Detecting multi-way epistasis in family-based association studies

Cheikh Loucoubar, Audrey Grant, Jean-François Bureau, Isabelle Casademont, Ndjido Ardo Bar, Avner Bar-Hen, Mamadou Diop, Joseph Faye, Fatoumata Diène Sarr, Abdoulaye Badiane, et al.

Briefings in Bioinformatics, 18(3), 2017, 394–402

doi: 10.1093/bib/bbw039
Advance Access Publication Date: 13 May 2016
Paper

Detecting multi-way epistasis in family-based association studies

Cheikh Loucoubar*, Audrey V. Grant*, Jean-François Bureau, Isabelle Casademont, Ndjido Ardo Bar, Avner Bar-Hen, Mamadou Diop, Joseph Faye, Fatoumata Diene Sarr, Abdoulaye Badiane, Adama Tall, Jean-François Trape, Freddy Cliquet, Benno Schwikowski, Mark Lathrop, Richard Edward Paul and Anavaj Sakuntabhai

Corresponding author: Anavaj Sakuntabhai, 25-28 Rue du Docteur Roux, 75015 Paris, France. Tel.: +33 1 44 38 91 03; E-mail: anavaj.sakuntabhai@pasteur.fr
*These authors contributed equally to this work.

Abstract

The era of genome-wide association studies (GWAS) has led to the discovery of numerous genetic variants associated with disease. Better understanding of whether these or other variants interact leading to differential risk compared with individual marker effects will increase our understanding of the genetic architecture of disease, which may be investigated using the family-based study design. We present M-TDT (the multi-locus transmission disequilibrium test), a tool for detecting family-based multi-locus multi-allelic effects for qualitative or quantitative traits, extended from the original transmission disequilibrium test (TDT). Tests to handle the comparison between additive and epistatic models, lack of independence between markers and multiple offspring are described. Performance of M-TDT is compared with a multifactor dimensionality reduction (MDR) approach designed for investigating families in the hypothesis-free genome-wide setting (the multifactor dimensionality reduction pedigree disequilibrium test, MDR-PDT). Other methods derived from the TDT or MDR to investigate genetic interaction in the family-based design are also discussed. The case of three independent biallelic loci is illustrated using simulations for one- to three-locus alternative hypotheses. M-TDT identified joint-locus effects and distinguished effectively between additive and epistatic models. We showed a practical example of M-TDT based on three genes already known to be implicated in malaria susceptibility. Our findings demonstrate the value of M-TDT in a

Cheikh Loucoubar is the Director of the Biostatistics, Bioinformatics and Modeling Group at the Institut Pasteur de Dakar.
Audrey V. Grant is a research associate in Statistical Genetics at the McGill University and Genome Quebec Innovation Center.
Jean-François Bureau is a research director at the Functional Genetics of Infectious Diseases Laboratory at the Pasteur Institute.
Isabelle Casademont is a research engineer in the Functional Genetics of Infectious Diseases Laboratory at the Pasteur Institute.
Ndjido Ardo Bar was a PhD student in Statistics and Machine Learning in the Functional Genetics of Infectious Diseases Laboratory at the Pasteur Institute.
Avner Bar-Hen is a professor in Applied Mathematics in the Department of Mathematics and Computer Science at Paris Descartes University.
Mamadou Diop is a research engineer in the Biostatistics, Bioinformatics and Modeling Group at the Institut Pasteur de Dakar.
Joseph Faye is an engineer in the Epidemiology of Infectious Disease Unit at Institut Pasteur de Dakar.
Fatoumata Diene Sarr is a staff scientist and medical doctor in the Epidemiology of Infectious Disease Unit at Institut Pasteur de Dakar.
Abdoulaye Badiane is a nurse in the Epidemiology of Infectious Disease Unit at Institut Pasteur de Dakar.
Adama Tall is a staff scientist and medical doctor in the Epidemiology of Infectious Disease Unit at Institut Pasteur de Dakar.
Jean-François Trape is a research director at the Institut de Recherche pour le Développement.
Freddy Cliquet is a postdoctoral researcher in the Systems Biology Laboratory at the Institut Pasteur.
Benno Schwikowski is the Director of the Systems Biology Laboratory at the Institut Pasteur.
Mark Lathrop is the Scientific Director of the McGill University and Genome Quebec Innovation Center.
Richard Edward Paul is a group leader in the Functional Genetics of Infectious Diseases Laboratory at the Pasteur Institute.
Anavaj Sakuntabhai is the Director of the Functional Genetics of Infectious Diseases Laboratory at the Pasteur Institute.
Submitted: 28 December 2015; **Received (in revised form):** 26 February 2016

hypothesis-driven context to test for multi-way epistasis underlying common disease etiology, whereas MDR-PDT-based methods are more appropriate in a hypothesis-free genome-wide setting.

Key words: family-based genetic association studies; multi-locus; epistasis; interaction; malaria; innate immunity

Introduction

The GWAS approach to genetic analysis has been successful in the effort to uncover common genetic variants underlying several common diseases, usually requiring large sample sizes, aided by disease-specific consortia and meta-analyses that have identified susceptibility loci with increasingly lower effect sizes, as exemplified by the study of Crohn's Disease [1]. Nonetheless, after consideration of GWAS hits, a large proportion of estimated heritability for most common diseases remains unexplained. The genetic architecture of common diseases likely involves a combination of common and rare causal variants, as well as the interplay of multiple variants and the environment. We focus here on approaches to characterize this complexity arising from multi-locus effects in the family-based study design.

In genetic association studies, epistasis denotes the nonadditive risk-modifying effect of combinations of alleles. Screening for epistasis within a GWAS context is plagued by two major issues: the high number of statistical tests generated with increasing orders of epistasis considered, and the sparsity of certain genotype configurations [2]. Data-mining methods including multifactor dimensionality reduction (MDR) [3–5] overcome the issue of sparsity at the expense of loss of precision by collapsing genotype classes, reviewed in [5]. Others provide algorithms that maximize the number of combinations of markers considered without covering all possibilities exhaustively so that large sets of markers may be analyzed in a reasonable time frame such as random forests [6]. Whereas most screening methods are more appropriate for population-based studies, recently, MDR has been extended to family-based studies of nuclear families of all sizes, by combining MDR with the pedigree disequilibrium test (PDT) in the multifactor dimensionality reduction pedigree disequilibrium test (MDR-PDT) for qualitative traits [4]. This was then extended in a flexible mixed modeling approach allowing for extended pedigrees, adjusting association signals for the presence of linkage [7]. Practical applications of family-based MDR include a study of hypertension among Africans [8].

In parallel to the development of efficient strategies for investigating epistasis in the GWAS context, the growing body of functional molecular data (e.g. the ENCODE project [9]) enables the formulation of highly specific hypotheses implicating several genetic variants. Thus, it is also of current interest to develop model-based statistical methods for the testing of epistatic effects among specific alleles. Although this is straightforward in population-based studies in the logistic or multiple regression frameworks, this is more challenging in family-based studies. The classical transmission disequilibrium test (TDT) [10] identifies distortions in the transmission of alleles from parents to affected offspring, and has been extended to the case of two loci. In one extension, a two-locus TDT first estimates marginal effects of the genotypes at each locus and then the interactive effect of the two loci, using a likelihood ratio test, but is limited to binary traits [11]. In another extension to two independent loci [12], over-transmission of a specific pair of alleles is tested, one from each locus. Herein, we show how this statistic may be generalized to three or more loci, allowing for any number of alleles at each locus for both qualitative and quantitative traits. A transmission is informative if one of the parents is heterozygous at one or more loci [13]. We present a model, the multi-locus transmission disequilibrium test (M-TDT), that does not entail collapsing of genotype classes, and thus allows maximal specificity with respect to patterns of transmitted alleles. To counter the issue of sparsity, we show how to compute empirical *P*-values. In other multi-locus approaches, the TDT has been extended to include consideration of haplotypes [14, 15] or the combination of cis-variants that are in linkage disequilibrium into a single test such as in the family-based association test that linearly combines multiple markers tests, T(LC), [16] or a test that contrasts the linkage disequilibrium between transmitted and non-transmitted genotypes [17]. Other non-TDT-based methods include an analysis of variance-based test for candidate genes [18].

We thus present a generalization of the TDT method to an arbitrary number of loci and an arbitrary number of alleles per locus. Based on simulated data involving three markers, corresponding to a number of different scenarios going from single-locus effects to three-way epistasis, we evaluate the statistical power of M-TDT and MDR-PDT in parallel. In a real data example, we then applied M-TDT using a candidate gene approach to a malaria cohort from Senegal. We identified an epistatic effect among three variants for malaria resistance.

The model

The family-based multi-way gene–gene interaction approach (M-TDT)

We present the M-TDT approach for qualitative traits (for an extension to quantitative traits, see Supplementary Appendix). Consider a set of *N* parent-affected offspring trios. Let L^1, L^2, …, L^K, be *K*-independent multi-allelic loci having l_1, l_2, …, l_K alleles, respectively, and denote the alleles of L^i by $a^i_1, a^i_2, \ldots, a^i_{l_i}$. The total possible combinations of alleles across the *K* loci (*K*-tuples of alleles) is as follows:

$$l = \prod_{k=1}^{K} l_k = l_1 \times l_2 \times \ldots \times l_K$$

Let the two inherited *K*-tuples of a given affected offspring be denoted as $\left(a^1_{u_p}, a^2_{v_p}, \ldots, a^K_{r_p}\right)$ and $\left(a^1_{u_m}, a^2_{v_m}, \ldots, a^K_{r_m}\right)$ with the subscripts '*p*' for paternal and '*m*' for maternal origin, e.g. the offspring's genotype at locus 1 is $a^1_{u_p}/a^1_{v_m}$ such that (u_p, u_m) take on values from {1, 2, …, l_1}. All illustrations will consider three biallelic loci ($K = 3$). We assume that each of the *K* loci is in linkage disequilibrium with corresponding disease loci, and that the set of *K* disease loci are also independent from one another. Even without independence, valid empirical *P*-values may be calculated as presented under Statistical methods.

Transmission counts

Similar to the classic TDT, the M-TDT statistic can be calculated based on a contingency table of transmitted–untransmitted informative *K*-tuple pairings among the *N* trios including fully genotyped parent pairs. Transmissions from parents homozygous at all *K* loci figuring along the diagonal are not informative. If both parents are heterozygous for the same pair of alleles at two or more loci, and if the child shares the same genotypes at these markers, each possible configuration is weighted equally.

Transmission probabilities

Let $S = (a_u^1, a_v^2, \ldots, a_r^K)$ be the transmitted set of alleles from a parent, whereas $S' = (a_{u'}^1, a_{v'}^2, \ldots, a_{r'}^K)$ is the non-transmitted set. Note that u and u' take on values from {1, 2, …, l_1}, v and v' from {1, 2, …, l_2}, etc., until r and r' from {1, 2, …, l_K}. Let $\pi_{ss'}$ be the probability that S is transmitted while S' is not transmitted from a parent to an offspring, such that $S \neq S'$.

Likelihood of the transmission model

Let $m_{ss'}$ be the sample frequency of informative parents with respect to the transmitted set S and untransmitted set S' of alleles. Let $n_{SS'}$ be the sample frequency of parents transmitting S and not S'. The likelihood of the joint transmission model is given by the following:

$$l(\alpha^1, \alpha^2, \ \ldots, \alpha^K) = \prod_{S \neq S'} (\pi_{SS'})^{n_{SS'}} \times (1 - \pi_{SS'})^{(m_{SS'} - n_{SS'})}$$

Where α^i is the vector of transmission intensities (α_j^i, for j in {1, 2, …, l_i}), or the proportion of transmitted alleles, for each allele j at each locus i among heterozygous parents bearing the j allele. Without epistasis, π is a function of allele-specific transmission intensities, and with epistasis, π is a function of transmission intensities for specific combinations of alleles.

The log-likelihood is then given by the following:

$$\log l(\alpha^1, \alpha^2, \ \ldots, \alpha^K) = \sum_{S \neq S'} (n_{SS'}) \times \log(\pi_{SS'}) + \sum_{S \neq S'} (m_{SS'} - n_{SS'}) \times \log(1 - \pi_{SS'})$$

The null hypothesis

Under the null hypothesis of no linkage or no association between the *K* independent markers and disease, $\pi_{SS'} = 1/2$.

The log-likelihood of the null model is then as follows:

$$\log l_0 = -\log(2) \times \sum_{S \neq S'} (m_{SS'})$$

The test statistic is $2 \times (\log l(\alpha^1, \alpha^2, \ldots, \alpha^K) - \log l_0) \sim \chi^2$ with the number of degrees of freedom (DF) depending on the alternative model to test.

Alternative hypotheses

M-TDT considers the following alternative hypotheses, and each is tested with a specific likelihood:

1. Only one of the K markers is linked and associated to its corresponding disease locus

To test for the linkage of a marker locus i alone to its corresponding disease locus, the transmission probabilities are given by the following:

$$\pi_{SS'} = \frac{\alpha_j^i}{\alpha_j^i \ + \ \alpha_{j'}^i}$$

among parents having genotype $a_j^i/a_{j'}^i$ and transmitting allele a_j^i at that locus i, alleles a_j^i and $a_{j'}^i$ belong to $\left\{a_1^i, \ a_2^i, \ \ldots, \ a_{l_i}^i\right\}$.

The χ^2 test statistic for the corresponding likelihood has $l_i - 1$ DF. This model is equivalent to the single-locus extended TDT for multiple alleles [19].

2. A total of p (p = 2, 3, …, K) out of the K marker loci are linked and associated to their corresponding disease loci

There are two main assumptions:

a. assuming additive effects across disease loci, the transmission probability for a given set of p alleles at p marker loci is a function of the product of the marginal risks for the individual alleles:

$$\pi_{'SS} = \frac{\alpha_u^1 \times \alpha_v^2 \times \ldots \times \alpha_t^p}{\left(\alpha_u^1 \times \alpha_v^2 \times \ldots \times \alpha_t^p\right) \ + \ \left(\alpha_{u'}^1 \times \alpha_{v'}^2 \times \ldots \times \alpha_{t'}^p\right)}$$

among parents having genotypes $a_u^1/a_{u'}^1$ at L^1, $a_v^2/a_{v'}^2$ at L^2, etc., until $a_t^p/a_{t'}^p$ at L^p and transmitting the set of alleles $(a_u^1, \ a_v^2, \ \ldots, \ a_t^p)$. The number of DF of the χ^2 statistic for the corresponding likelihood is $(l_1 - 1) + (l_2 - 1) + \ldots + (l_p - 1)$. The M-TDT statistic for the concurrent transmission of specific allele sets across loci is applied to the union of all informative individual locus transmissions, and thus may not be derived from the individual locus M-TDT statistics.

b. If there is epistasis among the disease loci, the risk of transmission of a set of p alleles at the p marker loci is denoted $\gamma_{u,v,\ldots,t}^{1,2,\ldots,p}$ corresponding to the transmission intensity, and is derived from the joint transmission counts for the set of marker alleles:

$$\pi_{SS'} = \frac{\gamma_{u,v,\ldots,t}^{1,2,\ldots,p}}{\gamma_{u,v,\ldots,t}^{1,2,\ldots,p} \ + \ \gamma_{u',v',\ldots,t'}^{1,2,\ldots,p}}$$

The number of DF of the χ^2 statistic for the corresponding likelihood is $l_1 \times l_2 \times \ldots \times l_p - 1$.

Note that (u, u') take on values from {1, 2, …, l_1}, (v, v') from {1, 2, …, l_2}, etc., until (t, t') from {1, 2, …, l_p}.

The number of alternative hypotheses or models, m, is a function of K, the total number of markers:

$$m = K + 2 \times \sum_{p=2,\ldots,K} \binom{K}{p} = 2^{K+1} - K - 2$$

The value of m increases with the number of loci analyzed, given that all markers individually and combinations of 2 to K markers are tested (as additive and epistasis models).

P-values are adjusted for multiple testing using the Benjamini–Hochberg method [20]. Alternative hypotheses are retained if adjusted *P*-values are at or below a false discovery rate (FDR) of 0.05. Two hypotheses that are significant are

contrasted by an empirical method, as presented in Worked example, testing for their likelihood difference, and if not significant, the hypothesis involving the lower number of loci or the additive versus epistasis model is retained. The use of empirical tests permits the comparison of two distinct hypotheses that are not necessarily nested.

MDR-PDT is a statistical algorithm combining the MDR method and the PDT for nuclear families of any size, identifying single-locus or joint effects from among multiple loci [4]. MDR-PDT tests one composite null hypothesis of no association and no interaction such that multi-locus effects can result from true interactions or multiple main effects without interaction. The MDR-PDT algorithm determines the empirical statistical significance for single loci or joint effects, evaluating all possible combinations, for each set of markers going from one up to a certain prespecified total number [4].

Simulations

A simulation study was designed to compare the performance of the M-TDT statistic with MDR-PDT across a range of scenarios. Three biallelic loci were simulated for each data set, and genotypes across affected child–parent trios were generated for 1000 replicate data sets per scenario. Underlying etiologies explaining affection status were set by considering one to three causal loci (see Supplementary Table S1). We use 'etiology' to refer to models introduced into simulated data sets. The M-TDT and MDR-PDT model yielding the lowest *P*-value was identified for each replicate, and power was calculated by counting the number out of 1000 replicates with $\alpha \leq 0.005$ for the model matching the simulated etiology. Similarly, the number out of the 1000 replicates with $\alpha \leq 0.005$ for any model incorrectly matching the simulated etiology were counted toward the family-wise error rate (FWER).

Evaluation of M-TDT statistical power

Global average statistical power and FWER across all scenarios tested were 0.88 and 0.03, respectively, for M-TDT and 0.57 and 0.07, respectively, for MDR-PDT. The overall greater robustness of M-TDT is apparent on inspection of the heat map display of power (Figure 1, Supplementary Table S2). As expected, across scenarios, higher sample size, higher frequency and higher effect size were associated with higher power. Regarding MDR-PDT, in runs with low sample sizes, power decreased with increasing allele frequency for the epistasis scenarios, and in runs with higher sample sizes (600 or 1000), power increased with increasing allele frequency until saturation across effect sizes.

Both methods showed similar power to detect two- and three-locus epistasis (Figure 1). Although three-locus epistasis was the most difficult effect to detect (for M-TDT, power ≥ 0.63 overall; excluding three-locus epistasis, power ≥ 0.72, Supplementary Table S2), under scenarios with frequencies ≥ 0.25, power reached or exceeded 0.85. Although MDR-PDT had marginally increased power to detect epistasis, this came at the cost of much higher FWER, which is most notable at the lower end of the parameter spectrum. As an example, for the scenario with parameters allele frequency = 0.15, sample size = 200, effect size = 2, for three-locus epistasis, power for MDR-PDT was 0.65 whereas this was 0.22 for M-TDT. FWER was substantial (0.26) for MDR-PDT whereas this was null for M-TDT. At the higher end of the spectrum with parameters allele frequency = 0.25, sample size = 600, effect size = 3, power of the two methods were close at 1 for MDR-PDT and 0.97 for M-TDT, and FWER was null for both methods. Also, for all scenarios without epistasis, M-TDT had higher statistical power than MDR-PDT. For the single-locus scenario, for example, at the higher end of the spectrum of parameters (allele frequency = 0.35, sample size = 1000), a power of 0.5 was reached with M-TDT, and only 0.07 with MDR-PDT. M-TDT was strikingly higher powered for the detection of additive models involving two or three single-nucleotide polymorphisms (SNPs) than MDR-PDT even in scenarios at the low end.

Overall, MDR-PDT showed higher FWER than M-TDT. For M-TDT, the FWER exceeded 0.10 only for the additive scenarios, with a maximum of 0.19. The maximum was reached for the three-locus additive scenario for the lowest sample size and lowest allele frequency, and only six total scenarios reached 0.10. On the other hand, for MDR-PDT, 36 scenarios reached a FWER of 0.10, and the maximum FWER attained was 0.44.

Worked example

We performed a candidate gene family-based association study in a data set of 147 malaria-resistant offspring in 73 nuclear families from two village cohorts from Dielmo and Ndiop in Senegal (Table 1 and Supplementary Materials).

Two well-known malaria-resistant variants [HbS mutation (or rs334) and *CR1 R1601G* (or rs17047661)] were selected, and 23 SNPs within the borders of the coding part of *SLC4A1* ± 500 kb were evaluated for selection of one among these (Supplementary Materials).

Family-based analyses were performed to test for association of the markers in *HBB*, *CR1* and *SLC4A1* with malaria resistance. Single-locus M-TDTs (equivalent to the classical TDT) were performed across all *SLC4A1* markers, and the most significant marker was selected. M-TDT was implemented on the three markers. In addition to calculating the asymptotic M-TDT test statistics, we also calculated empirical *P*-values by simulating transmissions from parents to offspring under the null hypothesis of equiprobable transmission across alleles at each locus, generating 100 000 replicates. The empirical *P*-values allow for valid statistical inferences for data sets that include multiple affected offspring or complex families, or where markers are in linkage disequilibrium, or where non-nested models are contrasted. This includes the test of epistasis *stricto sensu* comparing the additive to the epistasis models; this test was calculated as the proportion of times the likelihood difference between the two models in each replicate exceeded the difference in the observed data. Across replicates, for each of the alternative hypotheses, we calculated the empirical *P*-value. These empirical *P*-values were adjusted using Benjamini and Hochberg's method for controlling the FDR [20], and the type I error rate was set at 0.05 for the FDR-adjusted empirical *P*-values.

Associations between HbS, rs17047661 in CR1 and markers in SLC4A1 with malaria resistance in Dielmo and Ndiop

Results from single-SNP M-TDTs, along with the classical family-based association test (FBAT) [13], are presented in Table 2. Of the two fixed variants, HbS reached statistical significance (empirical $P = 0.047$), whereas rs17047661 did not (empirical $P = 0.139$). Two SNPs in *SLC4A1*, rs45497993 (empirical $P = 0.013$) and rs2074106 (empirical $P = 0.007$), exceeded the threshold of 0.05. SNP rs2074106, showing the lowest *P*-value, was selected for the three-locus analysis. The most significant SNP in *SLC4A1*, rs2074106, was also the most significant SNP in

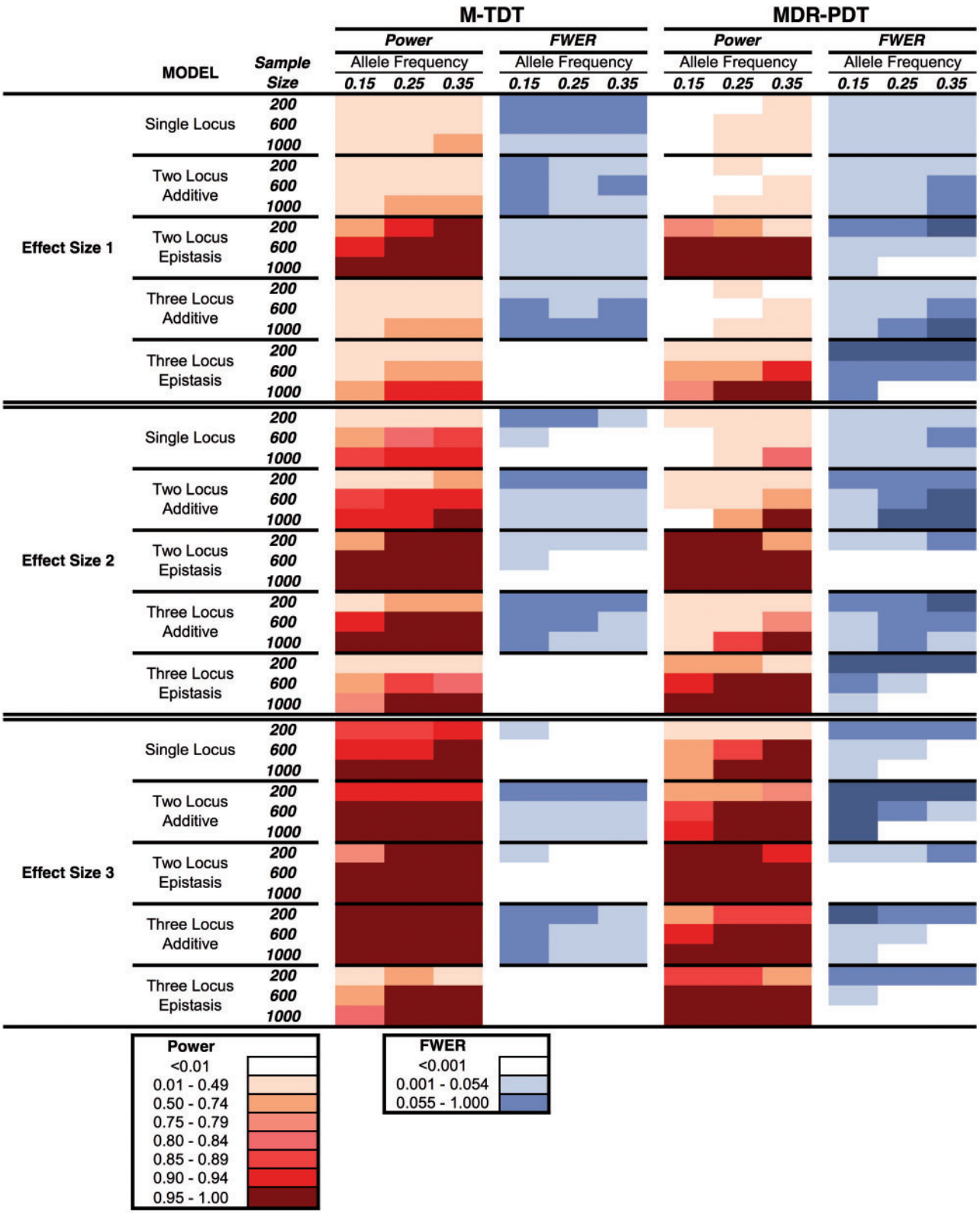

Figure 1. The figure illustrates power and FWER for M-TDT and the MDR-PDT for scenarios varying the risk allele frequency, sample size and effect size, under one-, two- or three-locus effect etiologies. Effect sizes are designated 1, 2 or 3, and represent different odds ratios depending on the etiology—see Supplementary Table S1 for details. Power (red) and FWER (blue) are based on 1000 simulated data sets per scenario using a heatmap with greater intensity of color representing higher power or FWER. Exact power and FWER values are presented in Supplementary Table S2.

the FBAT analysis (Table 2). Also, empirical *P*-values obtained using FBAT were similar to empirical *P*-values obtained using M-TDT.

M-TDT analysis demonstrated that the three-locus epistasis model yielded the highest likelihood (M-TDT-statistic = 39.4; nominal $P = 1.6 \times 10^{-6}$) (Tables 3 and 4), exceeding the three-locus additive model (M-TDT-statistic = 20.6; nominal $P = 0.0001$). The principal contribution to the epistasis was the second-order interaction involving rs2074106 in *SLC4A1* and rs17047661 in *CR1* (M-TDT-statistic = 27.7; nominal $P = 4.2 \times 10^{-6}$). However, the three-locus epistasis model gave a statistically significantly better fit to the data than the three-locus additive model (empirical $P = 0.002$) and a nominally significantly better fit to the data than the best two-locus epistasis models, *SLC4A1* and *CR1* (empirical $P = 0.04$).

Further inspection showed that the effects of every one of the three alleles depended on the identity of the alleles at the other loci, thus constituting three-way epistasis (Supplementary Materials).

Table 1. Family structure

Number of individuals in the analyzed sample	276
Number of parents	135
Number of offspring	147
Number of offspring also having the parent status	6
Number of independent families	11
Number of nuclear families	73
Mean number of offspring per nuclear family	2
Number of nuclear families with 1 offspring	28
Number of nuclear families with 2 offspring	26
Number of nuclear families with 3 offspring	10
Number of nuclear families with 4 offspring	8
Number of nuclear families with 5 offspring	1

Note. The 276 analyzed individuals are from Dielmo and Ndiop, Senegal, followed for clinical malaria attacks from 1990 to 2008. They were composed of resistant offspring and their parents.

Discussion

Approaches for the testing of epistasis in the family-based design are critical in the characterization of the genetic architecture of complex diseases. We presented the M-TDT statistic, extending from both the biallelic TDT [10] and the two-locus TDT [12] to a multi-locus multi-allelic TDT, for binary and quantitative traits in family-based studies for choosing the best

Table 2. Single-locus results using two TDT approaches (M-TDT and FBAT).

Gene symbol	Chr	rs Number	Position (GRCh38)	M/m	MAF	N	T	NT	M-TDT		FBAT	
									Nominal *P*-value	Empirical *P*-value	Nominal *P*-value	Empirical *P*-value
Pre-selected SNPs												
CR1	1	rs17047661	207609544	A/G	0.21	108	57.8	40.1	0.070	0.139	0.150	0.14
HBB	11	rs334	5227002	T/G	0.06	50	39.3	19.1	0.008	0.047	0.050	0.04
Screened SNP												
5′ UTR	17	rs9910055	44205669	C/T	0.40	15	4.3	11.5	0.070	0.067	0.120	0.12
5′ UTR	17	rs2071167	44210151	A/G	0.36	5	1.6	4.7	0.220	0.259	0.320	0.31
5′ UTR	17	rs9901595	44228331	A/G	0.17	3	2.6	0.3	0.150	0.484	0.300	0.50
5′ UTR	17	rs9906669	44228338	A/G	0.20	3	0.4	3.1	0.120	0.535	0.300	0.50
SLC4A1	17	rs8066822	44247988	C/T	0.33	7	5.6	2.0	0.180	0.291	0.280	0.26
SLC4A1	17	rs2857079	44248840	A/T	0.05	1	0.0	0.8	0.300	1.000	0.320	[a]
SLC4A1	17	rs1465204	44250109	C/T	0.20	7	2.1	5.2	0.250	0.347	0.350	0.35
SLC4A1	17	rs2072081	44250125	A/C	0.35	7	5.6	2.0	0.180	0.291	0.280	0.26
SLC4A1	17	rs45497993	44250506	T/C	0.33	152	52.2	89.2	0.002	0.013	0.016	0.02
SLC4A1	17	rs2857078	44252803	G/T	0.39	135	60.1	66.3	0.580	0.644	0.670	0.70
SLC4A1	17	rs45530735	44256644	A/G	0.25	106	42.6	55.2	0.200	0.302	0.320	0.34
SLC4A1	17	rs2074108	44258781	A/G	0.17	7	5.3	2.1	0.240	0.394	0.350	0.38
SLC4A1	17	rs2857082	44259723	A/G	0.30	15	10.2	5.6	0.240	0.262	0.310	0.30
SLC4A1	17	rs2074107	44260608	C/T	0.25	3	1.7	1.6	0.980	1.000	0.980	1.00
SLC4A1	17	rs5036	44261577	A/G	0.08	5	4.8	0.3	0.030	0.067	0.100	0.06
SLC4A1	17	rs2074106	44261935	A/C	0.16	84	28.8	60.8	6.3×10^{-4}	0.007	0.010	0.01
SLC4A1	17	rs9916116	44268611	A/G	0.46	11	8.7	3.5	0.130	0.164	0.210	0.20
3′ UTR	17	rs7222501	44277289	C/T	0.45	7	4.3	2.8	0.580	0.612	0.620	0.62
3′ UTR	17	rs6503366	44280405	A/G	0.26	10	3.6	6.6	0.330	0.455	0.410	0.46
3′ UTR	17	rs2879165	44283992	A/G	0.33	10	6.6	3.6	0.330	0.439	0.410	0.44
3′ UTR	17	rs10852960	44302585	A/G	0.18	10	4.0	5.3	0.660	0.702	0.690	0.72
3′ UTR	17	rs708386	44315254	A/G	0.22	10	4.0	5.3	0.660	0.719	0.690	0.72
3′ UTR	17	rs2074104	44316126	A/G	0.39	11	6.5	3.8	0.390	0.432	0.440	0.46

Note. Chr: Chromosome number; M/m: major/minor alleles; MAF: minor (variant) allele frequency; N: number of informative transmissions, i.e. transmission from parents heterozygous at the SNP; T and NT: transmission and non-transmission frequencies of the variant allele. Nominal *P*-values are based on asymptotic distributions of test statistics, and empirical *P*-values are based on 100 000 simulations.

(a): Insufficient number of informative families. Given that the study population is composed of interconnecting nuclear families with multiple offspring, single SNP analyses were also implemented using FBAT v2.0.3, which incorporates a combination of statistics accommodating multiplex families, unaffected offspring and missing genotype data [13], and we reported both the nominal *P*-values and permutation *P*-values. We then performed M-TDT single-SNP analysis providing empirical *P*-values based on simulations to accommodate for multiplex families as in FBAT.

Table 3. Multi-locus transmission counts

		Non-transmitted sets							
	{*HBB*, *SLC4A1*, *CR1*} set of alleles	{T,A,A}	{T,A,G}	{T,C,A}	{T,C,G}	{G,A,A}	{G,A,G}	{G,C,A}	{G,C,G}
Transmitted sets	{T,A,A}	116.0	28.4	21.0	17.7	5.8	0.3	5.1	0.2
	{T,A,G}	20.5	3.6	2.7	2.8		1.3		
	{T,C,A}	19.0	2.1	1.3		5.6			
	{T,C,G}	0.7	0.1		0.6				
	{G,A,A}	9.5	0.2	6.5	2.7				
	{G,A,G}	8.0	7.0	1.8					
	{G,C,A}	0.4	0.5		2.1				
	{G,C,G}	0.5							

Note. Minor alleles are underlined. Cells along the diagonal are uninformative and do not contribute to the test statistic.

Table 4. M-TDT results

Association with malaria-resistant phenotype									
Gene	**Model**	**Alleles or sets over$^{(+)}$/under$^{(-)}$ transmitted**	**N**	**LL1**	**M-TDT statistic**	**DF**	**M-TDT nominalP**	**M-TDT empiricalP**	**FDR-adjusted empiricalP**
HBB (rs334)	Single	$G^{+/}T^{-}$	49	−115.8	7.7	1	0.006	0.030	0.030
SLC4A1 (rs2074106)	Single	$A^{+/}C^{-}$	84	−113.8	11.7	1	6.2×10^{-4}	0.007	0.010
CR1 (rs17047661)	Single	$A^{+/}G^{-}$	95	−117.4	4.6	1	0.030	0.070	0.070
HBB, *SLC4A1*	Additive	$\{G,A\}^{+}$,$\{T,C\}^{-}$	115	−111.1	17.2	2	1.9×10^{-4}	0.005	0.010
HBB, *CR1*	Additive	$\{G,G\}^{+}$,$\{T,G\}^{-}$	124	−113.0	13.3	2	0.001	0.010	0.020
SLC4A1, *CR1*	Additive	$\{A,A\}^{+}$,$\{C,G\}^{-}$	148	−112.6	14.1	2	8.7×10^{-4}	0.009	0.010
HBB, *SLC4A1*	Epistasis		115	−112.2	15.0	3	0.002	0.010	0.020
HBB, *CR1*	Epistasis		124	−110.0	19.3	3	2.4×10^{-4}	0.002	0.008
SLC4A1, *CR1*	Epistasis		148	−105.8	27.7	3	4.2×10^{-6}	8×10^{-5}	4.4×10^{-4}
HBB, *SLC4A1*, *CR1*	Additive	$\{G,A,G\}^{+}$,$\{T,C,G\}^{-}$	173	−109.4	20.6	3	1.3×10^{-4}	0.004	0.010
HBB, *SLC4A1*, *CR1*	Epistasis		173	−100.0	39.4	7	1.6×10^{-6}	3×10^{-5}	3.3×10^{-4}

Note. N: number of informative transmissions, i.e. transmissions from parents heterozygous at least at one SNP; LL1: log-likelihood computed under the alternative model; M-TDT-statistic: $2 \times (LL1 - LL0)$ is the log likelihood-ratio-based test statistic where LL0: −119.7 is the log-likelihood computed under the null model; M-TDT nominalP: asymptotic *P*-value; M-TDT empiricalP: empirical *P*-value; FDR-adjusted empiricalP: empirical *P*-value corrected for multiple testing by the FDR method.

model, including epistasis. Sample sizes up to 1000 trios, and up to 3 causal biallelic loci with risk allele frequencies ranging from 0.15 to 0.35 using simulations were used to mimic what might be encountered in a research setting. Single-locus, two-locus and three-locus effects (both additive and with epistasis) were considered as alternative hypotheses using M-TDT. M-TDT was overall more powerful and had lower FWER than another family-based multi-locus test, MDR-PDT. In addition, through an empirical method, M-TDT distinguishes between additive and epistatic effects for the same loci and enables the identification of the specific allele sets driving associations. M-TDT was used to explore a candidate gene hypothesis: two variants known to impact on malaria resistance were considered (the sickle cell trait heterozygote genotype, HbS, and a non-synonymous variant in *CR1*, rs17047661 or R1601G), and *SLC4A1* was initially screened for variants based on single-locus results, leading to the identification of three SNPs involved in three-way epistasis including rs2074106 in *SLC4A1*.

The results of the simulation study comparing M-TDT with MDR-PDT highlighted some of the advantages of M-TDT: at sample sizes of 600 or more trios, high power of detection of the correct model was achieved across scenarios implicating one to three loci for the two higher effect sizes. Although three-locus epistasis was the most difficult effect to detect, under scenarios with frequencies > 0.25, power was > 0.85 and FWER was consistently low or undetectable (maximum FWER was 0.14). These data showed that M-TDT is particularly effective at identifying the causally implicated markers and both additive and epistatic effects. By contrast, using MDR-PDT at sample sizes of $\geq$600 trios and at the two higher effect sizes, power was null for some scenarios (e.g. two locus additive scenarios) and a FWER of up to 0.42 was recorded. This is not surprising, given that MDR-PDT was designed to identify general multi-locus effects under a composite null hypothesis of no association and no epistasis. Thus, although higher power was achieved for two- and three-locus epistasis using MDR-PDT, this was at the expense of higher FWER, and at the expense of lower power to detect multi-locus additive effects.

In spite of high power to select the correct model from among single-locus to all possible multi-locus combinations for a limited number of genetic markers, M-TDT has not been optimized for a genome-wide screen. In the aim of detecting epistasis in a genome-wide association study setting, we recommend MDR-based methods in a first screening phase. MDR-PDT scans a large number of markers quickly, evaluating all possible combinations from one up to a user-defined maximum number in joint effects. The high statistical power to detect epistasis using MDR-PDT would allow for markers involved in true epistasis to be detected in this phase, even though a non-negligible proportion of false positives would also be detected. Second, M-TDT would be valuable to fine-tune alternative hypotheses, testing for epistasis versus additive models including all combinations of SNPs, and to identify specific risk allele combinations. MDR-PDT identifies the markers involved given the algorithm's genotype-collapsing procedure, but not the specific risk alleles.

These features of M-TDT facilitate biological interpretation and provide a bridge toward functional experiments. The authors of MDR-PDT acknowledged the need for follow-up modeling after an MDR-PDT screening phase. A previous two-step framework, for both population- and family-based designs, was proposed with logistic regression following MDR. The authors clearly showed that this strategy was only appropriate for two-locus but not higher-order epistatic effects [14]. We showed that M-TDT performed well in the detection of three-locus epistasis, overcoming a key limitation of the previously employed framework and thus is the most appropriate choice for step 2 in a two-step framework.

A drawback of M-TDT is the rising number of alternative hypotheses as the number of loci increases, which substantially increases the corrected threshold of significance. M-TDT requires independence among markers, but we also propose a method to determine empirical *P*-values to bypass this restriction. Also, M-TDT does not accommodate missing genotype data. However, an increasing number of markers to be tested simultaneously increases the number of informative transmissions. M-TDT does not infer parental genotypes when these are missing using genotype data from affected and unaffected siblings as other TDT-derived test statistics such as sib-TDT [15].

The real data example, an application of M-TDT to malaria resistance, showed how three-way epistasis might be detected. Based on the single-locus results alone, the two SNPs to pass the FDR-corrected threshold of significance for M-TDT were HbS and rs2074106 (*SLC4A1*). We identified the common allele at rs2074106 (*SLC4A1*) (empirical $P = 0.007$) as associated with malaria resistance based on a single-SNP test (the biological interpretation of these results is discussed in the Supplementary Materials). In comparing results from the single-locus tests to multi-locus tests, two orders of magnitude on the log scale were gained, thanks to the three-way epistasis model (empirical $P = 3 \times 10^{-5}$). It is likely that many such combinations of alleles implicating several independent markers impact on risk of other diseases. The M-TDT approach presented here will facilitate their detection in family-based studies.

Key Points

- Although genome-wide association studies have met with success in the detection of individual marker effects impacting on common diseases, few combinations of markers showing epistatic effects—but no individual effects—have been identified.
- We present a family-based statistic (M-TDT) testing for association between a qualitative or quantitative phenotype and a set of independent multi-allelic markers. All alternative hypotheses of models including single marker effects and all combinations of markers are tested and empirical p-values allow for the selection of the best model.
- Power and false-discovery rates (FDR) calculated for M-TDT and the previously published method, MDR-PDT using simulated scenarios containing effects including one to three associated markers, showed higher power and lower FDR for M-TDT overall.
- Application of the method to malaria data allowed for the identification of interaction between the sickle cell trait, HbS, rs17047661 in *CR1* (*CR1-R1601G*, Complement Receptor 1) and rs2074106 in *SLC4A1* (coding Band 3).

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors thank Stevenn Volant and Bertrand Neron (Center for Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France) for advice on programming. They are grateful to the villagers of Dielmo and Ndiop for their participation and continued collaboration in this project. They also thank the administrative authority of Institut Pasteur of Dakar for their continuous support. The authors particularly thank the field workers for their sustained contribution to the project and for generating and maintaining the malaria databases.

Funding

This work was supported by the Institut Pasteur, the Institut Pasteur de Dakar and the Institut de Recherche pour le Développement. This study has received funding from the French Government's Investissement d'Avenir program, Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (grant no. ANR-10-LABX-62-IBEID) and 'ANR-11-BSV1-027-01'.

References

1. Franke A, McGovern DPB, Barrett JC, *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010;**42**:1118–25.
2. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–404.
3. Calle ML, Urrea V, Malats N, *et al.* mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. *Bioinforma Oxf Engl* 2010;**26**:2198–9.
4. Martin ER, Ritchie MD, Hahn L, *et al.* A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Gene Epidemiol* 2006;**30**:111–23.
5. Steen KV. Travelling the world of gene-gene interactions. *Brief Bioinform* 2012;**13**:1–19.
6. Kim Y, Wojciechowski R, Sung H, *et al.* Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proc* 2009;**3**(Suppl 7):S64.
7. De Lobel L, Thijs L, Kouznetsova T, *et al.* A family-based association test to detect gene–gene interactions in the presence of linkage. *Eur J Hum Genet* 2012;**20**:973–80.
8. Kimura L, Angeli CB, Auricchio MTBM, *et al.* Multilocus family-based association analysis of seven candidate polymorphisms with essential hypertension in an african-derived semi-isolated brazilian population. *Int J Hypertens* 2012; **2012**:859219.
9. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;**306**:636–40.
10. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;**52**:506–16.
11. Kotti S, Bourgey M, Clerget-Darpoux F. Power of the 2-locus TDT for testing the interaction of two susceptibility genes. *BMC Proc* 2007;**1**(Suppl 1):S65.

12. Morris A, Whittaker J. Generalization of the extended transmission disequilibrium test to two unlinked disease loci. *Genet Epidemiol* 1999; **17**(Suppl 1):S661–6.
13. Ma L, Han S, Yang J, *et al*. Multi-locus test conditional on confirmed effects leads to increased power in genome-wide association studies. *PloS ONE* 2010;**5**:e15006.
14. Abad-Grau MM, Medina-Medina N, Moral S, *et al*. Increasing power by using haplotype similarity in a multimarker transmission/disequilibrium test. *J Bioinform Comput Biol* 2013;**11**:1250014.
15. Abad-Grau MM, Medina-Medina N, Montes-Soldado R, *et al*. Sample reproducibility of genetic association using different multimarker TDTs in genome-wide association studies: characterization and a new approach. *PloS ONE* 2012;**7**:e29613.
16. Xu X, Rakovski C, Xu X, *et al*. An efficient family-based association test using multiple markers. *Genet Epidemiol* 2006;**30**:620–6.
17. Yu Z, Wang S. Contrasting linkage disequilibrium as a multilocus family-based association test. *Genet Epidemiol* 2011;**35**:487–98.
18. Rakovski CS, Xu X, Lazarus R, *et al*. A new multimarker test for family-based association studies. *Genet Epidemiol* 2007;**31**:9–17.
19. Sham PC, Curtis D. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* 1995;**59**:323–36.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995;**57**:289–300.