Ann. Zool. Fennici 45: 429–440 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 29 October 2008

Linear models for analysis of multiple single nucleotide polymorphisms with quantitative traits in unrelated individuals

Jian-Feng Meng[1] & Tasha E. Fingerlin[2,*]

[1)] *Pediatric Immunology Research Department, Children's Mercy Hospital, School of Medicine, University of Missouri at Kansas City, 2401 Gillham Road Kansas City, MO 64108, USA*
[2)] *Department of Preventive Medicine and Biometrics, School of Medicine, University of Colorado Denver, 4200 East Ninth Avenue, B-119, Denver, CO 80262, USA (corresponding author's e-mail: tasha.fingerlin@uchsc.edu)*

Received 30 Nov. 2007, revised version received 6 June 2008, accepted 8 June 2008

Meng, J. F. & Fingerlin, T. E. 2008: Linear models for analysis of multiple single nucleotide polymorphisms with quantitative traits in unrelated individuals. — *Ann. Zool. Fennici* 45: 429–440.

Population-based genetic association studies are increasingly used to explore the association between genetic polymorphisms and outcomes such as disease-status and disease-related quantitative traits. Because multiple polymorphisms are typically available, there are several statistical analysis strategies that might be appropriate depending on the goal of the study. In this paper, we compare several linear model parameterizations that might be used to perform a test of association between a genomic region defined by multiple SNPs and a quantitative trait. We compare via simulation the type I error and power of the omnibus F-test to detect association. As expected, there is no one most powerful test across the genetic models we considered, although tests based on simple parameterizations that do not rely on phase information can be as powerful as more complicated haplotype-based tests even when it is a haplotype that is truly associated with the trait.

Introduction

Many diseases of public health importance are complex genetic diseases, which likely result from the interaction of multiple genetic polymorphisms and environmental stimuli. Genetic association studies are increasingly used as a tool to search for genetic variants that influence susceptibility to diseases such as type 2 diabetes and to variability in disease-related quantitative traits such as insulin sensitivity. Advances in genotyping technology have dramatically reduced the cost and time of obtaining dense single-nucleotide polymorphism (SNP) data in large samples of individuals. Genetic association studies often test whether (1) the distribution of alleles or genotypes at one or more SNPs differs between samples of affected and unaffected individuals, or (2) the alleles or genotypes at one or more SNPs explain variability in a disease-related quantitative trait. These studies may be conducted using only a very small region of the genome as part of candidate gene study, but can now be conducted on a genome-wide scale.

Because multiple SNPs are typically available for genetic association studies, there are several statistical analysis strategies that might be appropriate depending on the goal of the study. The strategies most often used generally fall into two categories: those that use SNP genotypes and those that use haplotypes. Tests of association that use haplotypes are of interest, in part, because of the biological interpretation of a haplotype. A sequence of alleles along a single chromosome, a haplotype represents a portion of that chromosome and has the potential to capture multiple *cis*-acting variants. In addition, because of linkage disequilibrium between variants along a chromosome, a haplotype may be more correlated with an ungenotyped functional variant than any single genotyped variant, potentially increasing power to detect association (Schaid 2004). Although the parameter estimates from a SNP-based association study can be easily interpreted, models which include multiple SNPs do not model potentially relevant haplotype structure. Haplotype association analyses may increase power, but require stronger assumptions regarding the importance of multiple SNPs in defining the functional variant(s), and the effects of individual loci may not be as easily identified.

Several authors have compared the power of tests of association based on either single-SNP or haplotype tests of association for both qualitative and quantitative traits (e.g. Long & Langely 1999, Bader 2001, Akey & Xiong 2001, Conti & Gauderman 2004, Schaid 2004). As noted by Schaid (2004), more comprehensive investigations that compare additional models which include multiple SNPs and SNP interaction effects are warranted. In this paper, we compare several linear models that can be used to perform an initial test of association between a genomic region defined by multiple SNPs and a quantitative trait. Following a similar investigation conducted by Conti and Gauderman (2004) for a qualitative trait, we compare via simulation the type I error and power of the omnibus *F*-test to detect association between a genomic region and a quantitative trait. To illustrate the application of the different statistical models, we use data from the Insulin Resistance and Atherosclerosis Study.

Material and methods

We assume that genotypes for two or more SNPs in a candidate region are available on a set of unrelated individuals to test for association with a quantitative trait. Further, we make the simplifying assumption that haplotype phase is known, the implications of which we describe in the Discussion. We use a linear regression framework to model the relationship between one or more SNPs and the quantitative trait and assume that the quantitative trait (perhaps after appropriate transformation) follows a normal distribution conditional on the independent variables in the model. As described below, we assume the primary objective is to test for association between the candidate region and the quantitative trait and that subsequent analyses can be conducted to determine which SNP(s) or haplotype(s) might be most important.

Below we describe several models that have been proposed for testing for association between one or more SNPs and a quantitative trait and how we implemented them for this study.

Linear models for multilocus data

Models based on SNPs

The three most commonly used approaches that use SNP genotypes rather than haplotype information are (1) a single-SNP model, where each SNP is tested independently of the other SNPs, (2) a joint main effects model, where the omnibus *F*-test is used to test for association with all SNPs simultaneously, and (3) interaction models that include two or more of the SNPs of interest, where the omnibus *F*-test is again used to test for association. If there are N SNPs, N separate regression models are constructed and N *F*-statistics are computed for (1). The single SNP models have the following form for $n = (1, \ldots, N)$:

$$Y = \beta_0 + \beta_n X_n + \varepsilon \qquad (1)$$

where Y is a variable for the quantitative trait and X_n is a variable coding the genetic effect. The variable X_n may take on several forms, includ-

ing that of an indicator variable representing a dominant or recessive model for a given allele. Here, we let X_n = the number of rare alleles at marker n; i.e. $X_n = 2$ for genotype AA, $X_n = 1$ for genotype Aa and $X_n = 0$ for genotype aa under an additive model, where A is the rare allele and a is the common allele at one locus. A joint main effects model has the following form:

$$Y = \beta_0 + \sum_{n=1}^{N} \beta_n X_n + \varepsilon \quad (2)$$

To capture interaction effects, an interaction model can be built by adding interaction terms into the joint main effects model. Here we include only second-order interactions where $X_{m\times l} = X_m \times X_l$ for markers m and l.

$$Y = \beta_0 + \sum_{n=1}^{N} \beta_n X_n + \sum_{n=1}^{N} \sum_{l=n+1}^{N} \beta_{m\times l} X_{m\times l} + \varepsilon \quad (3)$$

For each of the multilocus models (2 and 3), we use an omnibus F-test to compare the full model with the model with no genetic effects as a test for association between the chromosomal region and the quantitative trait. A test for association between each SNP (or a subset of SNPs) and the trait can be obtained by a partial F-test in subsequent analyses. We apply a Bonferroni correction for the single-SNP models to correct for the multiple tests computed.

Models based on haplotypes

Similar to the multi-SNP models that assume an additive effect of each allele, for a set of H haplotypes, we model the additive haplotype effect for $H - 1$ of the haplotypes in a regression model as above, with

$$Y = \beta_0 + \sum_{h=2}^{H} \gamma_h X_h + \varepsilon \quad (4)$$

Here X_h is a variable for the number of haplotypes of type h in each individual. In the absence of an *a priori* putative at-risk haplotype, the most common haplotype is generally treated as the reference haplotype. Also similar to the multi-SNP analyses, an omnibus F-test is used as a test for association between the chromosomal region and the quantitative trait. A partial F-statistic can be calculated for each γ_h to test for association between each haplotype and the trait in subsequent analyses.

The SIMPle model

To take advantage of the simplicity of the SNP-based analyses while retaining the ability to model haplotype effects, Conti and Gauderman (2004) proposed a genotype-level analysis to jointly model SNPs via a SNP interaction model with phase information (SIMPle) to capture the underlying haplotype structure. Phase refers to whether or not the alleles are in *cis* (exist on the same haplotype). The SIMPle paradigm is very similar to the interaction model (3) above, but the interaction terms are modified. In the SIMPle model, the phase information between pair-wise SNPs is included by modifying the second order interaction terms in model (3) to reflect haplotypes. For example, for SNPs 1 and 2, with rare alleles A and B, respectively, $X'_{1\times 2}$ combines the phase information and interaction term in the following coding scheme given the two haplotypes for individual i, h_{i1} and h_{i2}:

$$X'_{1\times 2} \begin{cases} 2 \text{ if } X_1 \times X_2 = 4 \\ 1 \text{ if } X_1 \times X_2 = 2 \\ 1 = \text{if } X_1 \times X_2 = 1 \text{ and } h_{i1} \text{ or } h_{i2} \text{ is AB} \\ 0 \text{ if } X_1 \times X_2 = 0 \text{ and neither } h_{i1} \text{ nor } h_{i2} \text{ is AB} \\ 0 \text{ if } X_1 \times X_2 = 0 \end{cases}$$

where X_1 and X_2 are assigned a value under an additive model as described above in (1). Note that for those with heterozygous genotypes at both SNPs, the interaction term takes on a different value depending on which haplotypes the person carries. Because of this modification in the interaction term, the linear predictor for the SIMPle model distinguishes between individuals with haplotypes AB/ab and those with Ab/aB, whereas the linear predictors for those two haplotype combinations are identical in the interaction model (Table 1, rows 4 and 5). Again, an omnibus F-test is used as a test for association between the chromosomal region and the trait. The SIMPle parameterization has the advantage of being able to capture the importance of

pair-wise phase, whereas the SNP-based multilocus models (2 and 3) do not. Note that the SIMPle and haplotype models give the same fit to the data when there are only two SNPs. For more than two SNPs, the SIMPle phase/interaction terms as we have implemented them here capture only pair-wise phase effects, whereas the haplotype model captures phase information across all markers.

Simulation study

To compare the size and power of the *F*-test across the above linear models, we simulated multiple SNP (≥ 2 SNPs) data under the null hypothesis of no association between any of the SNPs and the quantitative trait (for size) and under several alternative hypotheses corresponding to at least one of the SNPs being associated with variation in the quantitative trait (for power). We generated haplotype information for each individual so that phase was known. We generated a quantitative trait value for each person (Y_0), drawn from a standard normal distribution with mean 0 and standard deviation (SD) 1. We assumed a 0.2 SD difference in means between genotypic groups of interest (more below), which accounts for approximately 4% of the variance of the quantitative trait if the true variant allele has a 50% population frequency. We generated data for a population size of 25 000, and selected a random sample of 500 individuals for each replicate. For each simulation condition, we generated 1000 replicates and calculated the omnibus *F*-test for each of the models described above. Size and power were determined by the proportion of replicates that had a $p \leq 0.05$.

In the first set of simulation conditions (simulation study I), we generated data for two SNPs so that there were only 4 possible haplotypes; SNP 1 with alleles A and a and SNP 2 with alleles B and b, each with frequency 50%. In scenario 1, we assumed that each copy of allele A from SNP 1 increased the trait value by 0.20 SD. In scenario 2, we assumed that neither allele A nor B were true trait-influencing variants, but could be used as proxies for an unknown quantitative trait variant within the candidate region. We included a third, unknown quantitative-trait-influencing variant, *D*, which was located between SNPs 1 and 2 and had the same r^2 value with both A and B ($r^2 = 0.85$). To model an additive haplotype effect, in scenario 3, we assumed that each copy of haplotype AB increased the trait value by 0.2 SD. In scenario 4, to simulate the case where multiple haplotypes may influence the trait, we assumed that the AB haplotype was associated with the largest increase in the trait but that the Ab and aB haplotypes also increased the mean trait value as compared with the ab haplotype (AB > Ab > aB; 0.1, 0.06, 0.04 SD, respectively). Finally, in scenario 5, we assumed that alleles A and B both increased the mean trait value and that the interaction of A and B (not necessarily on the same haplotype) was also associated with an increase in the quantitative trait value (A = B

Table 1. Linear predictors for multilocus models (adapted from Conti *et al.* 2004).

Haplotype profile	Genotype	SNP model		SIMPle	Haplotype
		Main	Interaction		
AB|AB	AA, BB	$\beta_0 + 2\beta_A + 2\beta_B$	$\beta_0 + 2\beta_A + 2\beta_B + 4\beta_{AB}$	$\beta_0 + 2\beta_A + 2\beta_B + 2\beta_{AB}$	$\gamma_0 + 2\gamma_{AB}$
AB|Ab	AA, Bb	$\beta_0 + 2\beta_A + \beta_B$	$\beta_0 + 2\beta_A + \beta_B + 2\beta_{AB}$	$\beta_0 + 2\beta_A + \beta_B + \beta_{AB}$	$\gamma_0 + \gamma_{AB} + \gamma_{Ab}$
AB|aB	Aa, BB	$\beta_0 + \beta_A + 2\beta_B$	$\beta_0 + \beta_A + 2\beta_B + 2\beta_{AB}$	$\beta_0 + \beta_A + 2\beta_B + \beta_{AB}$	$\gamma_0 + \gamma_{AB} + \gamma_{aB}$
AB|ab*	Aa, Bb	$\beta_0 + \beta_A + \beta_B$	$\beta_0 + \beta_A + \beta_B + \beta_{AB}$	$\beta_0 + \beta_A + \beta_B + \beta_{AB}$	$\gamma_0 + \gamma_{AB}$
Ab|aB*	Aa, Bb	$\beta_0 + \beta_A + \beta_B$	$\beta_0 + \beta_A + \beta_B + \beta_{AB}$	$\beta_0 + \beta_A + \beta_B$	$\gamma_0 + \gamma_{AB} + \gamma_{aB}$
Ab|Ab	AA, bb	$\beta_0 + 2\beta_A$	$\beta_0 + 2\beta_A$	$\beta_0 + 2\beta_A$	$\gamma_0 + 2\gamma_{Ab}$
aB|aB	aa, BB	$\beta_0 + 2\beta_B$	$\beta_0 + 2\beta_B$	$\beta_0 + 2\beta_B$	$\gamma_0 + 2\gamma_{aB}$
Ab|ab	Ab, bb	$\beta_0 + \beta_A$	$\beta_0 + \beta_A$	$\beta_0 + \beta_A$	$\gamma_0 + \gamma_{Ab}$
aB|ab	aa, Bb	$\beta_0 + \beta_B$	$\beta_0 + \beta_B$	$\beta_0 + \beta_B$	$\gamma_0 + \gamma_{aB}$
ab|ab	aa, bb	β_0	β_0	β_0	γ_0

* these two haplotype profiles result in the same genotype.

< AB; 0.05, 0.05, 0.10 SD, respectively). Within scenarios 1 and 3–5, we also varied whether A and B were in linkage disequilibrium. For scenarios 1.1, 3.1, 4.1 and 5.1, we assumed that A and B were in linkage *equilibrium*, so that the four haplotypes (AB, Ab, aB and ab) had equal frequencies (0.25). For scenarios 1.2, 3.2, 4.2 and 5.2, we assumed that A and B were in linkage *disequilibrium* with $r^2 = 0.85$, which corresponds to haplotype frequencies: AB = ab = 0.48 and Ab = aB = 0.02.

We also constructed a series of simulation conditions (simulation study II) to more broadly reflect the kinds of data we expect to see in data applications for $N = 2, 3$ and 4 SNPs in scenarios 6, 7 and 8. We again simulated 4 haplotypes with frequencies (ab = 0.05, Ab = 0.15, aB = 0.35 and AB = 0.45) for the two-SNP scenarios (6.1, 6.2, 6.3). To reflect the fact that haplotype analyses are generally only conducted in regions of relatively high LD, and therefore not all of the possible haplotypes are observed, we generated data for only 5 of the 8 possible haplotypes for the three-SNP scenarios (7.1, 7.2, 7.3), and for 8 of the possible 16 for the four-SNP scenarios (8.1, 8.2 and 8.3). Following Stram *et al*. (2003) we used the observed four-SNP haplotype frequencies from a study of the PGR gene. Since the frequency of the true quantitative trait variant is unknown, we varied which haplotype increased the trait value across scenarios 6, 7 and 8. We assumed a rare haplotype (6.1, 7.1, 8.1), a common haplotype (6.2, 7.2 and 8.2), or a random haplotype (each replicate had a randomly chosen haplotype; 6.3, 7.3 or 8.3) was the trait-influencing haplotype (Table 2). Finally, rather than fix the haplotype frequency distributions, we also randomly generated haplotype frequencies based on a multinomial distribution for each replicate, and then randomly chose one of those haplotypes to be the trait-influencing haplotype (6.4, 7.4 and 8.4).

Application: Association analysis of SNPs in *NFKBIA* and insulin sensitivity

Type 2 diabetes mellitus (T2DM) is a common, chronic disease characterized by hyperglycemia caused by defects in insulin secretion and insulin action. The regulatory mechanism behind the progress from normal glucose tolerance to T2DM is not well understood. Increasing insulin resistance has been observed prior to the development of T2DM (Lillioja *et al*. 1993, Haffner *et al*. 1995, Weyer *et al*. 1999) and amelioration of insulin resistance by thiazolidinediones in those at high risk has been shown to reduce risk of T2DM in some cases (Buchanan *et al*. 2000). Hence, reducing insulin resistance (increasing insulin sensitivity) has the potential to prevent T2DM. We examined several SNPs in *NFKBIA*, a candidate gene for influencing insulin sensitivity (Shoelson *et al*. 2003), and tested for association between these SNPs and a measure of insulin sensitivity (S_I) in the Insulin Resistance and Atherosclerosis Study (IRAS).

IRAS was a community-based epidemiological study of 1625 men and women designed to determine the correlates and predictors of insulin resistance and atherosclerosis. Detailed information about the study design and the measurements has been published elsewhere (Wagenknecht *et*

Table 2. Haplotype frequencies used for simulation study II.

	Haplotype	Frequency
Two SNPs		
1	00	0.05
2	10	0.15*
3	01	0.35
4	11	0.45**
Three SNPs		
1	000	0.266
2	001	0.202
3	110	0.131
4	011	0.075*
5	111	0.326**
Four SNPs		
1	0000	0.321
2	1000	0.163
3	0100	0.031
4	0001	0.143
5	1100	0.020*
6	1010	0.041
7	0101	0.041
8	1110	0.245**

* rare haplotype for simulation in 6.1, or 7.1 or 8.1.
** common haplotype for simulation in scenario 6.2, 7.2 and 8.2.
Note: 0 represents common allele and 1 represents rare allele for given SNP.

al. 1995). Briefly, a total of 165 individuals from San Antonio, TX, and 207 from the San Luis Valley, CO, subsequently consented to DNA studies and had a measure of insulin sensitivity (S_I) based on a frequently sampled intravenous glucose tolerance test (FSIGT). All studies were approved by the Institutional Review Boards of the respective institutions.

SNPs were selected from the SeattleSNP database (http://pga.gs.washington.edu/) in and around *NFKBIA*. A total of 11 SNPs were chosen for genotyping by requiring that each un-genotyped SNP had an $r^2 \geq 0.85$ with at least one genotyped SNP. Haplotype blocks were defined using Gabriel's definition (Gabriel *et al*. 2002) using the Haploview software (Barrett *et al*. 2005). Briefly, haplotype blocks are composed of groups of SNPs that are jointly in strong LD, such that there are fewer haplotypes observed than expected. Such blocks are identified for several reasons depending on the context; here the goal was to be able to obtain precise estimates of haplotype frequencies.

We computed the omnibus *F*-test for the SNP-based models using all SNPs. For each block, in addition to the SNP-based analyses (single-SNP, joint, interaction), we performed the analyses that required phase information (SIMPle and haplotype) using the SNPs in that block. Because haplotype phase was unknown, we used the most likely haplotype pair for each individual. The most likely pair of haplotypes was inferred for each individual based on the posterior probability associated with each pair for each person. These posterior probabilities were based on the estimated sample haplotype frequencies obtained via a Bayesian algorithm implemented in PHASE2.1 software (Stephens *et al*. 2001). Because the Bayesian algorithm requires that certain assumptions be made regarding population genetic parameters, we compared the haplotype frequency estimates obtained using the Bayesian algorithm to the maximum likelihood estimates obtained using an Expectation–Maximization (EM) algorithm using Haploview (Barrett *et al*. 2005). Both algorithms resulted in very similar haplotype frequency estimates. As the S_I distribution is skewed, the natural log transformation was used (Wagenknecht *et al*. 2003).

Results

Type I error rates and power comparison

The size of the *F*-test for each of the models considered was very close to the nominal 0.05 level (data not shown). In what follows, for ease of discussion, we refer to the power of the *F*-test associated with a certain model parameterization

Table 3. Power for 2-SNP scenarios (simulation study I).

Model (df)	Single-SNP 1 (1)	Single-SNP 2 (1)	Joint (2)	Interaction (3)	Haplotype (3)	SIMPle (3)
1.1	**0.81**	0.03	**0.80**	0.74	0.73	0.73
1.2	**0.81**	0.74	**0.81**	0.76	0.76	0.76
2	**0.74**	**0.75**	**0.74**	0.68	0.69	0.69
3.1	0.33	0.30	0.52	0.57	**0.66**	**0.66**
3.2	0.79	0.80	**0.82**	0.77	0.77	0.77
4.1	0.11	0.06	**0.16**	0.14	0.14	0.14
4.2	0.26	**0.28**	**0.28**	0.23	0.24	0.24
5.1	0.67	0.63	0.88	**0.89**	0.88	0.88
5.2	0.98	1.00	0.99	0.99	0.99	0.99
5.3	0.21	0.18	**0.33**	**0.34**	0.32	0.32
5.4	0.54	0.56	**0.57**	**0.55**	0.54	0.54

Note: Frequency of allele A at SNP 1 and allele B at SNP 2: 0.5. Haplotype frequencies are AB = Ab = aB = ab = 0.25 in 1.1, 3.1, 4.1, 5.1. Haplotype frequencies are AB = ab = 0.48 and Ab = aB = 0.02 in 1.2, 2, 3.2, 4.2 and 5.2. Trait variant is allele at SNP 1 for scenarios 1.1 and 1.2, is unobserved allele in LD with alleles from SNPs 1 and 2 for scenario 2, is haplotype AB for 3.1 and 3.2, is haplotype for 4.1 and 4.2, with haplotype AB most strongly associated (effect of AB > Ab > aB). For 5.1–5.4, interaction between alleles A and B influences trait. Values set in boldface indicate highest power for simulation scenario.

as the power for that model. The most important comparisons are within each simulation condition due to the artifactual differences in power across simulation conditions associated with the varying simulation conditions.

There are several important patterns to note when comparing the power for each of the models considered for the first two-SNP simulation scenarios (Table 3). First, since the SIMPle model is identical to the haplotype model for two SNPs, the two models have identical power for every simulation condition. The only case where the haplotype/SIMPle model was the most powerful was under the circumstance that only the AB haplotype was assumed to influence the trait (scenario 3.1). With the exception of scenario 3.1, the interaction model had essentially identical power to the haplotype/SIMPle model across the simulation conditions considered. In addition, with the exception of scenario 3.1, the joint model was as powerful as the most powerful model across the simulation scenarios.

When a single allele at SNP 1 was the only trait-influencing variant (scenarios 1.1, 1.2), the joint model and the single SNP model for SNP 1 were the most powerful models among those tested and had similar power due to the Bonferroni correction of the single-SNP p value. As expected, the interaction, haplotype and SIMPle models had lower power due to the added degree of freedom in the absence of a true haplotypic or interaction effect. When the true variant was unobserved, but was in LD with alleles at SNPs 1 and 2, respectively (scenario 2), both single-SNP models and the joint model had similar power. When the AB haplotype was associated with the largest increase in the trait value compared to the ab haplotype, but there were also trait differences associated with the other haplotypes (scenario 4.1, 4.2), the joint model was the most powerful unless there was strong LD between the alleles at the two SNPs, in which case both the single-SNP models had the same power as the joint model. When alleles at each SNP influenced the trait, and the interaction of the alleles also influenced variation in the quantitative trait, the joint model showed comparable power to the interaction model even though the test based on the interaction model might have been expected to be the most powerful (scenarios 5.1, 5.2, 5.3, 5.4). This similarity in power for the joint model did not persist when there was no main effect of either locus (data not shown). For the scenarios that assumed strong LD between the alleles at SNP 1 and 2 (1.2, 5.2) but did not explicitly assume a haplotypic effect, since the AB haplotype was so frequent, the A and B alleles were essentially observed together only in the context of that hap-

Table 4. Power for simulation study II.

	Model				
Scenario (# SNPs)	Single-SNP	Joint	Interaction	Haplotype	SIMPle
6.1 (2)	0.51	0.85	0.83	**0.86**	**0.86**
6.2 (2)	0.55	0.95	0.94	**0.97**	**0.97**
6.3 (2)	0.55	0.89	0.88	**0.90**	**0.90**
6.4 (2)	0.28	0.46	0.47	**0.52**	**0.52**
7.1 (3)	0.03	**0.44**	0.35	0.42	0.42
7.2 (3)	0.64	0.91	0.86	**0.94**	**0.94**
7.3 (3)	0.19	0.56	0.54	**0.67**	**0.67**
7.4 (3)	0.18	0.40	0.36	**0.42**	**0.42**
8.1 (4)	0.02	0.12	0.13	**0.21**	**0.21**
8.2 (4)	0.53	**0.89**	0.77	0.87	0.87
8.3 (4)	0.09	0.29	0.25	**0.34**	**0.34**
8.4 (4)	0.12	0.30	0.25	**0.32**	**0.32**

Note: rare haplotype influences trait in 6.1, 7.1, 8.1; common haplotype influences trait in 6.2, 7.2, 8.2, and randomly chosen haplotype (for each replicate) influences trait in 6.3, 7.3 and 8.3 (*see* Table 3). A randomly chosen haplotype with a randomly generated frequency influences trait in 6.4, 7.4, and 8.4. Values set in boldface indicate highest power for simulation scenario.

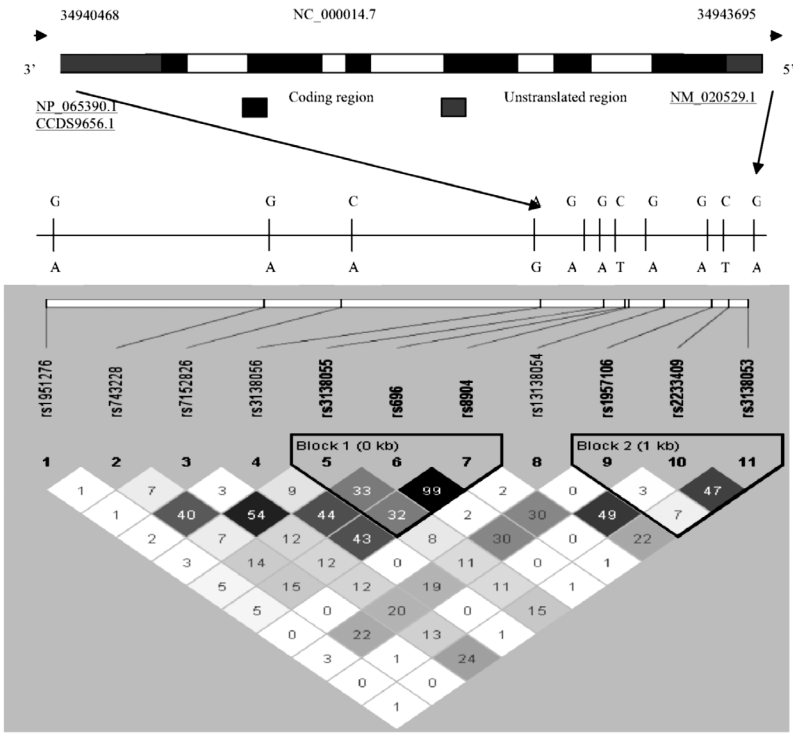

Fig. 1. LD blocks in *NFKBIA* gene in IRAS Hispanic sample. Numbers indicate squared correlation coefficient and shades of gray represent *D´*; Darker indicates higher *D´* value.

lotype. Hence, the AB haplotype was strongly associated with the trait and the haplotype model had correspondingly good power.

For the second set of simulation scenarios that considered a broader range of underlying genetic models, two comparisons are of particular interest (Table 4). Across the conditions we considered, the haplotype and SIMPle models are generally the most powerful among those we compared and have identical power even though they do not give the same fit to the data when there are more than two SNPs (scenarios 7 and 8; Table 4). It is important to note, however, that the joint model often had power similar to or slightly higher than these two models; in only two cases (7.3 and 8.1) was the power for the joint model appreciably lower than that for the haplotype and SIMPle models.

Application to *NFKBIA* and insulin resistance

The *NFKBIA* gene is located on chromosome 14 and is composed of six exons and five introns (Fig. 1). Genotypes were obtained for > 95% of subjects for each SNP. Based on Gabriel's block definition (Gabriel *et al*. 2002), two haplotype blocks were defined; Block 1 composed of rs3138055, rs696 and rs8904, and Block 2 composed of rs1957106, rs2233409 and rs3138053 (Fig. 1). Genotype frequencies at all SNPs were consistent with Hardy-Weinberg Equilibrium proportions (Table 5). Since SNPs rs696 and rs8904 were in near perfect LD, we used only rs696 in all models.

Variation in *NFKBIA* was not significantly associated with insulin resistance using the omnibus *F*-test for any of the SNP-based linear model parameterizations. In addition, when we analyzed SNPs from block two separately, none of the omnibus tests were significant. However, block 1 was associated with insulin sensitivity ($p = 0.03$) based on the *F*-tests from the joint, SIMPle and haplotype models. Because there were only two SNPs included, the SIMPLe and haplotype models would usually give the same fit to the data. However, the G allele at rs3138055 is essentially seen only on the same chromosome as the G allele at rs696, and so only three of the possible four haplotypes have frequency > 0.2% (Table 6). As such, the interaction term in the

Table 5. Allele frequency estimates for SNPs in *NFKBIA*.

SNP	Name	Position	HWE* *p*	Minor allele frequency	Minor allele
1	rs1951276	34923139	0.21	0.27	A
2	rs743228	34929825	0.18	0.405	G
3	rs7152826	34932155	0.42	0.419	A
4	rs3138056	34938165	0.67	0.48	G
5	rs3138055	34940205	0.84	0.448	G
6	rs696	34940844	0.22	0.292	A
7	rs8904	34940968	0.13	0.288	T
8	rs13138054	34942058	0.71	0.093	A
9	rs1957106	34943521	0.33	0.195	A
10	rs2233409	34944021	0.63	0.127	T
11	rs3138053	34944605	0.69	0.224	G

* Hardy-Weinberg Equilibrium.

SIMPle model is highly correlated with the term for SNP rs3138055, and having both terms in the model was not appropriate. Hence, in this case, the joint model and the SIMPle model were identical. Both the haplotype model and the joint model indicate that both rs3138055 and rs696 are associated with insulin sensitivity. Tests of the individual parameter estimates for the haplotype model indicate that the haplotype with the allele A at rs3138055 and allele G at rs696 is associated with an increase in insulin sensitivity as compared with the haplotype that that has allele G at rs3138055 and G at rs696 ($p = 0.02$). Even though the interaction model parameter estimates gave the same interpretation as the haplotype model, because it required three parameter estimates as compared with two for the haplotype model, the p value for the F-test for the interaction model was 0.08. The joint model was useful for testing purposes, but in this case was harder to interpret than the haplotype model since the effect, if real, appears to be haplotypic. In summary, these data illustrate the fact that the linkage disequilibrium patterns in a region of interest often dictate that the joint model has similar power to the haplotype model even when a haplotype effect is driving the association.

Discussion

Genetic association studies are an important tool in the search for disease-susceptibility variants that may influence risk of disease through a disease-related quantitative trait. Since many, perhaps densely-spaced, SNPs may define functional variation, statistical models that can be used to explore the association of these SNPs with quantitative traits are important for biomedical research. We have compared the power of the most common statistical testing frameworks used to test for association between a genomic region and a quantitative trait under several different genetic models for the relationship between the genomic region and the quantitative trait in a sample of unrelated individuals. We used a simulation approach to investigate the relative power of the various tests of association. This approach was particularly useful because it allowed us to compare the testing strategies using underlying genetic models that included non-trivial linkage

Table 6. Haplotype frequency estimates for blocks 1 and 2 in *NFKBIA*.

Haplotype*	Haplotype frequency
Block 1	
GGC	0.448
GAT	0.002
AGC	0.259
AAC	0.001
AAT	0.291
Block 2	
GCA	0.581
GCG	0.100
GTG	0.124
ACA	0.192
ATA	0.002

disequilibrium patterns among SNPs which can be very difficult or impossible to investigate when making analytic comparisons.

Conti and Gauderman (2004) compared the power of these statistical testing strategies for tests of association between multiple SNPs and a dichotomous trait. They found that the SIMPle model has comparable power to the haplotype model for detecting association between a genetic region represented by several SNPs and a dichotomous disease trait when the underlying functional variant is represented by a haplotype. Prior to this investigation, whether similar conclusions could be drawn with respect to quantitative traits was unknown.

Consistent with the observations of Conti and Gauderman (2004) for a qualitative trait, among the simulation scenarios we considered, the haplotype and SIMPle models have nearly identical power for a quantitative trait. When a specific haplotype is the true variant, both models are as powerful as the most powerful model. However, when the true variant is a single SNP or there are multiple haplotypes that influence a quantitative trait, the single-SNP models and joint model can be more powerful than both the haplotype and SIMPle models. Chapman *et al*. (2003) and Clayton *et al*. (2004) also found that tests based on a joint model (their main effects model) often had greater power than those based on haplotypes when a single variant, perhaps ungentoyped, was associated with a dichotomous outcome. The performance of the joint model is important as it allows joint testing of markers that are not necessarily in high linkage disequilibrium, while the current implementation of the haplotype model generally requires relatively high LD between the markers so that haplotype frequencies are estimated with reasonable accuracy and precision. (Fallin & Schork 2000).

Several others have investigated the relative merits of the single-marker (model 1) and haplotype (model 4) testing paradigms in the context of quantitative traits (e.g. Long & Langely 1999, Bader 2001, Schaid 2004). Our results are consistent with those investigations, indicating that the single-marker tests have the potential to be more powerful than haplotype tests when there are fewer SNPs than haplotypes and when one of the SNPs is either the causative SNP or in nearly perfect LD with the causative SNP. While Bader (2001) and others have noted the potential utility of jointly modeling the single-SNP effects and including interaction effects in quantitative trait analyses, those models were not within the scope of their investigations.

Our results indicate that, as expected, the assumptions about the number of true variant alleles, the LD between markers and the true trait variant(s), and the frequency of the marker and trait alleles influence the relative power among different statistical models and the absolute power for a single model across simulation conditions. For example, across the frequencies we considered, for a fixed effect of the allele (or haplotype), power is increased across models as the allele frequency increases. This can be illustrated by comparing scenarios 3 and 4. As the frequency of haplotype AB decreases, the power for all models was reduced since the effect of the haplotype was held constant. The power for each model decreases as the number of SNPs increases from two to four across all of the rare, common and random haplotype scenarios. As the number of SNPs increases, so does the number of single-SNP models and the degrees of freedom associated with the joint, interaction, and SIMPle models. The number of SNPs used for testing has also been noted as an important factor in determining power by Slager *et al*. (2000) and Morris and Kaplan (2002) for case-control studies. We considered up to four SNPs for use in analyses in what is presented here, although limited simulations show very similar results for up to eight SNPs. Since the number of potential haplotypes (and hence degrees of freedom) increases with each additional SNP, the single-SNP and/or joint tests will likely still be as or more powerful than haplotype-based tests when the true variant is a single SNP or there are multiple haplotypes that influence a quantitative trait.

We assumed that phase was known for our simulations. When phase is not known, ignoring the uncertainty in phase assignment may yield biased parameter estimates and inappropriately small estimates of the variability of those estimates (Schaid 2002, Zaykin *et al*. 2002, Kraft *et al*. 2005, Cordell 2006). As such, methods are available to appropriately account for unknown phase assignment for quantitative trait associa-

tion tests using unrelated individuals (Zaykin *et al*. 2002), but the properties of those methods are not well described for all of the parameterizations we explored. As genotyped SNPs are more densely spaced and as sequencing technology is increasingly feasible for large-scale studies, we expect that real data analyses will become more similar to phase-known than phase-unknown. However, investigating the impact of unknown phase on the relative power of the various parameterizations is a logical extension to the work presented here. We expect that in the phase-unknown case, the joint and single-SNP models may have even more of an advantage over the haplotype and SIMPle models when the true variant is a single SNP or there are multiple haplotypes that influence a quantitative trait. Tests based on either the haplotype or SIMPle model parameter estimates will be less efficient due to the variability induced by unknown phase, unlike those tests based on estimates from the single-SNP and joint models. The extent of the potential loss in power for the haplotype and SIMPle models will depend in part on the extent of LD among the SNPs included in the haplotype and the associated precision in haplotype frequency estimates. Clayton *et al*. (2004) found that in the dichotomous trait setting, for regions of high LD, the difference in power for haplotype tests of association between the phase known and phase unknown was modest.

We have investigated the power of the omnibus *F*-test for several linear models that might be used to test for association between a genomic region and a quantitative trait. We found that there is no one model that has best power across all underlying genetic mechanisms for influencing the trait, although the joint model that did not require phase information was almost always as powerful as the most powerful model across the scenarios that we considered. This is worth noting as the choice of a modeling strategy is often based, in part, on ease of implementation and interpretation. The joint and interaction models are generally more easily interpreted, and are are simple applications of widely known model parameterization strategies, as compared with the SIMPle and haplotype models. Finally, these results suggest that investigation of a model selection paradigm is warranted, as discussed and implemented by Conti and Gauderman (2004) for a qualitative trait.

Acknowledgements

The authors would like to thank the individuals who volunteered to participate in the IRAS study. This work was supported in part by an American Diabetes Association Junior Faculty Award to TEF.

References

Akey, J. & Xiong, M. 2001: Haplotypes *vs*. single marker linkage disequilibrium tests: what do we gain? — *European Journal of Human Genetics* 9: 291–300.

Bader, J. 2001: The relative power of SNPs and haplotypes as genetic markers for association tests. — *Pharmacogenomics* 2: 11–24.

Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. 2005: Haploview: analysis and visualization of LD and haplotype maps. — *Bioinformatics* 21: 263–265.

Buchanan, T. A., Xiang, A. H., Peters, R. K., Kjos, S. L., Berkowitz, K., Marroquin, A., Goico, J., Ochoa, C. & Azen, S. P. 2000: Response of pancreatic beta-cells to improved insulin sensitivity in women at high risk for type 2 diabetes. — *Diabetes* 49: 782–788.

Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. 2003: Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. — *Human Heredity* 56: 18–31.

Clayton, D., Chapman, J. & Cooper, J. 2004: Use of unphased multilocus genotype data in indirect association studies. — *Genetic Epidemiology* 27: 415–428.

Conti, D. V. & Gauderman, W. J. 2004: SNPs, haplotypes, and model selection in a candidate gene region: the SIMPle analysis for multilocus data. — *Genetic Epidemiology* 27: 429–441.

Cordell, H. J. 2006: Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. — *Genetic Epidemiology* 30: 259–275.

Fallin, D. & Schork, N. J. 2000: Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. — *American Journal of Human Genetics* 67: 947–959.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. & Altshuler, D. 2002: The structure of haplotype blocks in the human genome. — *Science* 296: 2225–2229.

Haffner, S. M., Miettinen, H., Gaskill, S. P. & Stern, M. P. 1995: Decreased insulin secretion and increased insulin resistance are independently related to the 7-year risk

of NIDDM in Mexican-Americans. — *Diabetes* 44: 1386–1391.

Kraft, P. & Hunter, D. 2005: Integrating epidemiology and genetic association: the challenge of gene-environment interaction. — *Philosophical Transactions of the Royal Society B: Biological Sciences* 360: 1609–1616.

Long, A. D. & Langley, C. H. 1999: The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. — *Genome Research* 9: 720–731.

Lillioja, S., Mott, D. M., Spraul, M., Ferraro, R., Foley, J. E., Ravussin, E., Knowler, W. C., Bennett, P. H. & Bogardus, C. 1993: Insulin resistance and insulin secretory dysfunction as precursors of non-insulin-dependent diabetes mellitus. Prospective studies of Pima Indians. — *New England Journal of Medicine* 329: 1988–1992.

Morris, R. W. & Kaplan, N. L. 2002: On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. — *Genetic Epidemiology* 23: 221–233.

Schaid, D. J. 2002: Relative efficiency of ambiguous *vs*. directly measured haplotype frequencies. — *Genetic Epidemiology* 23: 426–443.

Schaid, D. J. 2004: Evaluating associations of haplotypes with traits. — *Genetic Epidemiology* 27: 348–364.

Shoelson, S. E., Lee, J. & Yuan, M. 2003: Inflammation and the IKK beta/I kappa B/NF-kappa B axis in obesity- and diet-induced insulin resistance. — *International Journal of Obesterics and Related Metabolism Disorders* 27, suppl. 3: S49–S52.

Slager, S. L., Huang, J. & Vieland, V. J. 2000: Effect of allelic heterogeneity on the power of the transmission disequilibrium test. — *Genetic Epidemiology* 18: 143–156.

Stephens, M., Smith, N. J. & Donnelly, P. 2001: A new statistical method for haplotype reconstruction from population data. — *American Journal of Human Genetics* 68: 978–989.

Stram, D. O., Leigh Pearce, C., Bretsky, P., Freedman, M., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E. & Thomas, D. C. 2003: Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. — *Human Heredity* 55: 179–190.

Wagenknecht, L. E., Langefeld, C. D., Scherzinger, A. L., Norris, J. M., Haffner, S. M., Saad, M. F. & Bergman, R. N. 2003: Insulin sensitivity, insulin secretion, and abdominal fat: the Insulin Resistance Atherosclerosis Study (IRAS) family study. — *Diabetes* 52: 2490–2496.

Wagenknecht, L. E., Mayer, E. J., Rewers, M., Haffner, S., Selby, J., Borok, G. M., Henkin, L., Howard, G., Savage, P. J., Saad, M. F., Bergman, R. N. & Hamman R. 1995: The insulin resistance atherosclerosis study (IRAS) objectives, design, and recruitment results. — *Annals of Epidemiology* 5: 464–472.

Weyer, C., Bogardus, C., Mott, D. M. & Pratley, R. E. 1999: The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus. — *Journal of Clinical Investigation* 104: 787–794.

Zaykin, D. V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. & Ehm, M. G. 2002: Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. — *Human Heredity* 53: 79–91.