Journal of Bioinformatics and Computational Biology
Vol. 11, No. 2 (2013) 1250014 (22 pages)

DOI: 10.1142/S021972001250014X

Imperial College Press
www.icpress.co.uk

INCREASING POWER BY USING HAPLOTYPE SIMILARITY IN A MULTIMARKER TRANSMISSION/DISEQUILIBRIUM TEST

MARÍA M. ABAD-GRAU[*,§], NURIA MEDINA-MEDINA[*,¶], SERAFÍN MORAL[†,‖], ROSANA MONTES-SOLDADO[*,**], SERGIO TORRES-SÁNCHEZ[*,††] and FUENCISLA MATESANZ[‡,‡‡]

[*]*Department of Computer Languages and Systems − CITIC*
Universidad de Granada, Granada, 18071, Spain

[†]*Department of Computer Science and Artificial Intelligence − CITIC*
Universidad de Granada, Granada, 18071, Spain

[‡]*Instituto de Parasitología Lpoez Neyra*
Consejo Superior de Investigaciones Científicas
Granada, 18071, Spain
[§]*mabad@ugr.es*
[¶]*nmedina@ugr.es*
[‖]*smc@decsai.ugr.es*
[**]*rosana@ugr.es*
[††]*sergiot@ugr.es*
[‡‡]*lindo@ipb.csic.es*

Received 17 April 2012
Revised 17 May 2012
Accepted 21 May 2012
Published 10 July 2012

It is already known that power in multimarker transmission/disequilibrium tests may improve with the number of markers as some associations may require several markers to be captured. However, a mechanism such as haplotype grouping must be used to avoid incremental complexity with the number of markers. 2G, a state-of-the-art transmission/disequilibrium test, implements this mechanism to its maximum extent by grouping haplotypes into only two groups, high and low-risk haplotypes, so that the test has only one degree of freedom regardless of the number of markers. The test checks whether those haplotypes more often transmitted from parents to offspring are truly high-risk haplotypes. In this paper we use haplotype similarity as prior knowledge to classify haplotypes as high or low risk ones and start with those haplotypes in which the prior will have lower impact i.e. those with the largest differences between transmission and non-transmission counts. If their counts are very different, the prior knowledge has little effect and haplotypes are classified as low or high risk as 2G does. We show a substantial gain in power achieved by this approach, in both simulation and real data sets.

Keywords: Transmission/disequilibrium association methods; haplotype tree; group-based multimarker transmission/disequilibrium tests.

[§]Corresponding author. Departamento de Lenguajes y Sistemas Informáticos — c/ Periodista Daniel Saucedo Aranda s/n, Universidad de Granada 18071, Granada, Spain.

1. Introduction

Genome-wide association studies (GWASs) using polymorphic nucleotide markers have as their main goal the identification of new genetic factors conferring individual susceptibility to complex diseases. Case/control studies may yield inflated false positives due to population stratification and admixture. Moreover, haplotype reconstruction can be inaccurately inferred because there are not familial genotypes. Therefore, tests using more than one marker usually are based on genotypes[1–3] instead of haplotypes, and power may also be reduced as well. To overcome this issue, we use an alternative group of tests, which require nuclear families and compare differences in transmitted and non-transmitted haplotype counts from parents to affected offspring. The simplest test of this group, the Transmission/Disequilibrium Test (TDT),[4] is defined for a binary marker and considers that under the null hypothesis of no association or linkage, the only difference in the number of times an allele is transmitted or not should be due to random sampling. Therefore, the statistic is defined for heterozygous parents as:

$$TDT = \frac{(n_T - n_U)^2}{n_T + n_U}, \tag{1}$$

with n_T being the number of times one of the alleles is transmitted by a parent to the offspring and n_U the number of times it is not transmitted (i.e. the parent transmits the other allele). The statistic follows a χ^2 distribution with one degree of freedom (df) under the null hypothesis of no association or linkage.

TDT-based GWASs using only one marker at a time have revealed to be a very powerful approach to discover new genetic factors related to a trait when they are in allelic association — linkage disequilibrium — with an underlying risk allele (as long as it has a high relative risk to the trait). However, most complex diseases are polygenic traits which appear as a consequence of the interaction of several, maybe thousand, genetic variants, most of them with modest or very small effects, as well as the environment. This is the case of multiple sclerosis (MS)[5] or diabetes type 2.[6] In this situation, the basic TDT [Eq. (1)] has not enough power to detect those genetic factors with small relative risk, as it is a monomarker test. Different generalizations have been proposed to handle more than two markers and improve power.[7–11] Perhaps the most intuitive multimarker generalization of TDT is $mTDT$[7,8]:

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

with k being the number of different alleles/haplotypes and n_{iT}, n_{iU} being, respectively, the number of times an allele/haplotype i is transmitted and non-transmitted, considering only heterozygous parental genotypes. The measure has a limiting χ^2 with $k-1$ df (χ^2_{k-1}) under no linkage.[19]

However, the recent explosion of GWASs that are being performed has revealed an important problem in multimarker association tests: a high difficulty to reproduce

results when a different data set is used.[11] The problem intensifies with the number of markers tested at a time and makes most of multimarker association tests useless for more than two or three markers. This is a very discouraging result for the use of multimarker tests, even if it is well known that power increases with them. The main cause of this lack of test reliability is that they are built on methods that define an alternative hypotheses that is sample dependent, so that power is overestimated, a problem known as sample overfitting in the Machine Learning field. In this situation only very large samples would yield reliable results.

Solutions based on the distribution estimation of the results may be inaccurate. Therefore, those performing a uniform correction of p-values as a function of the total number of hypothesis, such as the the Bonferroni correction, perform an over-correction, yielding very low power tests. As an example, the number of different tests in the simple TDT applied using binary markers is 2^m, with m being the number of biallelic markers. Other less-strict corrections such as the stepwise procedures[12] consider the outcome of all the tests and not only the number of tests to reject an hypothesis. Thus, they compute a corrected p-value considering all raw p-values. However, they may be poorly applied i.e. the number of hypothesis is under-estimated, as when the overall hypothesis is tested by evaluating a complex structure of many other hypothesis,[13] so that false positive results due to sample overfitting will translate into a lack of test reproducibility when a different data set is used.[14]

A simple approach to overcome the issue of sample overfitting is to use multiple sampling to assess statistical significance. Therefore, in its simplest modality, called holdout, in which the data set is split into two parts, one to build the model/hypothesis and the other to compute the statistic, the test has the same df than when using another approach but only half of the data set. Although power will decrease due to a reduction in the sample size used to compute the statistic, the model built using half of the sample will fit to the other half only in the case of true association, as data used to build a model (for $mTDT$ the model is defined just with a set of haplotypes present in a data set), are never used to test it.[15] Therefore, this approach is simple and provides a highly reproducible solution, as associations found are mostly due to real associations and not to false positive results because of multiple testing.

Another problem of many multimarker tests such as $mTDT$ [Eq. (2)], $mTDT_S$[8,16] — a score-based method that modifies $mTDT$ to follow an exact χ^2_{k-1} distribution under the null — or other tests alike, is that df increases with the number of markers, so that their power is downwardly affected.[9,11] A common solution is to group haplotypes based on different criteria, such as ancestral proximity of haplotypes[17] or haplotype similarity.[9,13,14,18]

Reduction in df is led to its maximum when they do not change with the number of markers. This is the approach of TDT_1,[19] a χ^2_1 multimarker test which checks differences in transmissions between the haplotype with the largest differences and the rest of haplotypes. The test is very powerful in the case of no mutation occurring from the non-recombinant haplotype with the disease variant. Another χ^2_1 test

regardless the number of markers, $2G$,[15] improves power of TDT_1 by removing the assumption that the non-recombinant haplotype has never mutated. Therefore, instead of building a null hypothesis of no association for one risk haplotype, it tests the hypothesis of no association for the set of all risk haplotypes. $2G$ outperforms TDT_1 because it allows more complex models while keeping the same df as TDT_1. The model may represent multiple founders for a risk variant and more than one risk variant descending from the founder haplotype because of mutation/recombination. Similar ideas have been proposed for case/control or discordant-sib-pair studies but members of a risk group are in those studies not composed of haplotypes but of genotypes.[1–3]

Although $2G$ outperforms TDT_1 in power, low frequent haplotypes are highly unreliable and may reduce power. If several markers are used at a time, to disregard low frequent haplotypes is not a choice, as many haplotypes have very low counts. Thereby we would end up with very few data and thus a significant power reduction, close to the power of TDT_1, as the most recent mutations/recombinations from a high-risk haplotype (which usually have very low counts) that are still in association with the disease will be disregarded because of their low counts.

In this work we propose to use a method, $2GTree$, which modifies $2G$ by using the most reliable haplotypes to be classified as low- and high-risk haplotypes as prior information to compute the probability of the more ambiguous haplotypes to be high/low-risk ones. We assume that haplotype similarity (which at its maximum is called identity by state, IBS) means genetic similarity (which at its maximum is called identity by descend, IBD) so that given a haplotype i the most similar haplotype to it, j, is the one with the shortest time to their most recent common ancestor (MRCA): $t_{MRCA}(i,j) < t_{MRCA}(i,j')\forall j' \in \{1,\ldots,n; j' \neq i; j' \neq j\}$ and therefore it has a higher probability of being also a non-recombinant haplotype with the trait variant if j is considered to be a non-recombinant haplotype with the trait variant (a high-risk haplotype). Consequently, it favors models in which haplotype trees built on them show a low entropy i.e. haplotype branches are mostly composed of haplotypes belonging to the same low/high-risk group.

In Sec. 2 we give a detailed definition of the algorithm $2GTree$ and describe the data, simulated and real ones, used to test the algorithm. We show results in Sec. 3. Conclusions are provided in Sec. 4.

2. Methods

Before describing $2GTree$, we first describe $2G$, as our proposal, $2GTree$, is a modification of it.

2.1. *Overview of* $2G$

As the number of different haplotypes exponentially increases with the number of markers while sample sizes do not increase accordingly, $2G$ tries to keep power by

collapsing haplotypes. Therefore, $2G$ always divides a data set into two groups: high-risk haplotypes (g_1) and low-risk haplotypes (g_2). For the test to be highly powerful, neither homozygous parents or those parents with both haplotypes in the same group are used. Consequently, haplotypes in the same group are being considered equivalent since the test relies on the strong assumption of all haplotypes in a group having similar Relative Risk, $RR = \frac{\theta}{1-\theta}$ for group g_1, and $\frac{1}{RR} = \frac{1-\theta}{\theta}$ for group g_2, and θ being the probability for a parental genotype with one haplotype in each group of transmitting the haplotype in g_1 to the offspring. Under the null hypothesis of no association or linkage, RR is 1, so that θ is 0.5 in both groups.

There may be two possible explanations for having more than one high-risk haplotype. One is that all high-risk haplotypes share a common ancestor that mutated or was in linkage disequilibrium with a mutation associated with the trait under study. The other is that there may be more than one mutation at a genetic locus in association with a disease from which current haplotypes may have derived from, a situation known as founder heterogeneity.[20]

Even if the assumption of same relative risks may not be true in most of the populations, its simplicity compensates the accuracy of the assumption. Thereby it causes the test to especially outperform the other tests whenever the length of the haplotypes increases. Thus, while most of the tests increase model complexity (df) with the haplotype length and become powerless because of a limited sample size, $2G$ has a constant model complexity (df) regardless of the haplotype length.

Given groups g_1 and g_2 have been made up, the statistic is defined as:

$$2G = \frac{(n_{g_1g_2} - n_{g_2g_1})^2}{n_g}, \tag{3}$$

with $n_g = n_{g_1g_2} + n_{g_2g_1}$ being the total amount of parental heterozygous genotypes with one haplotype in each group. The test is a McNemar test following a χ^2_1 distribution under the null hypothesis of no linkage or association.[15]

Among the $\binom{k}{2} = \frac{k(k-1)}{2}$ different ways to make up groups g_1 and g_2 from a set of k different haplotypes, it is straightforward to show which is the solution achieving the largest power. In fact, this solution defines the groups by considering as high-risk haplotypes (g_1) those with more transmitted haplotypes in the population i.e. $\theta_i > 1/2$, and as low-risk haplotypes (g_2) those with more non-transmitted haplotypes in the population i.e. $\theta_i < 1/2$, with θ_i being the probability for a haplotype h_i of being transmitted. $2G$ uses this criterion:

$$g(h_i) = \begin{cases} g_1 & \text{if } \theta_i > 1/2 \\ g_2 & \text{if } \theta_i < 1/2 \\ \emptyset & \text{if } \theta_i = 1/2, \end{cases} \tag{4}$$

and estimates θ_i by using the maximum likelihood estimator (MLE) $\tilde{\theta}_i = \frac{n_{iT}}{n_{iT}+n_{iU}}$.

To avoid the problem of sample overfitting, which increases with k, $2G$ uses the holdout approach. Thus, a data subset used to choose the best hypothesis, the training data set does not share any genotype with the data subset used to assess statistical significance, the test data set.

The number k of different haplotypes exponentially increases with the number of markers. Therefore, for the test to be applied on a window of a few markers, many haplotypes in the test data set will not be found in groups g_1 or g_2. In this situation, a group will be assigned to a haplotype h_i by using the following criterion:

$$h_i \in \begin{cases} g_1 & \text{if } s_M(h_i, g_1) > s_M(h_i, g_2) \\ g_2 & \text{if } s_M(h_i, g_2) > s_M(h_i, g_1), \end{cases} \tag{5}$$

with $s_M(h_i, g_x), x \in \{1, 2\}$ being defined as the similarity between h_i and the haplotype in g_x most similar to h_i. In the case of same similarity to both groups, a group is randomly chosen. As similarity measure, $2G$ uses the length measure,[9,13] which equals the largest number of consecutive markers with matching alleles and which is also used in $mTDT_{LC}$ and $mTDT_{SR}$.[9] As an example for 4-snp-long haplotypes of biallelic markers 1010 and 1111, the length measure is 1, as the largest strand of matching alleles has size 1. Thus, there are two strands of size 1: the one with only the first allele (1) and the one with only the third allele (1). In a second example, the length measure between haplotypes 1101 and 1111 is 2, as there is only one strand of size 2 composed of the two former alleles: 10.

2.2. 2*GTree*

In this section we introduce 2*GTree*, a Bayesian modification of $2G$ which, in order to compute the probability of a haplotype of being a high-risk one, it considers as prior knowledge how similar it is to other high-risk haplotypes. 2*GTree* assumes IBS means IBD as prior knowledge. In general, to obtain prior knowledge, it assumes that haplotype similarity implies genetic linkage.

Under the null hypothesis of no association or linkage, the random variable X_{ij} representing the number of times a heterozygous parent with haplotypes $\{h_i, h_j\}$ transmits h_i to their offspring, follows a binomial distribution with parameters $(N_{ij}, 1/2)$, with N_{ij} being the number of heterozigotic parents with haplotypes $\{h_i, h_j\}$ in the sample. Therefore, since X_{ij} and X_{ik} are independent for any $k \neq j$, $X_i = \sum_{j \neq i} X_{ij}$ is binomial with parameters (N_i, θ_i), being $N_i = \sum_{j \neq i} N_{ij}$ and $\theta_i = 1/2$. Therefore, every situation with $\theta_i \neq 1/2$ for any X_i means an alternative hypothesis of association or linkage holds. Most of the multimarker TDTs, including $mTDT$ and $2G$, use the fact that for large samples, $2X_i$ is normal with mean and variance N_i and thus the square root of Y_i,

$$Y_i = \frac{(X_i - N_i)^2}{N_i}, \tag{6}$$

is standard normal and Y_i is χ^2_1.[16]

Our intention is to use a Bayesian approach to estimate θ_i in the training sample as the expectation of θ_i with respect to the posterior probabilities:

$$\hat{\theta}_i = E_{p(\theta_i|D)}(\theta_i) = \int \theta_i p(\theta_i \mid D) d\theta_i, \tag{7}$$

with D being the data set, so that h_i will be assigned to a group by using the same criterion as $2G$ [see Eq. (4)] but a Bayesian estimator $\hat{\theta}_i$ of θ_i.

As X_i is a binomial distribution with parameters (N_i, θ_i), we can assume the prior distribution for θ_i is a $Beta(\alpha_1, \alpha_2)$ distribution[21] which will have $\alpha_1 = \alpha_2$ in the case we believe the null hypothesis is true. Thereby, the larger the difference between α_1 and α_2, the larger the prior knowledge against the null hypothesis.

To obtain the hyperparameters α_1, α_2 for the *Beta* distribution, $2GTree$ uses information about haplotype similarities assuming the coalescence model with no recombination[22] since only a small number of consecutive markers are used.[23] Therefore it assumes all haplotypes descend from a common ancestor without recombination and they make up a haplotype tree, which have haplotypes at its nodes. A haplotype tree can be seen in Fig. 1. Similarity between haplotypes translates into a shorter distance between their respective nodes in the haplotype tree.

In the case of haplotypes with no differences in transmissions, the prior has a large effect and the haplotype will be assigned to the group to which the most similar

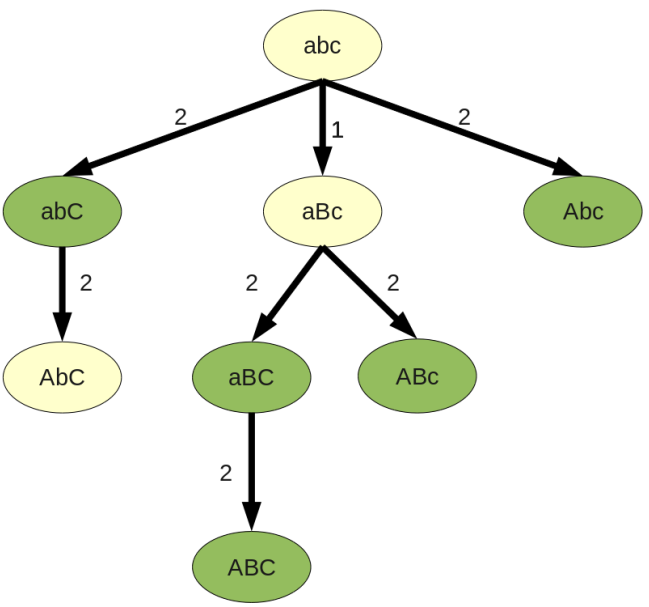


Fig. 1. Resulting haplotype tree built by 2GTree (Algorithm 1). Nodes in green represent low-risk haplotypes (g_2) and nodes in light yellow represent high-risk haplotypes (g_1). An arc between nodes a and b has been weighted by the length measure among the nodes connected.

haplotype belongs to. When differences increase, the prior decreases its effect on the posterior probability of a haplotype to be a low or a high-risk haplotype.

2*GTree* first sorts all different haplotypes $h_i, i = 1 \ldots k$ by descending effect difference $e(h_i)$ in a training data set, measured as:

$$e(h_i) = \frac{(n_{iT} - n_{iU})^2}{n_{iT} + n_{iU}}, \tag{8}$$

and for each haplotype h_i following this ordering criterion, it applies Eq. (4) to decide the group it belongs to. To make this decision, it estimates θ_i by using its Bayesian estimator $\hat{\theta}_i$ [Eq. (7)] instead of the MLE $\tilde{\theta}_i$ used by 2*G*. The purpose of sorting haplotypes by descending effect difference is to start by those haplotypes with less uncertainty about to which group they belong to. Thus, there will be more chances of a right decision at the former haplotypes and therefore, this information, which is used as a prior for the remaining haplotypes, will also increase chances of a right decision in the following haplotypes. Hyperparameters α_{i1} and α_{i2} are chosen for each haplotype by considering the current composition of groups g_1 and g_2 and by using the following rule:

$$\alpha_{i1} = \begin{cases} \alpha/2 & \text{if } |g_1| = 0 \quad \text{or} \quad |g_2| = 0 \\ \alpha/4 + (\alpha/2) r_i & \text{otherwise,} \end{cases}$$

with $|g_x|, x = 1, 2$ being the number of haplotypes in group g_x, $r_i = \frac{n_1(h_i)}{n_1(h_i)+n_2(h_i)}$ and $n_x(h_i), x \in \{1, 2\}$ being defined as the number of haplotypes in group g_x with the maximum similarity to h_i among all the haplotypes in $g_1 \cup g_2$:

$$n_x(h_i) = \sum_{\substack{h_j \in g_x \\ sim(h_i,h_j) = \\ max(s_M(h_i,g_1), s_M(h_i,g_2))}} n_j,$$

$n_j = n_{jT} + n_{jU}$ and $sim(a, b)$ being the similarity between haplotypes a and b. The intuition is to consider a largest prior probability to the group with a largest number of haplotypes with minimum distance to the one we are considering. r_i represents the proportion of haplotypes in group g_1. Therefore, both groups will have the same prior probability ($r_i = 0.5, \alpha_{i1} = \alpha_{i2}$) if both groups have the same number of haplotypes with minimal distance. In the extreme i.e. when there are no haplotypes in a group with minimal distance, $r = 1, \alpha_{i1} = 3\alpha_{i2}$ if $s_M(h_i, g_1) > s_M(h_i, g_2)$ as it means that $n_2(h_i) = 0$, and $r = 0, \alpha_{i2} = 3\alpha_{i1}$ in the case $s_M(h_i, g_1) < s_M(h_i, g_2)$, as it means that $n_1(h_i) = 0$.

The hyperparameter $\alpha = \alpha_{i1} + \alpha_{i2}, \forall\, i = 1, \ldots, k$ has been set to 2.[21] The hyperparameter α, sometimes called "precision," can be regarded as an equivalent sample size for the prior knowledge.[21] We chose it to be 2 so that a wrong prior will have a small effect in the final power unless the sample size is really small. We considered different sample sizes in our experiments from 125 trios to 1000 and even with the shortest data sets a wrong prior translated into very slight power decay of 2*GTree* compared with 2*G* (data not shown). In the case of at least a group being empty, a

$Beta(1,1)$ will be assumed so that every parameter in the likelihood distribution $p(D \mid \theta) \approx Bin(N_i, \theta)$ will have the same prior probability.

Figure 2 shows the $Beta$ distributions used when: (1) $s_M(h_i, g_1) > s_M(h_i, g_2)$: $Beta(1.5, 0.5)$ (top plots) so that there are larger prior support for haplotype h_i to belong to group g_1, (2) there are no haplotypes in at least one of the groups: $Beta(1,1)$ (left middle plot) so that there are not any prior knowledge about the probability of haplotype h_i to belong to a group (uniform distribution) and (3) $s_M(h_i, g_2) > s_M(h_i, g_1)$: $Beta(0.5, 1.5)$ (remaining plots) so that there are larger prior support for haplotype h_i to belong to group g_2.

As the $Beta(\alpha_{i1}, \alpha_{i2})$ distribution is the conjugate of the binomial distribution $Bin(n_1(h_i), \theta)$, the posterior probability is a $Beta(\alpha_{i1} + n_1(h_i), \alpha_{i2} + n_2(h_i))$ distribution and the Bayesian estimate of θ_i computed as the expectation of θ_i: $\hat{\theta}_i$ [Eq. (7)] has a close form solution [21]:

$$\hat{\theta}_i = \frac{\alpha_{i1} + n_{iT}}{\alpha_{i1} + \alpha_{i2} + n_{iT} + n_{iU}}. \tag{9}$$

Equations (4) and (9) are applied by 2*GTree* for all the haplotypes so that at each step a new haplotype will be added to a group. Table 1 shows the grouping strategy of 2*GTree* in pseudocode.

For each haplotype window $i = 1, \ldots, m$, 2*GTree* requires to loop over the $2n$ haplotypes in the sample to compute n_{iT} and n_{iU}, the starting point of its grouping strategy (see Table 1), as well as 2*G*, *mTDT*, $mTDT_S$ and TDT_1 do. To group haplotypes, it requires also to loop over the k different haplotypes found (it has to be noted that $k \ll n$ even for haplotype windows of several markers due to linkage disequilibrium) in order to compute $e(h_i)$, a measure required as well as input data by the grouping strategy. The first step in the grouping strategy (sorting haplotypes by $e(h_i)$) is known to be $O(k \log k)$ by the quicksort algorithm. The third and last time-consuming step requires to compute the length measure among all haplotypes already assigned to a group, so that $s \times k(k-1)/2$ operations, s being a constant (window size), will be performed. Therefore, 2*GTree* will have as time complexity upper bound $O(m(n + k^2))$, the same as $mTDT_S$ or, if we take into account $k \ll n$, we just have simply $O(mn)$ as the upper bound for 2*G*, 2*GTree*, *mTDT*, $mTDT_S$ and TDT_1 algorithms.

Table 2 shows an example of an input array built from all the heterozygous parents of a training data subset artificially generated. The second column shows the transmission counts of each different haplotype h_i in the training data set, the third column shows the non-transmission counts and the four column shows the effect difference $e(h_i)$. Table 3 sorts Table 2 by descending effect difference (fourth column) and add other columns with results that were required by the grouping strategy of 2*GTree*: $n_{iT} + n_{iU}$ (column 5), closest haplotypes in current g_1 (column 6), closest haplotypes in current g_2 (column 7), $s_M(h_i, g_1)$ (column 8), $s_M(h_i, g_2)$ (column 9), r_i

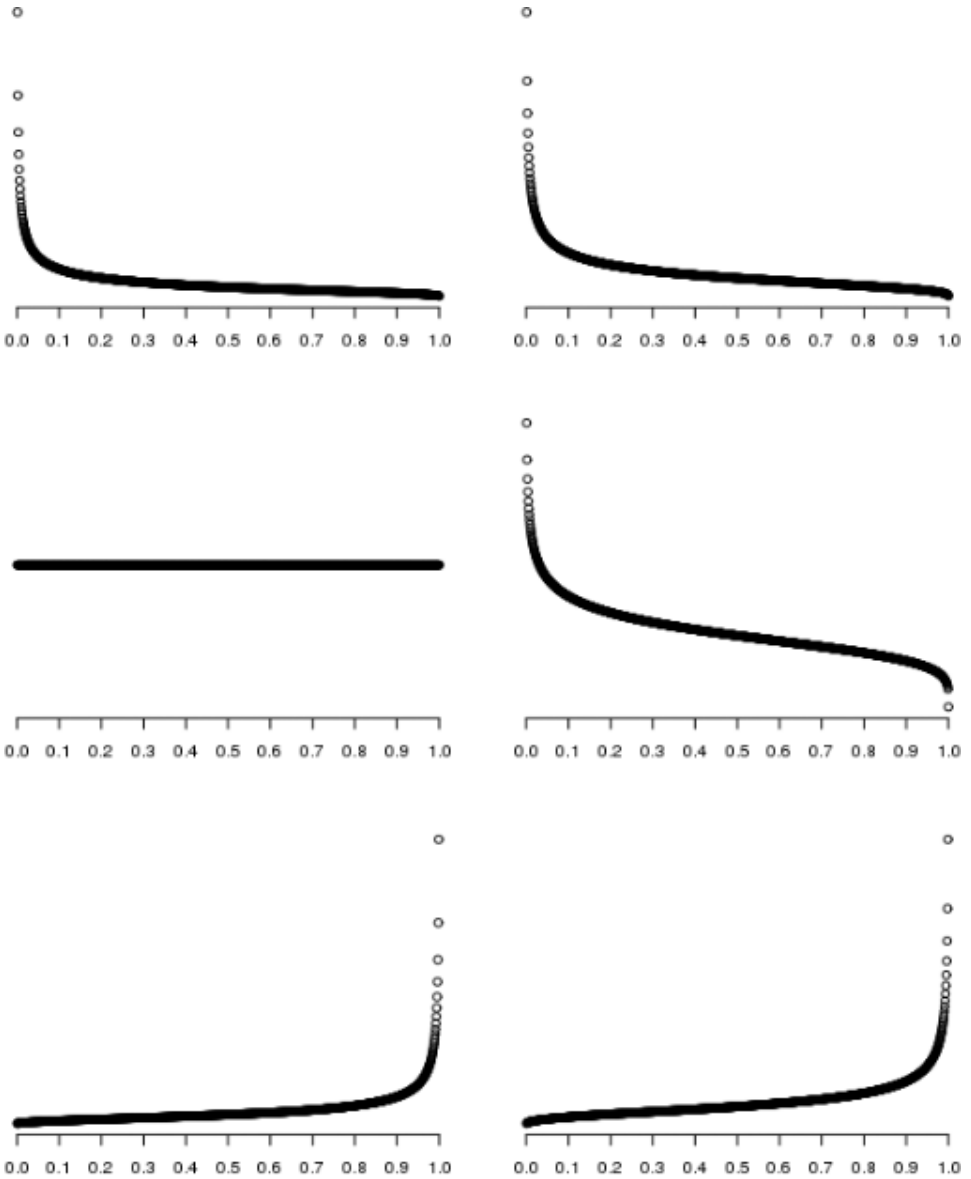


Fig. 2. Different $Beta\,(\alpha_1, \alpha_2)$ distributions, $\alpha_1 + \alpha_2 = 2$. Left column from top to bottom: α_1 values are $1.5, 1$ and 0.5. $\alpha_1 = 1$ is used when there is no prior knowledge about the group a haplotype should belong to. $\alpha_1 = 1.5$ and $\alpha_1 = 0.5$ is used when a haplotype is *a priori* considered as belonging to group g_1, g_2, respectively, because of similarity. Right column from top to bottom: α_1 values are $1.4, 0.8$ and 0.6. The larger the difference $|\alpha_1 - \alpha_2|$, the larger the prior evidence for the haplotype to belong to the group $g_x, x \in \{1, 2\}$ with the greatest α_x.

Table 1. 2GTree grouping strategy in pseudocode.

Input data: a $k \times 4$ array A with a row for each
different haplotype and four columns:
c1: haplotype value h_i,
c2: n_{iT},
c3: n_{iU},
c4: effect difference: $e(h_i)$
Output data: a list g_1 of high-risk haplotypes
and a list g_2 of low-risk haplotypes
Algorithm:

Sort rows of array A by descending order of column 4
$alpha \leftarrow 2$
For each row r_i in $A, i = 1 \ldots k$
  $alpha1 \leftarrow 0.25 \times alpha$
  if $g_1 \neq \emptyset$ and $g_2 \neq \emptyset$ then
    $h_i \leftarrow r_i[1]$
    $n_{iT} \leftarrow r_i[2]$
    $n_{iU} \leftarrow r_i[3]$
    $n1 \leftarrow 0$
    $n2 \leftarrow 0$
    $sim1 \leftarrow s_M(h_i, g_1)$
    $sim2 \leftarrow s_M(h_i, g_2)$
    $maxSim \leftarrow max(sim1, sim2)$
    For each row r_j in $A, j = 1 \ldots i-1$
      if ($r_j \in g_1$ and $sim(r_i, r_j) == maxSim$) $n1 \leftarrow n1 + 1$
      if ($r_j \in g_2$ and $sim(r_i, r_j) == maxSim$) $n2 \leftarrow n2 + 1$
    $r_i \leftarrow n1/(n1 + n2)$
    $alpha1 \leftarrow alpha1 + r$
  $alpha2 \leftarrow alpha - alpha1$
  $theta \leftarrow (alpha1 + n_{iT})/(alpha1 + n_{iT} + alpha2 + n_{iU})$
  if ($theta > 0.5$) $g_1 \leftarrow g_1 \cup h_i$
  if ($theta < 0.5$) $g_2 \leftarrow g_2 \cup h_i$

Table 2. An example of an array with haplotype counts required by *2GTree* to build groups g_1 and g_2. Counts must come from the heterozygous parents in the training data subset.

Haplotype	n_{iT}	n_{iU}	$e(h_i)$
AbC	13	7	1.8
aBC	5	15	5
ABC	4	4	0
ABc	9	14	1.09
abc	24	3	16.33
aBc	12	3	5.4
abC	2	13	8.07
Abc	11	21	3.13

Table 3. An extension of the example in Table 2 with partial and final results derived when applying algorithm in Table 1.

Haplotype	n_{iT}	n_{iU}	$e(h_i)$	$n_{iT}+n_{iU}$	$\arg s_M(h_i,g_1)$	$\arg s_M(h_i,g_2)$	$s_M(h_i,g_1)$	$s_M(h_i,g_2)$	r_i	α_{i1}	α_{i2}	$\hat{\theta}_i$	$g(h_i)$
abc	24	3	16.33	27	—	—	—	—	—	1	1	0.862	g_1
abC	2	13	8.07	15	—	—	—	—	—	1	1	0.176	g_2
aBc	12	3	5.4	15	abc	abC	1	1	0.5	1	1	0.765	g_1
aBC	5	15	5	20	aBc	abC	2	1	1	1.5	0.5	0.295	g_2
Abc	11	21	3.13	32	abc	abC	2	1	1	1.5	0.5	0.368	g_2
AbC	13	7	1.8	20	abc	abC, Abc	1	2	0	0.5	1.5	0.614	g_1
ABc	9	14	1.09	23	aBc	aBC, Abc	2	1	1	1.5	0.5	0.438	g_2
ABC	4	4	0	8	aBc	aBC, ABc	1	2	0	0.5	1.5	0.45	g_2

(column 10), α_{i1} (column 11), α_{i2} (column 12), $\hat{\theta}_i$ (column 13) and the final group the haplotype belongs to (column 14).

The haplotype tree which summarizes the steps conducted by 2*GTree* to build the groups is shown in Fig. 1. The most frequent haplotype h_i with $\hat{\theta}_i \neq 0$ is considered the root of the tree, and is the first haplotype added to a group. The haplotype with largest similarity (largest length measure) is considered the closest ancestor of a haplotype. In case of a tie, the haplotype in the same group is chosen. If all the most similar haplotypes are in the other group, the most frequent haplotype is chosen.

Once groups g_1 and g_2 have been made up with the training data set, 2*GTree* has no differences with 2*G* in the way the test is computed by using the test data set.[15] Therefore, each haplotype h_i in the test data set is assigned to a group by computing the function $g(h_i)$ and the statistic in Eq. (3).

2.3. *Simulation studies*

We have drawn haplotype samples of 500 familial trios under different standard configurations to check type-I errors under stratified and admixture populations and power.[9,18] In general, it means different haplotype lengths and different frequencies of common allele variants. For testing type-I errors, it means two stratified populations with different proportions among them ($1/2, 1/4$ and $1/6$) and different minor allele frequencies for one of the two subpopulations ($0.1, 0.3$ and 0.5) and 0.5 always for the other. For testing power, it means we considered scenarios with one and two disease susceptibility loci under different genetic models (additive, recessive and dominant for one locus; additive, recessive-or-recessive, dominant-or-dominant, dominant-and-dominant, threshold and modified for two loci).

However, we introduced several differences.[11,15] For both studies (population stratification/admixture and power) we considered a wide range of haplotype lengths, $1, 2, 5, 10, 15$ and 20. The studies above referred[9,18] only used one or two markers or sometimes three but as it has been said, the problem of model overfitting increases with the number of markers. Moreover, in the power study, we reduced relative risks to more realistic values $1.2, 1.6, 2.0, 2.4$ and 2.6. Populations were generated assuming the now standard coalescent approach.[22] The power study was enlarged by a locus-specificity study so that instead of testing a set of markers at one of the disease susceptibility locus (recombination fraction $\theta = 0$) we also considered an increase in $\theta = 0.00005, 0.0001, 0.00015, 0.0002$, in the way used in other works.[10,11] To increase sample reproducibility, all the tests were applied under the holdout approach, in which for each data set, 250 trios were randomly chosen to learn the model (the training data set) and 125 were randomly chosen to assess statistical significance i.e. to obtain p-values (the test data set). The same length measure [9,13] was used in all of them in order to assign haplotypes in the test data set to the models learned with the training data set.[15] A detailed description of the way simulations were performed can be seen at the supplementary website.

2.4. *Real data*

To test power, we have chosen nine different risk loci in which SNPs in association with a complex disease or in strong linkage disequilibrium have been reported, one of them related with Crohn disease[24] and the others with MS.[26] The Crohn data set (Crohn-affected) is a publicly available set that was originally used in 2001.[25] It consists of the genotype data of 103 SNPs typed in 129 trios with offspring having Crohn's disease.[24] The phenotype is the presence/absence of Crohn disease. The SNPs span across 500 kilobases at the $IBD5/SLC22A4$ locus (5q31), and the region contains 11 known genes. For MS disease, genotype information was obtained from a GWAS performed by the International Multiple Sclerosis Genetic Consortium with 334,923 SNPs in 931 family trios with affected offspring.[26]

To test specificity, we need data sets with all the family members being unaffected. To do that, we have used the same nine risk loci but family trios and their genotypes were obtained from the 30 Caucasian nuclear families (CEPH) used at Phase II of the International HapMap Project (IHMP).[27] For comparative purposes, and considering that different arrays were used for genetic sequencing in the data sets with affected offspring and in the IHMP data sets, we followed a procedure which mainly selected those markers present in both affected-offspring and IHMP data sets used for each risk loci.[15]

To reconstruct haplotypes from genotypes required to apply TDT or any of its generalizations, we used the common family-plus-E-M strategy,[9,18,28] in which family information is first used and, in those loci still unsolved, the E-M algorithm is used. Phasing errors are very unusual when using this method, as most positions are inferred without errors by using familial data.

To guarantee sample reproducibility, all the tests were applied under the holdout approach, in which for each data set, half of the trios were randomly chosen to learn the model (the training data set) and the other half were used to assess statistical significance i.e. to obtain p-values (the test data set).

3. Results

We have compared $2GTree$ with some state-of-the-art multimarker TDTs: $mTDT$, $mTDT_S$ — a score modification of $mTDT$[16] to guarantee that it asymptotically follows an exact χ^2_{k-1} under the null hypothesis of no linkage —, $mTDT_1$ and $2G$. We have chosen these tests because they have shown to be powerful tests with computational complexity lineal to the number of founders and the number of SNPs. As an example, $mTDT_S$, the slowest one, has upper bound time complexity $O(m(n+k^2))$, m being the number of windows being analyzed, n the sample size and k the number of different haplotypes in a window, with $k \ll n$ even for windows of several consecutive markers due to linkage disequilibrium. Therefore they are affordable to be used as tests to detect risk variants in GWASs. Other tests, such as the Length Contrast Test,[9] or the Signed Rank Test[9] have less power under a wide

Table 4. Type-I error rates in presence of population stratification and admixture.

α	MAFs	pp	l = 1	l = 5	l = 10	l = 15	l = 20
0.01	0.1	0.5	0.01	0.01	0.01	0.01	0.01
0.01	0.3	0.5	0.01	0.01	0.01	0.01	0.01
0.01	0.5	0.5	0.01	0.01	0.01	0.01	0.01
0.01	0.1	0.75	0.01	0.01	0.01	0.01	0.01
0.01	0.3	0.75	0.01	0.01	0.01	0.01	0.01
0.01	0.5	0.75	0.01	0.01	0.01	0.01	0.01
0.01	0.1	0.833	0.01	0.01	0.01	0.01	0.01
0.01	0.3	0.833	0.02	0.01	0.01	0.01	0.01
0.01	0.5	0.833	0.02	0.01	0.01	0.01	0.01
0.05	0.1	0.5	0.06	0.05	0.05	0.05	0.04
0.05	0.3	0.5	0.06	0.05	0.06	0.05	0.05
0.05	0.5	0.5	0.04	0.05	0.04	0.05	0.05
0.05	0.1	0.75	0.06	0.05	0.05	0.06	0.06
0.05	0.3	0.75	0.06	0.04	0.06	0.06	0.05
0.05	0.5	0.75	0.06	0.05	0.05	0.07	0.06
0.05	0.1	0.833	0.06	0.05	0.05	0.05	0.05
0.05	0.3	0.833	0.06	0.05	0.06	0.05	0.05
0.05	0.5	0.833	0.05	0.05	0.06	0.06	0.06

Note: Results for different minor allele frequencies (MAFs) in the second subpopulation (q) and different proportion of trios from the first subpopulation (pp), obtained by 2*GTree* for nominal levels $\alpha = 0.01$ and $\alpha = 0.05$ and haplotypes of length 1, 5, 10, 15 and 20 (columns 4 to 8 respectively).

range of scenarios and have computational complexity quadratic on the number of founders.[15]

By using simulations we first have shown that 2*GTree* is robust to population structure and admixture. Thus, Table 4 shows that type-I error rates are very close to the nominal α value used to reject the null hypothesis under different scenarios used to simulate population structure and admixture.

Once we have shown the test has the correct behavior under the null hypothesis of no linkage, even in the case of population structure and admixture, we have compared association rates between the state-of-the-art multimarker TDTs under a wide range of scenarios. Figure 3 shows results for haplotypes of length 20 and two disease susceptibility loci under the additive, dominant-and-dominant and recessive-or-recessive genetic models and nominal level $\alpha = 0.05$. Results for all the other configurations can be seen in Figs. S1 to S15 at the supplementary web site. As it can be observed in all the plots, 2*GTree* usually has the highest association rates at $\theta = 0$ i.e. the highest sensitivity rates (power), followed by 2*G*. Moreover, the test is also locus-specific, so that association rates experience a strong decay when recombination factor θ increases. This means that 2*GTree* will discriminate better between causal variants and those in linkage disequilibrium with them. Overall, and focusing on haplotypes of width 20 to make differences more clear, 2*GTree* outperforms in power (average sensitivity values from 100 data sets) 31 out of the 45 different scenarios all

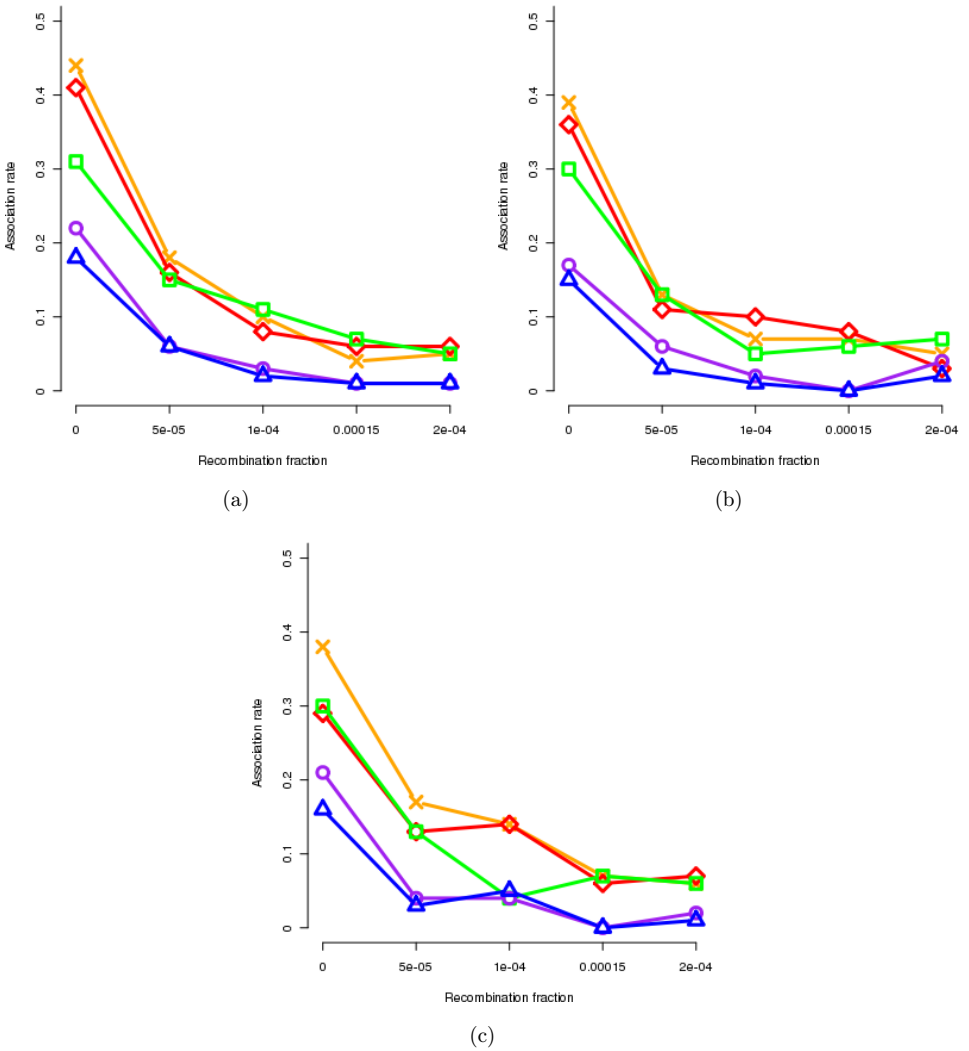


Fig. 3. Association rates under the holdout approach using a second data set to test reproducibility. Results for 100 simulations of 250 + 125 family trios as a function of the recombination rate using the (a) two-loci additive, (b) dominant-or-dominant and (c) recessive-or-recessive genetic models and haplotypes of length 20. A nominal level of $\alpha = 0.05$ and a relative risk of 1.6 were used for all plots. Results for $mTDT$, $mTDT_S$, $mTDT_1$, $2G$ and $2GTree$ are plotted in purple circles, blue triangles, green squares red diamonds and orange crosses, respectively.

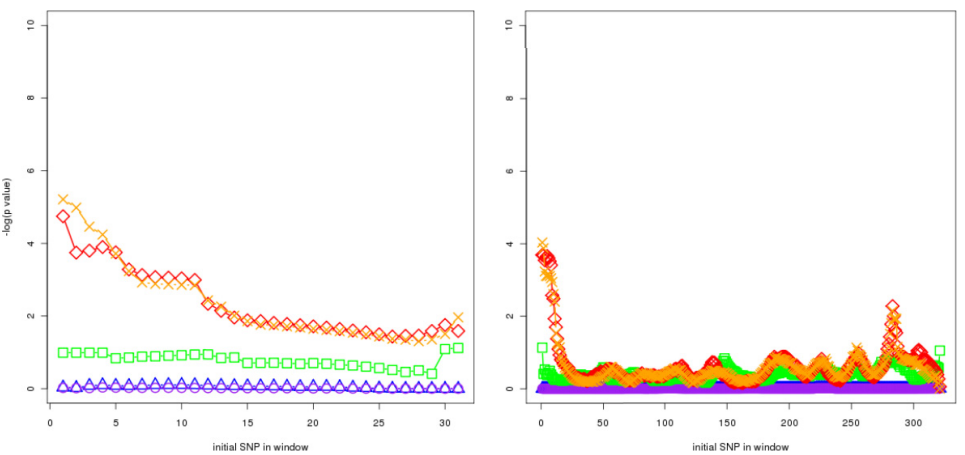
the other algorithms, $2G$ outperforms 8 out of 45 the other algorithms, TDT_1 only wins once (one-locus recessive model at relative risk 1.2) and $2G$ and $2GTree$ have 5 ties out of those 45 scenarios (see Table 5).

The same pattern can be seen when using real data sets. Figure 4 shows power results in sliding window maps (window size 15 and offset 1) for IL7R (a) and IL2R (b) locus, respectively. It is interesting to note that (1) $2GTree$ reaches the highest

Table 5. Summary of power results for haplotypes of length 20.

Algorithm	1 locus	2 loci A	2 loci B	Total
mTDT	0	0	0	0
$mTDT_S$	0	0	0	0
$mTDT_1$	1	0	0	1
2*G*	2	3	3	8
2*GTree*	11	10	10	31
Tie 2*G*-2*GTree*	1	2	2	5
Total	15	15	15	45

Note: Table cells show the number of times an algorithm has the largest power among all the algorithms used. All the scenarios above referred have been considered. Results are displayed by genetic models: one disease locus (column two), two disease loci A (additive, dominant-or-dominant and recessive-or-recessive genetic models) (column three) and two disease susceptibility loci B (dominant-and-dominant, threshold and modified genetic models) (column four).



(a) IL7R-Affected (b) IL2R-Affected

Fig. 4. Sliding window maps for the *IL7R* (a) and *IL2R* (b) affected data set. Window size is 15. TDTs used were 2*GTree* (orange crosses), 2*G* (red diamonds), $mTDT_1$ (green squares), *mTDT* (purple circles) and $mTDT_S$ (blue triangles).

association at the windows with the highest y-axis values, i.e. windows containing truly risk markers or markers in strong linkage disequilibrium with them (windows on the left in these plots) and (2) in other windows, i.e. windows that may show some degree of association just because of some extent of linkage disequilibrium, 2*GTree* is usually more monotonic and not the one with the highest association values. These results may respectively indicate that the test is (1) more powerful than the others and (2) more locus-specific as well. Figures S16 to S21 at the supplementary web site

show sliding windows for power and specificity results for all the data sets used. Figures S22 to S27 use comparative TDT (CTDT) maps[29] instead.

We have also shown that 2*GTree* is computationally affordable as a genome-wide association test, as it has lineal complexity to the number of positions to be analyzed, m, and the sample size n, with upper bound $O(m(n+k^2))$. As $k \ll n$ even for haplotype windows of several markers due to linkage disequilibrium, we can simplify its complexity in terms of m and n to be $O(mn)$.

4. Conclusion

We have introduced a Bayesian approach to improve 2*G*,[15] an algorithm to select loci in association with a disease by analyzing the genome of a set of nuclear families, affected offspring and their parents. The 2*G* algorithm is a very simple, efficient and highly reproducible multimarker Transmission/Disequilibrium method. Its simplicity is the cause of its high reproducibility, as it classifies haplotypes into only two groups: high- and low-risk haplotypes. However, power may decrease in presence of rare haplotypes. The measure presented in this work, 2*GTree*, uses a Bayesian approach so that prior knowledge is introduced in order to estimate the probability for an haplotype to be low or high risk. The prior knowledge is based on haplotype similarity, by assuming that the more similar the haplotypes are among them, the more recent their common ancestor is. Moreover, those haplotypes with more certainty about the group they belong to are the ones used to decide about the prior knowledge for those with higher uncertainty.

Simulation and real studies have shown that 2*GTree* usually reaches the same power as 2*G* does and many times it outperforms 2*G*. Those situations where power does not increase compared with 2*G* or it is sightly lower, may correspond to a wrong prior assumption. However, consequences in the final posterior distribution are small and power keeps very close to the one obtained by 2*G*.

We believe the proposed algorithm may be very helpful in genome-wide association studies. Although 2*G* and 2*GTree* converge with sample size, for small samples 2*GTree* may make a difference and help to discover causal variants in complex diseases. It may be especially important with the first data sets genotyped by using the next-generation sequencing technology, as the number of genotyped individuals may be small. The 2*G* technique is crucial as a method to choose and reduce the number of input variables when building genome-wide predictors of individual risk to complex diseases.[30,31] The improvement of 2*GTree* may be very important as well to increase the overall accuracy of these genetic predictors.

Web resources

A supplementary website has been created for this work at http://bios.ugr.es/*2GTree*, where Figs. S1−S27, the real data sets used, the software *trioSample*[11] implemented to obtain the samples upon which simulations were performed (scripts

for linux and software in c++) and 2*GTree*, the software used to implement the method, are available.

Acknowledgments

The authors were supported by the Spanish Research Program under project TIN2010-20900-C04-1, the Andalusian Research Program under project P08-TIC-03717 and the European Regional Development Fund (ERDF). We thank the reviewers for their helpful comments.

References

1. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH, Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *Am J Hum Genet* **69**:138−147, 2001.
2. Chung Y, Lee SY, Elston RC, Park T, Odds ratio based multifactor-dimensionality reduction method for detecting genegene interactions, *Bioinform* **23**(1):71−76, 2007.
3. Lee SY, Chung Y, Elston RC, Kim Y, Park T, Log-linear model-based multifactor dimensionality reduction method to detect gene gene interactions, *Bioinform* **23**(19):2589−2595, 2007
4. Spielman RS, McGinnis RE, Ewens WJ, Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM), *Am J Human Genet* **52**:506−516, 1993.
5. International Multiple Sclerosis Genetics Consortium, Evidence for polygenic susceptibility to multiple sclerosis — the shape of things to come, *Am J Human Genet* **86**:621−625, 2010.
6. Wray N, Goddard M, Visscher P, Prediction of individual genetic risk to disease from genome-wide association studies, *Genome Res* **17**:1520−1528, 2007.
7. BickeBöller H, Clerget-Darpoux F, Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers, *Genet Epidemiol* **12**:865−870, 1995.
8. Sham PC, Curtis D, An extended transmission/disequilibrium test (tdt) for multiallelic marker loci, *Annals Human Genet* **59**:323−336, 1995.
9. Yu K, Gu CC, Xiong C, An P, Province M, Global transmission/disequilibrium tests based on haplotype sharing in multiple candidate genes, *Genet Epidemiol* **29**:223−235, 2005.
10. Zhao J, Boerwinkle E, Xiong M, An entropy-based genome-wide transmission/ disequilibrium test, *Human Genet* **121**:357−367, 2007.
11. Abad-Grau M, Medina-Medina N, Montes-Soldado R, Moreno-Ortega J, Matesanz F, Genome-wide association filtering using a highly locus-specific transmission/disequilibrium test, *Human Genet* **128**:325−344, 2010.
12. Ge Y, Dudoit S, Speed T, Resampling-based multiple testing for microarray data analysis, *TEST* **12**:1−77 2003.
13. Sevon P, Toivonen H, Ollikainen V, Treedt: Tree pattern mining for gene mapping, *IEEE/ACM Trans Comput Biol Bioinf* **3**(2):174−185, 2006.
14. Moreno-Ortega JJ, Medina-Medina N, Montes-Soldado R, Abad-Grau MM, Improving reproducibility on tree based multimarker methods: Treedth. In *PACBB '11: Proc 5th Int Conf Practical Applications of Computational Biology and Bioinformatics* (Berlin,

Heidelberg, 2011), Rocha M, Corchado J, Fernández-Riverola F, Valencia A, eds., Vol. 1, Springer-Verlag, pp. 1−8.
15. Abad-Grau M, Medina-Medina N, Montes-Soldado R, Matesanz F, Bafna V, Sample reproducibility of genetic association using different multimarker tdts in genome-wide association studies: Characterization and a new approach, *PLoS ONE* doi:10.1371/journal.pone.0029613, 2012.
16. Sham PC, Transmission/disequilibrium tests for multiallelic loci, *Am J Human Genet* **61**:774−778, 1997.
17. Seltman H, Roeder K, Devlin B, Transmission/Disequilibrium test meets measured haplotype analysis: Family-based association analysis guided by evolution of haplotypes, *Am J Human Genet* **68**:223−235, 2001.
18. Zhang S, Sha Q, Chen H, Dong J, Jiang R, Transmission/Disequilibrium test based on haplotype sharing for tightly linked markers, *Am J Human Genet* **73**:566−579, 2003.
19. Ott J, *Analysis of Human Genetic Linkage*, John Hopkins, Baltimore, MD, 1999.
20. Yu K, Gu CC, Province M, Xiong C, Rao DC, Genetic association mapping under founder heterogeneity via weighted haplotpe similarity analysis in candidate genes, *Genet Epidemiol* **27**:182−191, 2004.
21. Heckerman D, Bayesian networks, *Science* **18**:1072−1079, 1995.
22. Hudson R, Generating samples under a wright-fisher neutral model of genetic variation, *Bioinformatics* **18**:337−338, 2002.
23. Powell J, Visscher P, Goddard M, Reconciling the analysis of ibd and ibs in complex trait studies, *Nat Rev Genet* **11**:800−805, 2010.
24. Daly M, Rioux J, Schaffner S, Hudson T, Lander E, High-resolution haplotype structure in the human genome, *Nat Genet* **29**:229−32 2001.
25. Rioux JD, Daly MJ, Silverberg MS *et al.*, Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to crohn disease, *Nature Genet* **29**:223−228 2001.
26. International Multiple Sclerosis Genetics Consortium, Compston A, Lander SSE, Daly M, Jager PD, de Bakker P, Gabriel S, DM, Pericak-Vance AIM, Gregory S, Rioux J, McCauley J, Haines J, Barcellos L, Cree B, Oksenberg J, Hauser S, Risk alleles for multiple sclerosis identified by a genomewide study, *New Engl J Med* **357**(9):851−862, 2007.
27. HapMap-Consortium TI, The international hapmap project, *Nat* **426**:789−796, 2003.
28. Abecasis GR, Martin R, Lewitzky S, Estimation of haplotype frequencies from diploid data, *Am J Human Genet* **69**:198, 2001.
29. Montes R, Abad-Grau MM, Biocase: Accelerating software development of genome-wide filtering applications, In *IWANN '09: Proc 10th Int Work-Conference on Artificial Neural Networks* (Berlin, Heidelberg, 2009), Omatu S, Rocha M, Bravo J, Corchado E, Eds., Vol. 5518, Springer-Verlag, Berlin, pp. 1097−1100.
30. Torres-Sánchez S, Medina-Medina N, Montes-Soldado R, Masegosa AR, Abad-Grau MM, Riskoweb: Web-based genetic profiling to complex disease using genome-wide snp markers, *Proc 5th Int Conf Practical Applications of Computational Biology & Bioinformatics (PACBB 2011)*, Rocha MP, Corchado JM, Fdez-Riverola F, Valencia A, (eds.) Vol. 1, pp. 1−8.
31. Abad-Grau M, Medina-Medina N, Masegosa A, Moral S, Haplotype-based classifiers to predict individual susceptibility to complex diseases: An example for multiple sclerosis, *Proceedings of Bioinform 2012*, pp. 360−366.

María M. Abad-Grau received her Ph.D. in Computer Science from the University of Murcia, Spain, in 2002. She is associate professor at the Department of Computer Languages and Systems, University of Granada and her main research interest is the application of Machine Learning to discover genetic bases of complex diseases.

Nuria Medina-Medina received the Ph.D. in Computer Science from the University of Granada in 2004. Since 2001, she works as researcher and associate professor in the Department of Computer Languages and Systems of this Spanish University. She is member of the GEDES research group in specification, development and evolution of software, http://www-lsi.ugr.es/∼gedes. Her main research interests include hypermedia systems, user modeling, user adaptation, and software evolution. In addition, recently she has worked on other topics such as Web browsing, refactoring for visually impaired and bioinformatics.

Serafín Moral received a Ph.D. in Fuzzy Information, Relationships between Possibility and Probability from the University of Granada in 1985. He is currently Professor of the Department of Computer Science and Artificial Intelligence at the University of Granada and member of the Uncertainty in Artificial Intelligence research group. His current areas of research interest are reasoning with imprecise probabilities, propagation algorithms in dependence graphs, relationships between uncertainty reasoning and non-monotonic logics.

Rosana Montes-Soldado holds a Ph.D. degree in Computer Science from the University of Granada in the field of Computer Graphics and Realistic Image based Rendering, though some of her research had been involved with eLearning and virtual 3D worlds. Over the last years, he has participated in several Longlife Learning European projects and now is the project coordinator of OERtest http://oer-europe.net. She is researcher at the Software Engineering Department and currently she is the Secretariat of the Virtual Learning Centre, University of Granada (Granada-Spain).

Sergio Torres-Sánchez received a M.Sc. in Computer Engineering in 2010 and in Software Development in 2011 from the University of Granada, Spain. During 2010 he joined the Bioinformatics Research Group at the Department of Languages and Computer Systems (University of Granada). Over the last year he has been working in this group and at present he is a Ph.D. student focusing on the automatic learning of genetic models to predict individual risk to polygenic diseases.

Fuencisla Matesanz received her Ph.D. in Biological Science from Autónoma University of Madrid, Spain in 1992. She is an associate professor since 2005 at the Instituto de Parasitología y Biomedicina López Neyra of the Consejo Superior de Investigaciones Científicas of Spain. Her main research interest is the determination of the environmental and genetic bases of Multiple Sclerosis aetiology.