5-2013

THE NATURAL AND ORTHOGONAL INTERACTION (NOIA) MODELS FOR QUANTITATIVE TRAITS (QTs) AND COMPLEX DISEASES

Feifei Xiao

THE NATURAL AND ORTHOGONAL INTERACTION (NOIA) MODELS FOR QUANTITATIVE TRAITS (QTs) AND COMPLEX DISEASES

by
Feifei Xiao, M.S.

APPROVED:

_____
Supervisory Professor: Christopher I. Amos, Ph.D.

_____
Ralf Krahe (On-site advisor), Ph.D.

_____
Jianzhong Ma, Ph.D.

_____
Yunxin Fu, Ph.D.

_____
Paul Scheet, Ph.D.

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

THE NATURAL AND ORTHOGONAL INTERACTION (NOIA) MODELS FOR QUANTITATIVE TRAITS (QTs) AND COMPLEX DISEASES

A

DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

MD Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Feifei Xiao, M.S.

Houston, Texas

May, 2013

To my dear mother who has suffered from cancer

ACKNOWLEDGEMENTS

I would never have been able to complete my dissertation without the guidance of my advisor, my academic committee members, and the emotional and moral support from my family and dear friends.

First, I would like to express my deepest gratitude to my advisor, Dr. Christopher I. Amos, for his excellent guidance, caring, patience and continuous support throughout my research process. He provided an excellent academic atmosphere for me to focus on my research. I am thankful for his instructions and encouragements that helped me shape my theoretical knowledge, computational and technical skills for research. I owe my thanks to my supervisory committee member, Dr. Jianzhong (David) Ma, who was always willing to help and give his best suggestions in both research and life. He guided my research for the past several years and helped me to develop my background in genetics, biostatistics and theoretical thinking in my research work. I owe many thanks to my on-site advisor, Dr. Ralf Krahe, for his continuous support on my study, my research work and suggestions on my presentation skills. I am also grateful to my supervisory committee members, Dr. Yunxin Fu and Dr. Paul Scheet, for their invaluable suggestions on my projects and revision of my dissertation. I would like to thank my advisory and candidacy committee members: Dr. Marsha L. Frazier, Dr. Shoudan Liang and Dr. Subrata Sen, for taking time to help me on my research and candidacy exam. Many thanks go to Dr. Bo Peng, Dr. Shenying Fang, Dr. Yafang Li, Emily Lu, Wei Chen, Dakai Zhu, and Yaji Xu in the Department of Genetics for their great help and advice during my Ph. D. study. My thanks also go to Anthony San. Lucas for his great help to improve the writing of this dissertation. I also want to thank Changlu Liu and Issac Wun, as friends and colleagues, for their great advice and support in both my life and study. I would also like to thank Dr. Victoria Knutson,

Ms. Lourdes Perez and other staff members from GSBS for their great help and support in the past several years.

I would also like to thank my dear parents, my elder brother for their endless love and encouragements which carried me through all difficult times. My research would never have been possible to accomplish without their support. Many thanks go to my beloved husband, Guoshuai Cai, who was always standing by me and encouraging me to believe in myself through both good and bad times. Finally, I would like to thank all my friends for their continuous support and friendship.

ABSTRACT

THE NATURAL AND ORTHOGONAL INTERACTION (NOIA) MODELS FOR QUANTITATIVE TRAITS (QTs) AND COMPLEX DISEASES

Publication No._____

Feifei Xiao, M.S.

Supervisory Professor: Christopher I. Amos, Ph.D.

My dissertation focuses on developing methods for gene-gene/environment interactions and imprinting effect detections for human complex diseases and quantitative traits. It includes three sections: (1) generalizing the Natural and Orthogonal interaction (NOIA) model for the coding technique originally developed for gene-gene (GxG) interaction and also to reduced models; (2) developing a novel statistical approach that allows for modeling gene-environment (GxE) interactions influencing disease risk, and (3) developing a statistical approach for modeling genetic variants displaying parent-of-origin effects (POEs), such as imprinting.

In the past decade, genetic researchers have identified a large number of causal variants for human genetic diseases and traits by single-locus analysis, and interaction has now become a hot topic in the effort to search for the complex network between multiple genes or environmental exposures contributing to the outcome. Epistasis, also known as gene-gene interaction is the departure from additive genetic effects from several genes to a trait, which means that the same alleles of one gene could display different genetic effects under different genetic backgrounds. In this study, we propose to implement the NOIA model for association studies along with interaction for human complex traits and diseases. We compare the performance of the new statistical models we developed and the usual functional model by both simulation study and real data analysis. Both simulation and real data analysis revealed higher power of the NOIA GxG interaction model for

detecting both main genetic effects and interaction effects. Through application on a melanoma dataset, we confirmed the previously identified significant regions for melanoma risk at 15q13.1, 16q24.3 and 9p21.3. We also identified potential interactions with these significant regions that contribute to melanoma risk.

Based on the NOIA model, we developed a novel statistical approach that allows us to model effects from a genetic factor and binary environmental exposure that are jointly influencing disease risk. Both simulation and real data analyses revealed higher power of the NOIA model for detecting both main genetic effects and interaction effects for both quantitative and binary traits. We also found that estimates of the parameters from logistic regression for binary traits are no longer statistically uncorrelated under the alternative model when there is an association. Applying our novel approach to a lung cancer dataset, we confirmed four SNPs in 5p15 and 15q25 region to be significantly associated with lung cancer risk in Caucasians population: rs2736100, rs402710, rs16969968 and rs8034191. We also validated that rs16969968 and rs8034191 in 15q25 region are significantly interacting with smoking in Caucasian population. Our approach identified the potential interactions of SNP rs2256543 in 6p21 with smoking on contributing to lung cancer risk.

Genetic imprinting is the most well-known cause for parent-of-origin effect (POE) whereby a gene is differentially expressed depending on the parental origin of the same alleles. Genetic imprinting affects several human disorders, including diabetes, breast cancer, alcoholism, and obesity. This phenomenon has been shown to be important for normal embryonic development in mammals. Traditional association approaches ignore this important genetic phenomenon. In this study, we propose a NOIA framework for a single locus association study that estimates both main allelic effects and POEs. We develop statistical (Stat-POE) and functional (Func-POE) models, and demonstrate conditions for orthogonality of the Stat-POE model. We conducted simulations for both quantitative and qualitative traits to evaluate the performance of the statistical and functional models with different levels of POEs. Our results showed that the newly proposed Stat-POE model, which ensures orthogonality of variance components if Hardy-Weinberg Equilibrium (HWE) or equal

minor and major allele frequencies is satisfied, had greater power for detecting the main allelic additive effect than a Func-POE model, which codes according to allelic substitutions, for both quantitative and qualitative traits. The power for detecting the POE was the same for the Stat-POE and Func-POE models under HWE for quantitative traits.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Genetic Association Analyses for Finding Causal Variants for Human Complex Traits and Diseases

For the past several years, searching for genetic factors that cause various human complex traits and diseases has become one of the most important and challenging goals for modern geneticists. Genome-wide association studies (GWASs) have contributed substantively to this goal [1-3]. In this approach, every locus is isolated and analyzed. Several hundred thousand single nucleotide polymorphisms (SNPs) in thousands of individuals are assayed, which has provided a powerful approach for investigating the underlying genetic architecture of human complex traits and diseases [4, 5]. GWASs have identified a large number of causal variants for human genetic diseases and traits, such as cancer, diabetes and heart diseases [6-8], and have provided valuable insights into the complexities of the human diseases. For example, about 90 loci have been identified for association with the common human trait, height, and have explained about 56% of the overall phenotypic variance [9]. GWAS also identified common variants which account for 32% of narrow-sense heritability of body mass index [10].

The goal of GWASs is to identify common variants for common diseases, but explaining a large and missing proportion of the heritability of most complex or multifactorial diseases and disorders is still a challenging task in the field of genetic epidemiology. A limitation of this approach that has been cited is that interactions between loci or between genes and environmental exposures are usually ignored [11, 12]. For this reason, more efforts are being made to characterize the complex network between multiple genes and environmental factors that contribute to disease outcome. Potential gene-gene or gene-environmental interactions have been indicated in recent years, but few

of them have been validated. The underlying biological pathways could be successfully elucidated as more and more interactions are uncovered. Aside from interactions, structural variation, such as copy number variations (CNVs), may account for some of the missing heritability if those variants contribute to the genetic basis of the human disease [13, 14]. Imprinting effects and rare variants may also account for part of the missing heritability too [15, 16]. Rare variants (minor allele frequency<0.5%) are not well captured by the GWA genotype arrays because of their small minor allele frequency, although they may have substantial effect sizes and contribute in aggregating to the burden of disease from genetic factors [17-19].

1.2 Usual Functional Models for Genotype-Phenotype Mapping

We first briefly review the usual functional models for genotype-phenotype mapping. In the usual approach for genotype-phenotype mapping of a quantitative trait locus (QTLs), if the trait is influenced by a single diallelic locus, with alleles A_1 and A_2, we let minor allele be A_2. Assume we have a sample with n individuals. For the i-th individual, let y_i be the observed trait phenotype and G_i^* be the genotypic value for specific locus. We use y to denote the vector of the observed trait which is normally distributed and $y = [y_1, y_2, \dots, y_n]^T$. We model the phenotype as $y_i = G_i^* + \varepsilon_i$. The vector $G^* = Z \cdot G$, where G denote the vector of genotypic values including G_{11}, G_{12} and G_{22} as the genotypic values for the three possible genotypes for alleles A_1 and A_2; the n rows of matrix Z represent the corresponding genotype. Therefore, the vector of the observed phenotypes G^* could be expressed as

$$\begin{pmatrix} G_1^* \\ G_2^* \\ \vdots \\ \vdots \\ G_n^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{pmatrix} \cdot \begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix}. \quad (1)$$

Several methods have been proposed for mapping a quantitative trait controlled by one locus with two alleles [20]. The vector of genotypic values G can be modeled as the product of genetic-effect design matrix S and the vector of genetic effect E.

$$G = S \cdot E. \tag{2}$$

Let X be the design matrix for the whole sample, $X = Z \cdot S$. Therefore, we obtain the regression of genetic effects as the form $y = X \cdot E + \varepsilon = Z \cdot S \cdot E + \varepsilon = Z \cdot G + \varepsilon$, where ε is the error term. Different mapping methods focus on the core design matrix, S. One of the usual regression models, which is referred to as a functional model, can be described as follows [20]:

$$G = \begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = S_F E_F = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} R \\ a \\ d \end{pmatrix}. \tag{3}$$

For an individual with genotype G_{11}, the coding will be the first row of the design matrix S_F, and for an individual with genotype G_{12}, the coding will be the second row of the design matrix S.

The inverse of equation (3) is

$$E_F = \begin{pmatrix} R \\ a \\ d \end{pmatrix} = {S_F}^{-1} G = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix}. \tag{4}$$

Here, the reference point R corresponds to the genotypic value of one of the two homozygotes, G_{11}. The additive effect, a, is half of the difference between the two homozygotes genotypic values. The dominance effect, d, is the difference of the heterozygote genotypic value and the average of the homozygotes genotypic values. Estimation of the genetic effects, a and d, could be performed by linear regression for quantitative trait or logistic regression for qualitative traits. The coding in equation (4) is referred to as Func-Usual modeling in our study. Another usual functional model codes the additive effect as (-1,0,1) for the three genotypes and the reference point corresponds to the average genotypic values of the two homozygotes [20]. These two usual functional models have the same estimators except the intercept term, and we therefore will not consider the second model in

what follows. These models are called functional models since they use natural effects of allele substitutions as parameters, mainly focusing on the biological properties [21].

1.3 The Natural and Orthogonal Interactions (NOIA) model and Its Advantage for Association Studies

A second approach to modeling, the "statistical model", referred as the NOIA statistical model, was proposed by Alvarez-Castro and Carlborg et al. for estimating genetic effects for a quantitative trait and gene-gene (GxG) interactions [21]. As shown in Ma et al. [22], G could be expressed as, in the NOIA model,

$$G = \begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix} = S_S E_S = \begin{pmatrix} 1 & -\bar{N} & -2p_{12}p_{22}/V \\ 1 & 1-\bar{N} & 4p_{11}p_{22}/V \\ 1 & 2-\bar{N} & -2p_{11}p_{12}/V \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix}, \quad (5)$$

which ensures orthogonality of the estimated parameters. Here, p_{ij} denotes the genotype frequencies of this locus in the population, where ij $= 11,12$ or 22. $\bar{N} = p_{12} + 2p_{22}$, $V = p_{12} + 4p_{22} - (p_{12} + 2p_{22})^2 = p_{11} + p_{22} - (p_{11} - p_{22})^2$. $\bar{N}$ is the expected value of N and V is the variance of N. N is the number of variant alleles (A_2, for example) which is equal to 0, 1 or 2 when the genotype is G_{11}, G_{12}, or G_{22}, respectively.

The inverse of equation (5) is

$$E_S = \begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix} = S_S^{-1} G = \begin{pmatrix} p_{11} & p_{12} & p_{22} \\ p'_{11} & p'_{12} & p'_{22} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{22} \end{pmatrix}, \quad (6)$$

with

$$p'_{ij} = p_{ij} \frac{N_{ij} - \bar{N}}{V}. \quad (7)$$

The genetic effects, E_S, are based on the genotype frequencies of this locus in the population. Alvarez-Castro et al. [23] noted that the statistical model is an orthogonal model that has uncorrelated estimates of the parameters, which was also reflected by variance components

decomposition [22]. The statistical model (Equation (5)) and the functional model (Equation (3)) can be transformed to each other using:

$$\begin{pmatrix} R \\ a \\ d \end{pmatrix} = \begin{pmatrix} 1 & \bar{N} & p_{12} \\ 0 & 1 & p'_{12} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \delta \end{pmatrix}. \tag{8}$$

We notice that these two models have the same estimators for the dominant effect and different estimators for the additive effect.

As pointed out by Alvarez-Castro and Carlborg, there are two main advantages for the orthogonal models [21]. First, it makes model selection straightforward as the estimates are consistent in reduced models. Second, it enables accurate variance component analysis because of independent estimation of the genetic effects. A model with design matrix X satisfying $X^T \cdot X$ being a diagonal matrix will be an orthogonal model [21]. That is

$$X^T \cdot X = \left(S^T \cdot Z^T\right) \cdot Z \cdot S = nS^T \cdot Q \cdot S,$$

where

$$Q = \begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 0 & p_{22} \end{pmatrix}.$$

And given that $S = \left(s_{ij}\right)$ with $s_{i1} = 1$, the criteria for orthogonality of the genetic regression model was derived by Cockerham and denoted in terms of our notation as following [21]. To attain orthogonality, one sets the off diagonal elements of the $X^T \cdot X$ matrix to be zero since it will then follow that $S^T \cdot Q \cdot S$ is a diagonal matrix.

$$s_{11}p_{11} + s_{22}p_{12} + s_{32}p_{22} = 0,$$

$$s_{13}p_{11} + s_{23}p_{12} + s_{33}p_{22} = 0,$$

$$s_{12}s_{13}p_{11} + s_{22}s_{23}p_{12} + s_{32}s_{33}p_{22} = 0.$$

The statistical model fulfills these criteria and shows orthogonality for detecting and estimating genetic effects, whereas some parameters of the functional model (Equation (3)) are confunded, which can cause issues in hypothesis testing when Wald-type tests are used (as we shall see later in

the dissertation). The statistical model uses average effects of allele substitutions in populations as parameters for the decompostion of genetic variance. Its statistical formulation provides an approach in which the estimates of the genetic effects remain orthogonal; that is, they are consistent in reduced and unconfouned in the full model. This holds true even if Hardy-Weinberg Equilibrium (HWE) is violated. The orthogonality of the NOIA model is attractive becuase it ensures that the estimated genetic effects are not statistically correlated, rendering a more meaningful calculation of heritability of a trait comparing to the traditional models. The orthogonality of the statistical formulation of NOIA framework become important when multiple loci are contributing to the outcome. This is also why we were motivated to do the following studies.

1.4 Testing Statistics of the Usual Functional Model and NOIA Statistical Model

To further understand the statistical characteristic of the usual functional model and the NOIA statistical model on testing the additive effect with or without dominant effect detection, we constructed the Wald test statistics for these two models before and after the dominance component is removed (details see Appendix 2.2). The Wald test statistic is $z = \frac{\hat{\beta}}{se(\hat{\beta})} = \frac{(X'X)^{-1}X'y}{se(\hat{\beta})}$ where $\hat{\beta}$ denotes the vector of the estimation of the genetic effects and $var(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

We constructed the test statistic of the functional model with both additive and dominance effect detection as following. $X'X = n\begin{pmatrix} 1 & \bar{N} & p_{12} \\ \bar{N} & p_{12} + 4p_{22} & p_{12} \\ p_{12} & p_{12} & p_{12} \end{pmatrix}$ which is not a diagonal matrix. The test statistic for the functional model is

$$\frac{1}{\sqrt{n\sigma^2 V}}\begin{pmatrix} \sqrt{\frac{V}{p_{11}}} & 0 & 0 \\ -\sqrt{\frac{p_{22}V}{p_{11}(1-p_{12})}} & 0 & \sqrt{\frac{p_{11}V}{p_{22}(1-p_{12})}} \\ -\sqrt{\frac{p_{12}p_{22}}{p_{11}}} & 2\sqrt{\frac{p_{11}p_{22}}{p_{12}}} & -\sqrt{\frac{p_{11}p_{12}}{p_{22}}} \end{pmatrix} Z'y,$$

with the linear combination of the second or third row and $Z'y$ for the additive effect testing or dominant effect testing, respectively. We also constructed the test statistic of the functional model with only additive effect detection as follows. $X'X = n\begin{pmatrix} 1 & \bar{N} \\ \bar{N} & \bar{N} + 2p_{22} \end{pmatrix}$. The test statistic for the additive functional model is

$$\frac{1}{\sqrt{n\sigma^2 V}}\begin{pmatrix} \sqrt{\bar{N} + 2p_{22}} & \frac{2p_{22}}{\sqrt{\bar{N}+2p_{22}}} & \frac{-p_{12}}{\sqrt{\bar{N}+2p_{22}}} \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \end{pmatrix} Z'y,$$

with the linear combination of the second row and $Z'y$ for the additive effect testing.

In what follows, we show the Wald test statistic of the NOIA model with both additive and dominance effect detection. $X'X = n\begin{pmatrix} 1 & 0 & 0 \\ 0 & V & 0 \\ 0 & 0 & \frac{4p_{11}p_{12}p_{22}}{V} \end{pmatrix}$ which fulfills the requirement of the orthogonality that we discussed in Section 1.3. The test statistic for the NOIA statistical model is

$$\frac{1}{\sqrt{n\sigma^2 V}} * \begin{pmatrix} \sqrt{V} & \sqrt{V} & \sqrt{V} \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \\ -\sqrt{\frac{p_{12}p_{22}}{p_{11}}} & 2\sqrt{\frac{p_{11}p_{22}}{p_{12}}} & -\sqrt{\frac{p_{11}p_{12}}{p_{22}}} \end{pmatrix} Z'y,$$

with the linear combination of the second or third row and $Z'y$ for the additive effect testing or dominant effect testing, respectively. After we remove the dominance component from the NOIA model, $X'X = n\begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix}$. Moreover, the test statistic for the additive NOIA statistical model after the dominance component is removed:

$$\frac{1}{\sqrt{n\sigma^2 V}}\begin{pmatrix} \sqrt{V} & \sqrt{V} & \sqrt{V} \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \end{pmatrix} * Z'y,$$

with the linear combination of the second row and $Z'y$ for the additive effect testing.

From above formulations, we can clearly state that the NOIA model has same test statistic with the usual functional model for additive effect detection when only additive effect testing is included. The NOIA statistical model also has same test statistic with the usual functional model for dominance effect detection. Obviously, the NOIA statistical model has consistent testing for additive

effect detection after the dominance component is added to the modeling, whereas the usual functional model loses power.

1.5 Gene-Gene Interactions and Gene-Environment Interactions Contributing to Human Complex Traits and Diseases

Unlike Mendelian diseases or traits in which single variants influence the outcome, multiple factors including genetic and environmental factors contribute to the complex diseases/traits. As stated in section 1.1, the interactions among different loci and environmental exposure are usually ignored in the usual GWAS. Accurate modeling of associations along with interactions remains a challeging task for geneticists. The term gene-gene (GxG) interaction, also called epistasis, has various definitions. The most common statistical definition of epistasis is a departure from additivity of genetic effects at each locus from two or more genes that influence a trait; thus, the same alleles of one gene could display different genetic effects in different genetic backgrounds. Epistasis has become a hot topic for genetic researchers in recent years. It was initially characterized in animal model in the early 1900's as playing an important role in determining some phenotypes. For common human diseases and disorders, such as anemia, cystic fibrosis and complex autoimmune diseases, the relevance of gene-gene interactions is still under exploration but became a more prominent explanation for the failure of GWAS to explain much of the variation in risk among individuals in the last decade [24-26]. Moreover, epistasis was recently revealed to be the main force in long-term molecular evolution [27]. To test for statistical interactions influencing quantitative traits, linear regression may be used including both main genetic effects and interaction effects. For binary outcomes, the usual approach for modeling uses a log odds scale that is fitted with logistic regression. Several methods have been developed for searching for the interactions when performing genetic association studies [28-32]. The major motivation of developing these approaches is to improve the power of detecting effects and to provide a more comprehensive assessment of genetic architecture influencing a trait [33]. The contribution of enviromental factors in determining human

complex diseases has been the provenance of epidemiologists. The role of gene-environment interactions in disease etioogy has engaged both geneticists and epidemiologists and there was a resurrection of interest in this area starting about a decade ago as geneticists tools became easier to use for large scale studies [34]. The interactions between the enviromental exposure and the genetic factor, which is called gene-environment (GxE) interactions, are believed to be able to play an important role in the genetic archtecture of most human complex traits and diseases. The definition of GxE interactions is similar to the GxG interactions. The same alleles of one gene could display different genetic effects in different environmental backgrounds. For example, the interaction between genetic factors and cigarete smoking exposure contributing to the lung cancer is among the most well-know examples of GxE interactions [35]. Individuals with variants of a specific gene may be more susceptible to lung cancer risk in smokers; individuals with the same variants may not be inclined to increased risk of lung cancer in non-smokers. Therefore, understanding the underlying mechanisms may give valuable insights on cancer prevention and possibly treatment. GxE interactions have been recently revealed to be play crucial roles on development of Parkinson's diseases, rheumatoid arthritis and lung cancer [36-38]. In recent years, for understanding the complexity of genotype-phenotype relationships along with the gene-environment interactions, several approaches and software have been developed [39-41]. Unwinding this complexity will help in explaining more of the heritability of human complex traits and diseases.

1.6 Imprinting Effect is Usually Ignored in Traditional Association Studies

Genetic imprinting frequently affects genes during embryogenesis and is the most well-known parent-of-origin effect (POE). Imprinting causes the differential expression of genes based on the parental origin of the chromosome [42].The same alleles transmitted from the father have different levels of transcription and thus may render a different effect on the phenotype compared with the alleles transmitted from the mother. Genetic imprinting has been shown to be important for normal embryonic development in mammals [43]. So far, approximately 200 imprinted genes have been

validated or predicted in humans (http://www.geneimprint.com). Imprinted genes have been implicated in several complex human disorders, including diabetes, breast cancer, alcoholism, and obesity [44-47]. Kong et al. identified several variants of known imprinted genes showing significant effects on the development of breast cancer, carcinoma and type II diabetes [48]. An allele on an imprinted region of chromosome 14q32 was recently identified to affect type I diabetes susceptibility by Wallace et al. [49].

Several statistical approaches have been developed for modeling POEs and imprinting effects. Shete et al. implemented a variance-components (VC) method for testing genetic linkage by incorporating an imprinting parameter [50]. They applied this framework to rheumatoid arthritis and gene expression data and found significant signals for linkage [51]. Gorlova et al. developed a QTL analysis test to evaluate both total and parent-specific linkage signals based on identity-by-descent sharing [16] . Ainsworth et al. also described an implementation of a family-based multinomial modeling methodology in which POE detection is considered using mothers and their offspring [52]. However, none of above approaches considered the orthogonality properties in the modeling of the genetic effects.

Most traditional association approaches assume that the two alleles from the parents contribute equally to the trait, thereby ignoring the important genetic phenomenon, POEs. These approaches estimate the main allelic effect, which could also be considered as the overall genetic effect, without considering POEs. As mentioned in Section 1.1, genetic imprinting affects expression of genes and may explain some of the missing heritability of human complex traits and diseases. It is important to develop new methods applicable to genome-wide scans that model the differential contribution of paternal and maternal alleles. It is desired that a method that allows for POE also maintain the power to detect the main allelic effect after adding one or more parameters related to POE to the model. Therefore, the proper and orthogonal decomposition of genetic variance renders the NOIA framework meaningful and useful to estimate main allelic effects along with the POE.

In conclusion, considering the advantages of the orthogonal NOIA model on detecting genetic effects, missing heritability and ignored GxG/GxE or imprinting effects by usual association approach, we propose to apply the NOIA orthogonal models to characterize the complex network between multiple genes, environmental factors and imprinting effect, for investigating the underlying architecture of human complex traits and diseases.

CHAPTER 2

Natural and Orthogonal Interaction Framework for Modeling Gene-Gene Interactions Applied to Cutaneous Melanoma

In this chapter, to evaluate the performance of the NOIA statistical model on detecting gene-gene interactions that was proposed by Alvarez-Castro and Carlborg [21], we applied the NOIA statistical model on both simulated and real data. For testing the gene-gene interactions in the association modeling, we added the interaction effect parameters into the modeling which always resulted in lost power to detect the main genetic effects, that is, the main additive and dominance effects. The usual functional one-locus model (equation (3)) uses natural substitution for the parameter estimations which renders a non-orthogonal model whenever a dominance component is modeled, which means that the hypothesis tests lose power when the interaction terms are incorporated into the modeling. However, the NOIA statistical model (equation (5)) overcomes this disadvantage because of its orthogonality [6]. That is, even when we add several additional parameters into NOIA modeling, the estimation of the original parameters will not be influenced. Therefore, we propose to formalize the NOIA statistical one-locus model in equation (5) to a two-locus model incorporating the detection of interactions. We also extended the usual functional one-locus model to compare the performance of this testing with the NOIA model. We evaluate the behavior of the NOIA statistical model over the usual functional model for detecting the genetic effects, through both simulation analyses and application on melanoma dataset.

With extensive simulation studies for both quantitative traits and case-control traits, we evaluated the performance of NOIA statistical model and usual functional model for detecting the main genetic effects and interaction effects. To evaluate the influence of the parameter setting on the

simulation results, we simulated different scenarios with positive, negative or zero values of the interaction terms. We also extended the NOIA statistical model to reduced models including the additive, dominant and recessive models. We evaluated the power and type I error for detecting the genetic effects. To further characterize the performance of the two models, we applied them to the melanoma dataset to search for casual variants and potential interaction effects influencing melanoma risk.

GWAS and family-based approaches have previously revealed several loci that influence CM risk. Several previous studies have shown that melanocortin 1 receptor (*MC1R*) located at 16q24.3, HERC2/ (HECT and RLD domain containing E3 ubiquitin protein ligase 2)/OCA2 at region 15q13.1 and cyclin-dependent kinase inhibitor 2A (*CDKN2A* or *p16*) at 9p21.3 are the most significant susceptibility genes for melanoma susceptibility [53, 54]. Although one-locus association studies have been applied widely to investigate melanoma risk widely, the gene-gene interactions underlying this disease have not been fully exploited. Understanding how these genetic loci and interactions influence the development of melanoma could provide important clues in the pathogenesis and treatment of melanoma.

In the following sections, we introduce the methodology development of the NOIA and usual functional two-locus interaction models and the design of the simulations studies. We describe the application of the newly developed methods on a genome-wide scale melanoma dataset.

2.1 Methods

2.1.1 Two-Locus Gene-Gene Interaction Models

We already described the one-locus NOIA statistical model in Section 1.3. To extend the model to a two-locus model allowing gene-gene interaction testing, we assumed that a quantitative trait is influenced by two diallelic loci, A and B. We use p_{ij} and q_{ij} to denote the genotype frequencies of genotype A_{ij} and B_{ij}, respectively. N_A is the number of reference allele A_2, which is equal to 0, 1 or 2

when the genotype is G_{11}, G_{12} or G_{22}, respectively. Similarly, N_B is the number of reference allele B_2. $\overline{N}_A$ and $\overline{N}_B$ denote the means of $N_A$and N_B, respectively, whereas $V_A$and V_B denote the variance of $N_A$and N_B, respectively. Therefore, $\overline{N}_A = p_{12} + 2p_{22}$, $V_A = p_{12} + 4p_{22} - (p_{12} + 2p_{22})^2$. Correspondingly, $\overline{N}_B = q_{12} + 2q_{22}$, $V_B = q_{12} + 4q_{22} - (q_{12} + 2q_{22})^2$.

For two-locus gene-gene interaction models, which were described by Alvarez-Castro and Carlborg [21], the vector of two-locus genotypic values, G_{AB}, can be built as follows:

$$G_{AB} = S_{AB} \cdot E_{AB} = (S_B \otimes S_A) \cdot E_{AB}, \tag{9}$$

if we assume that the two loci, A and B, and in linkage equilibrium. E_{AB} is the two-locus vector of genetic effects; S_{AB} is the two-locus genetic effect design matrix which is the Kronecker product of the design matrix of loci B and A. From NOIA one-locus statistical model (equation (5)), the two-locus modeling vectors G_{AB}, E_{AB} and design matrix S_{AB} can all be obtained by the Kronecker product of one-locus modeling as follows:

$$G_{AB} = \begin{pmatrix} G_{B_{11}} \\ G_{B_{12}} \\ G_{B_{22}} \end{pmatrix} \otimes \begin{pmatrix} G_{A_{11}} \\ G_{A_{12}} \\ G_{A_{22}} \end{pmatrix} = \begin{pmatrix} G_{B_{11}} \cdot G_{A_{11}} \\ G_{B_{11}} \cdot G_{A_{12}} \\ G_{B_{11}} \cdot G_{A_{22}} \\ G_{B_{12}} \cdot G_{A_{11}} \\ G_{B_{12}} \cdot G_{A_{12}} \\ G_{B_{12}} \cdot G_{A_{22}} \\ G_{B_{22}} \cdot G_{A_{11}} \\ G_{B_{22}} \cdot G_{A_{12}} \\ G_{B_{22}} \cdot G_{A_{22}} \end{pmatrix} = \begin{pmatrix} G_{1111} \\ G_{1211} \\ G_{2211} \\ G_{1112} \\ G_{1212} \\ G_{2212} \\ G_{1122} \\ G_{1222} \\ G_{2222} \end{pmatrix}, \tag{10}$$

$$S_{AB}{}^S = S_B{}^S \otimes S_A{}^S = \begin{pmatrix} 1 & -\overline{N}_B & -\frac{2q_{12}q_{22}}{V_B} \\ 1 & 1-\overline{N}_B & \frac{4q_{11}q_{22}}{V_B} \\ 1 & 2-\overline{N}_B & -\frac{2q_{11}q_{12}}{V_B} \end{pmatrix} \otimes \begin{pmatrix} 1 & -\overline{N}_A & -\frac{2p_{12}p_{22}}{V_A} \\ 1 & 1-\overline{N}_A & \frac{4p_{11}p_{22}}{V_A} \\ 1 & 2-\overline{N}_A & -\frac{2p_{11}p_{12}}{V_A} \end{pmatrix}, \tag{11}$$

$$E_{AB}{}^S = E_B{}^S \otimes E_A{}^S = \begin{pmatrix} 1 \\ \alpha_A \\ \delta_A \end{pmatrix} \otimes \begin{pmatrix} 1 \\ \alpha_B \\ \delta_B \end{pmatrix} = \begin{pmatrix} \mu \\ \alpha_A \\ \delta_A \\ \alpha_B \\ \alpha\alpha \\ \delta\alpha \\ \delta_B \\ \alpha\delta \\ \delta\delta \end{pmatrix}. \tag{12}$$

Therefore, the vector of genotypic values, G_{AB}, can be expressed as

$$\begin{pmatrix} G_{1111} \\ G_{1211} \\ G_{2211} \\ G_{1112} \\ G_{1212} \\ G_{2212} \\ G_{1122} \\ G_{1222} \\ G_{2222} \end{pmatrix} = \begin{pmatrix} 1 & -\bar{N}_B & -\frac{2q_{12}q_{22}}{V_B} \\ 1 & 1-\bar{N}_B & \frac{4q_{11}q_{22}}{V_B} \\ 1 & 2-\bar{N}_B & -\frac{2q_{11}q_{12}}{V_B} \end{pmatrix} \otimes \begin{pmatrix} 1 & -\bar{N}_A & -\frac{2p_{12}p_{22}}{V_A} \\ 1 & 1-\bar{N}_A & \frac{4p_{11}p_{22}}{V_A} \\ 1 & 2-\bar{N}_A & -\frac{2p_{11}p_{12}}{V_A} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_A \\ \delta_A \\ \alpha_B \\ \alpha\alpha \\ \delta\alpha \\ \delta_B \\ \alpha\delta \\ \delta\delta \end{pmatrix}. \quad (13)$$

Through this derivation, we obtain the coding matrix, ${S_{AB}}^S$, for two-locus association along with gene-gene interactions modeling testing by linear regression. For this model, there are nine parameters to be inferred, including one baseline term (μ), two additive terms ($\alpha_A$and α_B), two dominant terms (δ_A and δ_B), and four interaction terms ($\alpha\alpha$, $\delta\alpha$, $\alpha\delta$ and $\delta\delta$). This was a full model including both additive effects and dominant effects. Reduced models, including additive, dominant, and recessive models, were also extended (Appendix 2.1).

As described in Section 1.2, the one-locus genotypes are usually coded as (-1, 0, 1) or (0, 1, 2) for the additive effect in the usual approach. Dominance effect is sometimes added for full modeling. Both of these two models are called a functional model, as it reflects the functionality of the alleles at the locus. Unlike the statistical model, the genetic effects from this functional model are using natural substitutions rather than based on the population effects which depend upon genotype frequencies. Similarly, using the (0, 1, 2) coding approach in equation (3), the two-locus genetic effect design matrix can be obtained as the Kronecker product of the two design matrices,

$${S_{AB}}^F = {S_B}^F \otimes {S_A}^F = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 2 & 2 & 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 2 & 4 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (14)$$

Therefore, the genotypic values could be expressed as

$$\begin{pmatrix} G_{1111} \\ G_{1211} \\ G_{2211} \\ G_{1112} \\ G_{1212} \\ G_{2212} \\ G_{1122} \\ G_{1222} \\ G_{2222} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 2 & 2 & 2 & 0 & 0 & 0 \\ 1 & 2 & 0 & 2 & 4 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ d_A \\ a_B \\ aa \\ da \\ d_B \\ ad \\ dd \end{pmatrix}. \quad (15)$$

Herein we use Greek letters for the genetic effects to distinguish with those from the statistical model. Reduced models, including additive, dominant, and recessive models, were also extended for the usual functional model (Appendix 2.1). As in the one-locus functional model, the estimation of the parameters was not based on the genotype frequencies and therefore reflects the main and interaction effects in a different way compared with the NOIA model. This model is also not orthogonal. The relationship between the NOIA statistical model and usual functional model can be derived through ${S_{AB}}^{F}{E_{AB}}^{F} = {S_{AB}}^{S}{E_{AB}}^{S}$ [21] as follows:

$$\begin{pmatrix} R \\ a_A \\ d_A \\ a_B \\ aa \\ da \\ d_B \\ ad \\ dd \end{pmatrix} = \left({S_{AB}}^{F^{-1}} \cdot {S_{AB}}^{S}\right) \cdot \begin{pmatrix} \mu \\ \alpha_A \\ \delta_A \\ \alpha_B \\ \alpha\alpha \\ \delta\alpha \\ \delta_B \\ \alpha\delta \\ \delta\delta \end{pmatrix}. \quad (16)$$

2.1.2 Simulation Studies on Quantitative Traits and Qualitative Traits

We performed simulation analyses for both quantitative and case-control traits by applying the NOIA statistical GxG interaction model and the usual functional GxG interaction model.

To simulate samples of independent individuals with a quantitative trait controlled by two diallelic loci, we assumed that there was no linkage disequilibrium among the two markers. For locus A, a value of the minor allelic frequency (p) was given in the simulated population. Genotypes A_{11}, A_{12} and A_{22} were assigned to an individual with probabilities $(1-p)^2$, $2p(1-p)$and p^2 respectively. Similarly, the minor allelic frequency (p) was given to locus B. Genotype B_{11}, B_{12} and

B_{22} were assigned to an individual with probabilities $(1-q)^2$, $2q(1-q)$ and q^2 respectively. From a prespecified vector of parameters, $\vec{E}_F{}^T = [R, a_A, d_A, a_A, d_A, aa, ad, da, dd]$, we assigned each individual a genotypic value according to his/her assigned two-locus genotypes. Then, by randomly generating a value from a normal distribution with prespecified mean and variance (0 and σ_e^2), we generated an observed phenotype/trait by adding this residual to the previously assigned genotypic. We used data from 2000 individuals as a replicate and simulated 1000 replicates for each genetic model.

In this part of our investigation of quantitative traits, three scenarios were simulated with different interaction terms (Table 2.1). The minor allele frequencies for both SNPs were set to 0.3, and the residual variance σ_e^2 was 144.0. The true values of the nine parameters in these three scenarios are shown in Table 2.1.

To investigate whether the setting of allele frequency influences the testing of the effects, we also simulated another scenario for quantitative traits. The minor allele frequency was set to be 0.49. The pre-specified value for the other terms remained the same.

Ma et al. [22] thoroughly derived the formulation of the statistical model in quantitative traits and demonstrated that a similar statistical model could also be defined for a qualitative trait by handling the genetic effects as the logit scale of the outcome. Similarly, we performed a case-control simulation analysis in our study. We used logistic regression and Bayes theorem to set the genotypic values of each individual according to the prespecified genetic effect terms, $\vec{E}_F{}^T$. For each replicate, 1000 cases and 1000 controls were simulated, and a total of 1000 replicates were simulated. The minor allele frequency was set to 0.30. Three scenarios were simulated with different generating values for the interaction terms. The generated values of the parameters in the three different scenarios are shown in Table 2.1.

Table 2.1 Simulation parameter values of genetic effects for quantitative and case-control traits dataset. R is the intercept term; a_A and d_A are the additive and dominant effects of locus A; a_B and d_B are the additive and dominant effects of locus B; aa, ad, da and dd are the interaction effects between locus A and locus B. Interaction coefficients are positive values for scenario 1, negative for scenario 2, and zero for scenario 3 which means no interaction. Main additive effect and dominant effect all exists in every scenario for both traits.

	R	$\mathbf{a_A}$	$\mathbf{d_A}$	$\mathbf{a_B}$	$\mathbf{d_B}$	**aa**	**ad**	**da**	**dd**
Quantitative trait									
Scenario 1	100.00	1.50	0.40	1.10	0.50	0.80	0.23	0.32	0.12
Scenario 2	100.00	1.50	0.40	1.10	0.50	-0.80	-0.23	-0.32	-0.12
Scenario 3	100.0	1.50	0.40	1.10	0.50	0	0	0	0
Case-control trait									
Scenario 1	-2.0	0.50	0.30	0.40	0.37	0.15	0.08	0.10	0.04
Scenario 2	-2.0	0.50	0.30	0.40	0.37	-0.15	-0.08	-0.10	-0.04
Scenario 3	-2.0	0.50	0.30	0.40	0.37	0	0	0	0

2.1.3 Application on Melanoma Susceptibility

We applied the NOIA statistical model and the usual functional model to the Cutaneous Melanoma (CM) data, samples from a genome-wide case-control study including 1804 cases and 1026 controls. The SNPs were genotyped from Illumina Omni 1-Quad_v1-0_B array and 783,945 SNPs remained after the quality control and other filtering procedures were applied [8]. The CM samples were collected from patients treated at The University of Texas MD Anderson Cancer Center between 1998 and 2008, and the controls were collected from the friends of the patients with matched sex and age during the same period. All the participants were non-Hispanic whites. The details of the genome-wide case-control study have been described previously [8]. The initial goal of that study was to detect novel loci that predisposed whites to CM. The objective of the current study was to apply the newly developed methods to validate the already identified potential causal SNPs

and gene-gene interactions that contribute to melanoma risk. We also attempted to compare the performance of the NOIA statistical model with that of the usual functional model on genetic effects detection. Logistic regression was used for the genetic effects estimation, and the P values were obtained using the Wald test statistic with the null hypothesis that the coefficient was zero. The Manhattan plots for the P values tested for the additive, dominant and interaction effects were graphed by Haploview software.

2.2 Results

2.2.1 Simulation Studies on Quantitative Traits and Qualitative Traits

We performed the simulation analysis on both simulated quantitative traits and case-control datasets. For each trait, analyses of three scenarios were performed when there were positive, negative or zero values for the interaction coefficients (Table 2.1). In each case, the minor allele frequency of locus A and locus B was both 0.30, and the residual variance was 144.0 for the quantitative trait.

First, we performed simulation studies on a quantitative trait under three scenarios. Our first simulation exhibited both main effects of two genes and their interactions with the true effect values $\vec{E}_F^{\,T} = [\mathrm{R}, \mathrm{a_A}, \mathrm{d_A}, \mathrm{a_B}, \mathrm{d_B}, \mathrm{aa}, \mathrm{ad}, \mathrm{da}, \mathrm{dd}] = [\,100.00, 1.50, 0.40, 1.10, 0.50, 0.80, 0.50, 0.32, 0.12]$. Figure 2.1 illustrates the power of the NOIA statistical model and usual functional model on detecting the four main genetic effects including the additive effects and the dominant effects of locus A and locus B, and four interaction effects between locus A and locus B. For detecting the main genetic effects, the NOIA statistical model clearly had greater power than the usual functional model, especially for additive effects (Fig. 2.1, upper panel). The NOIA statistical model also exhibited slightly greater or equal power than the usual functional model for detecting the interaction effects except the dominance by dominance effect (Fig. 2.1, bottom panel). The density distributions of the parameters estimated from these replicates was shown in Figure S2.1. Clearly, the variance of

all the main genetic effects (a_A, d_A, a_B and d_B) and most of the interaction effects (aa, ad and da) estimated from the NOIA statistical model was much smaller than those from the usual functional model (Fig. S2.1). Furthermore, the estimations of the genetic effects were both accurate for the two models, as the peaks were all located around the simulated true values (Fig. S2. 1).

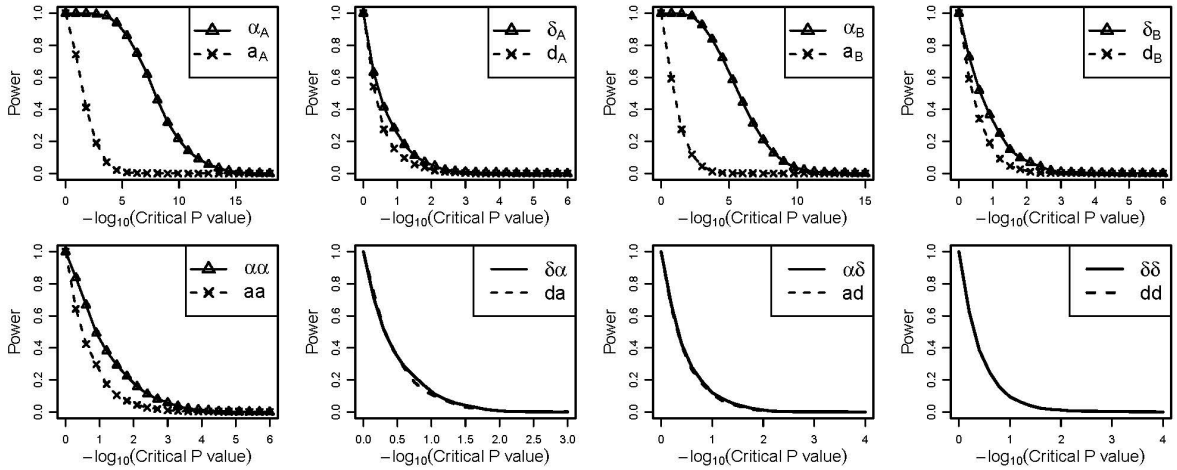

Figure 2.1 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation dataset under scenario 1 when the interaction terms were positive. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were $\vec{E}_F^{\,T} = [\,100.00, 1.50, 0.40, 1.10, 0.50, 0.80, 0.50, 0.32, 0.12]$. Corresponding values of the statistical genetic effects were $\vec{E}_S^{\,T} = [102.39, 2.35, 0.59, 1.97, 0.74, 1.04, 0.28, 0.37, 0.12]$.

To explore whether the values of the interaction terms influence the estimations of the parameters, we analyzed another scenario in which the interaction effect coefficients were set to be negative values and $\vec{E}_F^{\,T} = [100.00, 1.50, 0.40, 1.10,\ 0.50, -0.80, -0.23, -0.32, -0.12]$. A similar pattern with the first scenario was detected for the power of detecting the genetic effects; however, in this scenario the preference of the statistical NOIA model over the usual functional model in detecting the main effect of locus A and locus B was not obvious (Fig. 2.2). For some of the parameters, the usual functional model even showed slightly greater power than the NOIA statistical model.

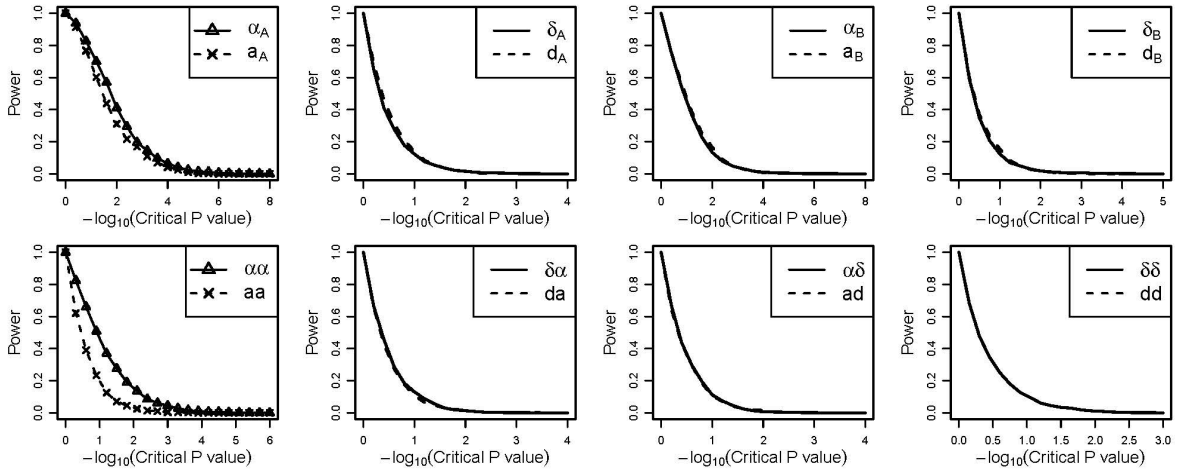

Figure 2.2 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation dataset under scenario 2 when the interaction coefficients were negative. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were
$\vec{E}_F{}^T = [100.00, 1.50, 0.40, 1.10,\ 0.50, -0.80, -0.23, -0.32, -0.12]$. Corresponding values of the statistical genetic effects were
$\vec{E}_S{}^T = [101.49, 0.97, 0.21, 0.63, 0.26, -1.04,\ -0.28, -0.37, -0.12]$.

We also analyzed a third scenario, in which there were no epistastic effects and only the main genetic effects from the two loci influence the trait (Fig. 2.3). In this scenario, the NOIA statistical model still had greater power for detecting the main genetic effects (Fig. 2.3, upper panel). The NOIA statistical and usual functional model yielded similar false positive rates for detecting the interaction effects, both of which were close to the nominal value (Fig. 2.3, bottom panel). The density distributions of the parameters estimated from these replicates in scenario 2 and scenario 3 of quantitative traits simulations are shown in Figure S2.2-S2.3.

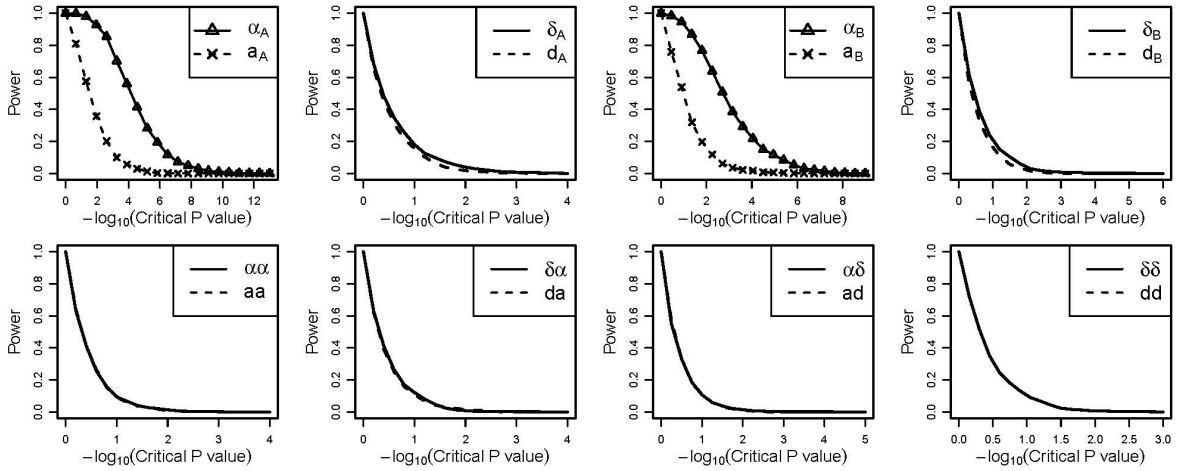

Figure 2.3 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation dataset under scenario 3 when no interaction effects present. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were $\vec{E}_F^{\,T} = [100.00, 1.50, 0.40, 1.10,\ 0.50, 0.0, 0.0, 0.0, 0.0]$. Corresponding values of the statistical genetic effects were $\vec{E}_S^{\,T} = [101.94, 1.66, 0.40, 1.30, 0.50, 0.0, 0.0, 0.0, 0.0]$.

Figures 2.4-2.6 show the results obtained from the case-control trait simulations. In Figure 2.4, the simulating values of the genetic effects were $\vec{E}_F^{\,T} = [-2.00, 0.50, 0.30, 0.40, 0.37, 0.15, 0.08, 0.10, 0.04]$, in which main genetic effects and interaction effects influence the outcome trait and the interaction coefficients were positive values. Similar to the simulation studies of the quantitative traits, the NOIA statistical model had greater power than the usual functional model for detecting most of the genetic effect terms. The parameter of the dominant-dominant interaction effect was exactly the same between these two models, which is expected from the equation of the models (equation (16)). We can see $\mathrm{dd} = \delta\delta$ after computation of the equation (16) which means that the parameters are identical. The test statistic for these two parameters should be identical too which can be implied from the test statistic of the dominance effect detection shown in Section 1.3. Interestingly, when we set the interaction terms to be negative values, where $\vec{E}_F^{\,T} = [-2.00, 0.50, 0.30, 0.40, 0.37, -0.15, -0.08, -0.10, -0.04]$, the power of

both models for detecting additive effects of locus A or locus B were similar to the power of these two models when the interaction terms were positive (Fig. 2.5).

For the third scenario, in which no interaction effects were present for the case-control trait, the power of the NOIA statistical model was still greater than that of the usual functional model to detect the main effects, while the false positive rates for detecting the interaction effects remained the same (Fig. 2.6). For all the scenarios we simulated, the density distributions of the eight parameters are presented in Figure S2.4-2.6. The estimation of the genetic effects was accurate, and the variance of the effects from the NOIA statistical modeling was less than that from the usual functional model for most parameters.

In the above analyses, we simulated the minor allele frequency of the two loci to be 0.3. We also studied setting the minor allele frequency to be 0.5 (Fig. S2.7-9). In most scenarios we simulated, the NOIA statistical model still had greater power than the usual functional model for detecting the main genetic effects and slightly greater power in detecting the interaction effects except for the scenarios when the interaction coefficients were negative values (Fig. S2.9). To evaluate the false positive rates of the two models, we also simulated a null scenario where no any effect existed. The false positive rates of the NOIA statistical model in the 0.05 significance level for detecting the eight genetic effects are: 0.051, 0.044, 0.054, 0.044, 0.048, 0.061, 0.055 and 0.058. The false positive rates of the usual functional model for detecting the eight genetic effects are: 0.042, 0.04, 0.04, 0.037, 0.051, 0.058, 0.052 and 0.058.

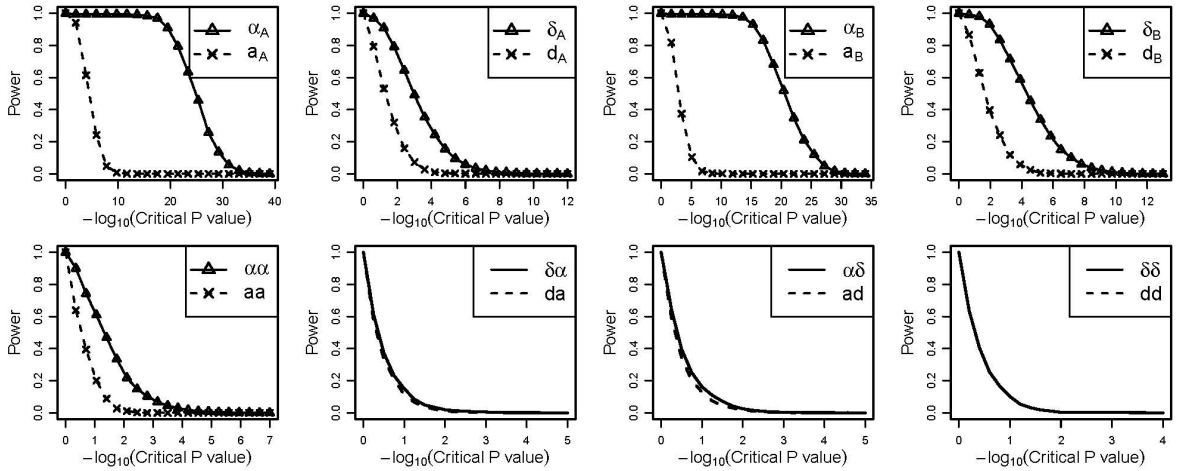

Figure 2.4 Power under different critical values of the P values obtained using the Wald test for the case-control simulation dataset under scenario 1 when positive interaction effects present. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were

$\vec{E}_F^{\,T} = [-2.00, 0.50, 0.30, 0.40, 0.37, 0.15, 0.08, 0.10, 0.04]$. Corresponding values of the statistical genetic effects were

$\vec{E}_S^{\,T} = [-1.07, 0.78, 0.36, 0.70, 0.45, 0.23, 0.10, 0.12, 0.04\ ]$.

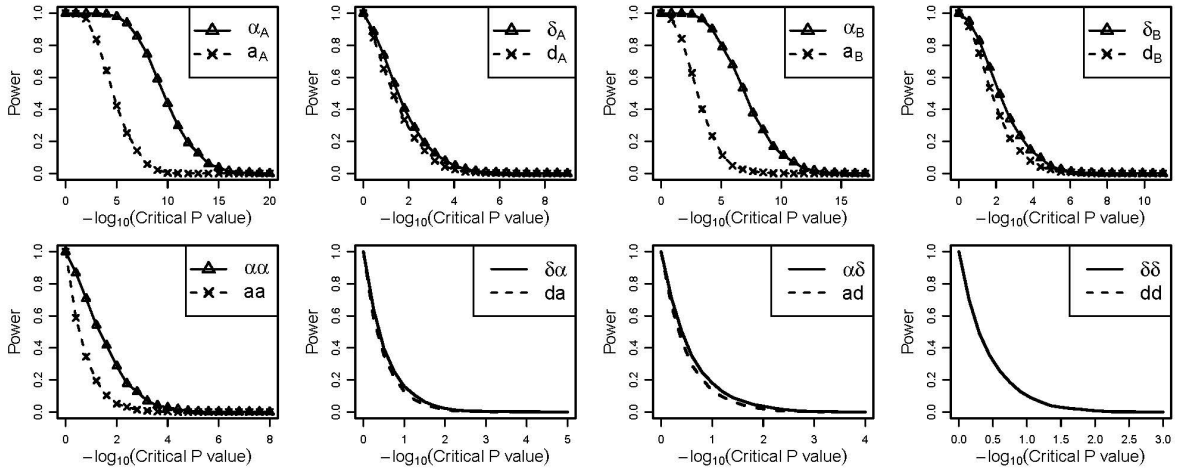

Figure 2.5 Power under different critical values of the P values obtained using the Wald test for the case-control simulation dataset under scenario 1 when negative interaction effects present. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were

$\vec{E}_F^{\,T} = [-2.00, 0.50, 0.30, 0.40, 0.37, -0.15, -0.08, -0.10, -0.04]$. Corresponding values of the statistical genetic effects were

$\vec{E}_S^{\,T} = [-1.29, 0.46, 0.24, 0.39, 0.29, -0.23, -0.10, -0.12, -0.04\ ]$.

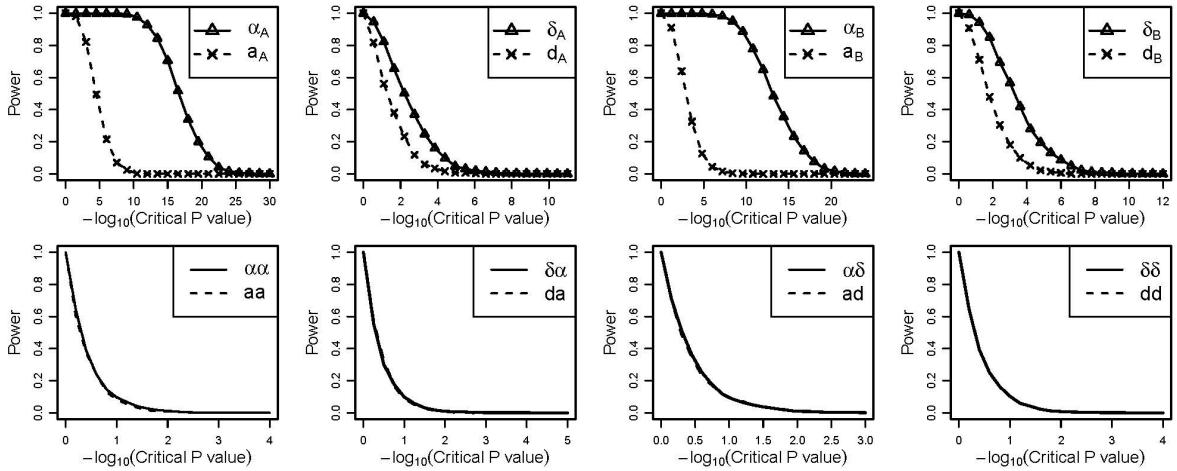

Figure 2.6 Power under different critical values of the P values obtained using the Wald test for the case-control simulation dataset under scenario 1 when negative interaction effects present. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were
$\vec{E}_F{}^T = [-2.0, 0.5, 0.3, 0.4, 0.37, 0.0, 0.0, 0.0, 0.0]$. Corresponding values of the statistical genetic effects were
$\vec{E}_S{}^T = [-1.18, 0.62, 0.30, 0.55, 0.37, 0.0, 0.0, 0.0, 0.0]$.

2.2.2 Application on a Real Dataset: Melanoma Susceptibility

To evaluate the performance of the NOIA statistical model and usual functional model, we carried out GWAS in the 2831 white participants, including 1805 cases and 1026 controls. To identify novel and verify the previously identified potential causal SNPs, we performed initial analyses using the one-locus NOIA statistical additive model. The Q-Q plot for the sample is shown in Figure S2.10. No obvious inflation of the test ($\gamma = 1.011$) was observed for the test statistic. Same estimations for the genetic effects were found for the one-locus usual functional additive model as expected (Section 1.4). Next, we applied the one-locus NOIA model with dominance component included to the melanoma dataset. SNPs with very few frequency of rare homozygotes (genotype cut off value was 0.005) were filtered and the Q-Q plot is shown in Figure S2.11 ($\gamma = 1.014$).The one-locus association results showed that 9 SNPs were significant at the genome-wide association level (5.0×10^{-8}) and 140 SNPs were significant at the 1.0×10^{-4} significance level (Table 2.2 and Table S2.1). Of the most significant SNPs that contribute to melanoma risk, two regions were found

to be genome-wide significant (Table 2.2). They are located on 15q13.1 (centered at the *HERC2*/*OCA2* region and 16q24.3 *MC1R* region). These two most significant SNPs located in these two regions are rs1129038 ($P = 3.73 \times 10^{-8}$, odds ratio [OR] = 0.70, 95% confidence interval [CI] =0.61-0.79) and rs4785751 ($P = 1.13 \times 10^{-10}$, OR=1.43, 95% CI=1.29 -1.60), respectively. The risk variants of these two SNPs were A and G, respectively. The SNPs located around *MTAP* were shown to be the third highly significant regions which are located at 9p21.3. The most significant SNP, SNP9-21789598 ($P = 4.15 \times 10^{-7}$), is located at the 5'-UTR of the *MTAP* gene, close to the *CDKN2A* gene.

Table 2.2 Top SNPs result from genome-wide association analysis of melanoma by NOIA statistical one-locus model using logistic regression ($p < 1.0 \times 10^{-6}$). The odds ratio (OR), confidence interval (CI) and P value are shown for the additive effect testing.

CHR	SNP	A1	A2	A2 freq	Position	OR(95%CI)	P value	Gene Symbol
16	rs4785751	A	G	0.53	88556918	1.43(1.29-1.60)	1.13E-10	DEF8
16	rs4408545	A	G	0.54	88571529	1.43(1.28-1.59)	3.81E-10	AFG3L1
16	rs11076650	A	G	0.46	88595442	1.40(1.26-1.56)	1.65E-09	DBNDD1
16	rs8051733	A	G	0.36	88551707	1.42(1.27-1.59)	2.66E-09	DEF8
16	rs7195043	A	G	0.50	88548362	0.72(0.64.0.80)	5.73E-09	DEF8
16	rs11648898	A	G	0.18	88573487	1.57(1.35-1.84)	1.46E-08	AFG3L1
15	rs1129038	A	G	0.22	26030454	0.70(0.61-0.79)	3.73E-08	HERC2
16	rs4785752	A	G	0.53	88562642	0.73(0.66-0.82)	4.14E-08	DEF8
16	rs4785759	A	C	0.53	88578381	0.73(0.66-0.82)	4.26E-08	AFG3L1
15	rs12913832	A	G	0.78	26039213	1.43(1.25-1.62)	6.15E-08	HERC2
16	rs10852628	A	G	0.31	88607428	1.40(1.24-1.58)	6.94E-08	DBNDD1
9	rs6475552	A	G	0.50	21691674	1.32(1.19-1.48)	3.71E-07	LOC402359
9	SNP9-21789598	A	G	0.49	21789598	0.75(0.68-0.84)	4.15E-07	MTAP
9	rs7848524	A	G	0.50	21691432	0.76(0.68-0.84)	4.28E-07	LOC402359
16	rs4238833	A	C	0.40	88578190	1.34(1.20-1.50)	4.56E-07	AFG3L1
9	rs2383202	A	G	0.49	21700215	1.32(1.19-1.47)	5.24E-07	LOC402359
9	rs12380505	A	G	0.50	21685893	0.76(0.68-0.85)	6.02E-07	LOC402359
9	rs1335500	A	G	0.49	21701675	1.32(1.18-1.47)	6.24E-07	LOC402359
9	rs1452658	A	G	0.50	21690795	1.32(1.18-1.47)	7.22E-07	LOC402359

To compare the performance of the NOIA statistical model with that of the usual functional model on a one-locus association study, we compared the top SNPs identified by these two models in a Manhattan plot (Fig. 2.7). The NOIA statistical model showed a highly significant signal in the *HERC2* regions (Fig. 2.7a) at 15q13.1 whereas the usual functional model did not (Fig. 2.7b). The

identification of the other two regions at 9p21.3 and 16q24.3 were similar for the two models. The results and signals we have reported so far are for the estimation of additive effect. No obvious signal for the dominance effects was identified by either model (data not shown).

We further applied the extended NOIA statistical model (equation (13)) and the usual functional model (equation (15)) on the two-locus association study in which gene-gene interactions testing were incorporated. Attempting to identify potential SNPs that interacted with the two significant genes (*HERC2* and *MC1R*) while contributing to the association with melanoma risk, we selected rs1129038 and rs4785751, the two most significant SNPs, as the reference SNPs for the two-locus scan, respectively. We then performed a genome-wide, two-locus scan by treating these two SNPs as reference SNPs separately and compared the performance of the NOIA statistical model with that of the usual functional model for detecting the main genetic and interaction effects.

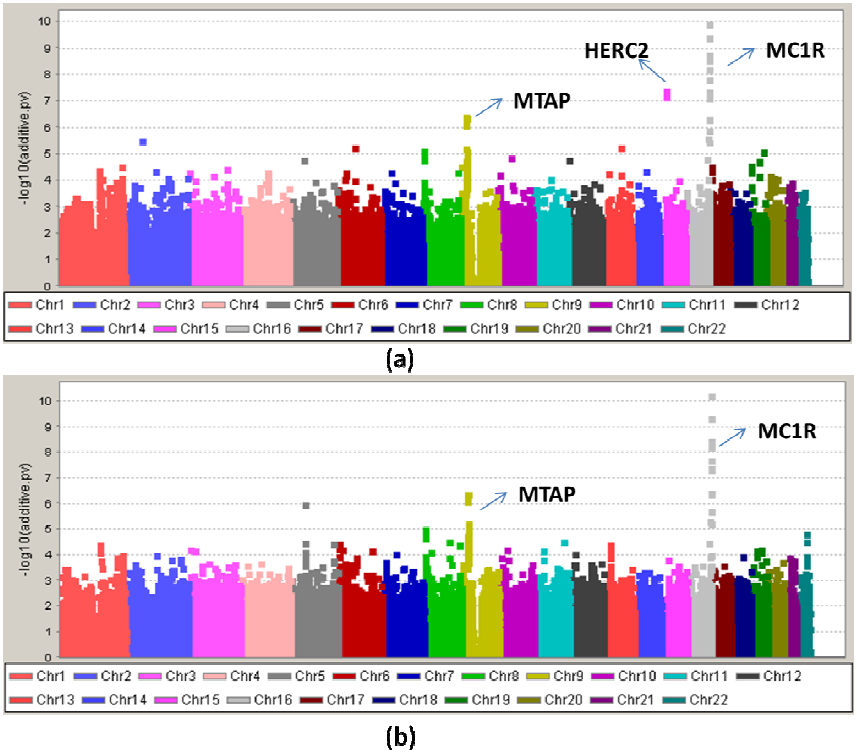

Figure 2.7 Manhattan plot for the genome-wide association studies of the CM susceptibility by one-locus scan. Detection of the additive effect through (a) the NOIA statistical model and (b) the usual functional model.

First, we performed the analysis for SNP rs1129038 in *HERC2* region. For the additive effects evaluation, the NOIA statistical model still showed a strongly significant signal with P value in the 1×10^{-10} significance level on the two significant regions adjacent to the MTAP and around MC1R genes, whereas the functional model had no obvious signal (Fig. 2.8a-b). Compared to the one-locus scan (Fig. 2.7a), the overall power for detecting the additive effect did not decrease in the NOIA two-locus model when more parameters were added in the model (Fig. 2.7a; 2.8a). This advantage did not emerge for the functional model (Fig. 2.7b; 2.8b). Moreover, no significant signal was observed for the dominant effects by either model (data not shown). For the four interaction terms, except the dominant-by-additive (da) interaction term, no obvious signal was identified by either model. A series of significant SNPs around gene *IL31RA* (interleukin-31 receptor A) and *DDX4* on chromosome 5 were identified by the NOIA statistical model for interaction with rs1129038 at the dominant-by-additive interaction term (Fig. 2.8c), where the da term means the interaction between the additive effect of the rs1129038 and the dominant effect of the candidate interacted SNP. These signals were not identified by the usual functional model (Fig. 2.8d). We then checked the linkage disequilibrium (LD) status between the significant SNPs around the *IL31RA* gene and the significant SNPs around the *DDX4* gene, showing that the two genes are in strong LD.

Table 2.3 presents the top SNPs interacted with rs1129038 at the da interaction term analyzed by the NOIA statistical two-locus interaction model. Four SNPs near *IL31RA* and three SNPs near *DDX4* were showing significant interaction with rs1129038 at the 1.0×10^{-6} significance level. However, other than the da interaction effect and the main additive effect from rs1129038, no main effects from the candidate interacted SNPs were identified.

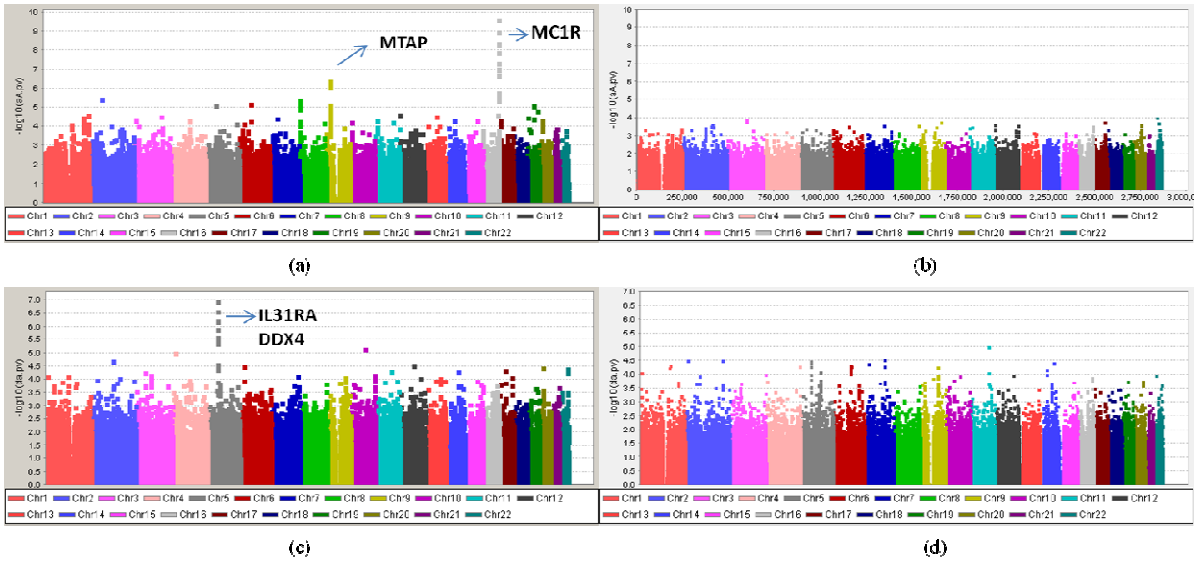

Figure 2.8 Manhattan plot for the genome-wide association studies of the CM susceptibility by two-locus scan for rs1129038. Detection of the additive effect (and) through (a) the NOIA statistical model and (b) the usual functional model; detection of the dominant-additive interaction effect (and) through (c) the NOIA statistical model and (d) the usual functional model.

Similarly, we compared the performance of the gene-gene interaction models for detecting genes that interacted with rs4785751 located around *MC1R* gene. Comparing the performance of the NOIA statistical model and that of the usual functional model on detection of the main additive effects, the former still remained the signal for those identified strongly associated regions while the latter did not (Fig. 2.9). No interaction effects were identified by either model for rs4785751 (data not shown).

Table 2.3 P values for the main effects and interaction effects when rs1129038 are used for reference SNP in the two-locus association analysis by NOIA statistical model ($p < 10^{-6}$). Add=additive effect, Dom=dominant effect, Add-Add=additive-additive interaction effect, Dom-Add=dominant-additive interaction effect; Add-Dom=additive-dominant interaction effect; Dom-Dom=dominant-dominant interaction effect. Locus B is the reference SNP, rs1129038; Locus A is the candidate interacted SNP that scanned from the whole genome.

CHR	SNP	Coordinate	Gene Symbol	P value							
				Add_A	Dom_A	Add_B	Dom_B	Add-Add	Dom-Add	Add-Dom	Dom-Dom
5	rs6871296	55175024	LOC402216	0.07	0.02	5.78E-08	0.46	0.45	**1.06E-07**	0.90	0.91
5	rs3857290	55182187	IL31RA	0.08	0.03	6.26E-08	0.45	0.45	**2.57E-07**	0.87	0.83
5	rs6876491	55181692	IL31RA	0.04	0.02	4.19E-08	0.33	0.47	**5.37E-07**	0.86	0.92
5	rs10042075	55178483	IL31RA	0.03	0.01	3.01E-08	0.35	0.51	**6.03E-07**	0.86	0.73
5	rs327240	55216666	IL31RA	0.02	0.25	2.61E-08	0.34	0.61	**1.20E-06**	0.75	0.27
5	rs3843458	55090435	DDX4	0.03	0.12	6.69E-08	0.26	0.46	**2.60E-06**	0.70	0.60
5	rs957459	55118231	DDX4	0.07	0.05	6.65E-08	0.28	0.46	**2.92E-06**	1.00	0.36
5	rs10035707	55098280	DDX4	0.06	0.05	6.43E-08	0.28	0.46	**4.07E-06**	0.99	0.33
10	rs12775320	78584174	KCNMA1	0.61	0.42	1.95E-07	0.00	0.50	**6.58E-06**	0.78	0.51
4	rs6825100	17305532	KIAA1276	0.27	0.10	7.30E-08	0.15	0.49	**9.49E-06**	0.04	0.28

Finally, we also applied the reduced NOIA statistical model and the reduced usual functional model, the additive models (details shown in Appendix 2.1), for detecting the gene-gene interactions that contribute to melanoma risk. For the second reference SNP, rs4785751, significant SNPs were identified for the additive-by-additive interaction effect (p value=7.07×10^{-6}) by both the NOIA and usual additive models (Table 2.4). These SNPs are located in chromosome 4 close to gene *PGRMC2*.

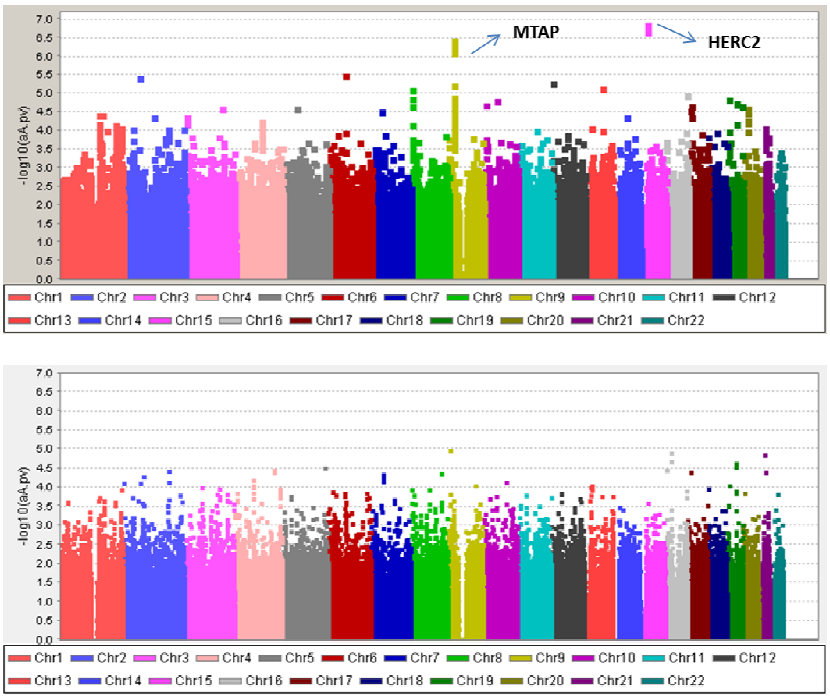

Figure 2.9 Manhattan plot for the genome-wide association studies of the CM susceptibility by two-locus scan for rs4785751. Detection of the additive effect (and) through the NOIA statistical model and (b) the usual functional model.

Table 2.4 p values and estimates for the main effects and interaction effects when rs4785751 was used for reference SNP in the two-locus association analysis by NOIA additive statistical model ( ).

CHR	SNP	Coordinate	Gene Symbol	Estimates			P value		
				Add_A	Add_B	Add-Add	Add-A	Add_B	Add-Add
4	rs10009093	143849384	FLJ44477	0.08	0.37	-0.55	0.32	5.29E-11	**7.07E-06**
4	rs10019366	129459033	PGRMC2	0.02	0.36	-0.44	0.76	1.09E-10	**9.23E-06**
4	rs4975181	129466845	PGRMC2	0.10	0.37	-0.35	0.09	8.91E-11	**9.71E-06**
4	rs11723210	129500895	PGRMC2	0.06	0.36	-0.45	0.44	1.24E-10	**9.84E-06**

2.3 Discussion

In most scenarios we simulated, the NOIA statistical model presented greater power for detecting additive effects and some interaction effects compared with the usual functional model. The NOIA model also yielded more precise estimators. Moreover, the investigation of type I error showed no significant difference of these two models. Real data analyses on the melanoma dataset also showed preference of the NOIA statistical model by one-locus and two-locus genome-wide scan. The epistasis analyses on the melanoma dataset showed that the NOIA statistical model preserved power for detecting the main genetic effects. The functional model lost power when multiple loci were jointly modeled. The NOIA statistical model identified potential epistasis between the rs1129038 (located around *HERC2* gene) and *IL31RA* gene while the functional model did not. Another significant region interacting with rs4785751 near *MC1R* gene was also identified, PGRMC2 located at 4q26, by the NOIA statistical additive model allowing gene-gene interactions.

On the other hand, by applying the additive one-locus NOIA model and usual functional model, we found that their performance for detecting the additive effect was the same. This can be explained by the theoretical evidence in Appendix 2.2. To further explore the reason why the NOIA model has preserved power on detecting the additive effect when the dominance component is included, whereas the usual model does not, we constructed the test statistic for each model in Appendix 2.2. From the test statistics shown there, we can clearly see the underlying mechanism is still related with the fact that the covariance between the additive and dominance component is not zero if the usual model is applied.

We also found another significant characteristic of the NOIA one-locus framework. Both the full NOIA and functional models are ill-conditioned if the SNP has only two or less genotypes in the population. This arose because of value of the determinant of the design matrix (X^TX) is equal to zero when the dominance component is included in the testing. For the additive models, the determinant of the design matrix is not equal to zero (Appendix 2.2). When the design matrix is noninvertible, one needs to use generalized inverses as another alternative in setting up the tests

rather than the inverse procedure that in R programming software. It has always been a problem in R that the inverse procedures are not robust. Therefore, in our Q-Q plots, the λ values were less than 1(0.86~0.92) when performing NOIA or usual functional one-locus models with dominance component testing on melanoma dataset. After we removed the SNPs with minor genotype frequency less than 0.005, the λ values were 1.014 and 1.023 respectively (Fig. S2.11). According to our analyses, we suggest to apply the NOIA full model in two stratifications for one-locus scan of real data. NOIA full model is preferred for SNPs with three types of genotypes to identify potential dominant effects while maintaining the power to detect additive effects. Additive model would be better to be applied for those SNPs to get right distribution when they do not have all three genotypes.

In this section, we compared the NOIA statistical model and usual functional model for analyzing epistasis, or gene interactions, for quantitative traits and dichotomous diseases. These two models were able to be transformed to each other and they had different meaning for their parameters. The NOIA statistical model focuses on the population properties whereas the usual functional model focuses on the biological properties. The methodology of the NOIA statistical model was developed early in 2007 [21], however their performance on detecting gene-gene interactions has not been tested or compared with the other models.

For the real data analyses, the NOIA statistical one-locus model provided confirmatory evidence of the association of three previously identified causal regions with melanoma risk, *HERC2* at 15q13.1, *MC1R* at 16q24.3 and *CDKN2A* at 9p21.3. Compared to the NOIA statistical model, the usual functional one-locus model did not detect the most significant region, the *HERC2* gene, which has been well characterized in the previous studies [8]. When we compare our analyzed results from the full usual functional model to those from the usual functional additive model (Fig. 2.7), we found that the *HERC2* signal was detected very clearly by the usual additive model but not by the usual full model. Thus we conclude the NOIA full model has greater power than the usual functional model in one-locus genome-wide scans.

The epistasis analyses showed that the power of the NOIA statistical model was greater than that of the usual functional model for detecting main genetic effects when interactions are included. When two loci and epistasis were modeled together, the usual functional model presented decreased power while the statistical model maintained its power for detecting the main genetic effects (Fig. 2.8a). This result reflects one of the important properties of the NOIA model, orthogonality. Using orthogonal models for quantitative traits analysis or binary diseases yields consistent genetic effect estimation in reduced models. Here, we clearly see that the functional model had no consistent genetic effect estimation when multiple loci were modeled together.

Moreover, the NOIA statistical model identified potential epistasis between the rs1129038 (located around *HERC2*) and a region at chromosome 5, whereas the functional model did not. This associated region is located in the 5'-UTR of *IL31RA* gene located at 5q11.2 and the intron of the gene *DDX4* located at 5p15.2-p13.1. The expression of *IL31RA* is induced in activated monocytes and is constitutively expressed in epithelial cells. The interesting aspect of this interaction is that no main genetic effect was found for these SNPs. The interaction is based on the dominant-by-additive interaction term. Although it is hard to interpret the dominant-additive interaction term here, it is possible that only the reference locus (rs1129038) has a main effect while a significant interaction effect exists for gene-gene interaction models. Another significant region that interacted with rs4785751, *PGRMC2* located at 4q26, was also identified, by the NOIA additive model. This is the first report of the implication of potential genes and regions that were shown to interact with SNPs associated with melanoma risk. If these interactions are confirmed by validation studies, there will be no doubt that the NOIA statistical model is preferred for epistasis detection compared to the usual functional model.

Whether there are factors that influence interaction effects without playing marginal/main effects has been a critical issue in genetic association studies [55]. In single-locus analysis, each locus is considered separately. Therefore, factors that influence interaction effects but not marginal effects will be missed, as they do not lead to marginal correlation between the genotype and outcome

phenotype. Our results highlight the application of NOIA interaction models for detecting both main and interaction effects, which could explain more heritability of human complex diseases. The usual functional models do not have this advantage because they may lose power when more parameters are added to the modeling.

Beyond two-locus interactions, we may also expect interaction of multiple loci, for instance, three-locus interactions. One may simply extend the full and additive NOIA statistical models by straightforwardly applying Kronecker products to additional loci. The NOIA three-locus interaction models on the significant SNPs contributing to the melanoma risk showed no signal for higher dimensional interactions. We also applied the three-locus interaction models on the significant SNPs contributing to lung cancer (the dataset we will use in next section) and we did not detect any three dimensional interactions. This may be because even less power is available to detect higher-order models. Large datasets will be required to estimate these parameters accurately. Interpreting the interactions is also complicated even for two-locus interactions. Validation from replication analysis and experiments to explain how these factors interact with each other is a challenging task. The underlying mechanism of the interactions may also difficult to explain.

The difference between the NOIA statistical model and the usual functional model lies in their focus. The former is characterized by orthogonal parameters that denote average effects of allele substitutions over population, whereas the latter focuses on the natural allele substitutions for parameter estimation. They are different viewpoints of a similar analysis. Nonetheless, when investigating the epistasis or gene-environment interactions, choosing the most appropriate framework is still important. We still recommend using the NOIA statistical model for epistasis study because of its greater power and its desired statistical properties.

CHAPTER 3

Natural and Orthogonal Interaction Framework for Modeling Gene-Environment Interactions with Application to Lung Cancer

GWASs have not been effective in identifying much of the heritability of the human diseases as single SNP is isolated and analyzed by this approach. Much heritability has not been able to be explained especially when multiple loci or binary environmental exposure are jointly influencing disease risk. In the previous section, we formulized the NOIA orthogonal model on modeling joint contribution of multiple loci to the diseases development or quantitative levels of a trait. In this section, we propose to develop a statistical approach for modeling effects from genetic factors and environmental exposure, from the orthogonal NOIA model. We included a binary environmental exposure and its interaction with gene in the modeling of a quantitative trait. We evaluated the performance of the newly developed NOIA gene-environment (GxE) interaction model by comparing its statistical behavior with the usual models on simulated datasets for quantitative traits.

> We also explored 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. Used with permission from S. Karger AG, Basel, Ma J. et al: Hum Hered 2012; 73: 185–194. [22]

We then applied the NOIA GxE modeling on a real lung cancer dataset. As widely known, smoking is by far the main risk factor of lung cancer as well as genetic factors. Previous studies have identified three chromosomal regions at 15q25, 5p15 and 6p21 as being significantly associated with the susceptibility of lung cancer [8, 56-58]. Replication studies then indicated interaction of smoking

and 15q25 variants in predisposing white populations to lung cancer susceptibility [59]. Our motivation is to validate the potential causal variants of lung cancer risk and identify potential interactions between these variants and smoking through the newly developed approach. In following sections, the performance of the usual functional model and NOIA statistical model will also be evaluated for the testing of effects from genetic factors and smoking exposure in full models and reduced models.

3.1 Methods

3.1.1 Methodology Development of the NOIA Gene-Environment Interaction Model

One-locus Model

We already introduced the one-locus usual functional model and NOIA statistical model in Chapter One.

> As shown in Alvarez-Castro and Carlborg, 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. [22]
>
> For the analysis of case-control data sampled according to a qualitative trait such as a 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 long 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
>
> $$\mathrm{Var}(\bar{\beta}) = \sigma^2(\chi^{\mathrm{T}}\chi)^{-1},$$
>
> 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. However, for logistic regression, the variance of estimates of parameters is
>
> $$\mathrm{Var}(\bar{\beta}) = \sigma^2(\chi^{\mathrm{T}}\nu\chi)^{-1},$$
>
> where ν is a diagonal matrix with elements
>
> $$\pi_{G_i}(1-\pi_{G_i})$$
>
> for the i-th individual in the sample with π_{G_i} the probability of being affected given the values of repressor for the individual. It can be shown that
>
> $$\mathrm{X}^{\mathrm{T}}\mathrm{V}\,\mathrm{X} = \mathrm{n}\mathrm{S}^{\mathrm{T}}\mathrm{D}'\mathrm{S},$$
>
> where
>
> $$\mathrm{D}' = \begin{pmatrix} \pi_{11}(1-\pi_{11})\mathrm{p}_{11} & 0 & 0 \\ 0 & \pi_{12}(1-\pi_{12})\mathrm{p}_{12} & 0 \\ 0 & 0 & \pi_{22}(1-\pi_{22})\mathrm{p}_{22} \end{pmatrix},$$

and S is a design matrix. This means that, for logistic regression, the statistical model defined in equation (5) 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. Used with permission from S. Karger AG, Basel, Ma J. et al: Hum Hered 2012; 73: 185–194. [22]

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. We denote the unexposed frequency by m. A functional model for this environmental exposure is

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

with effects defined as

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

For a two-level factor, following Alvarez-Castro and Carlborg, the criterion for orthogonality can be derived as follows: from the regression model

$$\begin{pmatrix} M_1^* \\ M_2^* \\ \vdots \\ M_n^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = S\vec{E} = X\vec{E},$$

orthogonality requires that

$$X^T X = nS^T Z^T Z S = S^T D S$$

is diagonal, where

$$D = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix},$$

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

$$X^T X = n \begin{pmatrix} m_1 s_{11}^2 + m_2 s_{21}^2 & m_1 s_{11} s_{12} + m_2 s_{21} s_{22} \\ m_1 s_{11} s_{12} + m_2 s_{21} s_{22} & m_1 s_{12}^2 + m_2 s_{22}^2 \end{pmatrix}.$$

It follows that the model S is orthogonal when

$$m_1 s_{11} s_{12} + m_2 s_{21} s_{22} = 0.$$

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

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

with effects defined as

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

Applying the Kronecker product rule, we have the following non-orthogonal functional model for the gene-environment interaction:

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

and the following statistical model:

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

The relation between the statistical and functional models is

$$\begin{pmatrix} \mu \\ \alpha_G \\ \delta_G \\ \alpha_M \\ \alpha\alpha \\ \delta\alpha \end{pmatrix} = \begin{pmatrix} 1 & \bar{N} & p_{12} & 1-m & (1-m)\bar{N} & (1-m)p_{12} \\ 0 & 1 & p'_{12} & 0 & 1-m & (1-m)p'_{12} \\ 0 & 0 & 1 & 0 & 0 & 1-m \\ 0 & 0 & 0 & 1 & \bar{N} & p_{12} \\ 0 & 0 & 0 & 0 & 1 & p'_{12} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_G \\ d_G \\ a_M \\ aa \\ da \end{pmatrix}. \quad [22]$$

The formulation of the three reduced genetic models of the statistical and functional GxE and their relationships are shown in the supplementary text in Ma et al. Used with permission from S. Karger AG, Basel, Ma J. et al: Hum Hered 2012; 73: 185–194. [22]

3.1.2 **Simulation Studies on Quantitative Traits and Qualitative Traits**

The simulation methods we are using in this section are similar to the simulation methods we mentioned in Chapter 2 for GxG interactions. Here, we set the exposed frequency m to be 0.22 for the simulated population. The allele frequencies (p) for the SNP were set to 0.30. Genotype 111, 121, 221, 112, 122, or 222 were assigned to an individual with probabilities $(1-p)^2(1-m)$, $2p(1-p)(1-m)$, $p^2(1-m)$, $(1-p)^2m$ $2p(1-p)m$ or p^2m respectively. From a prespecified vector of parameters, $\vec{E}_F{}^T = [R, aG, dG, aM, aa, da]$, we assigned each individual a phenotypic value

according to his/her assigned one locus genotypes and exposure status. Then, by randomly generating a value from a normal distributions with prespecified mean and variance (0 and σ_e^2), we generated an observed phenotype/trait by adding this residual to the previously assigned phenotypic value. The residual variance σ_e^2 was 144.0. We used data from 2000 individuals as a replicate and simulated 1000 replicates for each genetic model.

In our investigation of quantitative traits, two scenarios were simulated with different effects terms.

> 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
>
> $$\Pr(d = 1|i) = \frac{1}{1+\exp(-G_i)},$$
>
> 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 distributions of the six genotypes in the cases as follows
>
> $$\Pr(i|d = 1) = \frac{P_i/(1+\exp(-G_i))}{\sum_j P_j/(1+\exp(-G_j))},$$
>
> where P_i is the frequency of genotype i in the population, given by $(1-p)^2(1-m)$, $2p(1-p)(1-m)$, $p^2(1-m)$, $(1-p)^2m$ $2p(1-p)m$ or 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:
>
> $$\Pr(i|d = 0) = \frac{P_i/(1+\exp(-G_i))}{\sum_j P_j/(1+\exp(-G_j))}.$$
>
> 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. Used with permission from S. Karger AG, Basel, Ma J. et al: Hum Hered 2012; 73: 185–194. [22]

For the case-control trait, two scenarios were simulated with different effects terms. The minor allele frequencies for the markers were set to 0.25. The unexposed frequency was set to 0.22. The residual variance σ_e^2 was 144.0. We used data from 2000 individuals as a replicate and simulated 1000 replicates for each genetic model.

3.1.3 **Application on Lung Cancer Susceptibility**

> We applied the NOIA statistical model and the usual functional model to the ILCCO (International lung cancer consortium) data, 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 [59]. 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 17836 Caucasians with 7392 cases and 10444 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. Used with permission from S. Karger AG, Basel, Ma J. et al: Hum Hered 2012; 73: 185–194. [22]

A wald test was performed for the null hypothesis test that there is no association. The testing models can be generally shown as following,

$$\text{Logit (qualitative trait)} = \beta_0 + \beta_1 * G + \beta_2 * E + \beta_3 * GxE + \text{sex} + \text{age} + D_1 + \ldots + D_n.$$

In the above expression, G denotes the genetic value; E denotes the environmental exposure value which is cigarettes smoking status here (0 for non-smokers and 1 for smokers); GxE is for the interaction values between the genetic effect and environmental effect. D_i (i=1, 2…16) is the dummy variable for the independent study groups. For both the NOIA statistical models and the usual functional models, logistic regression was performed, with sex, age and study group as covariates. We applied the full models including GxE interactions testing, the reduced models including GxE interactions testing, the full models without GxE interactions testing and the additive models without GxE interactions testing to the ILCCO dataset.

3.2 Results

3.2.1 Simulation Studies on Quantitative Traits and Qualitative Traits

We conducted extensive simulation analyses on both simulated quantitative traits and case-control traits. First, we performed simulation studies on a quantitative trait under two scenarios. The pre-specified minor allele frequency and exposure frequency was 0.30 and 0.22 respectively. The simulating residual variance was 144.0.

Our first simulation exhibited both effects from the gene and the environmental factor along with the interactions. The true effect values were $\vec{E}_F^{\,T} = [R, a_G, d_G, a_M, aa, da] = [\,100.0,\ 3.0,\ 1.0,\ 2.0,\ 1.5,\ 1.0]$. The corresponding statistical genetic effects values $\vec{E}_S^{\,T}$ could then be calculated by the minor allele frequency, exposure frequency and the actual functional terms. Then, we got $\vec{E}_S^{\,T} = [\mu, \alpha_G, \delta_G, \alpha_M, \alpha\alpha, \delta\alpha] = [101.75,\ 4.18,\ 1.22,\ 2.71,\ 2.2,\ 1.0]$. Figure 3.1 illustrates the power of the NOIA statistical model and usual functional model on detecting the effects from the genetic factor including the additive and dominant effects, the environmental exposure, and the additive-by-environment interaction. For detecting the main genetic effects and environmental effects, the NOIA statistical model clearly showed greater power than the usual functional model, especially for additive effects (Fig. 3.1, upper panel and left bottom panel). The NOIA statistical model also exhibited slightly greater power than the usual functional model for detecting the interaction effects (Fig. 3.1, right bottom panel). The density distribution of the parameters estimated from these replicates was shown in Figure S3.1. Clearly, the variance of the genetic additive effect and interaction effect estimated from the NOIA statistical model was much smaller than that from the usual functional model (Fig. S3.1). The variance of the other effects estimated from these two models was very close to each other. Furthermore, the estimations of the genetic effects were both accurate for the two models, as the peaks were all located around the simulated true values (Fig. S3. 1).

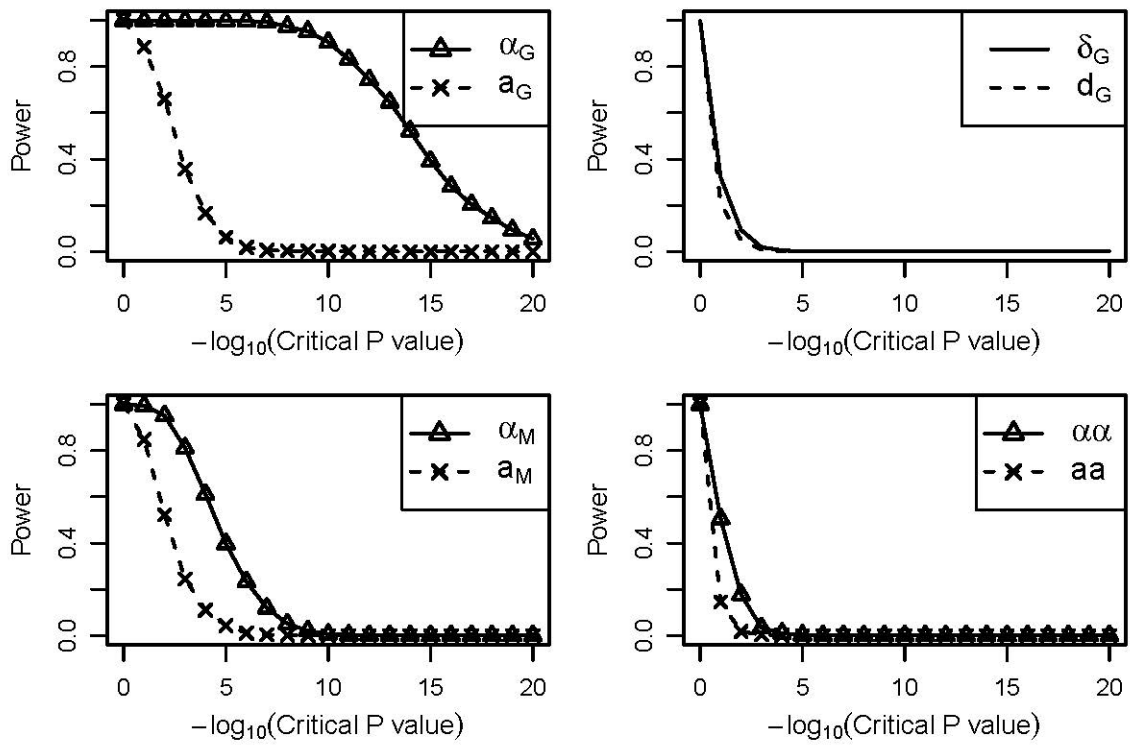

Figure 3.1 Power under different critical values of the P values obtained using the Wald test for the simulated data with a quantitative trait influenced by a genetic factor and an environmental factor. The pre-specified minor allele frequency and exposure frequency was 0.30 and 0.22 respectively. The simulating residual variance was 144.0. The values of the six parameters of the genetic effects were $\vec{E}_F{}^T = [\,100.0,\ 3.0,\ 1.0,\ 2.0,\ 1.5,\ 1.0]$. The corresponding statistical genetic effects were $\vec{E}_S{}^T = [101.75,\ 4.18,\ 1.22,\ 2.71,\ 2.2,\ 1.0]$.

Another simulation was performed for a scenario where a quantitative trait is only influenced by a genetic factor. The true effect values were $\vec{E}_F{}^T = [\,100.0,\ 3.0,\ 1.0,\ 0.0,\ 0.0,\ 0.0]$. The corresponding statistical genetic effects were $\vec{E}_S{}^T = [101.16,\ 3.70,1.00,0.00,0.00,0.00]$. Figure 3.2 shows the power of the NOIA statistical model and usual functional model on detecting the genetic effects, the environmental effect, and additive-by-environment interaction effect. For detecting the genetic additive effect, the NOIA statistical model clearly had greater power than the usual functional model (Fig. 3.2, upper left panel). For the other two parameters, the false positive rates were very close to the nominal level for both the NOIA statistical model and usual model (Fig. 3.2, bottom panel). The density distribution of the parameters estimated from these replicates was shown in Figure S3.2.

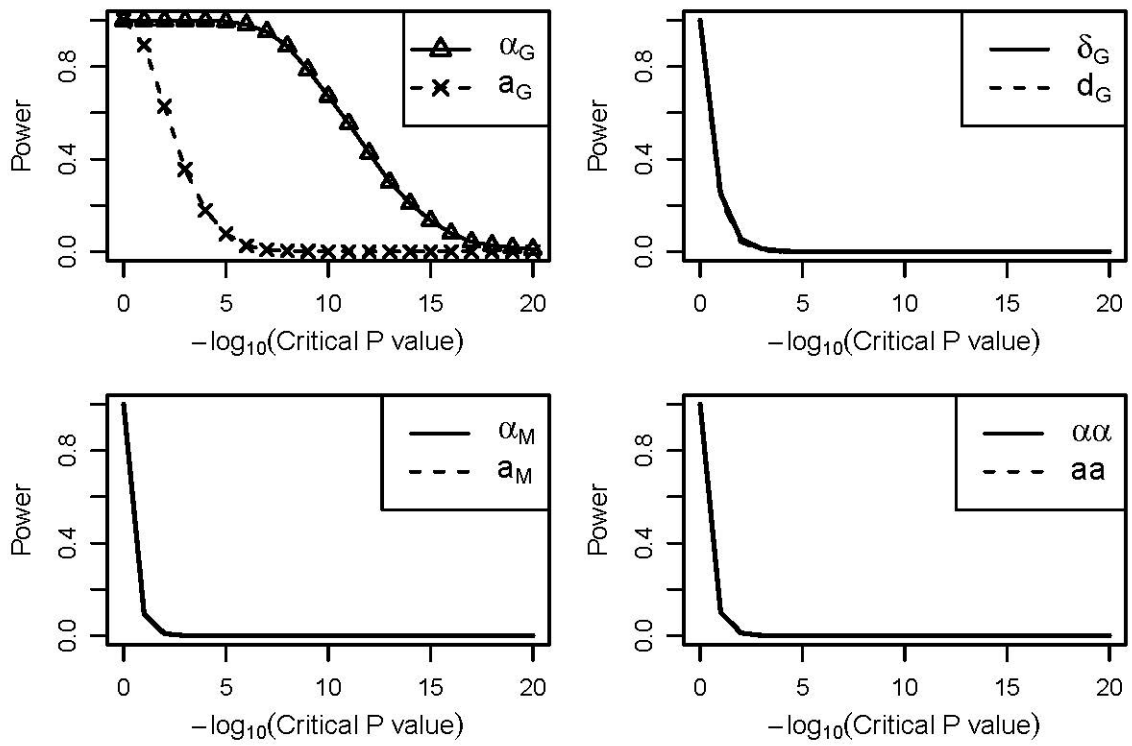

Figure 3.2 Power under different critical values of the P values obtained using the Wald test for the simulated data with a quantitative trait influenced by a genetic factor. The pre-specified minor allele frequency and exposure frequency was 0.30 and 0.22 respectively. The simulating residual variance was 144.0. The values of the six parameters of the genetic effects were $\vec{E}_F^{\,T} = [\,100.0,\ 3.0,\ 1.0,\ 0.0,\ 0.0,\ 0.0]$. The corresponding statistical genetic effects were $\vec{E}_S^{\,T} = [101.16,\ 3.70,1.00,0.00,0.00,0.00]$.

Figures 3.3-3.4 show the results obtained from the case-control trait simulations. For the scenario when both a genetic factor and an environment factor influence the trait, the true effect values were $[-2.0,\ 0.3,\ 0.1,\ 0.2,\ 0.1,\ 0.04]$. The corresponding statistical genetic effects were $[-1.75,\ 0.38,\ 0.11,\ 0.27,\ 0.12,\ 0.04]$. Figure 3.3 shows that for detecting the main genetic effects and environmental effects, the NOIA statistical model clearly had greater power than the usual functional model, especially for additive effects (Fig. 3.3). The NOIA statistical model also exhibited slightly greater power than the usual functional model for detecting the interaction effects (Fig. 3.3, right bottom panel). The density distribution of the parameters estimated from these replicates was shown in Figure S3.3. Still, the variance of the genetic additive effect, environmental effect and

additive-by-environment interaction effect estimated from the NOIA statistical model was much smaller than that from the usual functional model (Fig. S3. 3).

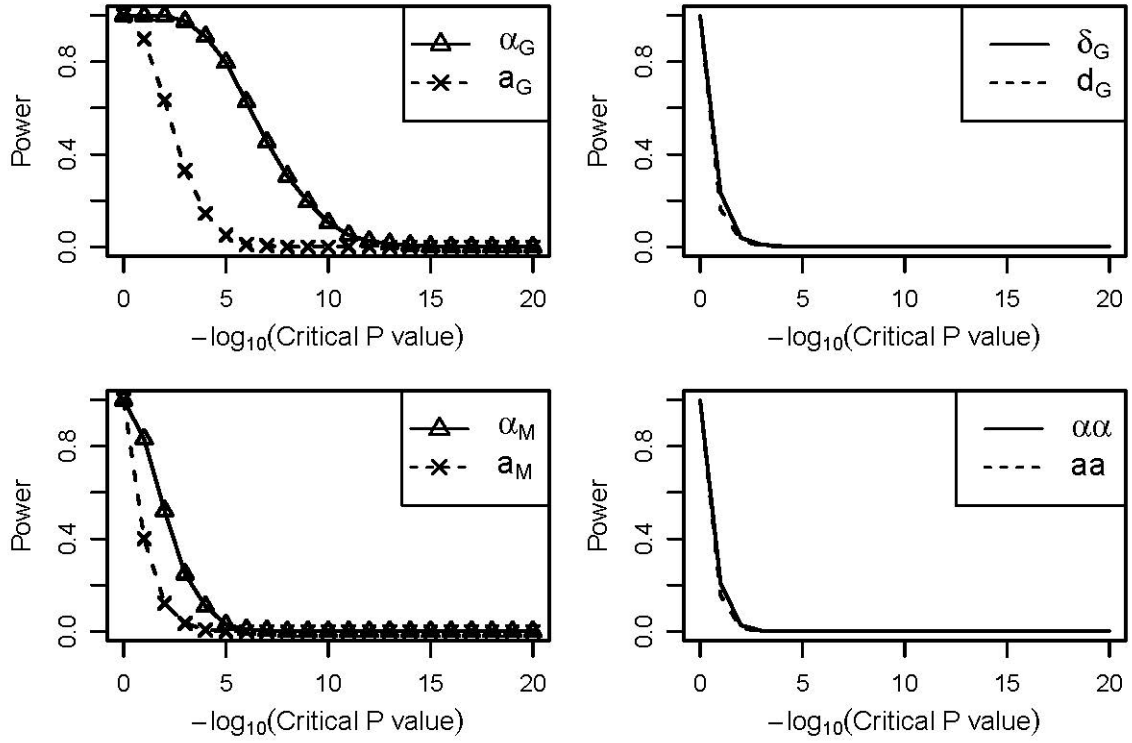

Figure 3.3 Power under different critical values of the P values obtained using the Wald test for the simulated data with a case-control trait influenced by a genetic factor and an environmental factor. The pre-specified minor allele frequency and exposure frequency was 0.25 and 0.22 respectively. The values of the six parameters of the genetic effects were $\vec{E}_F{}^T = [-2.0, 0.3, 0.1, 0.2, 0.1, 0.04]$. The corresponding statistical genetic effects were $\vec{E}_S{}^T = [-1.75, 0.38, 0.11, 0.27, 0.12, 0.04]$.

For the scenario when only a genetic factor influences the case-control trait (Fig. 3.4), the true effect values were $[-2.0, 0.4, 0.2, 0.0, 0.0, 0.0]$. The corresponding statistical genetic effects were $[-1.73, 0.5, 0.2, 0.0, 0.0, 0.0]$. For detecting the genetic effect, the NOIA statistical model clearly had greater power than the usual functional model, especially for the additive effect (Fig. 3.4, upper panel). For the other two parameters, the false positive rates were very close to the nominal level for both the NOIA statistical model and usual model (Fig. 3.4, bottom panel). The density distribution of the parameters estimated from these replicates was shown in Figure S3.4.

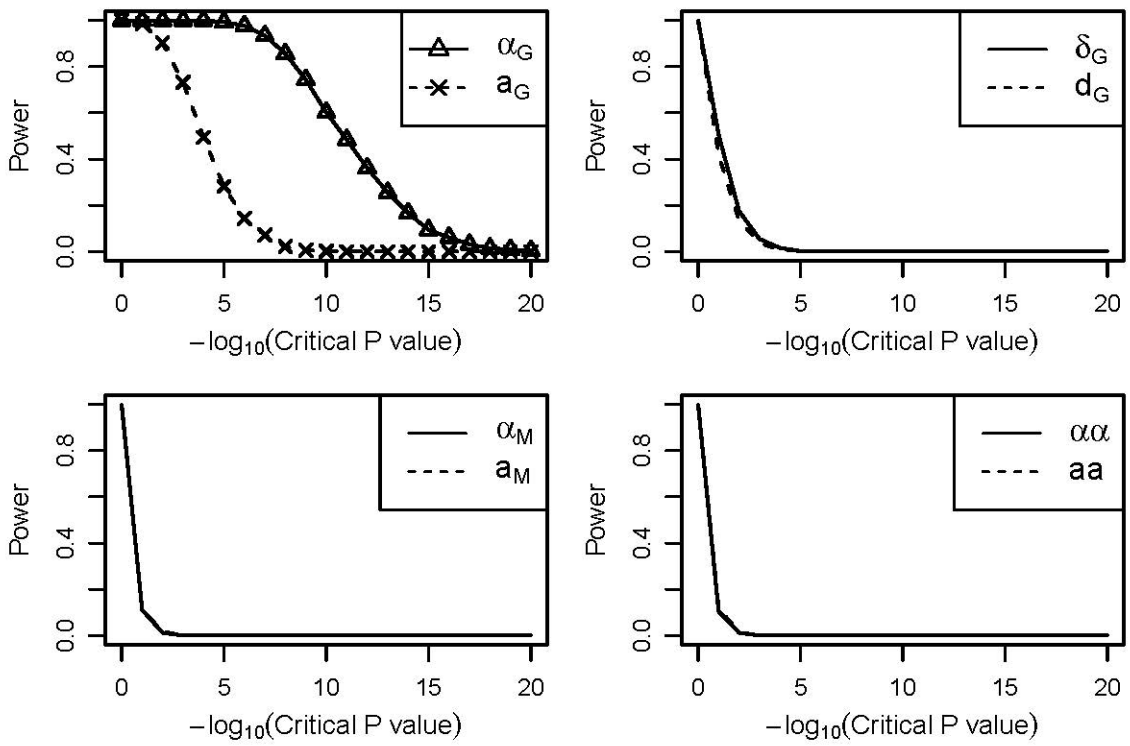

Figure 3.4 Power under different critical values of the P values obtained using the Wald test for the simulated data with a case-control trait influenced by a genetic factor. The pre-specified minor allele frequency and exposure frequency was 0.25 and 0.22 respectively. The values of the six parameters of the genetic effects were $\vec{E}_F{}^T = [-2.0,\ 0.4,\ 0.2,\ 0.0,\ 0.0,\ 0.0]$. The corresponding statistical genetic effects were $\vec{E}_S{}^T = [-1.73,\ 0.5,\ 0.2,\ 0.0,\ 0.0,\ 0.0]$.

To evaluate the false positive rate of the two models, we also simulated a null scenario where no any effect influences the quantitative trait. The false positive rates of the NOIA statistical model in the 0.05 significance level for detecting the five effects are: 0.058, 0.053, 0.043, 0.052 and 0.049. The false positive rates of the usual functional model for detecting the five effects are: 0.058, 0.058, 0.061, 0.055 and 0.05. They are very close to the nominal level for both models. For the qualitative traits, the false positive rates of the NOIA statistical model in the 0.05 significance level for detecting the five effects are: 0.047, 0.045, 0.047, 0.052 and 0.052. The false positive rates of the usual functional model for detecting the five effects are: 0.06, 0.05, 0.059, 0.058 and 0.052. The functional model has slightly higher false positive rates than the NOIA statistical model when applied on qualitative traits.

3.2.2 Application of the NOIA Model on Lung Cancer Susceptibility

Next, we conducted real data analyses by applying the NOIA statistical model and usual functional model to the ILCCO dataset. The statistics of the samples and the summary of the SNPs that we used in ILCCO dataset are shown in Table S3.1 and Table S3.2, respectively. There are about 41% females and 59% males in cases, 48% females and 52% males in controls. The distribution of the age and smoking status among cases and controls are also shown in Table S3.1. The risk allele and the minor allele frequencies for the 6 SNPs under investigation are shown in Table S3.2. We mainly focus on the three chromosomal regions (5p15, 6p21 and 15q25), including 6 potential causal variants. First, to evaluate the performance of the NOIA statistical model and usual functional model, we applied the additive models with no interaction effects testing to the dataset. The odds ratio with 95% confidence intervals and P values estimated from the NOIA statistical model and usual functional model are shown in Table 3.1. The two models had similar performance on detecting the genetic additive effect and smoking exposure effect. Consistent with previously published results, both models detected significant additive effects from 5p15 and 15q25. Table 3.2 shows the results when additive-by-smoking (Add-SM) interaction effect is incorporated in the models. Both models identified potential Add-SM interactions from the rs2256543 on 6p21 and the two SNPs on 15q25 predisposing to the lung cancer risk. The NOIA statistical (NOIA-Stat) model showed greater power on detecting the additive effect and same power on detecting the Add-SM interaction effect than the usual functional (Usual-Func) model. Comparing Table 3.1 and Table 3.2, we clearly state that after the Add-SM interaction term is added into the modeling, the additive Usual-Func model has extremely larger P values on detecting the additive effect than those obtained before, whereas the additive NOIA-Stat model preserved the power.

We then attempted to apply the full NOIA-Stat model with dominance component and the full Usual-Func model with dominance component to the ILCCO dataset. First, we performed the full models without GxE interactions testing. Table 3.3 illustrates the odds ratio with 95% confidence

intervals and P values. The full NOIA-Stat model and Usual-Func model had close coefficient estimators and P values on testing the significant additive effects from 5p15 and 15q25. Still, the P values estimated from the full NOIA-Stat model were slightly smaller than those from the full Usual-Func model. Moreover, we performed the full NOIA-Stat and Usual-Func models in which testing for additive effects, dominant effects and GxE interactions were incorporated on the real dataset (Table 3.4). Only the NOIA-Stat model identified potential Add-SM interactions from the rs2256543 on 6p21 contributing to the lung cancer risk. Both models identified the Add-SM interactions from the two SNPs on 15q25. Again, the full NOIA-Stat model showed greater power for detecting the additive effect compared to the full Usual-Func model.

To further validate the potential interaction effects of the two SNPs on 15q25 with smoking exposure, we performed stratified analyses on this dataset. First, we stratified the samples into two sub-populations: non-smokers and smokers. Then, we used (0, 1, 2) coding for the additive effect testing. We used sex, ages, study groups as covariates. The 95% confidence interval of the OR and P values are shown in Table 3.5. We can clearly state that the two SNPs on 15q25 are extremely significant in smokers (P value=$2.08 \times 10^{-23}$for rs16969968, P value=2.50×10^{-21} for rs8034191) contributing to lung cancer susceptibility. And the two SNPs are not significant in non-smokers.

For all SNPs and all models, the smoking effect was extremely significant. None of the SNPs had significant dominant effect or dominant-smoking interaction effect by any model. For all models, the estimation of the effect of sex and age were consistent because they were modeled in the same way. Furthermore, to explore whether there are any potential GxG interactions among these six SNPs predisposing white population to lung cancer, we performed the GxG models that mentioned in Chapter 2 on the ILCCO dataset. No interactions were identified (data not shown).

Table 3.1 Odds ratio and P values estimated from additive models when there were no interactions modeled [a].

SNPs	Model [b]	OR(95% CI)				p value			
		Add	SM	Sex	Age	Add	SM	Sex	Age
rs2736100_5p15	Func/Stat	**1.17(1.11-1.23)**	5.99(5.42-6.63)	1.26(1.16-1.36)	1.01(1.01-1.02)	**2.9E-10**	8.9E-268	8.8E-09	1.5E-16
rs402710_5p15	Func/Stat	**0.86(0.81-0.90)**	6.02(5.42-6.69)	1.26(1.16-1.36)	1.01(1.01-1.02)	**7.7E-09**	1.5E-244	6.8E-09	3.1E-37
rs2256543_6p21	Func/Stat	1.03(0.98-1.08)	5.99(5.43-6.62)	1.25(1.16-1.35)	1.01(1.01-1.02)	0.23	1.2E-271	8.2E-09	6.0E-16
rs4324798_6q21	Func/Stat	1.12(0.95-1.32)	5.93(5.37-6.55)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.28	3.4E-270	3.0E-08	6.1E-29
rs16969968_15q25	Func/Stat	**1.25(1.19-1.31)**	5.86(5.31-6.48)	1.23(1.14-1.32)	1.02(1.01-1.02)	**1.3E-19**	9.3E-266	3.5E-08	4.7E-28
rs8034191_15q25	Func/Stat	**1.29(1.22-1.36)**	5.11(4.50-5.81)	1.29(1.19-1.40)	1.02(1.02-1.03)	**1.4E-19**	1.2E-136	2.6E-09	3.2E-33

a. The study group has been used as covariates. Add=Additive effect; SM=smoking;
b. Func=The usual functional model; Stat=The NOIA statistical model.

Table 3.2 Odds ratio and P values estimated from additive models when interactions were modeled [a]

SNPs	Model [b]	OR(95% CI)					p value				
		Add	SM	Add-SM	Sex	Age	Add	SM	Add-SM	Sex	Age
rs2736100_5p15	Usual-Func	**1.23(1.10-1.39)**	6.02(5.45-6.67)	0.94(0.82-1.07)	1.26(1.16-1.36)	1.01(1.01-1.02)	**0.0005**	1.6E-265	0.33	9.0E-09	1.6E-16
	NOIA-Stat	**1.18(1.12-1.24)**	6.00(5.43-6.64)	0.94(0.82-1.07)	1.26(1.16-1.36)	1.01(1.01-1.02)	**1.9E-10**	2.2E-267	0.33	9.0E-09	1.6E-16
rs402710_5p15	Usual-Func	**0.88(0.76-1.00)**	5.96(5.31-6.71)	0.97(0.84-1.13)	1.26(1.17-1.36)	1.02(1.02-1.03)	0.055	1.2E-195	0.73	7.0E-09	3.1E-37
	NOIA-Stat	**0.86(0.81-0.91)**	6.01(5.41-6.68)	0.97(0.84-1.13)	1.26(1.17-1.36)	1.02(1.02-1.03)	**2.3E-08**	4.3E-244	0.73	7.0E-09	3.1E-37
rs2256543_6p21	Usual-Func	0.92(0.81-1.03)	6.10(5.52-6.76)	**1.15(1.01-1.31)**	1.26(1.16-1.36)	1.01(1.01-1.02)	0.15	2.4E-267	**0.036**	6.6E-09	5.2E-16
	NOIA-Stat	1.02(0.97-1.07)	6.00(5.44-6.64)	**1.15(1.01-1.31)**	1.26(1.16-1.36)	1.01(1.01-1.02)	0.47	1.2E-271	**0.036**	6.6E-09	5.2E-16
rs4324798_6q21	Usual-Func	1.07(0.64-1.64)	5.80(5.22-6.47)	1.05(0.66-1.80)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.60	3.0E-66	0.31	3.0E-08	6.2E-29
	NOIA-Stat	1.04(0.95-1.12)	5.94(5.38-6.56)	1.12(0.89-1.41)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.42	2.8E-270	0.31	3.0E-08	6.2E-29
rs16969968_15q25	Usual-Func	1.00(0.88-1.13)	6.29(5.66-7.01)	**1.31(1.14-1.49)**	1.23(1.14-1.32)	1.02(1.02-1.02)	0.99	4.0E-249	**0.0001**	3.7E-08	2.7E-28
	NOIA-Stat	**1.23(1.17-1.29)**	5.87(5.32-6.49)	**1.31(1.14-1.50)**	1.23(1.14-1.32)	1.02(1.02-1.02)	**9.2E-16**	4.9E-266	**0.0001**	3.7E-08	2.7E-28
rs8034191_15q25	Usual-Func	1.02(0.86-1.21)	5.43(4.74-6.24)	**1.29(1.08-1.55)**	1.29(1.19-1.41)	1.02(1.02-1.03)	0.79	1.9E-128	**0.0057**	2.3E-09	1.9E-33
	NOIA-Stat	**1.26(1.19-1.33)**	5.09(4.48-5.80)	**1.29(1.08-1.55)**	1.29(1.19-1.41)	1.02(1.02-1.03)	**3.1E-15**	1.8E-136	**0.0057**	2.3E-09	1.9E-33

a. The study group has been used as covariates. Add=Additive effect; SM=smoking.
b. Usual-Func=The usual functional model; NOIA-Stat=The NOIA statistical model.

Table 3.3 Odds ratio and P values estimated from full models when no interactions were modeled [a]

SNPs	Model [b]	OR(95% CI)					p value				
		Add	Domi	SM	Sex	Age	Add	Domi	SM	Sex	Age
rs2736100_5p15	Usual-Func	**1.17(1.11-1.23)**	1.00(0.93-1.07)	5.99(5.42-6.63)	1.26(1.16-1.36)	1.01(1.01-1.02)	**3.4E-10**	0.95	1.0E-267	8.8E-09	1.5E-16
	NOIA-Stat	**1.17(1.11-1.23)**	1.00(0.93-1.07)	5.99(5.42-6.63)	1.26(1.16-1.36)	1.01(1.01-1.02)	**2.9E-10**	0.95	1.0E-267	8.8E-09	1.5E-16
rs402710_5p15	Usual-Func	**0.86(0.81-0.90)**	0.99(0.91-1.07)	6.01(5.42-6.69)	1.26(1.16-1.36)	1.01(1.01-1.02)	**4.0E-07**	0.79	1.5E-244	6.8E-09	3.1E-37
	NOIA-Stat	**0.86(0.81-0.90)**	0.99(0.91-1.07)	6.01(5.42-6.69)	1.26(1.16-1.36)	1.01(1.01-1.02)	**7.7E-09**	0.79	1.5E-244	6.8E-09	3.1E-37
rs2256543_6p21	Usual-Func	1.03(0.98-1.08)	1.02(0.95-1.09)	5.99(5.43-6.62)	1.25(1.16-1.35)	1.01(1.01-1.02)	0.28	0.55	1.1E-271	8.1E-09	5.9E-16
	NOIA-Stat	1.03(0.98-1.08)	1.02(0.95-1.09)	5.99(5.43-6.62)	1.25(1.16-1.35)	1.01(1.01-1.02)	0.23	0.55	1.1E-271	8.1E-09	5.9E-16
rs4324798_6q21	Usual-Func	1.12(0.95-1.32)	0.92(0.77-1.11)	5.93(5.37-6.55)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.19	0.38	3.6E-270	2.9E-08	5.9E-29
	NOIA-Stat	1.05(0.97-1.13)	0.92(0.77-1.11)	5.93(5.37-6.55)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.27	0.38	3.6E-270	2.9E-08	5.9E-29
rs16969968_15q25	Usual-Func	**1.23(1.17-1.30)**	1.06(0.99-1.14)	5.86(5.31-6.48)	1.23(1.14-1.32)	1.02(1.01-1.02)	**1.3E-15**	0.10	1.0E-265	3.9E-08	5.7E-28
	NOIA-Stat	**1.25(1.19-1.31)**	1.06(0.99-1.14)	5.86(5.31-6.48)	1.23(1.14-1.32)	1.02(1.01-1.02)	**9.6E-20**	0.10	1.0E-265	3.9E-08	5.7E-28
rs8034191_15q25	Usual-Func	**1.27(1.20-1.35)**	1.05(0.97-1.13)	5.11(4.50-5.81)	1.29(1.19-1.40)	1.02(1.02-1.03)	**1.5E-16**	0.26	9.7E-137	2.7E-09	3.6E-33
	NOIA-Stat	**1.29(1.22-1.36)**	1.05(0.97-1.13)	5.11(4.50-5.81)	1.29(1.19-1.40)	1.02(1.02-1.03)	**1.1E-19**	0.26	9.7E-137	2.7E-09	3.6E-33

a. The study group has been used as covariates. Add=Additive effect; Dom=Dominant effect; SM=smoking.
b. Usual-Func=The usual functional model; NOIA-Stat=The NOIA statistical model.

Table 3.4 Odds ratio and P values estimated from full models when interactions were modeled [a]

		OR(95% CI)							p value						
SNPs	**Model**	**Add**	**Domi**	**SM**	**Add_SM**	**Domi_SM**	**Sex**	**Age**	**Add**	**Domi**	**SM**	**Add_SM**	**Domi_SM**	**Sex**	**Age**
rs2736100_5p15	Usual-Func	1.24(1.10-1.40)	1.03(0.87-1.23)	6.14(5.37-7.05)	0.93(0.82-1.07)	0.96(0.80-1.16)	1.25(1.16-1.36)	1.01(1.01-1.02)	**0.0005**	0.71	3.24E-149	0.3105	0.67	9.07E-09	1.66E-16
	NOIA-Stat	1.18(1.12-1.24)	1.00(0.93-1.07)	6.00(5.43-6.64)	0.94(0.82-1.07)	0.96(0.80-1.16)	1.25(1.16-1.36)	1.01(1.01-1.02)	**1.9E-10**	0.97	3.50E-267	0.3212	0.67	9.07E-09	1.66E-16
rs402710_5p15	Usual-Func	0.87(0.75-1.02)	1.00(0.82-1.23)	6.02(5.08-7.16)	0.98(0.83-1.16)	0.98(0.79-1.22)	1.26(1.17-1.36)	1.02(1.02-1.03)	0.0880	0.97	1.35E-93	0.8131	0.88	6.98E-09	3.13E-37
	NOIA-Stat	0.86(0.81-0.91)	0.99(0.91-1.07)	6.01(5.41-6.68)	0.97(0.84-1.13)	0.98(0.79-1.22)	1.26(1.17-1.36)	1.02(1.02-1.03)	**2.29E-08**	0.82	4.43E-244	0.7312	0.88	6.98E-09	3.13E-37
rs2256543_6p21	Usual-Func	0.92(0.82-1.04)	0.95(0.80-1.13)	5.84(5.09-6.71)	1.14(1.00-1.30)	1.09(0.91-1.32)	1.26(1.16-1.36)	1.01(1.01-1.02)	0.1950	0.55	3.99E-138	0.0579	0.36	6.56E-09	4.90E-16
	NOIA-Stat	1.02(0.97-1.07)	1.01(0.95-1.09)	6.00(5.44-6.64)	1.15(1.01-1.31)	1.09(0.91-1.32)	1.26(1.16-1.36)	1.01(1.01-1.02)	0.4594	0.68	1.18E-271	**0.0370**	0.36	6.56E-09	4.90E-16
rs4324798_6q21	Usual-Func	1.07(0.64-1.64)	0.86(0.53-1.49)	6.11(3.83-10.46)	1.05(0.66-1.80)	1.08(0.61-1.82)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.7826	0.57	8.06E-13	0.8370	0.78	2.91E-08	5.87E-29
	NOIA-Stat	1.04(0.95-1.12)	0.92(0.76-1.11)	5.94(5.38-6.56)	1.12(0.89-1.41)	1.08(0.61-1.82)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.4027	0.37	2.90E-270	0.3322	0.78	2.91E-08	5.87E-29
rs16969968_15q25	Usual-Func	1.00(0.87-1.14)	1.01(0.84-1.22)	6.10(5.25-7.11)	1.29(1.11-1.49)	1.06(0.87-1.30)	1.23(1.14-1.32)	1.02(1.02-1.02)	0.9511	0.91	3.69E-121	**0.0007**	0.56	4.13E-08	3.48E-28
	NOIA-Stat	1.23(1.17-1.29)	1.06(0.98-1.14)	5.87(5.32-6.49)	1.31(1.14-1.50)	1.06(0.87-1.30)	1.23(1.14-1.32)	1.02(1.02-1.02)	**7.4E-16**	0.13	6.94E-266	**0.0001**	0.56	4.13E-08	3.48E-28
rs8034191_15q25	Usual-Func	1.03(0.85-1.23)	0.98(0.76-1.26)	5.23(4.31-6.40)	1.27(1.05-1.54)	1.07(0.82-1.40)	1.29(1.19-1.41)	1.02(1.02-1.03)	0.7671	0.89	2.01E-60	**0.0165**	0.60	2.43E-09	2.19E-33
	NOIA-Stat	1.26(1.19-1.33)	1.04(0.96-1.13)	5.09(4.48-5.80)	1.29(1.08-1.55)	1.07(0.82-1.40)	1.29(1.19-1.41)	1.02(1.02-1.03)	**2.23E-15**	0.35	2.10E-136	**0.0055**	0.60	2.43E-09	2.19E-33

a. The study group has been used as covariates. Add=Additive effect; Dom=Dominant effect; SM=smoking.
b. Func=The usual functional model; NOIA-Stat=The NOIA statistical model.

Table 3.5 Analyses after stratification by smoking status for rs16969968 and rs8034191 in 15q25 [a]

	Non-smokers			**Smokers**		
SNPs	Add	Sex	Age	Add	Sex	Age
OR (95% CI) [b]						
rs16969968	**1.00(0.88-1.13)**	1.41(1.17-1.70)	1.00(1.00-1.01)	**1.31(1.24-1.38)**	1.19(1.09-1.29)	1.02(1.02-1.03)
rs8034191	**1.02(0.86-1.22)**	1.95(1.52-2.52)	1.02(1.01-1.03)	**1.32(1.25-1.40)**	1.21(1.11-1.33)	1.02(1.02-1.03)
P-value						
rs16969968	**0.95**	0.0003	0.42	**2.08E-23**	3.29E-05	1.41E-30
rs8034191	**0.81**	2.05E-07	0.003	**2.50E-21**	2.78E-05	2.90E-29

a The study group has been used as covariates. Add=Additive effect.
b OR=Odds Ratio; CI=Confidence Interval

3.3 Discussion

In this chapter, we described a new extension of the existing NOIA framework which was originally developed for testing GxG interactions for quantitative traits. We extended it to include a binary environment trait and explored the testing of GxE interactions along with the main effects. The NOIA model was also extended to case-control analyses, although some of the important properties of the NOIA model did not hold true. One is that estimates of the parameters from logistic regression for the case-control trait are no longer statistically uncorrelated under the alternative model that there is an association. Also, the estimates of the parameters from logistic regression of the full NOIA model are not consistent with those from the reduced (additive) NOIA model. We also showed the evidence of the second point by the real data analyses by comparing the results in Table 3.1 and Table 3.3.

By simulation studies, we stated that NOIA statistical models were usually more powerful than the functional models for detecting main effects and interaction effects for both quantitative traits and binary traits. This point is consistent with the results from real data analyses. When we performed additive models without interaction effect testing integrated, the NOIA-Stat model showed same performance with the Usual-Func model (Table 3.1). After the interaction effect was tested along with the additive effect, the NOIA-Stat preserved the power for identifying the additive effects whereas the Usual-Func did not (Table 3.2). When we performed full models without interaction effect testing integrated, the NOIA-Stat model presented similar performance with the Usual-Func model (Table 3.3). Similarly, after the interaction effect was tested along with the main effects, the NOIA-Stat model preserved the power for identifying the main genetic effects, whereas the Usual-Func model had greatly larger P values compared with those tested without interaction effects incorporated (Table 3.4).

The application of the NOIA-Stat model confirmed four SNPs in 5p15 and 15q25 region to be significantly associated with lung cancer susceptibility in Caucasians population: rs2736100,

rs402710, rs16969968 and rs8034191. The full Usual-Func model failed to identify them or with larger P values. We also validated that rs16969968 and rs8034191 in 15q25 region are significantly interacted with smoking in Caucasians population by stratification analyses. Potential interactions of SNP rs2256543 on 6p21 with smoking on contributing to lung cancer risk are indicated in our study which is the first report. It is interesting that no main effects were found for this SNP, however which happens in reality [60]. Such cases that display interaction but no marginal effect are usually ignored in usual one-locus analyses. However, this interaction needs more evidence to be validated.

Comparing the performance of the NOIA statistical model and usual functional model, we can clearly state the preference of the NOIA model on modeling GxE interactions for both quantitative traits and qualitative traits. 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.

CHAPTER 4

The NOIA Model Integrating Parent-of-Origin Effects (POEs) for Association Study of QTLs and Complex Diseases

In this chapter, we propose to implement the NOIA framework by incorporating POEs. The highly significant genetic markers identified via GWAS have explained only a proportion of the heritability of most human diseases [13, 61]. Genetic imprinting affects expression of genes and may explain some of the missing heritability. Both the orthogonal NOIA statistical model and usual functional model ignored the important genetic phenomenon, imprinting effects. We propose that more disease-associated genes could be detected by incorporating POEs with orthogonal models than by using traditional models, and that the NOIA POE model would fulfill the requirement of maintaining the power to detect the main allelic effect for complex diseases when multiple loci contribute to disease risk. The orthogonality of the statistical formulation of NOIA framework is important, especially when multiple loci are contributing to the outcome. We also proposed that using Kroneker product rule, our one-locus NOIA POE formulation can be easily extended to the general case of multiple loci (and environmental factors) to model general GxG/GxE interactions in the presence of imprinting effect, making NOIA a unified framework for detecting GxG/GxE interactions along with imprinting effects. Here we focus on one-locus association analysis for quantitative trait, implementing NOIA into a POE integrated framework by re-parameterization.

From the NOIA statistical model without POE (Stat-Usual, equation (5)) and the traditional functional model without POE (Func-Usual, equation (3)), we derived the formulas of several different quantitative trait association models, including a statistical POE (Stat-POE) model and a functional POE (Func-POE) model. Then, we evaluated the performance of the Stat-POE and Func-

POE models. We also compared the performance of the POE models (Stat-POE and Func-POE) with that of the models without POE incorporated (Stat-Usual and Func-Usual). These studies were all performed for both a simulated quantitative trait dataset and a qualitative trait dataset. We found that the incorporation of POE into the statistical model did not affect the estimation of the main allelic effect. Although our methods are currently developed and evaluated for single locus association study, they can be readily extended to gene-gene interaction or gene-environment interaction models.

In following sections, considering the orthogonal property of the NOIA statistical model and the non-orthogonal functional model mentioned in Chapter 1, we introduce our methodology extension of these two models by integrating POE detection. We also sought to evaluate the performance of these extended models in detecting both the overall genetic effect and POEs.

4.1 Methodology Development of the POE Models

Instead of three genotypic values in usual models without POE incorporation, the vector of genotypic values G has four components: G_{11}, G_{12}, $G_{21}$and G_{22}, in which the first allele represented by the first digit in the subscript is transmitted from the mother, and the second allele from the father. We used N_1 and N_2 to denote the number of maternal and paternal reference allele A_2, respectively. N1 and N2 are independent variables with binomial distributions, respectively. That is,

$$N_1 = \begin{cases} 0 & \text{if } G = G_{11} \text{ or } G_{12} \\ 1 & \text{if } G = G_{21} \text{ or } G_{22} \end{cases}, \tag{17a}$$

$$N_2 = \begin{cases} 0 & \text{if } G = G_{11} \text{ or } G_{21} \\ 1 & \text{if } G = G_{12} \text{ or } G_{22} \end{cases}. \tag{17b}$$

Similar to equation (1), the vector of the observed phenotypes G^* can be expressed as $G^* = Z_2 \cdot G$ and

$$\begin{pmatrix} G_1^* \\ G_2^* \\ \vdots \\ \vdots \\ G_n^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \cdot \begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix},$$

where the new n rows of matrix Z_2 represents the corresponding genotype for each individual.

First, we extended the usual functional (Func-Usual, equation (3)) and statistical model (Stat-Usual, equation (5)) by decomposing the additive effects into paternal and maternal additive effects (see Appendix 3.1). In the process of extension of the statistical model, our motivation was to incorporate POE detection while still maintaining its orthogonality. Next, these models were transformed into an equivalent but more comprehensive framework, which was straightforward for detecting the main allelic additive effect and POE simultaneously. The main allelic effects denote the overall additive effect on the outcome trait conferred by this allele, and the POE is defined as the imprinting effect of the allele with paternal origin over the same allele with maternal origin. The following subsections depict the developed models, and Appendix 3.1 shows the details about how we derived these new models.

4.1.1 The POE Functional (Func-POE) Model

First, we defined r_1 and r_2 as the main allelic additive effect and POE of the locus, or gene, respectively. Then, we extended the functional model (3) to the following,

$$G = R + \frac{N_1+N_2}{2} r_1 + \frac{N_2-N_1}{2} r_2 + \varepsilon d, \tag{18}$$

which could be also expressed as

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = S_{F_2} E_{F_2} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 1 \\ 1 & \frac{1}{2} & -\frac{1}{2} & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} R \\ r_1 \\ r_2 \\ d \end{pmatrix}. \tag{19}$$

The inverse is

$$E_{F_2} = \begin{pmatrix} R \\ r_1 \\ r_2 \\ d \end{pmatrix} = {S_{F_2}}^{-1} G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix}. \quad (20)$$

4.1.2 The POE Statistical (Stat-POE) Model

Here we let γ_1 and γ_2 denote the main allelic additive effect and POE of the locus, or gene, respectively. Similarly, we extended the orthogonal statistical model (5) to following.

$$G = \mu + \frac{N_1 + N_2 - (\bar{N}_1 + \bar{N}_2)}{2}\gamma_1 + \frac{N_2 - N_1 - (\bar{N}_2 - \bar{N}_1)}{2}\gamma_2 + \varepsilon\delta, \quad (21)$$

where $\bar{N}_1$and $\bar{N}_2$ denote the means of $N_1$and N_2, respectively, whereas $V_1$and V_2 denote the variance of $N_1$and N_2, respectively. In the models without POE incorporated (Func-Usual and Stat-Usual), p_{12} relates to the probability of an allele that has an allele A1 from either parent. In our new models, the meaning of p_{12} is different since it relates to the probability of an allele that has allele A1 from the mother and allele A2 from the father. p_{21} denotes the probability of an allele that has allele A2 from the mother and allele A1 from the father. Thus,

$$\bar{N}_1 = p_{21} + p_{22},$$

$$\bar{N}_2 = p_{12} + p_{22},$$

$$V_1 = (p_{21} + p_{22})(p_{11} + p_{12}) = \bar{N}_1(1 - \bar{N}_1),$$

$$V_2 = (p_{12} + p_{22})(p_{11} + p_{21}) = \bar{N}_2(1 - \bar{N}_2).$$

Therefore, according to equation (21), the vector of genotypic values can be expressed as

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = S_{S_2} E_{S_2} = \begin{pmatrix} 1 & \frac{-(\bar{N}_1 + \bar{N}_2)}{2} & \frac{-(\bar{N}_2 - \bar{N}_1)}{2} & \varepsilon_{11} \\ 1 & \frac{1-(\bar{N}_1 + \bar{N}_2)}{2} & \frac{1-(\bar{N}_2 - \bar{N}_1)}{2} & \varepsilon_{12} \\ 1 & \frac{1-(\bar{N}_1 + \bar{N}_2)}{2} & \frac{-1-(\bar{N}_2 - \bar{N}_1)}{2} & \varepsilon_{21} \\ 1 & 1 - \frac{\bar{N}_1 + \bar{N}_2}{2} & \frac{-(\bar{N}_2 - \bar{N}_1)}{2} & \varepsilon_{22} \end{pmatrix} \begin{pmatrix} \mu \\ \gamma_1 \\ \gamma_2 \\ \delta \end{pmatrix}, \quad (22)$$

where

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \end{pmatrix} = \begin{pmatrix} -2p_{12}p_{21}p_{22}/D \\ 2p_{11}p_{21}p_{22}/D \\ 2p_{11}p_{12}p_{22}/D \\ -2p_{11}p_{12}p_{21}/D \end{pmatrix}, \quad (23)$$

and

$$D = p_{12}p_{21}p_{22} + p_{11}p_{21}p_{22} + p_{11}p_{12}p_{22} + p_{11}p_{12}p_{21}. \tag{24}$$

The inverse is then $E_{S_2} = {S_{S_2}}^{-1}G$ which can be expressed as

$$\begin{pmatrix} \mu \\ \gamma_1 \\ \gamma_2 \\ \delta \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & p_{21} & p_{22} \\ p''_{11}+p'_{11} & p''_{12}+p'_{12} & p''_{21}+p'_{21} & p''_{22}+p'_{22} \\ p''_{11}-p'_{11} & p''_{12}-p'_{12} & p''_{21}-p'_{21} & p''_{22}-p'_{22} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix}, \tag{25}$$

If we define

$$\begin{cases} p'_{ij} = (-1)^{N_{1(ij)}-1} p_{ij}\left[(1-\bar{N}_2)^{1-N_{2(ij)}}(\bar{N}_2)^{N_{2(ij)}} - p_{ij}\right]{\bar{N}_2}^{1-N_{2(ij)}}(1-\bar{N}_2)^{N_{2(ij)}}/D \\ p''_{ij} = (-1)^{N_{2(ij)}-1} p_{ij}\left[(1-\bar{N}_1)^{1-N_{1(ij)}}(\bar{N}_1)^{N_{1(ij)}} - p_{ij}\right]{\bar{N}_1}^{1-N_{1(ij)}}(1-\bar{N}_2)^{N_{1(ij)}}/D \end{cases}, \tag{26}$$

where $N_{1(ij)}$ and $N_{2(ij)}$ denoted $N_1$and N_2 value of the genotype A_{ij}, respectively. From equations (25) and (26), each column of ${S_{S_2}}^{-1}$ is independent of the others therefore the parameters are orthogonal.

${S_{S_2}}^{-1}$ can also be expressed as

$$\begin{pmatrix} p_{11} & p_{12} & p_{21} & p_{22} \\ -\frac{p_{11}(p_{12}\bar{N}_1+p_{21}\bar{N}_2)}{D} & \frac{p_{12}(p_{11}\bar{N}_1-p_{22}(1-\bar{N}_2))}{D} & -\frac{p_{21}(p_{22}(1-\bar{N}_1)-p_{11}\bar{N}_2)}{D} & \frac{p_{22}(p_{21}(1-\bar{N}_1)+p_{12}(1-\bar{N}_2))}{D} \\ -\frac{p_{11}(p_{12}\bar{N}_1-p_{21}\bar{N}_2)}{D} & \frac{p_{12}(p_{11}\bar{N}_1+p_{22}(1-\bar{N}_2))}{D} & -\frac{p_{21}(p_{22}(1-\bar{N}_1)+p_{11}\bar{N}_2)}{D} & \frac{p_{22}(p_{21}(1-\bar{N}_1)-p_{12}(1-\bar{N}_2))}{D} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix}. \tag{27}$$

The POE functional model (Func-POE) and statistical model (Stat-POE) are related by

$$\begin{pmatrix} \mu \\ \gamma_1 \\ \gamma_2 \\ \delta \end{pmatrix} = \begin{pmatrix} 1 & \frac{N_1+N_2}{2} & \frac{N_2-N_1}{2} & p_{12}+p_{21} \\ 0 & 1 & 0 & p''_{12}+p''_{21}+p'_{12}+p'_{21} \\ 0 & 0 & 1 & p''_{12}+p''_{21}-(p'_{12}+p'_{21}) \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ r_1 \\ r_2 \\ d \end{pmatrix}, \tag{28}$$

where

$$p''_{12} + p''_{21} - (p'_{12} + p'_{21}) = 2p_{11}p_{22}(p_{12} - p_{21})/D,$$

which means $\gamma_2 = r_2$ for the case of equal frequency of the two types of heterozygote ($p_{12} = p_{21}$).

4.2 Results

4.2.1 Orthogonality of the Stat-POE Model

We have previously showed that the Stat-Usual model was orthogonal in the sense that the estimates of the four parameters were uncorrelated [22]. As stated in the previous section, from equations (25) and (26), the lack of correlation of the column values of ${S_{S_2}}^{-1}$ implies that the Stat-POE model is also orthogonal. The fact that the variance of G can be decomposed into two independent additive components and one dominance component also reflect the orthogonality of the statistical imprinting model. To prove the orthogonality, we performed the following decomposition. From Equation (21), we have

$$\mathrm{V_G} = \mathrm{Var}\left[\frac{\mathrm{N_1} + \mathrm{N_2} - (\bar{\mathrm{N}}_1 + \bar{\mathrm{N}}_2)}{2}\gamma_1\right] + \mathrm{Var}\left[\frac{\mathrm{N_2} - \mathrm{N_1} - (\bar{\mathrm{N}}_2 - \bar{\mathrm{N}}_1)}{2}\gamma_2\right] + \mathrm{Var}(\varepsilon\delta)$$

$$+2\mathrm{Cov}\left[\frac{\mathrm{N_1}+\mathrm{N_2}-(\bar{\mathrm{N}}_1+\bar{\mathrm{N}}_2)}{2}\gamma_1, \frac{\mathrm{N_2}-\mathrm{N_1}-(\bar{\mathrm{N}}_2-\bar{\mathrm{N}}_1)}{2}\gamma_2\right]. \quad (29)$$

Note that

$$\mathrm{Cov}\left[\frac{\mathrm{N_1}+\mathrm{N_2}-(\bar{\mathrm{N}}_1+\bar{\mathrm{N}}_2)}{2}\gamma_1, \varepsilon\delta\right] = \gamma_1\delta\mathrm{Cov}\left(\frac{\mathrm{N_1}+\mathrm{N_2}}{2}, \varepsilon\right) = 0,$$

and similarly,

$$\mathrm{Cov}\left[\frac{\mathrm{N_2}-\mathrm{N_1}-(\bar{\mathrm{N}}_2-\bar{\mathrm{N}}_1)}{2}\gamma_2, \varepsilon\delta\right] = 0.$$

Also, $\mathrm{Var}(\varepsilon\delta) = \delta^2\mathrm{var}(\varepsilon) = 4\mathrm{p_{11}p_{12}p_{21}p_{22}}\delta^2/\mathrm{D}$. Therefore, we can express the additive and dominant variance components as

$$\mathrm{V_\gamma} = {\gamma_1}^2\mathrm{Var}\left(\frac{\mathrm{N_1}+\mathrm{N_2}}{2}\right) + {\gamma_2}^2\mathrm{Var}\left(\frac{\mathrm{N_2}-\mathrm{N_1}}{2}\right) + 2\gamma_1\gamma_2\mathrm{Cov}\left(\frac{\mathrm{N_1}+\mathrm{N_2}}{2}, \frac{\mathrm{N_2}-\mathrm{N_1}}{2}\right), \quad (30)$$

$$\mathrm{V_\delta} = 4\mathrm{p_{11}p_{12}p_{21}p_{22}}\delta^2/\mathrm{D}. \quad (31)$$

To show that the additive variance, $\mathrm{V_\gamma}$, can be decomposed to be two parts that are dependent on only two additive effects ($\gamma_1$and γ_2) respectively, $\mathrm{Cov}\left(\frac{\mathrm{N_1}+\mathrm{N_2}}{2}, \frac{\mathrm{N_2}-\mathrm{N_1}}{2}\right) = 0$ needs to be satisfied. And, as we know

$$\mathrm{Cov}\left(\frac{\mathrm{N_1}+\mathrm{N_2}}{2}, \frac{\mathrm{N_2}-\mathrm{N_1}}{2}\right) = \frac{1}{4}\,\mathrm{E}\big((\mathrm{N_1}+\mathrm{N_2})(\mathrm{N_2}-\mathrm{N_1})\big) - \frac{1}{4}\mathrm{E}(\mathrm{N_1}+\mathrm{N_2})\mathrm{E}(\mathrm{N_2}-\mathrm{N_1})$$

$$= \frac{1}{4}V_2 - \frac{1}{4}V_1 = \frac{1}{4}(p_{12} + p_{22})(p_{11} + p_{21}) - \frac{1}{4}(p_{21} + p_{22})(p_{11} + p_{12})$$

$$= \tfrac{1}{4}(p_{11} - p_{22})(p_{12} - p_{21}), \quad (32)$$

which is indeed equal to 0 under the condition that $p_{12} = p_{21}$ or $p_{11} = p_{22}$. In this way, we divided the additive variance component into two independent parts as follows:

$$V_1 = (p_{21} + p_{22})(p_{11} + p_{12}) = \bar{N}_1(1 - \bar{N}_1),$$

$$V_2 = (p_{12} + p_{22})(p_{11} + p_{21}) = \bar{N}_2(1 - \bar{N}_2),$$

$$V_{\gamma_1} = \frac{{\gamma_1}^2}{4}\mathrm{Var}(N_1 + N_2) = \frac{{\gamma_1}^2}{4}\big((p_{21} + p_{22})(p_{11} + p_{12}) + (p_{12} + p_{22})(p_{11} + p_{21})\big), \quad (33)$$

$$V_{\gamma_2} = \frac{{\gamma_2}^2}{4}\mathrm{Var}(N_2 - N_1) = \frac{{\gamma_2}^2}{4}(p_{11} - p_{22})(p_{12} - p_{21}). \quad (34)$$

And $V_G = V_{\gamma_1} + V_{\gamma_2} + V_\delta$.

The two additive variance components V_{γ_1} and V_{γ_2} are related only to the additive effects parameters $\gamma_1$and γ_2, respectively, one due to overall genetic effect and the other due to POEs. The dominance variance component V_δ is only related to the dominance effect δ. Of the fact that the variance components can be decomposed into two independent additive components and one dominant component suggests the notion that the transformed POE statistical model is orthogonal. We also proved that the Stat-POE model is orthogonal before transformation (Appendix 3.2). On the other hand, by checking whether $X^T \cdot X$ is a diagonal matrix, we showed that the transformed Stat-POE model was orthogonal (Appendix 3.3). However, for the transformed Func-POE model, the variance components could not be decomposed into three independent parts, indicating that the Func-POE model is not completely orthogonal (Appendix 3.4).

4.2.2 Simulation Methods

We performed simulation analysis for both quantitative traits and qualitative traits (case-control) using an approach similar to that used in [22], and the simulated data were analyzed using the four aforementioned models: Stat-POE, Func-POE, Stat-Usual and Func-Usual. We already derived the test statistics of the Stat-Usual and Func-Usual in Appendix 2.2. Similarly, the Wald test statistic for

the Stat-POE model is

$$\frac{1}{\sqrt{n\sigma^2 D}}\begin{pmatrix} \sqrt{D} & \sqrt{D} & \sqrt{D} & \sqrt{D} \\ \frac{-(\bar{N}_1+\bar{N}_2)\sqrt{p_{12}+p_{21}}}{2} & \frac{(1-(\bar{N}_1+\bar{N}_2))\sqrt{p_{12}+p_{21}}}{2} & \frac{(1-(\bar{N}_1+\bar{N}_2))\sqrt{p_{12}+p_{21}}}{2} & \frac{(2-(\bar{N}_1+\bar{N}_2))\sqrt{p_{12}+p_{21}}}{2} \\ 0 & \frac{\sqrt{-(\bar{N}_1+\bar{N}_2)^2+(\bar{N}_1+\bar{N}_2)+2p_{22}}}{2} & -\frac{\sqrt{-(\bar{N}_1+\bar{N}_2)^2+(\bar{N}_1+\bar{N}_2)+2p_{22}}}{2} & 0 \\ -\sqrt{\frac{p_{12}p_{21}p_{22}}{p_{11}}} & \sqrt{\frac{p_{11}p_{21}p_{22}}{p_{12}}} & \sqrt{\frac{p_{11}p_{12}p_{22}}{p_{21}}} & -\sqrt{\frac{p_{11}p_{12}p_{21}}{p_{22}}} \end{pmatrix}.$$

$Z_2{}' \cdot y$ with the second to fourth rows for the main additive effect, POE and dominant effect testing, respectively. Since the functional model is not orthogonal (Appendix S3.4), it is difficult to obtain an expression for the test statistic of the Func-POE model.

4.2.2.1 Simulation of Data with a Quantitative Trait

To simulate samples of independent individuals with a quantitative trait controlled by a diallelic locus, we assumed that the gene is under HWE. The case that a gene is not under HWE will be investigated in our future work. For a given value of the minor allelic frequency (p) in the population, genotype 11, 12, 21, 22 were assigned to an individual with probabilities $(1-p)^2$, $p(1-p)$, $p(1-p)$ and p^2 respectively. We assumed the genotype frequencies of the two types of heterozygotes were the same in the population. We also assumed the phenotype was influenced by a main allelic additive effect, a POE, and a dominant effect. From a prespecified vector of parameters (${E_F}^T = [R, a_1, a_2, d\,]$), we assigned each individual a genotypic value according to his/her assigned genotypes. Then, by randomly generating a value from a normal distributions with prespecified mean and variance (0 and σ_e^2), we generated an observed phenotype/trait by adding this residual to the previously assigned phenotypic value. We used data from 2000 individuals as a replicate and simulated 1000 replicates for each genetic model.

In the simulation study of a quantitative trait, three scenarios were simulated with different levels of POE (Table 4.1). The minor allele frequency p was set to 0.28, and the residual variance σ_e^2 was 144.0. The true values of the four parameters in these three scenarios are shown in Table 4.1. For the sample size with 2000 individuals, the computation speeds of the four models on quantitative traits

analysis are: 22 seconds for the Stat-POE model, the Stat-Usual model and Func-Usual model, respectively; 24 seconds for the Func-POE model.

4.2.2.2 Simulation of Data with a Qualitative Trait

Ma et al. [22] previously derived the formulation of the statistical model without POE incorporated in quantitative traits and demonstrated that a similar statistical model could also be defined for a qualitative trait by handling the genetic effects as the logit function of the disease. As mentioned in previous sections, the orthogonality of that model does not exist for the qualitative trait under the alternate hypothesis when there is a genetic effect, but is valid under the null hypothesis of no effect. It is not difficult to show that it is still the case for our Stat-POE model. Here we performed simulations to evaluate the performance of our POE-related models in a case-control study design.

Briefly, we used the logistic model and Bayes' theorem to set the genotype of each individual according to the prespecified genetic effect terms, $E_F{}^T = [R, a_1, a_2, d]$. The disease penetrance for each genotype was determined by

$$\Pr(d = 1|ij) = \frac{1}{1+\exp(-G_{ij})},$$

where d denotes the disease status with value 1 for patients and 0 for control. G_{ij} was the genotypic value when the genotype was ij with $ij = 11,12,21 \text{ or } 22$. Then the distributions of the four genotypes in the cases was determined by

$$\Pr(ij|d = 1) = \frac{P_{ij}/\left(1+\exp(-G_{ij})\right)}{\sum_{kl} P_{kl}/\left(1+\exp(-G_{kl})\right)}.$$

As in the simulation study for a quantitative trait, P_{ij} is the genotype frequency of 11, 12, 21 and 22 in the population, determined by $(1-p)^2$, $p(1-p)$, $p(1-p)$ and p^2, respectively. For simulating controls in the population, we used a similar distribution as follows

$$\Pr(ij|d = 0) = \frac{P_{ij}/\left(1+\exp(G_{ij})\right)}{\sum_{kl} P_{kl}/\left(1+\exp(G_{kl})\right)}.$$

For each replicate, 1000 cases and 1000 controls were generated, and a total of 1000 replicates were simulated. The minor allele frequency p was set to 0.28. Two scenarios were simulated with different levels of POE (Table 4. 1). The simulating values of the parameters in the two different scenarios are shown in Table 4.1.

Table 4.1 Simulation true values of genetic effects for quantitative and qualitative traits datasets. R denotes intercept; r_1 and r_2 denote overall genetic effect and POE effect, respectively; and d denotes dominant effect. Three scenarios with strong, medium and weak POEs were simulated for quantitative traits; two scenarios with strong and weak POEs were simulated for qualitative traits.

	R	r_1	r_2	d
Quantitative trait				
Scenario 1	90.0	3.0	-3.0	1.2
Scenario 2	90.0	3.0	-2.0	1.2
Scenario 3	90.0	3.0	-1.0	1.2
Qualitative trait				
Scenario 1	100.0	2.0	-2.0	0.5
Scenario 2	100.0	2.0	-0.6	0.5

To determine whether the setting of the MAF value influence the performance of the models, we also simulated two additional scenarios with different MAF values (0.03 and 0.48) for both quantitative traits and qualitative traits.

4.2.3 Results for simulated data

First we performed a simulation study for a quantitative trait in three scenarios with strong, moderate, and weak imprinting effect while the main allelic additive effect remained the same (Table 4.1). The true values of the four parameters in these three scenarios are shown in Table 4.1. The density distributions of all four effects after analyzing 1000 replicates in scenario 1 with strong imprinting effect is shown in Figure 4.1. The estimates of all four parameters were accurate for both the Stat-POE and Func-POE models. Compared with the Func-POE model, the Stat-POE model had

smaller variance in most cases for detecting the intercept and main allelic additive effect terms. The estimates for the POE term and dominant effect term were the same between the Func-POE and Stat-POE models. Similar patterns could be detected for the other two scenarios (data not shown).

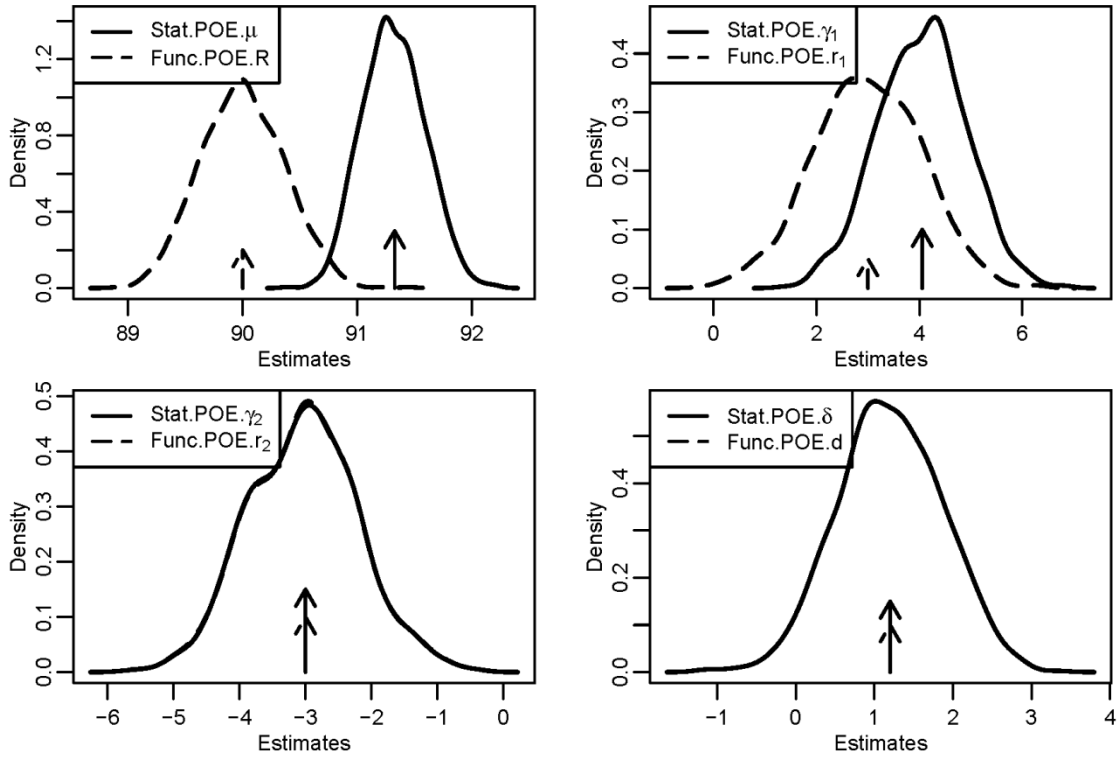

Figure 4.1 Density distributions of the estimates of the parameters from a simulated data analysis with a quantitative trait influenced by a genetic factor and by strong POE (Scenario 1). The pre-specified minor allele frequency was 0.28. The values of the four parameters were $E_F{}^T = [90.0, 3.0, -3.0, 1.2]$ and $E_S{}^T = [91.3, 4.05, -3, 1.2]$ for the functional POE (Func-POE) model and the statistical POE (Stat-POE) model, respectively. The solid arrows denote the true simulated values of the parameters for Stat-POE model and the dashed arrows denote those for the Func-POE model.

To evaluate the performance of these models in detecting a main allelic additive effect and POE, we calculated the statistical power of four models under different critical values of P values obtained using a Wald test (Fig. 4.2). Figure 4.2 shows the power for detecting the main allelic additive effect for scenario 1 with strong POE. The power of both statistical models (Stat-POE and Stat-Usual) for detecting additive effects was greater than that of both functional models (Func-POE and Func-Usual). The power of detecting additive effect was the same for the Stat-POE and Stat-Usual models.

It was also the same for the Func-POE and Func-Usual models. In the other two scenarios in which medium or weak POE was simulated, identical results(data not shown) were obtained for the main genetic effect term as shown in Figure 4.2a, since the main allelic additive effect was set to the same value, 3.0 (Table 4.1). These results indicated that the power for detecting the main allelic effect did not change even if a POE parameter was integrated into the analysis model. The performance of these four models for detecting dominant effects was the same in three scenarios (data not shown), which was consistent with the formulations.

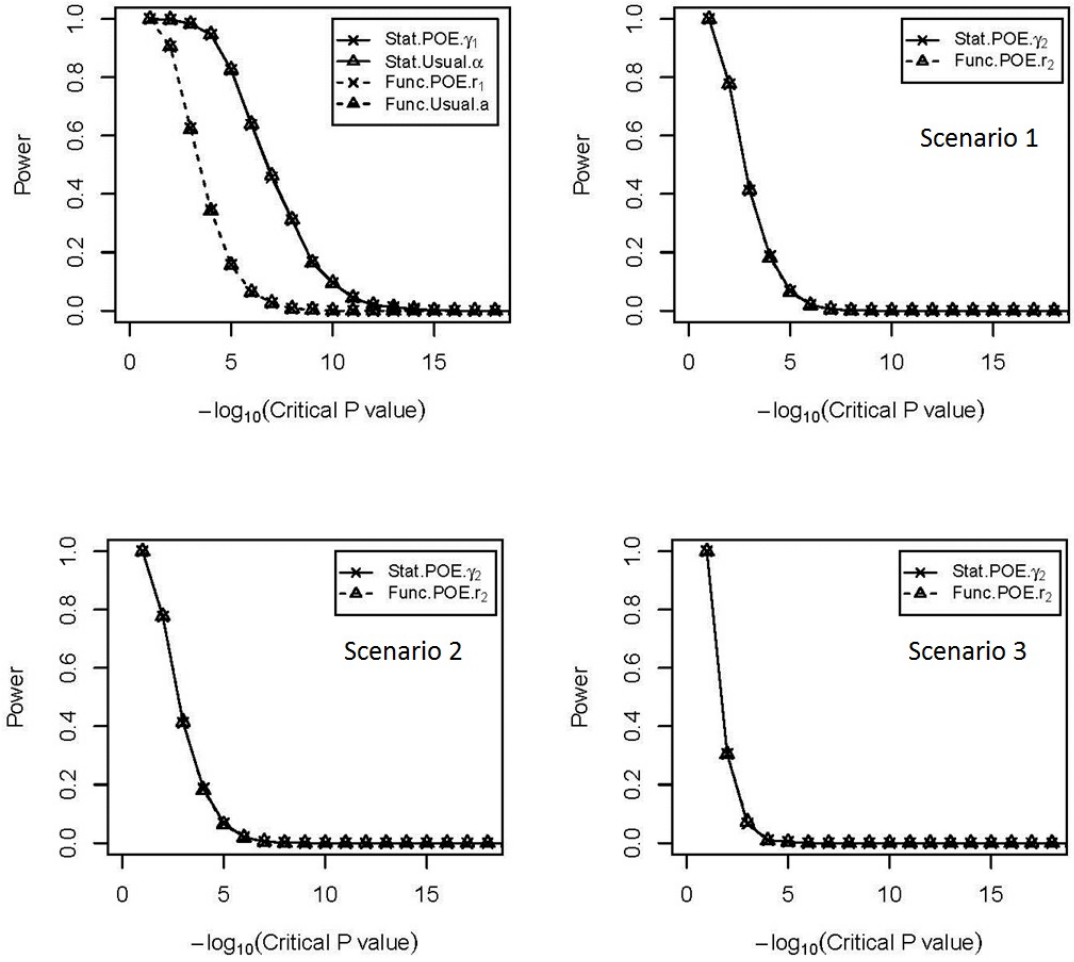

Figure 4.2 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation data shown in Table 4.1. (a) Power for detecting the main allelic additive effect in scenario 1 when strong POE exists. Power for detecting POE of the Stat-POE and Func-POE models was compared for scenario 1 (b), scenario 2 (c), and scenario 3 (d).

Figure 4.2b-d shows the power of the Stat-POE and Func-POE models for detecting the POE in three scenarios. The performance of the two POE models remained the same for the three scenarios. This is because in our simulation, the genotype frequency values for the two types of heterozygotes were set at the same value which is valid when HWE hold true. This results in $p''_{12} + p''_{21} - (p'_{12} + p'_{21}) = \frac{2p_{11}p_{22}(p_{12}-p_{21})}{D} = 0$. Therefore, according to equation (28), the POE repressor in the Stat-POE model was equivalent to that in the Func-POE model. When the assumption that the genotype frequencies for the two heterozygotes are the same is violated, it will result in different performance of the Stat-POE and Func-POE models for detecting POE. Additionally, the overall power decreases when the POE decreases (Fig. 4.2b-d).

To evaluate whether the MAF influences the estimation of the genetic effects by these models, we also performed analyses for quantitative traits when the MAF was 0.03 and 0.48, respectively (Fig. S4.1-S4.2). Figure S1 shows that when strong POE existed, the Stat-POE model still presented extremely greater power than the Func-POE model in detecting main additive effect for rare variants (MAF=0.03). Figure S4.2 shows that when strong POE existed, the Stat-POE model presented slightly greater power than the Func-POE model in detecting main additive effect for variants with MAF as 0.48.

Similarly, we also performed analyses for simulated case-control data. The simulating values for each of the two scenarios are shown in Table 4.1. Figure 4.3 shows the density distributions of all four effects after analyzing 1000 replicates in scenario 1. Similar patterns were detected for the distributions of the four parameters as in the quantitative trait. And the estimates were all accurate for both the Stat-POE and Func-POE models, except for the intercept term. The differential estimation of the intercept term arose from non-random sampling in our simulation. The variance of the main allelic additive effect was still smaller via analysis using the Stat-POE model than via analysis using the Func-POE model. And the estimate distributions are very close or the same for these two models for detecting POE and dominant effect.

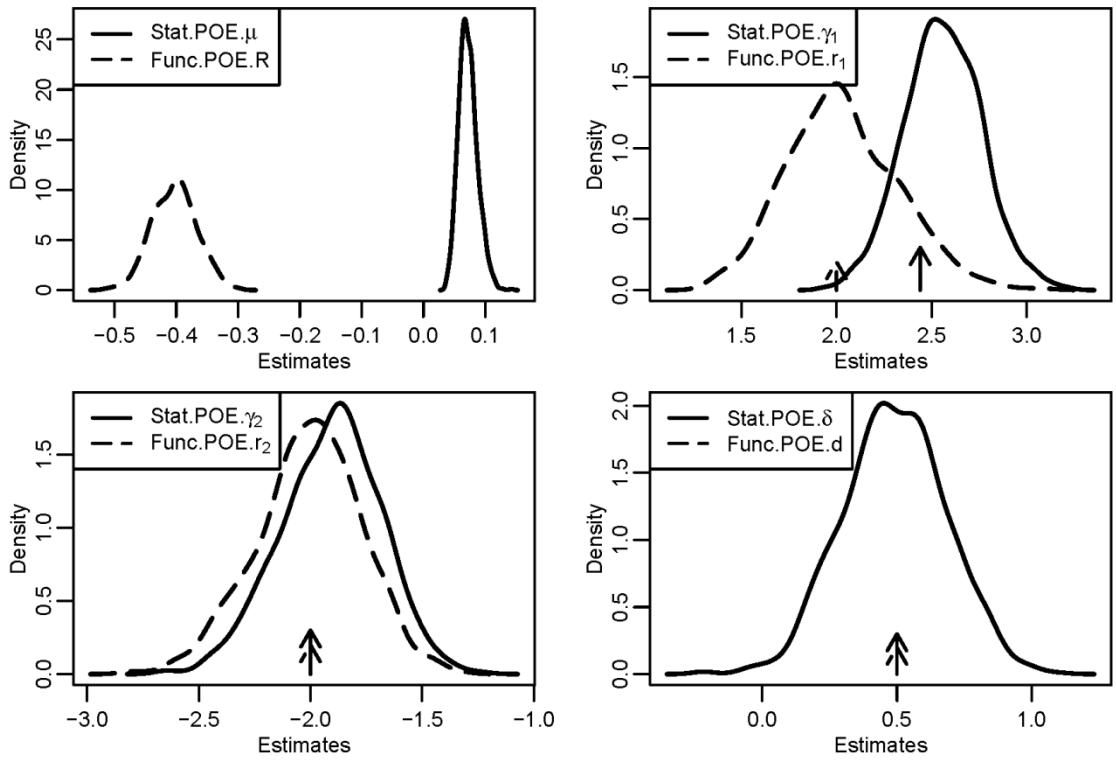

Figure 4.3 Density distributions of the estimates of all four parameters from a simulated data analysis with a qualitative trait influenced by a genetic factor and by strong POE. The pre-specified minor allele frequency was 0.28; the true values of the four parameters were $E_F{}^T = [100.0, 2.0, -2.0, 0.5]$ and $E_S{}^T = [100.0, 2.44, -2, 0.5]$ for the Func-POE and the Stat-POE models, respectively. The solid arrows denote the true simulated values of the parameters for Stat-POE model and the dashed arrows denote those for the Func-POE model.

Figure 4.4 shows the power of the four models for detecting the main allelic additive effect, POE and dominant effect when the trait was affected by relatively strong POE for case-control data. The performance of the Stat-POE model was slightly better than that of the Stat-Usual model, and the performance of both was better than that of the functional models, Func-POE and Func-Usual (Fig. 4.4a). The Stat-POE and Func-POE models had the same power for detecting POE (Fig. 4.4b, 4.4c). Interestingly, both POE models (Stat-POE and Func-POE) had higher power for detecting dominance effect than the usual models, Stat-Usual and Func-Usual (Fig. 4.4c).

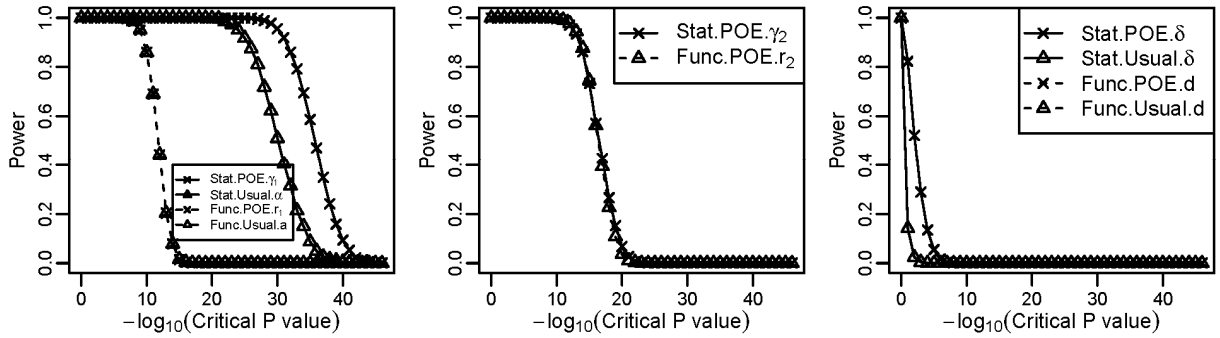

Figure 4.4 Power under different critical values of the P values obtained using the Wald test for the case-control simulation data influenced by a genetic factor with strong POE (scenario 1). The minor allele frequency was 0.28.

Another simulation was performed with a moderate POE for case-control data (Table 4.1, scenario 2; Fig. 4.5). Interestingly, the performance of the Stat-POE model was not much better than that of the Stat-Usual model (Fig. 4.5a) for detecting the main allelic additive effect (Fig. 4.4a). For detecting the main allelic additive effects, the statistical models (Stat-POE and Stat-Usual) had much higher power than the functional models, Func-POE and Func-Usual. The statistical models and functional models had the same or very close power with and without the incorporation of POE, respectively. The Stat-POE and Func-POE models had the same or very close power for detecting POE and dominant effect (Fig. 4.5b, 4.5c).

Simulations were also performed when MAF was set as 0.03 and 0.48 for case-control traits, respectively (Fig. S4.3-S4.4). For rare variants (MAF=0.03), the Stat-POE model presented extremely greater power than the Func-POE model in detecting main additive effect, although slightly greater power was observed for Func-POE model in detecting the POE (Fig. S4.3). For variants with MAF=0.48, the Stat-POE model presents extremely greater power than the Func-POE model in detecting main additive effect and dominant effect (Fig. S4.4). The power of the Stat-POE model was even higher than that of the Stat-Usual model in detecting the main additive effect (Fig. S4.4a).

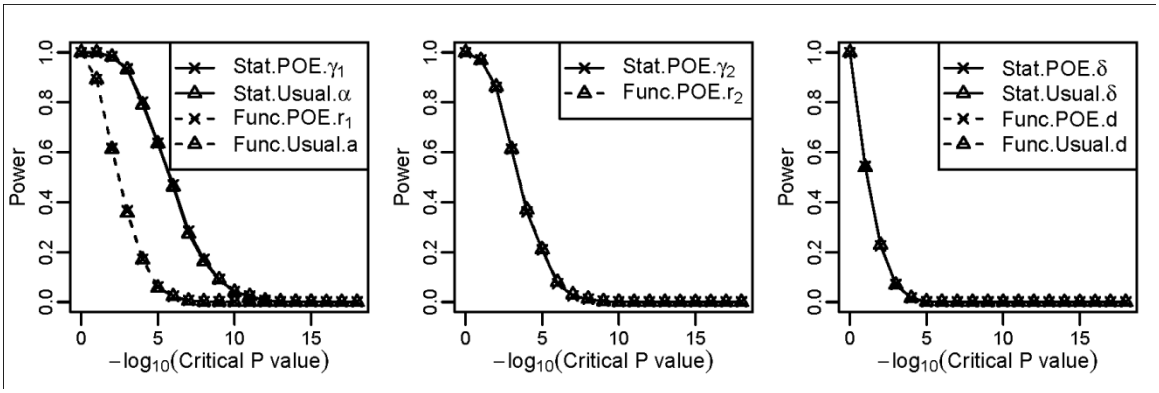

Figure 4.5 Power under different critical values of the P values obtained using the Wald test for the case-control simulation data influence by a genetic factor with moderate POE (scenario 2). The minor allele frequency was 0.28.

Type I error was also inspected for both the quantitative trait and the qualitative trait by simulating a null scenario where there was no main genetic effect or POE. We estimated the type I error for the main additive effect, POE and dominant effect for both quantitative traits and case-control traits when the MAF was set as 0.03, 0.28 or 0.48 (Table 4.2). The false positive rate for detecting the additive effect was almost the same for the statistical and functional POE models in most scenarios we simulated (around 0.05 or less for the 1000 replicates). The false positive rate for detecting the additive effect was smaller estimated from the Func-POE model than that from the Stat-POE model, when MAF was set as 0.03 for case-control traits. For detecting POE, these two models usually had very close false positive rates for both quantitative and case-control traits.

Table 4.2 Type I error for simulation of quantitative and case-control traits data sets. False positive rates for the genetic effects estimated from the Stat-POE, Func-POE, Stat-Usual and Func-Usual models under different minor allele frequency settings. Add= overall genetic additive effect; Dom=dominant effect; MAF=minor allele frequency.

	MAF=0.03			**MAF=0.28**			**MAF=0.48**		
Models/MAF	**Add**	**POE**	**Dom**	**Add**	**POE**	**Dom**	**Add**	**POE**	**Dom**
Quantitative trait									
Stat-POE	0.047	0.037	0.059	0.055	0.038	0.048	0.053	0.043	0.043
Func-POE	0.055	0.036	0.059	0.056	0.037	0.048	0.052	0.042	0.043
Stat-Usual	0.048		0.06	0.056		0.048	0.053		0.044
Func-Usual	0.048		0.06	0.056		0.048	0.053		0.044
Case-control trait									
Stat-POE	0.044	0.062	0.017	0.05	0.045	0.046	0.047	0.049	0.039
Func-POE	0.01	0.063	0.017	0.049	0.047	0.046	0.049	0.048	0.039
Stat-Usual	0.045		0.017	0.047		0.047	0.047		0.038
Func-Usual	0.045		0.017	0.047		0.047	0.047		0.038

Table 4.3 Summary of the power of the Stat-POE and Func-POE models in different simulation scenarios for both quantitative traits and case-control traits. Add: overall genetic additive effect; Dom=dominant effect. Threshold of the P value was 0.001.

		MAF=0.03		MAF=0.28		MAF=0.48	
		Strong POE	**Weak POE**	**Strong POE**	**Weak POE**	**Strong POE**	**Weak POE**
Quantitative traits							
Add	**Stat-POE**	0.98	0.98	0.93	0.94	0.79	0.78
	Func-POE	0.01	0.01	0.36	0.35	0.75	0.75
POE	**Stat-POE**	0.4	0.1	0.61	0.1	0.77	0.03
	Func-POE	0.4	0.1	0.61	0.1	0.77	0.02
Dom	**Stat-POE**	0.005	0.007	0.07	0.06	0.12	0.15
	Func-POE	0.005	0.007	0.07	0.06	0.12	0.15
Case-Control traits							
Add	**Stat-POE**	0.73	0.73	1	1	1	1
	Func-POE	0.001	0.001	1	1	1	1
POE	**Stat-POE**	0.8	0.33	1	0.33	1	0.61
	Func-POE	1	0.41	1	0.38	1	0.73
Dom	**Stat-POE**	0.04	0.05	0.29	0.35	0.8	0.9
	Func-POE	0.04	0.05	0.29	0.35	0.8	0.9

4.3 Discussion

In this chapter, we extended the NOIA framework, which was initially developed for epistasis quantitative traits analyses by incorporating POE for genetic association analysis. Herein, we propose a unified framework for one-locus association study that allows for both main allelic additive effect and POE estimation via linear regression. By simulation study, we illustrated the statistical properties of this implemented framework on one-locus association study. We summarized the detailed comparison of the performance of the Stat-POE and Func-POE models in Table 4.3. In most scenarios we simulated, the Stat-POE model had greater power than the Func-POE model in detecting the main additive effect. For testing imprinting effect, the Stat-POE model had same power as the Func-POE model for quantitative traits whereas the former presented slightly worse power than the latter for qualitative traits.

We used genetic variance decomposition to show that the Stat-POE model was orthogonal when either HWE or equal minor and major allele frequencies is satisfied for quantitative traits (equations 29-33). Thus, even when the POE was absent, estimating of the main allelic additive effect was not affected when a new parameter was added in the analytic model. This was not true for the Func-POE model, as demonstrated by simulation results for quantitative traits (Fig. 4.2a). Although the Func-POE model was not orthogonal (Appendix 3.4), the same performance of the Func-POE and Func-Usual models for detecting the main allelic additive effect in Figure 4.2a could still held true as the power between these two models was only slightly different. Another reason might be because the term $\frac{r_1 d}{2}\mathrm{Cov}(N_1 + N_2, \varepsilon)$ in equation (D3) (Appendix 3.4) is rather small. The Stat-POE and Func-POE models we proposed could also be applied to qualitative traits via logistic regression although the property of orthogonality would no longer exist under the alternative model [22]. When orthogonality exists under the null, the subsequent tests have appropriate type I error rates, but the failure of orthogonality under the alternate model can lead to improper estimates of heritability, although the estimators may be less biased than those that are obtained from the functional models.

Using simulations, we demonstrated that the statistical models, including the Stat-POE and Stat-Usual models, had better performance for detecting the main allelic additive effect than the functional models, Func-POE model and Func-Usual for both quantitative traits and qualitative traits. And the same power of these two POE models on detecting the POE arose when $p_{12} = p_{21}$ was true. Stat-POE model had better performance on detecting the main allelic additive effect than the Stat-Usual model for qualitative traits when strong POE exists. The power was the same for detecting the main allelic effect even if a POE parameter was integrated into the analysis model and supported orthogonality of the Stat-POE model (Fig. 4.2a, Fig. S4.1 and S4.2). The performance of our framework was not exactly the same in quantitative and qualitative trait simulation studies. The simulation study for both quantitative and qualitative traits showed that the estimates of all four parameters were accurate for both the Func-POE and Stat-POE models. However, the performance of these two models for detecting the main allelic effect and dominance effect presented a different pattern in qualitative traits (Fig. 4.4 and 4.5). In qualitative traits, for detecting the main allelic effect, the statistical (Stat-POE and Stat-Usual) models, still had greater power than did the functional (Stat-Usual and Func-Usual) models in most cases, regardless of the strength of the POE, which is consistent with the findings of the quantitative traits simulation study (Fig. 4.2). However, the power of the Stat-POE and Stat-Usual models was not usually the same for the qualitative trait simulation in different scenarios (Fig. 4.4a, 4.5a), which varied from the findings of the quantitative traits simulation (Fig. 4.2). The performance of the four models on detecting the dominance effect is also different in the simulation analysis for a qualitative trait (Fig. 4.4c; Fig. S4.3 and S4.4), which shows that the POE models (including the Stat-POE and Func-POE) usually have greater power than the usual models. This difference in performance arises because the test statistics used for logistic and linear regression differ.

We can also illustrate the reason why the proposed model could detect more disease-associated genes than the traditional models in model setting as follows. First, the orthogonal (Stat-Usual) model proposed by Alvarez-Castro et al. has an advantage of orthogonalizing the estimating of the

additive effect and dominant effect but the usual model (Func-Usual) does not. We constructed the test statistic of the Stat-Usual and Func-Usual models for quantitative traits with dominance component and without dominance effect (Section 1.4 and Appendix 2.2). The test statistic for estimating the additive effect did not change if the dominance component was removed for the Stat-Usual model. However, the test statistic was not consistent if the dominance component was removed from the Func-Usual model. Thus, the Stat-Usual model is preferred than the Func-Usual model in association studies when dominance component is incorporated. Second, we also compared the test statistic of the Stat-Usual and our newly developed Stat-POE models. We found that the test statistic of the main additive effect was the same for the two models, which was consistent with the simulation studies. And even in simulation studies for a case-control trait, we found that the Stat-POE had greater power for detecting the main additive effect than the usual orthogonal model (Stat-Usual). Comparing the test statistic of the Func-POE and Fun-Usual models, we found that the estimation of the main genetic effect was not consistent, and the power decreased when POE testing was included. Therefore, Stat-POE model could detect more significant additive effect signals than the Func-POE model.

Several recent studies have incorporated POEs in association analyses for quantitative traits. Genome-wide rapid association using mixed model and regression (GRAMMAR) and its extension are a recently developed approach that is based on a measured genotype approach and has been shown to have greater power than the transmission disequilibrium test (TDT)-based tests [62]. A maximum likelihood test was also developed for detecting POEs using haplotypes [63]. Ainsworth et al. also described an implementation of a family-based multinomial modeling approach that allows for imprinting detection [52]. This method used family data, case-mother duos or case-parents trios, to look for departures in observed genotypes distributions from expected distributions among affected offspring, given the genotypes of their parents. The mechanism of this approach is still more related to the TDT test. However, to our knowledge, our approach is the only one that has the advantage of orthogonality on the effects estimation for association studies of detecting POE. NOIA

was previously proposed and formularized for gene-gene interaction analysis models of quantitative traits and was further implemented and extended by Ma et al. [22] to r reduced genetic models and estimating effects from both genetic and binary environmental exposure. However, neither of these models had the potential to detect POE. We already found that when POE was not incorporated, the power of the statistical model (Stat-Usual) was greater than that of the functional model (Func-Usual) in most cases. This finding held true in our study for detecting main effects even when POE was integrated. Our study exemplifies another significant implementation of NOIA that adopts the orthogonal property of the statistical model if the family data are available or if phasing is plausible for obtaining the parental transmitting status of the candidate disease-associated locus. Because alleles of different parental origins can exert different effects, the effect contributing to the disease outcome may be masked in usual models that can detect only the main allelic additive effect. The methodology and simulation study used for our extension of NOIA yielded a plausible means of detecting more genes that contribute to complex diseases or quantitative traits that were not detected in routine GWASs.

Although our extension significantly contributes to disease gene mapping, pedigree data are needed for our framework to be used to estimate transmitting information of each heterozygotes or homozygotes locus. Obtaining the transmitting status of one locus is more difficult for non-informative pedigrees than for informative pedigrees which need to be determined by nearby linked loci or haplotype phasing. This limits the application of our model in GWAS. However, with the development of genotyping technique, it will be possible to obtain pedigree data with sufficient sample size in the near future. Another direction of our next step will be generalizing our formulation to the case of non-deterministic genotypes, e.g. with probabilistic parental information or missing data, to incorporate the phasing uncertainty of genotypes.

The motivation of our implemented framework was based on the orthogonality property of NOIA which allows model selection and variance component analysis more straightforward. A next step is to extend the formulation proposed here to multi-locus and/or environment factor case, including

gene-gene interaction and gene-environment interaction analyses while POE is integrated. Conceptually, this generalization should fairly straightforward by applying the Kronecker product rule as in [21], if we assume linkage equilibrium between loci and no association between a genetic locus and an environment factor. However, it would probably be challenging to deal with and properly interpret a large number of interaction terms. The extension, nevertheless, would be attractive as imprinting effects of one locus may indeed have complex interaction with main effects of other loci. We are currently working along this direction.

CONCLUSIONS

To investigate the properties and application of the NOIA framework on association studies, we implemented it in three directions. First, we generalized the NOIA coding technique to the full model and reduced models (including additive, dominant and recessive) allowing for GxG interaction testing. Through extensive simulation studies, we demonstrated greater statistical power of the NOIA model comparing with the usual approach. The newly developed methods were applied to melanoma datasets. Through real data analyses, we confirmed that NOIA model had obviously greater power on detecting the main genetic effects and interaction effects compared to the usual approach. We also validated several previously identified causal variants of melanoma and found some novel gene-gene interactions. Further experiments need to be carried out to verify these interactions.

To explore the extension of the NOIA framework for detecting GxE interactions, we developed a novel statistical approach that allows us to model effects from a genetic factor and binary environmental exposure that are jointly influencing disease risk. Through extensive simulation studies, we demonstrated greater statistical power of the NOIA GxE model on detecting the main effects and interaction effects comparing with the usual approach. To evaluate the performance of the newly developed method, we applied it on lung cancer datasets. Our results of identifying the causal variants were consistent with previous studies. Moreover, we also found some novel gene-environment interactions for lung cancer risk.

We also developed a statistical approach for modeling genetic effects due to imprinting effects in the orthogonalized framework. The POEs are usually ignored in traditional approaches (for example, GWAS), which were designed to only detect the overall genetic effect, resulting in weaker tests and

lower estimate of heritability of human complex diseases and traits. We believe that incorporations of POEs detecting into association studies could solve this problem to some extent, and estimating the effects and testing for imprinting is important for further delineating the complex genetic architecture for human diseases and traits. Extensive simulation studies demonstrated the statistical performance of the new methodology that we have developed. We found that Stat-POE model had better performance on detecting the main allelic additive effect than the Func-POE model for both quantitative traits and qualitative traits. We also found that Stat-POE model had better performance on detecting the main allelic additive effect than the Stat-Usual model for case-control traits when strong POE exists.

We believe that the new methods that we have developed will be useful in further understanding the impact of gene-gene interactions, gene-environment interactions and imprinting effects on human complex traits and diseases. Orthogonal methods are useful for improving estimation of effects particularly when multiple loci or environmental factors are jointly contributing to the outcome and when GxG/GxE interactions are investigated, especially when imprinting effects are incorporated into the modeling. Through our implementation, the NOIA framework could be a more unified and comprehensive system for detecting GxG/GxE interactions and even POE, which will provide invaluable insight into the efforts for finding the "missing heritability". And the revealed interactions will be useful to help explain the underlying mechanism of the development of lung cancer and melanoma.

Next, we will extend our newly developed orthogonal models to higher dimensional interactions which could be easily reached by applying the Kronecker product, for example, GxGxG or GxGxE interactions modeling or GxG/GxE interactions testing with consideration of imprinting effects. The more than two dimensional interactions have not been widely investigated; however, they may indeed explain some of the heritability of human complex diseases and traits. We will also explore to explore the application of our models to the cases when the loci are not in HWE or the parent origin information is missed.

APPENDIX

Appendix 1: Supplementary figures and tables

Figure S2.1 Density distributions of the estimates of the parameters from a simulated data analysis with a quantitative trait influenced by two loci and positive interaction coefficients.

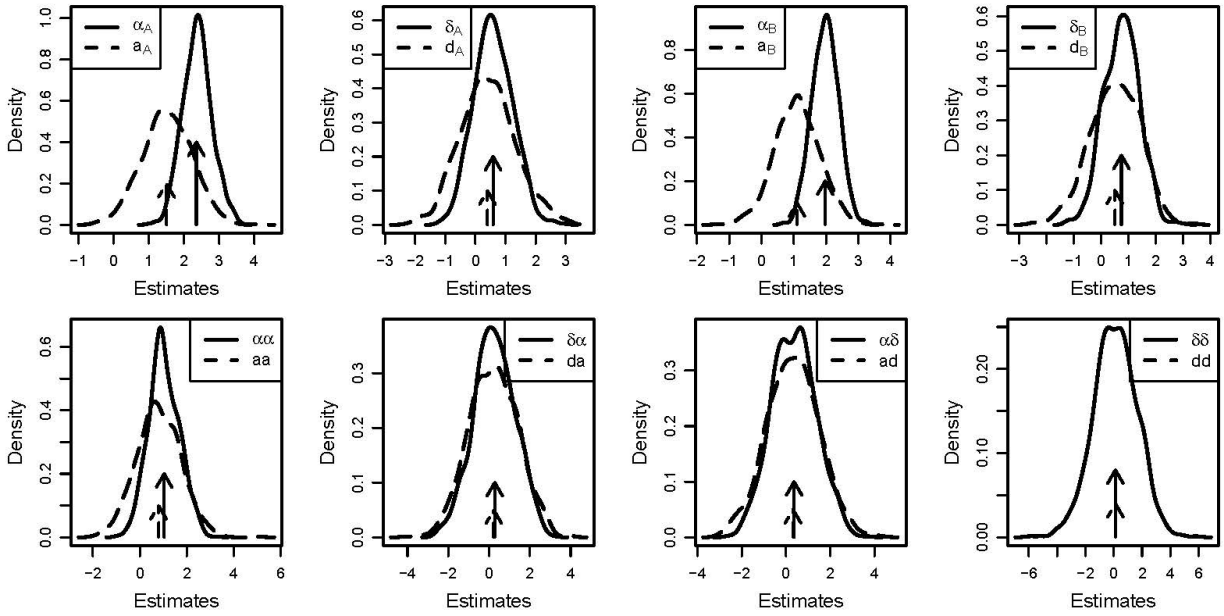

Figure S2.2 Density distributions of the estimates of the parameters from a simulated data analysis with a quantitative trait influenced by two loci and negative interaction coefficients.

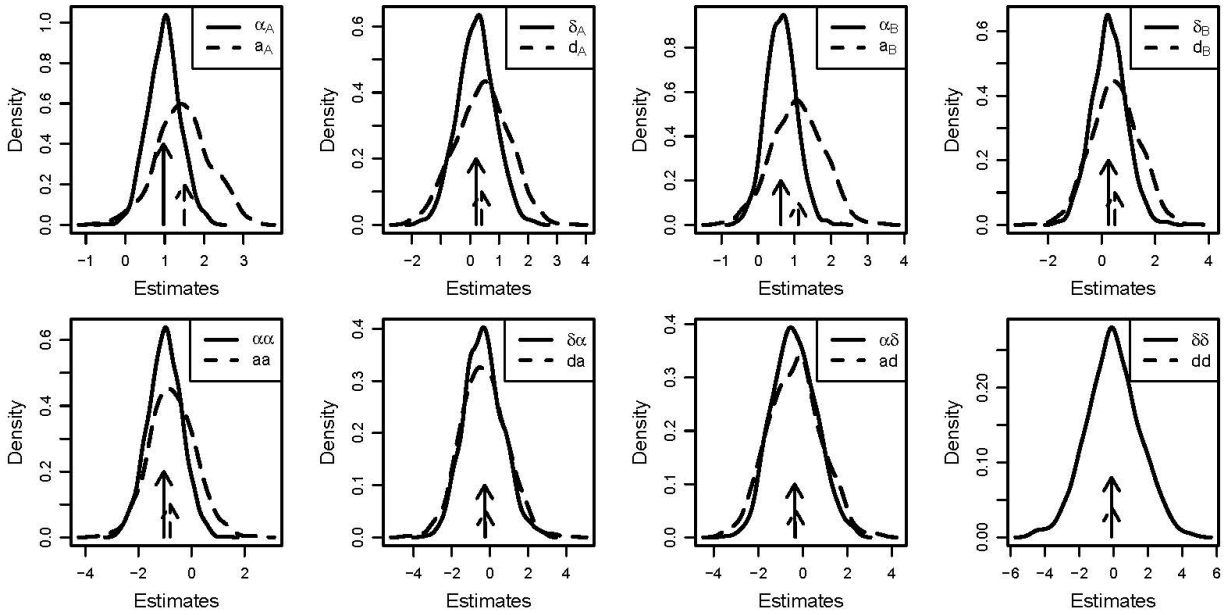

Figure S2.3 Density distributions of the estimates of the parameters from a simulated data analysis with a quantitative trait influenced by two loci and no gene-gene interactions.

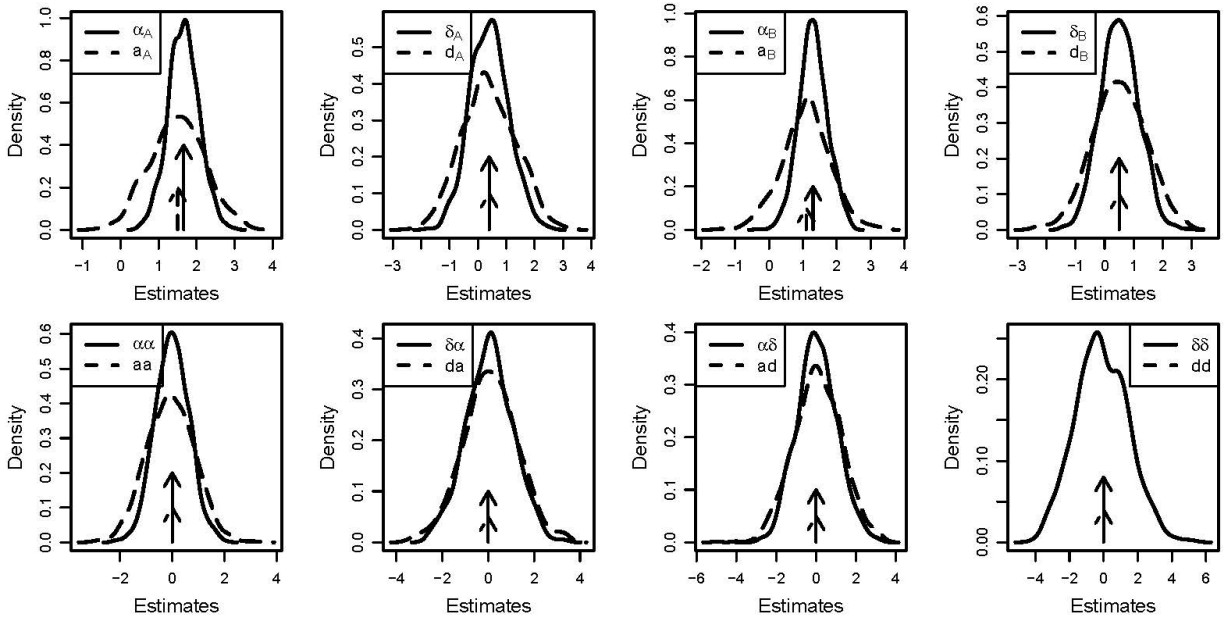

Figure S2.4 Density distributions of the estimates of the parameters from a simulated data analysis with a case-control trait influenced by two loci and positive g interaction coefficients.

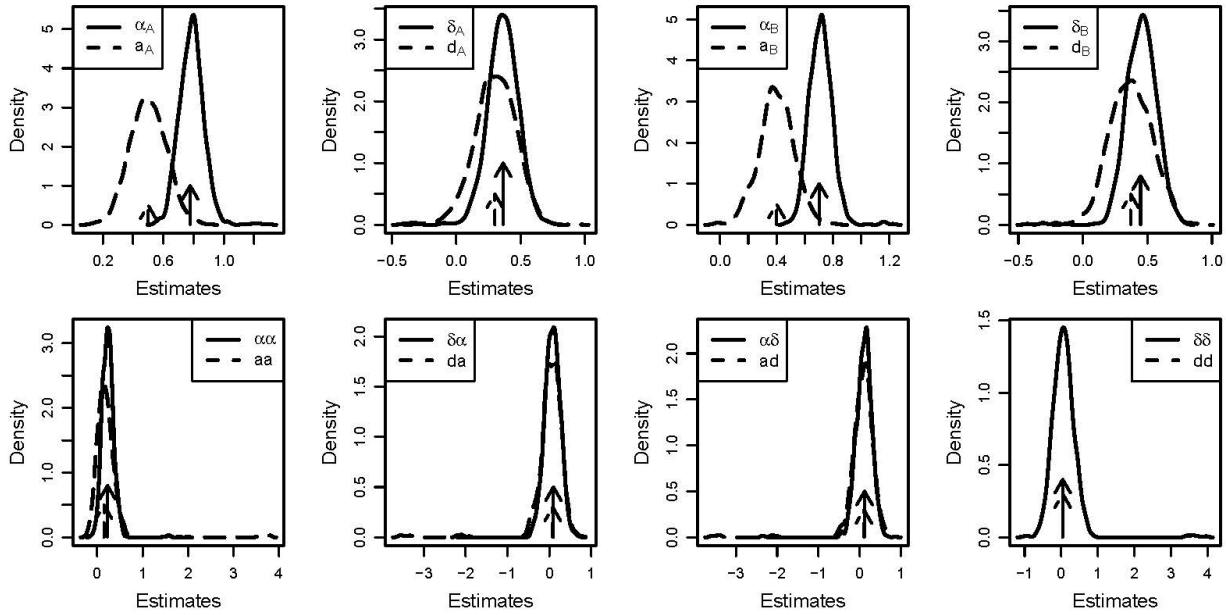

Figure S2.5 Density distributions of the estimates of the parameters from a simulated data analysis with a case-control trait influenced by two loci and negative interaction coefficients.

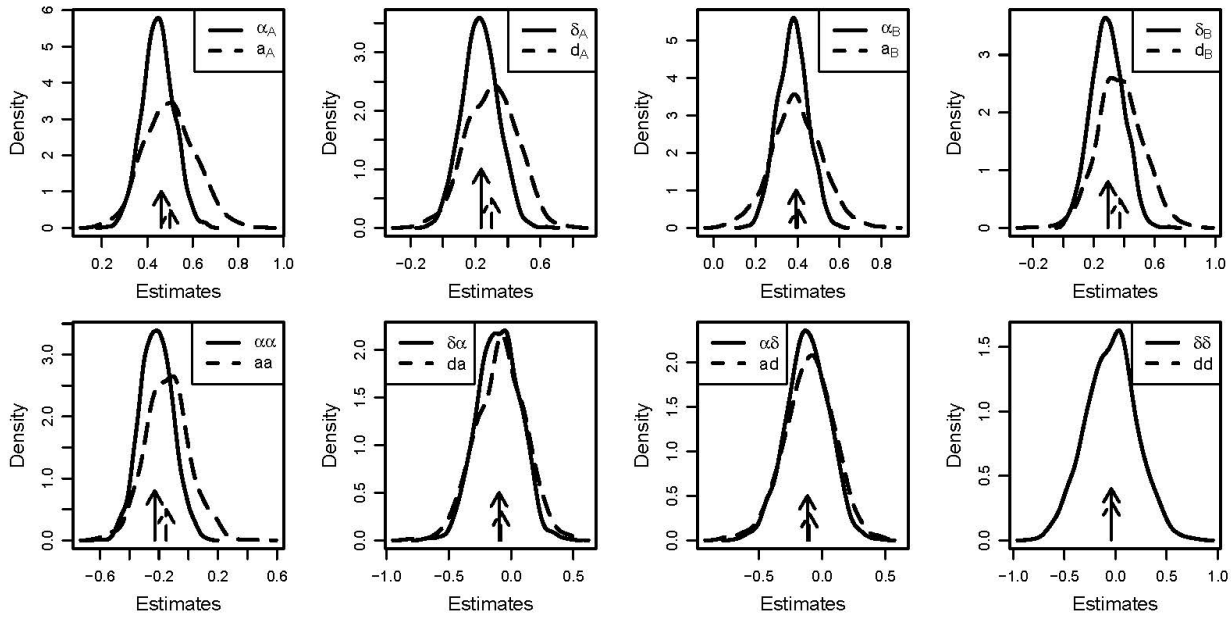

Figure S2.6 Density distributions of the estimates of the parameters from a simulated data analysis with a case-control trait influenced by two loci and no gene-gene interactions.

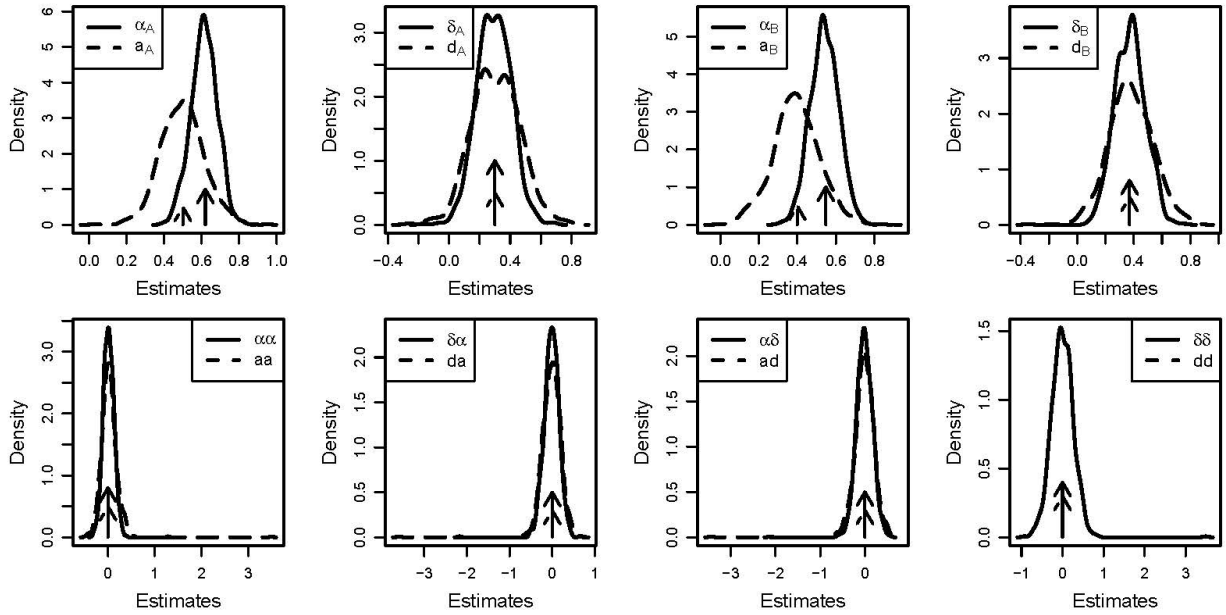

Figure S2.7 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation data influence by two loci and positive interaction coefficients. The minor allele frequency was 0.50. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were $\vec{E}_F^{\,T} = [100.00, 1.50, 0.40, 1.10,\ 0.50, 0.80, 0.23, 0.32, 0.12]$. Corresponding values of the statistical genetic effects were $\vec{E}_S^{\,T} = [104.16, 2.46, 0.69, 2.02, 0.88, 0.8, 0.23, 0.32, 0.12]$.

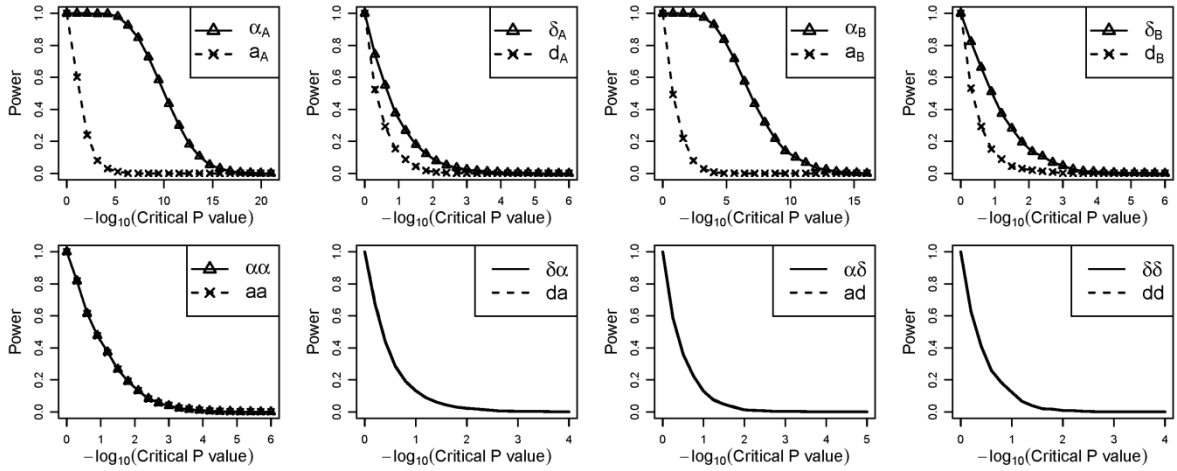

Figure S2.8 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation data influence by two loci and negative interaction coefficients. The minor allele frequency was 0.50. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were $\vec{E}_F^{\,T} = [100.00, 1.50, 0.40, 1.10,\ 0.50, -0.80, -0.23, -0.32, -0.12]$. Corresponding values of the statistical genetic effects were $\vec{E}_S^{\,T} = [101.95, 0.54, 0.11, 0.18, 0.12, -0.80, -0.23, -0.32, -0.12]$.

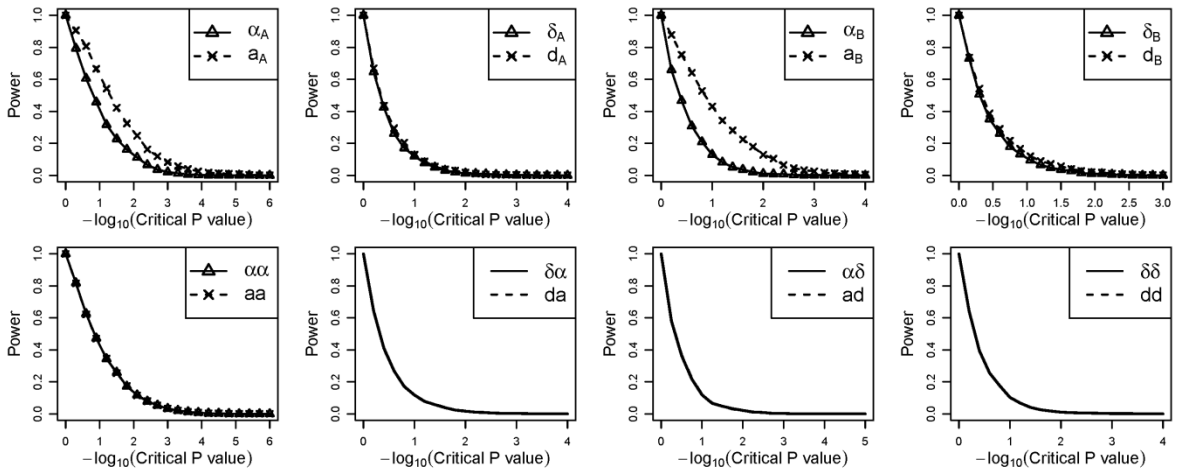

Figure S2.9 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation data influence by two loci no interaction effects. The minor allele frequency was 0.50. The upper panel is for the additive effects and dominant effects of locus A and locus B, respectively. The bottom panel is for the interaction effect between locus A and locus B. The simulating values of the genetic effects were $\vec{E}_F^{\,T} = [100.00, 1.50, 0.40, 1.10,\ 0.50, 0.0, 0.0, 0.0, 0.0]$. Corresponding values of the statistical genetic effects were $\vec{E}_S^{\,T} = [103.05, 1.50, 0.40, 1.10, 0.50, 0.0, 0.0, 0.0, 0.0]$.

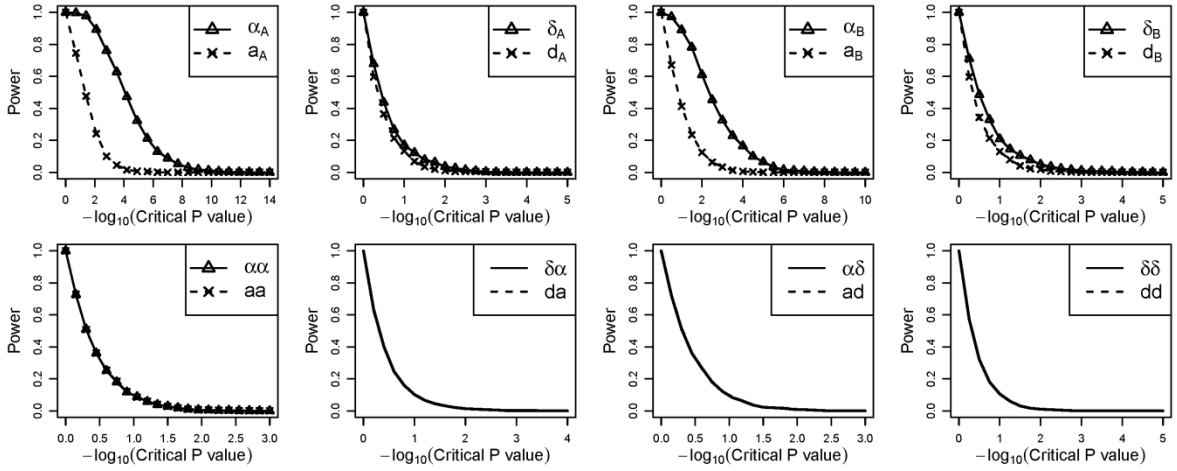

Figure S2.10 Q-Q plot for P values of genotyped SNPs obtained from NOIA statistical model on additive effect estimation. lambda=1.011.

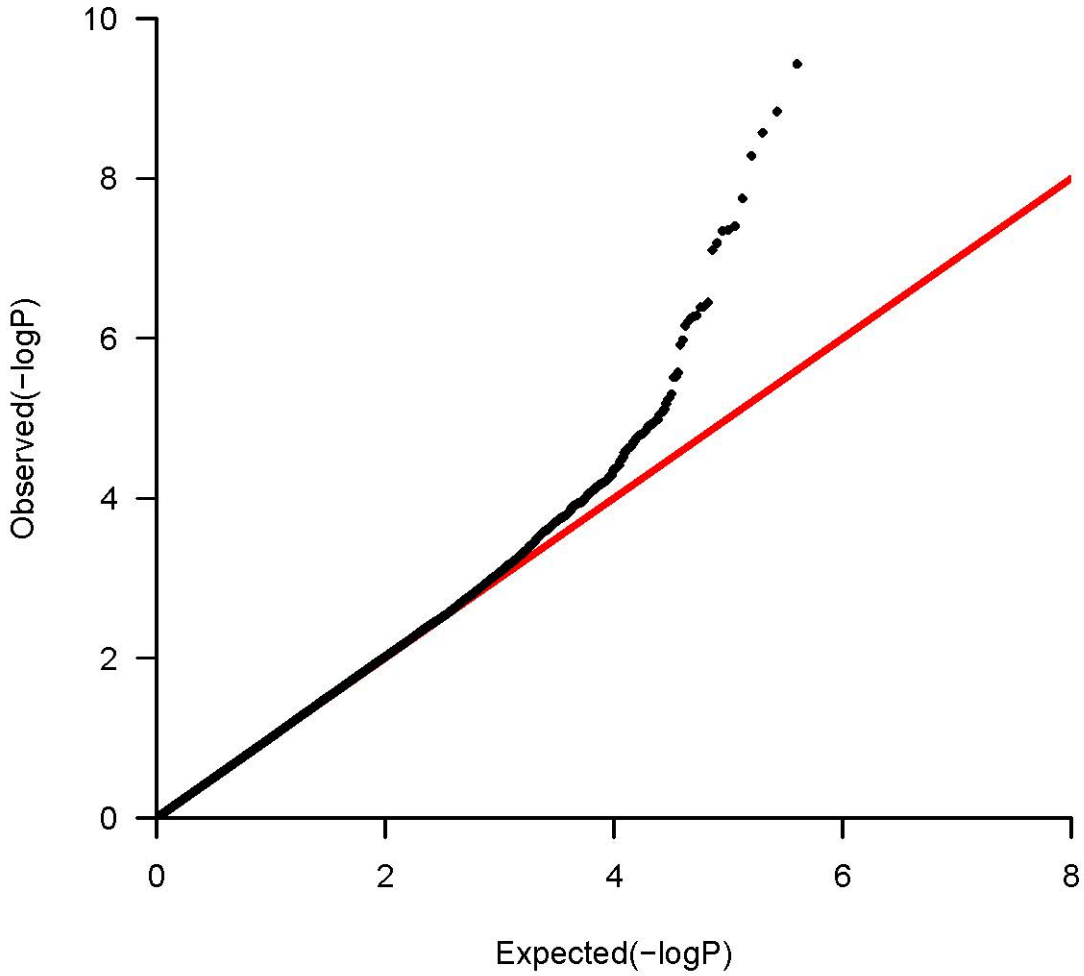

Figure S2.11 Q-Q plot for P values of genotyped SNPs obtained from NOIA statistical model with dominance component detection on additive effect estimation, lambda=1.014. SNPs with genotype frequency of any homozygote less than 0.005 were filtered.

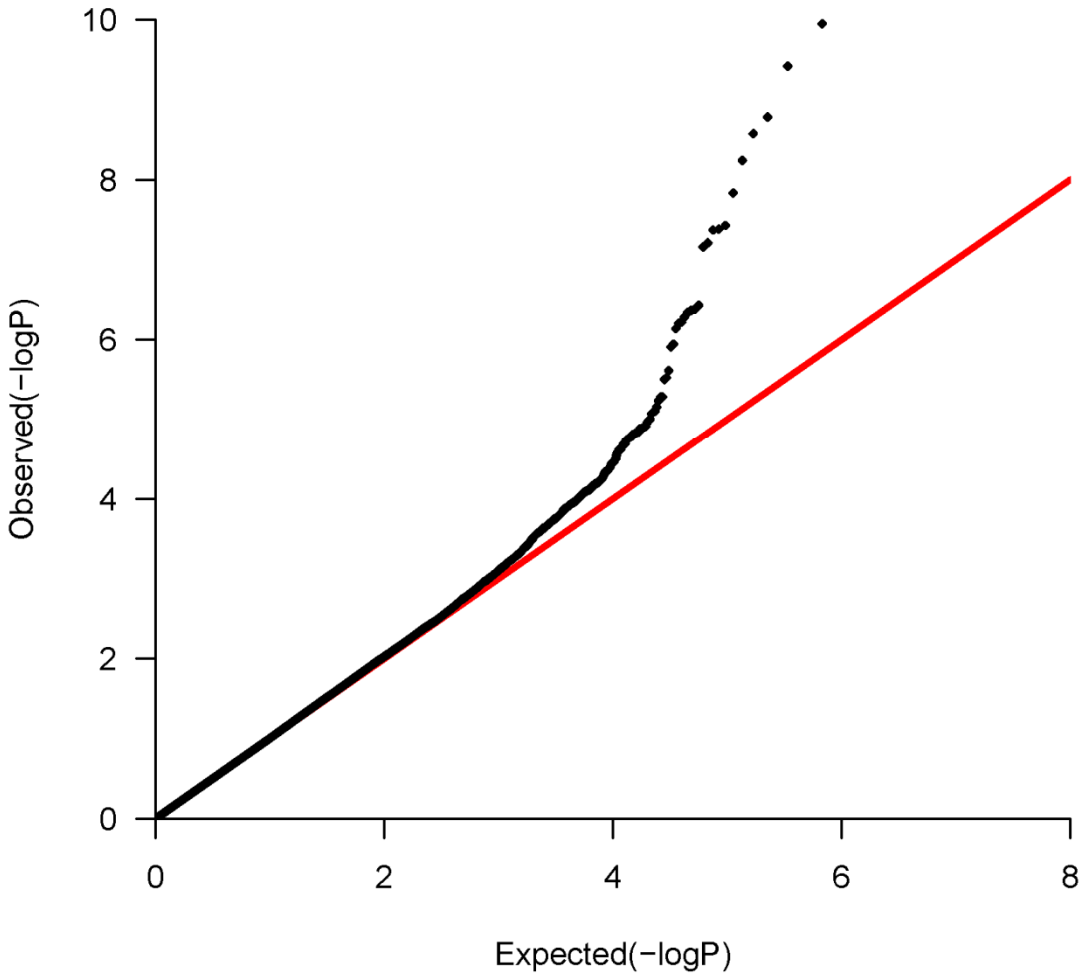

Table S2.1 Results from genome-wide association analysis of melanoma by NOIA statistical one-locus model using logistic regression ($p < 1.0 \times 10^{-4}$).

CHR	SNP	A1	A2	A2 freq	Coordinate	OR	2.5% CI	97.5% CI	P	Gene Symbol
1	rs2089427	A	G	0.45	213637204	1.24	1.11	1.39	9.96E-05	LOC643536
1	rs6693552	A	C	0.18	238280547	1.34	1.16	1.55	9.39E-05	LOC645884
1	rs12733694	A	G	0.18	238284757	1.34	1.16	1.56	7.69E-05	FMN2
1	rs11204754	A	G	0.5	149227878	0.8	0.72	0.89	6.43E-05	ANXA9
1	rs12753507	A	G	0.43	213446975	1.24	1.11	1.39	9.81E-05	KCNK2
1	rs1722784	A	G	0.5	149228493	1.26	1.13	1.4	3.59E-05	ANXA9
1	rs11506	A	G	0.07	163898130	0.67	0.55	0.82	8.57E-05	ALDH9A1
1	rs10926064	A	G	0.64	238240621	1.27	1.14	1.42	2.67E-05	LOC645884
2	rs2060167	A	G	0.75	166397588	1.29	1.14	1.46	7.46E-05	TTC21B
2	rs12471713	C	G	0.88	65957786	1.48	1.26	1.75	2.98E-06	FLJ16124
2	rs3791511	A	G	0.88	239745477	0.7	0.58	0.83	7.71E-05	HDAC4
2	rs2083244	A	G	0.39	119377691	0.79	0.71	0.88	4.14E-05	MARCO
3	rs4643673	A	G	0.36	148129994	0.79	0.7	0.88	3.52E-05	PLSCR5
3	rs3912449	A	G	0.14	7658162	0.73	0.63	0.85	5.98E-05	GRM7
3	rs9790140	A	C	0.92	96158387	1.49	1.23	1.82	5.97E-05	WDR82P1
3	rs6549877	A	G	0.87	28720143	1.37	1.17	1.61	9.04E-05	C3orf53
3	rs1872396	A	G	0.86	7655452	1.38	1.18	1.61	4.46E-05	GRM7
4	rs17035512	C	G	0.88	106509509	1.4	1.19	1.65	4.50E-05	PPA2
4	rs6811159	A	G	0.12	106542507	0.72	0.61	0.84	6.12E-05	PPA2
4	rs17035584	A	G	0.88	106574235	1.39	1.18	1.64	7.01E-05	PPA2
4	rs6823995	A	G	0.13	106483130	0.72	0.61	0.85	6.51E-05	PPA2
4	rs17035553	A	G	0.84	106540872	1.35	1.17	1.56	5.22E-05	PPA2
4	rs3898404	A	G	0.16	106575844	0.75	0.64	0.86	8.05E-05	PPA2
4	rs6812270	A	T	0.12	106285899	0.72	0.62	0.85	9.73E-05	KIAA1546
5	rs10940474	A	G	0.34	54916458	1.3	1.16	1.47	1.49E-05	FLJ90709
6	rs4431416	A	G	0.43	68146608	1.24	1.12	1.39	9.43E-05	LOC728052
6	rs7769019	A	G	0.04	33706620	0.57	0.43	0.74	4.39E-05	ITPR3
6	rs2495971	A	C	0.16	34037043	1.38	1.18	1.61	5.47E-05	GRM4
6	rs3087617	A	T	0.08	31664635	1.54	1.24	1.91	8.64E-05	LST1
6	rs9454109	A	G	0.72	68004709	0.75	0.67	0.85	5.18E-06	RCADH5
7	rs10245068	A	C	0.89	37331759	0.68	0.56	0.82	4.53E-05	ELMO1
8	rs2248448	A	G	0.5	4528471	0.81	0.72	0.9	9.68E-05	CSMD1
8	rs4909616	A	G	0.31	135569749	0.78	0.7	0.88	4.53E-05	ZFAT1
8	rs10094500	A	C	0.51	4558681	1.28	1.15	1.43	8.41E-06	CSMD1
8	rs2724961	A	G	0.53	4547635	0.79	0.7	0.88	1.52E-05	CSMD1
8	rs2617014	A	G	0.55	4537866	0.78	0.7	0.87	6.99E-06	CSMD1
9	rs7023954	A	G	0.61	21806758	1.28	1.15	1.43	1.22E-05	MTAP
9	rs11792508	A	C	0.92	243594	1.52	1.25	1.84	2.78E-05	DOCK8
9	SNP9-21803495	A	G	0.52	21803495	0.78	0.7	0.88	1.48E-05	MTAP
9	SNP9-21803518	A	G	0.51	21803518	0.8	0.72	0.89	5.37E-05	MTAP
9	rs1987458	A	G	0.44	21694873	0.8	0.72	0.89	6.46E-05	LOC402359
9	SNP9-21803241	A	G	0.49	21803241	1.25	1.12	1.4	5.78E-05	MTAP
9	SNP9-21816516	A	G	0.38	21816516	0.78	0.7	0.87	1.29E-05	MTAP

9	SNP9-21816940	A	C	0.62	21816940	1.28	1.15	1.43	1.29E-05	MTAP
9	SNP9-21808674	A	G	0.41	21808674	0.79	0.7	0.88	1.68E-05	MTAP
9	SNP9-21786791	A	G	0.63	21786791	1.29	1.15	1.44	9.87E-06	MTAP
9	SNP9-21817406	A	G	0.38	21817406	0.78	0.7	0.88	1.64E-05	MTAP
9	rs10811615	A	G	0.55	21772164	1.25	1.12	1.39	8.34E-05	LOC402359
9	SNP9-21761241	A	G	0.39	21761241	0.79	0.7	0.88	2.43E-05	MTAP
9	SNP9-21818242	A	G	0.52	21818242	0.79	0.71	0.88	2.23E-05	MTAP
9	SNP9-21806637	A	G	0.41	21806637	0.78	0.7	0.87	1.03E-05	MTAP
9	SNP9-21796564	A	G	0.62	21796564	1.28	1.14	1.43	1.42E-05	MTAP
9	rs1561650	A	G	0.39	21742358	0.8	0.71	0.89	6.87E-05	LOC402359
9	SNP9-21778782	A	G	0.61	21778782	1.27	1.14	1.42	1.78E-05	MTAP
9	SNP9-21775139	A	T	0.39	21775139	0.78	0.7	0.88	1.80E-05	MTAP
9	SNP9-21803718	A	G	0.47	21803718	0.78	0.7	0.87	8.04E-06	MTAP
9	SNP9-21775018	A	G	0.61	21775018	1.27	1.14	1.42	2.45E-05	MTAP
9	SNP9-21799077	A	T	0.62	21799077	1.27	1.14	1.42	2.26E-05	MTAP
9	SNP9-21777262	A	G	0.61	21777262	1.28	1.14	1.42	1.72E-05	MTAP
9	SNP9-21818110	A	C	0.38	21818110	0.78	0.7	0.87	1.11E-05	MTAP
9	rs10965144	A	G	0.62	21798913	1.26	1.13	1.41	3.74E-05	MTAP
9	SNP9-21774758	A	T	0.61	21774758	1.27	1.14	1.42	2.07E-05	MTAP
9	SNP9-21778481	A	G	0.39	21778481	0.79	0.7	0.88	2.03E-05	MTAP
9	rs3928894	A	G	0.49	21808310	1.25	1.12	1.39	7.84E-05	MTAP
9	SNP9-21806646	A	G	0.51	21806646	0.8	0.72	0.89	6.39E-05	MTAP
9	rs10965133	A	G	0.61	21778656	1.27	1.14	1.42	2.06E-05	LOC402359
9	SNP9-21760951	A	G	0.39	21760951	0.79	0.71	0.88	3.26E-05	MTAP
9	rs1335503	A	G	0.39	21727822	0.8	0.72	0.9	8.94E-05	LOC402359
9	SNP9-21783177	A	C	0.48	21783177	0.79	0.71	0.89	3.40E-05	MTAP
9	SNP9-21768660	A	G	0.55	21768660	1.25	1.12	1.39	7.80E-05	MTAP
9	SNP9-21764467	A	G	0.48	21764467	0.8	0.72	0.9	8.78E-05	MTAP
9	rs1335500	A	G	0.49	21701675	1.32	1.18	1.47	6.24E-07	LOC402359
9	SNP9-21751440	C	G	0.52	21751440	1.24	1.11	1.39	9.04E-05	MTAP
9	SNP9-21761756	A	G	0.48	21761756	0.8	0.72	0.89	6.86E-05	MTAP
9	SNP9-21762267	A	G	0.48	21762267	0.8	0.72	0.9	8.11E-05	MTAP
9	rs12380505	A	G	0.5	21685893	0.76	0.68	0.85	6.02E-07	LOC402359
9	SNP9-21780669	C	G	0.52	21780669	1.26	1.13	1.4	4.27E-05	MTAP
9	SNP9-21794693	A	G	0.48	21794693	0.79	0.71	0.89	3.14E-05	MTAP
9	rs10811582	A	G	0.4	21682017	1.29	1.16	1.45	5.80E-06	LOC402359
9	SNP9-21775304	C	G	0.52	21775304	1.24	1.11	1.38	9.91E-05	MTAP
9	SNP9-21765061	A	G	0.48	21765061	0.8	0.72	0.9	8.23E-05	MTAP
9	SNP9-21780142	C	G	0.52	21780142	1.25	1.12	1.39	6.32E-05	MTAP
9	rs896655	A	G	0.39	21696571	0.8	0.72	0.89	6.69E-05	LOC402359
9	SNP9-21763167	C	G	0.48	21763167	0.8	0.72	0.89	4.74E-05	MTAP
9	rs2383202	A	G	0.49	21700215	1.32	1.19	1.47	5.24E-07	LOC402359
9	SNP9-21780067	A	G	0.48	21780067	0.8	0.72	0.89	6.90E-05	MTAP
9	SNP9-21778523	A	G	0.48	21778523	0.8	0.72	0.9	8.15E-05	MTAP
9	SNP9-21755601	A	G	0.52	21755601	1.24	1.11	1.39	9.27E-05	MTAP
9	SNP9-21765957	A	G	0.52	21765957	1.25	1.12	1.39	7.01E-05	MTAP
9	SNP9-21759412	A	G	0.52	21759412	1.24	1.11	1.39	9.77E-05	MTAP
9	SNP9-21792469	A	G	0.47	21792469	0.8	0.71	0.89	3.49E-05	MTAP
9	rs7866787	A	G	0.48	21750639	0.8	0.72	0.9	7.94E-05	LOC402359

9	SNP9-21778081	A	C	0.48	21778081	0.8	0.72	0.89	7.24E-05	MTAP
9	rs7848524	A	G	0.5	21691432	0.76	0.68	0.84	4.28E-07	LOC402359
9	rs7023329	A	G	0.47	21806528	0.79	0.71	0.88	1.27E-05	MTAP
9	rs6475552	A	G	0.5	21691674	1.32	1.19	1.48	3.71E-07	LOC402359
9	rs1345026	A	G	0.58	21735756	0.79	0.7	0.88	2.43E-05	LOC402359
9	rs1452658	A	G	0.5	21690795	1.32	1.18	1.47	7.22E-07	LOC402359
9	SNP9-21789598	A	G	0.49	21789598	0.75	0.68	0.84	4.15E-07	MTAP
10	rs2797272	A	G	0.24	9125309	0.77	0.68	0.88	5.74E-05	LOC389936
10	rs11001702	A	G	0.09	53773431	0.67	0.56	0.8	1.25E-05	DKK1
11	rs644817	A	C	0.93	69749575	0.62	0.48	0.78	7.88E-05	FADD
12	rs12826471	A	G	0.84	3455203	0.71	0.61	0.83	1.53E-05	PRMT8
13	rs2202561	A	G	0.61	70442966	1.26	1.13	1.41	5.69E-05	LOC647277
13	rs7995083	A	G	0.13	23043162	1.41	1.2	1.67	4.96E-05	TNFRSF19
13	rs17691655	A	G	0.76	70445908	1.34	1.18	1.52	5.23E-06	LOC647277
14	rs7150290	A	G	0.14	52220859	1.4	1.19	1.64	4.10E-05	ERO1L
15	rs8030574	A	C	0.23	71415267	0.77	0.68	0.88	9.47E-05	HCN4
15	rs1129038	A	G	0.22	26030454	0.7	0.61	0.79	3.73E-08	HERC2
15	rs12913832	A	G	0.78	26039213	1.43	1.25	1.62	6.15E-08	HERC2
16	rs11648898	A	G	0.18	88573487	1.57	1.35	1.84	1.46E-08	AFG3L1
16	rs4238833	A	C	0.4	88578190	1.34	1.2	1.5	4.56E-07	AFG3L1
16	rs10852628	A	G	0.31	88607428	1.4	1.24	1.58	6.94E-08	DBNDD1
16	rs258322	A	G	0.88	88283404	0.63	0.53	0.76	1.13E-06	CDK10
16	rs164741	A	G	0.65	88219799	0.76	0.67	0.85	2.44E-06	DPEP1
16	rs4785751	A	G	0.53	88556918	1.43	1.29	1.6	1.13E-10	DEF8
16	rs17827507	A	G	0.27	83202987	0.77	0.68	0.86	1.43E-05	COTL1
16	rs4785752	A	G	0.53	88562642	0.73	0.66	0.82	4.14E-08	DEF8
16	rs352935	A	G	0.52	88176081	1.25	1.12	1.39	8.26E-05	CPNE7
16	rs4785759	A	C	0.53	88578381	0.73	0.66	0.82	4.26E-08	AFG3L1
16	rs8051733	A	G	0.36	88551707	1.42	1.27	1.59	2.66E-09	DEF8
16	rs4408545	A	G	0.54	88571529	1.43	1.28	1.59	3.81E-10	AFG3L1
16	rs4785763	A	C	0.63	88594437	0.75	0.67	0.84	1.23E-06	AFG3L1
16	rs11076650	A	G	0.46	88595442	1.4	1.26	1.56	1.65E-09	DBNDD1
16	rs7195043	A	G	0.5	88548362	0.72	0.64	0.8	5.73E-09	DEF8
16	rs9939542	A	C	0.3	88580549	1.33	1.18	1.51	3.15E-06	AFG3L1
17	rs3744578	A	G	0.21	11589057	1.32	1.15	1.52	8.33E-05	DNAH9
17	rs9904264	A	G	0.21	11587008	1.34	1.17	1.54	3.96E-05	DNAH9
17	rs16957962	A	G	0.33	9000751	1.29	1.14	1.45	2.85E-05	NTN1
19	rs868878	A	G	0.88	7737047	0.69	0.57	0.82	4.60E-05	CLEC4M
19	rs2285963	A	G	0.88	5542735	1.42	1.21	1.68	2.57E-05	SAFB2
19	rs934433	A	G	0.6	36000617	0.79	0.7	0.88	1.86E-05	ZNF536
19	rs934432	A	C	0.4	36000424	1.27	1.14	1.42	1.66E-05	ZNF536
19	rs6510181	A	C	0.39	35984559	1.25	1.12	1.4	6.43E-05	ZNF536
19	rs1549951	A	G	0.61	35984184	0.8	0.71	0.89	6.85E-05	ZNF536
19	rs3745542	A	G	0.3	56279455	1.32	1.17	1.49	7.87E-06	KLK14
20	rs2284271	A	G	0.91	43038835	1.47	1.21	1.78	8.12E-05	STK4
20	rs17730901	A	C	0.88	16198822	1.39	1.18	1.63	7.78E-05	C20orf23
20	rs4814466	A	G	0.88	16201819	1.39	1.18	1.64	6.24E-05	C20orf23
20	SNP20-31969319	C	G	0.89	31969319	0.68	0.57	0.82	7.25E-05	CHMP4B

Figure S3.1 Density distributions of the estimates of the parameters from a simulated data analysis, illustrated in Figure 3.1. The pre-specified minor allele frequency and exposure frequency was 0.30 and 0.22, respectively. The simulating residual variance was 144.0. The values of the six parameters of the genetic effects were $\vec{E}_F^{\,T} = [100.0, 3.0, 1.0, 2.0, 1.5, 1.0]$. The corresponding statistical genetic effects were $\vec{E}_S^{\,T} = [101.75, 4.18, 1.22, 2.71, 2.2, 1.0]$. The solid arrows denote the true simulated values of the parameters for the NOIA statistical model and the dashed arrows denote those for the usual functional model.

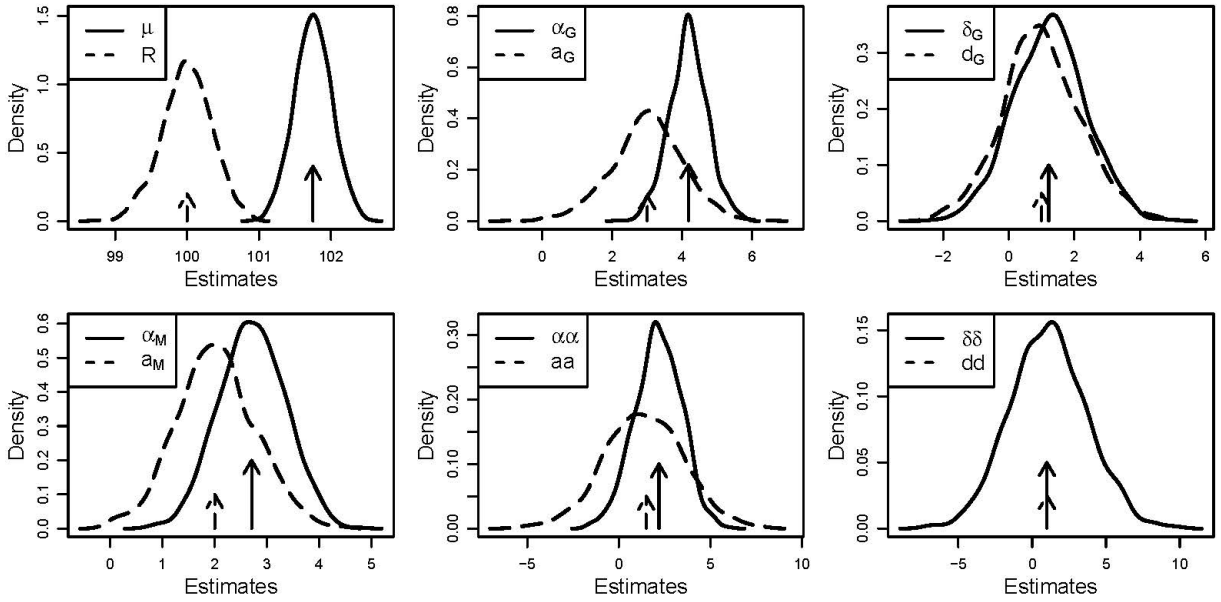

Figure S3.2 Density distributions of the estimates of the parameters from a simulated data analysis, illustrated in Figure 3.2. The pre-specified minor allele frequency and exposure frequency was 0.30 and 0.22, respectively. The simulating residual variance was 144.0. The values of the six parameters of the genetic effects were $\vec{\mathrm{E}}_{\mathrm{F}}{}^{\mathrm{T}} = [\,100.0,\ 3.0,\ 1.0,\ 0.0,\ 0.0,\ 0.0]$. The corresponding statistical genetic effects were $\vec{E}_S{}^{T} = [101.16,\ 3.70,1.00,0.00,0.00,0.00]$. The solid arrows denote the true simulated values of the parameters for the NOIA statistical model and the dashed arrows denote those for the usual functional model.

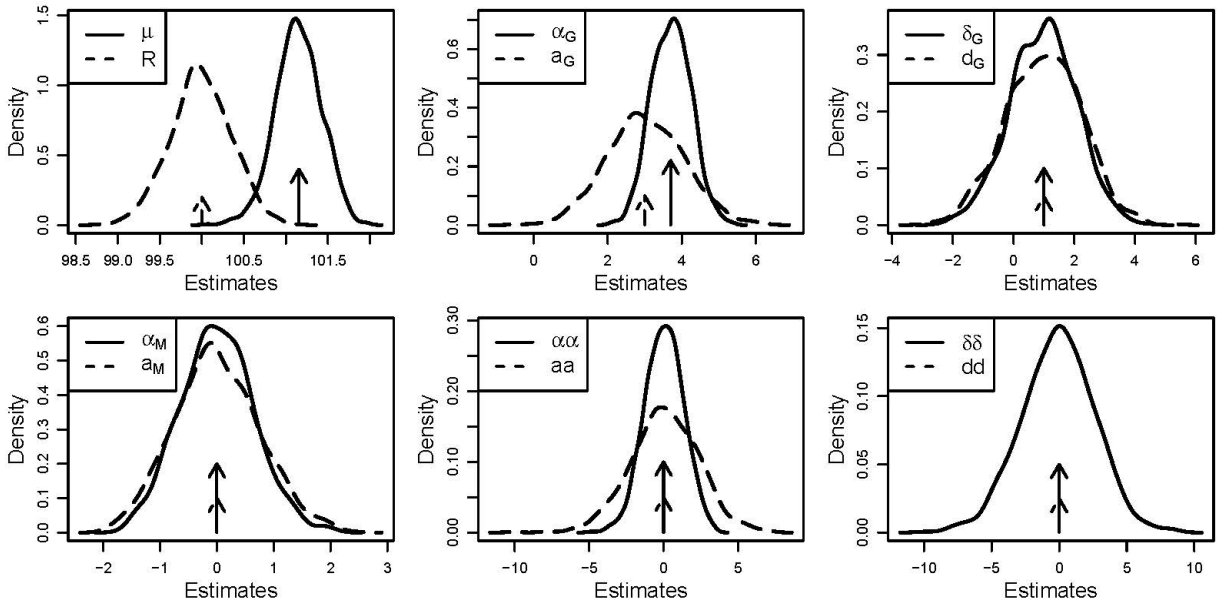

Figure S3.3 Density distributions of the estimates of the parameters from a simulated data analysis, illustrated in Figure 3.3. The pre-specified minor allele frequency and exposure frequency was 0.25 and 0.22, respectively. The values of the six parameters of the genetic effects were $\vec{\mathrm{E}}_{\mathrm{F}}^{\ \mathrm{T}} = [-2.0,\ 0.3,\ 0.1,\ 0.2,\ 0.1,\ 0.04]$. The corresponding statistical genetic effects were $\vec{E}_S^{\ T} = [-1.75,\ 0.38,\ 0.11,\ 0.27,\ 0.12,\ 0.04]$. The solid arrows denote the true simulated values of the parameters for the NOIA statistical model and the dashed arrows denote those for the usual functional model.

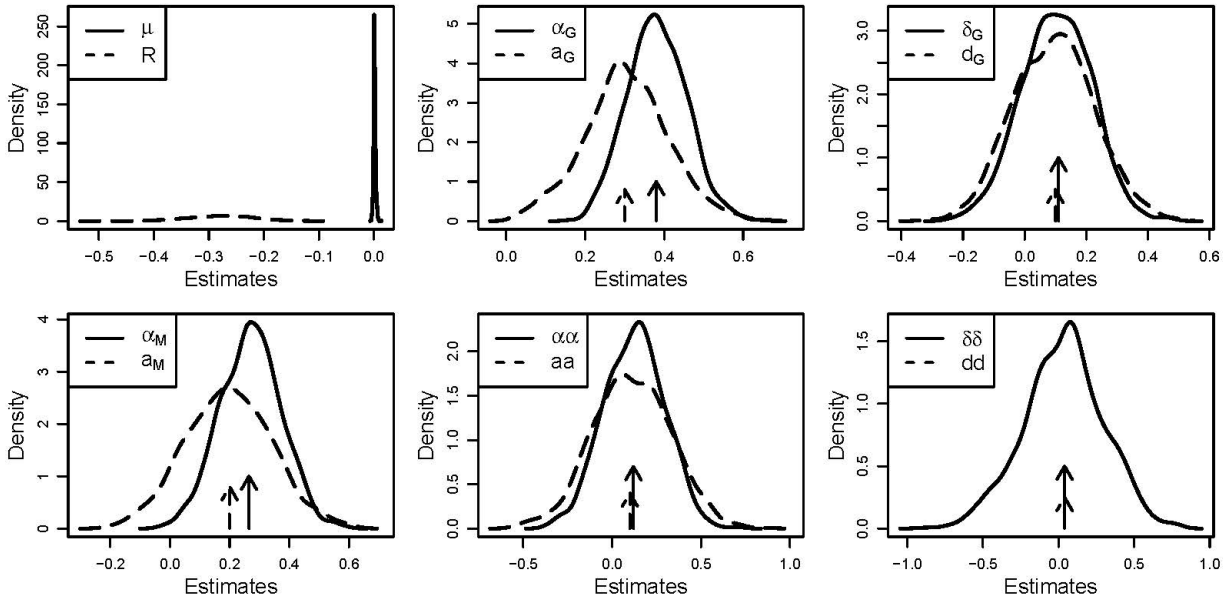

Figure S3.4 Density distributions of the estimates of the parameters from a simulated data analysis, illustrated in Figure 3.4. The pre-specified minor allele frequency and exposure frequency was 0.25 and 0.22, respectively. The values of the six parameters of the genetic effects were $\vec{E}_F{}^T = [-2.0,\ 0.4,\ 0.2,\ 0.0,\ 0.0,\ 0.0]$. The corresponding statistical genetic effects were $\vec{E}_S{}^T = [-1.73,\ 0.5,\ 0.2,\ 0.0,\ 0.0,\ 0.0]$. The solid arrows denote the true simulated values of the parameters for the NOIA statistical model and the dashed arrows denote those for the usual functional model.

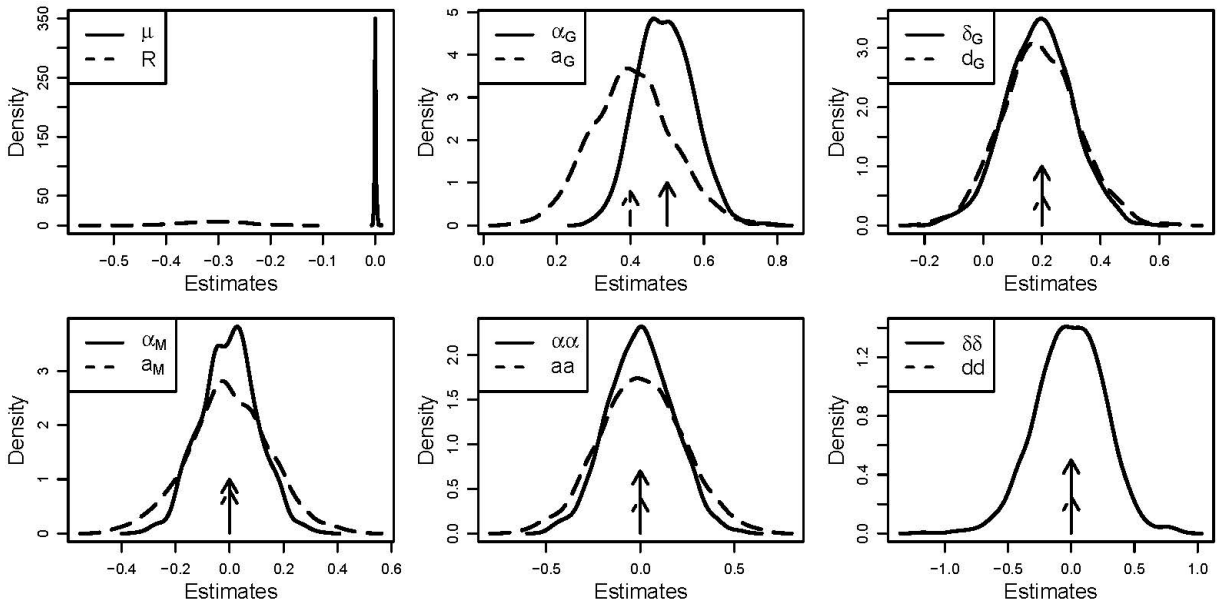

Figure S4.1 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation data influence by a genetic factor with strong POE (scenario 1). The minor allele frequency was 0.03. Power for detecting (a) the main allelic additive effect, (b) the POE and (c) the dominant effect.

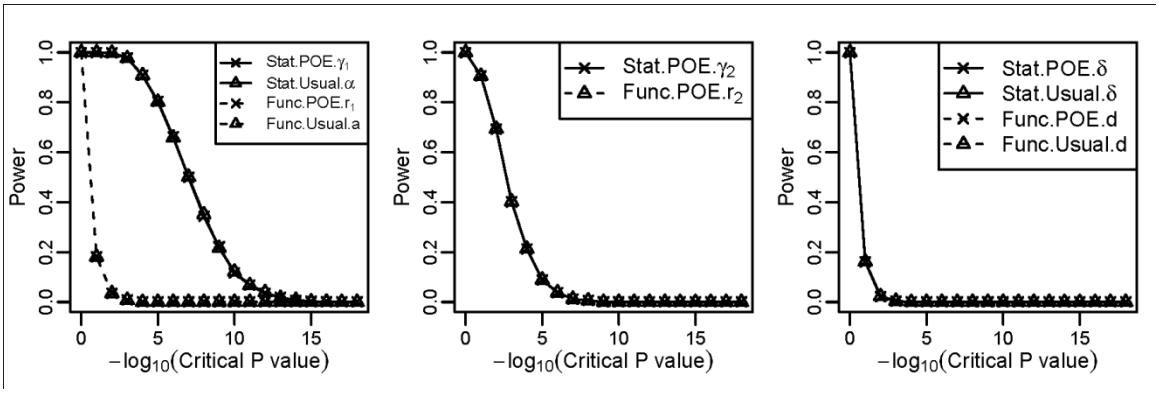

Figure S4.2 Power under different critical values of the P values obtained using the Wald test for the quantitative simulation data influence by a genetic factor with strong POE (scenario 1). The minor allele frequency was 0.48. Power for detecting (a) the main allelic additive effect, (b) the POE and (c) the dominant effect.

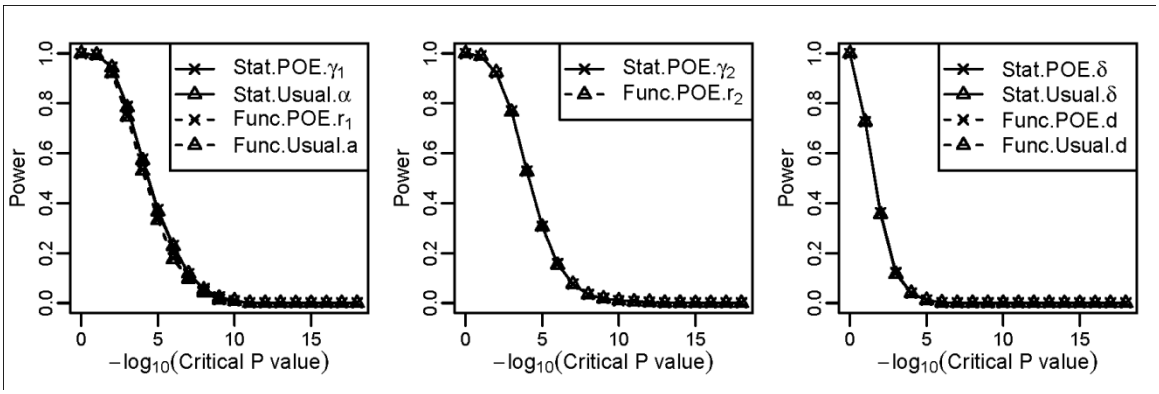

Figure S4.3 Power under different critical values of the P values obtained using the Wald test for the case-control simulation data influence by a genetic factor with POE. The minor allele frequency was 0.03. Power for detecting (a) the main allelic additive effect, (b) the POE and (c) the dominant effect.

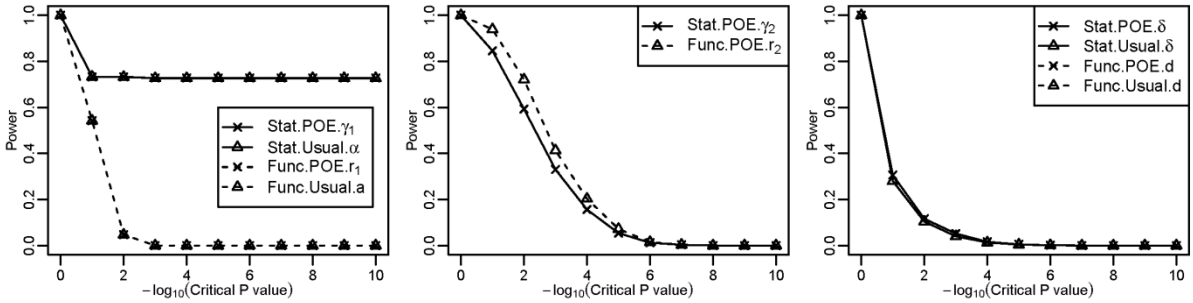

Figure S4.4 Power under different critical values of the P values obtained using the Wald test for the case-control simulation data influence by a genetic factor with POE. The minor allele frequency was 0.48. Power for detecting (a) the main allelic additive effect, (b) the POE and (c) the dominant effect.

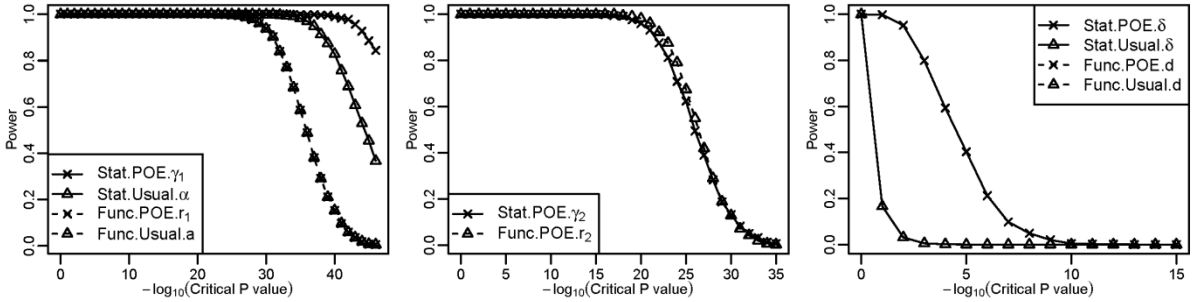

Table S3.1 Distributions of selected demographic variables of the ILCCO dataset

Variables	Case patients, No. (%)	Control subjects, No. (%)
Sex		
Male	4334(59)	6054(58)
Female	3058(41)	4390(42)
Age(year)		
<50	1043(14)	2172(21)
50-59	1823(25)	2704(26)
60-69	2594(35)	3430(33)
70-79	1682(23)	1969(19)
>=80	206(3)	141(1)
Smoking status		
never	735(10)	3344(32)
former smokers	3033(41)	3873(37)
current smokers	3524(48)	2772(27)
ever smokers	100(1)	455(4)

Table S3.2 Summary of the SNPs that we used in ILCCO dataset [a]

Chromosomal locus and variant	Risk Allele	MAF	Case	Control	Non.smoker.freq	SNPs
rs2736100_chr5p15	C	0.51	6684	9558	0.24	0=AA 1=AC 2=CC
rs402710_chr5p15	G	0.65	6682	8054	0.22	0=GG 1=GA 2=AA
rs2256543_chr6p21	A	0.44	6781	9621	0.24	0=GG 1=GA 2=AA
rs4324798_chr6q21	A	0.09	7283	10163	0.23	0=GG 1=GA 2=AA
rs16969968_chr15q25	A	0.35	7186	10070	0.23	0=GG 1=GA 2=AA
rs8034191_chr15q25	G	0.35	5070	8316	0.19	0=AA 1=AG 2=GG

a Non.smoker.freq=the frequency of non-smokers;

Appendix 2: One-Locus and Two-Locus Scan by the NOIA and Usual Functional model

Appendix 2.1: Gene-gene interaction: Reduced models

Ma et.al [22] stated the reduced one-locus models for NOIA statistical and usual models, including the additive, recessive, dominant models. Hereafter we are showing the corresponding two-locus models.

If the gene is additive, we have the following two-locus statistical model

$$\begin{pmatrix} G_{1111} \\ G_{1211} \\ G_{2211} \\ G_{1112} \\ G_{1212} \\ G_{2212} \\ G_{1122} \\ G_{1222} \\ G_{2222} \end{pmatrix} = \begin{pmatrix} 1 & -\bar{N}_B \\ 1 & 1-\bar{N}_B \\ 1 & 2-\bar{N}_B \end{pmatrix} \otimes \begin{pmatrix} 1 & -\bar{N}_A \\ 1 & 1-\bar{N}_A \\ 1 & 2-\bar{N}_A \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix} = \begin{pmatrix} 1 & -\bar{N}_A & -\bar{N}_B & \bar{N}_A\bar{N}_B \\ 1 & 1-\bar{N}_A & -\bar{N}_B & -(1-\bar{N}_A)\bar{N}_B \\ 1 & 2-\bar{N}_A & -\bar{N}_B & -(2-\bar{N}_A)\bar{N}_B \\ 1 & -\bar{N}_A & 1-\bar{N}_B & -\bar{N}_A(1-\bar{N}_B) \\ 1 & 1-\bar{N}_A & 1-\bar{N}_B & (1-\bar{N}_A)(1-\bar{N}_B) \\ 1 & 2-\bar{N}_A & 1-\bar{N}_B & (2-\bar{N}_A)(1-\bar{N}_B) \\ 1 & -\bar{N}_A & 2-\bar{N}_B & -\bar{N}_A(2-\bar{N}_B) \\ 1 & 1-\bar{N}_A & 2-\bar{N}_B & (1-\bar{N}_A)(2-\bar{N}_B) \\ 1 & 2-\bar{N}_A & 2-\bar{N}_B & (2-\bar{N}_A)(2-\bar{N}_B) \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix},$$

and functional model:

$$G_{AB} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 0 & 2 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 4 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix}.$$

They are related as

$$\begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix} = \begin{pmatrix} 1 & \bar{N}_A & \bar{N}_B & -\bar{N}_A\bar{N}_B \\ 0 & 1 & 0 & \bar{N}_B \\ 0 & 0 & 1 & \bar{N}_A \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix}.$$

If the gene is recessive, we have the following two-locus statistical model

$$\begin{pmatrix} G_{1111} \\ G_{1211} \\ G_{2211} \\ G_{1112} \\ G_{1212} \\ G_{2212} \\ G_{1122} \\ G_{1222} \\ G_{2222} \end{pmatrix} = \begin{pmatrix} 1 & -q_{22} \\ 1 & -q_{22} \\ 1 & 1-q_{22} \end{pmatrix} \otimes \begin{pmatrix} 1 & -p_{22} \\ 1 & -p_{22} \\ 1 & 1-p_{22} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix} = \begin{pmatrix} 1 & -p_{22} & -q_{22} & p_{22}q_{22} \\ 1 & -p_{22} & -q_{22} & p_{22}q_{22} \\ 1 & 1-p_{22} & -q_{22} & -(1-p_{22})q_{22} \\ 1 & -p_{22} & -q_{22} & p_{22}q_{22} \\ 1 & -p_{22} & -q_{22} & p_{22}q_{22} \\ 1 & 1-p_{22} & -q_{22} & -(1-p_{22})q_{22} \\ 1 & -p_{22} & 1-q_{22} & -p_{22}(1-q_{22}) \\ 1 & -p_{22} & 1-q_{22} & -p_{22}(1-q_{22}) \\ 1 & 1-p_{22} & 1-q_{22} & (1-p_{22})(1-p_{22}) \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix},$$

and the functional model:

$$G_{AB} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix}.$$

They are related as

$$\begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix} = \begin{pmatrix} 1 & p_{22} & q_{22} & -p_{22}q_{22} \\ 0 & 1 & 0 & -q_{22} \\ 0 & 0 & 1 & -p_{22} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix}.$$

If the gene is dominant, we have the following two-locus statistical model

$$\begin{pmatrix} G_{1111} \\ G_{1211} \\ G_{2211} \\ G_{1112} \\ G_{1212} \\ G_{2212} \\ G_{1122} \\ G_{1222} \\ G_{2222} \end{pmatrix} = \begin{pmatrix} 1 & q_{11}-1 \\ 1 & q_{11} \\ 1 & q_{11} \end{pmatrix} \otimes \begin{pmatrix} 1 & p_{11}-1 \\ 1 & p_{11} \\ 1 & p_{11} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix} = \begin{pmatrix} 1 & p_{11}-1 & q_{11}-1 & (p_{11}-1)(q_{11}-1) \\ 1 & p_{11} & q_{11}-1 & p_{11}(q_{11}-1) \\ 1 & p_{11} & q_{11}-1 & p_{11}(q_{11}-1) \\ 1 & p_{11}-1 & q_{11} & q_{11}(p_{11}-1) \\ 1 & p_{11} & q_{11} & p_{11}q_{11} \\ 1 & p_{11} & q_{11} & p_{11}q_{11} \\ 1 & p_{11}-1 & q_{11} & q_{11}(p_{11}-1) \\ 1 & p_{11} & q_{11} & p_{11}q_{11} \\ 1 & p_{11} & q_{11} & p_{11}q_{11} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix},$$

and the functional model

$$G_{AB} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix}.$$

They are related as

$$\begin{pmatrix} \mu \\ \alpha_A \\ \alpha_B \\ \alpha\alpha \end{pmatrix} = \begin{pmatrix} 1 & p_{11}-1 & q_{11}-1 & -(1-p_{11})(1-q_{11}) \\ 0 & 1 & 0 & q_{11}-1 \\ 0 & 0 & 1 & p_{11}-1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_A \\ a_B \\ aa \end{pmatrix}.$$

Appendix 2.2: Test Statistics for Full and Reduced One-Locus Models

The wald test statistic is $z = \frac{\hat{\beta}}{se(\hat{\beta})}$, where

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} = \left(se(\hat{\beta})\right)^2$$

The n rows of matrix Z represent the corresponding genotype for individual i. And $X = Z * S$.

$$Z = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{pmatrix}, \text{ and } Z'Z = n\begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 0 & p_{22} \end{pmatrix}$$

$$X'X = (ZS)'ZS = S'Z'ZS$$

We recall that, $\bar{N} = p_{12} + 2p_{22}$, $V = p_{12} + 4p_{22} - \bar{N}^2 = 4p_{11}p_{22} + p_{11}p_{12} + p_{12}p_{22}$.

For NOIA statistical model with no dominance component modeled,

$$S = \begin{pmatrix} 1 & -\bar{N} \\ 1 & 1-\bar{N} \\ 1 & 2-\bar{N} \end{pmatrix}$$

$$X'X = S'Z'ZS = \begin{pmatrix} 1 & 1 & 1 \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \end{pmatrix} * n\begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 0 & p_{22} \end{pmatrix} * \begin{pmatrix} 1 & -\bar{N} \\ 1 & 1-\bar{N} \\ 1 & 2-\bar{N} \end{pmatrix} = n\begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix}$$

$$\det(X'X) = nV$$

$$(X'X)^{-1} = \frac{1}{nV}\begin{pmatrix} V & 0 \\ 0 & 1 \end{pmatrix}$$

$$\text{Then } z = \frac{\hat{\beta}}{se(\hat{\beta})} = \frac{(X'X)^{-1}X'y}{\sqrt{\sigma^2(X'X)^{-1}}} = \frac{\frac{1}{nV}\begin{pmatrix} V & 0 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} 1 & 1 & 1 \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \end{pmatrix} * Z'y}{\sqrt{\frac{\sigma^2}{nV}\begin{pmatrix} \sqrt{V} & 0 \\ 0 & 1 \end{pmatrix}}} = \frac{1}{\sqrt{n\sigma^2 V}}\begin{pmatrix} \sqrt{V} & \sqrt{V} & \sqrt{V} \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \end{pmatrix} * Z'y$$

For NOIA statistical model with dominance component,

$$S_S = \begin{pmatrix} 1 & -\bar{N} & -\frac{2p_{12}p_{22}}{V} \\ 1 & 1-\bar{N} & \frac{4p_{11}p_{22}}{V} \\ 1 & 2-\bar{N} & -\frac{2p_{11}p_{12}}{V} \end{pmatrix}$$

$$X'X = {S_S}'Z'ZS_S = n\begin{pmatrix} 1 & 1 & 1 \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \\ -2p_{12}p_{22}/V & 4p_{11}p_{22}/V & -2p_{11}p_{12}/V \end{pmatrix} * \begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 0 & p_{22} \end{pmatrix} * \begin{pmatrix} 1 & -\bar{N} & -\frac{2p_{12}p_{22}}{V} \\ 1 & 1-\bar{N} & \frac{4p_{11}p_{22}}{V} \\ 1 & 2-\bar{N} & -\frac{2p_{11}p_{12}}{V} \end{pmatrix} = n\begin{pmatrix} 1 & 0 & 0 \\ 0 & V & 0 \\ 0 & 0 & \frac{4p_{11}p_{12}p_{22}}{V} \end{pmatrix}$$

$$\det(X'X) = 4np_{11}p_{12}p_{22}$$

$$(X'X)^{-1} = \frac{1}{4np_{11}p_{12}p_{22}}\begin{pmatrix} 4p_{11}p_{12}p_{22} & 0 & 0 \\ 0 & \frac{4p_{11}p_{12}p_{22}}{V} & 0 \\ 0 & 0 & V \end{pmatrix}$$

Then we got the test statistic for the NOIA one-locus full mode as follows:

$$z = \frac{\hat{\beta}}{se(\hat{\beta})} = \frac{(X'X)^{-1}X'y}{\sqrt{\sigma^2(X'X)^{-1}}} = \frac{1}{\sqrt{n\sigma^2V}} * \begin{pmatrix} \sqrt{V} & \sqrt{V} & \sqrt{V} \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \\ -\sqrt{\frac{p_{12}p_{22}}{p_{11}}} & 2\sqrt{\frac{p_{11}p_{22}}{p_{12}}} & -\sqrt{\frac{p_{11}p_{12}}{p_{22}}} \end{pmatrix} Z'y.$$

For usual functional model with no dominance component modeled,

$$S = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$$

$$X'X = S'Z'ZS = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} * n\begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{12} & 0 \\ 0 & 0 & p_{22} \end{pmatrix} * \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} = n\begin{pmatrix} 1 & \bar{N} \\ \bar{N} & \bar{N}+2p_{22} \end{pmatrix}$$

$$\det(X'X) = nV$$

$$(X'X)^{-1} = \frac{1}{nV}\begin{pmatrix} \bar{N}+2p_{22} & -\bar{N} \\ -\bar{N} & 1 \end{pmatrix}$$

$$z = \frac{\hat{\beta}}{se(\hat{\beta})} = \frac{(X'X)^{-1}X'y}{\sqrt{\sigma^2(X'X)^{-1}}} = \frac{1}{\sqrt{n\sigma^2V}} * \begin{pmatrix} \sqrt{\bar{N}+2p_{22}} & \frac{2p_{22}}{\sqrt{\bar{N}+2p_{22}}} & \frac{-p_{12}}{\sqrt{\bar{N}+2p_{22}}} \\ -\bar{N} & 1-\bar{N} & 2-\bar{N} \end{pmatrix} * Z'y$$

It results in the same test statistic as the NOIA model for the additive effect estimating.

For usual functional model with dominance component modeled,

$$X'X = S_F{}'Z'ZS_F = n\begin{pmatrix}1 & 1 & 1\\ 0 & 1 & 2\\ 0 & 1 & 0\end{pmatrix} * \begin{pmatrix}p_{11} & 0 & 0\\ 0 & p_{12} & 0\\ 0 & 0 & p_{22}\end{pmatrix} * \begin{pmatrix}1 & 0 & 0\\ 1 & 1 & 1\\ 1 & 2 & 0\end{pmatrix} = n\begin{pmatrix}1 & \bar{N} & p_{12}\\ \bar{N} & p_{12}+4p_{22} & p_{12}\\ p_{12} & p_{12} & p_{12}\end{pmatrix}$$

$$\det(X'X) = 4np_{11}p_{12}p_{22}$$

$$(X'X)^{-1} = \frac{1}{4np_{11}p_{12}p_{22}}\begin{pmatrix}4p_{12}p_{22} & -2p_{12}p_{22} & -2p_{12}p_{22}\\ -2p_{12}p_{22} & p_{12}(1-p_{12}) & -p_{12}(1-\bar{N})\\ -2p_{12}p_{22} & -p_{12}(1-\bar{N}) & V\end{pmatrix}$$

$$z = \frac{\hat{\beta}}{se(\hat{\beta})} = \frac{(X'X)^{-1}X'y}{\sqrt{\sigma^2(X'X)^{-1}}} == \frac{1}{\sqrt{n\sigma^2 V}}\begin{pmatrix}\sqrt{\frac{V}{p_{11}}} & 0 & 0\\ -\sqrt{\frac{p_{22}V}{p_{11}(1-p_{12})}} & 0 & \sqrt{\frac{p_{11}V}{p_{22}(1-p_{12})}}\\ -\sqrt{\frac{p_{12}p_{22}}{p_{11}}} & 2\sqrt{\frac{p_{11}p_{22}}{p_{12}}} & -\sqrt{\frac{p_{11}p_{12}}{p_{22}}}\end{pmatrix}Z'y$$

From the second row of the matrix, we observe that the test statistic for the additive effect is different with those of the previous three models. And the test statistic z is smaller.

Appendix 3 POE models

Appendix 3.1: POE models before transformation

a_1 and a_2 (or α_1, α_2) denote the POE, from maternal and paternal origin, respectively. We extended the usual functional (Func-Usual) model (3) to the POE functional (Func-POE) model by incorporating a POE parameter into the model, and similar steps were carried out for the usual statistical model (Stat-Usual) (5) and resulted in a POE statistical model (Stat-POE).

Model 1: POE functional (Func-POE) model:

Under the usual coding approach, the genotypic value G could be expressed as

$$G = \begin{cases} G_{11} = R \\ G_{12} = R + a_2 + d \\ G_{21} = R + a_1 + d \\ G_{22} = R + a_1 + a_2 \end{cases}. \text{(A1)}$$

That is

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = S_F E_F = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} R \\ a_1 \\ a_2 \\ d \end{pmatrix}. \text{(A2)}$$

The inverse is

$$E_{F_1} = {S_{F_1}}^{-1} G,$$

$$\begin{pmatrix} R \\ a_1 \\ a_2 \\ d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix}. \text{(A3)}$$

Simply, we could express genotypic value G on the number of the paternal or maternal reference allele, as follows:

$$G = R + N_1 a_1 + N_2 a_2 + \varepsilon d. \text{(A4)}$$

Model 2: POE statistical (Stat-POE) model:

From multiple linear regression, $\hat{\beta} = \left({X_1}^T X_1\right)^{-1} {X_1}^T y$. $\hat{\beta}$ consists of three of the four regression parameters, μ, α_1 and α_2. X_1 is a $n \times 3$ vector of N_1 and N_2 information and

$$X_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ \dots & \dots & \dots \\ 1 & 0 & 0 \end{pmatrix}. \text{(A5)}$$

y is the observed trait phenotype and $y = Z_2 * G$ when those observations perfectly fit the genotypic values in ideal situations. Therefore, we could get the expression of the three parameter in $\hat{\beta}$, as $\hat{\beta} = \left({X_1}^T X_1\right)^{-1} {X_1}^T Z_2 G$. Additionally, the dominance effect $\delta = \frac{G_{12}+G_{21}}{2} - \frac{G_{11}+G_{22}}{2}$. After

combining the coding of the additive effects and dominant effects, adjusting the coding for the intercept term μ, finally we got

$$E_{S_1} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \delta \end{pmatrix} = {S_{S_1}}^{-1} G = \begin{pmatrix} p_{11} & p_{12} & p_{21} & p_{22} \\ p'_{11} & p'_{12} & p'_{21} & p'_{22} \\ p''_{11} & p''_{12} & p''_{21} & p''_{22} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix}. \text{(A6)}$$

${S_{S_1}}^{-1}$can also be expressed as

$$\begin{pmatrix} p_{11} & p_{12} & p_{21} & p_{22} \\ \frac{-p_{11}p_{21}\bar{N}_2}{D} & \frac{-p_{12}p_{22}(1-\bar{N}_2)}{D} & \frac{p_{11}p_{21}\bar{N}_2}{D} & \frac{p_{12}p_{22}(1-\bar{N}_2)}{D} \\ \frac{-p_{11}p_{12}\bar{N}_1}{D} & \frac{p_{11}p_{12}\bar{N}_1}{D} & \frac{-p_{21}p_{22}(1-\bar{N}_1)}{D} & \frac{p_{21}p_{22}(1-\bar{N}_1)}{D} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{pmatrix}. \text{(A7)}$$

The inverse is

$$G = S_{S_1} E_{S_1}$$

It could be also expressed as

$$G = \mu + (N_1 - \bar{N}_1)\alpha_1 + (N_2 - \bar{N}_2)\alpha_2 + \varepsilon\delta, \text{(A8)}$$

which is equal to

$$\begin{pmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{pmatrix} = S_{S_1} E_{S_1} = \begin{pmatrix} 1 & -\bar{N}_1 & -\bar{N}_2 & \varepsilon_{11} \\ 1 & -\bar{N}_1 & 1-\bar{N}_2 & \varepsilon_{12} \\ 1 & 1-\bar{N}_1 & -\bar{N}_2 & \varepsilon_{21} \\ 1 & 1-\bar{N}_1 & 1-\bar{N}_2 & \varepsilon_{22} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \delta \end{pmatrix}. \text{(A9)}$$

This relation between the functional model and the statistical model parameters is

$$\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \delta \end{pmatrix} = \begin{pmatrix} 1 & \bar{N}_1 & \bar{N}_2 & p_{12} + p_{21} \\ 0 & 1 & 0 & p'_{12} + p'_{21} \\ 0 & 0 & 1 & p''_{12} + p''_{21} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ a_1 \\ a_2 \\ d \end{pmatrix}. \text{(A10)}$$

Then, we transformed these two models to two equivalent models (18) and (21) by re-parameterization using $r_1 = a_1 + a_2$, $r_2 = a_2 - a_1$ for the functional model and $\gamma_1 = \alpha_1 + \alpha_2$, $\gamma_2 = \alpha_2 - \alpha_1$ for the statistical model. These two frameworks are equivalent to some extent,

whereas the transformed one is more straightforward than this original framework for nominating the overall genetic effect and POE separately.

Appendix 3.2: Orthogonality of the Stat-POE model before transformation

In the Stat-POE model, from Equation A7, we could decompose the variance of the phenotypic value as

$$V_G = \mathrm{Var}[(N_1 - \bar{N}_1)\alpha_1] + \mathrm{Var}[(N_2 - \bar{N}_2)\alpha_2] + \mathrm{Var}(\varepsilon\delta) + 2\mathrm{Cov}[(N_1 - \bar{N}_1)\alpha_1, (N_2 - \bar{N}_2)\alpha_2], \text{(B1)}$$

as

$$\mathrm{Cov}[(N_1 - \bar{N}_1)\alpha_1, \varepsilon\delta] = \alpha_1\delta\mathrm{Cov}(N_1, \varepsilon) = 0, \text{(B2)}$$

and similarly,

$$\mathrm{Cov}[(N_2 - \bar{N}_2)\alpha_2, \varepsilon\delta] = \alpha_2\delta\mathrm{Cov}(N_2, \varepsilon) = 0. \text{(B3)}$$

Moreover, $\mathrm{Var}(\varepsilon\delta) = \delta^2\mathrm{var}(\varepsilon) = 4p_{11}p_{12}p_{21}p_{22}\delta^2/D$. Therefore, we could express the additive and dominant variance components as

$$V_\alpha = {\alpha_1}^2\mathrm{Var}(N_1) + {\alpha_2}^2\mathrm{Var}(N_2) + 2\alpha_1\alpha_2\mathrm{Cov}(N_1, N_2), \text{(B4)}$$

$$V_\delta = 4p_{11}p_{12}p_{21}p_{22}\delta^2/D. \text{(B5)}$$

To show that the additive variance, V_α, could be decomposed to be two parts which are only dependent on two additive effects ($\gamma_1$and γ_2), respectively, $\mathrm{Cov}(N_1, N_2) = 0$ needs to be satisfied. And as we know $\mathrm{Cov}(N_1, N_2) = E(N_1N_2) - E(N_1)E(N_2) = p_{22} - \bar{N}_1\bar{N}_2 = p_{11}p_{22} - p_{21}p_{12}$ which indeed equals to 0 if the locus is in HWE. In this way,

$$V_{\alpha_1} = {\alpha_1}^2\mathrm{Var}(N_1) = {\alpha_1}^2V_1, \text{(B6)}$$

$$V_{\alpha_2} = {\alpha_2}^2\mathrm{Var}(N_2) = {\alpha_2}^2V_2. \text{(B7)}$$

The two additive variance components V_{α_1} and V_{α_2} is related only to the additive effects α_1 and α_2, respectively, with one due to maternal alleles and the other due to paternal alleles. The dominant variance component V_δ is only related with the dominant effect δ, This property of the variance component to be divided into two independent additive components and one dominant component supports the notion that the POE statistical model before transformation is orthogonal.

Appendix 3.3: Orthogonality of the Stat-POE model after transformation

Except the variance component decomposition approach, the orthogonality of the Stat-POE models could also be checked by showing $X^T * X$ is a diagonal matrix in which X is the $n \times 4$ design matrix for the sample. As described in the original NOIA paper[21],

$$X^T * X = n\, S^T * D * S \text{ (C1)}$$

needs to be satisfied, where

$$D = \begin{pmatrix} p_{11} & 0 & 0 & 0 \\ 0 & p_{12} & 0 & 0 \\ 0 & 0 & p_{21} & 0 \\ 0 & 0 & 0 & p_{22} \end{pmatrix}. \text{ (C2)}$$

Given that $S = (s_{ij})$ shown in equation (5) with $s_{i1} = 1$, from (C1) and (C2) we derive the criteria for orthogonality when POE incorporated as

$$p_{11}s_{12} + p_{12}s_{22} + p_{21}s_{32} + p_{22}s_{42} = 0, \text{ (C3)}$$

$$p_{11}s_{13} + p_{12}s_{23} + p_{21}s_{33} + p_{22}s_{43} = 0, \text{ (C4)}$$

$$p_{11}s_{14} + p_{12}s_{24} + p_{21}s_{34} + p_{22}s_{44} = 0, \text{ (C5)}$$

$$p_{11}s_{12}s_{13} + p_{12}s_{22}s_{23} + p_{21}s_{32}s_{33} + p_{22}s_{42}s_{43} = 0, \text{ (C6)}$$

$$p_{11}s_{12}s_{14} + p_{12}s_{22}s_{24} + p_{21}s_{32}s_{34} + p_{22}s_{42}s_{44} = 0, \text{ (C7)}$$

$$p_{11}s_{13}s_{14} + p_{12}s_{23}s_{24} + p_{21}s_{33}s_{34} + p_{22}s_{43}s_{44} = 0. \text{ (C8)}$$

Except equation (C6), all of these criteria are satisfied by S in equation (5). And for equation (C6), $p_{11}s_{12}s_{13} + p_{12}s_{22}s_{23} + p_{21}s_{32}s_{33} + p_{22}s_{42}s_{43} = \frac{(p_{21}-p_{12})(p_{22}-p_{11})}{4} = 0$ when $p_{21} = p_{12}$ or $p_{22} = p_{11}$ holds true.

Appendix 3.4: Orthogonality of the Func-POE models

We also checked the variance decomposition of the Func-POE models before and after transformation.

For the Func-POE model before transformation, as from Equation (A4),

$$V_G = a_1{}^2\mathrm{Var}(N_1) + a_2{}^2\mathrm{Var}(N_2) + \mathrm{Var}(\varepsilon d) + a_1a_2\mathrm{Cov}(N_1, N_2) + a_1\mathrm{dCov}(N_1, \varepsilon) + a_2\mathrm{dCov}(N_2, \varepsilon), \text{ (D1)}$$

where $\mathrm{Cov}(N_1, N_2)$=0 under HWE which I already showed previously in Appendix 3.2, and

$$\mathrm{Cov}(N_1, \varepsilon) = p_{21} - (p_{21} + p_{22})(p_{12} + p_{21})$$

$$\mathrm{Cov}(N_2, \varepsilon) = p_{12} - (p_{12} + p_{22})(p_{12} + p_{21}).$$

Neither of them was equal to 0 even when the locus was under HWE.

For the Func-POE model after transformation, as from Equation (10),

$$V_G = \frac{{r_1}^2}{4}\mathrm{Var}(N_1 + N_2) + \frac{{r_2}^2}{4}\mathrm{Var}(N_2 - N_1) + \mathrm{Var}(\varepsilon d) + \frac{r_1 d}{2}\mathrm{Cov}(N_1 + N_2, \varepsilon) + \frac{r_2 d}{2}\mathrm{Cov}(N_2 - N_1, \varepsilon), \text{ (D2)}$$

where $\mathrm{Cov}(N_1 + N_2, \varepsilon) = (p_{12} + p_{21})(p_{11} - p_{22})$ is not equal to 0, and $\mathrm{Cov}(N_2 - N_1, \varepsilon) = (p_{11} + p_{22})(p_{12} - p_{21}) = 0$ if it is under HWE.

$$V_G = \frac{{r_1}^2}{4}\mathrm{Var}(N_1 + N_2) + \frac{{r_2}^2}{4}\mathrm{Var}(N_2 - N_1) + \mathrm{Var}(\varepsilon d) + \frac{r_1 d}{2}\mathrm{Cov}(N_1 + N_2, \varepsilon). \text{ (D3)}$$

Therefore, even if the HWE assumption holds, the variance of the Func-POE models could not be expressed as the completely decomposed form as could the Stat-POE models, which confirms that the Func-POE models were not orthogonal.

BIBLIOGRAPHY

1. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 2007. **447**(7145): 661-78.
2. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R; SEARCH collaborators, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schürmann P, Dörk T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X; kConFab; AOCS Management Group, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE, Cox DR, Ponder BA. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 2007. **447**(7148): 1087-93.
3. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ,

Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 2007. **316**(5826): 889-94.

4. Hardy J and Singleton A. Genomewide association studies and human disease. *N Engl J Med*, 2009. **360**(17): 1759-68.

5. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean

G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 2007. **449**(7164): 851-61.

6. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews TD, Arbury H, Attwood A, Auton A, Ball SG, Balmforth AJ, Barrett JC, Barroso I, Barton A, Bennett AJ, Bhaskar S, Blaszczyk K, Bowes J, Brand OJ, Braund PS, Bredin F, Breen G, Brown MJ, Bruce IN, Bull J, Burren OS, Burton J, Byrnes J, Caesar S, Clee CM, Coffey AJ, Connell JM, Cooper JD, Dominiczak AF, Downes K, Drummond HE, Dudakia D, Dunham A, Ebbs B, Eccles D, Edkins S, Edwards C, Elliot A, Emery P, Evans DM, Evans G, Eyre S, Farmer A, Ferrier IN, Feuk L, Fitzgerald T, Flynn E, Forbes A, Forty L, Franklyn JA, Freathy RM, Gibbs P, Gilbert P, Gokumen O, Gordon-Smith K, Gray E, Green E, Groves CJ, Grozeva D, Gwilliam R, Hall A, Hammond N, Hardy M, Harrison P, Hassanali N, Hebaishi H, Hines S, Hinks A, Hitman GA, Hocking

L, Howard E, Howard P, Howson JM, Hughes D, Hunt S, Isaacs JD, Jain M, Jewell DP, Johnson T, Jolley JD, Jones IR, Jones LA, Kirov G, Langford CF, Lango-Allen H, Lathrop GM, Lee J, Lee KL, Lees C, Lewis K, Lindgren CM, Maisuria-Armer M, Maller J, Mansfield J, Martin P, Massey DC, McArdle WL, McGuffin P, McLay KE, Mentzer A, Mimmack ML, Morgan AE, Morris AP, Mowat C, Myers S, Newman W, Nimmo ER, O'Donovan MC, Onipinla A, Onyiah I, Ovington NR, Owen MJ, Palin K, Parnell K, Pernet D, Perry JR, Phillips A, Pinto D, Prescott NJ, Prokopenko I, Quail MA, Rafelt S, Rayner NW, Redon R, Reid DM, Renwick, Ring SM, Robertson N, Russell E, St Clair D, Sambrook JG, Sanderson JD, Schuilenburg H, Scott CE, Scott R, Seal S, Shaw-Hawkins S, Shields BM, Simmonds MJ, Smyth DJ, Somaskantharajah E, Spanova K, Steer S, Stephens J, Stevens HE, Stone MA, Su Z, Symmons DP, Thompson JR, Thomson W, Travers ME, Turnbull C, Valsesia A, Walker M, Walker NM, Wallace C, Warren-Perry M, Watkins NA, Webster J, Weedon MN, Wilson AG, Woodburn M, Wordsworth BP, Young AH, Zeggini E, Carter NP, Frayling TM, Lee C, McVean G, Munroe PB, Palotie A, Sawcer SJ, Scherer SW, Strachan DP, Tyler-Smith C, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Gough SC, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand W, Parkes M, Rahman N, Todd JA, Samani NJ, Donnelly P. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 2010. **464**(7289): 713-20.

7. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Smith AV, Tobin MD, Verwoert GC, Hwang SJ, Pihur V, Vollenweider P, O'Reilly PF, Amin N, Bragg-Gresham JL, Teumer A, Glazer NL, Launer L, Zhao JH, Aulchenko Y, Heath S, Sõber S, Parsa A, Luan J, Arora P, Dehghan A, Zhang F, Lucas G, Hicks AA, Jackson AU, Peden JF, Tanaka T, Wild SH, Rudan I, Igl W, Milaneschi Y, Parker AN, Fava C, Chambers JC, Fox ER, Kumari M, Go MJ, van der Harst P, Kao WH, Sjögren M, Vinay DG, Alexander M,

Tabara Y, Shaw-Hawkins S, Whincup PH, Liu Y, Shi G, Kuusisto J, Tayo B, Seielstad M, Sim X, Nguyen KD, Lehtimäki T, Matullo G, Wu Y, Gaunt TR, Onland-Moret NC, Cooper MN, Platou CG, Org E, Hardy R, Dahgam S, Palmen J, Vitart V, Braund PS, Kuznetsova T, Uiterwaal CS, Adeyemo A, Palmas W, Campbell H, Ludwig B, Tomaszewski M, Tzoulaki I, Palmer ND; CARDIoGRAM consortium; CKDGen Consortium; KidneyGen Consortium; EchoGen consortium; CHARGE-HF consortium, Aspelund T, Garcia M, Chang YP, O'Connell JR, Steinle NI, Grobbee DE, Arking DE, Kardia SL, Morrison AC, Hernandez D, Najjar S, McArdle WL, Hadley D, Brown MJ, Connell JM, Hingorani AD, Day IN, Lawlor DA, Beilby JP, Lawrence RW, Clarke R, Hopewell JC, Ongen H, Dreisbach AW, Li Y, Young JH, Bis JC, Kähönen M, Viikari J, Adair LS, Lee NR, Chen MH, Olden M, Pattaro C, Bolton JA, Köttgen A, Bergmann S, Mooser V, Chaturvedi N, Frayling TM, Islam M, Jafar TH, Erdmann J, Kulkarni SR, Bornstein SR, Grässler J, Groop L, Voight BF, Kettunen J, Howard P, Taylor A, Guarrera S, Ricceri F, Emilsson V, Plump A, Barroso I, Khaw KT, Weder AB, Hunt SC, Sun YV, Bergman RN, Collins FS, Bonnycastle LL, Scott LJ, Stringham HM, Peltonen L, Perola M, Vartiainen E, Brand SM, Staessen JA, Wang TJ, Burton PR, Soler Artigas M, Dong Y, Snieder H, Wang X, Zhu H, Lohman KK, Rudock ME, Heckbert SR, Smith NL, Wiggins KL, Doumatey A, Shriner D, Veldre G, Viigimaa M, Kinra S, Prabhakaran D, Tripathy V, Langefeld CD, Rosengren A, Thelle DS, Corsi AM, Singleton A, Forrester T, Hilton G, McKenzie CA, Salako T, Iwai N, Kita Y, Ogihara T, Ohkubo T, Okamura T, Ueshima H, Umemura S, Eyheramendy S, Meitinger T, Wichmann HE, Cho YS, Kim HL, Lee JY, Scott J, Sehmi JS, Zhang W, Hedblad B, Nilsson P, Smith GD, Wong A, Narisu N, Stančáková A, Raffel LJ, Yao J, Kathiresan S, O'Donnell CJ, Schwartz SM, Ikram MA, Longstreth WT Jr, Mosley TH, Seshadri S, Shrine NR, Wain LV, Morken MA, Swift AJ, Laitinen J, Prokopenko I, Zitting P, Cooper JA, Humphries SE, Danesh J, Rasheed A, Goel A, Hamsten A, Watkins H, Bakker SJ, van Gilst WH, Janipalli CS, Mani KR, Yajnik CS, Hofman A, Mattace-Raso FU, Oostra BA, Demirkan A, Isaacs A,

Rivadeneira F, Lakatta EG, Orru M, Scuteri A, Ala-Korpela M, Kangas AJ, Lyytikäinen LP, Soininen P, Tukiainen T, Würtz P, Ong RT, Dörr M, Kroemer HK, Völker U, Völzke H, Galan P, Hercberg S, Lathrop M, Zelenika D, Deloukas P, Mangino M, Spector TD, Zhai G, Meschia JF, Nalls MA, Sharma P, Terzic J, Kumar MV, Denniff M, Zukowska-Szczechowska E, Wagenknecht LE, Fowkes FG, Charchar FJ, Schwarz PE, Hayward C, Guo X, Rotimi C, Bots ML, Brand E, Samani NJ, Polasek O, Talmud PJ, Nyberg F, Kuh D, Laan M, Hveem K, Palmer LJ, van der Schouw YT, Casas JP, Mohlke KL, Vineis P, Raitakari O, Ganesh SK, Wong TY, Tai ES, Cooper RS, Laakso M, Rao DC, Harris TB, Morris RW, Dominiczak AF, Kivimaki M, Marmot MG, Miki T, Saleheen D, Chandak GR, Coresh J, Navis G, Salomaa V, Han BG, Zhu X, Kooner JS, Melander O, Ridker PM, Bandinelli S, Gyllensten UB, Wright AF, Wilson JF, Ferrucci L, Farrall M, Tuomilehto J, Pramstaller PP, Elosua R, Soranzo N, Sijbrands EJ, Altshuler D, Loos RJ, Shuldiner AR, Gieger C, Meneton P, Uitterlinden AG, Wareham NJ, Gudnason V, Rotter JI, Rettig R, Uda M, Strachan DP, Witteman JC, Hartikainen AL, Beckmann JS, Boerwinkle E, Vasan RS, Boehnke M, Larson MG, Järvelin MR, Psaty BM, Abecasis GR, Chakravarti A, Elliott P, van Duijn CM, Newton-Cheh C, Levy D, Caulfield MJ, Johnson T. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 2011. **478**(7367): 103-9.

8. Amos CI, Wang LE, Lee JE, Gershenwald JE, Chen WV, Fang S, Kosoy R, Zhang M, Qureshi AA, Vattathil S, Schacherer CW, Gardner JM, Wang Y, Bishop DT, Barrett JH; GenoMEL Investigators, MacGregor S, Hayward NK, Martin NG, Duffy DL; Q-Mega Investigators, Mann GJ, Cust A, Hopper J; AMFS Investigators, Brown KM, Grimm EA, Xu Y, Han Y, Jing K, McHugh C, Laurie CC, Doheny KF, Pugh EW, Seldin MF, Han J, Wei Q. Genome-wide association study identifies novel loci predisposing to cutaneous melanoma. *Hum Mol Genet*, 2011. **20**(24): 5012-23.

9. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*, 2011. **43**(6): 519-25.

10. Yang J, Ferreira T, Morris AP, Medland SE; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, Frayling TM, McCarthy MI, Hirschhorn JN, Goddard ME, Visscher PM. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*, 2012. **44**(4): 369-75, S1-3.

11. Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet*, 2002. **70**(2): 461-71.

12. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered*, 2003. **56**(1-3): 73-82.

13. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*, 2009. **461**(7265): 747-53.

14. Xu Y, Peng B, Fu Y, Amos CI. Genome-wide algorithm for detecting CNV associations with diseases. *BMC Bioinformatics*, 2011. **12**: 331.

15. Liu DJ and Leal SM. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. Am J Hum Genet, 2012. **91**(4): 585-96.

16. Gorlova OY, Lei L, Zhu D, Weng SF, Shete S, Zhang Y, Li WD, Price RA, Amos CI. Imprinting detection by extending a regression-based QTL analysis method. *Hum Genet*, 2007. **122**(2): 159-74.

17. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 2001. **69**(1): 124-37.

18. McCarthy MI and Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet*, 2008. **17**(R2): R156-65.

19. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 2008. **9**(5): 356-69.

20. Zeng ZB, Wang T, and Zou W. Modeling quantitative trait Loci and interpretation of models. *Genetics*, 2005. **169**(3): 1711-25.

21. Alvarez-Castro JM and Carlborg O. A unified model for functional and statistical epistasis and its application in quantitative trait Loci analysis. *Genetics*, 2007. **176**(2): 1151-67.

22. Ma J, Xiao F, Xiong M, Andrew AS, Brenner H, Duell EJ, Haugen A, Hoggart C, Hung RJ, Lazarus P, Liu C, Matsuo K, Mayordomo JI, Schwartz AG, Staratschek-Jox A, Wichmann E, Yang P, Amos CI. Natural and orthogonal interaction framework for modeling gene-environment interactions with application to lung cancer. *Hum Hered*, 2012. **73**(4): 185-94.

23. Alvarez-Castro JM, Le Rouzic A, and Carlborg O. How to perform meaningful estimates of genetic effects. *PLoS Genet*, 2008. **4**(5): e1000062.

24. Nagel RL. Pleiotropic and epistatic effects in sickle cell anemia. *Curr Opin Hematol.*, 2001. 8(2):105-10.

25. Wandstrat A and Wakeland E. The genetics of complex autoimmune diseases: non-MHC susceptibility genes. *Nat Immunol.*, 2001. 2(9):802-9.

26. Nagel RL. Epistasis and the genetics of human diseases. *C R Biol*, 2005. **328**(7): 606-15.

27. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. *Nature*, 2012. 490(7421):535-8.

28. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 2001. **69**(1): 138-47.

29. Hahn LW, Ritchie MD, and Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 2003. **19**(3): 376-82.

30. Chung Y, Lee SY, Elston RC, Park T. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics*, 2007. **23**(1): 71-6.

31. Gayán J, González-Pérez A, Bermudo F, Sáez ME, Royo JL, Quintas A, Galan JJ, Morón FJ, Ramirez-Lorca R, Real LM, Ruiz A. A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genomics*, 2008. **9**: p. 360.

32. Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C. Multivariate detection of gene-gene interactions. *Genet Epidemiol*, 2012. **36**(6): 622-30.

33. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*, 2007. **63**(2): 111-9.

34. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, 2003. **32**(1): p. 1-22.

35. Wei S, Wang LE, McHugh MK, Han Y, Xiong M, Amos CI, Spitz MR, Wei QW. Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. *Carcinogenesis*, 2012. **33**(8): 1531-7.

36. Chung SJ, Armasu SM, Anderson KJ, Biernacka JM, Lesnick TG, Rider DN, Cunningham JM, Ahlskog JE, Frigerio R, Maraganore DM. Genetic susceptibility loci, environmental exposures, and Parkinson's disease: A case-control study of gene-environment interactions. *Parkinsonism Relat Disord*, 2013.

37. Karlson EW, Ding B, Keenan BT, Liao K, Costenbader KH, Klareskog L, Alfredsson L, Chibnik LB. Association of environmental and genetic factors and gene-environment interactions with risk of developing rheumatoid arthritis. *Arthritis Care Res (Hoboken)*, 2013.

38. Wang J, Spitz MR, Amos CI, Wu X, Wetter DW, Cinciripini PM, Shete S. Method for evaluating multiple mediators: mediating effects of smoking and COPD on the association between the CHRNA5-A3 variant and lung cancer risk. *PLoS One*, 2012. **7**(10): e47705.

39. Chatterjee N, Kalaylioglu Z, and Carroll RJ. Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. *Genet Epidemiol*, 2005. 28(2): 138-56.

40. Mukherjee B and Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 2008. **64**(3): 685-94.

41. Li Y and Graubard BI. Pseudo semiparametric maximum likelihood estimation exploiting gene environment independence for population-based case-control studies with complex samples. *Biostatistics*, 2012. **13**(4): 711-23.

42. Reik W, Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet.*, 2001. 2(1):21-32.

43. Barlow DP. Gametic imprinting in mammals. *Science*, 1995. 270(5242):1610-3.

44. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Jonasdottir A, Sigurdsson A, Kristinsson KT, Jonasdottir A, Frigge ML, Gylfason A, Olason PI, Gudjonsson SA, Sverrisson S, Stacey SN, Sigurgeirsson B, Benediktsdottir KR, Sigurdsson H, Jonsson T, Benediktsson R, Olafsson JH, Johannsson OT, Hreidarsson AB, Sigurdsson G; DIAGRAM Consortium, Ferguson-Smith AC, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K. Parental origin of sequence variants associated with complex diseases. *Nature* , 2009: 462(7275):868-74.

45. Wyszynski DF, Panhuysen CI. Parental sex effect in families with alcoholism. *Genet Epidemiol*, 1999. 17 Suppl 1:S409-13.

46. Dong C, Li WD, Geller F, Lei L, Li D, Gorlova OY, Hebebrand J, Amos CI, Nicholls RD, Price RA. Possible genomic imprinting of three human obesity-related genetic loci. Am J *Hum Genet*, 2005. 76(3):427-37.

47. Guo YF, Shen H, Liu YJ, Wang W, Xiong DH, Xiao P, Liu YZ, Zhao LJ, Recker RR, Deng HW. Assessment of genetic linkage and parent-of-origin effects on obesity. *J Clin Endocrinol Metab.*, 2006. 91(10):4001-5.

48. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Jonasdottir A, Sigurdsson A, Kristinsson KT, Jonasdottir A, Frigge ML, Gylfason A, Olason PI, Gudjonsson SA, Sverrisson S, Stacey SN, Sigurgeirsson B, Benediktsdottir KR, Sigurdsson H, Jonsson T, Benediktsson R, Olafsson JH, Johannsson OT, Hreidarsson AB, Sigurdsson G; DIAGRAM Consortium, Ferguson-Smith AC, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K. Parental origin of sequence variants associated with complex diseases. *Nature*, 2009. **462**(7275): 868-74.

49. Wallace C, Smyth DJ, Maisuria-Armer M, Walker NM, Todd JA, Clayton DG. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat Genet*, 2010. **42**(1): 68-71.

50. Shete S and Amos CI, Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *Am J Hum Genet*, 2002. **70**(3): p. 751-7.

51. Zhou X, Chen W, Swartz MD, Lu Y, Yu R, Amos CI, Wu CC, Shete S. Joint linkage and imprinting analyses of GAW15 rheumatoid arthritis and gene expression data. *BMC Proc*, 2007. **1**(Suppl 1): S53.

52. Ainsworth HF, Unwin J, Jamison DL, Cordell HJ. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet Epidemiol*, 2011. **35**(1): 19-45.

53. Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, Randerson-Moor J, Aitken JF, Avril MF, Azizi E, Bakker B, Bianchi-Scarrà G, Bressac-de Paillerets B, Calista D, Cannon-Albright LA, Chin-A-Woeng T, Debniak T, Galore-Haskel G, Ghiorzo P, Gut I, Hansson J, Hocevar M, Höiom V, Hopper JL, Ingvar C, Kanetsky PA, Kefford RF, Landi MT, Lang J, Lubiński J, Mackie R, Malvehy J, Mann GJ, Martin NG, Montgomery GW, van Nieuwpoort FA, Novakovic S, Olsson H, Puig S, Weiss M, van Workum W, Zelenika D, Brown KM, Goldstein AM, Gillanders EM, Boland A, Galan P, Elder DE, Gruis NA, Hayward NK, Lathrop GM, Barrett JH, Bishop JA. Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet*, 2009. **41**(8): 920-5.

54. Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, Martin NG, Montgomery GW. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am J Hum Genet*, 2008. **82**(2): 424-31.

55. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 2009. **10**(6): 392-404.

56. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-de-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcátová I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, 2008. **452**(7187): 633-7.

57. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, McLaughlin J, Shepherd F, Montpetit A, Narod S, Krokan HE, Skorpen F, Elvestad MB, Vatten L, Njølstad I, Axelsson T, Chen C, Goodman G, Barnett M, Loomis MM, Lubiński J, Matyjasik J, Lener M, Oszutowska D, Field J, Liloglou T, Xinarianos G, Cassidy A; EPIC Study, Vineis P, Clavel-Chapelon F, Palli D, Tumino R, Krogh V, Panico S, González CA, Ramón Quirós J, Martínez C, Navarro C, Ardanaz E, Larrañaga N, Kham KT, Key T, Bueno-de-Mesquita HB, Peeters PH, Trichopoulou A, Linseisen J, Boeing H, Hallmans G, Overvad K, Tjønneland A, Kumle M, Riboli E, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*, 2008. **40**(12): 1404-6.

58. Wang Y, Broderick P, Webb E, Wu X, Vijayakrishnan J, Matakidou A, Qureshi M, Dong Q, Gu X, Chen WV, Spitz MR, Eisen T, Amos CI, Houlston RS. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*, 2008. **40**(12): 1407-9.

59. Truong T, Hung RJ, Amos CI, Wu X, Bickeböller H, Rosenberger A, Sauter W, Illig T, Wichmann HE, Risch A, Dienemann H, Kaaks R, Yang P, Jiang R, Wiencke JK, Wrensch M, Hansen H, Kelsey KT, Matsuo K, Tajima K, Schwartz AG, Wenzlaff A, Seow A, Ying C, Staratschek-Jox A, Nürnberg P, Stoelben E, Wolf J, Lazarus P, Muscat JE, Gallagher CJ, Zienolddiny S, Haugen A, van der Heijden HF, Kiemeney LA, Isla D, Mayordomo JI, Rafnar T, Stefansson K, Zhang ZF, Chang SC, Kim JH, Hong YC, Duell EJ, Andrew AS, Lejbkowicz F, Rennert G, Müller H, Brenner H, Le Marchand L, Benhamou S, Bouchardy C, Teare MD, Xue X, McLaughlin J, Liu G, McKay JD, Brennan P, Spitz MR. 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**(13): 959-71.

60. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions: a review. *Appl Bioinformatics*, 2006. **5**(2): 77-88.

61. Maher B. Personal genomes: The case of the missing heritability. *Nature*, 2008. **456**(7218): 18-21.

62. Belonogova NM, Axenovich TI, Aulchenko YS. A powerful genome-wide feasible approach to detect parent-of-origin effects in studies of quantitative traits. *Eur J Hum Genet*, 2010. **18**(3): 379-84.

63. Feng R, Wu Y, Jang GH, Ordovas JM, Arnett D. A powerful test of parent-of-origin effects for quantitative traits using haplotypes. *PLoS One*., 2011. 6(12):e28909.

VITA

Feifei Xiao, the daughter of Wenming Xiao and Hongying Liu, was born on July 18th, 1985 in Songzi, Hubei, People's Republic of China. She graduated from Songzi No.1 High School in 2002. Then she attended Wuhan University in Wuhan, Hubei, China and received her Bachelor Degree of Science in Biotechnology in 2006. Then from the same school, she received her Master Degree of Science in Microbiology in 2009. In August 2009, she entered the Graduate School of Biomedical Sciences, The University of Texas Health Science Center at Houston and The University of Texas M.D. Anderson Cancer Center. She expects to get her Doctor of Philosophy Degree in Biostatistics in May 2013.