Gene–environment interaction tests for family studies with quantitative phenotypes: A review and extension to longitudinal measures

Hortensia Moreno-Macias,[1,2] Isabelle Romieu,[3] Stephanie J. London[4] and Nan M. Laird[5]*

[1]Department of Environmental Health, Harvard School of Public Health, Boston, MA 02115, USA
[2]Universidad Autónoma Metropolitana Unidad Iztapalapa, México DF 09340, México
[3]International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon CEDEX 08, France
[4]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC 27709, USA
[5]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA
**Correspondence to*: Tel: +55 5804 6562; Fax: +55 5612 5682; E-mail: hmoreno@hsph.harvard.edu

Date received (in revised form): 22 April 2010

Abstract

Longitudinal studies are an important tool for analysing traits that change over time, depending on individual characteristics and environmental exposures. Complex quantitative traits, such as lung function, may change over time and appear to depend on genetic and environmental factors, as well as on potential gene–environment interactions. There is a growing interest in modelling both marginal genetic effects and gene–environment interactions. In an admixed population, the use of traditional statistical models may fail to adjust for confounding by ethnicity, leading to bias in the genetic effect estimates. A variety of methods have been developed to account for the genetic substructure of human populations. Family-based designs provide an important resource for avoiding confounding due to admixture. To date, however, most genetic analyses have been applied to cross-sectional designs. In this paper, we propose a methodology which aims to improve the assessment of main genetic effect and gene–environment interaction effects by combining the advantages of both longitudinal studies for continuous phenotypes, and the family-based designs. This approach is based on an extension of ordinary linear mixed models for quantitative phenotypes, which incorporates information from a case–parent design. Our results indicate that use of this method allows both main genetic and gene–environment interaction effects to be estimated without bias, even in the presence of population substructure.

Keywords: *gene–environment interaction, longitudinal phenotypes, power, bias, population substructure*

Introduction

In spite of the multiple efforts to find genetic factors conferring susceptibility to complex diseases, the success of genetic association studies is still hampered by the difficulty in replicating findings in different populations. Among the plausible explanations for this lack of replication is the fact that the effects of environmental factors, which can interact with genetic factors, are not always taken into consideration.[1] There is an increasing interest in studying different susceptibilities to environmental factors in subjects with different genotypes; however, power and bias issues with regard to the statistical estimation of gene–environment interaction effects persist.

High-quality information about individual environmental exposure is crucial for the assessment of gene–environment interactions.[2] Failure to measure changes in exposure levels over time could lead to an underestimation of the role of the environment in the interaction. Repeated measurements of the temporal relationship between an outcome and the exposure may overcome such a problem when both the endpoint and the exposure are time-dependent variables. In addition, potential misclassification due to ambiguity in the definition of complex diseases may be avoided through the measurement of quantitative disease-related phenotypes as the outcomes of interest. For example, quantifying the decrements in lung function over time through repeated spirometric tests may provide insights into the pathogenesis of chronic obstructive pulmonary disease (COPD) or asthma. Many disease 'predictor' phenotypes are thought to change within-subject because of both environmental and genetic factors, and of their potential interactions over time.

On the genetic side, population substructure is an important practical issue for genetic association studies. When the study population is not a collection of randomly mating individuals, several discrete subgroups that are genetically different may be identified; the collection of these subpopulations is referred to as population substructure or stratification.[3] Moreover, disease prevalence also tends to differ among these subgroups.[4] Consequently, without stratification adjustment, allele frequency can appear to be associated with the disease, regardless of whether the genotype has a functional effect on that health outcome or not. By contrast, when the genotype distribution is homogeneous among groups, population substructure may not be an issue. For example, if people are randomly assigned to treatment groups, it is expected that those groups will be genetically similar. If, additionally, there are no differences in the response to treatment among the different subgroups, bias due to population substructure is unlikely.

Another source of spurious associations is population admixture, which refers to the mixture of different ancestries; that is, people from different ethnic groups interbreed, so the genome of the new generations is a combination of genotypes of the original ancestry groups, and, consequently, in some genes, allele frequencies are not homogeneously distributed in the study population. For example, it has been recognised that Latino populations have varying proportions of African, Native American and European ancestry.[5] Like population substructure, if the risk of disease depends on ancestry, a high risk of disease may be erroneously associated with a high allele frequency; thus, in admixed populations, ethnicity may confound associations between genotype and outcome and assessment of gene by environment interactions. The direction of the confounding could be positive or negative. Therefore, to identify true associations, population substructure must be taken into account in the analysis.

With the increasing availability of genetic data, there is a growing interest in modelling both marginal genetic effects and gene–environment interactions. Inclusion of interactions, when they exist, can increase the statistical power of detecting both genetic and environmental effects.[6] Traditional statistical models for detecting significant main effects and interactions may not be completely adequate for studying genetics in admixed or stratified populations, however.

A variety of methods have been developed to account for the genetic substructure of human populations.[7] Family-based designs provide an important resource for avoiding confounding due to admixture.[8] The simplest design for testing association is the case–parent (or trio) design because it uses genotypes from an affected offspring, the case, and his/her two parents. The outcome is measured, however, only in the offspring. Many of these methods have been developed for cross-sectional designs, but can be applied to repeated measurements through the two-step modelling approach. The first step consists of calculating the slope between the longitudinal outcome and the time-dependent environmental exposure; thus, we calculate a single individual endpoint, the slope, for each subject. In the second step, the genetic methods for cross-sectional studies, where the slope is the single outcome, can be applied.[9]

In this paper, we first provide a short review of different approaches for studying gene–environment interactions for quantitative traits, and then propose a method that aims to improve the assessment of main and gene–environment interaction effects by combining the advantages of both longitudinal studies for continuous phenotypes and the family-based designs. This approach is based on an extension of ordinary linear mixed models (OLMM) for quantitative phenotypes which incorporates information from a case–parent design. We call the model the 'adjusted linear mixed model' (ALMM), and through simulation methods we show that even when population stratification is present, both main genetic and gene–environment interaction effects can be estimated without bias, and that this is more powerful than the two-step modelling approach.

The broad objectives of this paper do not extend to giving technical details about the family-based approach and its extensions, or to giving an extensive explanation about linear mixed models. Rather, we present what we consider to be a widely applicable method for correctly assessing the main genetic effect and gene–environment interactions for time-dependent quantitative traits in stratified populations. For this purpose, we use simulated repeated measurements of forced expiratory flow between 75 per cent and 25 per cent of vital capacity (FEF_{25-75}) ie (lung function) on asthmatic children exposed to ozone pollution, based on the observed distributions in a real cohort study conducted in Mexico City.[10]

In order to set the stage for our methodology, we first provide a brief overview of some existing ordinary linear regression (OLR) models for testing main genetic effects and gene–environment interactions in cross-sectional studies that incorporate information about parental genotype (case–parent or trio design), adjusting for admixture. We then briefly present the family-based association test (FBAT) approach, which, as a second step (after computing the slope between the outcome and the exposure), represents an alternative method for analysing genetic associations over time. We next review the ordinary linear mixed models (OLMM) which are a standard approach for the analysis of longitudinal data, and present the adjusted linear mixed models (ALMM) as an extension of OLMM combined with the adjusted cross-sectional regression models. In order to show that the two-step modelling approach provides a valuable alternative for analysing longitudinal data, we explain the relationship between this approach and the linear mixed models. Finally, we give details about the simulation procedures and present our results and discussion.

Methods

Models for cross-sectional data with a single quantitative measure for each subject

Existing methods for testing main genetic effect and gene–environment interactions with a single measured outcome include (1) OLR models and extensions that aim to adjust for ethnicity by including parental genotype information in a case–parent or trio design, and (2) the FBAT approach, which uses a score test based on a conditional likelihood.[11]

The OLR approach

In a genetic association analysis with quantitative traits that follow a linear model, the assessment of gene–treatment interactions may be conducted using standard linear regression models for independent subjects. Under the usual assumptions, the well known model for testing the interaction between two covariates is:

$$E(Y_i|X,Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i \quad \text{for } i = 1, 2, \ldots n; \tag{1}$$

where:

n is the number of subjects in the study;
X_i is a fixed variable that translates an offspring genotype into a numeric value; and
Z_i is an observed environmental covariate, either continuous or dichotomous.

X_i is a scalar whose value depends on the disease model. If the locus has two alleles, A and a, the additive model assumes that each copy of the variant allele* 'A' changes the outcome in an additive amount. Thus, X_i counts the number of 'A'

*Usually the variant allele is the less frequent.

alleles in the offspring genotype ($X_i = \{0, 1, 2\}$). In the recessive model, X_i is coded as an indicator variable for the AA genotype. As a special case, model (2) is used for testing main genetic effects adjusted for the environmental covariate:

$$E(Y_i|X, Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i \tag{2}$$

A rejection of the null hypothesis $H_0 : \beta_1 = 0$ means that the quantitative trait is associated with the alleles in the marker.

The case–parent or trio design

Unlike OLR models, family designs aim to avoid spurious associations due to population admixture. In the case–parent design, the proband is the offspring that identifies the family for the study; the genotypes of the candidate gene are measured for all members of the trio, but the quantitative trait is measured only in the offspring. The alternative form of the OLR approach for testing main genetic effect on quantitative outcomes, where the parental genotype information is included, was developed by Allison;[12] it is based on the simplest family-based design for testing associations, known as the transmission disequilibrium test (TDT).[13] The model is adjusted for the expected value of the offspring's genotype conditional on the parental genotypes; thus, the adjusted version of model (2) is:

$$E(Y_i|X, Z) = \beta_0 + \beta_1(X_i - E(X_i|g_{im}, g_{if})) + \beta_2 Z_i \tag{3}$$

where:

g_{im}, g_{if} are the parental genotypes (mother and father, respectively) and $E(X_i|g_{im}, g_{if})$ is calculated under segregation and independent assortment assumptions using Mendel's law. Its value depends on the mating type and the disease model.

The adjusted genotype $X_i - E(X_i|g_{im}, g_{if})$ is the subject's deviation from the family mean under Mendel's law. β_1 represents the within-family effect of the gene on the outcome. As a result of centring X_i by its expected value conditional on parental genotypes (g_{im}, g_{if}), ethnicity bias is avoided, since all possible genotypes — depending on the mating type — are taken into account, even those that were not transmitted to the affected offspring. This strategy does not, however, necessarily prevent bias due to other kinds of population stratification,[14] such as the one that occurs when parental mating type is highly correlated with the levels of exposure, for example. For this reason, Allison[12] and Ewens *et al.*[14] propose an alternative version of model (3) in which the intercept, representing the between-family component, depends on the mating type as a fixed effect:

$$\begin{aligned} E(Y_i|X, Z) &= \beta_{0M} + \beta_1(X_i - E(X_i|g_{im}, g_{if})) + \beta_2 Z_i \\ &= \tilde{\beta}_{0M} + \beta_1 X_i + \beta_2 Z_i \end{aligned} \tag{4}$$

where:

$$\tilde{\beta}_{0M} = \beta_{0M} - \beta_1 E(X_i|g_{im}, g_{if})$$

and M= 1, 2, 3 are the three possible mating types with at least one heterozygous parent, including informative families only.[†] Note that here and in subsequent equations, M depends upon i via the parental genotypes, but this is suppressed for simplicity.

It is noteworthy that since both β_{0M} and $E(X|g_{im}, g_{if})$ are constant within the mating type, the estimation of the main genetic effect (β_1) through model (4) is completely equivalent to using model (5):

$$E(Y_i|X, Z) = \beta_{0M} + \beta_1 X_i + \beta_2 Z_i \tag{5}$$

Following the same idea, Gauderman[15] proposed a likelihood ratio test (LRT) of β_1 in model (5), although, in order to increase the power for detecting main genetic effects, he included the whole set of families, regardless of the heterozygous condition. This is called the quantitative transmission disequilibrium test with mating type indicators (QTDT$_M$). An extension of the QTDT$_M$ to

[†] There are six different mating types: AAXAA, AAXAa, AAXaa, AaXAa, AaXaa, and aaXaa. When both parents are homozygous, the observed and conditional expected genotypes are equal and it is said that the family is not informative.

include the environmental covariate and the gene–environment interaction was also suggested by Gauderman[15] with the model:

$$E(Y_i|X,Z) = \beta_{0M} + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i \quad (6)$$

Where $M = 1,\ 2, \ldots, 6$ are the six possible mating-types.

In contrast to the relationship between models (4) and (5), model (6) is not equivalent to:

$$\begin{aligned} E(Y_i|X,Z) = \beta_{0M} &+ \beta_1[X_i - E(X_i|g_{im},g_{if})] \\ &+ \beta_2 Z_i + \beta_3 Z_i[X_i - E(X_i|g_{im},g_{if})] \end{aligned} \quad (7)$$

because, although β_{0M} and $E(X_i|g_{im},g_{if})$ are constant within mating type, the environmental covariate (Z_i) is not.

An alternative model to (7), (and an extension of Gauderman's idea), is given by the adjusted model that we call the adjusted quantitative transmission disequilibrium test with mating-type indicators ($\mathrm{AQTDT_M}$).

$$\begin{aligned} E(Y_i|X,Z) = \beta_{0M} + \beta_1[X_i - E(X_i|g_{im},g_{if})] + \beta_{2M} Z_i \\ + \beta_3 Z_i[X_i - E(X_i|g_{im},g_{if})] \end{aligned} \quad (8)$$

where now both the intercept and the slope for the environmental covariate depend on the mating type, which means that the model has been adjusted for a potential correlation between the exposure to environmental factors and the mating type. Note that now model (8) is equivalent to:

$$E(Y_i|X,Z) = \beta_{0M} + \beta_1 X_i + \beta_{2M} Z_i + \beta_3 Z_i X_i \quad (9)$$

The advantage of using models (5) and (9) is that the inclusion of an indicator variable for the mating type, rather than calculating the expected genotype, provides protection against population admixture while taking into account those situations where environmental exposure (Z) may depend on the mating type and other additional types of population substructure.[14] As usual, in (6), (7), (8) and (9), β_3 estimates the effect modification of the gene on the environmental effect Z_i and $H_0 : \beta_3 = 0$ is the null hypothesis that states the no interaction effect.

The FBAT approach

FBAT is the generalisation of the TDT for the trio design. It encompasses a broad class of statistical methods for testing genetic associations adjusting for potential admixture or stratification. Such methods are also based on extensions of the TDT and regression models, although the covariance between genotype and phenotype is the statistic of interest.[11] The general FBAT statistic has been explained elsewhere.[16] Briefly, for *n* nuclear families, one offspring in the family *i* and no covariates:

$$\chi^2_{FBAT} = \frac{U^2}{Var(U)} \quad (10)$$

where:

$$U = \sum[(Y_i - E(Y_i))*(\mathrm{X_i} - \mathrm{E}(\mathrm{X_i}|\mathrm{g_{im}}, \mathrm{g_{if}}))]$$
$$\text{for } i = 1, 2, \ldots n;$$
$$Var(U) = \sum_i (Y_i - E(Y_i))^2 * \mathrm{Var}(\mathrm{X_i}|\mathrm{g_{im}}, \mathrm{g_{if}}),$$

and $E(X_i|g_{im},g_{if})$ and $Var(X_i|g_{im},g_{if})$ are calculated under the null hypothesis of Mendel's law. This statistic follows a chi-square distribution with one degree of freedom. In addition, unlike regression models where the offspring's genotype is assumed to be fixed and observed, the general FBAT approach considers the offspring genotype as a random variable. The general idea is first to calculate a test statistic for the association between the trait and the marker locus, and then, as a second step, the distribution of this test statistic is derived from the distribution of the offspring genotype under the null hypothesis of no association. The distribution of the test statistic is computed conditioning on the sufficient statistic given by the parental genotype and the observed offspring's phenotype. Under these conditions, no assumptions about the allele frequency, the recombination fraction or the penetrance function are required. Due to the fact that the general FBAT statistic can only test main genetic effects, and since the test statistic is based the relative size of U with respect to its standard deviation, the genetic effect is not directly estimated.

Under the philosophy of the FBAT, Vansteelandt *et al.*[17] proposed an extension that permits the assessment of and testing for the gene–environment interaction without any assumptions about the genotype distribution and with no bias due to unmeasured ethnicity confounding through which Mendel's laws hold. Such an extension is based on G-estimation (causal inference) methodology and is called QBAT-I; this test statistic has the same form as the general FBAT (10), although the expected genotype and the U statistic have more complex expressions. A brief presentation of both FBAT and QBAT-I can be found in Appendix 1.

FBAT and QBAT-I statistics are available from PBAT free software at http://biosun1.harvard.edu/~clange/pbat.htm and in the library pbatR under the R package environment.

Models for longitudinal data

Until now, we have discussed methods for analysing gene–environment interactions when a single measurement of the phenotype is available. We now turn to a discussion of methods for the analysis of quantitative repeated phenotype data to evaluate the effects of a gene and the environment over time.

In longitudinal designs, the unit of study is not each individual or each measurement, but rather the sequence of measurements on each subject. This means that the major advantage of a longitudinal analysis is that the so-called cohort and age effects are estimated separately; that is, differences among people in their baseline levels (cohort effects), can be discriminated from the changes over time (ageing effects) within individuals. In other words, measurements across people and repeated values across time are sources of strength. Note that cross-sectional data provide information for assessing only the former effect; thus, longitudinal studies tend to be more powerful than cross-sectional studies.[18]

Although there are different approaches to longitudinal data analysis, we consider here the two most commonly used: OLMM and two-step methods.

OLMM

OLMM assumes that the vector of repeated measurements on each subject follows a linear regression model. Thus, each individual model may have subject-specific intercept and slope (the random effects) representing the different susceptibilities to the environmental exposure among subjects. For most outcomes, variability across individuals is greater than within-subject. This difference may be due the influence of genetic composition.[19]

In general, a linear mixed-effects model satisfies:

$$Y_{ij} = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_6 Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij} + b_{1i} + b_{2i} t_{ij} + e_{ij} \tag{11}$$

where:

$i = 1, 2, 3, \ldots, n$ subjects;
$j = 1, 2, 3, \ldots, m$ corresponding times at which the measurements are taken on each subject;
t_{ij} is the repeated time (or exposure) variable;
b_{1i} is the random subject intercept effect $(\alpha_0 + b_{1i})$, which varies among subjects;
b_{2i} is the random subject slope effect $(\alpha_1 + b_{2i})t_{ij}$, which varies among subjects;

e_{ij} is a random variable regarded as measurement or sampling errors;

$$E(b_{1i}) = E(b_{2i}) = E(e_{ij}) = 0$$

$$Var(b_{1i}) = \sigma^2_{b1}; \quad Var(b_{2i}) = \sigma^2_{b2}; \quad Var(e_{ij}) = \sigma^2; \quad Cov(b_{1i}, b_{2i}) = \sigma_{12},$$

and X_i and Z_i are as previously defined. Note that model (11) assumes that both genotype (X_i) and environmental exposure (Z_i) remain fixed over time. However, we can take advantage of the time variable (t_{ij}) to include an extra source of exposure that changes across time. For example, we are particularly interested in the assessment of the gene–environment interaction effect between the glutathione-S-transferase M1 gene (*GSTM1*) and antioxidant supplementation on lung function — FEF_{25-75} — of asthmatic children exposed to ozone.

Therefore, in model (11), X_i represents the individual *GSTM1* genotype and Z_i is a dichotomous variable that denotes the antioxidant supplementation group which was randomly assigned at baseline and remained fixed during follow-up, and which since ozone is a time-dependent variable, can be represented by t_{ij}. This is a model with random intercept and random slope on ozone. As for ordinary linear regression models in cross-sectional studies, OLMM assumes independent subjects and does not account for population admixture. Therefore, in order to avoid potential estimation bias, it is necessary to adjust for ethnicity.

ALMM

The ALMM form is a straightforward extension of the approaches presented for the cross-sectional designs. That is, following the ideas of Allison,[12] Ewens *et al.*[14] and Gauderman,[15] model (12) can be rewritten using the indicator variables for the mating type and the offspring's conditional expected genotype:

$$\begin{aligned}\mathrm{FEF}_{2575ij} = \alpha_{0M} &+ \alpha_{1M} t_{ij} + \alpha_2 Z_i \\ &+ \alpha_3 [X_i - E(X_i | g_{im}, g_{if})] \\ &+ \alpha_4 [X_i - E(X_i | g_{im}, g_{if})] Z_i \\ &+ \alpha_5 [X_i - E(X_i | g_{im}, g_{if})] t_{ij} + \alpha_{6M} Z_i t_{ij} \\ &+ \alpha_7 [X_i - E(X_i | g_{im}, g_{if})] Z_i t_{ij} \\ &+ b_{1i} + b_{2i} t_{ij} + e_{ij}\end{aligned} \tag{12}$$

where i, j, b_{1i}, b_{2i}, e_{ij} are as defined for (11); X_i represents the *GSTM1* genotype; Z_i denotes the antioxidant supplementation group; t_{ij} represents the ozone exposure and $M = \{1,2,\ldots,6\}$ are the six possible mating types. This model is an extension of (11) and is able to control for population admixture and for any dependence of the environmental exposure on mating type. Once again, using an indicator variable for the mating type prevents controls for other potential sources of population structure.

Note that, as in model (8), the simplest, but equivalent, expression for the above model that does not need the calculation of the expected genotype is given by:

$$\begin{aligned}FEF_{2575ij} = \alpha_{0M} &+ \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i \\ &+ \alpha_5 X_i t_{ij} + \alpha_{6M} Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij} \\ &+ b_{1i} + b_{2i} t_{ij} + e_{ij}\end{aligned} \tag{13}$$

Through this model, the main genetic and environmental effects and gene–environment interactions can be assessed using standard statistical software. Additional advantages of this model are that genetic and longitudinal effects are estimated simultaneously; thus, parameters are mutually adjusted for each other[9] and within- and between-individual variabilities are taken into account. Moreover, the model can be adjusted by other covariates, such as gender and age at baseline, or even by smooth functions of variables which have non-linear association with the outcome. In particular, in the example, two sources of environmental exposures — supplementation treatment (a constant) and ozone (which is time dependent) — are included. Computationally, longitudinal models require more complex algorithms than those corresponding to cross-sectional designs; however, this is no longer a problem for the users of statistical packages. Appendix 1 includes the R code for estimating the effects of model (13).

Two-step modelling approach

As the term implies, this approach includes two separate steps. In the first step, it is assumed that the repeated measures on subject i, are independent and follow a linear regression model. Therefore, the individual intercept and slope are estimated by an ordinary linear regression model:

$$Y_{ij} = \gamma_{0i} + \gamma_{1i} t_{ij} + e_{ij} \tag{14}$$

Therefore, different regression coefficients (γ_0 intercept and γ_1 slope) correspond to different individuals. For example, with the lung function data, the outcome is the repeated FEF_{25-75} and γ_1

represents individual susceptibility to ozone over time. Thus, the longitudinal observations are reduced to one summary statistic per subject.[20] The second step includes the genetic analysis where these slopes are the single outcome for each person. This approach has been frequently used in segregation, linkage and association analysis.[9] The R code for these models is included in Appendix 1.

It is interesting to note the close relationship between the coefficients in the two-step model and the respective coefficients in a linear mixed model. That is, by definition, from the ordinary linear regression model (14):

$$slope_i = \gamma_{1i} = \frac{\sum_j (Y_{ij} - \bar{Y}_i)(t_{ij} - \bar{t}_i)}{\sum_j (t_{ij} - \bar{t}_i)^2}$$

where Y_{ij} is given by (11) and

$$\begin{aligned}\bar{Y}_i = \alpha_0 + \alpha_1 \bar{t}_i + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i \bar{t}_i \\ + \alpha_6 Z_i \bar{t}_i + \alpha_7 X_i Z_i \bar{t}_i + b_{1i} + b_{2i} \bar{t}_i + \bar{e}_i\end{aligned}$$

Thus, it is straightforward to show that

$$\begin{aligned}slope_i &= \alpha_1 + \alpha_5 X_i + \alpha_6 Z_i + \alpha_7 X_i Z_i + \bar{e}_i \\ &\Rightarrow E(slope_i | X, Z) = \alpha_1 + \alpha_5 X_i \\ &\quad + \alpha_6 Z_i + \alpha_7 X_i Z_i\end{aligned}$$

This therefore enables one to relate the α coefficients of the linear mixed model to the β coefficients of the ordinary linear regression model given by (1) as follows: $\beta_1 = \alpha_5$ and $\beta_3 = \alpha_7$ representing the main genetic effect and the gene–treatment interaction on the slope, respectively.

This approach is remarkably similar to that of mixed models, although longitudinal and genetic effects are not jointly estimated, and time-dependent covariates need to be summarised in one measurement — the mean or the median, for example — in order to be included in the analysis. Computationally, this procedure is simpler than mixed models, and any elemental statistical package will suffice for conducting the analysis.

Table 1 summarises the different models included in this section.

Simulations

Here, we address differences in the analytical approaches, in terms of both bias and power, for detecting main genetic effects and gene–environment interactions using simulations.

In order to compare the different methods presented, we simulated data with similar characteristics to those in the study by Romieu *et al.*[10] Briefly, this study was a randomised trial using a double-blinded and longitudinal design, including antioxidant supplementation for asthmatic children who were residents of Mexico City and therefore exposed to ozone pollution. There were 12 repeated measures for both FEF_{25-75} and ozone. The deletion polymorphism of *GSTM1*, absent versus present, was determined for each child and, through a stratified analysis, evidence for interaction between the antioxidant treatment (dichotomous and fixed variable) and the *GSTM1* genotype was seen for the effect of ozone on lung function.

For our simulations, 12 repeated measures of lung function (FEF_{25-75}) were generated for each offspring with a mean vector given by a linear mixed model conditional on treatment, genotype and ozone level, assuming an additive (or recessive) disease model (15). An error vector was added to each mean by drawing from a multivariate normal distribution. The variance–covariance matrix of the errors was assumed to be equal to the observed variance–covariance matrix among the residuals in the real data, where model (15) was fit to the repeated FEF_{25-75} measurements.

For the purpose of using the family-based approach, samples of independent trios were simulated. Each parent was randomly assigned a genotype assuming the Hardy–Weinberg equilibrium, while each offspring was assigned a genotype assuming Mendel's law. Treatment (Z) was randomly and independently assigned for each subject *i* with a 50 per cent probability for supplement or placebo group. Both additive and recessive disease models were considered.

With regard to the population stratification, two different situations were considered: the first one assumed a homogeneous population (HP), where

Table 1. Regression models included in the paper . The 'Model' column refers to the number that identifies each model in the paper.

Models for cross-sectional data			
Model	**Independent subjects design**	**Case–parent design**	**Comments**
Main genetic effect			
(2)	$E(Y_i\|X,Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i$		Ordinary linear regression model
(3)		$E(Y_i\|X,Z) = \beta_0 + \beta_1(X_i - E(X_i\|g_{im}, g_{if})) + \beta_2 Z_i$	Adjusted version of (2). The model is adjusted by the expected value of the offspring's genotype conditional to the parental genotypes
(4)		$E(Y_i\|X,Z) = \beta_{0M} + \beta_1(X_i - E(X_i\|g_{im}, g_{if})) + \beta_2 Z_i$	Gauderman's model (QTDT$_M$) adjusted for the covariate Z
(5)		$E(Y_i\|X,Z) = \beta_{0M} + \beta_1 X_i + \beta_2 Z_i$	(4) equivalent to(5)
(10)		$\chi^2_{FBAT} = \frac{U^2}{Var(U)}$	FBAT statistic
Gene–environment interaction			
(1)	$E(Y_i\|X,Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i$		Ordinary linear regression model
(6)		$E(Y_i\|X,Z) = \beta_{0M} + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i$	Gauderman's model (QTDT$_M$)
(7)		$E(Y_i\|X,Z) = \beta_{0M} + \beta_1[X_i - E(X_i\|g_{im}, g_{if})] + \beta_2 Z_i + \beta_3 Z_i[X_i - E(X_i\|g_{im}, g_{if})]$	(6) is not equivalent to (7) when the environment covariate (Z_i) is not constant within mating type.
(8)		$E(Y_i\|X,Z) = \beta_{0M} + \beta_1[X_i - E(X_i\|g_{im}, g_{if})] + \beta_{2M} Z_i + \beta_3 Z_i[X_i - E(X_i\|g_{im}, g_{if})]$	Adjusted QTDT$_M$
(9)		$E(Y_i\|X,Z) = \beta_{0M} + \beta_1 X_i + \beta_{2M} Z_i + \beta_3 Z_i X_i$	(8) equivalent to (9)

Continued

Table 1. Continued

Models for longitudinal data			
Model	**Independent subjects design**	**Case–parent design**	**Comments**
Main genetic effect			
(19)	$E(Y_{ij}\|X,Y,t) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i$ $+ \alpha_5 X_i t_{ij} + \alpha_6 Z_i t_{ij}$		Ordinary linear mixed model (OLMM)
(20)		$E(Y_{ij}\|X,Y,t) = \alpha_{0M} + \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i$ $+ \alpha_5 X_i t_{ij} + \alpha_{6M} Z_i t_{ij}$	Adjusted linear mixed model (ALMM)
Gene–environment interaction			
(11)	$Y_{ij} = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij}$ $+ \alpha_6 Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij} + b_{1i} + b_{2i} t_{ij} + e_{ij}$		Ordinary linear mixed model
(12)		$\mathrm{FEF}_{2575ij} = \alpha_{0M} + \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 [X_i$ $- E(X_i\|g_{im}, g_{if})] + \alpha_4 [X_i - E(X_i\|g_{im}, g_{if})] Z_i +$ $\alpha_5 [X_i - E(X_i\|g_{im}, g_{if})] t_{ij} + \alpha_{6M} Z_i t_{ij} + \alpha_7 [X_i - E(X_i\|g_{im}, g_{if})] Z_i t_{ij} + b_{1i} + b_{2i} t_{ij} + e_{ij}$	Adjusted linear mixed model (ALMM)
(13)		$FEF_{2575ij} = \alpha_{0M} + \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij}$ $+ \alpha_{6M} Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij} + b_{1i} + b_{2i} t_{ij} + e_{ij}$	(13) is equivalent to (12)

X_i is a fixed variable that translates an offspring genotype to a numerical value; Z_i is an observed environmental covariate, either continuous or dichotomous; g_{im}, g_{if} are the parental genotypes (mother and father, respectively);$E(X_i|g_{im}, g_{if})$ is calculated under segregation and independent assortment assumptions using Mendel's law; $M = 1, 2, \ldots, 6$ are the six possible mating types; $i = 1, 2, 3, \ldots, n$ subjects; $j = 1, 2, 3, \ldots, m$ measurement occasions into the subject; t_{ij} is the repeated time (or exposure) variable;

b_{1i} is the random subject intercept effect; $(\alpha_0 + b_{1i})$ varies among subjects; b_{2i} is the random subject slope effect: $(\alpha_1 + b_{2i})t_{ij}$ varies among subjects; e_{ij} is a random variable regarded as measurement or sampling errors.

the observed allele frequency for *GSTM1*, $P(a) = 0.4$ and $P(A) = 0.6$,[‡] was used for simulating the genotypes. The generating model we used had the following form:

$$E(Y_{ij}|X, Z, t) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_6 Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij} \tag{15}$$

Where: t_{ij} represents the time-dependent ozone exposure; Z_i is a dichotomous variable representing the treatment group and X_i is either a continuous variable counting the copy variant number of *GSTM1* (0, 1 or 2) in the additive disease model or a dummy variable in the recessive case. The parameters are:

$$\alpha_0 = 1.8;\ \ \alpha_1 = -0.8;\ \ \alpha_2 = -0.2;\ \ \alpha_3 = -0.05;\ \ \alpha_4 = 0.2;\ \ \alpha_5 = 0.5;\ \ \alpha_6 = 0.6; \alpha_7 = 1.0$$

α_0, α_1, α_2, α_3, α_4 are equal to the corresponding estimates obtained through a linear mixed model with random intercept and random slope on ozone applied to the original dataset. With the only aim to simulate situations where the effects are clearly identified without using large samples but rather looking at the differences between the methods, α_5, α_6, α_7 were magnified. Assigned values to α_5, α_6, (α_7) try to mimic the respective effects found in the stratified analysis reported by Romieu *et al.*[10] Specifically, the genotype effect for each part per billion (ppb) of ozone was 0.8 ml/s in the placebo group ($n = 78$ subjects) and 0.16 ml/s in the supplement group ($n = 80$); thus, the main genetic effect was taken as the rounded weighted average effect, while ignoring the supplement effect. The same procedure was applied to calculating the main treatment effect. Finally, since the effect of antioxidants was stronger in the *GSTM1* null genotype group (0.95 ml/s), and there was no significant effect in the *GSTM1*-positive group, the coefficient for the gene–treatment interaction on the lung function–ozone relationship was rounded to 1.0. Table S1 summarises the models used in the simulation process and Table S2 shows the observed effects in the real cohort study conducted in Mexico City.

The second situation presumes a stratified population (AP) with a 50/50 mix coming from two populations with different allele frequencies and different susceptibilities: $P_1(A) = 0.4; P_1(a) = 0.6$ and $P_2(A) = 0.8; P_2(a) = 0.2$. Note that, on average, the combined population has the same allele frequencies as the homogeneous one. With the purpose of simulating differential susceptibilities to ozone and supplementation, although allowing the bias assessment for the main genetic and interaction effects, generating model (15) differs in β_0 and β_6 coefficients (based on the observed percentiles). That is, in the first population, the observed 95th percentile for FEF_{25-75} ($\alpha_0 = 3.3$) was used as the intercept, and no treatment effect on the slope $\alpha_6 = 0$ was assumed. On the other hand, in the second population, the 5th percentile ($\alpha_0 = 0.75$) was taken as the intercept, and a strong treatment effect on the slope was assumed, $\alpha_6 = 2$ (meaning that 20 ml/s/10 ppb decreased lung function, on average, in the placebo group in comparison with the supplement group). The variance–covariance matrix was constant over the different simulated samples.

Regarding the genetic effect, two different scenarios were considered. The first scenario represents situations where the variability in the outcome may be attributable just to the main effect of the gene and the treatment, meaning that there is no gene–treatment interaction effect ($\alpha_7 = 0$) in the generating model (15). The second scenario assumes that all genetic, treatment and gene–treatment interaction are present in the true model.

Assessment of main genetic effect

The first scenario, where there is no gene–treatment interaction ($\alpha_7 = 0$), was used for testing the main genetic effect, adjusted by treatment effect.

Under the two-step modelling approach, the slope between the outcome and ozone was first computed. In the second step, the ordinary linear regression model (16), the AQTDT_M (17) model

[‡] Although allele *A* was not the less frequent, it was considered the variant allele because, in the original study, children with genotype *AA* were classified as *GSTM1* null (no copy), and those with genotypes *aA* (one copy) and *aa* (two copies) were considered *GSTM1* positive.

Table 2. Bias results for main genetic effect assessment comparing ordinary statistical methods (OLR and OLMM) to family-based methods (AQTDT$_M$ and ALMM) under homogeneous (HP) and stratified (SP) populations. Each time, *n* cases were simulated with parameters $\beta_1 = \alpha_5 = 0.5$. Simulations are based on the additive genetic model. † = number that identifies each model in the paper.

n	$\hat{\beta}_1 - \beta_1$				$\hat{\alpha}_5 - \alpha_5$			
	Two-step models				Mixed models			
	OLR (16) †		AQTDT$_M$ (17)		OLMM (19)		ALMM (20)	
	HP	SP	HP	SP	HP	SP	HP	SP
100	−0.003	−0.356	−0.008	−0.017	−0.013	−0.360	−0.018	−0.026
200	−0.005	−0.356	−0.026	−0.035	−0.011	−0.350	−0.017	−0.027
300	0.008	−0.362	0.01	0.014	−0.002	−0.361	0.004	0.017
400	−0.006	−0.356	−0.005	0.005	0.001	−0.356	−0.004	0.008
500	0.005	−0.351	0.005	0.009	0.006	−0.351	0.006	0.003
600	−0.001	−0.355	0.006	0.002	0.000	−0.356	0.001	0.000
1000	−0.001	−0.363	−0.006	−0.008	−0.001	−0.364	−0.005	−0.007

OLR, ordinary linear regression; OLMM, ordinary linear mixed models; ALMM, adjusted linear mixed models; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators

and the FBAT statistic (18) were used for testing the null hypothesis of no main genetic effect $(H_0 : \beta_1 = 0)$.

$$E(slope_i|X, Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i, \qquad (16)$$

$$E(slope_i|X, Z) = \beta_{0M} + \beta_1 X_i + \beta_2 Z_i \qquad (17)$$

$$\chi^2_{FBAT} = \frac{\sum[(slope_i - E(slope_i)) *(\mathrm{X_i} - \mathrm{E}(\mathrm{X_i}|\mathrm{g_{im}}, \mathrm{g_{if}}))]}{\sum_i (slope_i - E(slope_i))^2 * \mathrm{Var}(\mathrm{X_i}|\mathrm{g_{im}}, \mathrm{g_{if}})} \qquad (18)$$

The corresponding null hypothesis when using longitudinal outcomes $H_0 : \alpha_5 = 0$ was tested

Table 3. Bias results for gene–environment interaction effect assessment comparing ordinary statistical methods (OLR and OLMM) to family-based methods (AQTDT$_M$ and ALMM) under homogeneous (HP) and stratified (SP) populations. Each time, *n* cases were simulated with parameters $\beta_3 = \alpha_7 = 1$. Simulations are based on the additive genetic model. † = number that identifies each model in the paper.

n	$\hat{\beta}_3 - \beta_3$				$\hat{\alpha}_7 - \alpha_7$			
	Two-step models				Mixed models			
	OLR (21) †		AQTDT$_M$ (22)		OLMM (23)		ALMM (24)	
	HP	SP	HP	SP	HP	SP	HP	SP
100	0.001	−0.687	−0.013	0.066	−0.013	−0.719	0.008	0.029
200	0.001	−0.745	0.034	−0.008	0.007	−0.738	−0.008	−0.007
300	−0.008	−0.713	0.007	0.004	−0.005	−0.716	0.011	0.009
400	0.002	−0.719	0.012	0.01	0.005	−0.711	0.014	0.016
500	0.01	−0.729	0.000	−0.027	−0.001	−0.725	0.011	−0.006
600	0.011	−0.697	0.014	0.008	0.012	−0.704	0.016	0.005
1000	0.003	−0.728	0.007	−0.018	0.052	−0.730	0.011	−0.075

OLR, ordinary linear regression; OLMM, ordinary linear mixed models; ALMM, adjusted linear mixed models; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators

through ordinary and adjusted linear mixed models; that is:

$$E(Y_{ij}|X, Y, t) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_6 Z_i t_{ij} \tag{19}$$

and

$$E(Y_{ij}|X, Y, t) = \alpha_{0M} + \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_{6M} Z_i t_{ij} \tag{20}$$

For the purpose of comparison, statistics based on model (16) will be referred to as OLR, those based on (17) will be referred to as the $\mathrm{QTDT_M}$, those based on (19) will be referred as OLMM and those based on (20) will be referred as ALMM.

Assessment of gene–environment interaction

Using the same idea, the second scenario, where $\alpha_7 = 1$ in the generating model, was used for assessing the interaction effect through a one degree of freedom test.

Assuming one outcome per individual, in the two-step modelling approach, the null hypothesis $H_0 : \beta_3 = 0$, was tested using (21), (22) and the QBAT-I statistic:

$$E(slope_i|X, Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 Z_i X_i \tag{21}$$

$$E(slope_i|X, Z) = \beta_{0M} + \beta_1 X_i + \beta_{2M} Z_i + \beta_3 Z_i X_i \tag{22}$$

For repeated measurements, the corresponding null hypothesis $H_0 : \alpha_7 = 0$ was tested using the OLMM and ALMM models respectively:

$$E(Y_{ij}|X, Z, t) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_6 Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij} \tag{23}$$

and

$$E(Y_{ij}|X, Y, t) = \alpha_{0M} + \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_{6M} Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij} \tag{24}$$

For the purpose of comparison, statistics based on model (21) will be referred to as OLR, those based on (22) will be referred to as the adjusted $\mathrm{AQTDT_M}$, those based on (23) will be referred as OLMM and those based on (24) will be referred as ALMM.

The empirical power for each test was estimated as the percentage of occasions on which the null hypothesis was rejected at a significance level $\alpha \leq 0.05$ for a two-sided test. In each simulation study, 1,000 independent replicate datasets were generated. Each dataset consisted of n ($n = 100$, 200, 300, 400, 500, 600 and 1,000) complete and independent trios.

Bias was calculated as the average of the difference between the estimator and the true parameter value ($\hat{\beta} - \beta$).

Results

Estimation bias

In order to look at the differences among methods, the estimation bias for the main genetic effects and gene–environment estimation was computed under both population conditions, homogeneous (HP) and stratified (SP) populations. Table 2 shows the resultant bias for the four methods, with two columns per method (HP and SP). When there is no ethnicity confounding (HP), all methods (OLR, $\mathrm{AQTDT_M}$, OLMM and ALMM) for estimating main effects are unbiased, regardless of the design analysis (independent subjects or trios) or the modelling approach (two-step or longitudinal data). By contrast, when the population is stratified, the selection of the design is crucial. In other words, while estimators obtained from the case–parent design are unbiased, models using independent subjects underestimate the effects by around 0.36 units. Note that, regardless of the modelling approach, both designs provide similar results.

As in the main genetic effect assessment, the estimation of the gene–environment interaction also depends strongly on the design of the study when there is population stratification. In this case, when using ordinary statistical methods, the interaction effect was also underestimated by about 0.72 units, but with a homogeneous population the design is not relevant in terms of bias (Table 3).

Empirical power

Since power comparisons among biased and unbiased methods cannot be fair, the power of OLR and OLMM under population stratification was not computed. Regarding genetically homogeneous populations, in both main genetic and gene–treatment interaction effects, ordinary regression models are the most powerful methods (Figures 1 and 2). Moreover, and as was expected, the use of repeated measures (OLMM) is more powerful than the use of a single measure (OLR). Note that, among the family-based models, ALMM is the most powerful and that FBAT or QBAT-I statistics and AQTDT$_M$ are equivalent with regard to power.

When the population is genetically mixed, all methods, regardless of the modelling approach or the design, lose power in comparison with the setting where the population is homogeneous (Tables 4 and 5). Once again, ALMM is the most powerful method for detecting main genetic or the interaction effects. In other words, the modelling

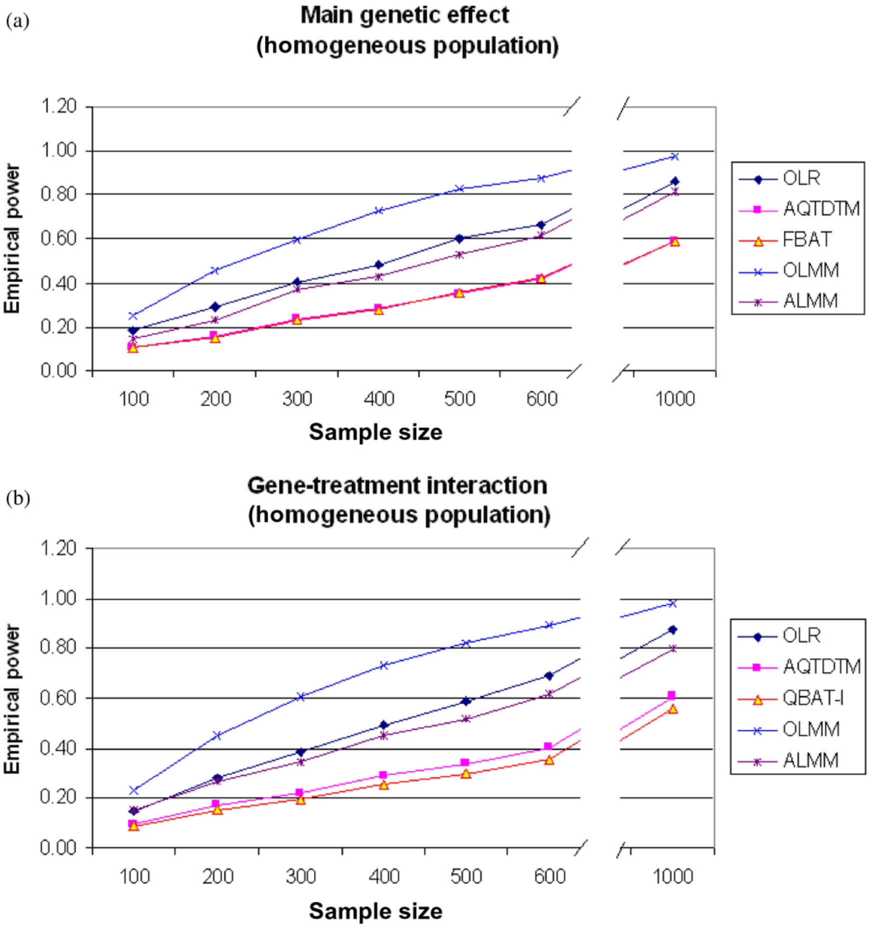

Figure 1. (a). Empirical power results for main genetic effect assessment comparing different methods under the assumption of a homogeneous population. Simulations are based on the sample size as indicated in the plot and the additive genetic model. (b) Empirical power results for gene–environment interaction effect comparing different methods under the assumption of a homogeneous population. Simulations are based on the sample size as indicated in the plot and the additive genetic model.

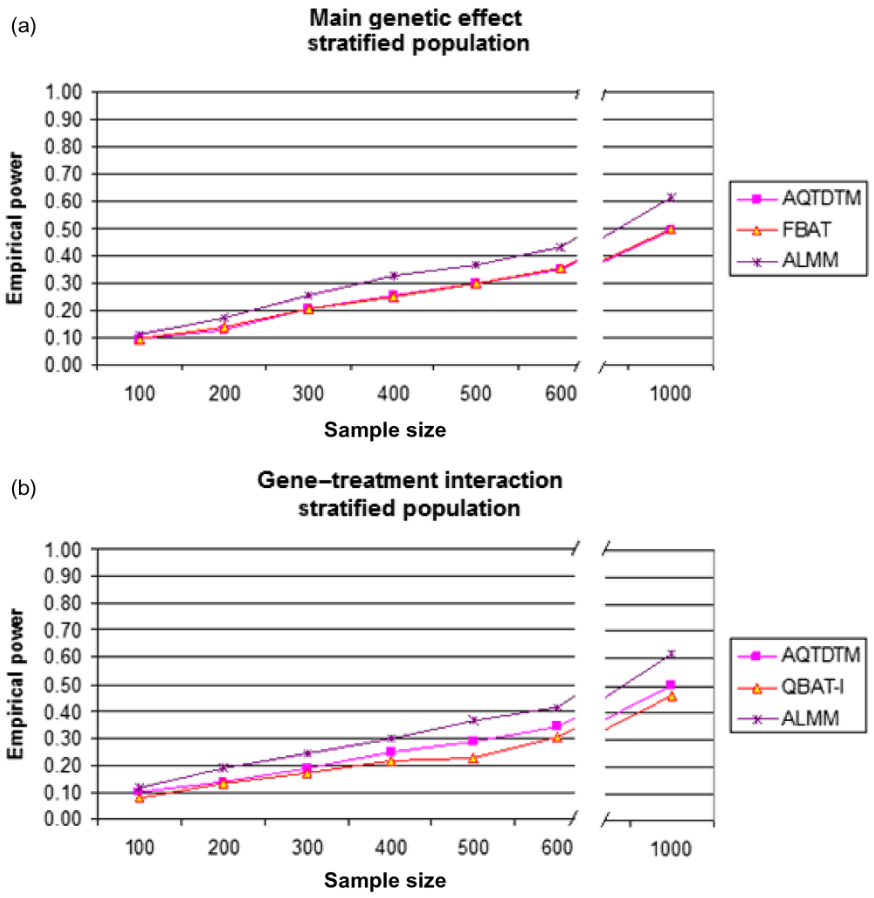

Figure 2. (a). Empirical power results for the main genetic effect assessment comparing different methods under the assumption of a stratified population. Simulations are based on the sample size as indicated in the plot and the additive genetic model. (b). Empirical power results for the gene–environment interaction effect comparing different methods under the assumption of a stratified population. Simulations are based on the sample size as indicated in the plot and the additive genetic model.

approach may be crucial, in terms of power, for conducting the association analysis.

Additive versus recessive disease models

In the recessive model, the same relationship among methods is observed, although the additive model is always more powerful than the recessive one, regardless of the testing approach. The data are shown in the supplementary tables. This may be related to the fact that in the additive model, X has a wider range of variation, whereas in the recessive one, X is an indicator variable. In addition, the number of informative families for FBAT statistics is always larger when the additive disease model is assumed, while the number with a causal genetic mutation is smaller under the recessive model assumption.

In summary, if the study population is genetically homogeneous, an independent subjects design provides unbiased genetic estimates, regardless of the modelling approach, and offers the most powerful tests as well. The models that use repeated measurements are even more powerful than those using one single outcome per subject, however. When the study population is stratified, using OLR or OLMM can result in spurious associations; therefore, in order to control for potential population admixture or stratification, family-based designs are strongly recommended. ALMM is more powerful than $QTDT_M$ and FBAT statistics.

Table 4. Empirical power results for main genetic effect assessment comparing ordinary statistical methods (OLR and OLMM) to family-based methods (AQTDT$_M$, FBAT and ALMM) under homogeneous (HP) and stratified (SP) populations. Each time, *n* cases were simulated with parameters $\beta_1 = \alpha_5 = 0.5$. Simulations are based on the additive genetic model. † = number that identifies each model in the paper.

n	$H_0 : \beta_1 = 0$						$H_0 : \alpha_5 = 0$			
	Two-step-models						**Mixed models**			
	OLR (16) †		**AQTDT$_M$ (17)**		**FBAT (18)**		**OLMM (19)**		**ALMM (20)**	
	HP	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**
100	0.184	-	0.107	0.093	0.108	0.095	0.254	-	0.147	0.113
200	0.289	-	0.156	0.129	0.155	0.140	0.456	-	0.232	0.171
300	0.406	-	0.237	0.206	0.234	0.207	0.595	-	0.371	0.253
400	0.481	-	0.285	0.256	0.281	0.25	0.726	-	0.429	0.325
500	0.606	-	0.353	0.297	0.355	0.296	0.831	-	0.531	0.365
600	0.663	-	0.415	0.346	0.422	0.351	0.874	-	0.614	0.429
10000	0.864	-	0.589	0.493	0.589	0.495	0.976	-	0.813	0.613

OLR, ordinary linear regression; OLMM, ordinary linear mixed model; ALMM, adjusted linear mixed model; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators

Table 5. Empirical power results for gene–environment interaction effect assessment comparing ordinary statistical methods (OLR and OLMM) to family-based methods (AQTDT$_M$, QBAT-I and ALMM) under homogeneous (HP) and stratified populations (SP). Each time, *n* cases were simulated with parameters $\beta_3 = \alpha_7 = 1$. Simulations are based on the additive genetic model. † = number that identifies each model in the paper.

n	$H_0 : \beta_3 = 0$						$H_0 : \alpha_7 = 0$			
	Two-step-models						**Mixed models**			
	OLR (21) †		**AQTDT$_M$ (22)**		**QBAT-I**		**OLMM (23)**		**ALMM (24)**	
	HP	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**
100	0.147	-	0.094	0.098	0.091	0.080	0.234	-	0.155	0.116
200	0.282	-	0.172	0.138	0.156	0.134	0.452	-	0.267	0.186
300	0.386	-	0.220	0.188	0.195	0.172	0.607	-	0.347	0.242
400	0.495	-	0.289	0.251	0.256	0.216	0.731	-	0.450	0.301
500	0.589	-	0.338	0.290	0.299	0.226	0.820	-	0.515	0.363
600	0.688	-	0.401	0.340	0.358	0.304	0.889	-	0.620	0.414
1000	0.875	-	0.604	0.503	0.561	0.465	0.978	-	0.794	0.618

OLR, ordinary linear regression; OLMM, ordinary linear mixed model; ALMM, adjusted linear mixed model; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators

Discussion

The ALMM method combines the characteristics of both longitudinal data analysis and case–parent design. While repeated measurements of quantitative phenotypes allow for the assessment of the effect of time-dependent environmental exposures, the use of the case–parent design with analysis based on ALMM eliminates the potential bias in the estimated coefficients associated with population admixture or stratification, provided that the linear model is correct. Unbiased tests also require valid estimates of standard errors, however.[8] Use of the ALMM model, which allows intercepts and environment effects to depend upon parental mating type, can help to address that issue.[14] Our results show that ALMM represents a valuable methodology for correctly assessing main and gene–environment interaction effects for quantitative traits in stratified populations. In addition, by taking advantage of the structure of the ordinary linear mixed-effects models, covariates may be included and balanced repeated measurements are not required. This model can be implemented using standard statistical software, including a linear mixed models module.

Since ALMM is based on a case–parent design, ethnicity bias is avoided because all possible genotypes are taken into account, even those that were not transmitted to the affected offspring. Including an indicator variable for the mating type allows one to use different intercepts; thus, differences within and across mating-types are considered in the genetic effects estimation. In order to account for a potential correlation between the exposure and the mating type, different levels of exposure, depending on the mating type, have been modelled in ALMM through the inclusion of the interaction between such an indicator variable and the exposure. In this manner, situations where, for example, the environmental exposure may depend on the mating type can be assessed without bias. Therefore, population stratification and admixture are no longer sources of estimation bias.

It can be the case that the study population is mixed, although the trait of interest does not vary within the subpopulations. In those situations, ethnicity is not a confounder; thus, genetic effects may be estimated without bias through the use of ordinary regression models. When there is an admixed population, and the exposure of interest does not depend on substructure, the indicator variable for the mating type (α_{1m}) can be omitted in model (13). In the case where Z is randomised, it is known that Z is independent of exposure and genetic background; allowing α_{6m} to depend on m ensures that the effect of treatment–genetic interaction is not confounded by different responses to treatment in the different subgroups. If it can be assumed that both coefficients do not depend upon m, a simpler model can be fit, which has two clear advantages. First, there are more degrees of freedom, and this is important in small studies, especially if some strata of mating types have few observations. Secondly, if parents are missing, instead of computing $E(X)$ conditional on parental genotypes, we can replace it by $E(X)$ conditional on the sufficient statistics for parental genotype.[8]

A disadvantage of case–parent designs is that parental genotypes are not easily accessible for late-onset disorders. In those cases, other family-based designs suggest using siblings rather than parents, although larger sample sizes are required in order to achieve comparable power.[21]

It is important to note that, because OLR and OLMM provide biased estimators under population stratification, power comparison against unbiased methods may not be completely fair. Power will be underestimated when the parameter is incorrectly estimated with values that are close to zero, although when the reverse occurs, the power will be magnified. For that reason, it was decided to exclude those methods in power comparisons under population stratification.

For testing both main genetic effect and gene–environment interaction effects, regardless of the composition of the population, ALMM was found to be more powerful than the two-step modelling approach where $\mathrm{AQTDT_M}$ and FBAT — or QBAT-I, in the gene–environment interaction assessment — were used in the second stage. This is because, while the longitudinal analysis approach takes advantage of both repeated values across time and measurements across people, the two-step procedure does not account for the relative degree of within- and between-subject variability. Nevertheless, there are weighting procedures that account for both sources of variability; thus, the summary statistic obtained in the first step can be adjusted for.[22] Although, methodologically, the linear mixed models represent an adequate approach for longitudinal data analysis, one should not forget about the two-step modelling approach because it represents an intuitively simpler procedure and the opportunity to use existing genetic software. It should be noted that the difference in power between both modelling approaches may strongly depend on the number of repeated measurements, the underlying true effect sizes and the frequency of the missing phenotypes.

It is evident that, with a homogeneous population, OLMM is the most powerful tool. The decrement in the power of ALMM relative to OLMM is related to the lessening of the variability in the genotype due to the centring procedure and to the extra parameters, for each mating type, to be estimated. When the population is not homogeneous, however, these factors should not represent a disadvantage when contrasted with the added advantage of an unbiased estimation.

In summary, in addition to comparing the longitudinal approach against the two-step modelling approach, we have also compared designs using independent subjects against family-based approaches under homogeneous and stratified populations. Assuming no population stratification, ordinary regression methods are valid and more powerful than the other methods. Nevertheless, the family-based approach is strongly recommended when the homogeneous ethnicity in the population is in doubt, in order to achieve unbiased estimators. ALMM now represents a powerful tool for assessing the main genetic effect and gene–environment interactions on time-dependent phenotypes under population stratification.

Acknowledgments

Dr London was supported by the Division of Intramural Research, National Institutes of Health, Department of Health and Human Services (ES049019). Dr Laird was supported by the National Institute of Mental Health. The authors thank Douglas Dockery, ScD and Diane Gold, ScD from the Environmental Health Department, Harvard School of Public Health, for their invaluable suggestions and comments.

References

1. Martinez, F.D. (2007), 'CD14, endotoxin, and asthma risk: Actions and interactions', *Proc. Am. Thorac. Soc.* Vol. 4, pp. 221–225.
2. Hunter, D. (2005), 'Gene-environment interaction in human diseases', *Nat. Rev. Genet.* Vol. 6, pp. 287–298.
3. Li, C.C. (1969), 'Population subdivision with respect to multiple alleles', *Ann. Hum. Genet.* Vol. 33, pp. 23–29.
4. Deng, H.W. and Chen, W. M. (2001), 'The power of the transmission disequilibrium test (TDT) with both case–parent and control-parent trios', *Genet. Res.* Vol. 78, pp. 289–302.
5. Choudhry, S., Seibold, M.A., Borrell, L.N., Tang, H. *et al.* (2007), 'Dissecting complex diseases in complex populations: Asthma in Latino Americans', *Proc. Am. Thorac. Soc.* Vol. 4, pp. 226–233.
6. Almasy, L. (2001), 'Introduction: Methods for detecting genotype X environment interaction', *Genet. Epidemiol.* Vol. 21, pp. S817–S818.
7. Tian, C., Gregersen, P.K. and Seldin, M.F. (2008), 'Accounting for ancestry: Population substructure and genome-wide association studies', *Hum. Mol. Genet.* Vol. 17(R2): R143–R150.
8. Rabinowitz, D. and Laird, N. (2000), 'A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information', *Hum. Hered.* Vol. 50, pp. 211–223.
9. Gauderman, W.J., Macgregor, S., Briollais, L., Scurrah, K. *et al.* (2003), 'Longitudinal data analysis in pedigree studies', *Genet. Epidemiol.* Vol. 52, pp. S18–S58.
10. Romieu, I., Sienra-Monge, J.J., Ramírez-Aguilar, M., Moreno-Macías, H.R. *et al.* (2004), 'Genetic polymorphism of *GSTM1* and antioxidant supplementation influence lung function in relation to ozone exposure in asthmatic children in Mexico City', *Thorax* Vol. 59, pp. 8–10.
11. Laird, N.M., Horvath, S. and Xu, X. (2000), 'Implementing a unified approach to family-based tests of association', *Genet. Epidemiol.* Vol. 19, pp. S36–S42.
12. Allison, D.B. (1997), 'Transmission-disequilibrium tests for quantitative traits', *Am. J. Hum. Genet.* Vol. 60, pp. 676–690.
13. Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993), 'Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *Am. J. Hum. Genet.* Vol. 52, pp. 506–516.
14. Ewens, W.J., Li, M. and Spielman, R.S. (2008), 'A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker', *PLoS Genet.* Vol. 26, p. e1000180.
15. Gauderman, W.J. (2003), 'Candidate gene association analysis for a quantitative trait, using parent-offspring trios', *Genet. Epidemiol.* Vol. 25, pp. 327–338.
16. Laird, N.M. and Lange, C. (2006), 'Family-based designs in the age of large-scale gene-association studies'. *Nat. Rev. Genet.* Vol. 7, pp. 385–394.
17. Vansteelandt, S., Demeo, D.L., Lasky-Su, J., Smoller, J.W. *et al.* (2008), 'Testing and estimating gene-environment interactions in family-based association studies', *Biometrics* Vol. 64, pp. 458–467.
18. Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994), *Analysis of longitudinal data*, Oxford University Press, Inc., New York, NY.
19. London, S.J. and Romieu, I. (2009), 'Gene by environment interaction in asthma', *Annu. Rev. Public Health* Vol. 30, pp. 55–80.
20. Fitzmaurice, G., Laird, N.M. and Ware, J. (2004), *Applied Longitudinal Analysis*, John Willey and Sons, Inc., Haboken, NJ.
21. Abecasis, G.R., Cardon, L.R. and Cookson, W.O. (2000), 'A general test of association for quantitative traits in nuclear families', *Am. J. Hum. Genet.* Vol. 66, pp. 279–292.
22. Korn, E.L. and Whittemore, A.S. (1979), 'Methods for analyzing panel studies of acute health effects of air pollution', *Biometrics* Vol. 35, pp. 795–802.

Appendix I

R code for ALMM

```
library(nlme)
mmodel < -lme(fef2575 ~ o3*tx*as.factor(mating_
type) + gstm1*tx*o3,      random = ~1 + o3|id,
method = "ML", data = base)
```

where:
fef2575 = outcome
o3 = ozone exposure (time-dependent)
tx = supplementation treatment (fixed)
gstm1 = genotype (takes values 0, 1, 2 if additive disease model; or 0,1 if recessive)
mating_type = vector with values 1, 2, 3, 4, 5, 6 classifying the different mating types.
id = individual identification

R code for the second stage in the two-step modelling approach

$QTDT_M$

```
qtdtm < -lm(slope ~ tx*as.factor(mating_type) +
tx*gstm1, data = base1)
```

FBAT

```
library (pbatR)
pbat.m(slope ~ tx | gstm1, ped = ped, phe = phe,
   fbat = "gee",min.info = 10, max.pheno = 1, scan.
   genetic = "additive")
```

QBAT-I

```
library (pbatR)
pbat.m(slope ~ mi(tx) | gstm1, ped = ped, phe =
   phe, fbat = "gee",min.info = 10, max.pheno = 1,
   scan.genetic = "additive")
```

where:
gstm1, tx and mating_type are defined as above

More details about the PBAT commands can be found in http://biosun1.harvard.edu/~clange/pbat.htm

General FBAT statistic

For N nuclear families, one offspring in the family *i* and no covariates

$$\chi^2_{FBAT} = \frac{U^2}{Var(U)}$$

where:

$$U = \sum[(Y_i - E(Y_i))*(\mathrm{X_i} - \mathrm{E}(\mathrm{X_i}|\mathrm{g_{im}}, \mathrm{g_{if}}))]$$
$$i = 1, 2, \ldots N;$$

$$Var(U) = \sum_i (Y_i - E(Y_i))^2 * \mathrm{Var}(\mathrm{X_i}|\mathrm{g_{im}}, \mathrm{g_{if}}),$$

and $E(X_i|g_{im}, g_{if})$ and $Var(X_i|g_{im}, g_{if})$ are calculated under the null hypothesis of Mendel's law. That is:

$$E(X_i|g_{im}, g_{if_i}) = \sum_g X(g)P(g)$$

and

$$Var(X_i|g_{im}, g_{if}) = \left[\sum_g X^2(g)P(g)\right] - \left[\sum_g X(g)P(g)\right]^2,$$

where g on the right hand side of these expectations indexes the possible offspring genotypes and $P(g)$ is the probability of a particular genotype given the parents' genotypes, calculated under the null hypothesis. Thus,

$$\chi^2_{FBAT} \sim \chi^2_{1df}.$$

If both parents are homozygous, $X_i = E(X_i|g_{im}, g_{if})$ and $Var(X_i|g_{im}, g_{if}) = 0$. Therefore, these triads do not add information to the FBAT statistic and they are referred to as non-informative families.

The test is robust against population stratification, as a result of centring X by its expected value conditional on parental genotypes (g_{im}, g_{if}) assuming Mendel's laws.

The statement that case selection was not based on their genotype information is the only assumption about the ascertainment process.

Since in U, $E(Y_i)$ is calculated under the null hypothesis, it can be estimated by $\bar{Y}$. Note that the test statistic is based on the relative size of U with respect to its standard deviation but not on the size of β_1 explicitly. Thus, the genetic effect is not directly estimated.

QBAT-I

This statistic[17] is based on the following regression model:

$$E(Y_i|X_i, Z_i, S_i) = \beta_0(Z_i, S_i) + \beta_1 X_i + \beta_2 X_i z_i$$

$$= \beta_0 + (\beta_1, \beta_2) * \begin{pmatrix} \mathrm{X_i} \\ \mathrm{X_i * Z_i} \end{pmatrix} \quad (14)$$

where:

$S_i = (g_{im}, g_{if})$ Sufficient statistic (parental genotypes)
$\beta_0(Z_i, S_i)$ encodes the dependence of the outcome on the environmental exposure and the parental genotypes
β_1 = main genetic effect
β_2 = gene–environment interaction effect

And X_i and z_i are as defined previously.

Note that there is no coefficient for the environmental effect, as this is subsumed in the intercept β_0. Assuming that the environmental exposure is independent of the candidate gene, and conditional on S_i, estimators for both β_1 and β_2 are obtained through the equation:

$$\sum_i U_i(\beta) = \sum_i \begin{pmatrix} X_i - E(X_i|g_{im}, g_{if}) \\ X_i * \mathrm{Z_i} - \mathrm{E(X_i|g_{im}, g_{if}) * Z_i} \end{pmatrix}$$

$$* \mathrm{e_i} = 0$$

where:

$$e_i = Y_i - E(Y_i|X_i, Z_i, S_i).$$

Under weak regularity conditions, the solution to this equation leads to consistent estimators for $\beta = (\beta_1, \beta_2)$ which are robust for population stratification.

The test statistic for the gene–environment interaction has the same form as the original FBAT statistic given in (12); that is:

$$QBAT - I = \frac{U^2}{Var(U)} \sim \chi_1^2 \quad (15)$$

where:

$$U = \sum_{i=1}^{N} \{X_i - E(X_i|g_{im}, g_{if})\}(Z_i - \hat{\mu}_Z) e_i((\hat{\beta}_1, 0))\}$$

with:

$$\hat{\beta}_1 = \left\{ \sum_i \{X_i - E(X_i|g_{im}, g_{if})\} X_i^t \right\}^{-1}$$

$$\times \left\{ \sum_i \{X_i - E(X_i|g_{im}, g_{if})\} e_{ij}(0) \right\}$$

and

$$\hat{\mu}_Z = \left\{ \sum_i Z_i h_i((\hat{\beta}_1, 0)) \{X_i - E(X_i)|g_{im}, g_{if})\} X_i^t \right\}$$

$$\times \left\{ \sum_{i,j} \{X_i - E(X_i|g_{im}, g_{if})\} X_i^t \right\}^{-1}$$

$\hat{\beta}_1$ is an estimate for the main genetic effect under the null hypothesis of no gene–environment interaction and $\hat{\mu}_Z$ is a weighted average of the environmental exposures that ensures QBAT-I $\sim \chi_1^2$.

Note that, in (14), the point of attention is on the genetic effect through the main and the gene–environment interaction. In other words, the parental genotype and the environment main effect are not of direct interest for estimation. In this sense, the test of $H_0 : \beta_3 = 0$ based on model (9) may be thought of as an equivalent test to QBAT-I

Table S1. List of models used in the simulation process. The column model refers to the number that identifies each model in the paper. X_i is a fixed variable that translates an offspring genotype to a numeric value; Z_i is an observed environmental covariate, either continuous or dichotomous; g_{im}, g_{if} are the parental genotypes (mother and father, respectively); $E(X_i|g_{im}, g_{if})$ is calculated under segregation and independent assortment assumptions using Mendel's law; $M = 1, 2, \ldots, 6$ are the six possible mating types; $i = 1, 2, 3, \ldots, n$ subjects; $j = 1, 2, 3, \ldots, m$ measurement occasions into the subject; t_{ij} is the repeated ozone exposure variable.

Generating models			
Model	**Linear mixed model**	**Parameters**	**Allele frequencies**
(15)	$E(Y_{ij} \mid X, Z, t) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij}$ $+ \alpha_6 Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij}$		
Homogeneous population			
		$\alpha_1 = -0.8; \alpha_2 = -0.2; \alpha_3 = -0.05; \alpha_4 = 0.2;$ $\alpha_6 = 0.6; \alpha_7 = 1.0$	P(A) = 0.6 P(*a*) = 0.4
Admixed population			
		$\alpha_1 = -0.8; \alpha_2 = -0.2; \alpha_3 = -0.05; \alpha_4 = 0.2;$ $\alpha_6 = 0; \alpha_7 = 1.0$	Population 1 P(A) = 0.4 P(*a*) = 0.6
		$\alpha_1 = -0.8; \alpha_2 = -0.2; \alpha_3 = -0.05; \alpha_4 = 0.2;$ $\alpha_6 = 2.0; \alpha_7 = 1.0$	Population 2 P(A) = 0.8 P(*a*) = 0.2
Assessing models			
Model	**Independent subjects design**	**Case-parent design**	**Comments**
Two-step modelling approach			
Main genetic effect			
(16)	$E(slope_i \mid X, Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i$		Taken from model (2)
(17)		$E(slope_i \mid X, Z) = \beta_{0M} + \beta_1 X_i + \beta_2 Z_i$	Taken from model (5)
(18)		$\chi^2_{FBAT} = \frac{\sum[(slope_i - E(slope_i)) * (X_i - E(X_i \mid g_{im}, g_{if}))]}{\sum_i (slope_i - E(slope_i))^2 * Var(X_i \mid g_{im}, g_{if})}$	Taken from model (10)

Continued

Table S1. Continued

Generating models			
Model	**Linear mixed model**	**Parameters**	**Allele frequencies**
Gene–environment interaction			
(21)	$E(slope_i \| X, Z) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 Z_i X_i$		Taken from model (1)
(22)		$E(slope_i \| X, Z) = \beta_{0M} + \beta_1 X_i + \beta_{2M} Z_i + \beta_3 Z_i X_i$	Taken from model (9)
Models for longitudinal data			
Main genetic effect			
(19)	$E(Y_{ij} \| X, Y, t) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_6 Z_i t_{ij}$		Taken from model (15) with $\alpha_7 = 0$
(20)		$E(Y_{ij} \| X, Y, t) = \alpha_{0M} + \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_{6M} Z_i t_{ij}$	Taken from model (13) with $\alpha_7 = 0$
Gene–environment interaction			
(23)	$E(Y_{ij} \| X, Z, t) = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_6 Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij}$		Taken from model (15)
(24)		$E(Y_{ij} \| X, Y, t) = \alpha_{0M} + \alpha_{1M} t_{ij} + \alpha_2 Z_i + \alpha_3 X_i + \alpha_4 X_i Z_i + \alpha_5 X_i t_{ij} + \alpha_{6M} Z_i t_{ij} + \alpha_7 X_i Z_i t_{ij}$	Taken from model (13)

Table S2. Observed effects in the real cohort study conducted in Mexico City. 95% CI = 95% confidence interval.

Group	Subgroup	*n*	Coefficient (95% CI*)
Placebo			
	GSTM1 null	29	−1.01 (−1.80, −0.22)
	GSTM1 positive	49	−0.21 (−0.77, 0.35)
	Genotype effect	78	0.80 (−0.15, 1.75)
Supplement			
	GSTM1 null	33	−0.06 (−0.81, 0.69)
	GSTM1 positive	47	0.10 (−0.60, 0.80)
	Genotype effect	80	0.16 (−0.88, 1.02)
GSTM1 null	Placebo	29	−1.01 (−1.80, −0.22)
	Supplement	33	−0.06 (−0.81, 0.69)
	Supplement effect	62	0.95 (−0.14, 2.04)
GSTM1 positive	Placebo	49	−0.21 (−0.77, 0.35)
	Supplement	47	0.10 (−0.60, 0.80)
	Supplement effect	96	0.31 (−0.58, 1.21)
Simulation parameters:			
$\alpha_5 = \frac{0.8(78) + 0.16(80)}{158} \approx 0.5$			
$\alpha_6 = \frac{0.95(62) + 0.31(96)}{158} \approx 0.6$			
$\alpha_7 = 0.95 \approx 1$			

This Table is based on results previously published in Thorax, Vol. 59 (2004)[10].

Table S3a. Bias results for main genetic effect assessment comparing ordinary statistical methods (OLR and OLMM) with family-based methods (AQTDT$_M$ and ALMM) under homogeneous (HP) and stratified (SP) m populations. Each time, *n* cases were simulated with parameters $\beta_1 = \alpha_5 = 0.5$. Simulations are based on the recessive genetic model. † = number that identifies each model in the paper.

n	$\hat{\beta}_1 - \beta_1$				$\hat{\alpha}_5 - \alpha_5$			
	OLR (16) †		**AQTDT$_M$ (17)**		**OLMM (19)**		**ALMM (20)**	
	HP	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**
100	−0.010	−0.505	−0.012	−0.049	−0.026	−0.508	−0.023	−0.052
200	−0.022	−0.507	−0.049	−0.040	−0.019	−0.494	−0.031	−0.023
300	−0.002	−0.500	−0.003	0.028	−0.010	−0.505	−0.004	0.025
400	−0.020	−0.499	−0.018	0.002	−0.008	−0.496	−0.012	0.006
500	0.007	−0.494	0.012	0.004	0.005	−0.495	0.006	−0.002
600	0.000	−0.496	0.010	0.003	−0.001	−0.499	0.001	0.001

OLR, ordinary linear regression; OLMM, ordinary linear mixed models; ALMM, adjusted linear mixed models; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators

Table S3b. Bias results for gene–environment effect assessment comparing ordinary statistical methods (OLR and OLMM) with family-based methods (AQTDT$_M$ and ALMM) under homogeneous (HP) and stratified (SP) populations. Each time, *n* cases were simulated with parameters $\beta_3 = \alpha_7 = 1$. Simulations are based on the recessive genetic model. † = number that identifies each model in the paper.

n	$\hat{\beta}_3 - \beta_3$				$\hat{\alpha}_7 - \alpha_7$			
	OLR (21) †		**AQTDT$_M$ (22)**		**OLMM (23)**		**ALMM (24)**	
	HP	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**
100	0.008	−0.962	0.001	0.048	0.006	−1.003	0.017	0.027
200	−0.008	−1.042	0.020	−0.007	−0.001	−1.024	0.019	−0.006
300	0.011	−1.005	0.048	0.001	0.009	−1.004	0.012	0.010
400	0.014	−1.001	0.019	0.022	0.001	−0.988	−0.009	0.034
500	0.011	−1.010	0.000	−0.028	−0.002	−1.010	−0.012	−0.006
600	0.016	−0.967	0.016	0.033	0.015	−0.973	0.008	0.032

OLR, ordinary linear regression; OLMM, ordinary linear mixed models; ALMM, adjusted linear mixed models; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators

Table S4a. Empirical power results for main genetic effect assessment comparing ordinary statistical methods (OLR and OLMM) with family-based methods (AQTDT$_M$, FBATand ALMM) under homogeneous (HP) and stratified (SP) populations. Each time, *n* cases were simulated with parameters $\beta_1 = \alpha_5 = 0.5$. Simulations are based on the recessive genetic model. † = number that identifies each model in the paper.

n	$H_0 : \beta_1 = 0$						$H_0 : \beta_5 = 0$			
	OLR (16) †		**AQTDT$_M$ (17)**		**FBAT (18)**		**OLMM (19)**		**ALMM (20)**	
	HP	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**	**HP**	**SP**
100	0.108	-	0.074	0.074	0.075	0.072	0.146	-	0.116	0.09
200	0.155	-	0.104	0.098	0.103	0.101	0.259	-	0.166	0.118
300	0.237	-	0.161	0.128	0.158	0.128	0.359	-	0.236	0.164
400	0.255	-	0.182	0.143	0.178	0.144	0.469	-	0.295	0.18
500	0.357	-	0.226	0.176	0.227	0.175	0.552	-	0.342	0.221
600	0.411	-	0.275	0.206	0.278	0.275	0.605	-	0.417	0.255

OLR, ordinary linear regression; OLMM, ordinary linear mixed models; ALMM, adjusted linear mixed models; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators

Table S4b. Empirical power results for gene−environment effect assessment comparing ordinary statistical methods (OLR and OLMM) with family-based methods (AQTDT$_M$, QBAT-I and ALMM) under homogeneous (HP) and stratified (SP) populations. Each time, *n* cases were simulated with parameters $\beta_3 = \alpha_7 = 1$. Simulations are based on the recessive genetic model. † = number that identifies each model in the paper.

n	$H_0 : \beta_3 = 0$						$H_0 : \beta_7 = 0$			
	OLR (21) †		**AQTDT$_M$ (22)**		**QBAT-I**		**OLMM (23)**		**ALMM (24)**	
	HP	**AP**	**HP**	**AP**	**HP**	**AP**	**HP**	**AP**	**HP**	**AP**
100	0.099	-	0.067	0.078	0.070	0.065	0.151	-	0.118	0.098
200	0.160	-	0.127	0.111	0.119	0.109	0.252	-	0.180	0.144
300	0.220	-	0.150	0.124	0.130	0.120	0.374	-	0.219	0.155
400	0.296	-	0.215	0.132	0.158	0.129	0.466		0.302	0.172
500	0.345	-	0.241	0.160	0.182	0.161	0.535	-	0.342	0.216
600	0.422	-	0.252	0.192	0.201	0.190	0.633	-	0.402	0.260

OLR, ordinary linear regression; OLMM, ordinary linear mixed models; ALMM, adjusted linear mixed models; AQTDT$_M$, adjusted quantitative transmission disequilibrium test with mating type indicators