Genetic Epidemiology 27: 118–130 (2004)

The Haplotype Runs Test: The Parent-Parent-Affected Offspring Trio Design

Ethan M. Lange,[1]* and Michael Boehnke[2]

[1]*Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Winston-Salem, North Carolina*
[2]*Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, Michigan*

The increasing availability of maps of dense polymorphic markers makes use of haplotype data in family-based association analyses an attractive alternative to single marker association tests. We describe a novel class of statistics designed to test for an association between marker haplotypes and a qualitative trait using the parent-parent-affected-offspring trio design. Our haplotype runs test (HRT) is based on consecutive allele-sharing between pairs of haplotypes. We assign weights according to the relative frequencies of the alleles for which the two haplotypes match. Herein, we compare the HRT to the maximum-identity-length-contrast (MILC) statistic, the single-locus transmission/disequilibrium test (TDT), and the generalized test of transmission disequilibrium for haplotype data, as implemented in the software TRANSMIT, using both simulated data and published haplotype data from the recessive disorder ataxia-telangiectasia. Our simulation results suggest that the HRT outperforms the MILC and that the HRT provides comparable power to the TDT and TRANSMIT when the number of distinct founder haplotypes with a disease susceptibility allele is small but substantially outperforms the TDT and TRANSMIT when the number of distinct founder haplotypes with a disease susceptibility allele is even of modest size. © 2004 Wiley-Liss, Inc.

Key words: haplotype; allele-sharing; linkage disequilibrium; family-based; trios

Contract grant sponsor: NIH; Contract grant numbers: HG00376, HG00040.
*Correspondence to: Ethan M. Lange, Ph.D., Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC, 27157-1063. E-mail: elange@wfubmc.edu
Received 31 July 2003; Revised 18 February 2004; Accepted 8 March 2004
Published online 26 May 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20010

INTRODUCTION

Haplotype-based association analysis is a powerful method for mapping disease susceptibility genes through linkage disequilibrium (LD). To avoid possible confounding due to population stratification and to reduce unknown haplotype phase, family-based methods are often employed. Recent advances have allowed researchers to saturate candidate chromosomal regions with high densities of genetic markers. The abundance of genetic information creates new analytical challenges to maximize the impact of the available information when conducting haplotype-based association analyses.

Traditional family-based haplotype association methods are likelihood-based and rigid by design. Clayton [1999] introduced a generalized test of transmission disequilibrium for haplotype data for the family-based study design. His approach uses the expectation-maximization (EM) algorithm to derive estimates of, and facilitate comparisons between, frequencies of transmitted and non-transmitted haplotypes. Likelihood-based haplotype association methods, such as Clayton's, can utilize only a limited number of markers due to computational constraints and non-robust haplotype frequency estimates due to rare haplotypes. For Clayton's approach, marker sets used in analyses must be specified *a priori* by the user. In practice, the optimal number of markers to be used when analyzing haplotype data is often unclear. Using large numbers of markers can create sparse distributions of haplotypes and unnecessarily increase the degrees of freedom used to evaluate the statistical significance of such tests. Further, recombination events, mutations, and genotype errors can cause two highly similar, although not identical, haplotypes derived from a common ancestor to compete against one another in the test statistic. These problems can decrease the power to detect LD when LD is present.

© 2004 Wiley-Liss, Inc.

Van der Meulen and te Meerman [1997a,b] introduced an alternative haplotype-based test of association using haplotype data for the parent-parent-affected offspring trio design that is based on the length of conserved allele sharing between pairs of haplotypes. Their test involves choosing a reference marker from an ordered marker map. For a given pair of haplotypes, a score is assigned equal to the number of consecutive identity-by-state (IBS) allele matches spanning the reference marker, where a match is defined as two haplotypes sharing the same allele at a marker. Their haplotype sharing statistic (HSS) is equal to the standard deviation of the length of the shared haplotype segments between all possible pairs of parental haplotypes, regardless of the transmission status of the haplotypes. The foundation of this approach is that the probability that two gametes are identical by descent at a locus increases as the number of markers surrounding that locus with IBS alleles increases [Meuwissen and Goodard, 2001; Nolte and te Meerman, 2002]. The major advantage of Meulen and te Meerman's approach over traditional haplotype methods is that their method can use all available marker data rather than being forced to select a small subset of markers. Unlike traditional approaches, two haplotypes that are highly concordant, but not identical, provide evidence for association.

Bourgain et al. [2002] modified Clayton and Jones' [1999] haplotype IBS sharing statistic in the context of the parent-parent-affected offspring trio design. Similar to Clayton and Jones' statistic, their maximum identity length (MILC) statistic is the total length of the contiguous region over which all markers are IBS. As in the case of Van der Meulen and te Meerman [1997a,b] and Clayton and Jones [1999], the definition of similarity is restricted to mandate inclusion of a reference or ''focal'' marker in the shared region. For the MILC approach, at each reference marker, a sharing statistic is calculated as the mean shared length from all pairings of non-transmitted haplotypes subtracted from the mean shared length from all pairings of transmitted haplotypes. The MILC statistic is the maximum observed value of these sharing statistics over the set of considered reference markers. Similar to Lange and Boehnke [1998], statistical significance is evaluated by a permutation test where scores for transmitted and non-transmitted haplotypes are randomly reassigned in each family trio.

In this report, we describe the haplotype runs test (HRT), a new set of parent-parent-affected-offspring trio non-parametric tests for association based on runs of shared alleles. Similar to the HSS and MILC, our tests evaluate shared consecutive IBS allele matches between haplotypes spanning a reference marker. We assign weights, based on allele frequencies, to shared haplotype segments. In contrast to the other proposed haplotype sharing statistics, calculation of the HRT test statistics only involves measures of sharing between the transmitted haplotypes. We address the issues of phase ambiguity and missing genotype data and allow for the possibility of genotype errors and marker allele mutations. We evaluate the significance of the HRT statistics by permuting transmitted haplotypes within family trios. Unlike traditional haplotype approaches, the HRT requires no allele or haplotype removal or lumping due to small sample sizes, excess degrees of freedom, or computational constraints.

We compare the power of our method to detect an association to other parent-parent-affected offspring trio association tests by computer simulation. We show that, for the situations we consider, our test frequently outperforms the MILC statistic, the single-locus transmission disequilibrium test (TDT) [Spielman et al., 1993], and Clayton's generalized transmission/disequilibrium test for uncertain haplotype transmission (as implemented in the computer software TRANSMIT) over multiple loci. We present strong evidence that suggests power for sharing statistics, such as the HRT and MILC, can be improved by focusing the calculation of the test statistics solely on the scores from the pairings of transmitted haplotypes as opposed to subtracting off the scores from the pairings of nontransmitted haplotypes. Finally, we illustrate the use of our test on published haplotype data from the mapping study for the rare recessive disorder ataxia-telangiectasia (A-T).

METHODS

THE CONSERVED HAPLOTYPE SHARING STATISTIC

We assume our sample is made up of N independent family trios, each consisting of two parents who may or may not be affected and a single affected offspring. Each family trio contains four parental haplotypes of M ordered genetic markers, providing a total of $4N$ parental haplotypes in the overall sample. We pair up each parental haplotype in the sample with each of the

remaining haplotypes, resulting in $\binom{4N}{2}$ haplotype pairs. We calculate a conserved haplotype sharing statistic (CHSS) for each haplotype pair centered about a chosen reference marker r, where $1 \le r \le M$. To calculate the CHSS for a pair of haplotypes, let $\hat{f}_j(A)$ be the estimated population-based allele frequency of allele A at marker j as determined by the entire sample of transmitted and nontransmitted alleles. Let A_j^k be the allele present at marker j ($1 \le j \le M$) on haplotype k (k=1,2) in the specified haplotype pair. For now, assume for a particular haplotype pair that the two haplotypes share an allele identical by state (IBS) at the reference marker r, i.e. $A_r^1 = A_r^2$. Define markers a, b, x and y ($1 \le a < b < r < x < y \le M$) such that b and a (x and y) are the first and second markers to the left (right) of the reference marker that fail to match alleles IBS between the two haplotypes.

To allow for the possibility of marker allele error or mutation, we introduce a penalty parameter, π, and allow for up to one allele error or mutation on each side of the reference marker. Allowing for up to one break on either side of the reference marker in the conserved marker-allele sequence creates four possible values of $CHSS^r$ for a particular haplotype pairing spanning the reference marker r. These are:

$$CHSS_1^r = \prod_{i=b+1}^{x-1} \left(\hat{f}_i(A_i^1)\right)^{-1}$$

(do not accept penalty at marker b or x)

$$CHSS_2^r = \prod_{j=a+1}^{b-1} \left(\hat{f}_j(A_j^1)\right)^{-1} \cdot \pi \cdot CHSS_1^r$$

(accept penalty at marker b, do not accept penalty at marker x)

$$CHSS_3^r = CHSS_1^r \cdot \pi \cdot \prod_{k=x+1}^{y-1} \left(\hat{f}_k(A_k^1)\right)^{-1}$$

(do not accept penalty at marker b, accept penalty at marker x)

$$CHSS_4^r = \prod_{j=a+1}^{b-1} \left(\hat{f}_j(A_j^1)\right)^{-1} \cdot \pi \cdot CHSS_1^r \cdot \pi \cdot \prod_{k=x+1}^{y-1} \left(\hat{f}_k(A_k^1)\right)^{-1}$$

(accept penalties at markers b and x). We define $CHSS^r = max\{1, CHSS_{i=1,4}^r\}$.

The four possible values of $CHSS^r$ are not always valid or considered. For example, when the two haplotypes in a pair are identical at every marker, then only $CHSS_1^r$ is relevant. Also, if the two haplotypes fail to match at the reference marker r, we initially multiply each $CHSS_i^r$ by π, thus allowing up to three mismatches in this case.

An example haplotype pair and the four corresponding $CHSS_i^r$ values are presented in Figure 1. In this example, marker r=7 is the reference marker. Note the two haplotypes share a conserved haplotype segment spanning markers 6 through 10 and that the pair of haplotypes also share alleles at markers 2, 3, 4, 12, and 14. In this case, the maximum $CHSS_i^7$ value depends on the estimated allele frequencies and the penalty value π.

PHASE AMBIGUITY AND MISSING DATA

Missing genotype data and ambiguous phase information can cause difficulties when analyzing haplotype data using allele sharing methods. To maintain an unbiased test, it is critical that there exists equal marker information on both the transmitted and non-transmitted haploytpe [Bourgain et al., 2002].

Bourgain et al.'s [2002] solution for addressing phase ambiguity, when using the MILC, is to treat markers with ambiguous phase as missing. We propose an alternative approach that uses the available marker data to either declare a mismatch or to assign a down-weighted score reflecting a contingency match at the marker. For a parent, when phase is known, only one possible allele at each marker is assigned to each parental haplotype. However, when phase is ambiguous at a marker in a parent, either parental allele at that marker must be considered as possible when calculating the CHSS values using the haplotypes from this parent. If neither possible allele matches the allele for the other haplotype in the haplotype pairing, then a mismatch is declared at the marker. Otherwise, if either of the two possible parental alleles for the haplotype with the phase ambiguity matches the allele at the marker for the haplotype with phase certainty, we then declare a match. However, because both alleles were considered, the match is down-weighted by a factor of $\frac{1}{2}$. If the allele for which the haplotypes match has an estimated frequency greater than 0.5, then, when taking into account the down-weighting factor, including the match would actually penalize the proposed match. To avoid this penalty, if the

	Marker													
	1	2	3	4	5	6	7[a]	8	9	10	11	12	13	14
Haplotype *A*:	1	**3**	**1**	**2**	1	**1**	**3**	**1**	**4**	**1**	1	**1**	1	**2**
Haplotype *B*:	2	**3**	**1**	**2**	2	**1**	**3**	**1**	**4**	**1**	2	**1**	2	**2**

$$\mathrm{CHSS}_1^7 = \left[\hat{f}_6(1)\cdot\hat{f}_7(3)\cdot\hat{f}_8(1)\cdot\hat{f}_9(4)\cdot\hat{f}_{10}(1)\right]^{-1}$$

$$\mathrm{CHSS}_2^7 = \left[\hat{f}_2(3)\cdot\hat{f}_3(1)\cdot\hat{f}_4(2)\right]^{-1}\cdot\pi\cdot\mathrm{CHSS}_1^7$$

$$\mathrm{CHSS}_3^7 = \mathrm{CHSS}_1^7\cdot\pi\cdot\left[\hat{f}_{12}(1)\right]^{-1}$$

$$\mathrm{CHSS}_4^7 = \left[\hat{f}_2(3)\cdot\hat{f}_3(1)\cdot\hat{f}_4(2)\right]^{-1}\cdot\pi\cdot\mathrm{CHSS}_1^7\cdot\pi\cdot\left[\hat{f}_{12}(1)\right]^{-1}$$

$$\mathrm{CHSS}_{AB}^7 = max\left\{1, \mathrm{CHSS}_{i=1,4}^7\right\}$$

[a] reference marker

Fig. 1. Calculation of CHSS_i^r values for an example haplotype pair. Haplotypes *A* and *B* share alleles (in bold) at markers 2, 3, 4, 6, 7, 8, 9, 10, 12, and 14. Marker *r*=7 is the reference marker in the calculation of CHSS_i^r values. Penalties are considered in CHSS_i^7 calculations for mismatches at markers 5 and 11.

estimated allele frequency of the matching allele is >0.5, we let CHSS values remain unchanged (in essence the marker is skipped). If phase is ambiguous at a marker in both haplotypes of a pairing, then all four possible allele pairings are considered and the sum of the inverse allele frequencies for the allele pairings that match are down-weighted by a factor of $\frac{1}{4}$. If the resulting value is <1, the marker is skipped when calculating the CHSS.

We address the issue of missing marker data using an algorithm very similar to the algorithm of Bourgain et al. [2002]. If marker genotypes are missing on both parents in a trio, we skip the marker when calculating the CHSS scores for all pairings of parental haplotypes that involve any of the four haplotypes from the trio in question. Marker genotypes are kept, when available, in each parent regardless of missing genotypes in the other parent or the offspring. When marker data are missing on one parent, marker phase in the other parent is determined by the marker genotype in the offspring if possible. To maintain an unbiased test, where the amount of marker information for both the transmitted and the non-transmitted haplotype in the parent is equal, the genotype for the parent with missing marker data remains coded missing even in the event that the allele for the transmitted haplotype from this parent can be determined unambiguously from the offspring. If the marker genotype in the offspring cannot unambiguously determine marker phase in the parent with marker data, phase is considered ambiguous and analyses are conducted as described in the preceding paragraph. One could in principle probabilistically infer the missing marker data, but this would increase computational complexity and only negligibly impact the results unless the proportion of missing data is large. In Figure 2, we provide the different CHSS values for an example haplotype pairing that includes both phase ambiguity and missing marker data. Similar to the previous example, the maximum CHSS_i^5 value depends on the estimated allele frequencies and the penalty value π.

THE HAPLOTYPE RUNS TEST (HRT) STATISTICS

We consider two different HRT statistics formed using the CHSS^r values:

$$\mathrm{HRT}_{ln}^r = \sum_{1\le i\le j\le 4N} I_i I_j\, ln(\mathrm{CHSS}_{ij}^r)$$

$$\mathrm{HRT}_{\mathrm{T}=t}^r = \sum_{1\le i\le j\le 4N} I_i I_j I_{\{\mathrm{CHSS}_{ij}^r \ge t\}}$$

	Marker											
	1	2	3	4	5[a]	6	7	8	9	10	11	12
				1				3			**1**	
				or				or			or	
Haplotype *A*:	**1**	1	1	**2**	**3**	**1**	**x**	**4**	**2**	1	**2**	3
Haplotype *B*:	**1**	2	2	**2**	**3**	**1**	**1**	**4**	**2**	2	**2**	4
								or		or	or	
								5		3	**1**	

$$\mathrm{CHSS}_1^5 = \left[max[(2\hat{f}_4(2))^{-1},1] \cdot \hat{f}_5(3)^{-1} \cdot \hat{f}_6(1)^{-1} \cdot max[(4\hat{f}_8(4))^{-1},1] \cdot \hat{f}_9(2)^{-1} \right]$$

$$\mathrm{CHSS}_2^5 = \pi \cdot \mathrm{CHSS}_1^5$$

$$\mathrm{CHSS}_3^5 = \mathrm{CHSS}_1^5 \cdot \pi \cdot max[(4\hat{f}_{11}(1) + 4\hat{f}_{11}(2))^{-1},1]$$

$$\mathrm{CHSS}_4^5 = \pi \cdot \mathrm{CHSS}_1^5 \cdot \pi \cdot max[(4\hat{f}_{11}(1) + 4\hat{f}_{11}(2))^{-1},1]$$

$$\mathrm{CHSS}_{AB}^5 = max\left\{1, \mathrm{CHSS}_{i=1,4}^5\right\}$$

[a] reference marker

x = missing genotype

a
or → phase ambiguous and either parental allele 'a' or 'b' can be assigned to haplotype
b

Fig. 2. Calculation of CHSS_i^r values for an example haplotype pair with phase ambiguity and missing marker data. Marker r=5 is the reference marker. Haplotypes *A* and *B* have unambiguous phase and share alleles (in bold) at markers 1, 5, 6, and 9. Data are missing on haplotype *A* at marker 7 and hence we skip this marker in calculation of CHSS_i^5 values. Haplotype *A* has phase ambiguity at marker 4, with one of the two possible alleles (allele "2") identical to the allele at marker 4 on haplotype *B*. Marker phase is ambiguous at markers 8 and 11 in both haplotypes, with one and two possible allele matches, respectively, between the two haplotypes at the two markers. Haplotype *B* has phase ambiguity at marker 10 with no possible allele matches with haplotype *A*. Penalties are considered in CHSS_i^5 calculations for mismatches at markers 3 and 10.

Here CHSS_{ij}^r is the value of CHSS^r for haplotypes i and j at reference marker r, $I_{i(j)}$ is one if haplotype $i(j)$ is transmitted to the affected offspring and zero otherwise, t is a threshold specified by the user as the minimum criterion for a pair of haplotypes to be considered likely to be identical by descent (IBD), and $I_{\{\mathrm{CHSS}_{ij}^r \geq t\}}$ is one if $\mathrm{CHSS}_{ij}^r \geq t$ and zero otherwise. Herein, we set t=100 or 10,000.

HRT_{ln}^r is the natural logarithm of the inverse of the estimated match probabilities over shared regions assuming linkage equilibrium (LE) between adjacent markers. $\mathrm{HRT}_{\mathrm{T}=t}^r$ is a threshold statistic that is designed to focus on haplotype pairs that are plausibly considered to be IBD. All pairings of haplotypes determined to meet the threshold of likely being IBD are given the same weight, while all pairings of haplotypes failing to meet this criterion are given no weight. In this manner, we make a clear distinction between the haplotypes that fail to match or match for only a few common alleles and the haplotypes that match over an extended set of markers. Importantly, $\mathrm{HRT}_{\mathrm{T}=t}^r$ avoids having the score(s) from a single or small number of haplotype pairs dominate the test statistic. Choosing a modest threshold value (e.g., t=100) results in a test that focuses on excess sharing of short or common haplotypes whereas choosing a high threshold (e.g., t=10,000) results in a test that focuses on excess sharing of rare or extended haplotypes. It should be noted that calculation of both HRT_{ln}^r and $\mathrm{HRT}_{\mathrm{T}=t}^r$ only involves CHSS^r scores from pairs of transmitted haplotypes. This

framework is different from the framework of the MILC, which contrasts the degree of haplotype sharing between transmitted and non-transmitted haplotypes.

SIGNIFICANCE ESTIMATION OF THE HRT STATISTICS

Linkage disequilibrium (LD) between closely linked markers and the lack of robust haplotype frequency estimates for even a modest number of loci make evaluation of statistical significance of the HRT statistics difficult using conventional statistical procedures. We surmount this problem by the use of a permutation test. By design, equal amounts of marker information are available for both haplotypes from each parental transmitted/non-transmitted haplotype pair. Specifically, for a given parent, if phase is ambiguous for the transmitted haplotype, then phase is also ambiguous for the non-transmitted haplotype and vice versa. The haplotype scores from parental transmitted/non-transmitted pairs are thus ''exchangeable'' [Good, 1994]. For each parent, we randomly choose, with equal probability, which one of the two parental haplotypes is transmitted to the affected offspring. Under the null hypothesis of no association, this approach results in permutations of the data that are equally likely. We evaluate the P values of the observed HRT test statistics as the proportion of permutations for which the permuted-haplotype test statistics are greater than the observed test statistics.

SIMULATION DESIGN

To verify that the permutation framework results in the correct nominal significance level and to compare statistical power and accuracy of gene localization for the HRT, the single-locus TDT, and haplotype-based generalized transmission/disequilibrium test proposed by Clayton [1999], we conducted computer simulations. For each simulation condition we generated R=1,000 replicate samples of N=250 parent-parent-affected-offspring trios. For each replicate sample, we then constructed 500 randomly permuted data sets to assess significance for that replicate sample.

We generated genotype data for five multi-allelic markers spaced 1 cM apart and 16 biallelic markers spaced evenly between the five multi-allelic markers; we placed a disease locus directly on top of the middle multi-allelic marker. The multi-allelic markers had six codominant alleles. Consistent with a complex genetic disorder, we set the disease prevalence to 8% and the sib-recurrence-risk ratio [Risch, 1987] λ_s to 2.0. We simulated trios under three models: no disease predisposing variant, a dominant (f_D=0.05, Pen(DD)=Pen(Dd)=0.4264, Pen(dd)=0.0426) predisposing variant, and a recessive (f_D=0.20, Pen(DD)=0.7588, Pen(Dd)=Pen(dd)=0.0517) predisposing variant.

We constructed a population of disease-associated and non-disease-associated haplotypes using an evolutionary, stochastic algorithm similar to the ones presented by Devlin and Risch [1995] and Calafell et al. [1998]. We assumed a founder population of F=100, 500, or 1,000 individuals. Each founder was assigned alleles to each of the 21 markers and the disease locus. To assess the impact of LD on the relative power of the different methods, we modeled two distinct types of founder populations. We generated samples with founder marker data randomly under conditions of either LE or strong LD. Models with founder LE were simulated from a single population using equal allele frequencies of 0.50 (heterozygosity H=0.50) for the di-allelic markers and 0.20, 0.20, 0.15, 0.15, 0.15, 0.15 (H=0.83) for the multi-allelic markers. We randomly assigned founder alleles for each marker, based on these allele frequencies, independent of other marker data. We simulated examples with founder LD by combining two equal-sized subpopulations with very different allele frequencies to create a single founder population with marker-marker LD. We simulated the first founder subpopulation using (ordered) allele frequencies of 0.80, 0.20 for the biallelic markers and 0.35, 0.25, 0.25, 0.05, 0.05, 0.05 for the multi-allelic markers and the second founder subpopulation with (ordered) allele frequencies of 0.20, 0.80 for the biallelic markers and 0.05, 0.05, 0.05, 0.25, 0.25, 0.35 for the multi-allelic markers. We simulated the allele at the disease susceptibility gene in the founder population independently of the marker alleles. When combining two founder subpopulations to create marker-marker LD, we assumed the frequency of the disease predisposing allele to be either equal in the two subpopulations or twice as frequent in the first subpopulation as in the second subpopulation. The latter choice increased LD between the disease locus and the individual marker loci. After construction of the founder population, we grew the population exponentially over 50 generations to reach a final population of 250,000 individuals.

To accomplish this desired population growth, we chose a pair of "parental" haplotypes with replacement at random using a uniform random number generator. From this pair of haplotypes, we formed a single new haplotype, allowing for recombination, mutation at the disease locus and at each biallelic marker (with probability 10^{-5}), and mutation at each of the five multi-allelic markers (with probability 10^{-4}). We constructed trios randomly from the final generation of haplotypes using the proposed genetic models. Finally, with probability equal to 0.01, we introduced marker allele misspecification error for each allele at each marker in each individual. Genotypes that were subsequently inconsistent with Mendelian inheritance were removed.

For each simulated data set, we performed the HRT, MILC, TDT, and TRANSMIT to test the *a priori* hypothesis of no association at the center multi-allelic marker. For the HRT and MILC, this center multi-allelic marker was set as the reference marker. The HRT, MILC, and TRANSMIT used marker data from the available multi-allelic and biallelic markers in addition to data provided by the reference marker. When analyzing data using the MILC, we reclassified markers with ambiguous phase as "missing" and used "score option 2" as described in Bourgain et al. [2002] to calculate the sharing statistic. We calculated the single-locus TDT at the center multi-allelic marker and evaluated statistical significance using a permutation procedure [Lazzeroni and Lange, 1998]. To maintain expected counts of at least 5 observations per haplotype among the transmitted haplotypes, we restricted analyses using TRANSMIT to haplotypes with estimated frequencies of at least 0.01. We based statistical significance estimates from TRANSMIT on asymptotic distribution theory.

Finally, to assess the impact of evaluating sharing only between transmitted haplotypes as opposed to contrasting the sharing between transmitted and non-transmitted haplotypes, we constructed the HRT statistics using the framework design of the MILC. We also constructed a new statistic, $\mathrm{HRT}_{\mathrm{MILC}}$, which was calculated in the identical fashion as the MILC statistic except that this new statistic is calculated only from the transmitted haplotypes. That is,

$$\mathrm{HRT}^{r}_{\mathrm{MILC}} = \sum_{1 \le i \le j \le 4N} I_i I_j \, length^{r}_{ij},$$

where $length^{r}_{ij}$ is the distance [Clayton and Jones, 1999; Bourgain et al., 2000] spanned by consecutive IBS matches for haplotypes i and j at reference marker r and $I_{i(j)}$ is one if haplotype $i(j)$ is transmitted to the affected offspring and zero otherwise. Thus, both the HRT and MILC statistics were evaluated by using our approach of only considering transmitted haplotypes and by Bourgain et al.'s [2000] approach of contrasting the difference in sharing between transmitted and non-transmitted haplotypes.

We analyzed two different sets of markers using TRANSMIT. The first set consisted of the middle multi-allelic marker (directly on top of the disease locus) along with the two nearest multi-allelic markers. Due to computational complexity and sparse haplotype counts, it was not feasible to include the remaining two multi-allelic markers. The second set consisted of the center multi-allelic marker along with the two nearest biallelic markers on each side of the middle multi-allelic marker. We carried out analysis with the HRT and the MILC first for the five multi-allelic markers and then for all 21 simulated markers. We assigned a value of 0.05 (see Discussion for the impact of this choice) to the mismatch penalty parameter, π, for the HRT analyses.

APPLICATION TO ATAXIA-TELANGIECTASIA

Ataxia-telangiectasia (A-T) is a rare autosomal recessive disorder characterized by progressive cerebellar ataxia, oculocutaneous telangiectasia, radiosensitivity, cellular and humoral immunodeficiency, and predisposition to malignancies [Gatti et al., 1991]. Twenty-seven A-T families from Costa Rica were collected and used for haplotype analysis to fine map the A-T locus [Uhrhammer et al., 1995] on chromosome 11q. Nineteen of the 27 families consisted of two parents and an affected offspring. We performed HRT, MILC, TDT, and TRANSMIT analyses on these 19 parent-parent-offspring trios using nine markers spanning an approximate 6-cM interval. Due to the small number of families, we conducted the TDT tests using Fisher's exact test and restricted TRANSMIT analyses to three marker haplotypes expected to occur a minimum of five times in the offspring. TRANSMIT was performed for all possible consecutive sets of three adjacent markers. Contrary to the approach of Bourgain et al. [2000], which calculates a "global" P value over a predefined set of reference markers, we calculated

statistical significance estimates for the MILC independently at each reference marker for comparison to the other tests.

RESULTS

SIMULATIONS

To assess type I error rates, we tested the HRT, MILC, TDT, and TRANSMIT statistics under the null hypothesis of no association using 1,000 randomly generated data sets from founder populations of size 100, 500, and 1,000. Results for the HRT, MILC, and the TDT obtained using the permutation tests were consistent with the nominal type I error rate at the α=0.05 level. The results using TRANSMIT, where statistical significance estimates were calculated based on asymptotic theory results, were also consistent with the nominal type I error rate (data not shown). Results from the power analyses suggest the HRT, MILC, TDT, and TRANSMIT all have good power when the population of disease-associated haplotypes is derived from a small number of founders (F=100) (Tables I and II). For the models we considered, the results demonstrate that among the methods, TRANSMIT is most powerful given 100 founders while the HRT is most powerful given a larger number (F=500 or 1,000) of founders. Power decreased precipitously for the MILC, TDT, and TRANSMIT as the number of disease founder haplotypes increased. Power decreased at a much slower rate for the HRT. For some HRT statistics, power actually increased when going from 100 to 500 founders (see Discussion). The HRT, MILC, and TRANSMIT all demonstrated significantly improved power when data were available on the biallelic markers.

Significant advantages were gained using $HRT_{T=t}$, even when not including the 16 biallelic markers for the larger founder populations (F=500, 1,000). $HRT_{T=100}$ outperformed the MILC, TDT, and TRANSMIT for founder population sizes of F=500 and 1,000 individuals under both the recessive and dominant models when analyses were restricted to the five multi-allelic markers. In the simulations with all 21 markers, $HRT_{T=10,000}$ clearly outperformed all other statistics for the larger founder populations regardless of disease model or presence versus absence of marker-marker LD. Results using HRT_{ln} were mixed, but for the cases considered, HRT_{ln} generally did not perform as well as $HRT_{T=t}$. The HRT_{ln} outperformed the MILC test in every simulation. Marker-marker LD in the founder population had

TABLE I. Estimated power for λ_s=2.0, significance level α=0.05, and founder population markers in linkage equilibrium

Number of founders	Disease model	Number of markers used[a]	HRT_{ln}	$HRT_{T=100}$	$HRT_{T=10,000}$	HRT_{MILC}	MILC	TDT[b]	TRANSMIT[c]
100	Dom	5	0.94	0.87	0.12	0.92	0.79	0.93	0.99
100	Dom	21	0.98	0.98	0.97	0.97	0.86	0.93	0.99
500	Dom	5	0.81	0.92	0.18	0.87	0.71	0.59	0.83
500	Dom	21	0.96	0.97	0.99	0.96	0.86	0.59	0.94
1,000	Dom	5	0.69	0.83	0.25	0.77	0.62	0.38	0.66
1,000	Dom	21	0.90	0.93	0.99	0.92	0.79	0.38	0.80
100	Rec	5	0.85	0.81	0.14	0.85	0.78	0.84	0.93
100	Rec	21	0.92	0.94	0.92	0.91	0.78	0.84	0.99
500	Rec	5	0.59	0.71	0.17	0.65	0.46	0.37	0.52
500	Rec	21	0.82	0.84	0.96	0.82	0.66	0.37	0.81
1,000	Rec	5	0.42	0.56	0.16	0.50	0.36	0.23	0.34
1,000	Rec	21	0.66	0.71	0.95	0.69	0.51	0.23	0.62

[a]Disease locus placed directly on top of center multi-allelic marker. The center multi-allelic marker is the one and only reference marker. Marker maps consist of (1) 5 multi-allelic markers with 6 alleles and (2) the same 5 multi-allelic markers and an additional 16 biallelic markers evenly spaced between the 5 multi-allelic markers.

[b]TDT using only middle multi-allelic marker for models including both 5 and 21 markers.

[c]TRANSMIT using middle 3 multi-allelic markers for model with 5 multi-allelic markers and middle multi-allelic marker plus the 4 closest biallelic markers for model with 21 markers.

TABLE II. Estimated power for λ_s=2.0, significance level α=0.05, and founder population markers in linkage disequilibrium (frequency of disease susceptibility allele equal in the two founder subpopulations)

Number of founders	Disease model	Number of markers used[a]	HRT_{ln}	$HRT_{T=100}$	$HRT_{T=10,000}$	HRT_{MILC}	MILC	TDT[b]	TRANSMIT[c]
100	Dom	5	0.92	0.85	0.13	0.89	0.74	0.93	0.98
100	Dom	21	0.93	0.92	0.94	0.92	0.80	0.93	0.99
500	Dom	5	0.77	0.84	0.19	0.82	0.63	0.58	0.81
500	Dom	21	0.77	0.67	0.98	0.83	0.67	0.58	0.92
1,000	Dom	5	0.61	0.71	0.23	0.66	0.48	0.39	0.63
1,000	Dom	21	0.58	0.41	0.97	0.66	0.48	0.39	0.75
100	Rec	5	0.83	0.78	0.15	0.82	0.66	0.85	0.94
100	Rec	21	0.86	0.81	0.91	0.85	0.69	0.85	0.98
500	Rec	5	0.54	0.59	0.19	0.55	0.39	0.39	0.52
500	Rec	21	0.53	0.39	0.89	0.57	0.40	0.39	0.74
1,000	Rec	5	0.35	0.41	0.17	0.38	0.27	0.23	0.32
1,000	Rec	21	0.33	0.24	0.82	0.40	0.28	0.23	0.50

[a]Disease locus placed directly on top of center multi-allelic marker. The center multi-allelic marker is the one and only reference marker. The Marker maps consist of (1) 5 multi-allelic markers with 6 alleles and (2) the same 5 multi-allelic markers and an additional 16 biallelic markers evenly spaced between the 5 multi-allelic markers.
[b]TDT using only middle multi-allelic marker for models including both 5 and 21 markers.
[c]TRANSMIT using middle 3 multi-allelic markers for model with 5 multi-allelic markers and middle multi-allelic marker plus the 4 closest biallelic markers for model with 21 markers.

no influence on which statistic had maximal power (compare Tables I and II). Power was greater for models with founder LD and the disease-predisposing allele twice as frequent in the first founder sub-population as in the second founder sub-population, but the orders of the statistics, in terms of power, were identical to the orders for the models that incorporated founder LD and equal founder disease susceptibility allele frequencies (data not shown). Specifically, for the recessive model with F=1,000 and using 21 markers, the estimated powers to detect significant deviations from the null hypothesis were 0.44, 0.25, 0.85, 0.49, 0.41, 0.35, and 0.58 for HRT_{ln}, $HRT_{T=100}$, $HRT_{T=10,000}$, HRT_{MILC}, MILC, TDT, and TRANSMIT, respectively.

Our modified MILC statistic, HRT_{MILC}, which only considered scores from the transmitted haplotype pairs, clearly outperformed the MILC statistic under all simulation models we considered (Tables I and II). Consistent with this result, all HRT statistics calculated from just the transmitted haplotypes pairs uniformly outperformed the corresponding HRT statistics calculated by subtracting off the scores of the non-transmitted pairs from the scores of the transmitted pairs (data not shown).

ATAXIA-TELANGIECTASIA

Results from the A-T example are illustrated in Figure 3. Due to computational constraints, P values for HRT and MILC are based on 10^6 permutations of the data. The single-locus TDT achieved very strong significance estimates ($P < 1.0 \times 10^{-6}$) at markers D11S1817, D11S1343, and D11S927. None of the permuted statistics were as large as the observed statistics for the 10^6 random permutations of the data. Interestingly, results were less significant at biallelic marker D11S384 ($P = 3.6 \times 10^{-4}$) and multi-allelic marker D11S1294 ($P = 3.1 \times 10^{-5}$), which immediately flank the A-T locus. Results from TRANSMIT based on the three adjacent marker haplotypes were generally not as significant as those for the single-locus TDT and were again less significant for the markers immediately flanking the A-T locus. Significance estimates for TRANSMIT are plotted in Figure 3 at the center marker for each three-marker haplotype. The most significant result ($P = 4.6 \times 10^{-6}$) was obtained using the proximal haplotype D11S1817-D11S1343-D11S1819. HRT_{ln}, $HRT_{T=100}$, HRT_{MILC}, and MILC gave the most significant findings across the region with an estimated $P < 1.0 \times 10^{-6}$ at all

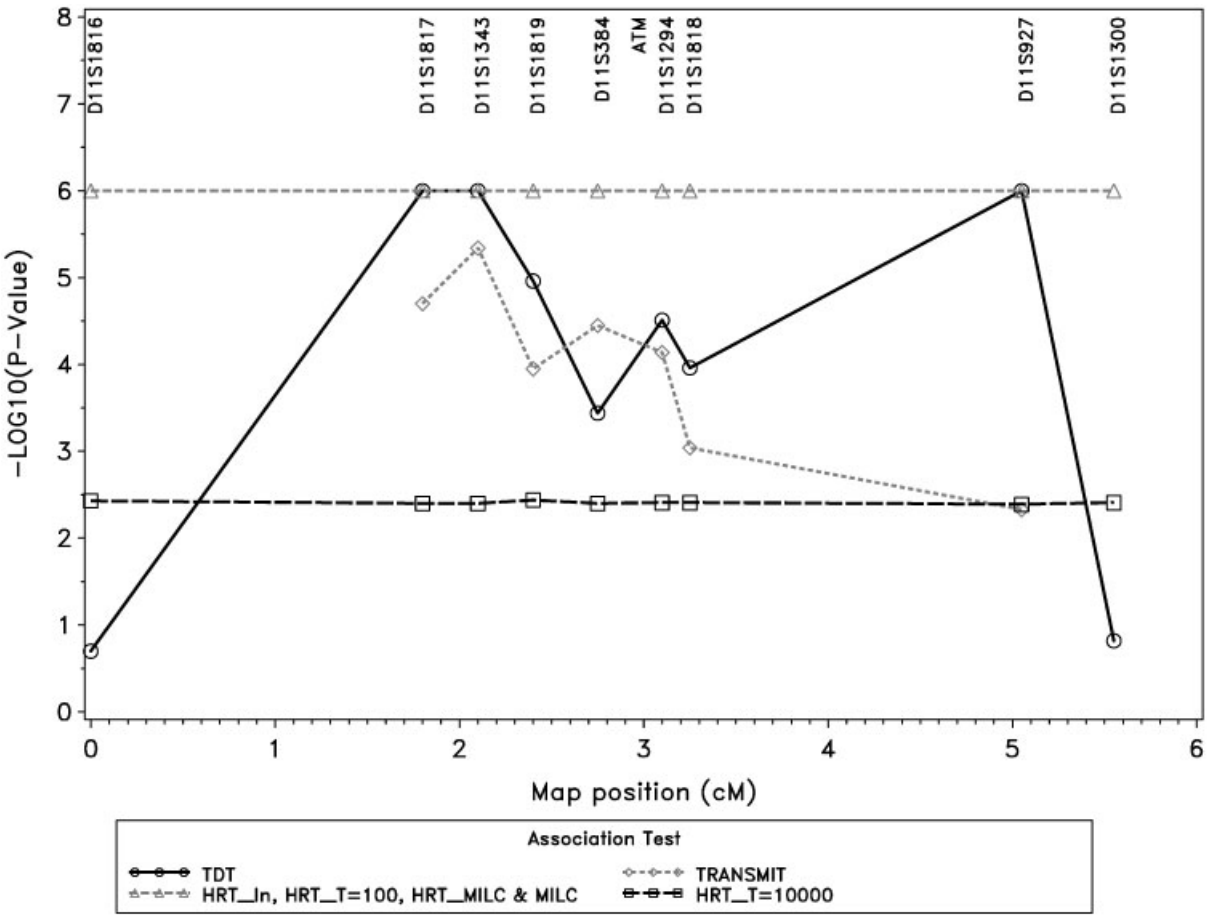

Fig. 3. Association results for the ataxia-telangiectasia (A-T) example. The A-T gene (ATM) was localized between biallelic marker D11S384 and multi-allelic marker D11S1294. Due to computational constraints, a significance ceiling was imposed such that $-\log_{10}(P$ value$) \leq 6.0$. Statistical significance estimates for TRANSMIT are plotted at the center marker in each haplotype. Results for each test are connected by a line to aid visualization of results and the lines are not intended to be used to infer statistical significance at markers not analyzed.

marker locations. $HRT_{T=10,000}$ gave estimated P values of approximately 3.5×10^{-3} at all marker locations.

DISCUSSION

Herein we introduce a novel class of statistics, the haplotype runs test (HRT), to test for an association between marker haplotype data and a qualitative trait of interest. Our test is based on consecutive allele matches between pairs of haplotypes, with weights assigned according to the frequencies of alleles for which the two haplotypes match. Our results demonstrate the utility of this approach and suggest that the HRT will be of particular value when the disease at-risk haplotypes are not derived from a very small number of founder individuals. In particular, for the models we considered, our threshold statistic $HRT_{T=t}$ greatly outperformed the MILC, the single-locus TDT and the multilocus transmission/disequilibrium test implemented in TRANSMIT when a founder population size of 500 or 1000 individuals was considered. These results are of particular interest because common complex traits are unlikely to be the consequence of a single or small number of common founder mutations/variants.

The $HRT_{T=t}$ statistics were particularly powerful when all available data were analyzed. Two unrelated haplotypes may be IBS for a few consecutive markers by chance alone, but with additional marker information only truly IBD

chromosome segments are likely to hold up to this strict criterion of a match. Herein, we chose to consider a modest (t=100) and a stringent (t=10,000) threshold value and we demonstrated clear differences in power between the two choices depending on the available marker data. For smaller marker sets including just five multiallelic markers, $HRT_{T=100}$ on average outperformed $HRT_{T=10,000}$. Reducing the threshold value, t, to 10 resulted in dramatic loss of power (data not shown). These results suggest that the optimal choice of the threshold value t will depend on the number and density of markers in addition to marker heterozygosity and population history. A real strength of the HRT is its ability to put to use many tightly-linked markers, thus allowing consideration of a high threshold value. The utility of the larger threshold value, t, clearly depends on the availability of large, dense sets of markers. The feasibility of a large value of t for a particular reference marker can be partially evaluated by looking at the total distribution of CHSS values between all pairings of haplotypes, blind to transmission status. If the total number of haplotype pairs that have a CHSS value that exceeds the threshold value is small, the choice of the threshold value is likely too large to ensure acceptable power. Our other HRT statistic, HRT_{ln}, generally did not perform as well as $HRT_{T=t}$ in our simulations. Under our simulation conditions, both HRT_{ln} and $HRT_{T=t}$ clearly outperformed the MILC statistic regardless of the number of founders or the chosen disease model. An area of future research is to explore additional haplotype sharing statistics and to try to design an algorithm that tailors the choice of the sharing statistic to be optimal for a given data set.

In addition to the threshold parameter t, the investigator also must specify a mismatch penalty π. In the analyses described, we chose π=0.05. When simulating error-free data, negligible differences were observed between the results using π=0.05 or π=0.00 (data not shown). Doubling the penalty parameter to π=0.10 in our simulation models resulted in a slight decrease in power. Based on these results, the choice of π=0.05 appears reasonable. It is also possible to assign different mismatch penalties to different markers. For example, one might assign greater values of π to microsatellite markers than to SNPs due to their greater mutability. In practice, the mismatch penalty π will only become relevant if the number of markers analyzed is sufficiently large to allow the imposed penalty to be overcome. If the number of markers is small or a mismatch occurs near the end of an ordered map of markers, then CHSS will likely take the value $CHSS_1$.

For the models we simulated, the HRT had slightly less power than TRANSMIT when the founder sample size was small (100 individuals) and actually resulted in lower power in some cases as compared to the results when applying HRT to a founder sample size of 500 individuals. An explanation of this small deficit in power for the HRT when using a very small founder sample size centers about the allele frequency estimates we used in calculation of the HRT statistics. For the HRT, we use estimated allele frequencies calculated from both the transmitted and non-transmitted alleles in our sample. Under the alternative hypothesis, this approach has the potential to make the test overly conservative due to the likely overestimation of the frequencies of the alleles that appear on the disease-associated haplotypes. This problem of overestimating allele frequencies is greatest when there is a small number of disease-associated founder haplotypes. To alleviate this problem when dealing with a small founder population, we could increase the density of markers, reduce the threshold value, or estimate allele frequencies from an outside well-matched sample to better estimate the allele frequencies in the general population. The latter approach would retain the correct nominal significance level and could increase the power of the test, particularly if one suspects a single common founder haplotype to dominate among the disease-associated haplotypes. In practice, however, this is not a concern as this scenario is unlikely for complex diseases. In place of weighting our statistics by estimated allele frequencies, we also considered using conditional marker-allele frequencies, conditioning on the allele at the adjacent marker nearest the reference marker, to account partially for any marker-marker LD. We found that using estimated conditional marker-allele frequencies from the analyzed collected parent-parent-offspring sample also made the test unnecessarily conservative.

One major difference between the HRT and MILC statistics is that the HRT statistic only considers sharing scores for pairs of transmitted haplotypes while the MILC statistic subtracts off the sharing scores for pairs of non-transmitted haplotypes. Under all simulation conditions, we calculated both the HRT and MILC statistics using both mechanisms of scoring. Both the HRT and MILC statistics were consistently more powerful

when calculated only using the scores from pairs of transmitted haplotypes. We believe there are two possible explanations for the profound differences in power observed between the two approaches. If there is minimal haplotype sharing between the non-transmitted haplotypes, then the two test statistics, corresponding to the two different approaches, will be very similar in magnitude. However, the variance for the approach using both transmitted and non-transmitted haplotype sharing would be considerably larger than the variance for the approach using just the transmitted haplotypes. On the other hand, if there is considerable sharing in both the transmitted and non-transmitted haplotype pairings, then under the alternative hypothesis, there will be heterogeneity of the haplotype frequencies with respect to the transmitted and non-transmitted haplotypes. Considering the differences in the amount of sharing between the two groups actually diminishes the impact of the heterogeneity. While not a rigorous proof, the intuition of this argument can be illustrated by a simple example. Consider a sample of five parent-parent affected trios where there are ten "A" transmitted haplotypes and ten "B" non-transmitted haplotypes. Using a threshold statistic such that A paired with A gives a score of 1, B with B gives a score of 1, and A with B gives a score of 0, then the HRT statistic is equal to $\binom{10}{2}=45$ when only considering the transmitted haplotypes and $\binom{10}{2}-\binom{10}{2}=0$ when subtracting off the non-transmitted haplotype scores from the transmitted scores. The first approach would clearly result in a significant finding while the latter approach clearly would not. Thus, by taking the difference between the scores from the transmitted haplotype pairings and the non-transmitted haplotype pairings, we have removed the impact of having sets of different transmitted and non-transmitted haplotypes.

The simulations we presented herein focused on a single *a priori* chosen reference marker. In practice, it is unlikely that application of the HRT would be performed using a single reference marker, but would instead be performed repeatedly using closing linked reference markers. Inherent to this application is the potential bias caused by uncorrected multiple testing. The HRT statistics and corresponding significance estimates between neighboring markers are likely to be highly correlated and consequently a Bonferroni correction would be overly conservative. The level of dependence between results at adjacent markers depends on a number of factors not easily measured, including the distance between the markers, the age of the population, allele frequencies, and the degree of LD between the markers. To test the overall significance of the HRT over a set of reference markers, one can apply the randomization procedure described for the MILC [Bourgain et al., 2000] or adapt the randomization approach described by Lazzeroni and Lange [1998]. The approach of Bourgain et al. [2000] takes the maximum value of the test statistics over all reference markers as the final test statistic. A randomization procedure is performed and a *P* value is calculated as the probability of observing a maximum sharing statistic over all reference markers that is at least as large as the observed final test statistic. The approach of Lazzeroni and Lange [1998] is to calculate the *P* value at each marker and use the smallest *P* value as the final test statistic. A randomization procedure can then be performed to determine the overall probability of observing a *P* value as small or smaller than the observed minimum *P* value. Assessing the relative merit of the different approaches for addressing multiple testing is beyond the scope of this report, but clearly is an important consideration for future research.

As we demonstrated in the simulations, the HRT can be particularly useful given a dense extended set of markers. The development and availability of dense sets of single nucleotide polymorphisms (SNPs) will make this an increasingly attractive feature of our method. We foresee the HRT being particularly useful when a specific region of interest has been identified, usually through linkage analysis, and a follow-up analysis of the region is conducted by saturating the region with additional markers. As with any other method designed to fine-map genes through LD, choice of the population being studied for the trait of interest is still likely crucial to success.

In addition to the observed increase in power for these simulations, the HRT has several properties that make it a highly flexible, useful approach. Unlike TRANSMIT, the HRT requires no allele or haplotype removal or lumping due to small sample size, excess degrees of freedom, or computational constraints. No restrictions, computational or otherwise, are made on the number of markers that can be considered. In theory, the more markers included in the analyses, the more power there will be to detect an association when present. Finally, because the HRT computes

a score for each haplotype pairing, it is trivial to identify likely related haplotypes from a sample of family trios. This option readily allows the user to delimit the location of a putative disease susceptibility locus by identifying key individual haplotypes, thus allowing quick identification of breaks in common founder haplotypes.

Software for the HRT is available and can be requested by e-mail: elange@wfubmc.edu.

ACKNOWLEDGMENTS

This work was supported by NIH grant HG00376 to M.B. E.M.L. was previously supported by NIH training grant HG00040. The authors thank Dr. Richard Gatti for sharing haplotype data from the ataxia-telangiectasia research project. We also thank Dr. Carl Langefeld and two anonymous reviewers for a helpful review of the manuscript, and Dr. Linda Green for programming assistance.

REFERENCES

Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F. 2000. Search for multifactorial disease susceptibility genes in founder populations. Am J Hum Genet 64:255–265.

Bourgain C, Genin E, Ober C, Clerget-Darpoux F. 2002. Missing data in haplotype analysis: a study on the MILC method. Ann Hum Genet 66:99–108.

Calafell F, Grigorenko EL, Chikanian AA, Kidd KK. 1998. Haplotype evolution and linkage disequilibrium: a simulation study. Hum Hered 51:85–96.

Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. Am J Hum Genet 65:1170–1177.

Clayton D, Jones H. 1999. Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65:1161–1169.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Gatti RA, Boder E, Vinters HV, Sparkes RS, Norman A, Lange K. 1991. Ataxia-telangiectasia: An interdisciplinary approach to pathogenesis. Medicine 70:99–117.

Good P. 1994. Permutation tests. New York: Springer.

Lange E, Boehnke M. 1998. A novel approach for identification of common ancestral haplotypes associated with complex phenotypes. Am J Hum Genet 63:A17.

Lazzeroni LC, Lange K. 1998. A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48:67–81.

Meuwissen THE, Goddard ME. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 33:605–634.

Nolte IM, te Meerman GJ. 2002. The probability that similar haplotypes are identical by descent. Ann Hum Genet 66: 195–209.

Risch N. 1987. Assessing the role of HLA-linked and unlinked determinants of disease. Am J Hum Genet 40:1–14.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–516.

Uhrhammer N, Lange E, Porras O, Naeim A, Chen X, Sheikhavandi S, Chiplunker S, Yang L, Dandekar S, Liang T, Patel N, Teraoka S, Udar N, Calvo N, Concannon P, Lange K, Gatti RA. 1995. Sublocalization of an ataxia-telangiectasia gene distal to D11S384 by ancestral haplotyping in Costa Rican families. Am J Hum Genet 57:103–111.

Van der Meulen MA, te Meerman GJ. 1997a. Association and haplotype sharing due to identity by descent, with an application to genetic mapping. In: Edwards JH, Pawlowitzki IH, Thompson EA, editors. Genetic mapping of disease genes. London: Academic Press Ltd.

Van der Meulen MA, te Meerman GJ. 1997b. Haplotype sharing analysis in affected individuals from nuclear families with at least on affected offspring. Genet Epidemiol 14:915–919.