Am. J. Hum. Genet. 61:1189–1199, 1997

Detection of Gene-Environment Interactions in Joint Segregation and Linkage Analysis

W. James Gauderman[1] and Cheryl L. Faucett[2]

[1]University of Southern California and [2]University of California, Los Angeles

Summary

We compare approaches for analysis of gene-environment (G × E) interaction, using segregation and joint segregation and linkage analyses of a quantitative trait. Analyses of triglyceride levels in a single large pedigree demonstrate the two methods and show evidence for a significant interaction ($P = .015$ when segregation analysis is used; $P = .006$ when joint analysis is used) between a codominant major gene and body-mass index. Genotype-specific correlation coefficients, between triglyceride levels and body-mass index, estimated from the joint model are $r_{AA} = .72$, $r_{Aa} = .49$, and $r_{aa} = .20$. Several simulation studies indicate that joint segregation and linkage analysis leads to less-biased and more-efficient estimates of a G × E–interaction effect, compared with segregation analysis alone. Depending on the heterozygosity of the marker locus and its proximity to the trait locus, we found joint analysis to be as much as 70% more efficient than segregation analysis, for estimation of a G × E–interaction effect. Over a variety of parameter combinations, joint analysis also led to moderate (5%–10%) increases in power to detect the interaction. On the basis of these results, we suggest the use of combined segregation and linkage analysis for improved estimation of G × E–interaction effects when the underlying trait gene is unmeasured.

1. Introduction

Many human traits (e.g., blood pressure, lung cancer, and breast cancer) appear to depend on both genetic and environmental factors, possibly interacting. Detection of gene-environment (G × E) interactions is important for several reasons. First, identifying an interaction will increase our understanding of the mechanisms through which the gene and the environmental agent act to control expression of the trait. From a public-health perspective, identifying an interaction will make it possible to target prevention measures at persons who are at particularly high risk. An example is the well-known interaction between phenylalanine exposure and the gene responsible for causing the recessive disorder phenylketonuria, in which dietary restrictions are necessary only for individuals homozygous for the disease gene. From a statistical standpoint, ignoring an existing G × E interaction in an analysis can, erroneously, make the main effects of the gene and the environmental factor appear nonsignificant (Ottman et al. 1990), and thus important risk factors for the trait may be overlooked. Finally, failing to model a G × E interaction in a segregation analysis can lead to incorrect conclusions with respect to determination of the mode of inheritance (Tiret et al. 1993) and estimation of the magnitude of genetic effects and allele frequencies (Eaves 1984).

In the context of pedigree studies, there are two primary methods for examination of G × E interactions. The first is to stratify the sample of pedigrees into two or more groups, on the basis of one of the factors (e.g., environmentally exposed vs. unexposed) and then, within each stratum, to analyze the relationship between the other factor (e.g., the gene) and the trait. This method relies on the ability to classify an entire pedigree as exposed or unexposed, which will not be feasible for many types of exposures (e.g., sex, if gene-sex [G × S] interaction is of interest). Furthermore, reducing the sample size through subsetting the data will result in lower statistical power for identification of interactions than will a single analysis using the entire data set.

The second method for analysis of G × E interactions is to directly model them in a segregation-analysis framework, an approach that has been used by several investigators (Moll et al. 1984; Gueguen et al. 1989; Rebbeck et al. 1989; Konigsberg et al. 1991; Gauderman et al. 1997). For example, using the 337 lung cancer pedigrees analyzed by Sellers et al. (1990, 1992), Gauderman et al. (1997) specified a proportional-hazards model for the joint effect of smoking and a major gene and for their

Received January 2, 1997; accepted for publication September 2, 1997; electronically published October 29, 1997.

Address for correspondence and reprints: Dr. W. James Gauderman, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033. E-mail: jimg@rcf.usc.edu

0002-9297/97/6105-0026-$02.00

interaction on lung cancer risk. Direct modeling of interactions is commonly done in the analysis of measured environmental factors from epidemiological studies, and several software packages are available for this type of analysis (e.g., SAS and BMDP). These packages can also be used to analyze interactions between environmental factors and measured genes. However, if the gene is unmeasured, some form of segregation analysis is typically used, incorporating the interaction in the penetrance function for the trait. Direct modeling of interactions has the advantage that the entire data set can be used in a single analysis.

It is well known that detecting an interaction between two measured covariates is less powerful than detecting each component main effect (Breslow and Day 1987). Power for detecting a G × E interaction when the trait gene is unmeasured will depend in part on the underlying trait-gene distribution for each person. In a segregation-analysis setting, this distribution depends on the pattern of trait phenotypes in the family and can be quite vague in the case of a multifactorial trait. The addition of a linked marker in a joint segregation and linkage analysis may provide additional information about the person-specific trait-gene distributions and thus may lead to an increase in power for detection of G × E interactions.

A general regressive model for joint segregation and linkage analysis was proposed by Bonney et al. (1988), and analysis using this model has been used to investigate a variety of traits (Tiret et al. 1992; Martinez et al. 1995; Craig et al. 1996). Software has been developed for fitting Bonney et al.'s model to allow it to incorporate gene-covariate interactions (Demenais and Lathrop 1994). Markov-chain Monte Carlo methods for joint segregation and linkage analysis have also been developed (Guo and Thompson 1992; Thomas and Cortessis 1992).

In this paper, we will use simulation studies to compare the efficiency for estimation and the power for detection of G × E interactions in segregation analysis alone versus the efficiency and power obtained with use of joint segregation and linkage analysis. We will investigate relative efficiency (RE) and power over a variety of inheritance modes, parameter values, and data structures. We will also compare the efficiency and power from these approaches versus what would have been obtained if the trait gene could be measured, as a barometer of the loss in power that we can expect compared with the optimal case.

In section 2, we describe the models and assumptions underlying the analyses. Section 3 summarizes a real-data analysis of a G × E interaction for triglyceride (Tg) levels in a single large pedigree. The results demonstrate the difference, in estimates and hypothesis tests, that can arise from the use of the different analytic techniques, which will motivate the simulation studies. Simulation methods and results are described in section 4, and concluding remarks are given in section 5.

2. Models

Let $i = 1, ..., I$ index the set of subjects and let $f = 1, ..., F$ index the set of families in a given data set. The information collected for each subject includes a trait phenotype Y_i, one or more measured covariates Z_i, and a marker phenotype M_i. Any of these data may be missing for some subjects. We assume that M_i is determined by a fully penetrant gene (m_i) with an arbitrary number of alleles and corresponding allele frequencies q_m. We further assume that Y_i depends on a single diallelic partially penetrant major gene g_i with alleles A and a and population frequency q_A. We define G_i to be a "genetic covariate" based on genotype g_i and an assumed mode of inheritance. For example, under dominant inheritance, $G_i = 1$ if g_i = AA or Aa, and $G_i = 0$ otherwise. Let $\mathbf{X}_i$ denote the vector of covariates for subject i, including Z_i, G_i, and, possibly, their interaction(s).

We concentrate on a continuous trait and consider a penetrance function based on the linear model: $Y_i = \alpha + \beta'\mathbf{X}_i + \epsilon_i$, where α is the intercept and β is a vector of regression coefficients corresponding to the covariates in the model. If we assume that the random errors ϵ_i, $i = 1, ..., I$, are independent and normally distributed with mean 0 and variance σ_e^2, then the penetrance function for the ith subject is the normal density

$$f(Y_i|\Omega) = \frac{1}{\sigma_e\sqrt{2\pi}} \exp[-(Y_i - \alpha - \beta'\mathbf{X}_i)^2/(2\sigma_e^2)] \ ,$$

where $\Omega = \{\alpha, \beta, \sigma_e^2\}$ is the set of penetrance-model parameters. An assumption in this model is that the major gene accounts for all of the intrafamily correlation in the trait phenotype. This restriction can be relaxed to account for additional intrafamily dependence—for example, by use of the regressive model (Bonney 1984; Bonney et al. 1988) or the mixed model (Morton and Maclean 1974).

If the major gene can be observed, the likelihood is simply the product of the individual-specific penetrance functions; that is,

$$L(\Omega) = \prod_i f(Y_i|\Omega) \ .$$

When the major gene is unobserved, the likelihood is formed by summing over all possible combinations of joint genotypes in each family. In a segregation-analysis setting, the likelihood for family f is given by

$$L_f(\Omega, q_A) = \sum_{g_f} [P(g_f|q_A) \prod_{i \in f} f(Y_i|\Omega)] \ , \qquad (1)$$

where the first factor is a function of the trait-allele frequency for founders and of Mendelian-transmission probabilities for nonfounders. The total likelihood for all pedigrees is the product of these family-specific contributions; that is, $L(\Omega, q_A) = \prod_f L_f(\Omega, q_A)$. In a joint segregation and linkage analysis, the family-specific likelihood is written as

$$L_f(\Omega, q_A, q_m, \theta) = \sum_{g_f, m_f} [P(g_f, m_f | q_A, q_m, \theta) \prod_{i \in f} f(Y_i|\Omega) P(M_i|m_i)] , \quad (2)$$

where the first factor now also depends on the marker-allele frequencies (q_m) for founders and on the recombination fraction (θ) for nonfounders. The factor $P(M_i|m_i)$ is the marker-penetrance function, which is assumed to be 1 for all m_i consistent with M_i and to be 0 otherwise (Bonney et al. 1988).

Computation of the likelihoods in equations (1) and (2) is accomplished by use of the peeling algorithm (Elston and Stewart 1971; Lange and Elston 1975). To test for G × E interaction, these likelihoods can be maximized both with and without a G × E–interaction term in the penetrance model, and the likelihood-ratio test can be used to test the null hypothesis that the corresponding regression coefficient is 0. Alternatively, a Wald test can be computed as the estimated interaction coefficient divided by its standard error. All analyses reported in this paper were performed by use of the Genetic Analysis Package (1997).

3. A Motivating Example

We present an analysis of G × E interaction, using the HGAR1 pedigree distributed to the 8th Genetic Analysis Workshop (Bailey-Wilson and Elston 1993). A diagram of this pedigree has been given by Olshen and Wijsman (1996). The pedigree consists of 232 subjects and includes measurements of blood lipid levels, including high-density lipids, low-density lipids, Tg (in mg/dl), and total serum cholesterol. Also collected were age, height (in inches), and weight (in pounds). Phenotype and covariate data are available for 190 (82%) of the subjects. Previous analyses of these data (Bailey-Wilson et al. 1993) showed evidence for linkage of Tg to the marker KM on chromosome 2p. Sznajd et al. (1989) analyzed a different data set and showed that Tg level was positively correlated with body-mass index (BMI), although the magnitude of the correlation was fairly weak. We will present segregation and joint segregation and linkage analyses aimed at testing whether there is an interaction between a major gene and BMI, in their effect on Tg.

The distribution of Tg in these data is statistically nonnormal when the Kolmogorov test is used, and so, as others have done, we consider instead the square root of Tg ($Y = \sqrt{\text{Tg}}$). A standard segregation analysis of $\sqrt{\text{Tg}}$ was performed by fitting the general "ousiotype" model (Cannings et al. 1978) and comparing several nested alternatives, by use of likelihood-ratio tests. The main effects of age and BMI were included in the penetrance model. The alternative transmission models tested were the Mendelian codominant, dominant, and recessive models; the sporadic model; and the pure environmental model (i.e., a single discrete type with frequency q_A, identically distributed among all subjects). The Mendelian dominant ($P = .002$), Mendelian recessive ($P = .0006$), sporadic ($P < .0001$), and environmental ($P = .006$) models all fit the data significantly worse than did the general model. However, the Mendelian codominant model could not be rejected ($P = .12$) and was the most parsimonious by Akaike's information criterion.

On the basis of these findings, we used a linear regression of the following form, to model gene-BMI interaction (G × BMI):

$$\begin{aligned} Y_i = {} & \alpha + \beta_{\text{Age}}(\text{Age}_i) + \beta_{\text{BMI}}(\text{BMI}_i) + \beta_{\text{AA}}(G_{\text{AA}_i}) \\ & + \beta_{\text{Aa}}(G_{\text{Aa}_i}) + \beta_{\text{AA}\times\text{BMI}}(G_{\text{AA}_i} \times \text{BMI}_i) \\ & + \beta_{\text{Aa}\times\text{BMI}}(G_{\text{Aa}_i} \times \text{BMI}_i) + \epsilon_i , \end{aligned} \quad (3)$$

where G_{AA_i} is 1 if g_i is AA and is 0 otherwise, and similarly for G_{Aa_i}. BMI was computed as 100(weight/height^2), and both age and BMI were standardized by subtracting their sample means (28.5 and 3.1, respectively).

On the basis of the model in equation (3), the genotype-specific variance of Y for a person of a specific age is given by

$$\sigma^2_{Y|g} = \begin{cases} \sigma_e^2 + (\beta_{\text{BMI}} + \beta_{\text{AA}\times\text{BMI}})^2 \sigma^2_{\text{BMI}} & \text{if } g = \text{AA} \\ \sigma_e^2 + (\beta_{\text{BMI}} + \beta_{\text{Aa}\times\text{BMI}})^2 \sigma^2_{\text{BMI}} & \text{if } g = \text{Aa}, \\ \sigma_e^2 + \beta^2_{\text{BMI}} \sigma^2_{\text{BMI}} & \text{if } g = \text{aa} \end{cases}$$

where σ^2_{BMI} is the variance of BMI. Genotype-specific correlation coefficients between Y and BMI are given by

$$\rho_g = \sqrt{\frac{\sigma^2_{Y|g} - \sigma_e^2}{\sigma^2_{Y|g}}} .$$

The maximum-likelihood estimator r_g of ρ_g is obtained by substituting the corresponding maximum-likelihood estimates of the variances and regression coefficients into the equations above.

Table 1 shows parameter estimates and standard errors from a segregation analysis using the codominant

Table 1

Main Effects and Interaction Models for Tg, With and Without Linkage to Marker KM, in HGAR1 Pedigree

	Maximum-Likelihood Estimate (Standard Error)			
	Segregation[a]		Joint Segregation and Linkage[b]	
	Model 1 (Main Effects)	Model 2 (Interaction)	Model 3 (Main Effects)	Model 4 (Interaction)
α	7.677 (.237)	7.616 (.228)	7.638 (.260)	7.507 (.237)
β_{Age}	.026 (.009)	.026 (.009)	.025 (.009)	.023 (.010)
β_{BMI}	.903 (.218)	.491 (.366)	.900 (.218)	.364 (.368)
β_{AA}	5.067 (.589)	4.690 (.447)	5.097 (.574)	4.719 (.465)
β_{Aa}	1.507 (.320)	1.419 (.305)	1.534 (.311)	1.484 (.305)
$\beta_{AA \times BMI}$	…	1.232 (.461)	…	1.411 (.448)
$\beta_{Aa \times BMI}$	…	.411 (.516)	…	.584 (.466)
σ_e^2	1.707 (.264)	1.585 (.250)	1.675 (.266)	1.504 (.241)
q_A	.315 (.077)	.349 (.071)	.318 (.075)	.362 (.067)
q_B	…	…	.383 (.060)	.383 (.060)
θ_{KM}	…	…	.183 (.282)	.100 (.113)
Log likelihood	−378.28	−374.06	−458.16	−452.97

[a] $\chi^2 = 8.44$; $P = .015$ (obtained from likelihood-ratio test of null hypothesis that both interaction coefficients are zero).
[b] $\chi^2 = 10.38$; $P = .006$ (obtained from likelihood-ratio test of null hypothesis that both interaction coefficients are zero).

model, both without (model 1) and with (model 2) a G × BMI interaction. The likelihood-ratio test, obtained by comparison of the log likelihoods from models 1 and 2, gives $\chi^2_{(2)} = 8.44$ ($P = .015$), indicating a significant G × BMI interaction. Genotype-specific correlation coefficients, based on model 2 estimates, are $r_{AA} = .70$, $r_{Aa} = .46$, and $r_{aa} = .27$.

We reanalyzed the data, using joint segregation and linkage analysis, to include estimation of θ between the trait locus and the marker KM. This marker is diallelic, with allele B dominant to allele b. The allele frequency q_B was also estimated in these analyses. Table 1 shows parameter estimates from this joint analysis, without (model 3) and with (model 4) the G × BMI interaction term. The estimated θ between g and KM is $\hat{\theta} = .18$ in the model without an interaction (model 3) and is $\hat{\theta} = .10$ when the interaction is included (model 4). The estimated interaction coefficients are larger in the joint analysis ($\hat{\beta}_{AA \times BMI} = 1.41$, and $\hat{\beta}_{Aa \times BMI} = .58$; model 4) than in the segregation analysis alone ($\hat{\beta}_{AA \times BMI} = 1.23$, and $\hat{\beta}_{Aa \times BMI} = .41$; model 2). The likelihood-ratio test of the interaction coefficient from joint models 3 and 4 gives $\chi^2_{(2)} = 10.38$ ($P = .006$), which is larger than the χ^2 statistic based on segregation models 1 and 2.

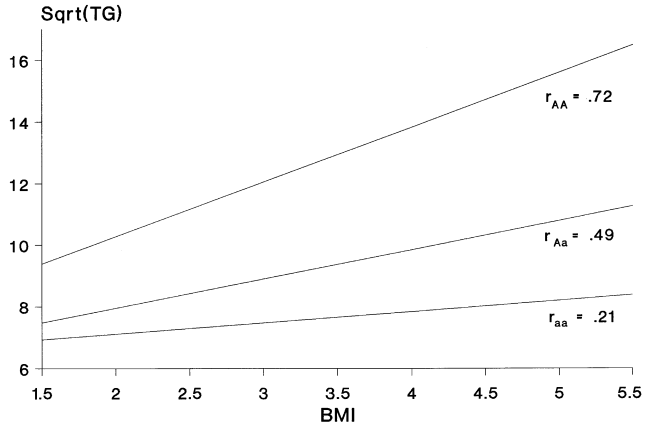

Figure 1 Genotype-specific regressions of square-root Tg (in mg/dl) on BMI (in lbs/100 in^2), for joint segregation and linkage analysis, on the basis of estimates shown in table 1, model 4.

Figure 1 graphically depicts the G × BMI interaction on the basis of the estimated coefficients from model 4. Also included in figure 1 are the corresponding estimated genotype-specific correlation coefficients between $\sqrt{Tg}$ and BMI that are based on this model. The overall Pearson correlation coefficient between $\sqrt{Tg}$ and BMI in these data is $r = .39$. The strongest genotype-specific correlation is predicted for homozygous carriers ($r_{AA} = .72$), estimated to represent approximately $q_A^2 = 13\%$ of the population. The model predicts moderate correlation for heterozygotes ($r_{Aa} = .49$; 46% of the population) and weak correlation for homozygous noncarriers ($r_{aa} = .21$; 41% of the population).

Both the difference, in G × E–interaction estimates, between the two types of analyses and the larger χ^2 statistic obtained in the joint analysis motivated us to investigate more generally the effect that inclusion of a linked marker had on the precision and efficiency of G × E–interaction estimates and on the power to detect an interaction.

4. Simulation Studies

We performed three sets of simulations to determine whether adding a linked marker to a segregation analysis leads to improved efficiency for estimating—and power

for detecting—a G × E interaction. In the first simulation, we investigated the effect of adding a linked marker while varying both the true strength of the interaction and the θ between the trait and marker loci. In the second study, we assessed how the level of heterozygosity (H) at the marker locus affects the relative performance of joint segregation and linkage analysis, compared with segregation analysis alone. In the third simulation study we examined the sensitivity of our findings to the inheritance mode and allele frequency at the trait locus and to the types of pedigree structures being analyzed.

4.1. *Data Generation and Analysis*

In each simulation study, 200 independent replicate data sets were generated. Each data set consisted of a single 232-member pedigree with the same structure as the HGAR1 pedigree described above. Genotypes at a trait locus (g) and at a fully codominant marker locus (m) were simulated for each subject, under the assumption of Hardy-Weinberg and linkage equilibrium. Trait and marker genotypes were randomly sampled for founders, conditional on the assumed allele frequencies, and then were "dropped" through the pedigree, according to Mendelian-transmission rules and the assumed θ.

An environmental covariate (Z) was generated randomly and independently for each subject, from the standard normal distribution. Conditional on the assigned trait genotype and covariate, a continuous phenotype was randomly sampled for each subject i from the normal distribution, with mean $\mu_Y = \beta_Z Z_i + \beta_G G_i + \beta_{GZ}(G_i * Z_i)$, where G codes for dominant inheritance and where the error variance σ_e^2 is set to 1.0. Trait and covariate data for subjects with missing data in the original HGAR1 pedigree ($n = 42$) were considered as missing in the simulated data sets. Marker phenotypes, however, were simulated for all subjects, since analysis of replicate data sets with missing marker data would have been computationally infeasible.

According to the logic of Tiret et al. (1993), the total variance of Y can be written as

$$\begin{aligned}\sigma_Y^2 &= \beta_Z^2\mathrm{Var}(Z) + \beta_G^2\mathrm{Var}(G) \\ &\quad + \beta_{GZ}[\beta_{GZ}\mathrm{Var}(GZ) + 2\beta_Z\mathrm{Cov}(Z, GZ)] + \sigma_e^2 \\ &= \sigma_Z^2 + \sigma_G^2 + \sigma_{GZ}^2 + \sigma_e^2 \ ,\end{aligned}$$

where the terms in the latter summation are the variances due to the environmental covariate, the major gene, the interaction, and the random error, respectively. Dividing each of the component variances by σ_Y^2 yields the percent of variance explained by each term in the model. In all simulations, σ_Z^2/σ_Y^2 was set to 10%, and σ_G^2/σ_Y^2 was set to 25%.

Three types of analysis were performed on each data set: (1) segregation analysis ignoring the marker locus, (2) joint segregation and linkage analysis estimating all model parameters except the marker allele frequencies, and (3) linear regression treating the simulated trait genotypes as if they were observed. For each analysis, maximum-likelihood estimates and the corresponding values of the log likelihood were obtained both with (L_1) and without (L_0) the G × E term in the penetrance model. A likelihood-ratio test was then computed as $\chi^2 = -2(L_0 - L_1)$, which is approximately χ^2 distributed with 1 df under the null hypothesis of no interaction. The empirical power for detecting the G × E interaction in each type of analysis was computed as the percent of data sets in which $\chi^2 > 3.84$—that is, significance at the .05 level, for a two-sided test. We also report the efficiency of each type of analysis, relative to segregation analysis, computed as the ratio of the estimated mean squared errors for the interaction parameter, β_{GZ}. An RE >1.0 for either joint segregation and linkage or measured gene analysis indicates greater efficiency for estimation of β_{GZ}, compared with segregation analysis. Finally, we report the mean of $\hat{\beta}_{GZ}$ across replicates and the bias in these estimates as a percent of the true value.

4.2. *Simulation 1: The Effect of the Interaction Strength and θ*

In this simulation, we assume, at the trait locus, dominant inheritance and, to maximize heterozygosity, true allele frequency $q_A = .50$. We vary the interaction variance to reflect three situations: strong ($\sigma_{GZ}^2/\sigma_Y^2 = 25\%$), weak ($\sigma_{GZ}^2/\sigma_Y^2 = 15\%$), and no ($\sigma_{GZ}^2/\sigma_Y^2 = 0\%$) G × E interaction. Table 2 shows the parameter values for the three generating models and also shows the corresponding genotype-specific correlation coefficients between Y and Z. We consider a diallelic marker locus with true allele frequency $q_B = .50$ (marker $H = .50$) and vary θ to reflect loose ($\theta = .20$) and tight ($\theta = .001$) linkage with the trait locus.

Simulation results for each model condition and method of analysis are summarized in table 3. For a strong interaction effect, estimates from segregation analysis were biased upward (by 5.7% on average), whereas estimates from joint analysis were less biased, for both loosely linked (4.3% on average) and tightly

Table 2

Parameter Values for Three Data-Generation Models

Strength of Interaction	σ_{GZ}^2/σ_Y^2	σ_Y^2	β_Z	β_G	β_{GZ}	$\rho_{AA,Aa}$	ρ_{aa}
Strong	.25	2.5	.50	1.83	.541	.72	.45
Weak	.15	2.0	.45	1.63	.327	.61	.41
None	.00	1.5	.39	1.43	.000	.37	.37

Table 3

Estimation and Detection of G × E Interaction, from Segregation, Joint Segregation and Linkage, and Observed-Gene Analysis, with Use of Diallelic Marker with $H = .50$, from 200 Replicates of Simulated Data

Interaction Strength (True β_{GZ}) and Parameter	Segregation	Joint Segregation and Linkage for $\theta =$		Observed-Gene Analysis
		.2	.001	
Strong (.541):				
$\hat{\beta}_{GZ}$ (Bias)[a]	.572 (5.7%)	.564 (4.3%)	.552 (2.1%)	.538 (−.6%)
RE[b]	1.0	1.14	1.26	2.69
Power[c]	69.5%	71.0%	74.5%	86.0%
Weak (.327):				
$\hat{\beta}_{GZ}$ (Bias)[a]	.360 (10.1%)	.351 (7.3%)	.347 (6.1%)	.323 (−1.2%)
RE[b]	1.0	1.04	1.34	2.50
Power[c]	30.0%	33.0%	32.5%	46.0%
None (.000):				
$\hat{\beta}_{GZ}$ (Bias)[a]	.002 (ND)	−.010 (ND)	.019 (ND)	−.007 (ND)
RE[b]	1.0	1.63	2.18	5.23
Power[c]	6.0%	5.5%	3.5%	4.5%

[a] Average estimate (bias as a percent of the true value) across 200 data replicates. ND = not defined.
[b] Mean squared error from segregation analysis, divided by mean squared error from other analysis.
[c] Percent of data replications in which H_0:$\beta_{GZ} = 0$ is rejected at the 5% significance level when the likelihood-ratio test is used.

linked (2.1% on average) markers. Estimates of efficiency relative to segregation analysis were higher for joint analysis with both a loosely linked marker (RE 1.14) and a tightly linked marker (RE 1.26). Power was also increased, from 69.5% in segregation analysis to 71.0% in joint analysis with a loosely linked marker and to 74.5% in joint analysis with a tightly linked marker. However, the efficiency and power in a joint analysis are still substantially less than those in the optimal case of observation of the trait gene (RE 2.69, power 86%).

When the interaction effect was weak, the upward bias in the segregation estimate was 10.1% on average, compared with 7.3% and 6.1% in joint analysis with a loosely linked and a tightly linked marker, respectively. Joint analysis with a tightly linked marker led to 34% greater efficiency, compared with segregation analysis. Under the null case of no interaction ($\beta_{GZ} = 0$), all methods produced nearly unbiased estimates, on average, and empirical powers within sampling variability of the nominal 5% significance level. However, efficiency was still improved with the inclusion of either a loosely linked (RE 1.63) or a tightly linked (RE 2.18) marker.

4.3. *Simulation 2: The Effect of* H *at the Marker Locus*

In this step of the study, we focused on the strong-interaction, tight-linkage model described in simulation 1. We performed several simulations, considering different H levels (.10–.95). $H < .50$ was achieved by simulating a diallelic marker with $q_B < .50$. More-informative markers were obtained by increasing the number of alleles at the marker locus (to 3, 5, 10, or 20) and assuming that all alleles were equally frequent in the population.

Table 4 shows percent bias, RE, and power, for various numbers of alleles (N) and H values at the marker locus. Estimates of bias from all joint analyses were less than those from segregation analysis and tended to decrease with increasing H. The estimated RE was 1.10 when $H = .10$ and generally increased with increasing

Table 4

Percent Bias, RE, and Power for a Strong Interaction Effect, Based on N and H at Marker Locus, from 200 Replicates of Simulated Data

		Joint Segregation and Linkage ($\theta = .001$)								Observed-Gene Analysis
		$N = 2$				$N = 3$,	$N = 5$,	$N = 10$,	$N = 20$,	
	Segregation	$H = .10$	$H = .20$	$H = .35$	$H = .50$	$H = .67$	$H = .80$	$H = .90$	$H = .95$	
Bias[a]	5.7%	3.8%	2.6%	2.4%	2.1%	2.1%	.8%	.7%	.9%	−.6%
RE[b]	1.00	1.10	1.19	1.28	1.26	1.37	1.49	1.68	1.72	2.69
Power[c]	69.5%	71.5%	74.0%	74.0%	74.5%	73.5%	74.5%	79.0%	78.0%	86.0%

[a] Mean bias as a percent of the true value (.541), across 200 replicate data sets.
[b] Mean squared error from segregation analysis, divided by mean squared error from other analysis.
[c] Percent of data replications in which H_0:$\beta_{GZ} = 0$ is rejected at the 5% significance level when the likelihood-ratio test is used.

H, to a high of 1.72 when $H = .95$. The power for detecting the interaction in joint analysis ranged from 71.5%, when $H = .10$, to 79%, when $H = .90$, compared with 69.5% in segregation analysis alone. Figure 2 graphically shows a comparison of mean squared errors from the joint analyses and segregation analysis ($H = 0$). The simulated conditions (denoted by asterisks) are connected to show the inferred mean squared errors over the range of H values. The addition of any linked marker reduces the mean squared error, compared with that for segregation analysis alone. However, even for a very informative marker ($H = .95$), the mean squared error is substantially higher than that for the optimal case of regression analysis of a measured trait gene.

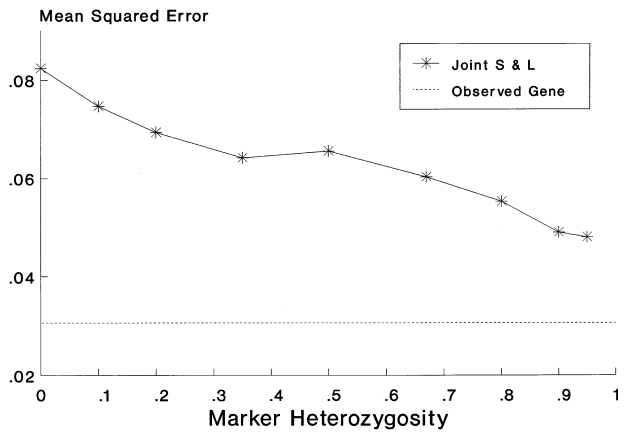

Figure 2 Estimated mean squared error, by H at marker locus, for a strong interaction effect, from segregation analysis ($H = 0$) and joint segregation and linkage analysis (Joint S & L), with a tightly linked marker ($\theta = .001$). The broken line gives the estimated mean squared error for the optimal case of analysis of a measured-trait gene.

4.4. *Simulation 3: The Effect of Varying Trait Models and Data Structures*

The simulation design described above considers a trait segregating in a single large pedigree, according to a dominant-inheritance model with allele frequency $q_A = .50$. We investigated the sensitivity of the findings to these assumptions, by performing several additional simulations. For the dominant model, we considered two additional allele frequencies, $q_A = .30$ and $q_A = .10$. We also simulated recessive inheritance, with $q_A = .70$ and $q_A = .50$, and additive inheritance, with $q_A = .50$ and $q_A = .30$. Each of these models was simulated in the single large pedigree described above. Additionally, each model was simulated in the pedigrees of the Berkeley Lipid Data Set, which was also distributed to participants of the 8th Genetic Analysis Workshop (Krauss et al. 1993). This data set includes 420 subjects in 27 pedigrees with 5–120 members (mean 14.2 members). We eliminated 36 childless spouses of family members, leaving 384 subjects. In the real data, 141 subjects are missing either lipid measurements or BMI; these subjects are considered to have missing data in the simulations as well.

In all of the sensitivity simulations, we generated data according to the regression coefficients of the strong-interaction model (table 2). Thus, within an inheritance model, the genotype-specific regressions of Y on Z are the same, but the proportion of subjects in each genotype group differ, depending on the assumed allele frequency. Across inheritance models, the definition of G differs; that is, $G = 1$ for $g = AA$ and is 0 otherwise in the recessive model, and $G = 1$ for $g = AA$, $G = .5$ for $g = Aa$, and is 0 otherwise in the additive model.

Table 5 shows the percent bias in estimates of the interaction effect in all the models and data structures considered. In almost all cases, estimates from joint analysis were less biased than those from segregation analysis. As an example, segregation-analysis estimates for the additive model were upwardly biased by ~10% in the case of a single large pedigree and by ~25% in the case of mixed-size pedigrees. The addition of a linked marker in a joint analysis virtually eliminated this bias for the large pedigree and reduced it 4% for mixed-size pedigrees. Table 6 shows estimates of power to detect the interaction, in all the simulations. Adding a linked marker did not improve power in the dominant model with $q_A = .30$ but did improve power in all other models. Power gains were modest, 1.5%–7%. In some of the models' conditions, power was already high when segregation analysis was used, leaving little room for improvement. For a given mode of inheritance and allele frequency, power with the mixed-size pedigrees was always higher than with the single large pedigree, because of the larger sample size in this data set.

Figure 3*a* and figure 3*b* show the relative efficiencies of joint and observed-gene analysis, respectively, compared with segregation analysis. In all models and data structures, adding a linked marker improved efficiency (fig. 3*a*), from a 5% increase (for the additive model, $q_A = .5$, mixed-size pedigree) to a 40% increase (for the dominant model, $q_A = .30$, single large pedigree). For both the single large pedigree and the mixed-size pedigrees, the largest improvement was observed under the dominant model, and the smallest was observed under the additive model. In general, efficiency gains for joint analysis were greater with a single large pedigree than with several mixed-size pedigrees. This is consistent with an overall finding of the 10th Genetic Analysis Workshop—that large pedigrees provide more linkage information per subject than do several smaller pedigrees, all other conditions being equal (Wijsman and Amos, in press).

Relative efficiencies when the gene was observed (fig.

Table 5

Percent Bias in Estimates of Strong G × E Interaction, from Segregation, Joint Segregation and Linkage, and Observed-Gene Analysis, for Various Modes of Inheritance, Trait-Allele Frequencies, and Data Structures

Mode of Inheritance and q_A	Bias in Estimates of Strong G × E Interaction (%)					
	Single Large Pedigree			Mixed-Size Pedigrees		
	Segregation	Joint Segregation and Linkage[a]	Observed-Gene Analysis	Segregation	Joint Segregation and Linkage[a]	Observed-Gene Analysis
Dominant:						
.10	4.9	2.1	−2.0	2.0	.4	.2
.30	.0	−1.0	−2.1	4.9	3.2	2.9
Additive:						
.30	10.1	.0	−5.8	27.6	23.1	5.4
.50	9.5	2.1	−2.6	24.4	19.9	4.4
Recessive:						
.50	−1.9	−7.9	−3.3	11.7	8.4	3.8
.70	2.9	−.9	.2	6.2	3.2	1.3

[a] Simulated marker is diallelic, with $H = .50$ and true $\theta = .001$.

3*b*) ranged from 1.7 (for the dominant model, $q_A = .30$, mixed pedigrees) to 5.3 (for the additive model, $q_A = .30$, mixed pedigrees). Relative efficiencies were larger in the additive model than in the dominant or recessive model, indicating that, by segregation analysis, obtaining accurate estimates for an additive locus may be more difficult than doing so for a dominant or recessive locus.

5. Discussion

In this paper, we have demonstrated that including a linked marker in a joint segregation and linkage analysis leads both to less bias and increased efficiency for estimation of G × E–interaction effects and to greater power for detection of interactions, compared with segregation analysis alone. Improvements in efficiency and power are greater with a closely linked marker, compared with a loosely linked marker, and with an increase of H at the marker locus. These findings are consistent with the hypothesis that addition of linkage information reduces variance in the distribution of the unmeasured trait gene within a family, leading to more-precise estimates of its effect. Sensitivity analyses indicated that improvements, in power and RE, by use of joint analysis can be expected over a range of inheritance modes, trait-allele frequencies, and pedigree structures. Including even a fairly uninformative marker ($H = .10$) led to improvements in power and efficiency, compared with segregation analysis alone. This indicates that the analyst can utilize any linked marker and expect some gains in efficiency; he or she does not necessarily have to focus on highly polymorphic markers that may be computationally time consuming to analyze if there are missing data.

Table 6

Power for Detection of G × E Interaction, from Segregation, Joint Segregation and Linkage, and Observed-Gene Analysis, for Various Modes of Inheritance, Trait-Allele Frequencies, and Data Structures

Mode of Inheritance and q_A	Power for Detection of G × E Interaction (%)					
	Single Large Pedigree			Mixed-Size Pedigrees		
	Segregation	Joint Segregation and Linkage[a]	Observed-Gene Analysis	Segregation	Joint Segregation and Linkage[a]	Observed-Gene Analysis
Dominant:						
.10	59	61	75.5	68.5	72	88
.30	84	84	93.5	95	95	99.5
Additive:						
.30	34.5	37	55.5	45	46.5	80
.50	36.5	43.5	71	57.5	60	86.5
Recessive:						
.50	60	62	80.5	79.5	81	92.5
.70	80.5	84	94	92.5	95	99.5

[a] Simulated marker is diallelic, with $H = .50$ and true $\theta = .001$.

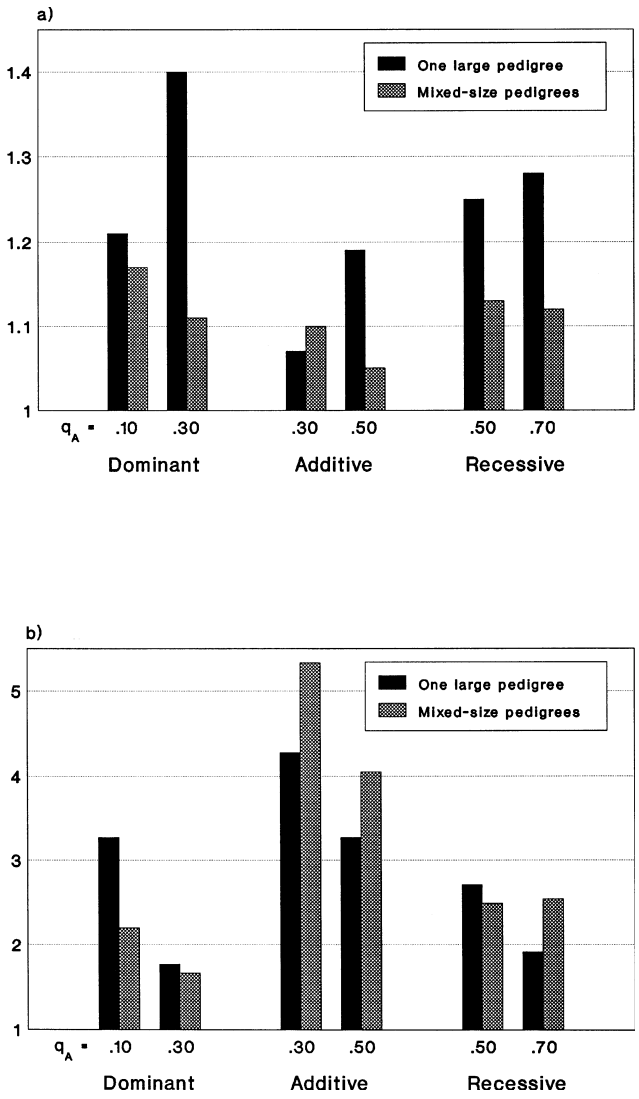

Figure 3 RE for estimation of strong G × E–interaction effect, by use of joint segregation and linkage analysis (*a*) or observed-gene analysis (*b*), compared with segregation analysis, for varying modes of inheritance, trait-allele frequencies, and data structures.

The focus of this paper was on estimation and detection of G × E interactions. An equally important issue is the effect of ignoring a G × E interaction in the detection of linkage and in the estimation of θ. In a general way, failure to model an important interaction is a form of penetrance-model misspecification and thus, on the basis of the results of Clerget-Darpoux et al. (1986), should lead to both reduced power to detect linkage and biased estimates of the θ. Towne et al. (in press) have analyzed pedigree data that were simulated with a G × S interaction effect and have found that ignoring the interaction did lead to lower LOD scores and to reductions in power to detect linkage, compared with inclusion of the interaction in the penetrance model. G × E interactions are likely to be important in the etiology of many complex traits and should be carefully considered in linkage studies of such traits.

Our simulation findings were based on analysis of a quantitative trait, and, although it is likely that the results are generalizable to disease and other qualitative outcomes, additional research is required. If pedigrees are sampled at random from the population, as is often the case when quantitative traits are of primary interest, both trait and linkage model parameters can be jointly estimated without ascertainment correction. However, for disease traits, families are typically sampled on the basis of the status of one or more of their members (probands), thus requiring an ascertainment correction if one is to obtain unbiased estimates of the disease allele-frequency and penetrance parameters. Hsu et al. (in press) have proposed, for disease outcomes, a class of population-based study designs that facilitate ascertainment correction in the context of joint segregation and linkage analysis. In some cases, though, when linkage analysis is the primary goal of the study, heavily disease-loaded families are collected, making ascertainment correction impossible and thus precluding the use of joint analysis.

In a typical progression of study, segregation analysis is first used to determine the mode of inheritance at the trait locus and to estimate corresponding penetrances and allele frequencies. Linkage analysis is then used to localize the trait gene, fixing the trait model parameters to their maximum-likelihood estimates from the segregation analysis. Several authors have demonstrated that increased power for detection of linkage derives from joint estimation of the trait and linkage model parameters (Tiret et al. 1992; Martinez et al. 1995; Craig et al. 1996; Gauderman et al., in press). Similarly, the results of the present study suggest that, once linked markers are identified, the analyst can use this information to improve estimates of the interaction and, possibly, other trait model parameters. Alternatively, joint segregation and linkage analysis can be used from the start, providing a unified and more powerful framework for estimating the effects of trait genes and for finding their locations.

In the past, the increased computational demands of joint segregation and linkage analysis may have been a barrier to its general use. However, the current availability of affordable fast computers reduces the importance of this factor. For example, each joint segregation and linkage analysis of the HGAR1 data required <2 min on a Pentium 166 personal computer, making comparison of several alternative models feasible. If computation of the joint likelihood is simply infeasible because of model or pedigree complexity (e.g., for an inbred pedigree), Monte Carlo techniques (Guo and

Thompson 1992; Thomas and Cortessis 1992; Faucett et al. 1993; Gauderman et al. 1995) can be applied.

In our analysis of Tg levels in the HGAR1 pedigree, we found a significant interaction between a major gene and BMI. A joint analysis, including a loosely linked ($\theta = .10$) diallelic marker, KM, that was missing in 18% of subjects, led to both reduced variance in the estimated G × BMI–interaction effect and a larger χ^2 statistic for the interaction-hypothesis test, compared with segregation analysis alone. Our estimates of the interaction effect indicate that, for the majority of the population, there is a weak-to-moderate correlation between BMI and Tg levels but that, for ~13% of subjects homozygous at a major locus, there is a strong correlation ($r = .72$). It is premature to apply this result from a public-health perspective, since determination of genotypic status is not feasible at this time. However, if a major gene that influences Tg levels is discovered, its interaction with BMI should be considered, and, if it is still indicated, dietary and other measures aimed at control of obesity could be targeted at actual or suspected gene carriers.

Acknowledgments

This work was supported by National Institute of Health grants CA 52862 and CA 58860.

References

Bailey-Wilson JE, Elston RC (1993) The HGAR1 familial hypercholesterolemia pedigree. Genet Epidemiol 10:529–532

Bailey-Wilson JE, Wilson AF, Bamba V (1993) Linkage analysis in a large pedigree ascertained due to essential familial hypercholesterolemia. Genet Epidemiol 10:665–669

Bonney GE (1984) On the statistical determination of major gene mechanisms in continuous human traits: regressive models. Am J Med Genet 18:731–749

Bonney GE, Lathrop GM, Lalouel J-M (1988) Combined linkage and segregation analysis using regressive models. Am J Hum Genet 43:29–37

Breslow NE, Day NE (1987) Statistical methods in cancer research. Vol 2: The design and analysis of cohort studies. International Agency for Research on Cancer, Lyon

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Prob 10:26–61

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Craig JE, Rochette J, Fisher CA, Weatherall DJ, Marc S, Lathrop GM, Demenais F, et al (1996) Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. Nat Genet 12:58–64

Demenais F, Lathrop M (1994) REGRESS: a computer program including the regressive approach into the LINKAGE programs. Genet Epidemiol 11:291

Eaves LJ (1984) The resolution of genotype × environment interaction in segregation analysis of nuclear families. Genet Epidemiol 1:215–228

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Faucett CL, Gauderman WJ, Thomas DC, Ziogas A, Sobel E (1993) Combined segregation and linkage analysis of late-onset Alzheimer's disease in Duke families using Gibbs sampling. Genet Epidemiol 10:489–494

Gauderman WJ, Faucett CL, Morrison JL, Carpenter CL. Joint segregation and linkage analysis of a quantitative trait compared to separate analyses. Genet Epidemiol (in press)

Gauderman WJ, Morrison JL, Carpenter CL, Thomas DC (1997) Analysis of gene-smoking interaction in lung cancer. Genet Epidemiol 14:199–214

Gauderman WJ, Witte JS, Faucett CL, Morrison JL, Thomas DC (1995) Genetic epidemiologic analysis of quantitative phenotypes using Gibbs sampling. Genet Epidemiol 12: 747–752

Genetic Analysis Package (1997) Genetic-epidemiology analysis package, release 1.0. Computer program available from Epicenter Software, Pasadena

Gueguen R, Visvikis S, Steinmetz J, Siest G, Boerwinkle E (1989) An analysis of genotype effects and their interactions by using the apolipoprotein E polymorphism and longitudinal data. Am J Hum Genet 45:793–802

Guo SW, Thompson EA (1992) A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Genet 51:1111–1126

Hsu L, Davidov O, Holte S, Quiaoit F, Zhao LP. A population based family study of a common oligogenic disease. I. Design. Genet Epidemiol (in press)

Konigsberg LW, Blangero J, Kammerer CM, Mott GE (1991) Mixed model segregation analysis of LDL-C concentration with genotype-covariate interaction. Genet Epidemiol 8: 69–80

Krauss RM, Williams PT, Blanch PJ, Cavanaugh A, Holl LG, Austin MA (1993) Lipoprotein subclasses in genetic studies: the Berkeley data set. Genet Epidemiol 10:523–528

Lange K, Elston RC (1975) Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. Hum Hered 25:95–105

Martinez M, Abel L, Demenais F (1995) How can maximum likelihood methods reveal candidate gene effects on a quantitative trait? Genet Epidemiol 12:789–794

Moll PP, Sing CF, Lussier-Cacan S, Davignon J (1984) An application of a model for a genotype-dependent relationship between a concomitant (age) and a quantitative trait (LDL cholesterol) in pedigree data. Genet Epidemiol 1: 301–314

Morton NE, MacLean CJ (1974) Analysis of family resemblance. III. Complex segregation of quantitative traits. Am J Hum Genet 26:489–503

Olshen AB, Wijsman EM (1996) Pedigree analysis package vs. MIXD: fitting the mixed model on a large pedigree. Genet Epidemiol 13:91–106

Ottman R (1990) An epidemiologic approach to gene-environment interaction. Genet Epidemiol 11:75–86

Rebbeck TR, Turner ST, Michels VV, Moll PP (1989) Evidence for a single locus with age- and sex-specific genotypic effects

on red cell sodium-lithium countertransport levels. Am J Hum Genet Suppl 45:A247

Sellers TA, Bailey-Wilson JE, Elston RC, Wilson AF, Elston GZ, Ooi WL, Rothschild H (1990) Evidence for mendelian inheritance in the pathogenesis of lung cancer. J Natl Cancer Inst 82:1272–1279

Sellers TA, Bailey-Wilson JE, Potter JD, Rich SS, Rothschild H, Elston RC (1992) Effect of cohort differences in smoking prevalence on models of lung cancer susceptibility. Genet Epidemiol 9:261–272

Sznajd J, Rywik S, Furberg B, Pajak A, Kurjata P, Williams OD, Sznajderman-Ciswiska M, et al (1989) Poland and U.S. collaborative study on cardiovascular epidemiology. II. Correlates of lipids and lipoproteins in men and women aged 35–60 years from selected Polish rural, Polish urban, and U.S. samples. Am J Epidemiol 130:446–456

Thomas DC, Cortessis V (1992) A Monte Carlo Bayesian method for genetic linkage analysis. Hum Hered 42:63–76

Tiret L, Abel L, Rakotovao R (1993) Effect of ignoring genotype-environment interaction on segregation analysis of quantitative traits. Genet Epidemiol 10:581–586

Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F (1992) Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I–converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet 51:197–205

Towne B, Siervogel RM, Blangero J. Effects of genotype-by-sex interaction on quantitative trait linkage analysis. Genet Epidemiol (in press)

Wijsman EM, Amos CL. Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol (in press)