS

The Human Biological Data Interchange is a repository for cell lines of nuclear families with unaffected parents and at least two affected siblings with type 1 diabetes (for further information, see www.hbdi.org). Genes in the HLA region (termed *IDDM1*) contribute the major gene effect for this complex disease [Todd, 1995]. A total of 283 Caucasian families

has been typed for six antigen-presenting HLA genes (listed in order of their map location): DPB1, DQB1, DQA1, DRB1, B, and A [Valdes et al., 1999b; Noble et al., 2000].

As in Schipper et al. [1998], families in which there was a recombination event among the affected sibs were not considered in the analyses. However, inclusion of the 29 recombinant families does not alter the major conclusions of this study. From the remaining 254 families, the oldest affected sib was chosen for analysis so that estimation could be done in a sample (the "full sample") of 254 unrelated individuals for whom the haplotypic phase of the six HLA genes could be assigned unambiguously.

In the full sample of 254 oldest affected sibs, the three most frequent DR-DQ haplotypes in patients, all of which are high risk, are the DR3 haplotype (DRB1*0301 DQA1*0501 DQB1*02) and two DR4 haplotypes, called DR4A and DR4B (DRB1*0401 and DRB1*0404 with DQA1*03 DQB1*0302) with respective frequencies 31.6, 24.4, and 8.7% in patients, and 10.0, 3.2, and 2.9% in the AFBAC (affected family-based controls) sample of parental haplotypes never transmitted to the affected sib pair [Thomson, 1995b]. The AFBAC sample fills the role of a control sample. As seen in many other data sets, DR3/DR4 heterozygotes with the high-risk DR3 and DR4 DR-DQ haplotypes listed above show increased risk of disease compared with the respective homozygotes. This results in a significant deviation from HWP in the patient population in the direction of excess heterozygosity (Table I). One goal of this study was to document the robustness of HF estimation to these deviations from HWP.

## METHODS

Departure from HWP for individual loci was tested for statistical significance using the exact test of Guo and Thompson [1992] in the full sample. The DQB1, DQA1, and DRB1 loci showed significant deviation from HWP, in the form of excess heterozygosity (Table I). Global LD, summing contributions of all the haplotypes in a multi-allelic two-locus system, was measured with Hedrick's [1987] $D'_L$ statistic, using the products of allele frequencies at the loci $p_i$ and $q_j$ as weights:

$$D'_L = \sum\sum p_i q_j \left| D'_{ij} \right|,$$

where $D'_{ij} = D_{ij}/D_{max}$. The coefficients of LD for the full sample are listed in Table I.

**TABLE I. HWP and Overall Linkage Disequilibrium in the Full Sample**

| Locus name | No. of alleles | HWP[a] | Overall linkage disequilibrium[b] | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | DPB1 | DQB1 | DQA1 | DRB1 | B | A |
| DPB1 | 20 | 0.88 | — | | | | | |
| DQB1 | 15 | <0.001 | 0.328 | — | | | | |
| DQA1 | 7 | <0.001 | 0.296 | 0.946 | — | | | |
| DRB1 | 27 | <0.001 | 0.408 | 0.963 | 0.997 | — | | |
| B | 40 | 0.38 | 0.417 | 0.633 | 0.626 | 0.680 | — | |
| A | 24 | 0.30 | 0.291 | 0.376 | 0.368 | 0.435 | 0.566 | — |

[a]The *P*-value for the exact test of HWP [Guo and Thomson, 1992] was computed for the full sample of 254 oldest affected sibs.
[b]$D'_L$ was computed using known haplotype frequencies for the full sample of 254 oldest affected sibs. Loci are listed in map order.

True HFs were computed by dividing the total number of occurrences of each haplotype by 2*n*, where *n* is the number of individuals in the sample. Estimated HFs, when phase information was ignored, were obtained using the EM algorithm as implemented by ARLEQUIN. Multiple (200) starting conditions were used to minimize the possibility of local maxima being reached by EM iterations using a convergence criterion of $10^{-7}$ [Excoffier and Slatkin, 1995; Long et al., 1995].

The accuracy of HF estimates was assessed using measures that compare the true HFs, $h_i$, and the frequencies estimated by the EM algorithm, $\hat{h}_i$.

$$I_f = 1 - \frac{1}{2}\sum\left|\hat{h}_i - h_i\right|$$

is a similarity index, where summation is over true and estimated haplotypes, that takes a value of one if the true and estimated frequencies are equal.

$$I_h = \frac{2(n_{true} - n_{missed})}{n_{true} + n_{estimated}}$$

is a measure of haplotype identification, where $n_{true}$ represents the true number of different haplotypes present, $n_{estimated}$ is the number of distinct haplotypes identified in the estimation, and $n_{missed}$ is the number of true haplotypes not identified in the estimation. It takes a value of 1 if the set of estimated haplotypes is identical to the set of true haplotypes in the sample. For the purposes of this statistic, a haplotype is considered to have been identified if the estimated HF is at least as large as the threshold value of 1/(2*n*) [Excoffier and Slatkin, 1995].

Two-locus HF estimation was assessed in the full sample. Generally, better estimates were obtained for locus pairs that had higher LD. For those locus pairs with the highest levels of LD, the estimation was nearly perfect. Due to the high level of LD among the DQB1, DQA1, and DRB1 loci, two-locus HFs for pairs of these loci were estimated very accurately and not considered for further analysis. Detailed analysis was done for the estimation of *-A haplotypes (* indicates all other loci) in subsamples from the full sample. Three-locus HF estimates were collapsed over a third locus to provide a new set of two-locus HF estimates for the same pair of loci. The accuracy of these estimates was then compared with that of the original two-locus estimates. These analyses were replicated 100 times for samples of size $n = 50$ and $n = 100$. Each of the 100 replicates involved a separate subsample without replacement from the full sample.

## RESULTS

### Collapsing Over a Locus

Figure 1 provides a graphic example for one of the 100 replications with $n = 50$ of the true and estimated DQB1-A HFs and the effect of collapsing over different loci to estimate the DQB1-A haplotypes. The percentage of replications in which each locus was identified as the best to collapse over was determined using $I_f$ as a measure. For each sample size and haplotype combination that did not include the B locus, the B locus was identified with the greatest frequency as the best locus to collapse over. Collapsing over the B locus led to the highest value of $I_f$ in 39, 62, 48, and 55% of the replications for estimation of DPB1-A, DQB1-A, DQA1-A, and DRB1-A HFs, respectively, for $n = 50$. The corresponding percentages for $n = 100$ were 48, 76, 66, and 74%, respectively. For the B-A haplotypes, in which neither locus deviates signifi-

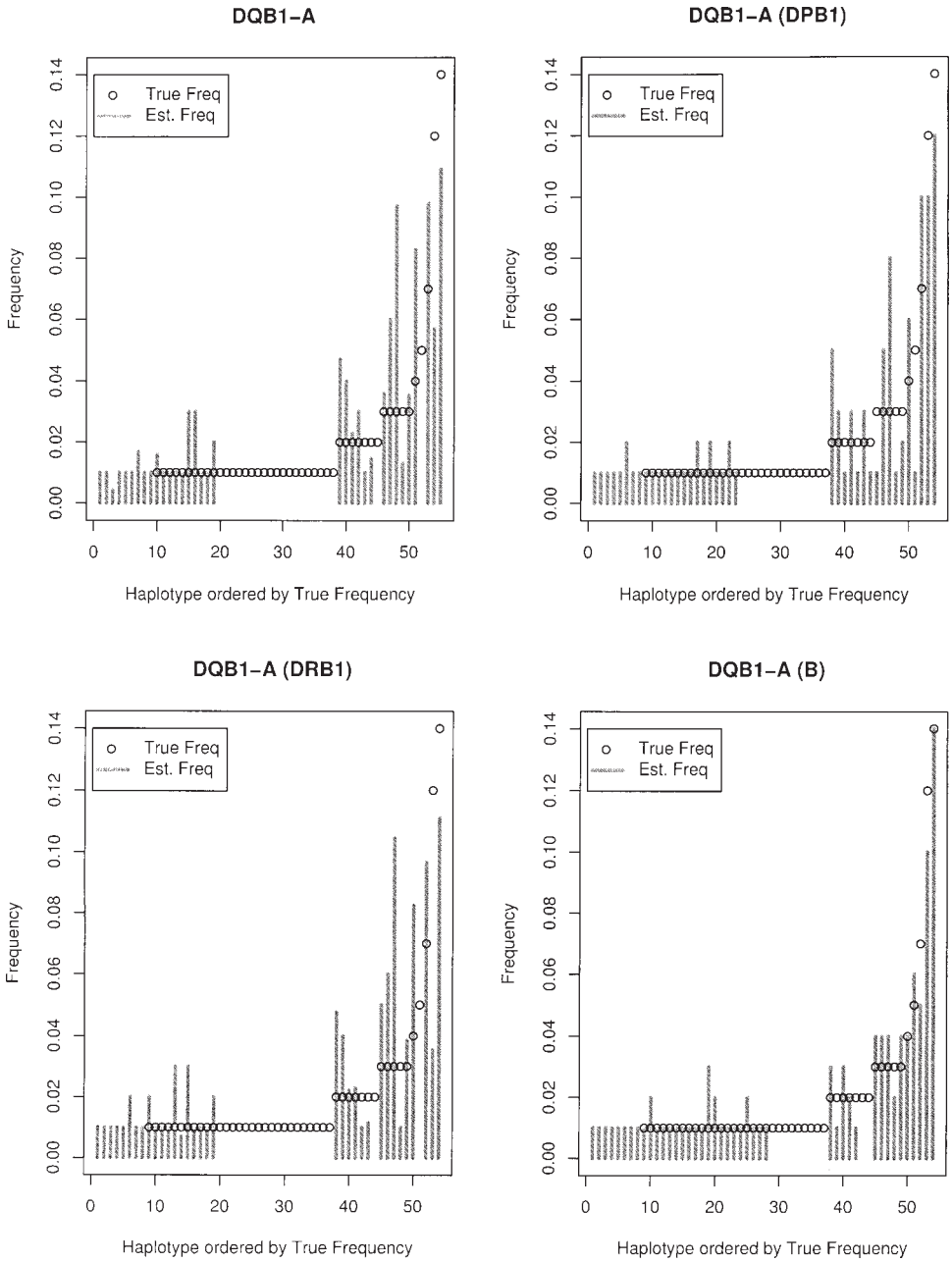


Fig. 1. Plots of true and estimated frequencies for DQB1-A haplotypes based on two- and three-locus estimation collapsed over the third locus in a sample of size 50. In each graph, the same set of true HFs is plotted along with the estimated HFs for that scenario. The graph in the upper left corner shows results from the two-locus HF estimation ($I_f$ = 0.61). In clockwise order, the results are displayed for three-locus estimation collapsed over the DPB1 locus ($I_f$ = 0.70), B locus ($I_f$ = 0.80), and DRB1 locus ($I_f$ = 0.60). Here the B locus is seen to be the best to collapse over. Collapsing over a locus refers to the process of estimating three-locus haplotype frequencies and summing over the appropriate haplotype frequencies to obtain two-locus haplotype frequencies.

cantly from HWP, the choice of a best locus to collapse over was less definitive with no single locus selected in more than 30% of the replications.

Since the above results indicate that the B locus is most frequently the best to collapse over, further results focus on three-locus estimation collapsed over this locus. Columns three to six of Table II contain summary statistics on the performance indices for both two- and three-locus estimation collapsing over the B locus. Averages were computed over the 100 replicates. For $n = 50$, the average value of $I_f$ was between 2.7 and 9.5% higher when HFs were computed by collapsing over the B locus. Collapsing over the B locus led to an average increase in $I_h$ of between 5.6 and 11.6% for $n = 50$. The increase in accuracy due to collapsing over the B locus was smallest for the DPB1-A haplotypes, in which neither locus deviates from HWP in the full sample. Similar results were seen for $n = 100$ (Table II).

## Effect of HWP on the Accuracy of Estimation

As seen in the graph of the two-locus estimation of Fig. 1, individual HFs can be largely over- or underestimated, even for common haplotypes. To identify the effect of deviations from HWP on estimation accuracy, samples were created using a shuffling procedure in which the true haplotypes were shuffled among individuals. For each replicate sample, a corresponding shuffled version was generated, resulting in a new set of genotypes with the same haplotypic composition. Shuffling the haplotypes maintains the original HFs and acts as a surrogate for the removal of significant deviation from HWP. The last four columns of Table II list performance measures in the shuffled samples. The average value of $I_f$ for two-locus (three-locus collapsed) estimation was higher by between 0% (0%) and 9.5% (3.8%) due to shuffling for $n = 50$. Shuffling led to an average increase in $I_h$ of between 1.4% (1.3%) and 11.6% (5.2%) for $n = 50$. Again, the increase in accuracy was smallest for DPB1-A haplotypes.

To measure the effect of deviations from HWP on the occurrence of large over- and underestimates of HFs, we defined a new statistic. Maxdiff = $\max |\hat{h}_i - h_i|$ is the maximum over all haplotypes (true and estimated) for a locus pair of the absolute value of the difference between the true and estimated frequencies. Since the observed and shuffled samples

**TABLE II. Mean Value of Performance Indices for Two-Locus and Collapsed Three-Locus Haplotype Frequency Estimation**

| | | Observed data | | | | Shuffled data[a] | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $I_h$ | | $I_f$ | | $I_h$ | | $I_f$ | |
| $n$ | Haplotypes | 2-Locus | collapsed[b] | 2-Locus | collapsed[b] | 2-Locus | collapsed[b] | 2-Locus | collapsed[b] |
| 50 | DPB1-A | 0.71 | 0.75 | 0.75 | 0.77 | 0.72 | 0.76 | 0.75 | 0.77 |
| | DQB1-A | 0.69 | 0.77 | 0.74 | 0.81 | 0.77 | 0.81 | 0.81 | 0.84 |
| | DQA1-A | 0.75 | 0.81 | 0.77 | 0.82 | 0.79 | 0.83 | 0.82 | 0.85 |
| | DRB1-A | 0.71 | 0.76 | 0.76 | 0.79 | 0.74 | 0.78 | 0.79 | 0.82 |
| 100 | DPB1-A | 0.74 | 0.78 | 0.80 | 0.82 | 0.74 | 0.78 | 0.79 | 0.80 |
| | DQB1-A | 0.70 | 0.80 | 0.77 | 0.84 | 0.80 | 0.85 | 0.85 | 0.87 |
| | DQA1-A | 0.76 | 0.83 | 0.82 | 0.86 | 0.82 | 0.87 | 0.87 | 0.88 |
| | DRB1-A | 0.73 | 0.79 | 0.79 | 0.83 | 0.78 | 0.81 | 0.83 | 0.85 |

[a]Haplotypes in each of the 100 replicate samples for each haplotype/sample size combination were shuffled to create new genotypes with no expected deviation from HWP.

[b]Three-locus haplotype frequencies were estimated using the two loci and the B locus and then collapsed over the B locus to yield estimates of the two-locus haplotype frequencies. Means were calculated over the 100 replicate samples of size either 50 or 100, with random sampling from the 254 oldest affected sibs.

are paired, the difference in maxdiff values between each observed and shuffled replicate pair was considered. When one locus deviates significantly from HWP, the improvement (decrease) in the maxdiff statistic due to shuffling was significant, as seen in the three middle boxplots in Fig. 2. When neither locus deviates from HWP, the decrease in the maxdiff statistic was not significant with boxplots centered near zero. The decrease (percentage decrease) in the maxdiff statistic due to shuffling, averaged over the 100 replicates, was 0.016 (24.4%), 0.018 (26.4%), and 0.009 (11.5%), respectively for DQB1-A, DQA1-A, and DRB1-A haplotype estimation for $n = 50$. Similar results were seen for $n = 100$.

## DISCUSSION

Fallin and Schork [2000] found that departures from HWP in the form of excess heterozygosity led to a loss of accuracy, whereas departures due to excess homozygosity had little effect on the accuracy of the estimation. The latter result is believed to be due to the balancing effect of a gain in accuracy due to fewer ambiguous genotypes compatible with several haplotypic configurations (ambiguous haplotypes) due to excess homozygosity. The HLA data that we analyzed had several loci with significant deviations from HWP in the form of excess heterozygosity. In this situation, estimating HFs for three or more loci and collapsing over loci to generate two-locus haplotypes can improve the accuracy of the estimation.

The B locus was most frequently indicated as the best locus to collapse over for each of the two-locus haplotypes that did not include B. This result is most likely due to the degree of LD between the B locus and the two loci that constitute the given haplotype of interest and the large number of alleles at the B locus. In the full sample, the B locus has the highest degree of LD with the A locus. While this will not necessarily be the case in every subsample, it will be so in the majority. It is of interest to consider the DPB1-A haplotypes since the B locus was selected as the best locus to collapse over among only 39 and 48% of the replicates for $n = 50$ and $n = 100$, respectively. For these haplotypes, neither locus deviates significantly from HWP in the full sample. When comparing all *-DPB1 haplotypes, the B-DPB1 shows the greatest LD, yet conversely, when considering all *-B haplotypes, the B-DPB1 shows the least LD. This leads us to conclude that, when possible, the choice of a locus to collapse over should be based on the degree of LD between that locus and the loci in the given haplotype of interest, where higher LD is better. The collapsing procedure appears to be most beneficial when one of the loci in the two-locus haplotype of interest deviates significantly from HWP. In contrast, estimates for DPB1-A haplotypes had the least improvement due to collapsing in the observed data and showed little to no improvement due to shuffling. Also, the benefit due to collapsing was smaller in the shuffled data than in the observed data for the other three locus pairs. Analyses using *-B haplotypes (results not shown) led to similar findings.

Better performance was found in the larger sample size for two- and three-locus collapsed estimation, although the difference was not great. Also, the relative difference between the performance of collapsed and two-locus estimation was not greatly affected by sample size.

The results of the shuffling procedure on the maxdiff statistic for two-locus estimation indicated that violations of HWP may be responsible for the large over- and underestimates that can occur even among the more frequent haplotypes in a sample. Some of the improve-

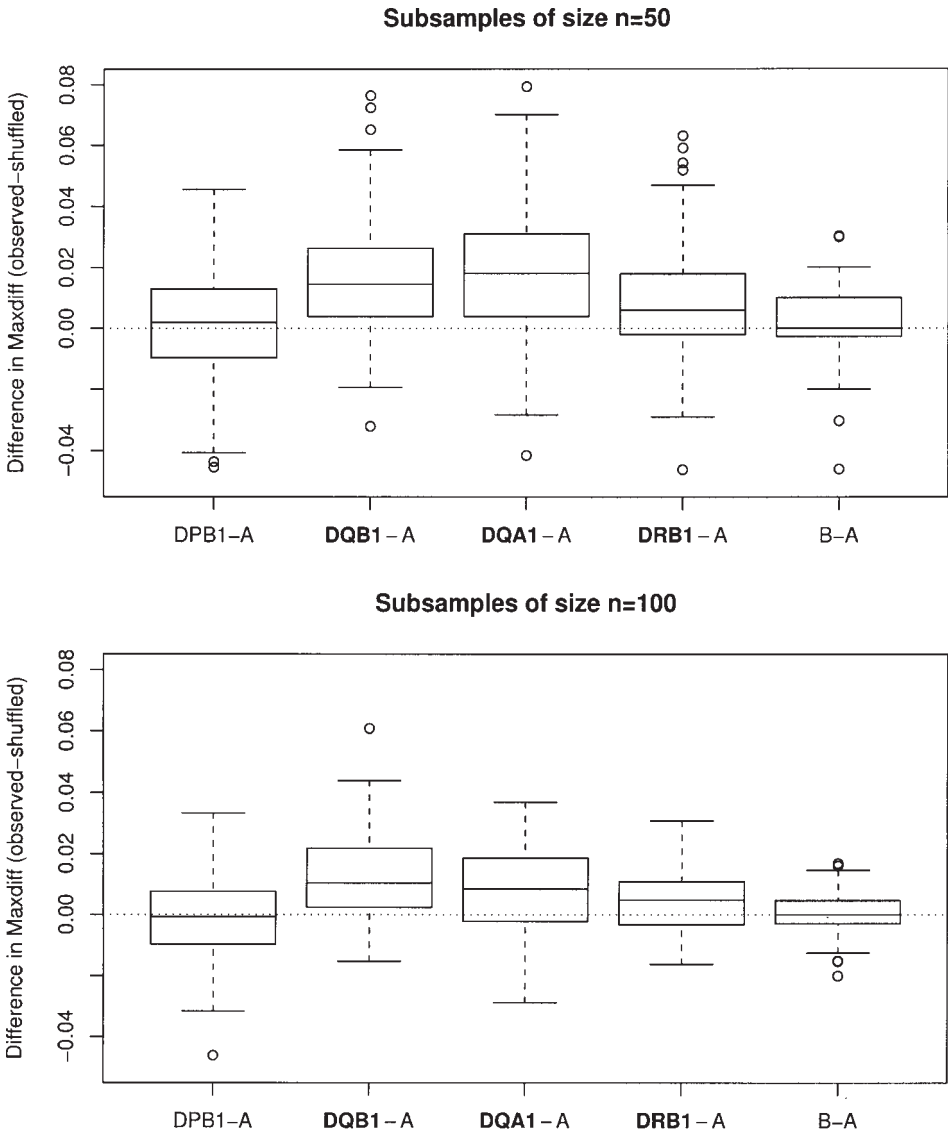


Fig. 2. Boxplots of the difference in maxdiff measures between each observed and shuffled replicate pair ($\text{maxdiff}_{\text{observed}} - \text{maxdiff}_{\text{shuffled}}$). Loci listed in bold type deviated significantly from HWP in the full sample (see Table I). For $n = 50$, a difference in maxdiff of 0.02 for a given sample indicates that shuffling reduced the largest estimation error by two haplotypes. The boxes indicate the middle 50% of the differences in 100 replications 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.

ment (reduction) in maxdiff was due to the smaller percentage of ambiguous haplotypes in the shuffled sample compared with the observed sample. Restricting analyses to replicate pairs (observed/shuffled) that had roughly the same percentage of ambiguous haplotypes still led to a significant reduction in maxdiff due to shuffling for the DQB1-A, DQA1-A, and

DRB1-A haplotypes. Thus, the larger estimation errors for specific haplotypes in the observed samples are not explained by the percentage of ambiguous haplotypes alone.

Estimation of HFs via the EM algorithm in the highly polymorphic HLA system is in general quite accurate. The assumption of HWP is relevant to the performance of the EM algorithm. This is especially true for data in which there is excess heterozygosity. Thus, we recommend testing for deviation from HWP before interpreting the results of HFs estimated by the EM algorithm. When two-locus haplotypes will be estimated by collapsing over a third locus, as implemented in some HF estimation programs, a choice of the best locus to collapse over can be based on the degree of LD and deviation from HWP. Future research will indicate whether collapsing over a larger number of loci may lead to further increases in accuracy. These increases, however, would be mediated by the fact that the number of rare and unique haplotypes will increase as the number of loci in the original estimation is increased.

## ACKNOWLEDGMENTS

This work was supported by NIH grants GM35326, CA84497 (R.M.S., D.M., J.A.H., G.T.), and DK46626 (J.A.N., H.A.E.) and by an American Diabetes Association Career Development Award (J.A.N.). We thank two anonymous reviewers and Peg Boyle for their helpful comments.

## REFERENCES

Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, Thomson G. 1997. Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet 61:734–47.

Baur MP, Danilovs J. 1980. Population genetic analysis of HLA-A, B, C, DR, and other genetic markers. In: Terasaki PI, editor. Histocompatibility testing 1980. Los Angeles: UCLA Tissue Typing Laboratory, p 955–93.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–7.

Fallin D, Schork NJ. 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–59.

Guo SW, Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–72.

Hawley ME, Kidd KK. 1995. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–11.

Hedrick PW. 1987. Gametic disequilibrium measures: proceed with caution. Genetics 117:331–41.

Johnson GC, Todd JA. 2000. Strategies in complex disease mapping. Curr Opin Genet Dev 10:330–4.

Long JC, Williams RC, Urbanek M. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810.

Noble JA, Valdes AM, Thomson G, Erlich HA. 2000. The HLA class II locus DPB1 can influence susceptibility to type 1 diabetes. Diabetes 49:121–5.

Schipper RF, D'Amaro J, de Lange P, Schreuder GM, van Rood JJ, Oudshoorn M. 1998. Validation of haplotype frequency estimation methods. Hum Immunol 59:518–23.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin: a software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.

Thomson G. 1995a. Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–98.

Thomson G. 1995b. Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. Am J Hum Genet 57:474–86.

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK. 2000. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518–22.

Todd JA. 1995. Genetic analysis of type 1 diabetes using whole genome approaches. Proc Natl Acad Sci U S A 92:8560–5.

Valdes AM, Thomson G. 1997. Detecting disease-predisposing variants: the haplotype method. Am J Hum Genet 60:703–16.

Valdes AM, McWeeney SK, Thomson G. 1999a. Evidence for linkage and association to alcohol dependence on chromosome 19. Genet Epidemiol 17:S367–72.

Valdes AM, Thomson G, Erlich HA, Noble JA. 1999b. Association between type 1 diabetes age of onset and HLA among sibling pairs. Diabetes 48:1658–61.