



Published in final edited form as:

Hum Hered. 2012 ; 73(4): 185–194. doi:10.1159/000339906.

Natural and Orthogonal Interaction framework for modeling gene-environment interactions with application to lung cancer

Jianzhong Ma^a, Feifei Xiao^a, Momiao Xiong^b, Angeline S Andrew^c, Hermann Brenner^d, Eric J. Duell^e, Aage Haugen^f, Clive Hoggart^g, Rayjean J. Hung^h, Philip Lazarusⁱ, Changlu Liu^a, Keitaro Matsuo^j, Jose Ignacio Mayordomo^k, Ann G. Schwartz^l, Andrea Staratschek-Jox^m, Erich Wichmannⁿ, Ping Yang^o, and Christopher I. Amos^a

^aDepartment of Genetics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

^bHuman Genetics Center, University of Texas School of Public Health, Houston, Texas, USA

^cDepartment of Community and Family Medicine, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH, USA

^dDivision of Clinical Epidemiology and Aging Research, German Cancer Research Center, Im Neuenheimer Feld 581, 69120, Heidelberg, Germany

^eUnit of Nutrition, Environment and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO-IDIBELL), Barcelona, Spain

^fThe National Institute of Occupational Health, P.O. Box 8149 Dep. N-0033 Oslo 1, Norway

^gEpidemiology Unit, London School of Hygiene and Tropical Medicine, UK

^hSamuel Lunenfeld Research Institute, 60 Murray St. Toronto ON M5T 3L9 Canada

ⁱDepartments of Pharmacology and Public Health Sciences, Penn State College of Medicine

^jDivision of Epidemiology and Prevention, Aichi Cancer Center Research Institute, 1-1 Kanokoden, Chikusa-ku, Nagoya 464-8681, Japan

^kServicio de Oncología Medica, Hospital Clinico Universitario, Av. San Juan Bosco, 15 50009 Zaragoza, Spain

^lKarmanos Cancer Institute and Wayne State University School of Medicine, Department of Oncology, 4100 John R, Detroit, MI 48201, USA

^mLife and Medical Sciences Bonn, Genomics and Immunoregulation, University of Bonn, Bonn, Germany

ⁿHelmholtz Zentrum Munchen, Deutsches Forschungszentrum fur Gesundheit und Umwelt (GmbH), Ingolstadter Landstr. 1, 85764 Neuherberg, Germany

^oMayo Clinic Cancer Center, 200 First street SW, Rochester, MN 55905, USA

Abstract

Objectives—We aimed at extending the natural and orthogonal interaction (NOIA) framework, developed for modeling gene-gene interactions in the analysis of quantitative traits, to allow for reduced genetic models, dichotomous traits, and gene-environment interactions. We evaluate the performance of the NOIA statistical models using simulated data and lung cancer data.

Methods—The NOIA statistical models are developed for the additive, dominant, recessive genetic models, and a binary environmental exposure. Using the Kronecker product rule, a NOIA statistical model is built to model gene-environment interactions. By treating the genotypic values as the logarithm of odds, the NOIA statistical models are extended to the analysis of case-control data.

Results—Our simulations showed that power for testing associations while allowing for interaction using the statistical model is much higher than using functional models for most of the scenarios we simulated. When applied to the lung cancer data, much smaller P-values were obtained using the NOIA statistical model for either the main effects or the SNP-smoking interactions for some of the SNPs tested.

Conclusion—The NOIA statistical models are usually more powerful than the functional models in detecting main effects and interaction effects for both quantitative traits and binary traits.

Keywords

Statistical power; Genetic association studies; Case-control association analysis; Gene-environment interaction; Environmental risk factor; Association mapping; Orthogonal modeling

Introduction

Although genome-wide association studies (GWAS) have been successful in identifying disease susceptibility loci, it remains a challenging goal for statistical geneticists to identify and characterize effects of genetic and environmental factors that influence common complex traits [1, 2]. The significant SNP associations identified by GWAS are estimated to account for only a few percent of the genetic variance [3, 4]. Since most of these studies have used a single-locus analysis strategy, there is increasing interest in genome-wide interaction analysis in the efforts of finding the missing heritability. An important issue is how one can properly quantify the genetic and environmental effects in a unified manner such that correct estimation can be achieved for the relative contributions of different factors. This is the case especially when there exist gene-gene or gene-environment interactions.

The term GxE interaction has various different meanings. The nature of an interaction may be biological or merely statistical. A statistical interaction may be defined as departure from an additive model on some scale. For a quantitative trait, if the contributions of a genetic locus and an environment factor are additive on the scale of this trait, we say that there is no interaction between the locus and the environment factor. Any deviation from this additivity will be referred to as a GxE interaction. For a binary trait (such as a disease), we define the GxE interaction as deviation from additivity of a genetic effect and an environment effect on the log-odds scale.

The statistical formulation of the natural and orthogonal interaction (NOIA) model of genetic effects, were recently developed [5] and provides a framework in which estimates of genetic effects for a quantitative trait remain orthogonal even under the departure from Hardy-Weinberg proportions (HWP) of the loci. The orthogonal estimates do not change in a reduced model and hence are very convenient for model selection for finding the genetic architecture of the traits[6]. More importantly, the NOIA framework directly leads to a proper and orthogonal decomposition of genetic variance and hence a more meaningful calculation of the heritability of the trait, while facilitating the modeling of multiple genetic factors along with their interactions.

In this paper, we evaluate the performance of the orthogonal models, compared to the usual models, in both testing for interaction between two factors and testing for association while allowing for interaction. We extend the NOIA framework in several different ways. First, we derive the NOIA formulation to all possible reduced genetic models, including the additive, dominant and recessive models. We then extend the NOIA framework to include a binary environmental exposure and its interaction with a gene with a reduced or saturated genetic model. Finally, we also explore the possibility of generalizing the orthogonal models to the analysis of binary traits, such as diseases. We found that the meaning of orthogonality is somewhat different on the log-odds scale than its original meaning for a quantitative trait: although the estimators are no longer orthogonal, the variance decomposition remains orthogonal when the log-odds are simply treated as genetic effects under the alternative hypothesis of an effect in the NOIA formulation. Our simulation results showed that for both quantitative and qualitative traits, the statistical models have higher power than the usual functional ones in most of the scenarios we have tested. We also illustrate the usage of the formulations using real data.

Methods

One-locus models: quantitative traits

If a quantitative trait is influenced by a single diallelic gene, or locus,

$$Y = G + e, \quad (1)$$

the genotypic value can be modeled using linear regression on the number of a reference allele, say allele 2, as follows:

$$G = \mu + (N - \bar{N})\alpha + \varepsilon\delta, \quad (2)$$

where $G = G_{11}$, G_{12} and G_{22} are the genotypic values when the number of reference alleles is, $N = 0, 1$ and 2 , respectively, and

$$\varepsilon = \begin{cases} -2p_{12}p_{22}/V \\ 4p_{11}p_{22}/V \\ -2p_{11}p_{12}/V \end{cases} \quad (3)$$

with genotype frequencies p_{11} , p_{12} and p_{22} , respectively. Here \bar{N} and V denote the mean and variance of N

$$\begin{aligned} \bar{N} &= p_{12} + 2p_{22} \\ V &= p_{12} + 2^2p_{22} - (p_{12} + 2p_{22})^2 = p_{11} + p_{22} - (p_{11} - p_{22})^2, \end{aligned}$$

respectively. The genetic effect vector, \vec{E}_s , consists of the three regression parameters, $\mu = \bar{G}$, $\alpha = \text{COV}(N, G)/V$ and $\delta = G_{12} - (G_{11} + G_{22})/2$, and can be expressed as

$$\vec{E}_s \equiv \begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & p_{22} \\ p_{11} & p_{12} & p_{22} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix}, \quad (4)$$

if we define

$$p'_{ij} = p_{ij} \frac{N_{ij} - \bar{N}}{V}. \quad (5)$$

The vector of genotypic values, \vec{G} , can be expressed by

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = S_s \vec{E}_s = \begin{pmatrix} 1 & -\bar{N} & -2p_{12}p_{22}/V \\ 1 & 1 - \bar{N} & 4p_{11}p_{22}/V \\ 1 & 2 - \bar{N} & -2p_{11}p_{12}/V \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix}. \quad (6)$$

Here, S_s is the design matrix. In a regression analysis, we make inference on parameters (μ , α and δ). Therefore S_s can also be referred to as coding matrix for sampled data, because each row of S_s includes the code for regression on the parameters for an individual with the corresponding genotype. For example, if an individual has genotype 12, data for this individual will be coded according to the second row of S_s as follows:

$$G = \mu + (1 - \bar{N})\alpha + \frac{4p_{11}p_{22}}{V}\delta. \quad (7)$$

Unlike the genotypic values, \vec{G} , the genetic effects, \vec{E}_s , depend on genotype frequencies and the model based on these parameters are referred to as statistical model [5]. As shown in [5], this statistical model is orthogonal, meaning that estimates of these parameters are uncorrelated. The orthogonality of the statistical model is also reflected by the fact that the variance of G can be decomposed into those of the additive and dominant components

$$V_G = \text{Var}[(N - \bar{N})\alpha] + \text{Var}[\varepsilon\delta] \quad (8)$$

because $\text{COV}[(N - \bar{N})\alpha, \varepsilon\delta] = 0$. The additive and dominant variances can be expressed as

$$V_\alpha = \alpha^2 V \quad (9)$$

$$V_\delta = 4p_{11}p_{12}p_{22}\delta^2(p_{12}p_{22} + 4p_{11}p_{22} + p_{11}p_{12})/V. \quad (10)$$

Traditionally, the one-locus genotypes are coded in one of the following two ways: $(-1, 0, 1)$ and $(0, 1, 2)$. In either case, the genetic effects do not depend on allele frequencies and are determined merely by the functionality of the locus. So, both coding methods correspond to functional models. For these two coding schemes, the genotypic values are expressed as

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} R \\ a \\ d \end{pmatrix} = S_1 \vec{E}_1, \quad (11)$$

and

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} R \\ a \\ d \end{pmatrix} = S_2 \vec{E}_2, \quad (12)$$

respectively. These two functional models are related to the statistical model through

$$\begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix} = \begin{pmatrix} 1 & p_{22} - p_{11} & p_{12} \\ 0 & 1 & p_{12} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a \\ d \end{pmatrix}, \quad (13)$$

and

$$\begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix} = \begin{pmatrix} 1 & \bar{N} & p_{12} \\ 0 & 1 & p'_{12} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a \\ d \end{pmatrix}, \quad (14)$$

respectively. In contrast to the statistical model, neither of these two functional models is orthogonal.

One-locus models: qualitative traits

For the analysis of cas-control data sampled according to a qualitative trait such as disease, we can define a similar statistical model by treating the genotypic values and the genetic effects as the logit (i.e., logarithm of the odds) of the disease. However, two important features of the orthogonal models may no longer be valid here. First, the estimates of parameters using logistic regression are not uncorrelated. Recall that the variance of estimates of parameters for linear regression can be expressed as

$$\text{Var}(\vec{\beta}) = \sigma^2 (\chi^T \chi)^{-1}, \quad (15)$$

where χ is the design matrix, as far as the error terms for all samples are independent and identically distributed with variance σ^2 , which can be shown to be diagonal for the statistical model [5]. However, for logistic regression, the variance of estimates of parameters is

$$\text{Var}(\vec{\beta}) = (\chi^T \nu \chi)^{-1} \quad (16)$$

where ν is a diagonal matrix with elements

$$\pi_{G_i}(1 - \pi_{G_i}) \quad (17)$$

for the i th individual in the sample with π_{G_i} the probability of being affected given the values of regressor for the individual. It can be shown that

$$\chi^T \nu \chi = n S^T D' S \quad (18)$$

where

$$D' = \begin{pmatrix} \pi_{11}(1 - \pi_{11})p_{11} & 0 & 0 \\ 0 & \pi_{12}(1 - \pi_{12})p_{12} & 0 \\ 0 & 0 & \pi_{22}(1 - \pi_{22})p_{22} \end{pmatrix}, \quad (19)$$

and S is a design matrix. This means that, for logistic regression, the statistical model defined in (6) has no orthogonal estimates as in the case of linear regression, unless the gene is not associated with the disease (π_G would then assume the same values for all genotypes). Second, as will be shown later, the estimates of main effects for a full interaction model is no longer the same as the corresponding effects of the reduced models, i.e., the single-locus model and the environment-only models. Nevertheless, the orthogonal decomposition of variance is still valid here on the log-odds scale. We will therefore apply this model to the

analysis of case-control data. We will hereafter use a common terminology, statistical model, for both quantitative and qualitative trait, and evaluate its performance in simulation studies. We do not explicitly model the influence of the genotype frequencies on the variance of the regression parameters in logistic regression.

We extended the formulations for the statistical and functional models to the following three reduced genetic models: additive, dominant, and recessive.

Gene-environment interaction

Suppose we have a binary environmental exposure, M , with phenotypic values M_1 and M_2 for unexposed and exposed individuals, respectively. Denote the unexposed frequency by m . A functional model for this environmental exposure is

$$\vec{M} = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_M \end{pmatrix}, \quad (20)$$

with effects defined as

$$\vec{E}_M = \begin{pmatrix} R \\ a_M \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}. \quad (21)$$

For a two-level factor, following [5], the criterion for orthogonality can be derived as follows. From the regression model

$$\begin{pmatrix} M_1^* \\ M_2^* \\ \vdots \\ M_n^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \overline{Z}S\vec{E} = \chi\vec{E} \quad (22)$$

orthogonality requires that

$$\chi^T\chi = nS^T\overline{Z}^T\overline{Z}S = S^TDS \quad (23)$$

is diagonal, where

$$D = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix}, \quad (24)$$

and $m_1 = m$ and $m_2 = 1 - m$ are the exposure frequencies. Since

$$\chi^T\chi = n \begin{pmatrix} m_1s_{11}^2 + m_2s_{21}^2 & m_1s_{11}s_{12} + m_2s_{21}s_{22} \\ m_1s_{11}s_{12} + m_2s_{21}s_{22} & m_1s_{12}^2 + m_2s_{22}^2 \end{pmatrix} \quad (25)$$

it follows that the model, S , is orthogonal when

$$m_1s_{11}s_{12} + m_2s_{21}s_{22} = 0. \quad (26)$$

Using this criterion, we find that the functional model given above is not orthogonal.

The orthogonal (or, statistical) model for the binary environmental factor is

$$\vec{M} = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} 1 & m-1 \\ 1 & m \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_M \end{pmatrix}, \quad (27)$$

with effects defined as

$$\vec{E}_M = \begin{pmatrix} \mu \\ \alpha_M \end{pmatrix} = \begin{pmatrix} m & 1-m \\ -1 & 1 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}. \quad (28)$$

Using the Kronecker product rule [5], we have the following non-orthogonal functional model for the gene-environment interaction

$$\vec{G}_{GM} = \begin{pmatrix} G_{11M_1} \\ G_{12M_1} \\ G_{22M_1} \\ G_{11M_2} \\ G_{12M_2} \\ G_{22M_2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} R \\ a_G \\ d_G \\ a_M \\ aa \\ da \end{pmatrix}, \quad (29)$$

and the statistical model

$$\vec{G}_{GM} = \begin{pmatrix} 1 & -\bar{N} & -2p_{12}p_{22}/V & m-1 & -(m-1)\bar{N} & -2(m-1)p_{12}p_{22}/V \\ 1 & 1-\bar{N} & 4p_{11}p_{22}/V & m-1 & (m-1)(1-\bar{N}) & 4(m-1)p_{11}p_{22}/V \\ 1 & 2-\bar{N} & -2p_{11}p_{12}/V & m-1 & (m-1)(2-\bar{N}) & -2(m-1)p_{11}p_{12}/V \\ 1 & -\bar{N} & -2p_{12}p_{22}/V & m & -m\bar{N} & -2mp_{12}p_{22}/V \\ 1 & 1-\bar{N} & 4p_{11}p_{22}/V & m & m(1-\bar{N}) & 4mp_{11}p_{22}/V \\ 1 & 2-\bar{N} & -2p_{11}p_{12}/V & m & m(2-\bar{N}) & -2mp_{11}p_{12}/V \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_G \\ \delta_G \\ \alpha_M \\ \alpha\alpha \\ \delta\alpha \end{pmatrix}. \quad (30)$$

The relation between the statistical and functional models is

$$\begin{pmatrix} \mu \\ \alpha_G \\ \delta_G \\ \alpha_M \\ \alpha\alpha \\ \delta\alpha \end{pmatrix} = \begin{pmatrix} 1 & \bar{N} & p_{12} & 1-m & (1-m)\bar{N} & (1-m)p_{12} \\ 0 & 1 & p'_{12} & 0 & 1-m & (1-m)p'_{12} \\ 0 & 0 & 1 & 0 & 0 & 1-m \\ 0 & 0 & 0 & 1 & \bar{N} & p_{12} \\ 0 & 0 & 0 & 0 & 1 & p'_{12} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_G \\ d_G \\ a_M \\ aa \\ da \end{pmatrix}. \quad (31)$$

One of the desired properties of the statistical models is that the marginal effects of the gene or the environmental exposure are the same as the corresponding main effects of the full GxE model. This is the case for the linear regression of a quantitative trait. For a quantitative trait determined by a gene and a binary exposure, from the vector of genotypic values,

$\vec{G} = (G_{11M_1}, G_{12M_1}, G_{22M_1}, G_{11M_2}, G_{12M_2}, G_{22M_2})^T$, the marginal model for the gene is

$$G'_{11} = mG_{11M_1} + (1-m)G_{11M_2} \quad (32)$$

$$G'_{12} = mG_{12M_1} + (1-m)G_{12M_2} \quad (33)$$

$$G'_{22} = mG_{22M_1} + (1 - m)G_{22M_2}. \quad (34)$$

It can be shown that the corresponding marginal genetic effects is

$$\begin{pmatrix} \mu' \\ \alpha' \\ \delta' \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & p_{22} \\ p'_{11} & p'_{12} & p'_{22} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G'_{11} \\ G'_{12} \\ G'_{22} \end{pmatrix} = \begin{pmatrix} \mu_G \\ \alpha_G \\ \delta_G \end{pmatrix}. \quad (35)$$

A similar expression can be obtained for the marginal effects of the environmental exposure. However, this may not be the case for the statistical model under the alternate model of association defined on the log-odds scale for a qualitative trait. This is because the same relation for the genotypic values, equations (32),(33), and (34), is valid on the scale of penetrance, and the penetrances are nonlinearly related to the log-odds.

For the three reduced genetic models, the formulations for statistical and functional models for GxE and their relationships are given in Supplementary Note.

Simulation methods

Simulation of data with a quantitative trait—It is straightforward to simulate data of independent individuals randomly sampled with a quantitative trait influenced by a diallelic gene and a binary exposure. For given values of the variant allele frequency, p , and the exposure prevalence, m , assuming independency between the gene and the exposure and Hardy-Weinberg proportion for the gene, an individual was assigned joint genotype 111, 121, 221, 112, 122, or 222 with probabilities, $p^2(1 - m)$, $2p(1 - p)(1 - m)$, $(1 - p)^2(1 - m)$, p^2m , $2p(1 - p)m$ or $(1 - p)^2m$, respectively. Each individual was then assigned a value of the quantitative trait according to the genotypes using the genotypic values determined from a pre-specified vector of genetic effects: $\vec{E}^T = [R, a_G, d_G, a_M, aa, da]$. A residual was then added to the simulated quantitative trait by generating a random number from a normal distribution with pre-specified mean (0) and variance (σ^2). Data so simulated for 1000 individuals served as a replicate. For each genetic model, 1000 replicates were simulated.

Simulations of case-control data—We simulated case-control data with both main and interaction effects using the logistic models. If the risk of disease is determined by a diallelic gene and a binary exposure, we assume that the penetrance model is given by

$$\Pr(d=1|i) = \frac{1}{1 + \exp(-G_i)}, \quad (36)$$

where $d = 1$ denotes that fact that an individual is affected and G_i is the genotypic value when the joint genotype is i with $i = 111, 121, 221, 112, 122$, or 222 . Using Bayes' theorem, we have the distribution of the six genotypes in the cases as follows

$$\Pr(i|d=1) = \frac{P_i/[1 + \exp(-G_i)]}{\sum_j P_j/[1 + \exp(-G_j)]}, \quad (37)$$

where P_i is the frequency of genotype i in the population, given by $p^2(1 - m)$, $2p(1 - p)(1 - m)$, $(1 - p)^2(1 - m)$, p^2m , $2p(1 - p)m$ or $(1 - p)^2m$, respectively, as in the simulation of a quantitative trait. Given the genotypic values and the frequencies of the joint genotypes, this

expression was used for simulating joint genotypes of cases. For the simulation of controls, we have a similar expression:

$$\Pr(i|d=0) = \frac{P_i/[1+\exp(G_i)]}{\sum_j P_j/[1+\exp(G_j)]} \quad (38)$$

The genotypic values were determined from pre-specified genetic effects, \vec{E} . It should be noted that, unlike the simulated data for a quantitative trait, not only the allele frequencies, but also the genetic effects, in the simulated case-control data are usually different from the corresponding pre-specified values (population parameters) because of ascertainment bias.

Results

Results of simulations

Our first simulation exhibited both main effects of a gene and a binary exposure and their interaction on a quantitative trait, representing a general scenario:

$\vec{E}^T = [R, a_G, d_G, a_M, aa, da] = [100.0, 4.0, 1.0, 3.0, 2.0, 1.5]$. The residual variance was 144.0, and the allele frequency and exposure frequency were 0.15 and 0.22 respectively. The vector of genetic effects in the statistical model was then [102.33, 5.37, 1.33, 3.98, 3.05, 1.5]. Figure S1 shows the distribution of the estimates of all six effects in all the 1000 replicates. It can be seen that for both the statistical and the functional models, the vectors of genetic effects were estimated accurately. Figure 1 shows the power as a function of critical values of the Wald test P-values for the four parameters: the additive effects of the gene, the dominant effects of the gene, the environment effect, and the interaction effect between the additive effect of the gene and the environment effect for both statistical models and functional models. For the dominant by environment interaction effects, the estimates and the test P-values were identical for the statistical and functional models. Also shown is the power for the additive and dominant effects of the gene using the gene-only statistical models. It is clear that for the additive effect of the gene, both the GxE model and the gene-only model had much higher power than the functional models. For the environment effect, the statistical model had higher power than the functional model.

Another simulation was performed for a scenario where only the gene was responsible to the trait: $\vec{E}^T = [100.0, 4.0, 2.0, 0.0, 0.0, 0.0]$. The residual variance was 144.0 and the allele frequency and exposure frequency were 0.15 and 0.22 respectively. The vector of genetic effects in the statistical model was then [101.71, 5.40, 2.00, 0.00, 0.00, 0.00]. Again, the estimates of parameters were accurate, as shown in Figure S2, for both the statistical and functional models. For the additive effect of the gene and the additive by environment interaction effect, the estimate of the statistical model had less variations among the replicates than that of the functional model. Power for detecting the additive effect was again much higher in the statistical model than that in the functional model (Figure 2). Power for detecting the dominant effects using both statistical and functional models were very low because of the small simulating value compared to the residual variance (2 vs. 144). For the other two parameters, for which the simulating values were zero, the false positive rates were very close to the nominal value for both statistical and functional models.

Similar scenarios were simulated for a qualitative trait. For a generic scenario, we used a vector of genetic effects given by [-2.0, 0.3, 0.1, 0.2, 0.1, 0.04]. The allele frequency and exposure frequency were set to 0.25 and 0.25, respectively. Because we simulated case-control data set, the ascertainment of equal number of cases and controls altered the proportions of being affected for different genotypes and exposure status. The resulting

genetic effects, averaged over all 1000 replicates in the sample were thus estimated as $[-0.26, 0.28, 0.10, 0.18, 0.27, -0.15]$, referred to as the actual functional effects. The statistical effects were then calculated using the actual allele frequency and exposure frequency and the actual functional effects: $[0.00, 0.38, 0.06, 0.26, 0.20, -0.15]$. Figure S3 shows that this actual (or, sample) statistical effects were indeed located in the corresponding center of the distribution of estimated parameters using logistic regression. Also shown in Figure S3 is that the functional model correctly estimated the actual genetic effects. Figure 3 shows that power of detecting the main additive effect of the gene was much higher than that of the functional model. Power of detecting the exposure effect using the statistical model was also significantly higher than the functional model. For the dominant effect and the interaction effect, both models had the similarly low power.

Similar to the case of quantitative trait, we also simulated case-control data a scenario where only the gene was associated to the disease with simulating values of genetic effects given by $[-2.0, 0.4, 0.2, 0.0, 0.0, 0.0]$. The allele frequency and exposure frequency were still set to 0.25 and 0.25, respectively. Again because of ascertainment, the sample genetic effects were actually $[-0.27, 0.35, 0.23, -0.08, -0.02, 0.05]$ and the corresponding statistical effects were $[0.00, 0.45, 0.25, -0.07, 0.00, 0.05]$. Figure S4 shows the distributions of estimates of the parameters from analyses using the statistical and functional models.

In the four simulations given above, the difference of powers between the NOIA statistical models and the functional models were extremely large. For instance, in Figure 1, it can be seen that the power for testing $\alpha = 0$ was about 95% and the power for testing for $a = 0$ was only around 30% at the 1% significance level. The power of tests depends on many factors, including allele frequencies, sample sizes, genetic effects, and significance level. We plotted power for a wide range of significance level, which allows one to compare tests of powers of tests across a broad range. We use these simulations to illustrate how the transformation from the usual functional model to the statistical model may improve power of detecting a genetic factor.

For both quantitative and qualitative traits, we also simulated the null scenario, that is, a scenario without any effects either from the gene or the environmental factor (data not shown), and found that all the false positive rates were around the nominal level.

Application to real dataset

We applied the NOIA statistical model and the usual functional model to the ILCCO (International lung cancer consortium) data [7], consisting of 17 independent case-control studies (most but not all of the original studies agreed to participate in this study). The objectives of the consortium are to share data to increase statistical power, reduce duplication of research efforts, replicate novel findings, and realize substantial cost savings. Details of the participating studies have been described previously [7]. Our goal here was to examine how genetic variants, which have been identified through GWAS, may interact with smoking in determining the risk of lung cancer by pooling the datasets. Here, we focused on six SNPs in three regions: rs2736100 and rs402710 (5p15), rs2256543 and rs4324798 (6p21), and rs16969968 and rs8034191 (15q25). Our analysis included 95468 Caucasians with 39686 cases and 55752 controls after quality control. For both NOIA statistical model and the usual functional model, logistic regression was performed with sex, age and study group as covariates.

Table 1 and Table S1 show the P-values and estimates and their 95% confidence intervals of the analysis, respectively. For the four SNPs located on 5p15 and 15q25, the NOIA statistical model detected genetic additive effects, while the functional model did not or showed larger P-values. For the additive-smoking interaction effects, the NOIA statistical

model was more powerful than the functional one for the two SNPs on 15q25 and SNP rs2256543 on 6q21. For all SNPs and all models, the smoking effect was extremely significant. But, here the NOIA statistical models showed much smaller P-values than the functional models. For the covariates, sex and age, results from statistical and functional models were the same, as they were modeled in the same way. None of the SNPs had significant dominant effects, or dominant-smoking interaction, under either statistical model or functional model. It is interesting that the SNP rs2256543 on 6p21 was detected to predict lung cancer risk through an interaction with smoking, but no main effect.

Discussion

Multicollinearity occurs naturally in genetic regression analysis using functional models between the additive component and the dominance component and becomes even more complicated between the main effects and interaction effects when two or more genes or environmental factors are involved. When multicollinearity is present, the standard errors can become large and thus coefficients need to be very large in order to be statistically significant. In the NOIA framework, we solve the collinearity problem by orthogonalizing the dominant regressor with respect to the additive regressor, in order to keep the *natural* meaning of the coefficient of the additive regressor, i.e. the effect of allele substitutions. As a result, the NOIA statistical and functional models have identical additive regressor (i.e. the number of variant allele) and dominance coefficients, but different additive coefficient and dominant regressor terms. This strategy is exactly the same as the Gram-Schmidt process in mathematics for orthonormalising a set of vectors. This orthogonalizing procedure assigns all the shared variance of the additive and dominant components in the functional model to the additive component in the statistical model, thus usually making the power for detecting the additive effects higher. Our simulations and real-data analysis confirmed this anticipation. We found that the statistical model usually showed higher power in detecting main and/or interaction effects for both linear regression for quantitative traits and logistic regression for binary traits.

However, caution has to be exercised in interpreting the results of the statistical model. Specifically, the meaning of additive effect (α) in the statistical model is different from that in the functional model (a). The statistical effect, α , is determined not only by the true additive effect, a , but also by the dominance effect, d , and allele frequency. Nevertheless, both tests for α and for a give information on whether there exists a genetic factor for a quantitative trait or the risk of a disease. Our results shown in the figures indicated that transformation from the parameters used in the usual functional model to those in the statistical model leads to a more powerful test for the existence of a genetic factor while allowing for a dominant effect and a GxE interaction.

Some of the important properties of the NOIA framework for linear regression of quantitative traits are not always valid for logistic regression of qualitative traits, when we generalize the statistical model to the later case by treating the logit of the disease as genotypic values and genetic effects. Under the alternate model, when there is an association between the genotypes or environmental factors, the estimates of logistic regressing parameters are no longer uncorrelated. Also, under the alternate model, the main effects of a full interaction model are not the same as the corresponding main effects of the reduced single-gene model or the environment-only model. Nevertheless, we still advocate the application of the statistical model in analyzing case-control data, because it is more powerful in most of the cases.

Application of the NOIA statistical model to the ILCCO data confirmed the associations of the following loci with lung cancer through main effects: rs2736100, rs402710, rs16969968,

and rs8034191. The main effects of these loci under the usual functional model were not significant (or had a larger P value) while allowing for gene-smoking interaction. Furthermore, the gene-smoking interaction was more significant under the statistical model for loci rs2256543, rs16969968, and rs8034191. Specifically, the statistical model revealed that the locus rs2256543 plays a role in the development of lung cancer through interaction with smoking, but not with a main effect.

Finally, the advantage of statistical model over the usual functional model is not limited to the study of interaction effects. We propose that even for one-locus genetic analysis, such as GWAS, one should consider applying the statistical model, since it orthogonalizes the additive and dominant effects and hence improves power of detecting genetic effects. Although, the genetic effects in the statistical model usually are determined not only by the biological mechanisms but also the population properties, proper explanations of the genetic effects can be achieved through transformations established in the NOIA model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health grants R01CA134682, CA121197, CA134682, CPRIT RP100443, U19 CA148127 and R01AR057120. We are grateful to Simone Benhamou, Paolo Boffetta, Paul Brennan, Neil Caporaso, Nilanjan Chatterjee, M. Dawn Teare, Yun-Chul Hong, Bart Kiemeneij, Loic Le Marchand, Thorunn Rafnar, Gadi Rennert, John R. McLaughlin, Adeline Seow, Margaret Spitz, Maria Tere Landi, Paolo Vineis, John Wiencke, Juergen Wolf, and Zuo-Feng Zhang for providing us with their data.

References

1. Donnelly P. Progress and challenges in genome-wide association studies in humans. *Nature*. 2008; 456:728–731. [PubMed: 19079049]
2. Amos C, Wu X, Broderick P, Gorlov I, Gu J, et al. Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics*. 2008; 40:616–622. [PubMed: 18385676]
3. Maher B. Personal genomes: The case of the missing heritability. *Nature*. 2008; 456:18–21. [PubMed: 18987709]
4. Manolio T, Collins F, Cox N, Goldstein D, Hindorf L, et al. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–753. [PubMed: 19812666]
5. Álvarez Castro J, Carlborg O. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*. 2007; 176:1151–1167. [PubMed: 17409082]
6. Álvarez Castro J, Le Rouzic A, Carlborg O. How to perform meaningful estimates of genetic effects. *PLoS Genet*. 2008; 4:e1000062. [PubMed: 18451979]
7. Truong T, Hung R, Amos C, Wu X, Bickeböller H, et al. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: A pooled analysis from the international lung cancer consortium. *J Natl Cancer Inst*. 2010; 102:959–971. [PubMed: 20548021]

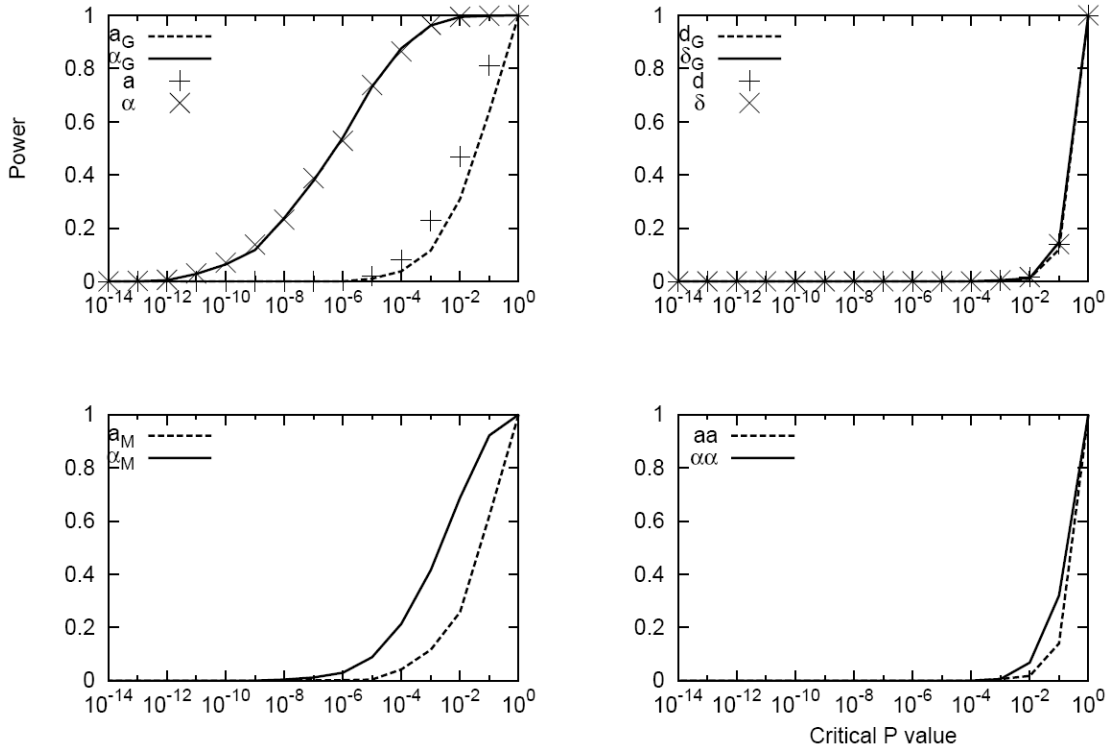


Fig. 1. Power vs. critical values of the Wald testing P-values as a test statistic for a simulated data set with a quantitative trait influenced by a gene and an environmental factor. The simulating values of the genetic effects were $\vec{E}^T = [100.0, 4.0, 1.0, 3.0, 2.0, 1.5]$. The corresponding statistical genetic effects were [102.33, 5.37, 1.33, 3.98, 3.05, 1.5]. The allele frequency and exposure frequency were 0.15 and 0.22, respectively. The simulating residual variance is 144.0.

\$watermark-text

\$watermark-text

\$watermark-text

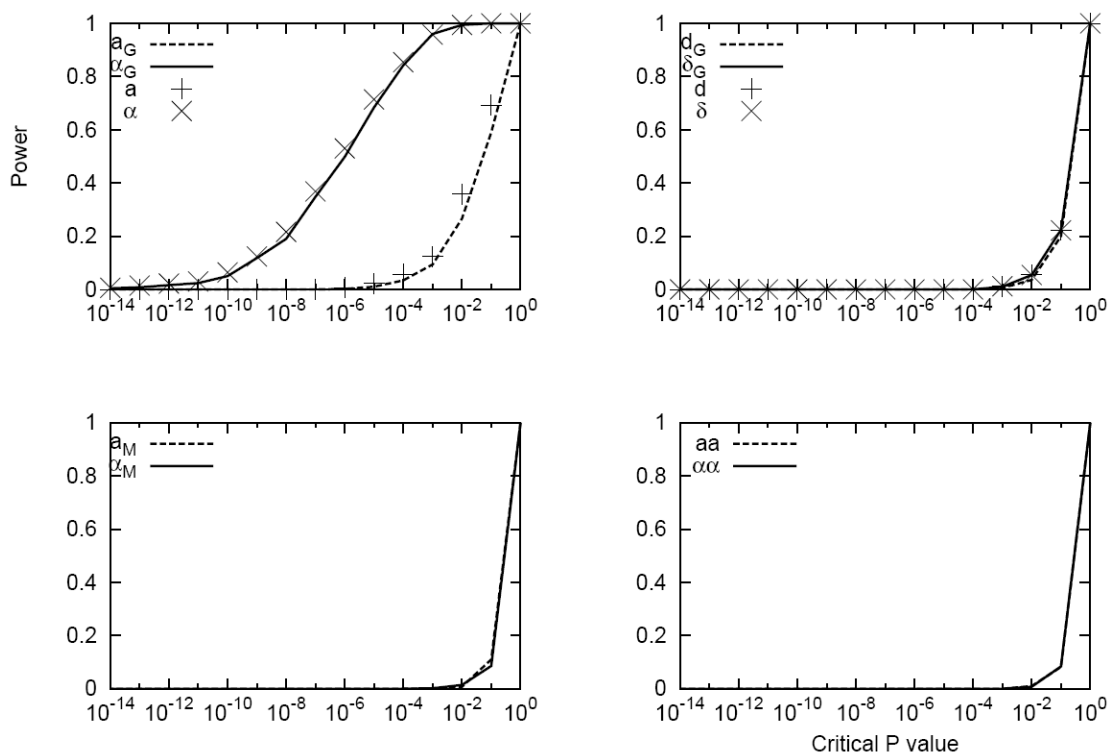


Fig. 2. Power vs. critical values of the Wald testing P-values as a test statistic for a simulated data set with a quantitative trait influenced by a gene and an environmental factor. The simulating values of the genetic effects were $\vec{E}^T = [100.0, 4.0, 2.0, 0.0, 0.0, 0.0]$. The corresponding statistical genetic effects were $[101.71, 5.40, 2.00, 0.00, 0.00, 0.00]$. The allele frequency and exposure frequency were 0.15 and 0.22, respectively. The simulating residual variance is 144.0.

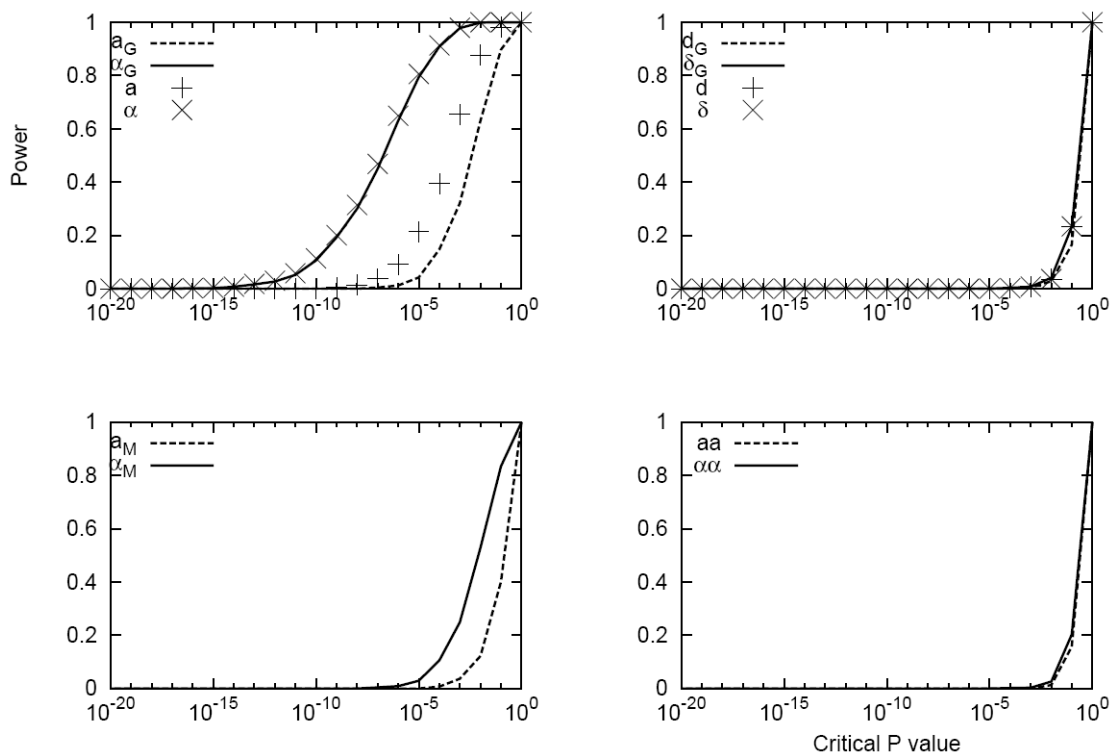


Fig. 3. Power vs. critical values of the Wald testing P-values as a test statistic for a simulated case-control data set with GxE interaction for all replicates. The simulating values of the genetic effects were $\vec{E}^T = [-2.0, 0.3, 0.1, 0.2, 0.1, 0.04]$. The actual sample values of the genetic effects were $[-0.28, 0.30, 0.10, 0.20, 0.11, 0.03]$. The corresponding statistical genetic effects were $[-0.01, 0.38, 0.10, 0.27, 0.13, 0.03]$. The allele frequency and exposure frequency were 0.25 and 0.25 respectively.

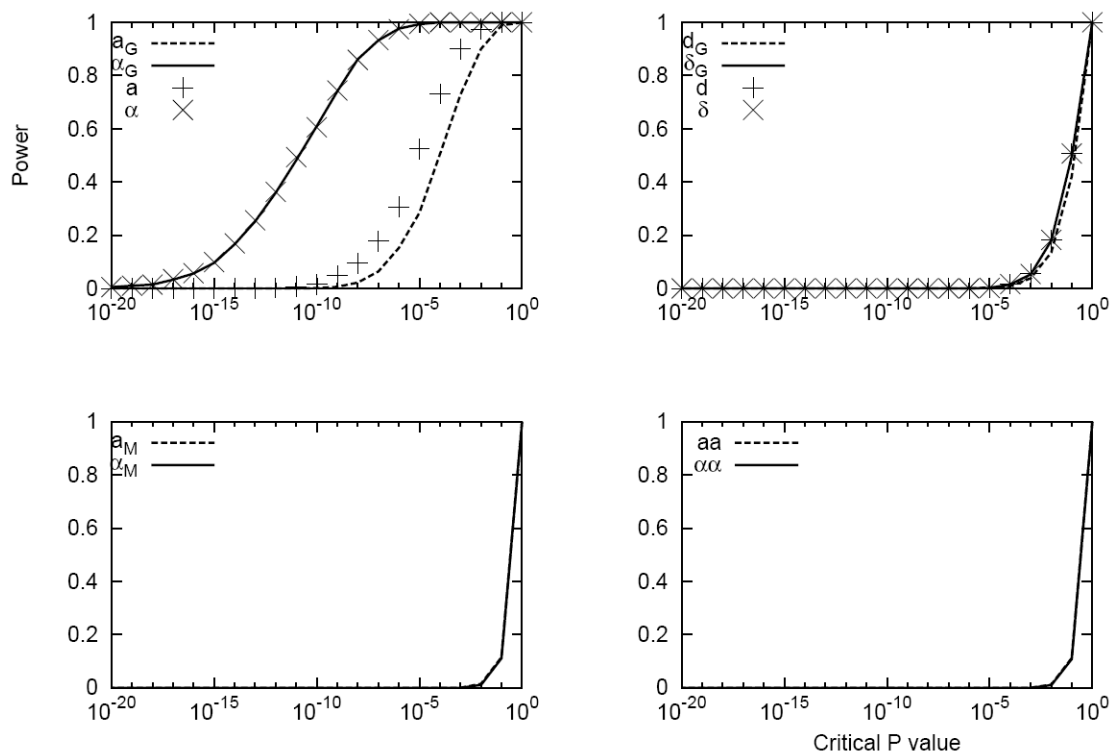


Fig. 4. Power vs. critical values of the Wald testing P-values as a test statistic for a simulated case-control data set with GxE interaction for all 1000 replicates. The simulating values of the genetic effects were $\vec{E}^T = [-2.0, 4.0, 0.2, 0.0, 0.0, 0.0]$. The actual sample values were $[-0.31, 0.40, 0.19, 0.00, 0.01, -0.01]$. The corresponding statistical genetic effect were $[0.00, 0.49, 0.19, 0.00, 0.00, -0.01]$. The allele frequency and exposure frequency were 0.25 and 0.25, respectively.

Table 1

P-values for the main effects and interactions^a

	Model	Add	Dom	SM	Add-SM	Dom-SM	Sex	Age
rs2736100	Func	0.0005	0.7	1.6e-72	0.3	0.7	9.1e-09	1.7e-16
5p15	Stat	1.9e-10	1	3.5e-267	0.3	0.7	9.1e-09	1.7e-16
rs402710	Func	0.09	1	3.2e-125	0.8	0.9	7.0e-09	3.1e-37
5p15	Stat	2.3e-08	0.8	4.4e-244	0.7	0.9	7.0e-09	3.1e-37
rs2256543	Func	0.2	0.6	1.3e-84	0.06	0.4	6.6e-09	4.9e-16
6p21	Stat	0.5	0.7	1.2e-271	0.04	0.4	6.6e-09	4.9e-16
rs4324798	Func	0.8	0.6	5.6e-227	0.8	0.8	2.9e-08	5.9e-29
6p21	Stat	0.4	0.4	2.9e-270	0.3	0.8	2.9e-08	5.9e-29
rs16969968	Func	1	0.9	6.0e-96	0.0007	0.6	4.1e-08	3.5e-28
15q25	Stat	7.4e-16	0.1	6.9e-266	9.5e-05	0.6	4.1e-08	3.5e-28
rs8034191	Func	0.8	0.9	1.1e-45	0.02	0.6	2.4e-09	2.2e-33
15q25	Stat	2.2e-15	0.3	2.1e-136	0.006	0.6	2.4e-09	2.2e-33

^aStudy group has been used as covariate.