Hum Genet (2010) 128:325–344
DOI 10.1007/s00439-010-0854-z

ORIGINAL INVESTIGATION

Genome-wide association filtering using a highly locus-specific transmission/disequilibrium test

María M. Abad-Grau · Nuria Medina-Medina · Rosana Montes-Soldado · José Moreno-Ortega · Fuencisla Matesanz

Received: 26 March 2010 / Accepted: 21 June 2010 / Published online: 6 July 2010

Abstract Multimarker transmission/disequilibrium tests (TDTs) are powerful association and linkage tests used to perform genome-wide filtering in the search for disease susceptibility loci. In contrast to case/control studies, they have a low rate of false positives for population stratification and admixture. However, the length of a region found in association with a disease is usually very large because of linkage disequilibrium (LD). Here, we define a multimarker proportional TDT ($mTDT_P$) designed to improve locus specificity in complex diseases that has good power compared to the most powerful multimarker TDTs. The test is a simple generalization of a multimarker TDT in which haplotype frequencies are used to weight the effect that each haplotype has on the whole measure. Two concepts underlie the features of the metric: the 'common disease, common variant' hypothesis and the decrease in LD with chromosomal distance. Because of this decrease, the frequency of haplotypes in strong LD with common disease variants decreases with increasing distance from the disease susceptibility locus. Thus, our haplotype proportional test has higher locus specificity than common multimarker TDTs that assume a uniform distribution of haplotype probabilities. Because of the common variant hypothesis, risk haplotypes at a given locus are relatively frequent and a metric that weights partial results for each haplotype by its frequency will be as powerful as the most powerful multimarker TDTs. Simulations and real data sets demonstrate that the test has good power compared with the best tests but has remarkably higher locus specificity, so that the association rate decreases at a higher rate with distance from a disease susceptibility or disease protective locus.

Introduction

Genome-wide genotyping of single-nucleotide polymorphisms (SNPs) can yield a few hundred thousand binary markers in a single chip array, providing a relatively unbiased examination of the entire genome for common risk variants. Many loci have been determined to be associated with multifactorial diseases using this new technology. However, in most cases, the information provided is not enough to localize the causal variant of the association. Nonetheless, genome-wide association studies yield useful information for better identification of an associated region that facilitates fine mapping of the region with a reduced number of markers.

There are two main types of genome-wide data association analyses: case–control studies and family-based studies. Although case–control association studies are the most common, they have high type I errors because of population stratification (Spielman et al. 1993; Zhang et al. 2003). In family-based studies, transmission/disequilibrium tests (TDTs) are powerful tests requiring only family trios with both parents and one affected offspring. In contrast to case–control studies, TDTs are known to be robust for population structures. Therefore, they are an interesting

M. M. Abad-Grau and N. Medina-Medina have contributed equally to this paper.

M. M. Abad-Grau (✉) · N. Medina-Medina · R. Montes-Soldado · J. Moreno-Ortega
Departamento de Lenguajes y Sistemas Informáticos, CITIC, Universidad de Granada, Granada 18071, Spain
e-mail: mabad@ugr.es

F. Matesanz
Instituto de Parasitología y Biomedicina López Neyra, Consejo Superior de Investigaciones Científicas, Granada, Spain

alternative to case–control studies when family trios can be genotyped. The classic single-marker biallelic TDT can detect association due to linkage. Multimarker generalizations of the classic TDT enhance it by detecting marker interactions, such as when a trait does not depend on a single marker but there is association when considering more than one marker together, which may point to linkage disequilibrium (LD) or gene–gene interaction (epistasis). This may be the case for genome-wide genotyping in which a disease susceptibility locus cannot be genotyped but some markers in LD with the locus can be. Thus, the power of a multimarker TDT can significantly enhance that reached by a single TDT.

Different approaches have been used to define multimarker TDTs, each of them computing statistical significance in a different way. The most widely used are: (1) TDTs that are straightforward extensions of the classic single-marker biallelic TDT; (2) TDTs that group haplotypes to reduce the degrees of freedom (*df*); and (3) TDTs based on haplotype similarities to reduce *df* and improve the test power.

The idea behind the first of the approaches is simple. In nuclear families with one affected child, there must be a difference between the counts for non-transmitted and transmitted haplotypes if they are directly associated with the disease or in linkage with a susceptibility locus. The most commonly used test in this approach is the classic *multimarker TDT (mTDT)* (Spielman and Ewens 1996; Lazzeroni and Lange 1998), a straightforward extension of the biallelic monomarker TDT that can be used by considering each haplotype as a particular allele (Sham 1997; Bourgain et al. 2001). Using this approach, we can also consider introducing some non-linear transformation to the transmitted/non-transmitted haplotype counts, such as TDT_E (Zhao et al. 2007), which is based on the concept of entropy. More specific tests have also been defined to improve power for uncertain transmission cases (Clayton 1999; Zhao et al. 2000) or genotyping errors (Gordon et al. 2001). The main problem with tests using this approach is that the *df* of the approximate χ^2 distribution increase with the number of haplotypes and thus permutation tests to determine the null distribution may be required for sparse data.

The second approach tries to reduce the *df* by grouping haplotypes using different criteria such as haplotype distance (Li et al. 2001) or a haplotype evolutionary relationship (Seltman et al. 2001). These tests are very time-consuming when used in genome-wide searches, as they have to first infer a model to group the haplotypes. As an example, a cladogram for which it is assumed that there are no recurrent disease mutations and no recombination or gene conversion must be estimated. Violation of these strong assumptions may decrease the general accuracy of the test.

The third approach also tries to reduce the *df* using haplotype similarities. However, instead of counts for the haplotype groups, similarity metrics are used, such as the length measure used in the length contrast test (TDT_{LC}) (Yu et al. 2005) and the signed rank test (TDT_{SR}) (Yu et al. 2005) and other metrics such as those used in the maximum identity length contrast (MILC) test (Bourgain et al. 2001) and the haplotype-sharing TDT (HS-TDT) (Zhang et al. 2003). For the TDT_{LC} and TDT_{SR} tests it is assumed that there must be less variation among haplotypes transmitted to affected offspring than among non-transmitted haplotypes, as they distinguish the sign of the difference in the measure between transmitted and non-transmitted data sets. However, TDTs based on this assumption are more specific than multimarker TDTs because they do not detect statistically significant differences in haplotype similarities when these are greater among non-transmitted haplotypes. This may occur when a haplotype is not in linkage with a disease susceptibility gene but with a protective gene, so that it will be more frequent in healthy individuals. There is a more important issue in similarity-based TDTs: similarity measures are computed by pairwise comparisons between individuals. Thus, their computational complexity is a quadratic function of the number of founders, in contrast to most of TDT measures, which use sample counts and increase linearly with the number of individuals. For current genotype samples with up to a few thousand individuals, similarity-based TDTs are thus a real burden.

Our goal was to define a computationally feasible multimarker measure, named a *proportional mTDT* ($mTDT_P$), with high power and high robustness for population admixture and stratification with high locus specificity as an association test. Therefore, association rates are expected to quickly decrease with distance from a disease susceptibility or protective locus. The measure belongs to the first of the approaches and is a generalization of *mTDT* that weights partial results for each haplotype by its probability frequency. The success of the measure in improving locus specificity is based on two assumptions: (1) according to the decrease in LD with chromosomal distance, the frequency of haplotypes in linkage with a disease haplotype is higher at shorter distances from the disease locus; and (2) according to the 'common disease, common variant' (CDCV) hypothesis, disease susceptibility variants are quite common in complex diseases and a combination of several genes, rather than a single gene, together with environmental factors, causes the disease. A consequence of these assumptions is that haplotypes in very strong LD with a disease or protective variant are common and their frequency will notably decrease with chromosomal distance.

Therefore, under both extremes of the expectrum of chromosomal distances (the null hypothesis of no linkage and no distance to the disease locus), there must be little difference between $mTDT_P$ and $mTDT$; as we depart from these, differences between the two tests arise: association detected by $mTDT_P$ will decrease more rapidly as we depart from the disease locus.

In "Methods", after analysis of *mTDT* and the reasons why it cannot be considered a highly locus-specific test, we propose $mTDT_P$, a modification of *mTDT* that considers differences in haplotype frequencies to improve both specificity and sensitivity. "Simulation studies" compares different multimarker TDTs for different genetic models, relative risks, haplotype lengths and total disease susceptibility loci. As mentioned above, our goal was not only to study test power and robustness under different configurations, but also to observe the rate at which statistical significance decreases with chromosomal distance. Simulations to study association rates at different chromosomal distances from a disease susceptibility locus have been performed for single-marker TDTs (Zhao et al. 2007). The "Simulation studies" compare sensitivity, specificity and robustness for some state-of-the-art multimarker TDTs defined under different approaches. In "Real data sets", we compare the power and locus specificity of our test ($mTDT_P$) with other TDTs using real trio samples for Crohn and multiple sclerosis (MS) diseases and robustness using control trio samples of individuals from the International Hapmap Project (IHP) (HapMap-Consortium 2003), and finally "Discussion".

Methods

Assume that the data represent M nuclear families in which one child is affected and that L SNPs are genotyped for all the family members. As an example, for $L = 2$ and assuming biallelic SNPs, there will be only $k = 4$ different haplotypes: AB, Ab, aB and ab. Consider a sample composed of all transmitted and non-transmitted haplotypes when the parents are heterozygotic. Let n be the sample size. Thus, subsamples S_T and S_U of transmitted and non-transmitted haplotypes, respectively, both contain $n/2$ haplotypes.

Analysis of *mTDT*

mTDT (Spielman and Ewens 1996) was first proposed as a multiallelic extension of the simple biallelic TDT. However, by considering haplotypes instead of alleles, the test can also be used as a multimarker TDT. The test is defined as:

$$mTDT = \frac{k-1}{k}\sum_{i=1}^{k}\frac{(n_{iT}-n_{iU})^2}{n_{iT}+n_{iU}},$$

where k is the number of different alleles/haplotypes and n_{iT} and n_{iU} are the number of times allele/haplotype i is transmitted or not transmitted, respectively, considering only heterozygous parental genotypes. The measure asymptotically follows a χ^2 distribution with $k-1$ *df* (χ^2_{k-1}) under no linkage if all heterozygous parental genotypes have the same frequencies. A modification of *mTDT*, $mTDT_s$, was defined to guarantee it follows a χ^2_{k-1} distribution under the null hypothesis for every frequency for heterozygous parental genotypes (Stuart 1955; Sham 1997).

Both *mTDT* and $mTDT_s$ give all haplotypes the same weight, regardless of their frequencies, as each summand is the square of a standard normal distribution under the null hypothesis. Even under the null hypothesis, the variability in haplotype frequency is usually very high, with some haplotypes very frequent and others very rare. Therefore, the assumption that differences in transmission of multimarker haplotypes follow a χ^2 distribution under the null hypothesis of no linkage leads to a test that is too simplistic and unrealistic. The larger the haplotypes, the greater is the departure of the true null distribution from a χ^2_{k-1} distribution, as there are more differences among haplotype frequencies.

We explore the consequences of this simplification once we introduce a generalization of *mTDT* that considers differences in haplotype frequencies.

Definition of $mTDT_P$

The test we propose here comprises a simple change in *mTDT*, with weighting of the summand of each haplotype by the haplotype frequency $\frac{n_i}{n}$. Thus, $mTDT_P$ is defined as:

$$mTDT_P = \sum_{i=1}^{k}\frac{n_i}{n}\frac{(n_{iT}-n_{iU})^2}{n_i} = \sum_{i=1}^{k}\frac{(n_{iT}-n_{iU})^2}{n},$$

where n is the overall number of haplotypes in parental heterozygous genotypes (i.e., twice the number of heterozygous parents).

Factors n_i/n, $\forall i \in 1,\ldots, k$ weight haplotypes according to their frequencies, which means that differences in transmission for the most frequent haplotypes have a greater effect on the measure.

Taking into account that haplotype counts are correlated, the asymptotic variance of $mTDT_P$ under the null hypothesis is derived in Appendix 1.

It is already known (Sham 1997) that *mTDT* follows a χ^2_{k-1} distribution under the null in the case of equal parental genotype frequencies. Therefore, it is straightforward to show that $mTDT_P$ under the same situation of equal parental genotype frequencies is equal to $mTDT/(k-1)$ so that it follows a scaled χ^2_{k-1}:

$(k-1)mTDT_P \sim \chi^2_{k-1}$.

Under different genotype frequencies, the variance (Appendix 1) is larger than $\frac{2}{k-1}$, so that, as it occurs with *mTDT* (Sham 1997), TDT_P will tend to be anticonservative. A feature of this measure is that it reduces the impact of random effects due to rare haplotypes without the need of imposing a lower bound in haplotype counts for haplotypes to be used, as is usually done by *mTDT* (Sham and Curtis 1995).

But the main feature of $mTDT_P$ is that, in contrast to most multimarker TDTs which lack either in power or in locus specificity, $mTDT_P$ has both: a high power and a high locus specificity to detect disease susceptibility or disease protective loci in complex diseases. The reason for the measure to be comparable in power to the powerful *mTDT* is that, assuming the CDCV hypothesis, the impact that non-recombinant haplotypes have on the measure is high when chromosomal distance to a disease locus is very short, as their frequencies are high and so are their weights. As we depart from the disease locus, the recombination factor increases, non-recombinant haplotypes will be less frequent in haplotypes transmitted to affected children and their impact in the whole measure will decrease faster than when weighting is not used, as in *mTDT*.

In order to characterize the distribution of $mTDT_P$ under the null hypothesis of no linkage to avoid using permutation tests to assess statistical significance we will first consider the simpler but unrealistic situation of haplotype counts being obtained from independent samples ("Independent random variables: characterization and approximation of a weighted χ^2 distribution") as a starting point to consider dependencies among them ("Dependent random variables: approximation of $mTDT_P$ under the null hypothesis").

Independent random variables: characterization and approximation of a weighted χ^2 distribution

Under the null hypothesis of no linkage, $Y_i^2 = \frac{(n_{iT}-n_{iU})^2}{n_i}$ follows a χ^2_1 distribution. If Y_i^2 were independent distributions, $mTDT_P$, which is defined as weighted summands, would asymptotically follow a weighted χ^2 distribution $W_{k,\mathbf{w}}$ of k independent χ^2_1 distributions:

$$W_{k,\mathbf{w}} = \sum_{i=1}^{k} w_i \chi^2_1,$$

with weights $\mathbf{w} = (\frac{n_1}{n}, \ldots, \frac{n_k}{n})$.

It is straightforward to show that $W_{k,\mathbf{w}} = (w_1, \ldots, w_k)$ can be considered a generalization of $\frac{\chi^2_k}{k}$ (χ^2_k being a sum of $k\chi^2_1$ distributions) in which each χ^2_1 are weighted with the only restriction $\sum_{k=1}^{k} w_i = 1$. As $mTDT_P$ imposes the weights to be $\mathbf{w} = (\frac{n}{n_1}, \ldots \frac{n}{n_k})$, in the case of equal parental genotype frequencies and ignoring dependencies among haplotypes (we will consider dependencies in the "Dependent random variables: approximation of $mTDT_P$ under the null hypothesis"), it would follow a $\frac{\chi^2_k}{k}$ distribution under the null hypothesis of no linkage, a distribution whose variance is $2k$. Therefore, $mTDT_P$ for equal parental genotype frequencies would have variance $\frac{2}{k}$, as $Var(\frac{X_k}{k}) = \frac{2k}{k^2} = \frac{2}{k}$. In general, the variance of a weighted χ^2_1 distribution $W_{k,\mathbf{w}} = (w_1, \ldots, w_k)$ is known to be (Johnson et al. 1994):

$$Var(W_k) = 2\sum_{i=1}^{k} w_i^2.$$

The computation of the distribution function of $W_{\mathbf{w}} = (w_1, \ldots, w_k)$ is very complicated because of numerical integration (Solomon and Stephens 1977; Gabler and Wolff 1987). As we are interested in a TDT for genome-wide association filtering, permutation tests should be avoided and an easily computable approximation of the asymptotic test distribution under the null hypothesis is required.

Several approximations (Solomon and Stephens 1977; Gabler and Wolff 1987; Castao-Martínez and López-Blázquez 2005) are available for a weighted sum of independent χ^2 distributions $W_{k,\mathbf{w}} = (w_1, \ldots, w_k)$. The one used here is based on two limiting distributions that are identical to W in the first three moments, with only minor differences in higher moments (Gabler and Wolff 1987). Given a statistic s, $Pr(W \leq s)$ is computed by choosing the shortest value from the two limiting distributions:

$$p(W \leq s) = \min \begin{cases} G(s) = \sum w_i \gamma\left(\frac{1}{2w_i}, \frac{s}{2w_i}\right) \\ U(s) = \gamma\left(\frac{k}{2}, \frac{s}{2\delta}\right) = Pr(\chi^2_k \leq s/\delta), \end{cases} \tag{1}$$

where $\gamma(a, b)$ is the normalized lower incomplete gamma function (Abramowitz and Stegun 1972), also called the incomplete gamma function, and $\delta = \prod {w_i}^{1/k}$.

It is straightforward to show that in the case of equal weights ($w_i = \frac{1}{k}, \forall i\{1, \ldots, k\}$), $\delta = \frac{1}{k}$ and the approximation turns out to be a true weighted χ^2 distribution, as the three distribution functions are exactly the same.

Dependent random variables: approximation of $mTDT_P$ under the null hypothesis

As each individual carries a pair of haplotypes, haplotype counts are not obtained from independent samples. Therefore, $Y_i^2, i \in \{1, \ldots, k\}$ are not independent χ^2_1 variables and thus $mTDT_P$ under the null is not $W_{k,\mathbf{w}=(\frac{n_1}{n}, \ldots, \frac{n_k}{n})}$. Therefore, the exact distribution needs to be assessed.

As it was said above, under the null hypothesis of no linkage and when the frequencies of all parental heterozygous genotypes are equal, *mTDT* asymptotically follows a χ^2_{k-1} distribution and, therefore, $mTDT_P = (k-1)mTDT$ a scaled χ^2_{k-1}. For $k = 2$, the asymptotic variance is 2. Moreover, for $k = 2$, $mTDT_P$ also reduces to the simple (i.e., monomarker, monoallelic) TDT.

To use the approximation of the weighted sum of χ^2 distributions *W* considered above (Gabler and Wolff 1987) in order to obtain the distribution of $mTDT_P$ under the null hypothesis and considering that the χ^2 distributions are not independent, we have modified the limiting distributions *G* and *U* so that it can be easily shown they will be exactly a scaled χ^2_{k-1} with scale factor $k-1$ under equal genotype heterozygous frequencies.

Therefore, the approximation will be:

$$Pr(W \le mTDT_P) = \min\begin{cases} \sum w_i \gamma\left(\frac{1}{2w_i}\frac{(k-1)}{k}, \frac{mTDT_P}{2w_i}\frac{(k-1)}{k}\right) \\ Pr\left(\chi^2_{k-1} \le mTDT_P\frac{(k-1)}{k}/\delta\right), \end{cases} \quad (2)$$

where $w_i = \frac{n_i}{n}$ and $\delta = \prod_{i=1}^{k} w_i^{1/k}$. A pseudo code describing how to compute *p* values is given at Table 1 and a computer program to compute it is provided at the supplementary website.

In order to check whether $mTDT_P$ follows a weighted χ^2 distribution in the more general case of different parental heterozygous genotype frequencies, we performed permutations in "Simulation studies" (Zhang et al. 2003; Yu et al. 2005) and we did not find significant differences (data not shown).

Table 1 Pseudo code describing how to compute *p* values for $mTDT_P$ using the approximation given in Eq. 2 (Gabler and Wolff 1987)

```
Inputs:
k: the number of different haplotypes in the sample
weights: a list of k weights
HP: the value of statistics $mTDT_P$ for the current sample
Output:
result: p value
Description:
$result = 0$
$DS = 1$
$R1 = 0$
$df = k - 1$
Foreach haplotype $i = 1,\ldots,k$
    $dZero = 0.5/weights(i)$
    $R1 = R1 + weights(i)*gammai(dZero, HP*dZero)$
    $DS = DS*weights(i)$
$R2 = pValTestChiSquare(HP/DS^{1/k}, k)$
$result = max(R1, 1 - R2)$
```

gammai(*a*, *b*) is the normalized lower incomplete gamma function, *pValTestChiSquare*(*a*, *b*) computes the *p* value for *a* using χ^2_b, i.e., $p(\chi^2_k \le a)$, and *max*(*a*, *b*) returns the maximum of *a* and *b*

Simulation studies

We compared the performance of our solution $mTDT_P$ with several state-of-the-art multimarker TDTs, such as the classic *mTDT* and other TDTs based on different approaches: the similarity-based tests $mTDT_{LC}$ and $mTDT_{SR}$, the entropy-based $mTDT_E$ and the group-based $mTDT_{T1}$. $mTDT_{1T}$ (Ott 1999) is a χ^2_1 test under the null hypothesis of no linkage that checks differences between the haplotype with more significant differences $n_{iT} - n_{iU}$ and the rest of the haplotypes in a sample.

We also modified *mTDT* using some well-known corrections of χ^2 tests to improve the specificity by reducing random errors due to low frequencies and some modifications of these (Appendix 2), such as the Yates (1934) correction $mTDT_Y$, its modification $mTDT_{YP}$ and the Laplace corrections $mTDT_{L1}$ and $mTDT_{L2}$.

Besides robustness to population stratification and power, we are interested in measuring locus specificity. Thus, the decrease in the rate of associations detected with incremental linkage distance or recombination rates (θ) was assessed considering the extreme points from $\theta = 0$ for which all associations detected are true positive associations (power) and from $\theta = 0.0002$ for which most associations detected are type I errors.

Statistical significance levels were obtained using a permutation procedure for $mTDT_{LC}$, $mTDT_{SR}$ and $mTDT_E$ (Zhang et al. 2003; Yu et al. 2005). For $mTDT_P$, the approximation of a weighted χ^2 with weights being the haplotype frequencies was used ("Independent random variables: characterization and approximation of a weighted χ^2 distribution"). For the remaining tests, the exact χ^2 distribution was used.

Simulation set-up

We tried to reproduce the same simulations used in several studies to check TDT accuracy (Zhang et al. 2003; Yu et al. 2005) and explained in the following subsections.

As our main goal is to have a useful test to perform genome-wide association filtering, computational complexity is a main issue and a linear relationship between computational complexity and the number of SNPs is highly desirable. Therefore, we applied the tests in a very feasible way in which only consecutive or overlapping clusters of SNPs (known as sliding windows) were tested together. For simulations of a cluster as suggested by Crawford et al. (2004), we assumed that recombination rates among all the markers tested is very low, which is equivalent to assuming that they belong to the same low-recombination block (Daly et al. 2001). The recombination fraction within blocks (θ_B) for a common population with exponential growth, such as an African population, has been estimated as 0.000088 (Hinds et al. 2005) and we used this value in the simulations.

We also modified the method for introducing a disease mutation compared to other studies (Sham 1997; Zhang et al. 2003; Yu et al. 2005). Instead of considering only one ancestral chromosome with the disease-causing mutation, or the improvement of using two ancestral chromosomes (Zhang et al. 2003), a more realistic simulation of inheritance of complex diseases was used, in which the number of ancestral disease chromosomes can change according to the coalescent model, as any other gene does.

Populations were drawn using msHOT (Hellenthal and Stephens 2007), a program for generating samples based on the coalescent model that incorporates recombination. The samples for all the populations were obtained using *trio-Sampling*, a computer program available on the supplementary website. In the following subsections, we describe the simulations in detail and highlight any departures from the set-up commonly used (Sham 1997; Zhang et al. 2003; Yu et al. 2005). A more detailed explanation of the simulations performed can be accessed on the supplementary website.

Robustness

To check the robustness to population stratification, simulations were performed as described by Zhang et al. (2003) and Yu et al. (2005). Therefore, we considered stratified populations. However, instead of using samples of 200 nuclear families (Zhang et al. 2003; Yu et al. 2005), we produced samples with 500 nuclear families. Moreover, we used recombination fraction from the markers to the disease locus $\theta = 0.5$ to represent a true null. Association rates were estimated based on 1, 000 replications. Families were randomly sampled by choosing haplotypes with the disease mutation and randomly choosing the haplotypes transmitted to children considering recombinations. For the first subpopulation, the minor allele frequency (MAF) for the markers was 0.5 and the probability of the disease mutation in parents p_D was 0.2. For the second subpopulation, different MAFs q for the markers were used: $q \in \{0.1, 0.3, 0.5\}$ and p_D was 0.3. Different proportions of individuals from the first sample were used, $pp \in \{1/2, 1/4, 1/6\}$. Therefore, by varying pp and q, nine different scenarios where considered to test the robustness.

Locus specificity and sensitivity

Simulations for power (sensitivity), i.e., assuming no recombination between the disease susceptibility locus and the markers tested, were similar to those used in several studies assuming one founder disease haplotype (Lam et al. 2000; Zhang et al. 2003; Yu et al. 2005), except that SNPs used were assumed to be in high LD, i.e., they belong to the same low-recombination block (Daly et al. 2001). Therefore, we performed simulation analyses using haplotype data sets for 200 nuclear families (family trios with both parents and an affected child). Association rates were estimated based on 100 replications of the simulations described below (Sham 1997; Zhang et al. 2003; Yu et al. 2005).

Four parameters were taken into account to generate samples from populations (one for each population). Table 2 shows the parameters and their values. The first parameter, the relative risk of being homozygous for the risk allele, *RR*, was varied from 2 to 10 in steps of 2 in the simulations. The second parameter is the number of disease loci used: one and two different disease susceptibility loci were considered. The third parameter is the genetic disease model. Affected and non-affected individuals were drawn by considering different genetic models for one and two disease susceptibility loci (Yu et al. 2005). Additive, dominant and recessive models were considered for only one locus. Additive, domAndDom, domOrDom, recOrRec, threshold and modified models were considered for two loci. Different relative genotype risks (RR) of having genotype *DD*, defined as $Pr(disease \mid DD)/Pr(disease \mid dd)$ (one disease locus) and of having joint genotypes *DD* and *EE*, defined as $Pr(disease \mid DD, EE)/Pr(disease \mid dd, ee)$ (two disease loci), where d and e are the normal alleles and D and E the disease alleles, were used. Relative risks for all other genotypes were computed based on RR (Fan and Xiong 2001; Yu et al. 2005) (see Table S1 on the supplementary website).

The fourth parameter checks the decrease in association rate due to chromosomal distance. We considered five different recombination fractions (θ) from the markers to the disease susceptibility locus, ranging from perfect LD (no recombination) to $\theta = 0.0002$. Use of the recombination fraction to choose markers for the samples forced us to modify the pattern of population growth to simulate the LD decrease with distance in a more realistic way in a human population (Kruglyak 1999; Crawford et al. 2004). For greater consistency with real populations and complex diseases in which different numbers of founders can carry the disease loci, we used the coalescent model (Nordborg 2001) to draw populations with a variable number of founder haplotypes and population growth as explained

Table 2 Values used to configure sample parameters used in specificity/sensitivity simulations

Relative risk	2, 4, 6, 8, 10
Genetic model	Additive, recessive, dominant
θ to disease loci	0, 5e−05, 1e−04, 1.5e−04, 2e−04
Haplotype length	1, 2, 4, 6, 8, 10

above. Any position can be a disease susceptibility locus. Disease founder haplotypes were chosen by selecting one SNP with a mutant allele with frequency in the interval [0.2, 0.4] to mimic a common disease (Yu et al. 2005).

We later produced a second set of simulations with more realistic relative risks (1.2, 1.6, 2.0, 2.4 and 2.6) and samples of 500 nuclear families and focused only in the most powerful statistics which were also highly efficient (computational complexity linear to the number of families).

In order to know how frequencies of the disease mutation affect $mTDT_P$ and the other measures, we generated a third set of simulations with same parameters as the second one but considering the frequency of the disease mutation in the interval [0.1, 0.2].

Simulation results

The sensitivity and specificity of the tests were analyzed by counting rates of association for different chromosomal distances from markers to disease loci.

Table 3 shows type I error results for $mTDT_P$ in the presence of population stratification and admixture for nominal levels of $\alpha = 0.01$ and $\alpha = 0.05$. Values shown are rates of samples in which association was found to be statistically significant, for all configurations of pp and q values used. As TDT_P is a scaled χ^2 distribution only under equal parental genotype frequencies and its variance is larger without this constraint, the measure tends to be anticonservative, so that p values may upward deviates from the nominal value. However, $mTDT_P$ is mainly proposed to perform genome-wide search to be more locus specific than the current alternatives. Moreover, association p values will be averaged in a sliding window approach and only associations found at few consecutive windows (considering enough marker density) will be considered to perform a further fine mapping.

Table 3 Type I error rates in presence of population stratification and admixture and recombination factor the the disease locus 0.5 based on 1,000 simulations

α	MAF	pp	
0.01	0.1	0.5	0.009
0.01	0.3	0.5	0.012
0.01	0.5	0.5	0.013
0.01	0.1	0.75	0.012
0.01	0.3	0.75	0.016
0.01	0.5	0.75	0.015
0.01	0.1	0.833	0.011
0.01	0.3	0.833	0.013
0.01	0.5	0.833	0.013
0.05	0.1	0.5	0.054
0.05	0.3	0.5	0.063
0.05	0.5	0.5	0.071
0.05	0.1	0.75	0.060
0.05	0.3	0.75	0.061
0.05	0.5	0.75	0.052
0.05	0.1	0.833	0.055
0.05	0.3	0.833	0.056
0.05	0.5	0.833	0.058

Results for different MAF in the second subpopulation (q) and different proportion of trios from the first subpopulation (pp), obtained by TDT_P for nominal levels $\alpha = 0.01$ and $\alpha = 0.05$

Results for sensitivity ($\theta = 0$) show that $mTDT$, $mTDT_{1T}$ and $mTDT_P$ achieve the best results under all scenarios tested, with little differences among the three of them, whereas locus specificity results ($\theta \in \{0.00005, 0.0001, 0.00015, 0.0002\}$) show that $mTDT_P$ has better performance than all the other methods. Therefore, association rates decrease faster with $mTDT_P$ than with the other methods whenever recombination fraction θ to the disease locus increases. These differences are more appreciable when we increase RR and haplotype length.

Therefore, Fig. 1 shows results ($\alpha = 0.05$) for haplotypes of length 4 and $RR = 6$ for one (first column) and two disease loci (second and third columns) under different disease models. For clarity, in this plot $mTDT_P$ was only compared with the two other TDTs that showed the highest power in all our simulations, $mTDT$ and $mTDT_{1T}$. Figures 2 and 3 show results for the same configurations used in Fig. 1 except that haplotype length is 10 and RR is 4 and 8, respectively. In general, differences between the tests increase with haplotype length and relative risk, with greater differences for haplotypes of length 10 and $RR = 8$ (Fig. 3) than for smaller haplotype length and/or RR (Fig. 1).

Results for $\alpha = 0.05$ and haplotype lengths of 1, 2, 4, 6, 8 and 10 for one locus are available on the supplementary web site (Figures S1–S6). Results for two loci and disease models Additive, DomOrDom and RecOrRec (Figures S7–S12) and for two loci and disease models DomAndDom, Threshold and Modified) Figures S13–S18) are available on the supplementary web site. We also used the corrections to the small data problem mentioned in Appendix 2 (Figs. S19–S36). As expected, the same pattern was always observed: all the corrections improved the specificity at a cost of a reduction in sensitivity. The higher the correction, the stronger was this pattern. It should be noted that for haplotypes of length 1, i.e., only one marker, $mTDT$, $mTDT_{1T}$ and $mTDT_P$ are equivalent and therefore yield the same results. Differences among them increase with haplotype length.

As $mTDT$ and $mTDT_P$ showed a constant pattern of higher power than the other statistics for all the scenarios provided, we focused in them together with $mTDT_Y$, the measure that performs the lightest correction to the small

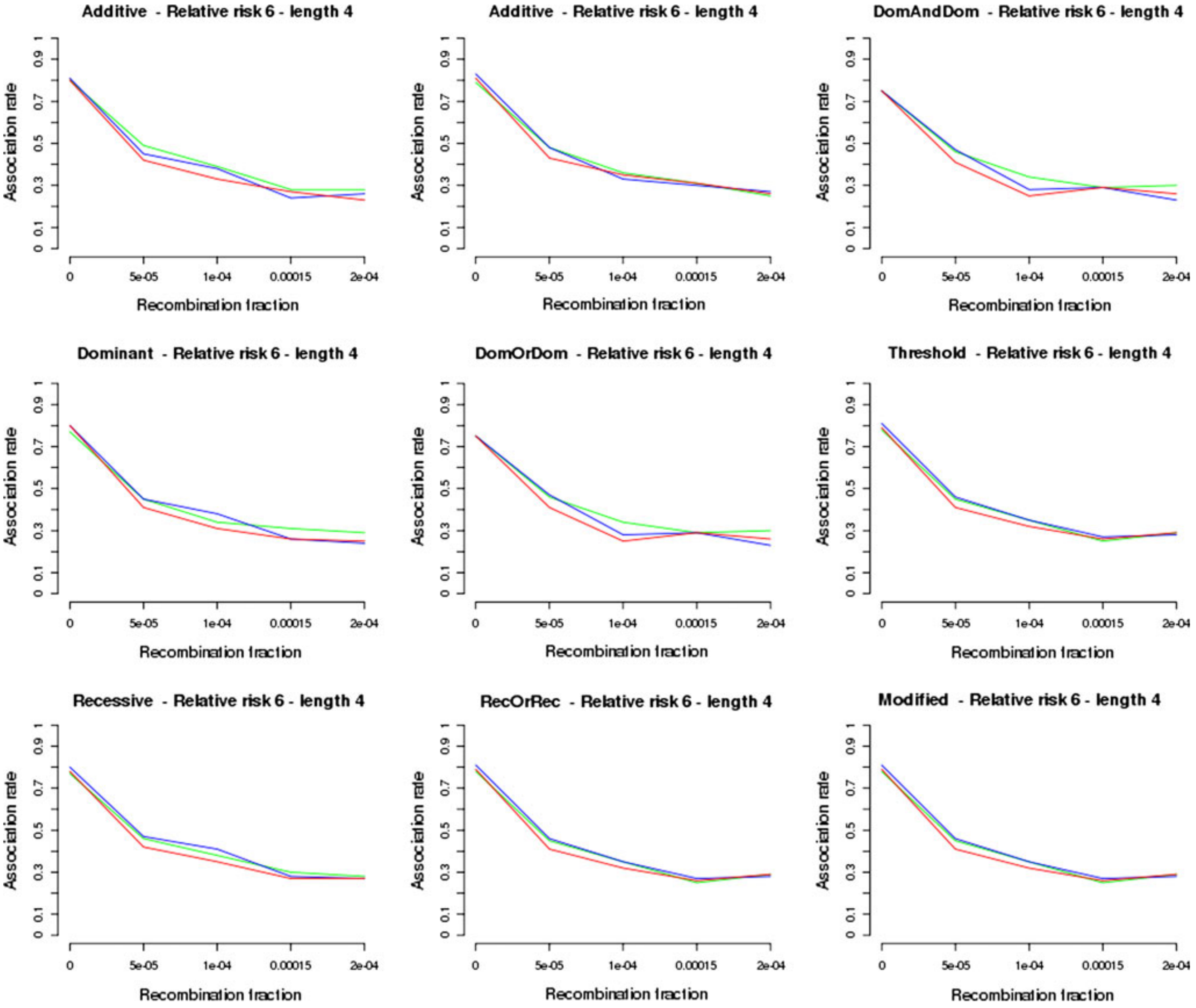

Fig. 1 Association rate based on 100 simulations of 200 family trios as a function of the recombination rate using haplotypes of length 4 and different genotype models (*rows*). The *first column* shows results for one disease susceptibility locus and the *second* and *third* show results for two disease loci. A nominal level of $\alpha = 0.05$ and a relative risk of 6 were used for all plots. Results for $mTDT_P$, $mTDT$ and $mTDT_Y$ are plotted in *red*, *blue* and *green*, respectively

data problem. Disregarding $mTDT_{LC}$ and $mTDT_{SR}$ made feasible to perform a second and third set of simulations using a larger number of nuclear families: 500. We did not use $mTDT_{1T}$ because it chooses the haplotype with the highest power and therefore it requires multitesting correction. When we used Bonferroni correction (data not shown) the measure was not competitive any more, in agreement with the already referred over-correct association results (Tang et al. 2009).

Using the second set of simulations, Figs. 4, 5 and 6 show association rates of $mTDT$, $mTDT_P$ and $mTDT_Y$ (blue, red and green, respectively) for nominal level $\alpha = 0.05$ and relative risks of 2, 1.6 and 2.4 and haplotypes of lengths 4, 10 and 10, respectively. By increasing the number of samples, the power increases and associations can be detected even with lower and more realistic relative risks. Differences among the three tests can still be observed for all the scenarios used.

Results for $\alpha = 0.05$ and haplotype lengths of 1, 2, 4, 6, 8 and 10 for one locus are available on the supplementary web site (Figs. S37–S42). Results for two loci and disease models Additive, DomOrDom and RecOrRec (Figs. S43–S48) and for two loci and disease models DomAndDom, Threshold and Modified (Figs. S49–S54) are available on the supplementary web site.

The third set of simulations was produced in order to analyze how the frequency of the disease mutation affects power in the three measures. Figure 7 shows association

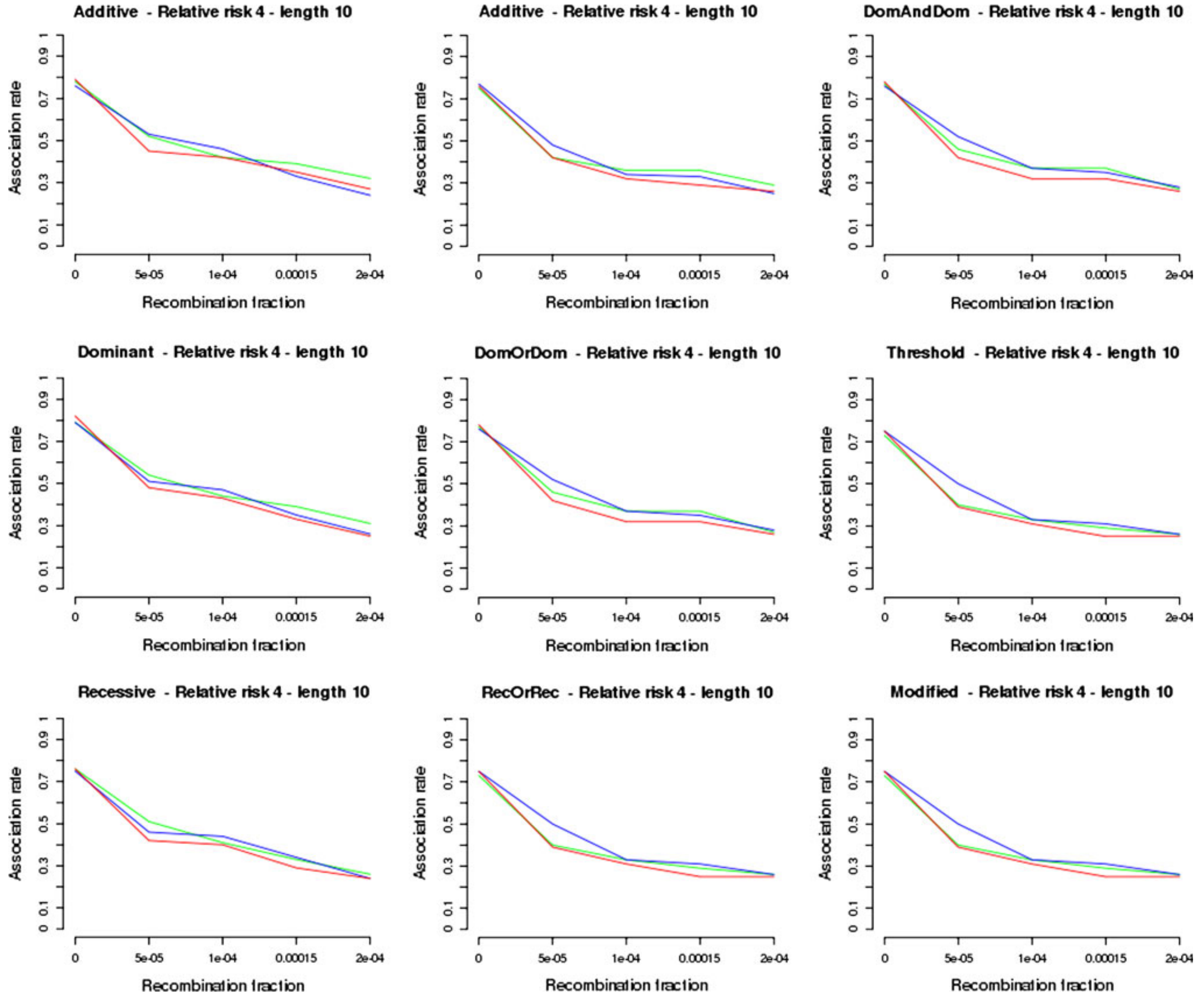

Fig. 2 Association rate based on 100 simulations of 200 family trios as a function of the recombination rate using haplotypes of length 10 and different genotype models (*rows*). The *first column* shows results for one disease susceptibility locus and the *second* and *third* show results for two disease loci. A nominal level of $\alpha = 0.05$ and a relative risk of 4 were used for all plots. Results for $mTDT_P$, *mTDT* and $mTDT_Y$ are plotted in *red*, *blue* and *green*, respectively

rates (100 simulations of 500 family trios each were used) for haplotypes of length 10 and different relative risks (*x*-axis). Simple lines show results for disease mutation frequencies in the interval [0.2, 0.4] while lines with diamonds show results for disease mutation frequencies in the interval [0.1, 0.2]. On light of these plots, two main results derive: for equal relative risk (1) power is lower with larger frequencies and (2) *mTDT* approaches $mTDT_P$ with low frequencies and even outperforms it. A possible reason for the first result, i.e., a higher power with lower mutation frequencies and equal relative risks, is that lower frequencies usually mean more recent mutations and there are less chances of recombinations between the disease variant and the neighboring haplotype so that larger differences would exist between transmitted and non-transmitted counts. It has to be noted that we used one and two disease loci simulations. The more loci were involved, the higher the chances of having a sample with no individuals with the disease variant at one locus, as relative risk can be still high because of the presence of a disease variant at a different locus and therefore power would decrease. The second result can be explained by considering again that lower disease variant frequencies usually mean more recent mutations. Hence, most of the neighboring SNPs already mutated and many different haplotypes arose so that there are less chances for the disease mutation to occur at a chromosome with a very common haplotype. Thus, a non-recombinant haplotype with a disease variant will have a lower frequency. Therefore, weighting transmission disequilibrium by haplotype frequencies will reduce power compared with the lineal mTDT.

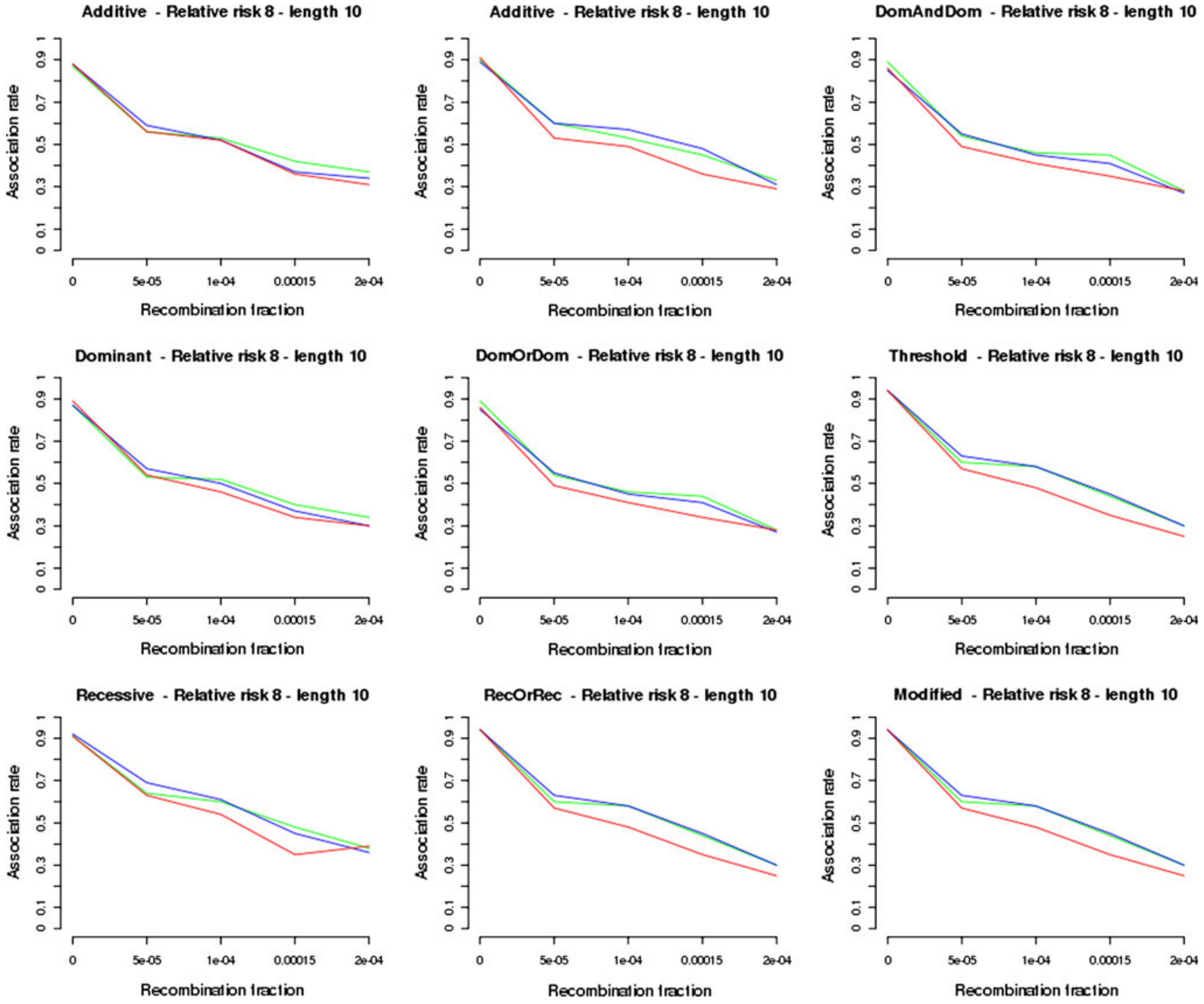

Fig. 3 Association rate based on 100 simulations of 200 family trios as a function of the recombination rate using haplotypes of length 10 and different genotype models (*rows*). The *first column* shows results for one disease susceptibility locus and the *second* and *third* show results for two disease loci. A nominal level of $\alpha = 0.05$ and a relative risk of 8 were used for all plots. Results for $mTDT_P$, $mTDT$ and $mTDT_Y$ are plotted in *red*, *blue* and *green*, respectively

Real data sets

As in the simulation study, besides *mTDT* and tests designed to cope with the problem of small data ($mTDT_Y$, $mTDT_{YP}$, $mTDT_{L1}$ and $mTDT_{L2}$), we used the same tests for state-of-the-art data sets for comparison with $mTDT_P$: $mTDT_{1T}$, $mTDT_E$, $mTDT_{LC}$ and $mTDT_{SR}$. We added a further test for the real data sets. $mTDT_{1U}$ is the same as $mTDT_{1T}$ but uses the most frequent non-transmitted instead of the most frequent transmitted haplotype. Our purpose was to consider whenever a disease is more common in the absence of a protective disease locus in affected individuals, a situation for which $mTDT_{1T}$ would be powerless.

A multimarker TDT for genome-wide association searches requires a very efficient exploration approach for the method to be feasible. A possible approach would consist of dividing the SNP sequence into blocks of low recombination using an algorithm based on confidence intervals (Gabriel et al. 2002). However, we chose to split regions in a block-free way because a low-recombination block has sensible differences depending on the definition used by the algorithm to split a region in blocks (Halldórsson et al. 2004). Thus, we used sliding windows (Daly et al. 2001) to apply the test to very small subsets of consecutive markers, such as 6, 8 or 10 markers. Each subset is a window and windows can share markers.

We used sliding windows of 1, 2, 4, 6, 8 and 10 SNPs per window and an offset of 1 to compute *p* values. Significance levels were computed for each sliding window using standard permutation tests (1,000 permutations) for

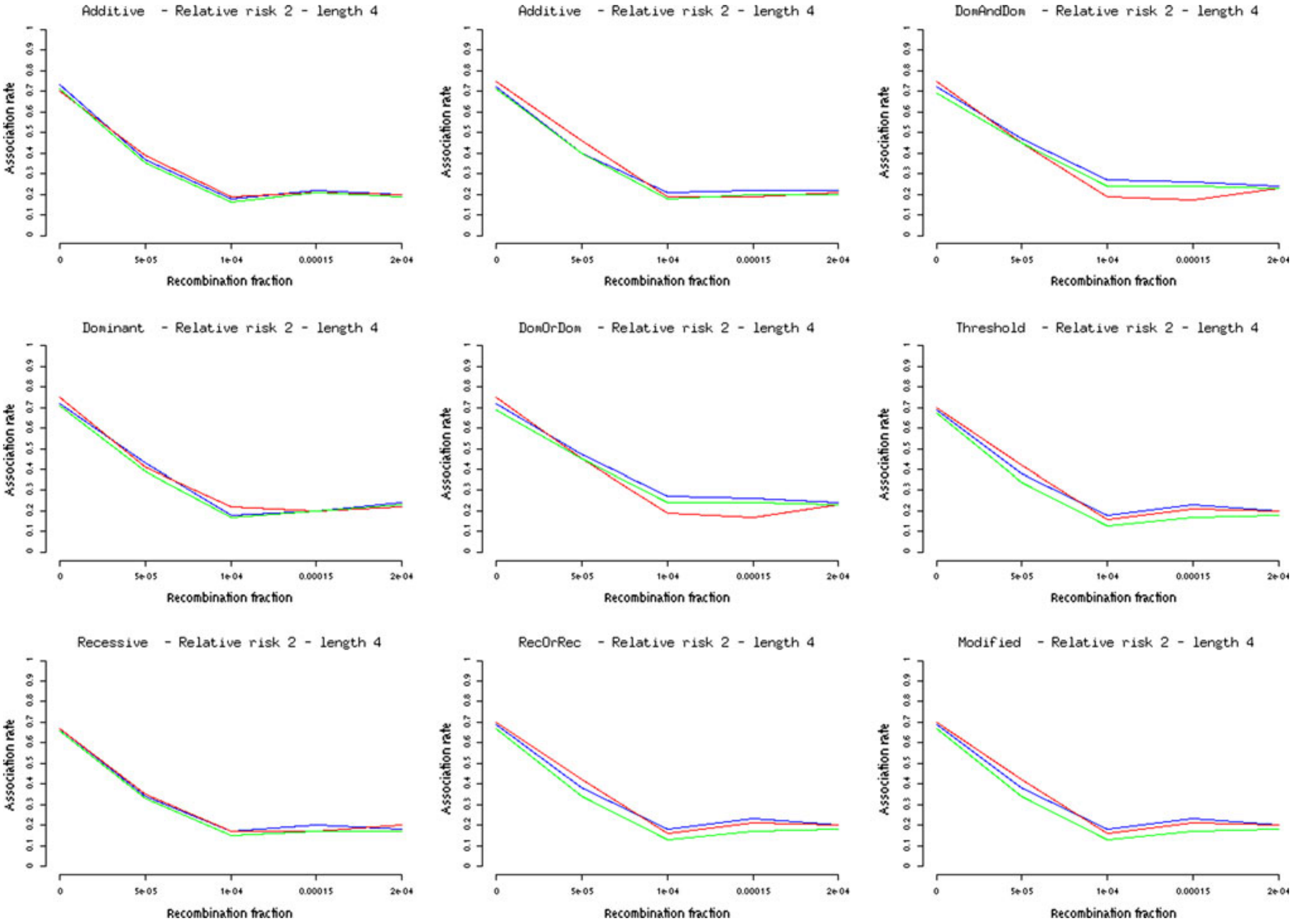

Fig. 4 Association rate based on 100 simulations of 500 family trios as a function of the recombination rate using haplotypes of length 4 and different genotype models (*rows*). The *first column* shows results for one disease susceptibility locus and the *second* and *third* show results for two disease loci. A nominal level of $\alpha = 0.05$ and a relative risk of 2 were used for all plots. Results for $mTDT_P$, $mTDT$ and $mTDT_Y$ are plotted in *red*, *blue* and *green*, respectively

when the null distribution is unknown. For all tests for which the null distribution or its approximation is known, we used that distribution to compute p values.

Phase reconstruction

We inferred haplotype frequencies using all the information from the family (Yu et al. 2005; Rinaldo et al. 2005). Those haplotypes that were unsolved using family information, were inferred using the E-M algorithm under the restriction of family information (Abecasis et al. 2001; Yu et al. 2005).

To avoid inaccurate haplotype reconstruction, E-M algorithm is usually applied within a low recombination block (Niu et al. 2002). However, despite we first performed a preliminary division of the chromosome in blocks of low recombination by using some of the several algorithms proposed to do that (Gabriel et al. 2002), we finally decided to use sliding windows because of the following two reasons.

On one hand, results from different block building algorithms are very distinct (Halldórsson et al. 2004) and they may bias results from TDT measures. Moreover, the chances of an haplotype of few SNPs to cover more than one block are being reduced with the increase in the number of sequenced SNPs. As an example, with a current genome-wide SNP array of about 500,000 SNP markers, and considering the estimation of 20,700 bp as the average block size in Caucasian populations (Hinds et al. 2005) it means about 20 SNPs per block. For windows of length 10, there are few chances for the haplotype to span through more than one block.

On the other hand, in trio samples the E-M algorithm is used under the restriction of family information (Zhang et al. 2003; Yu et al. 2005) and, therefore, it is more accurate than the simple E-M to infer the phase, even beyond block boundaries, as the only positions whose transmission/non-transmission alleles cannot be solved using family information are those for which the three

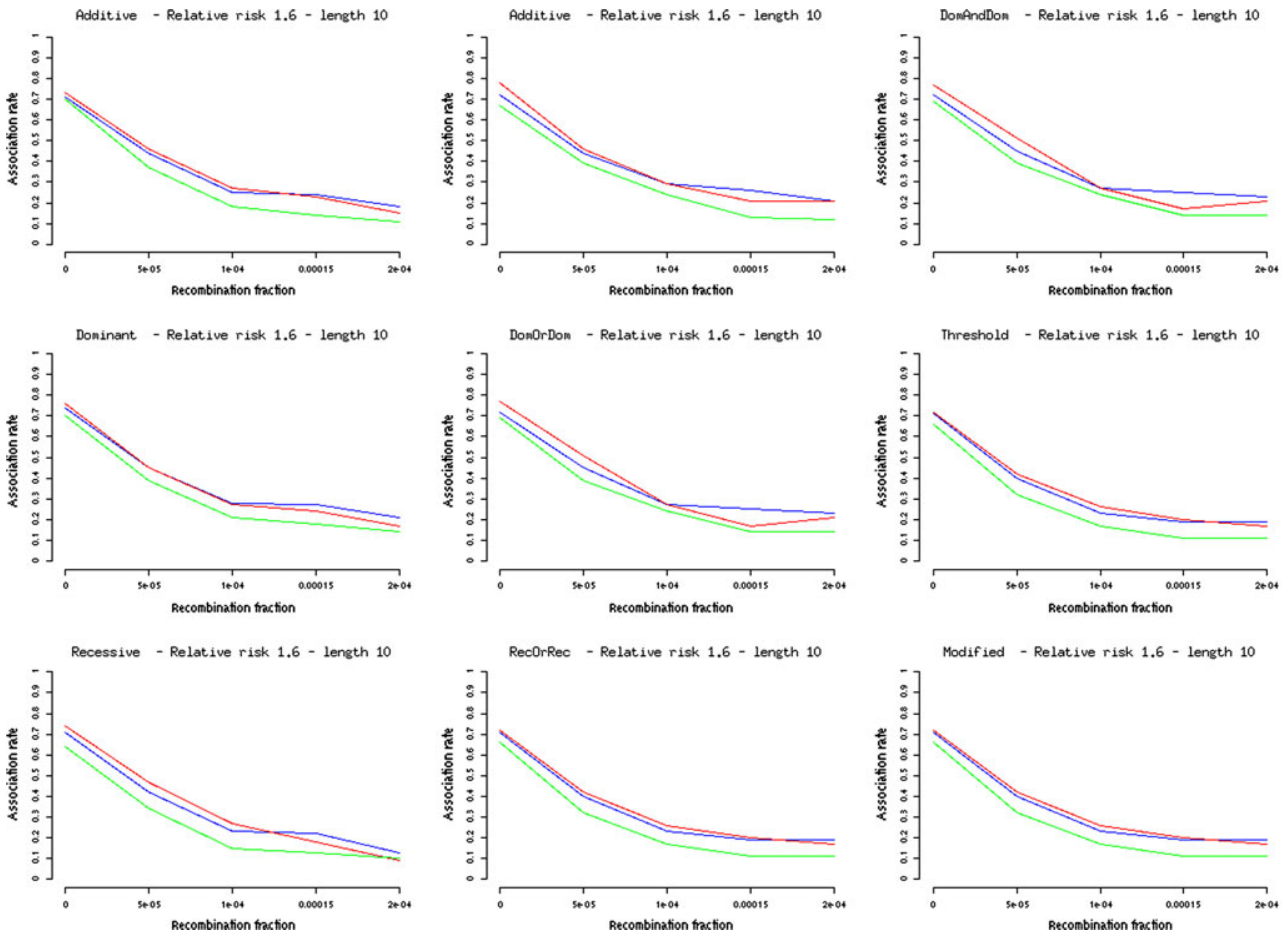

Fig. 5 Association rate based on 100 simulations of 500 family trios as a function of the recombination rate using haplotypes of length 10 and different genotype models (*rows*). The *first column* shows results for one disease susceptibility locus and the *second* and *third* show results for two disease loci. A nominal level of $\alpha = 0.05$ and a relative risk of 1.6 were used for all plots. Results for $mTDT_P$, $mTDT$ and $mTDT_Y$ are plotted in *red*, *blue* and *green*, respectively

family members are heterozygotic (Sebastiani et al. 2004). We compared (data not shown) results of two main ways to proceed within each family: (1) to choose the most likely phase according with the E-M algorithm under the restriction of family information or (2) to use weighted phases using as weights the frequencies reported by the algorithm (Zhang et al. 2003; Yu et al. 2005) and, in agreement with these works, found no significant differences among the two methods. Therefore, we opted for using the first one of the two choices, for being the one with lower computational complexity.

Data sets used

We used nine data sets of trio genotypes, one with individuals with Crohn’s disease (affected-Crohn) and the others with individuals with MS. The Crohn data set is a publicly available set originally used by Rioux et al. (2001).

Table 4 provides information about the MS data sets. Eight regions corresponding to risk loci for MS previously determined in well-powered studies were chosen. Genotype information for these regions was obtained from a genome-wide association study performed for the International Multiple Sclerosis Genetic Consortium. A DNA microarray (GeneChip Human Mapping 500K Array Set, Affymetrix) was used to examine 334,923 common genetic variants in 931 family trios consisting of a patient with MS and both parents (International Multiple Sclerosis Genetics Consortium et al. 2007).

For all the sets used, we prepared data sets for unaffected individuals from data publicly available at the website of the IHMP (HapMap-Consortium 2003) consisting of genotype data for 30 family trios (HapMap Phase II) typed in the CEPH population, who are Utah residents with ancestry from Northern and Western Europe. The tests for unaffected trios are used as a control, since an association found in unaffected individuals may point out to a

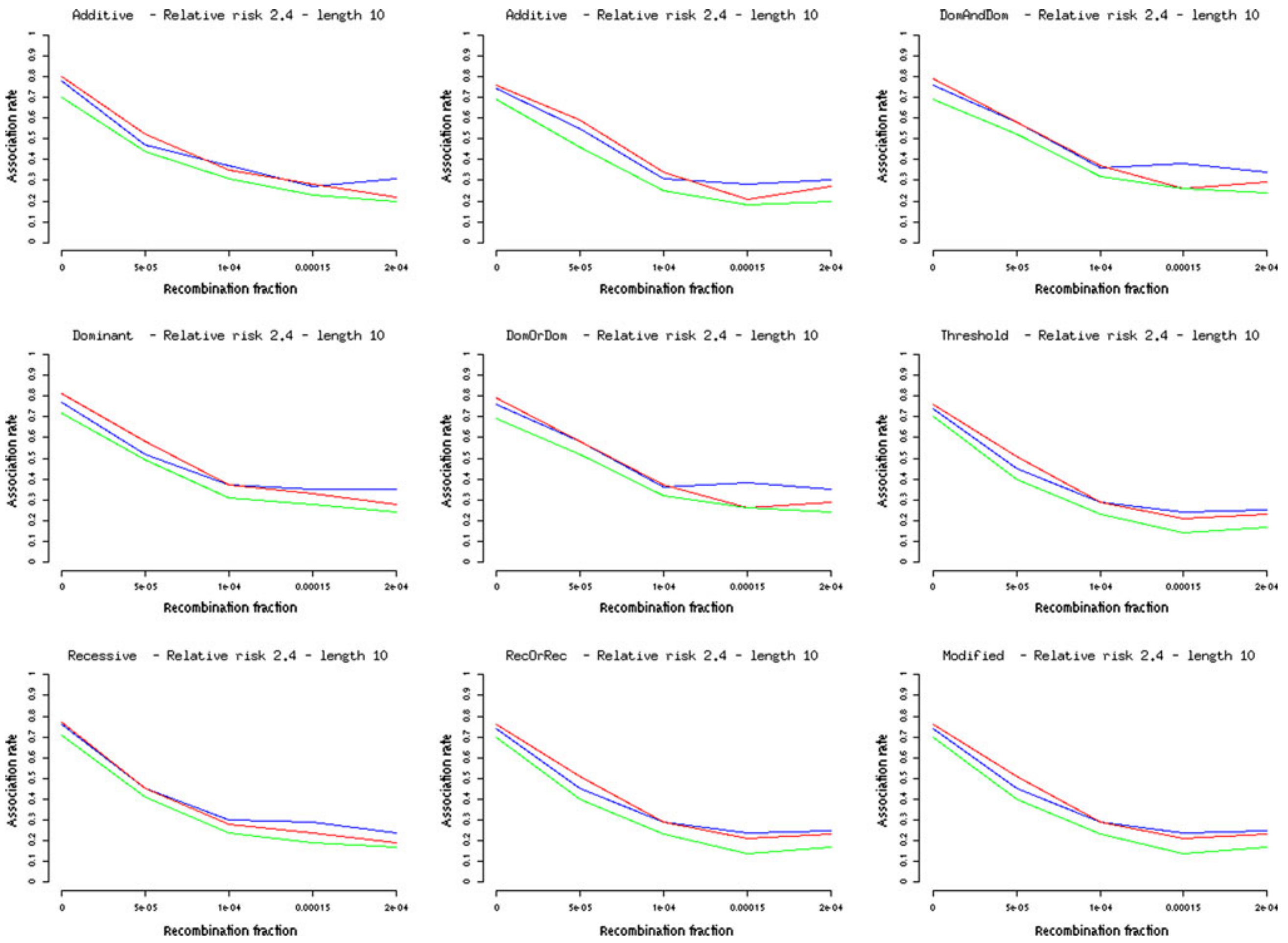

Fig. 6 Association rate based on 100 simulations of 500 family trios as a function of the recombination rate using haplotypes of length 10 and different genotype models (*rows*). The first column shows results for one disease susceptibility locus and the *second* and *third* show results for two disease loci. A nominal level of $\alpha = 0.05$ and a relative risk of 2.4 were used for all plots. Results for $mTDT_P$, *mTDT* and $mTDT_Y$ are plotted in *red*, *blue* and *green*, respectively

disease protective locus, genotypic errors or changes in Hardy–Weinberg equilibrium.

Crohn affected and unaffected data sets from the IHMP are all available on the supplementary website.

Results for real data sets

In general, $mTDT_P$ seems to be more locus-specific than the other tests used, with competitive power (see Fig. 8 for loci KIAA0350 and IRF5 for a window width of 10 and Figs. S38–S43 on the supplementary website for all loci and window widths 1, 2, 4, 6, 8 and 10).

To show these results we used *comparative TDT (CTDT) maps*, which are drawn by averaging the *p* values for each sliding window covering the same marker. A computer program to construct these maps was built using BioCASE (Montes and Abad-Grau 2009). Each row in a CTDT map represents sample results obtained from a different TDT. The height of the colored bar for each marker represents the range of the *p* value. If the *p* value is greater than 0.05, there is no color for that marker position, meaning that association is not significant. If the *p* value is less than 0.01, the colored bar has maximum height.

The association of the KIAA0350/ CLEC16A locus with MS was reported by the IMSGC genome-wide association study (International Multiple Sclerosis Genetics Consortium et al. 2007), however it did not reached genome-wide significance. Later on, it was replicated in several populations and now is considered a risk factor for MS (Martínez et al. 2010; M et al. 2009). Our results for the KIAA0350 locus (Fig. 8a) reveal that $mTDT_P$ detected a strong association (maximum height bar) from locus rs28087 to locus rs248836. Compared with *mTDT* and the alternative corrections for coping with the small data problem, $mTDT_P$ is more specific, as the range of markers with maximum association is smaller. The other tests were not able to detect association, with *p* values less than 0.01.

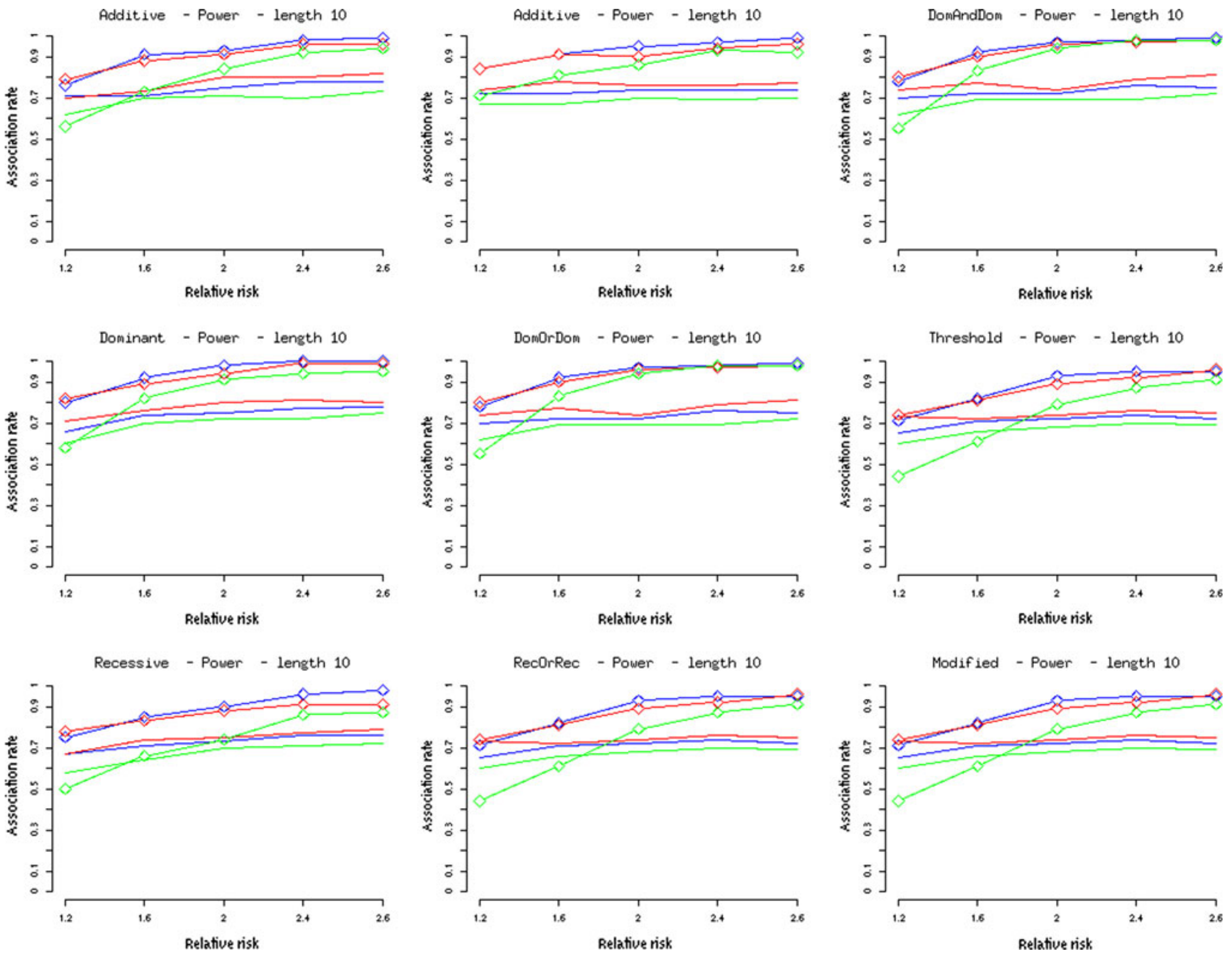

Fig. 7 Association rate based on 100 simulations of 500 family trios as a function of the recombination rate using haplotypes of length 10 and different genotype models (*rows*). The *first column* shows results for one disease susceptibility locus and the *second* and *third* show results for two disease loci. A nominal level of $\alpha = 0.05$ and a haplotype length of 10 were used for all plots. Results for $mTDT_P$, *mTDT* and $mTDT_Y$ are plotted in *red*, *blue* and *green*, respectively. *Simple lines* show values for frequency of the disease mutation in the interval [0.2, 0.4] while *lines* with *diamonds* show results for mutation frequencies in the interval [0.1, 0.2]

Table 4 Real data sets

Data set	ch.	First SNP	Last SNP	SNPs
EVI5	1	92388330	93651891	93
IL2R	10	6103680	7715013	353
IL7R	5	35847586	35991293	35
HLA	6	30736061	33163225	468
KIAA0350	16	11050221	11226546	26
CD226	18	65550188	65997985	38
CD58	1	116677600	116983610	19
IRF5	7	128055671	128309250	15

The first and last SNPs columns show the physical SNP position

Interferon regulatory factor 5 (IRF5) has been found to be associated with MS in a cadidate gene study in several population (Kristjansdottir et al. 2008). Results for IRF5 (Fig. 8b) show an interesting pattern in $mTDT_P$ and $mTDT_{1T}$: there is a locus with maximum association (rs3807306), which may mean that the actual disease susceptibility locus is somewhere between this marker and its left and right neighbors, and a continuous decrease with distance from the marker at maximum association either to the left or to the right along the chromosome. This pattern only applies to the right side of the locus, with maximum association for other *mTDT* measures. Thus, $mTDT_P$ again yields the maximum information: the power is maximum for a shorter region and significantly decreases with distance from this region.

However, results obtained by $mTDT_P$ do not always show a narrower region of association. Sometimes the region is as wide or even wider than that detected by *mTDT*. This is the case for the human leukocyte antigen (HLA) locus (see the

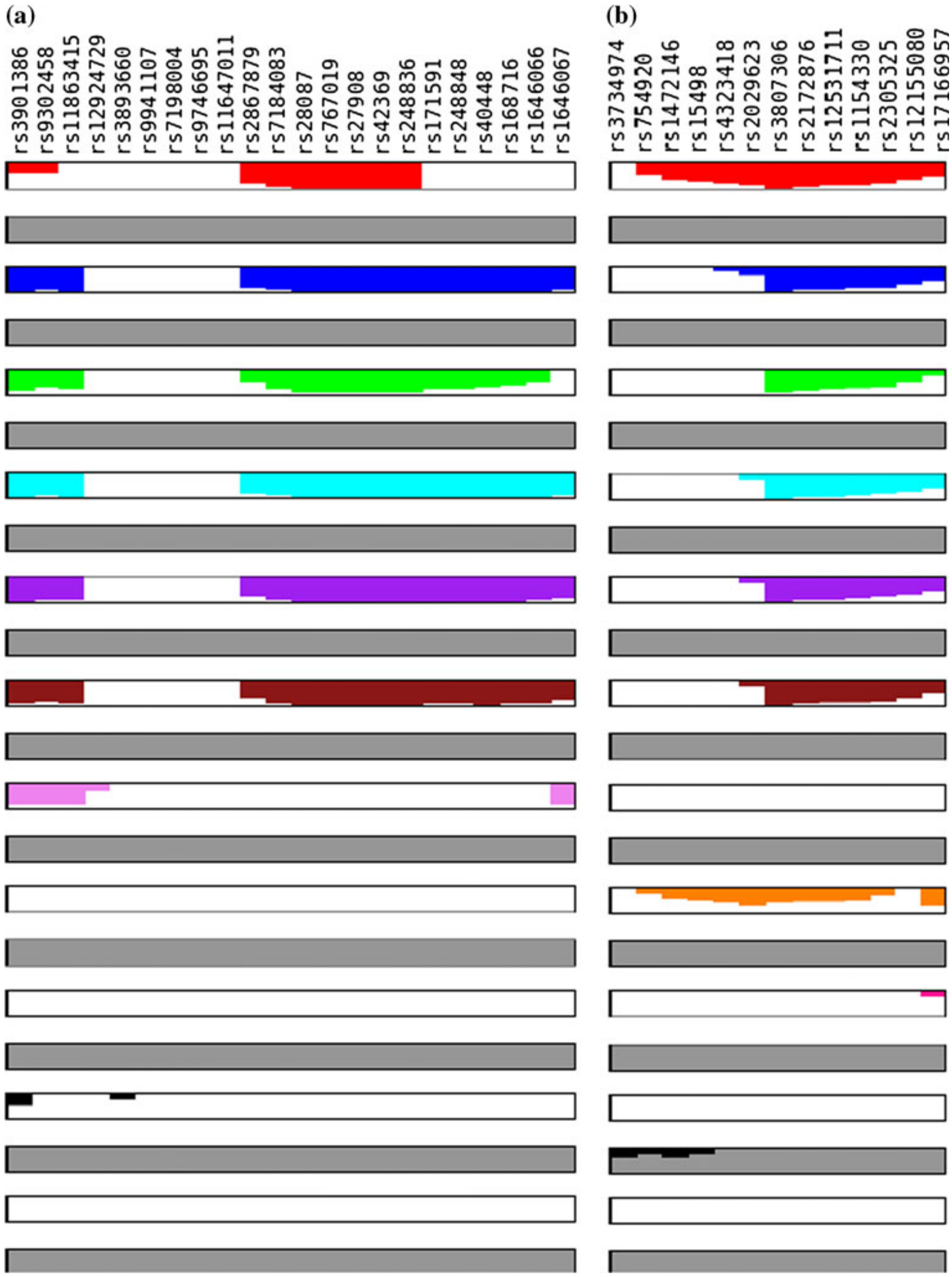

Fig. 8 Comparative TDT maps for loci, **a** KIAA0350 and **b** IRF5 data sets using sliding windows of width 6 and offset 1. *Rows* in *gray* below each TDT map (colored on a white background) show results for the IHMP data sets as a control test. Results correspond to the following TDTs from top to bottom: $mTDT_P$ (*red*), *mTDT* (*blue*), $mTDT_Y$ (*green*), $mTDT_{YP}$ (*cyan*), $mTDT_{L1}$ (*purple*), $mTDT_{L2}$ (*scarlet*), $mTDT_E$ (*violet*), $mTDT_{1T}$ (*orange*), $mTDT_{1U}$ (*pink*), $mTDT_{LC}$ (*black*) and $mTDT_{SR}$ (*brown*)

fifth CTDT map in Figs. S38–S43 on the supplementary website). This would mean that there is no single gene associated with the disease at that locus and other associations were detected as a result of linkage, but many of them along the HLA locus can influence disease onset. This is consistent with other studies suggesting that the HLA class II genes (HLA-DRB1) are the major determinants of MS risk in the major histocompatibility complex (MHC) region. Despite the recognized effect of HLA class II genes on risk, it is not clear what contributions other genes in this region may make. The MHC region has extensive LD spanning several megabases (Mb) and high levels of variability, with the HLA genes having hundreds of alleles. The MS data set analyzed here has not been genotyped at a sufficiently marker density across the entire MHC region to model the class II effects appropriately to be confident that the associations are not attributable to either the class II loci themselves or other (untyped) loci within the region.

Discussion

With current SNP genotype samples for family trios of a few hundred or thousand trios, the locus specificity of a test has become as important as its power, as it is very common to find associations due to linkage in loci at a considerable distance from the disease susceptibility locus. These associations usually cannot be replicated in other samples from close populations, as they are at some distance from the disease susceptibility locus and their haplotypes may have departed from the common ancestors in the first sample used due to recombination. A lack in locus specificity means they may detect association at considerably large chromosomal distance to the disease susceptibility locus. These associations can be considered spurious associations, as they do not point out to a susceptibility locus or positions very close to it and they will be hardly replicated in a lightly different sample. Thus, more than two markers may be used so that power will increase with a lower risk of low specificity. Therefore, it is very important to consider the locus specificity of TDTs to increase the chances of finding truly risky haplotypes, i.e., those actually at the disease susceptibility locus or at a very short distance from it, and thus the chances of replicating the results in other samples. With this goal, we proposed $mTDT_P$, which is based on *mTDT*, one of the first multimarker TDTs. *mTDT*, together with $mTDT_{1T}$ and $mTDT_P$ has the highest power under a wide range of scenarios in light of our simulations. Because $mTDT_P$ is based on *mTDT*, the new assumption used to define $mTDT_P$ is crucial to improve locus specificity without risking the high power of *mTDT*. Therefore, the new assumption and thus the modification introduced by the test had to be as simple as possible for the test to be as generic as *mTDT* and to focus on reducing association rates with chromosomal distance to the disease susceptibility locus at a faster rate. To achieve this, the new assumption was very specific: association decreases with chromosomal distance from a specific locus because of recombinations. As a consequence, haplotypes in phase with a disease variant at the time at which a variant appeared would recombine more often with other haplotypes with increasing distance from the disease locus. Thus, in a sample of trios with affected offspring, the frequency of these non-recombinant haplotypes will be lower than if the haplotype were closer to the disease locus. Therefore, by weighting each summand in *mTDT* by the haplotype frequency, we reduce the effect that haplotypes at some chromosomal distance to a disease locus can have on the measure because of linkage. Moreover, in positions close to the disease locus, and assuming the CDCV hypothesis, there would be very few, but common, haplotypes with strong association with the risk variant, so that the weighting procedure will not reduce the power.

We performed simulations under a wide range of population and disease variables, such as the number of disease loci, the disease model, the relative risk of a genotype, haplotype length, etc. Simulations confirmed the correctness of the assumptions and the improvement in locus specificity achieved by $mTDT_P$ without reducing the power. We also used several real trio data sets with affected offspring.

As these TDTs are to be applied to genome-wide data sets. a multiple testing correction should be performed. Multiple testing correction for GWAS is currently a very active research topic (Betensky and Rabinowitz 2000; Wei et al. 2009; Gorlov et al. 2009), as most of the current approaches do not consider LD between different markers and they usually over-correct association results and therefore true-effect associations may be missed. As the objective in the simulations performed was to compare power and locus specificity from different tests, we did not perform multiple testing corrections in any of the tests and *p* values were directly compared. Moreover, $mTDT_{1T}$ and $mTDT_{1U}$, which choose the haplotype with the lowest *p* value, were not competitive when the Bonferroni correction was applied. Current real genome-wide data usually have hundred thousand markers. We considered using sliding windows and comparative TDT maps as visual tools for genome-wide screening, including also the use of IHMP samples as controls. In these two visual tools, instead of a unique *p*-value for each window with multiple testing correction, average *p*-values for all the windows a marker belongs to are drawn in order to reduce the chances of spurious associations. Therefore, we chose a simple approach to detect association decay with distance in order to select a region to perform a further fine-mapping study including a more dense screening over the selected region and sample replication for which multiple testing correction may be required.

The results obtained using $mTDT_P$ analysis for the MS data set showed more precise definition of MS implicated variants among the loci analyzed. KIAA0350/CLEC16A has been associated with several autoimmune diseases in genome-wide association and replication studies (International Multiple Sclerosis Genetics Consortium et al. 2007; Todd et al. 2007; Márquez et al. 2009). Fine mapping of the region for type 1 diabetes (T1D) by resequencing of exons and flanking regions and SNP genotyping for the surrounding genes revealed that the most probable causal variant would be localized at the 3′ end of the KIAA0350/CLEC16A gene. Results for the $mTDT_P$ CTDT map of the KIAA0350/CLEC16A locus using MS data reveal that the region with greatest association is the last 3′ 60 Kbp of the gene, whereas the other TDTs extend the association to the intergenic 3′ region. These $mTDT_P$ results pointed to the 3′ end of the KIAA350 gene as the causative association region in MS as described for T1D. We also observed for some other loci that the *mTDTP* map extends to a larger region than the other TDT maps. This is the case for the IRF5 locus. The most probable causal variant for association of the IRF5 locus with MS is a functional 5-bp biallelic insertion–deletion polymorphism that differentially binds the SP1 transcription factor to the IRF5 promoter (Kristjansdottir et al. 2008). The $mTDT_P$ map revealed maximal association at IRF5 and extended it to the 5′ region, including the IRF5 promoter, whereas the other maps did not reveal any association with the IRF5 promoter. In designing a fine mapping of the IRF5 locus based on $mTDT$, $mTDT_Y$, $mTDT_{YP}$, $mTDT_{L1}$ or $mTDT_{L2}$ results, we would be erroneously focusing on the middle of the gene instead of on the promoter, where the most probable causative variants are located.

An interesting question arises about whether $mTDT_P$ would be still useful when disease-susceptibility variants have very low frequencies, i.e., under the ‘common disease, many rare variants’ (CDMRV) hypothesis. In general, GWAS are not suitable to capture rare variants and other techniques, such as DNA resequencing of candidate genes are often used (Bodner and Bonilla 2008). However, it is being recently claimed that many of the associations found by GWAS are due to ‘synthetic associations’ between very rare variants and less rare alleles, such as SNP markers (Dickson et al. 2010) on the basis that what is usually tested are not the causative genes but SNP markers around them. Under this hypothesis, we believe $mTDT_P$ may have less power than $mTDT$ if we consider results from our simulations (Fig. 7 and supplementary Figures S55–S59): using usual mutation frequencies in common diseases (interval [0.2, 0.4]) $mTDT_P$ outperforms $mTDT$ in power; if we reduce mutation frequencies to be in the interval [0.1, 0.2], still high to be considered a rare variant, differences in power between the two test converge and even $mTDT$ outperforms $mTDT_P$ under several scenarios.

Our ultimate goal is to have a multimarker test that: (1) requires little computational time, as $mTDT$ or $mTDT_{HE}$; (2) provides high power under very different circumstances, as $mTDT$ or $mTDT_{1T}$; (3) performs stronger filtering than state-of-the-art TDTs so that it can detect association in narrower regions when used as a first genome-wide step in searching for disease susceptibility or protective genes. $mTDT_P$ achieves these three goals better than all the other tests we used. Moreover, by producing highly informative *Comparative TDT (CTDT)* maps using different low-complexity TDT measures with very different specificity and sensitivity behaviors and using IHMP samples as both control and test validators, we provide a robust tool for visual exploration that may assist molecular biologists in decisions about the regions to choose for fine mapping.

In conclusion, we believe $mTDT_P$ can benefit genome-wide association studies as its higher locus specificity may be crucial to improve chances of detecting only associations close to a disease susceptibility or protective locus and therefore its chances of being replicated in different samples.

Web source

A supplementary website has been created for this study at http://bios.ugr.es/TDTP, where Figures S1–S43, Table S1, a detailed explanation of the simulations performed and the source code in c++ of the software developed for this work are available.

Acknowledgments The authors were supported by the Spanish Research Program under project TIN2007-67418-C03-03, the Andalusian Research Program under project P08-TIC-03717 and the European Regional Development Fund (ERDF). We acknowledge the International Multiple Sclerosis Genetics Consortium (IMSGC) for giving us access to their data repository.

Appendix 1: Variance of $mTDT_P$

The variance of $mTDT_P$ can be obtained using a slight modification of the procedure used by Sham (1997) for the variance of $mTDT$.

Let N_{ij} be the count for heterozygotic parents with haplotypes i, j transmitting haplotype i to their child, and N_{ji} the count for parents with the same genotype but transmitting haplotype j to their child. Let n_{ii} be the count for homozygotic parents for haplotype i. Let n_{ij} be $N_{ij} + N_{ji}$. Consider N_{ij} as a realization of the random variable X_{ij}, $i = 1,\ldots, k$, $j = i + 1,\ldots, k$, N_{ji} as a realization of the random variable X_{ji} and N_{ii} as a realization of the random variable X_{ii}. Thus, $X_{ji} = N_{ij} - X_{ij}$ holds. The counts n_{iT} and n_{iU} in $mTDT_P$ are then realizations of the random variables

$$X_{iT} = \sum_{j=1}^{k} X_{ij} - X_{ii}$$

and

$$X_{iU} = \sum_{j=1}^{k} X_{ji} - X_{ii}.$$

Moreover, $n_i = n_{iT} + n_{iU}$ and n is the total count haplotype count for heterozygotic parents: $n = \sum_{i=1}^{k} n_i$.

The variance of $mTDT_P$ is therefore:

$$V(mTDT_P) = Var\left[\sum_{i=1} k\frac{X_{iT} + X_{iU}}{n} Y_i\right],$$

with Y_i defined as:

$$Y_i = \frac{(X_{iT} - X_{iU})^2}{X_{iT} + X_{iU}} = \frac{\left[\sum_{j=i} 2X_{ij} - n_i\right]^2}{n_i}.$$

As shown by Sham (1997), under the null hypothesis of no linkage, Y_i is χ_1^2 and

$$Cov(Y_i, Y_j) = Cov\frac{n_{ij}^2}{n_i n_j} Var\left[\frac{(2X_{ij} - _{ij})^2}{n_{ij}}\right] = \frac{2n_{ij}^2}{n_i n_j}.$$

Therefore, the variance of $mTDT_P$ is:

$$\begin{aligned} V(mTDT_P) &= \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 \left[Var(Y_i) + \sum_{j \neq i} Cov(Y_i, Y_j)\right] \\ &= 2\sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 + 2\sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 \sum_{j \neq i} \frac{n_{ij}^2}{n_i n_j}. \end{aligned}$$

Appendix 2: Corrections to the small data problem

There is a well-known condition that must hold for a χ^2 test to be appropriately used as a test of independence: the expected value for each level of the variable cannot be very low. For *mTDT* this means that no haplotype count can be less than 10. In haplotype populations this is an important issue for haplotypes of a few SNP in length, as there are usually many rare haplotypes in a sample. The problem remains when a permutation test is used instead of the χ^2, as the definition of the measure does not change. Thus, an upward bias for association cannot be avoided owing to the high variances for low-frequency haplotypes. To the best of our knowledge, the consequences of using multimarker *mTDT* for a small number of data, which is a very common problem, have not been studied. The most widely used solution is to disregard haplotypes with a total count of less than 10 (Sham and Curtis 1995).

In the present study we considered two different approaches to the problem of small numbers of data for *mTDT* instead of disregarding low-count haplotypes. The first is based on the Yates (1934) correction and the second on the Laplace correction. It should be noted that all these corrections improve locus specificity at a cost of power. Therefore, when used for loci very close to the disease susceptibility locus (recombination rates close to 0) differences between transmitted and non-transmitted haplotypes, which are mainly due to true effects, will also be reduced.

One method for solving the problem of small numbers of data in χ^2 distributions is the Yates (1934) correction, which is straightforward to apply to *mTDT*, i.e., small numbers of data for low-frequency haplotypes, so that the new test $mTDT_Y$ is defined as:

$$mTDT_Y = \frac{k-1}{k} \sum_{i=1}^{k} \frac{[|n_{iT} - n_{iU}| - y]^2}{n_{iT} + n_{iU}},$$

with $y = 0.5$.

The aim of subtracting 0.5 is to reduce the random effect of very low-frequency haplotypes. However, when analyzing positions close to a disease susceptibility locus, most differences between transmitted and non-transmitted haplotype counts will be due to a true effect and the correction will lead to a power reduction. A straightforward generalization of $mTDT_Y$ is that in which y can be any value. Changing y by values greater than 0.5 will reduce the effect of random errors to a greater degree in the case of very low-frequency haplotypes. However, the power will also decrease to a greater extent. We denote the statistic for which $y = 1$ as $mTDT_{Y1}$.

Instead of a constant reduction in the module, a reduction proportional to the haplotype frequency seems to be a better choice to yield a higher correction for less frequent haplotypes. Based on this idea, we define $mTDT_{YP}$ as:

$$mTDT_{YP} = \frac{k-1}{k} \sum_{i=1}^{k} \frac{[|n_{iT} - n_{iU}| - 1/(2 * n_i)]^2}{n_{iT} + n_{iU}}$$

The correction is the same as TDT_Y for haplotypes with a frequency of 1 and is lower for more frequent haplotypes, with very little effect for high haplotype frequencies. The correction may outperform TDT_{Y1} in terms of locus specificity because it yields greater correction for haplotypes with lower frequencies. However, the correction may lead to a higher power reduction because, even for low-frequency haplotypes, differences between transmitted and non-transmitted haplotype counts when markers are very close to the disease susceptibility locus may be due to true effects.

Another way to proceed instead of reducing the numerator of each summand is to increase the denominator of each summand using the Laplace correction, which adds a constant value f to the count of each haplotype, yielding:

$$\begin{aligned} mTDT_L &= \frac{k-1}{k} \sum_{i=1}^{k} \frac{(n_{iT} + f - n_{iU} - f)^2}{n_{iT} + n_{iU} + 2f} \\ &= \frac{k-1}{k} \sum_{i=1}^{k} \frac{(n_{iT} - n_{iU})^2}{n_{iT} + n_{iU} + 2f}. \end{aligned}$$

Here we used $f = 1$ and $f = 2$ ($mTDT_{L1}$ and $mTDT_{L2}$, respectively).

References

Abecasis GR, Martin R, Lewitzky S (2001) Estimation of haplotype frequencies from diploid data. Am J Hum Genet 69:198

Abramowitz M, Stegun I (1972) Handbook of mathematical functions. Dover, New York

Betensky R, Rabinowitz D (2000) Simple approximations for the maximal transmission/disequilibrium test with a multi-allelic marker. Ann Hum Genet 64:567–574

Bodner W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701

Bourgain C, Genin E, Holopainen P, Mustalahti K, Mki M, Partanen J (2001) Maximum identity length contrast: a powerful method for susceptibility gene detection in isolated populations. Am J Hum Genet 68:154–159

Castao-Martínez A, López-Blázquez F (2005) Distribution of a sum of weighted central chi-square variables. Commun Stat Theory Methods 34:515–524

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

Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700–706

Daly M, Rioux J, Schaffner S, Hudson T, Lander E (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Dickson S, Wang K, Krantz I, Hakonarson H, Goldstein D (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:1000, 294

Fan RZ, Xiong MM (2001) Linkage transmission disequilibrium test of two unlinked disease loci. Adv Appl Stat 1:277–308

D'Netto MJ, Ward H, Morrison K, DeLuca S, Handunnetthi L, Sadovnick A, Ebers G (2009) Risk alleles for multiple sclerosis in multiplex families. Neurology 72:1984–1988

Gabler S, Wolff C (1987) A quick and easy approximation to the distribution of a sum of weighted chi-square variables. Stat Hefte Stat Pap 28:317–325

Gabriel S, Schaffner S, Nguyen H, Moore J, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander E, Daly M, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296

Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 69:371–380

Gorlov IP, Gallick GE, Gorlova OY, Amos C, Logothetis CJ (2009) Gwas meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? prostate cancer as an example. PLoS ONE 4(8):e6511

Halldórsson B, Bafna V, Lippert R, Schwartz R, de La Vega F, Clark A, Istrail S (2004) Optimal haplotype block-free selection of tagging snps for genome-wide association studies. Genome Res 14:1633–1640

HapMap-Consortium TI (2003) The international hapmap project. Nat Biotechnol 426:789–796

Hellenthal G, Stephens M (2007) mshot: modifying hudson's ms simulator to incorpore crossover and gene conversion hot spots. Bioinformatics 23:520–521

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common dna variation in three human populations. Science 18:1072–1079

International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, Sawcer S, Lander ES, Daly M, Jager PD, de Bakker P, Gabriel S, Mirel D, Ivinsonand A, Pericak-Vance M, Gregory S, Rioux J, McCauley J, Haines J, Barcellos L, Cree B, Oksenberg J, Hauser S (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357(9):851–62

Johnson N, Kotz S, Balakrishnan N (1994) Continuous univariate distributions. Wiley, New York

Kristjansdottir G, Sandling J, Bonetti A, IM IR, L LM, C CW, Gustafsdottir S, Sigurdssonand S, Lundmark A, K PTKK, Elovaara I, Pirttil T, Reunanen M, L LP, Saarela J, Hillert J, Olsson T, Landegren U, Alcina A, Fernández O, Leyva L, Guerrero M, Lucas M, Izquierdo G, Matesanz F, Syvnen A (2008) Interferon regulatory factor 5 (irf5) gene variants are associated with multiple sclerosis in three distinct populations. J Med Genet 45:362–369

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–142

Lam J, Roader K, Devlin B (2000) Haplotype fine mapping by evolutionary trees. Am J Hum Genet 66:659–673

Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. Human Heredity 48:67–81

Li J, Wannng D, Dong J, Jiang R, Zhang K, Zhang S, Zhao H, Sun F (2001) The power of transmission disequilibrium tests for quantitative traits. Genet Epidemiol 18 (Supp 1):632–637

Márquez A, Varadé J, Robledo G, Martínez A, Mendoza J, Taxonera C, Fernández-Arquero M, Díaz-Rubio M, Gómez-García M, Lpez-Nevot M, de la Concha E, Martín J, Urcelay E (2009) Specific association of a clec16a/kiaa0350 polymorphism with nod2/card15(-) crohn's disease patients. Eur J Hum Genet 17(10):1304–1308

Martínez A, Perdigones N, Espino MCL, Varadé J, Lamas J, J JS, Fernández-Arquero M, de la Calle H, Arroyo R, de la Concha E, B BFG, Urcelay E (2010) Chromosomal region 16p13: further evidence of increased predisposition to immune diseases. Ann Rheum Dis 69:309–11

Montes R, Abad-Grau MM (2009) Biocase: Accelerating software development of genome-wide filtering applications. In: IWANN '09: Proceedings of the 10th international work-conference on artificial neural networks. Springer, Berlin, pp 1097–1100

Niu T, Qin Z, XU X, Liu J (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169

Nordborg M (2001) Coalescent theory. Wiley, Chichester, pp 179–212

Ott J (1999) Analysis of human genetic linkage. John Hopkins, Baltimore

Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K (2005) Characterization of multilocus linkage disequilibrium. Genet Epidemiol 28:193–206

Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, Kulbokas EJ, O'Leary S, Winchester E, Dewar K, Green T, Stone V, Chow C, Cohen A, Langelier D, Lapointe G, Gaudet D, Faith J, Branco N, Bull SB, McLeod RS, Griffiths AM, Bitton A, Greenberg GR, Lander ES, Siminovitch KA, Hudson TJ (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to crohn disease. Nat Genet 29:223–228

Sebastiani P, Grau MMA, Alpargu G, Ramoni MF (2004) Robust transmission disequilibrium test for incomplete family genotypes. Genetics 168:2329–2337

Seltman H, Roeder K, Devlin B (2001) Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 68:223–235

Sham PC (1997) Transmission/disequilibrium tests for multiallelic loci. Am J Hum Genet 61:774–778

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multiallelic marker loci. Ann Hum Genet 59:323–336

Solomon H, Stephens MA (1977) Distribution of a sum of weighted chi-squared variables. J Am Stat Assoc 72:881–885

Spielman RS, Ewens WJ (1996) The tdt and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

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

Stuart A (1955) A test for homogeneity of the marginal distributions in a two-way classification. Biometrika Trust 42:412–416

Tang R, Feng T, Sha Q, Z S (2009) A variable-sized sliding-window approach for genetic association studies via principal component analysis. Ann Hum Genet 73:631–637

Todd J, Walker N, Cooper J, Smyth D, Downes K, Plagnol V, Bailey R, Nejentsev S, Field S, Payne F, Lowe C, Szeszko J, Hafler J, Zeitels L, Yang J, Vella A, Nutland S, Stevens H, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink L, Healy B, Burren O, Lam A, Ovington N, Allen J, Adlem E, Leung H, Wallace C, Howson J, Guja C, Ionescu-Trgovite C (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39:857–864

Wei Z, Sun W, Wang K, Hakonarson H (2009) Multiple testing in genome-wide association studies via hidden Markov models. Bioinformatics 25(21):2802–2808, http://www.bioinformatics.oxfordjournals.org/cgi/content/abstract/25/21/2802, http://www.bioinformatics.oxfordjournals.org/cgi/reprint/25/21/2802.pdf

Yates F (1934) Contingency table involving small numbers and the χ^2 test. J R Stat Soc 1:217–235

Yu K, Gu CC, Xiong C, An P, Province M (2005) Global transmission/disequilibrium tests based on haplotype sharing in multiple candidate genes. Genet Epidemiol 29:223–235

Zhang S, Sha Q, Chen H, Dong J, Jiang R (2003) Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. Am J Hum Genet 73:566–579

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK (2000) Transmission/disequilibrium tests using multiple tighly linked markers. Am J Hum Genet 67:936–946

Zhao J, Boerwinkle1 E, Xiong M (2007) An entropy-based genome-wide transmission/disequilibrium test. Hum Genet 121:357–367