

# Improved Use of SNP Information to Detect the Role of Genes

A.-S. Jannot,<sup>1,2\*</sup> L. Essioux,<sup>2</sup> M.G. Reese,<sup>2</sup> and F. Clerget-Darpoux<sup>1</sup>

<sup>1</sup>INSERM U535, Le Kremlin-Bicêtre, France

<sup>2</sup>ValiGen, Arcueil, France

A topical question in genetic association studies is the optimal use of the information provided by genotyped single-nucleotide polymorphisms (SNPs) in order to detect the role of a candidate gene in a multifactorial disease. We propose a strategy called “combination test” that tests the association between a quantitative trait and all possible phased combinations of various numbers of SNPs. We compare this strategy to two alternative strategies: the association test that considers each SNP separately, and a multilocus genotype-based test that considers the phased combination of all SNPs together. To compare these three tests, a quantitative trait was simulated under different models of correspondence between phenotype and genotype, including the extreme case when two SNPs interact with no marginal effects of each SNP. The genotypes were taken from a sample of 290 independent individuals genotyped for three genes with various number of SNPs (from 5–8 SNPs). The results show that the “combination test” is the only one able to detect the association when the two SNPs involved in disease susceptibility interact with no marginal effects. Interestingly, even in the case of a single etiological SNP, the “combination test” performed well. We apply the three tests to Genetic Analysis Workshop 12 (Almasy et al. [2001] *Genet. Epidemiol.* 21:332–338) simulated data, and show that although there was no interactions between the etiological SNPs, the “combination test” was preferable to the two other compared methods to detect the role of the candidate gene. *Genet. Epidemiol.* 25:158–167, 2003. © 2003 Wiley-Liss, Inc.

**Key words:** association studies; intragenic markers; interaction

\*Correspondence to: Anne-Sophie Jannot, Unité INSERM 535—Génétique épidémiologique et structure des populations humaines, Hôpital Paul Brousse—Bâtiment Leriche (1er étage), Secteur Jaune, Porte 18, B.P. 1000, 94817 Villejuif Cedex, France.

E-mail: jannot@vjf.inserm.fr

Received for publication 26 September 2002; Revision accepted 3 March 2003

Published online in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/gepi.10256

## INTRODUCTION

The candidate gene approach is increasingly used for the study of multifactorial diseases. Until recently, only very few polymorphisms were available on a candidate gene region. The common hypothesis was that there was only one etiological polymorphism per gene, i.e., directly involved in the disease process. Genomic efforts provide larger sets of markers, most of them single-nucleotide polymorphisms (SNPs). Besides, different results in human and other species show that not only nonsynonymous SNPs (giving an amino-acid change) in coding region can be etiological [Di Paola et al., 2002; Pagani et al., 2002], which impacts the number of markers to consider when planning a candidate gene study. The challenge is now to find strategies to optimally use this information in order to detect the role of a candidate gene.

No information is available about the underlying complexity of the relationship between genotypes and associated phenotype. Polymorphisms may interact, as was shown for alcohol dehydrogenase expression in *Drosophila melanogaster* [Stam and Laurie, 1996]. In the extreme, interaction can be so strong that there are no marginal effects of each polymorphism taken separately. Such situations are rarely found in the literature, because interaction is usually tested once main effects have been discovered, but this does not prove that this situation is not common. For example, genes interacting with no main effects were discovered for susceptibility to cancer in mice [Fijneman et al., 1996; van Wezel et al., 1996], using a study design that could not be applied to humans. This situation can explain why some nonspurious associations are not replicated in a second population: when testing one given polymorphism interacting with another one, their frequencies in the second population may lead to

no marginal effect of each of these polymorphisms, while such was not the case for the original population where the association was detected.

Several association test strategies were proposed in the literature, with some considering each SNP separately and others considering the pair of haplotypes formed by all the typed SNPs. Under the assumption that a single polymorphism contributes to the variation of the trait and that this polymorphism is in very strong disequilibrium with one of the typed SNPs, testing each SNP separately is a more powerful strategy than testing the association with the haplotype formed by the whole set of alleles at the genotyped SNPs [Bader, 2001; Long and Langley, 1999]. This result is not valid in some other situations when, for example, there is allelic heterogeneity [Longmate, 2001], or when the marginal effect on each typed SNP is weak.

Here, we compare the power of three methods to detect the role of a candidate gene under different situations, notably when two SNPs interact with no marginal effect of each SNP. The aim is to find a powerful test valid for the largest number of plausible situations including this last situation.

## METHODS AND MODELS

### DESCRIPTION OF THREE COMPARED TESTS

The three compared tests are genotype-based; we do not consider allelic or haplotypic tests. Haplotype-based tests rely on the assumption that the two haplotypes in an individual do not interact. Studies of some diseases showed that this hypothesis is not always true. For example, the individuals *DR3/DR4* are at higher risk than the two homozygotes *DR3/DR3* and *DR4/DR4* in type I diabetes [Thomson, 1983]. For such situations, allelic or haplotypic tests are not the most powerful.

#### TEST 1: SINGLE SNP TEST

This method tests the association between the trait and the genotype of each SNP locus, using an analysis of variance (F-test). The number of F-tests performed is equal to the number of selected SNPs. For each test, a *P*-value is obtained (we take the *P*-value because the different performed tests have different degrees of freedom, because some loci have 2 genotypes, and other 3 genotypes). The null hypothesis is rejected if at least one test is significant, so the statistic considered is the

minimum of the *P*-values obtained for all the performed tests. To account for multiple testing and correlation between the markers, the critical value of the statistic for a given type-I error rate is estimated via Monte Carlo simulations.

#### TEST 2: MULTILOCUS GENOTYPE TEST

This method tests the association between the trait and the "multilocus genotype." For a given candidate gene, the multilocus genotype is the pair of multilocus haplotypes observed on the candidate gene. We consider that we are able to reconstruct haplotypes with no uncertainty using familial information. Therefore, the phase of the pair of haplotypes is assumed to be known. Each haplotype of the pair is formed by the alleles of all the typed SNPs on the same chromosome. Let *J* be the number of observed genotypes, and *N* the number of individuals in the sample. The test of association performed is an F-test with (*J*-1, *N*-*J*) as degrees of freedom. This F-test establishes whether the distribution of the trait is homogeneous among the different genotypic groups. The statistic considered is the *P*-value of the performed F-test.

#### TEST 3: COMBINATION SNP TEST

We propose here as an alternative an association test consisting of several dependent F-tests testing the association between the trait and different genotypes defined as below. For each F-test, a subset of SNPs is chosen, that means one or several typed SNPs. For a candidate gene with *n* typed SNPs, all possible subsets from one SNP (*n* possible subsets) to *n* SNPs (one possible subset) are considered. For a given subset, the genotype considered for this F-test is the pair of haplotypes formed by the allele of all SNPs from the chosen subset. As for test 2, we consider that the phase of the pair of haplotypes is known. Such an F-test is performed for all possible subsets of SNPs: *n* typed SNPs require  $2^n - 1$  F-tests. The performed F-tests consider genotypes with variable numbers of SNPs. Therefore, they have different degrees of freedom. As for test 1, the statistic considered is the minimum of the *P*-value, and the critical value of the statistic for a given type-I error rate is estimated via Monte Carlo simulations.

### SIMULATION STUDY

**Genotypic data used.** To compare the three methods, we developed a simulation study based

on observed SNPs of three genes from a population of 290 independent individuals for which parental phase was known. This enables us to base our study on real patterns of linkage disequilibrium within those genes, as linkage disequilibrium within a gene cannot be modeled easily [Tiret et al., 2002]. For each individual, familial information was available. Therefore, haplotypic phases could be

inferred. For each individual, the most likely haplotypic configuration was retained, considering no recombination in the candidate gene.

Three genes were genotyped in exonic and perixonic regions: *CETP* (6 SNPs), *LICT* (7 SNPs), and *GPC6* (9 SNPs). More details about the frequencies of the different haplotypes and distances between SNPs can be found in Table I.

**TABLE I. Description of haplotypes of three genes (allele 1 is represented by a dot to simplify notation) and distances between SNPs (in bp)**

<i>CETP</i> (only haplotypes with frequency greater than 2%)									
SNP number	1	2	3	4	5	Frequency (%)			
Distance (bp)	979	68	9	19,809					
Haplotypes	.	.	.	.	.	36.55			
	.	2	2	.	.	22.24			
	.	2	2	.	2	19.14			
	.	.	.	.	2	6.38			
	.	2	.	.	2	5.34			
	2	.	.	.	.	2.93			
	2	.	.	2	.	2.24			
<i>LICT</i> (only haplotypes with frequency greater than 2%)									
SNP number	1	2	3	4	5	6	7	Frequency (%)	
Distance (bp)	32	110,287	748	3,216	53	28			
Haplotypes	.	.	.	.	.	.	.	21.90	
	.	.	.	2	.	.	.	13.28	
	.	.	.	.	.	2	2	11.72	
	.	.	.	2	.	2	2	10.52	
	2	.	.	2	.	.	.	7.07	
	2	.	.	.	.	.	.	6.72	
	.	.	.	.	.	.	2	5.86	
	2	.	.	.	.	2	2	5.00	
	.	.	.	2	2	2	2	3.45	
	.	.	.	.	2	2	2	2.41	
	2	.	.	.	2	2	2	2.24	
<i>GPC6</i> (only haplotypes with frequency greater than 2%)									
SNP number	1	2	3	4	5	6	7	8	Frequency (%)
Distance (bp)	315,438	5,520	21,670	241	13,298	253,821	528,346		
Haplotypes	.	.	.	.	.	.	.	.	6.21
	.	.	2	.	.	.	2	.	5.86
	.	.	.	.	.	.	2	.	5.69
	.	.	.	.	.	2	.	.	5.69
	.	.	2	.	.	.	.	.	3.97
	.	.	.	.	.	2	2	.	3.79
	.	2	.	.	.	2	.	.	3.28
	.	2	.	.	.	.	2	.	3.28
	2	.	.	.	.	2	.	.	3.28
	.	.	2	.	2	.	2	.	2.93
	.	2	.	.	.	.	.	.	2.93
	.	.	2	.	2	.	.	.	2.76
	2	.	.	.	.	.	2	.	2.76
	2	2	.	.	.	.	.	.	2.41
	.	2	.	.	.	2	2	.	2.24
	2	.	2	.	.	.	2	.	2.24
	.	2	2	.	.	.	2	.	2.07

A first step of selection was necessary for *CETP* and *GPC6*, because two SNPs show complete association, so only 5 SNPs and 8 SNPs, respectively, were considered for our study.

**Models of correspondence between phenotype and genotype (see Table II).** Phenotypic trait distributions were simulated under four types of situations on the sample of 290 individuals, keeping their SNP typing information. For the four situations, the genotyped SNPs contain the etiological ones. For the  $j$ th individual in genotypic group  $i$ , the phenotypic trait  $\mu_{ij}$  is taken from a normal distribution, with variance equal to 1 and with mean  $\mu_i$  (the expected mean of genotypic group  $i$ ). Let  $\mu$  be the mean of the total sample,  $g$  the number of genotypic groups for the considered model, and  $n_i$  the number of individuals in the genotypic group  $i$ . The percentage of variance explained by the model can be expressed as:

$$\omega^2 = \frac{\sum_{i=1}^g n_i (\mu_i - \mu)^2}{\sum_{i=1}^g \sum_{j=1}^{n_i} (\mu_{ij} - \mu)^2}$$

The first model (M1) considers that a single typed SNP contributes to the variation of the trait. Three different phenotypic trait distributions are observed in the population. If  $A$  and  $a$  are the alleles of the etiological SNP, then the phenotypic

trait of the individuals with the genotype  $aa$  follows a normal distribution with mean proportional to  $3\omega^2$  and variance 1,  $Aa$  with mean proportional to  $\omega^2$  and variance 1, and  $AA$  with mean 0 and variance 1. In this model, the effect of each allele is not additive. In many cases, this additive hypothesis does not hold. For example, the risk conferred by the different alleles of *APOE* are not additive for Alzheimer's disease [Bickeboller et al., 1997].

The second model (M2) considers that two typed SNPs contribute independently to the variation of the trait. If  $A$  and  $a$  are the alleles of one of the SNPs involved and  $B$  and  $b$  the alleles of the other SNP, individuals with genotype  $aabb$  have a phenotypic trait normal distribution with mean 0 and variance 1. Then each allele  $A$  increases the mean by  $k\omega^2$ , and each allele  $B$  by  $2k\omega^2$ ,  $k$  being a constant of proportionality.

The third model (M3) considers that two typed SNPs contribute to the variation of the trait, one of the SNPs masking the expression of the other SNP. For example, one can assume a polymorphism (allele  $A/a$ ) in the promoter region with a regulatory effect on expression, and another SNP (allele  $B/b$ ) located in an exon giving an amino-acid change that stops the production of the protein. Each haplotype acts with an additive effect, with haplotype  $ab$  having a production with mean  $k\omega^2$ ,  $k$  being a constant of proportionality. Each allele  $A$  increases the mean by  $k\omega^2$ . For each haplotype with  $B$ , the production has mean 0.

The fourth model (M4) corresponds to two typed SNPs interacting with no marginal effect of each SNP, meaning that the trait distribution is the same for the different genotypes at a given SNP. This situation can be found, for example, when the two SNPs give amino acids that are part of an active site, giving a strong interaction of the two SNPs. The parameters (means of the different genotypes) corresponding to models with no marginal effects depend not only on the effect size but also on the frequency of the two etiological alleles and their linkage disequilibrium (see Appendix). The number of possible models filling the constraints given in the Appendix is infinite. Table II shows the models we choose for the three genes.

For model M1, the selected SNPs are the second in the *CETP* gene, the seventh in the *LICT* gene, and the third in the *GPC6* gene. They were randomly selected among frequent markers. For models M2, M3, and M4, the selected SNPs are the second and fifth in the *CETP* gene, the first and

TABLE II. Trait mean for model M2, M3, and M4<sup>a</sup>

Model	Genotype for SNP-A	Genotype for SNP-B		
		<i>bb</i>	<i>bB</i>	<i>BB</i>
2	<i>aa</i>	0	2x	4x
	<i>Aa</i>	x	3x	5x
	<i>AA</i>	2x	4x	6x
3	<i>aa</i>	2x	x	0
	<i>Aa</i>	3x	x/2x <sup>b</sup>	0
	<i>AA</i>	4x	2x	0
4 ( <i>CETP</i> )	<i>aa</i>	-0.6x	2.5x	5.5x
	<i>Aa</i>	x	0	0
	<i>AA</i>	1.4x	0	0
4 ( <i>LICT</i> )	<i>aa</i>	-0.885x	x	1.12x
	<i>Aa</i>	1.05x	0	0
	<i>AA</i>	1.18x	0	0
4 ( <i>GPC6</i> )	<i>aa</i>	-1.77x	x	1.41x
	<i>Aa</i>	1.21x	0	0
	<i>AA</i>	1.42x	0	0

<sup>a</sup> $a/A$  and  $b/B$  are two SNPs involved. For each model and genotype, trait is normally distributed with variance 1 and mean depending on  $x=k\omega^2$ ,  $k$  being a constant and  $\omega^2$  the proportion of variance explained by model. Choice for M4 parameter values are justified in Appendix.

<sup>b</sup>Two different phased genotypes:  $x$  for  $ab/AB$ , and  $2x$  for  $Ab/aB$ .

sixth in the *LICT* gene, and the third and seventh in the *GPC6* gene, respectively. They were selected because they did not show strong linkage disequilibrium between each other. Indeed, if the two selected SNPs were strongly linked, the situation for M2 and M3 would have been similar to that for M1.

**Estimation of critical value for a global type I error of 5%.** Tests 1 and 3 perform a large number of tests (depending on the number of typed SNPs) that are not independent. As the structure of correlation between these test values is unknown, we used a Monte Carlo simulation procedure to estimate the critical value corresponding to a global type-I error rate of 5%. We first simulated the trait under the null hypothesis (100 replicates). For each replicate, the critical value was assessed using a randomization procedure (1,000 permuted sets) by the lowest 50th value obtained. Estimation of the critical value was then taken as the mean of the critical values over the 100 simulated replicates. Using this strategy, the critical value led to a global type-I error rate between 4.9–5.1%.

**Computation of power of three tests under four models considered.** We assessed the power for each test at the same level of 5% by the proportion of the simulated-set statistics that matched or exceeded the estimated critical value. This strategy is less demanding in computing resources than a true permutation test (for which the threshold would be recomputed for each simulated set), but it assumes that the distribution of the *P*-values is not dependent on the model underlying the causality. The validity of this hypothesis was checked by simulating alternative hypotheses, leading to a mixture of normal distributions far from the normal distribution. The results show that even if the simulated trait distribution is very far from a normal distribution, the confidence interval of the threshold found is about the same as for the normal distribution. For each candidate gene and each situation, the power was computed for an effect of the gene varying from 0 to 10–15% of the variance explained by the model (500 simulated replicates).

## RESULTS

The power is shown in Figures 1–3 as a function of  $\omega^2$  for the three candidate genes and the four

models of correspondence between the genotype and the trait.

When two SNPs interact with no marginal effects, the combination SNP test has the greatest power. For the *GPC6* gene, with  $\omega^2=8\%$ , the SNP combination test has 80% power compared to 5% and 10%, respectively, for the single SNP test and the multilocus genotype test.

Interestingly, the combination SNP test keeps good power for the model considering a single etiological polymorphism compared to the single SNP test, which fits the model. In this case, for the *LICT* gene, the decrease of power compared to the “single SNP test” (test 1) is only of 10% (75% compared to 85%) when  $\omega^2=5\%$ . The “combination SNP test” (test 3) remains an appropriate strategy for a single etiological SNP.

The pattern of results is similar for the three studied genes. Due to the different profiles of linkage disequilibrium in these genes, different powers are obtained for the same approach, but the combination SNP test always remains more powerful for models with no marginal effects. When the number of typed SNPs increases, the number of genotypic categories increases dramatically for the “multilocus genotype test” (test 2), which explains the decrease of power observed for *LICT* and *GPC6* compared to *CETP* (for the model with a single etiological SNP and  $\omega^2=10\%$ , the power is, respectively, 50% and 10% compared to 85%).

## APPLICATION TO GAW 12 SIMULATED DATA

To confirm the utility of the “combination SNP test,” we applied it to the Genetic Analysis Workshop 12 (GAW 12) simulated data set [Almasy et al., 2001] for gene 2 and quantitative trait 5. Under the simulated model, this gene explains 37% of the variance of quantitative trait 5, which is also influenced by sex and an environmental factor E1. All variants located in the exons and in the regulatory region have an additive effect. We used the 50 replicates extracted from the general population. For each replicate, we selected only unrelated individuals for whom we previously reconstructed the phased genotype, using Allegro [Gudbjartsson et al., 2000]. The quantitative trait was adjusted for age, sex, and the environmental factor E1. For computing-time reasons, we decide to select eight SNPs, using the following selection criteria: frequency higher than 5%, located either in the exons or in a regulatory

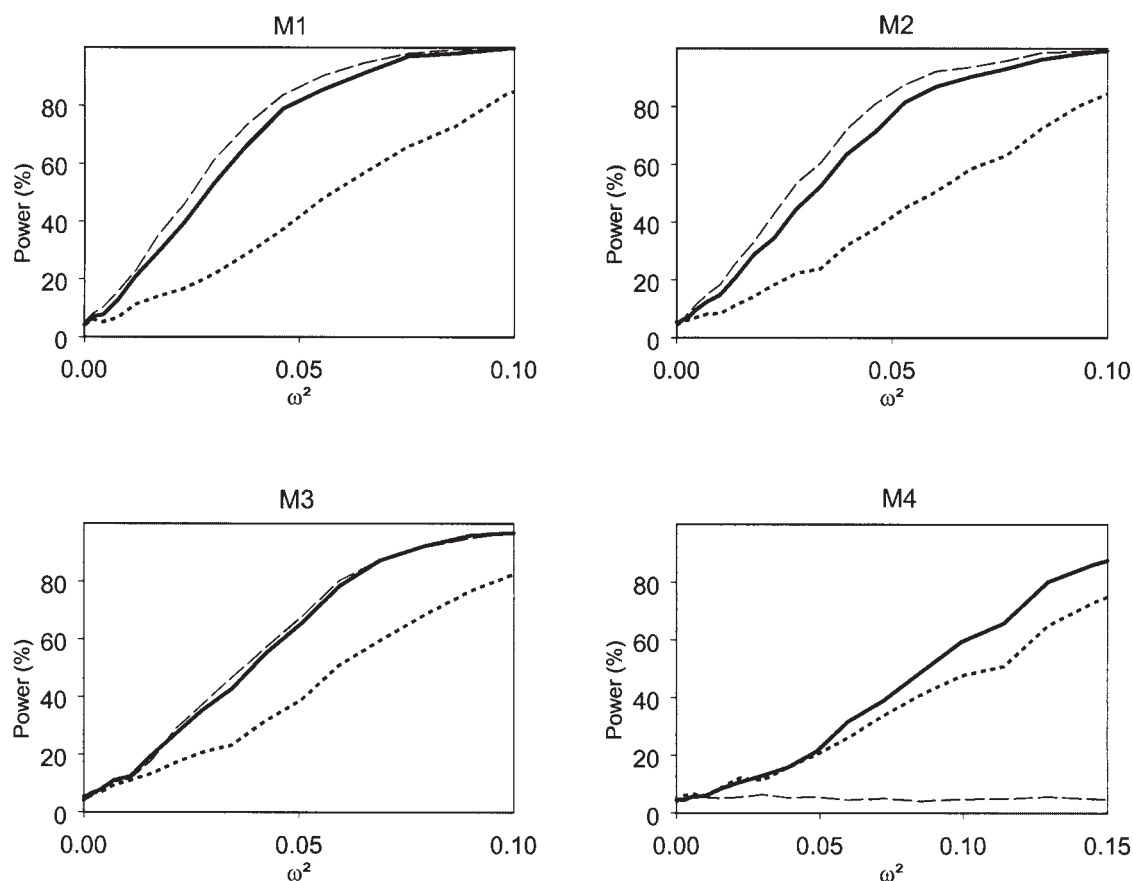


Fig. 1. Power to detect effect of gene, using *CETP* genotypes under four models of correspondence between genotype and phenotype for test 1 (dashed line), test 2 (dotted line), and test 3 (solid line).

region, and discarding SNPs showing strong linkage disequilibrium with any other. We found significant results in 24 (48%) replicates using the “combination SNP test,” 9 (18%) replicates using the “single SNP test,” and 10 (20%) replicates using the “multilocus genotype test.” For this additive model, the “single SNP test” and the “multilocus genotype test” have lower power compared to the “combination SNP test.” The low power achieved by the single SNP test can be explained by the fact that each SNP has only a small effect on the trait value, and by considering the SNPs independently, the association cannot be detected. The low power achieved by the “multilocus genotype test” is due to the large number of categories with sparse individuals, which do not provide information while adding one degree of freedom for each category. This shows that the “combination SNP test” can achieve a better power than the two other methods, even in the case of no interactions between the different SNPs.

## DISCUSSION

In this paper, we compared a new test to detect the role of a candidate gene, the combination SNP test, with two commonly used tests, the single SNP test and the multilocus genotype test. We showed that the combination SNP test is able to detect the role of a candidate gene in situations where the two other tests cannot. In addition, for other situations considered in this paper, even if the “combination SNP test” is not always the most powerful, it still has very close power to the best test. It also achieves the best power for GAW12 data among the three tests considered, when all the typed SNPs have small and additive effects.

The combination SNP test is the only one able to detect the role of the gene for the model with no marginal effects for each SNP (M4). This may be surprising, because the number of tests performed in this approach is often extremely high. However, the power remains good because firstly the correlation between the different tests performed

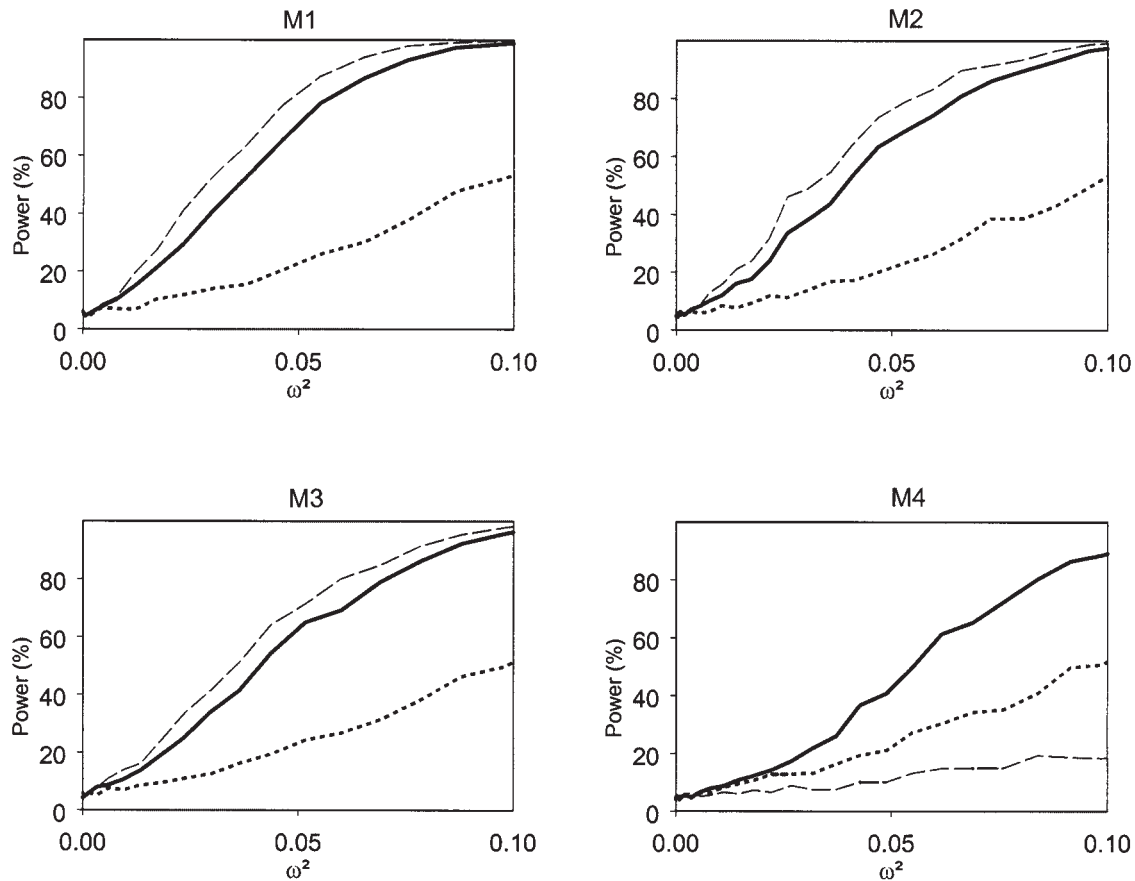


Fig. 2. Power to detect effect of gene, using *LICT* genotypes under four models of correspondence between genotype and phenotype for test 1 (dashed line), test 2 (dotted line), and test 3 (solid line).

is correctly taken into account, and secondly because the etiological genotype is tested. The simulation study shows that the power is fairly the same for the single SNP test and the combination SNP test for the three first models (M1, M2, and M3). Each involved SNP has a nonnegligible individual effect on the trait, so the effect is detectable by the single SNP approach.

For the three tests, power depends highly on the number of multilocus genotypes and the number of etiological SNPs. When few SNPs are etiological compared to the number of multilocus genotypes, many genotypic categories have the same trait level, leading to an inflation of the number of degrees of freedom. In this case, the multilocus genotype test has low power, while the combination SNP test keeps a better power. It is the same case for *GCP6* and *LICT*. On the contrary, for *CETP*, the number of multilocus genotypes is smaller than for *LICT* and *GCP6* (see Table III), leading to similar power for the two tests.

The calculation requirements for the combination SNP test is the number of tests needed for the combination test times the number of permutation tests needed to assess the critical value. Therefore, it can become rapidly prohibitive with an increasing number of typed SNPs available, as the algorithm is exponential with the number of markers. When the computation time becomes prohibitive, we could use an extension of the combination test that tests only combinations containing up to  $k$  SNPs,  $k$  being fixed by computational limitations.

When the sample size is small compared to the number of typed SNPs, the number of genotypic categories with sparse individuals when performing the combination SNP test and the multilocus genotype test is large. We checked that the F-test is robust to small categories.

The simulation study considers three distinct patterns of linkage disequilibrium and numbers of SNPs. For *CETP*, three major haplotypes can be clearly identified (see Table I), whereas there are

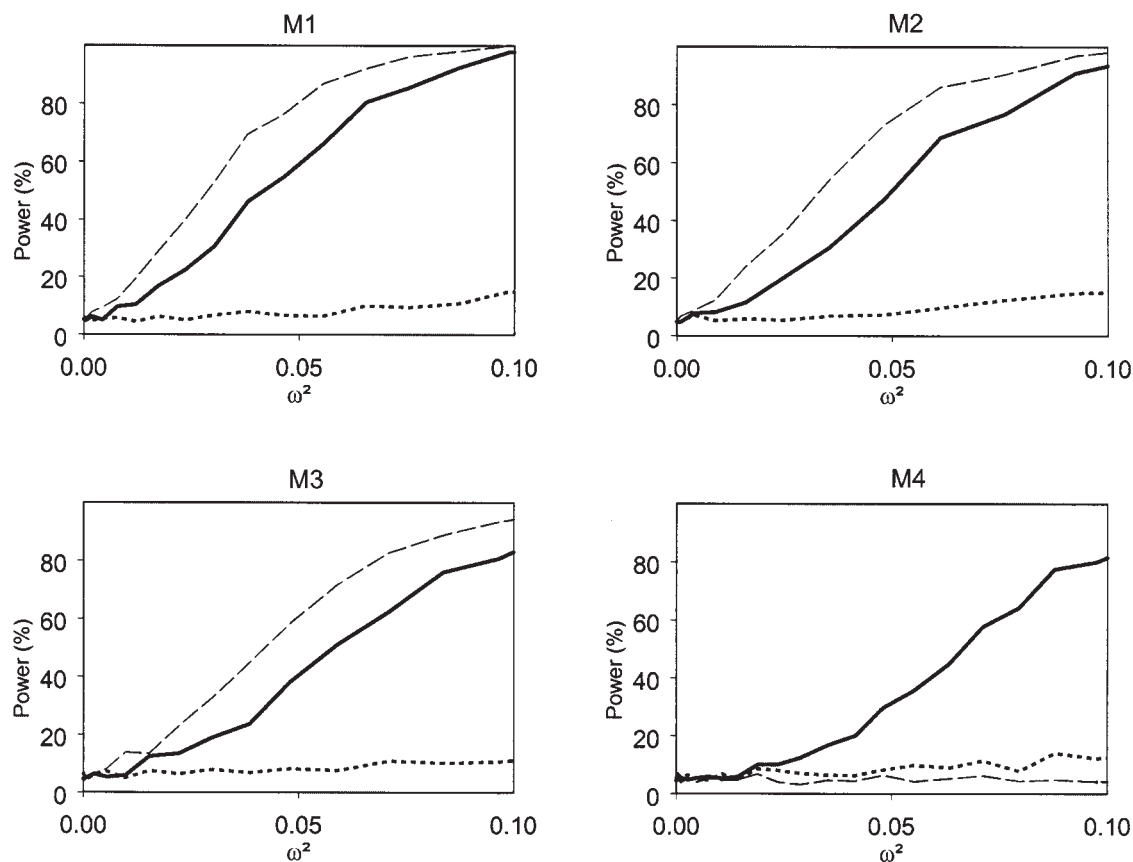


Fig. 3. Power to detect of effect of gene, using *GPC6* genotypes under four models of correspondence between genotype and phenotype for test 1 (dashed line), test 2 (dotted line), and test 3 (solid line).

TABLE III. Frequency and mean parameters taken for construction of model with no marginal effect of each SNP

SNP <i>b/B</i>	<i>bb</i>	<i>bB</i>	<i>BB</i>	Marginal of <i>A</i>
SNP <i>a/A</i>				
<i>aa</i>	$f_{11}, \mu_{11}$	$f_{12}, \mu_{12}$	$f_{13}, \mu_{13}$	$f_a^2, \mu_{aa}$
<i>aA</i>	$f_{21}, \mu_{21}$	$f_{22}, \mu_{22}$ <sup>a</sup>	$f_{23}, \mu_{23}$	$2f_a(1-f_a), \mu_{aA}$
<i>AA</i>	$f_{31}, \mu_{31}$	$f_{32}, \mu_{32}$	$f_{33}, \mu_{33}$	$(1-f_a)^2, \mu_{AA}$
Marginal of <i>B</i>	$f_b^2, \mu_{bb}$	$2f_b(1-f_b), \mu_{bB}$	$(1-f_b)^2, \mu_{BB}$	$1, \mu$

<sup>a</sup>There are two possibilities for haplotypes corresponding to these genotypes at each locus: either we have haplotypes *ab* and *AB*, or we have haplotypes *aB* and *Ab*.

no major haplotypes for *GPC6*. The advantage of the SNP combination test for models with weak marginal effects is valid for these three patterns of linkage disequilibrium.

All the simulated models consider that the etiological polymorphism(s) is/are typed. The etiological polymorphism can remain untyped or

not included. Some studies [Akey et al., 2001; Fallin et al., 2001; Verpillat et al., 2001] showed that, under the hypothesis that the etiological marker is not among those typed, a haplotype-based test has better power than a single marker test. In this case, the role of the gene can only be detected through preferential association of the etiological allele with some haplotypes, and thus not taking into account the association of some alleles on the same haplotype (the linkage disequilibrium information) decreases the power to detect the effect of the gene. The ‘‘combination SNP test’’ should be of high interest in this case, because it optimally uses the preferential association of alleles on the same haplotype.

Methods using the information provided by combinations of various numbers of SNPs like the combination SNP test have already been proposed. One strategy [Nelson et al., 2001] considers all possible partitions of multilocus genotypes. This strategy aims to find the partition that explains the largest part of the variance of

the phenotype. However, as stated by the authors, the number of tests involved is not computationally tractable for more than two SNPs, which limits the use of this approach. Our proposed SNP combination test has, on the contrary, a reasonable computing time for eight typed SNPs.

Multiple regression methods were also shown to be promising to deal with multiple intragenic SNPs. For the application to the GAW 12 data, this is the most powerful method (90% power), because it fits the simulated model. We tested (results not shown) that the multiple regression methods considering the effect of each SNP in a linear additive model have no power when models with weak marginal effects are considered. The power remains surprisingly low when interactions of the second order are considered (e.g., for *GPC6*, model M4 and  $\omega^2=7\%$ , the power is about 30%, compared to 80% for the "combination SNP test"). This is due to the large number of terms in the linear model considered in that case.

In our study, it must be noted that we compared the three methods for genotypic information. However, the advantage of test 3 compared to the other tests is likely to hold when using haplotypic information.

In this study, haplotypic phase is assumed to be available. Meanwhile, even when using a familial design, there may be phase uncertainty. The extension of the combination SNP test for uncertain phase is straightforward. The idea is to replace the genotype by all the possible genotypes with their probability.

The fact that the single SNP test and the multilocus genotype test give low power for the model with no marginal effects of each SNP was already noticed in another study [Culverhouse et al., 2002]. In that study, the authors showed that an association study for models displaying no main effect is not able to detect the effect of the gene, and they concluded that a linkage study is more powerful. Using the combination SNP test, we have a powerful nonparametric alternative approach that, without testing specifically the interaction between polymorphisms, enables detection of the role of the gene even when the etiological SNPs have no marginal effects.

## REFERENCES

- Akey J, Jin L, Xiong M. 2001. Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* 9:291–300.
- Almasy L, Terwilliger JD, Nielsen D, Dyer TD, Zaykin D, Blangero J. 2001. GAW12: simulated genome scan, sequence, and family data for a common disease. *Genet Epidemiol* 21:332–338.
- Bader JS. 2001. The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics* 2:11–24.
- Bickeboller H, Campion D, Brice A, Amouyel P, Hannequin D, Didierjean O, Penet C, et al. 1997. Apolipoprotein E and Alzheimer disease: genotype-specific risks by age and sex. *Am J Hum Genet* 60:439–446.
- Culverhouse R, Suarez BK, Lin J, Reich T. 2002. A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* 70:461–471.
- Di Paola R, Frittitta L, Miscio G, Bozzali M, Baratta R, Centra M, Spampinato D, et al. 2002. A variation in 3' UTR of hPTP1B increases specific gene expression and associates with insulin resistance. *Am J Hum Genet* 70:806–812.
- Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res* 11:143–151.
- Fijneman RJ, de Vries SS, Jansen RC, Demant P. 1996. Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. *Nat Genet* 14:465–467.
- Gudbjartsson DF, Jonasson K, Frigge ML, Kong A. 2000. Allegro, a new computer program for multipoint linkage analysis. *Nat Genet* 25:12–13.
- Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 9:720–731.
- Longmate JA. 2001. Complexity and power in case-control association studies. *Am J Hum Genet* 68:1229–1237.
- Nelson MR, Kardia SL, Ferrell RE, Sing CF. 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 11:458–470.
- Pagani F, Buratti E, Stuani C, Bendix R, Dork T, Baralle FE. 2002. A new type of mutation causes a splicing defect in ATM. *Nat Genet* 30:426–429.
- Schneider S, Roessli D, Excoffier L. 2000. A software for population genetics data analysis. Version 2.000. Geneva: Genetics and Biometry Lab, Department of Anthropology, University of Geneva.
- Stam LF, Laurie CC. 1996. Molecular dissection of a major gene effect on a quantitative trait: the level of alcohol dehydrogenase expression in *Drosophila melanogaster*. *Genetics* 144:1559–1564.
- Thomson G. 1983. Investigation of the mode of inheritance of the HLA associated diseases by the method of antigen genotype frequencies among diseased individuals. *Tissue Antigens* 21: 81–104.
- Tiret L, Poirier O, Nicaud V, Barbaux S, Herrmann SM, Perret C, Raoux S, et al. 2002. Heterogeneity of linkage disequilibrium in human genes has implications for association studies of common diseases. *Hum Mol Genet* 11:419–429.
- van Wezel T, Stassen AP, Moen CJ, Hart AA, van der Valk MA, Demant P. 1996. Gene interaction and single gene effects in colon tumour susceptibility in mice. *Nat Genet* 14: 468–470.
- Verpillat P, Bouley S, Campion D, Hannequin D, Dubois B, Belliard S, Puel M, et al. 2001. Use of haplotype information to test involvement of the LRP gene in Alzheimer's disease in the French population. *Eur J Hum Genet* 9:464–468.

## APPENDIX

## CONSTRUCTION OF MODEL WITH NO MARGINAL EFFECT OF EACH SNP

We consider two SNPs  $a/A$  and  $b/B$ , with frequency  $f_a$  for  $a$ , and  $f_b$  for  $b$ .

Following the notation of Table III, the genotypic means are:

$$\mu_{aa}=(f_{11}\mu_{11}+f_{12}\mu_{12}+f_{13}\mu_{13})/f_a^2$$

$$\mu_{aA}=(f_{21}\mu_{21}+f_{22}\mu_{22}+f_{23}\mu_{23})/2f_a(1-f_a)$$

$$\mu_{AA}=(f_{31}\mu_{31}+f_{32}\mu_{32}+f_{33}\mu_{33})/(1-f_a)^2$$

$$\mu_{bb}=(f_{11}\mu_{11}+f_{21}\mu_{21}+f_{31}\mu_{31})/f_b^2$$

$$\mu_{bB}=(f_{12}\mu_{12}+f_{22}\mu_{22}+f_{32}\mu_{32})/2f_b(1-f_b)$$

$$\mu_{BB}=(f_{13}\mu_{13}+f_{23}\mu_{23}+f_{33}\mu_{33})/(1-f_b)^2.$$

No marginal effects for each SNP taken separately give the following equalities:

$$\mu_{aa}=\mu_{aA}=\mu_{AA} \text{ and } \mu_{bb}=\mu_{bB}=\mu_{BB}.$$

This system of four equalities is equivalent to the following equations:

$$(f_{11}\mu_{11}+f_{12}\mu_{12}+f_{13}\mu_{13})/f_a^2$$

$$=(f_{21}\mu_{21}+f_{22}\mu_{22}+f_{23}\mu_{23})/2f_a(1-f_a)$$

$$=(f_{31}\mu_{31}+f_{32}\mu_{32}+f_{33}\mu_{33})/(1-f_a)^2$$

$$(f_{11}\mu_{11}+f_{21}\mu_{21}+f_{31}\mu_{31})/f_b^2$$

$$=(f_{12}\mu_{12}+f_{22}\mu_{22}+f_{32}\mu_{32})/2f_b(1-f_b)$$

$$=(f_{13}\mu_{13}+f_{23}\mu_{23}+f_{33}\mu_{33})/(1-f_b)^2.$$

To find parameter values compatible with no marginal effect for each SNP, we need to solve this system of equations. This system of equalities has an infinite number of solutions, because there are 4 equalities with 9 parameters to estimate, and so there are 5 degrees of freedom. To reduce the number of parameters, we put the constraint that the phenotypic trait means of the heterozygotes lies between the two phenotypic trait means of the homozygotes (this constraint is observed if there is no over/underdominance effect), and that one of the haplotypes has a dominant effect, but we could fix constraints of any type to obtain a model with weak marginal effects. We put the further constraint that the two genotypes  $Ab/aB$  and  $AB/ab$  have the same trait mean equal to zero.