Published in final edited form as:
Stat Med. 2012 September 28; 31(22): 2565–2568. doi:10.1002/sim.5495.

A Discussion of Gene-Gene and Gene-Environment Interactions and Longitudinal Genetic Analysis of Complex Traits

Ruzong Fan, **Paul S. Albert**, and **Enrique F. Schisterman**
6100 Executive Blvd, Room 7B05, MSC 7510, Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics, and Prevention, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD 20852

Background

In the last decades, there has been a lot effort on genome-wide association study (GWAS) of complex diseases and the widespread availability of high-throughput genotyping technology has made large scale GWAS possible. In many of the genetic association studies, the strategy is to scan entire genome of millions of single nucleotide polymorphisms (SNPs) for single SNP associations. Although GWAS seemed to be promising and the resources already expended, the discoveries of GWAS have been underwhelming. GWAS does not help a lot to elucidate the genetic architecture of complex diseases and it has not been able to explain a majority of the genetic risks. Our knowledge of the genetic architecture of complex diseases is still very limited. The identification and characterization of susceptibility genes of common complex human diseases remains a great challenge for human geneticists, genetic epidemiologists, and statisticians. This is partly because of the complex and non-linear relationship between genotype, environment factors, and phenotype. These complex relationships are believed to be a consequence of temporal dynamical interaction between the genetic determinants and the environmental factors. That is, the gene-gene and gene-environment interactions are time or age dependent and the temporal dynamical interactions need to be properly modelled [1, 2].

Due to the temporal dynamical nature, the interactions lead to phenotype changes over time. A trait that varies as a function of a continuous variable, such as age, is referred to as a function-valued trait. Longitudinal data that are collected on a certain number of occasions for each individual can be used to study function-valued traits, i.e., the repeated measures data. Genetic studies of function-valued traits are a natural extension of classical genetics, and are powerful tools for identifying the complex genetic structures of common diseases and for discovering strategies for disease prevention. Longitudinal genetic analysis of function-valued traits should be an integral part of genetic theory. The understanding of temporal trends can provide insights about efficient designs to detect genetic and environment effects, gene-gene, and gene-environment interactions of complex traits. It is important to develop statistical models and methods that make better use of the longitudinal genetic data, that may reflect temporal trends and familial structure simultaneously, and that may explain gene and environment effects, as well as gene-gene and gene-environment interactions.

In literature, most research focuses on interactions rather than the temporal trends. The existing methods are mainly for analysis of one phenotype an individual, and little research is available for multiple phenotype measurement data. There has been great enthusiasm to

Ruzong Fan: fanr@mail.nih.gov, Paul Albert: albertp@mail.nih.gov, and Enrique Schisterman: schistee@mail.nih.gov.

detect and to characterize gene-gene and gene-environment interactions of complex diseases in recent years [3-6]. Two interesting papers contribute to the topics in the framework of logistic regression. The first one discusses efficient designs for detecting gene-environment interactions [7], while the second develops sensitivity analysis to assess gene-environment interactions in the presence of an unmeasured confounder [8]. Mukherjee et al [9] proposes methodology for the analysis of longitudinal genetic data, a topic that has not received much attention in the literature and that needs more intensive and in depth research.

Gene-Gene and Gene-Environment Interactions

Based on analytical formulae of the regression coefficients and simulation studies, Chen et al. [7] assesses the genetic or environment effects in the presence of gene-environment interactions. The work suggests that one should supplement genetic data from the controls if one wants to test the genetic effect and to supplement environment data from the additional controls if one wants to assess the environment effect in the presence of gene-environment interactions. This work is both novel and intuitive. The paper does not explicitly discuss the detection of gene-gene and gene-environment interactions in the presence or absence of main genetic and environment effects. It is well-known that the power to detect the interactions can be low even the main genetic and environment effects are present [2]. It would be interesting to see more research and debate about novel methods to detect these interactions.

Vanderweele et al. [8] performs an interesting sensitivity analysis to assess the impact of unmeasured confounding for gene-environment interactions. Under an additive and multiplicative model, bias formulae are provided for sensitivity analysis of interactions when one or more unmeasured confounding factors exist. This is unique work since most research has focused on sensitivity analysis for the main genetic and environment effects and there is little research that investigates the sensitivity of the interactions. Interestingly, it shows under what condition unmeasured confounding does not induce bias at all. We believe that this paper will stimulate further interest and investigation on the issue, especially the utilization of directed acyclic graphs. An interesting question is how unmeasured confounding affects the power to detect the interactions.

Methodology of detecting interactions (gene-gene and gene-environment) continues to pose important and exciting problems for methodologists. To use the traditional linear and log-linear regressions, one needs to detect the main genetic and environment effects before detecting the statistical interaction [10]. In a biological sense, the interactions may exist in the absence of the main effects [1]. This may lead to complex and non-linearity analysis. In fact, nonlinear methods have received some attention in the genetics community. For instance, a non-parametric multifactor dimensionality reduction (MDR) is proposed and user-friendly algorithms and software are available [11-15]. Although MDR has been criticized by some statisticians, it is a popular and useful technique for genetics community [16]. In Park and Hastie [16], logistic regression was used to model the main effects and interactions; but it does not model non-linear relations between genotypes and phenotypes. In the absence of main effects, logistic regressions are unlikely to be effective. For instance, the logistic regression models failed to converge due to the sparse nature of bladder cancer data in Andrew et al. [17].

To model the non-linear relations, we need new and novel approaches. It is well-known that information theory based on entropy function is widely used to study nonlinear problems and complex system [18]. The entropy function is a non-linear transformation of interested variables. The entropy is commonly used in information theory to measure the uncertainty of random variables [18]. The entropy-based approach is likely to be very useful to study the

nonlinear relationship between genotypes, environmental-factors, and phenotypes, and to interpret the gene-gene and gene-environment interactions of diseases. In literature, there are some good references in recent years [19-21]. More research is necessary and welcome to advance understanding of these biological mechanisms.

Longitudinal Genetic Analysis

Although longitudinal genetic analysis is very important, the statistical models of function-valued genetic traits have not been well developed. For instance, there is no combined linkage and association mapping methods of longitudinal phenotypic data. Due to the lack of handy software and algorithms and novel statistical models, investigators of Framingham Heart Study collapsed the multiple measurements of an individual to be a single value by taking sample average and performed classical linkage analysis [22]. Ten years later, there is still little progress in developing novel statistical methods in the area. In Mukherjee et al. [9], the simple approach of averaging multiple response measurements of the same individual was proposed to analyze the longitudinal genetic traits. The phenotype traits are usually varying with age, and so there are temporal effects and variations. Besides, if the environment covariates such as body mass index are measured several times as the phenotypes, the gene-environment interactions can be time-dependent. After collapsing the multiple measurements to be a single value, no temporal effects and variations and time-dependent interactions can be detected in an analysis and the power can be low [23]. This type of analysis may not ideally make the best use of the available longitudinal data.

Although there is not much progress in terms of developing statistical methodology to analyze longitudinal genetic traits, there are some efforts to develop novel techniques to tackle the problem. de Andrade et al. [24, 25] extended variance components approach to incorporate temporal trends and longitudinal pedigree data analysis, but the methods may have a lot of parameters since the number of variance-covariance terms grow rapidly with the number of longitudinal measurements. Zhang and Zhong [26] proposed parametric variance component models for linkage analysis of longitudinal data, but it may not handle the temporal trend effectively. Wang et al. [27] proposed semi-parametric association models of longitudinal traits by reduced rank smoothing, which can deal with the temporal trend well, but it does no model the linkage information and no handy software is available to implement method.

To successfully model longitudinal genetic traits, we need to carefully parameterize the models so that they do not depend on the number of multiple measurements of a subject. Unlike de Andrade et al. [24, 25], the number of parameters should be fixed after carefully specifying regression models and the variance-covariance structure if variance component models are used. In model development, one should also consider the temporal trends and family structures. It is important to develop a general framework for longitudinal analysis of function-valued traits that can be quantitative or qualitative in either parametric or non-parametric framework. Algorithms and user-friendly software should be natural results of the research. Applications to real and simulated data will be necessary to investigate the properties of these methods for estimating the temporal trend of longitudinal genetic traits as well as to investigate statistical tests for gene-gene and gene-environment interactions involving these longitudinal genetic traits.

Conclusions

The gene-gene and gene-environment interactions and longitudinal genetic analysis are important and interesting research areas. In the future studies, the temporal trends and interactions need to be handled properly. One approach is to handle them simultaneously.

Further insight into the differences between statistical and biological interactions would be useful. As always, effective statistical methods and user-friendly algorithms and software are keys to successfully analyze the longitudinal genetic traits and genetic interactions.

So far, many studies use standard statistical methods to analyze genetic data. One advantage of using traditional statistical models to analyze genetic data is that the related theory is very mature and the user-friendly software is available. For instance, linear regressions, ANOVA, and log-linear models are standard procedure in SAS and R for data analysis and model selection. However, the available statistical software may not be able to effectively detect temporal trends of genetic traits, gene-gene, and gene-environment interactions. The traditional statistical methods are usually designed to detect the main effects as the first step in an analysis, and then to detect the statistical interactions [10]. Because of the complexity and nonlinearity between complex traits, genetic determinants, and environmental factors, the main effects may not be significantly present but the biological interactions may still exist[1]. The traditional statistical models may not properly fit the nonlinear relationship between genotypes, environmental-factors, and phenotypes in the absence of main effects, and may not be necessarily useful to model biological interactions efficiently.

Acknowledgments

This study was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Maryland, USA.

Reference List

1. Bateson, W. Mendel's Principles of Heredity. Cambridge University Press; Cambridge: 1909.
2. Moore JH, Williams SM. Epistasis and its implications for personal genetics. Am.J.Hum.Genet. 2009; 85(3):309–320. [PubMed: 19733727]
3. Frankel WN, Schork NJ. Who's afraid of epistasis? Nat.Genet. 1996; 14(4):371–373. [PubMed: 8944011]
4. Mahdi H, Fisher BA, Kallberg H, Plant D, Malmstrom V, Ronnelid J, Charles P, Ding B, Alfredsson L, Padyukov L, Symmons DP, Venables PJ, Klareskog L, Lundberg K. Specific interaction between genotype, smoking and autoimmunity to citrullinated alpha-enolase in the etiology of rheumatoid arthritis. Nat.Genet. 2009; 41(12):1319–1324. [PubMed: 19898480]
5. Mechanic LE, Chen HS, Amos CI, Chatterjee N, Cox NJ, Divi RL, Fan R, Harris EL, Jacobs K, Kraft P, Leal SM, McAllister K, Moore JH, Paltoo DN, Province MA, Ramos EM, Ritchie MD, Roeder K, Schaid DJ, Stephens M, Thomas DC, Weinberg CR, Witte JS, Zhang S, Zollner S, Feuer EJ, Gillanders EM. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. Genet.Epidemiol. 2011
6. van der Woude D, Alemayehu WG, Verduijn W, de Vries RR, Houwing-Duistermaat JJ, Huizinga TW, Toes RE. Gene-environment interaction influences the reactivity of autoantibodies to citrullinated antigens in rheumatoid arthritis. Nat.Genet. 2010; 42(10):814–816. [PubMed: 20877316]
7. Chen J, Kang G, Vanderweele T, Zhang C, B Mukherjee B. Efficient designs of gene-environment interactions studies: implications of Hardy-Weinberg equilibrium and geneenvironment independence. 2012
8. Vanderweele T, Mukherjee B, Chen JB. Sensitivity analysis for interactions under unmeasured confounding. 2012
9. Mukherjee B, Ko Y, Vanderweele T, Roy A, Park SK, Chen JB. Principal interactions analysis for repeated measures data: application to gene-gene, gene-environment interactions. 2012
10. Fisher RA. The correlations between relatives on the supposition of Mendelian inheritance. Trans Royal Soc Edinburgh. 1918; 52:399–433.

11. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003; 19(3):376–382. [PubMed: 12584123]
12. Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J.Theor.Biol. 2006; 241(2):252–261. [PubMed: 16457852]
13. 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):138–147. [PubMed: 11404819]
14. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC.Bioinformatics. 2003; 4:28. [PubMed: 12846935]
15. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet.Epidemiol. 2003; 24(2):150–157. [PubMed: 12548676]
16. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics. 2008; 9(1):30–50. [PubMed: 17429103]
17. Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, Tosteson TD, Schned AR, Karagas MR. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. Carcinogenesis. 2006; 27(5):1030–1037. [PubMed: 16311243]
18. Shannon CE. A mathematical theory of communications. The Bell System Technical Journal. 1948; XXVII:379–423. 623–656.
19. Dong C, Chu X, Wang Y, Wang Y, Jin L, Shi T, Huang W, Li Y. Exploration of genegene interaction effects using entropy-based methods. Eur.J.Hum.Genet. 2008; 16(2):229–235. [PubMed: 17971837]
20. Fan R, Zhong M, Wang S, Zhang Y, Andrew A, Karagas M, Chen H, Amos CI, Xiong M, Moore JH. Entropy-based information gain approaches to detect and to characterize genegene and gene-environment interactions/correlations of complex diseases. Genet.Epidemiol. 2011; 35(7):706–721. [PubMed: 22009792]
21. Kang G, Yue W, Zhang J, Cui Y, Zuo Y, Zhang D. An entropy-based approach for testing genetic epistasis underlying complex diseases. J.Theor.Biol. 2008; 250(2):362–374. [PubMed: 17996908]
22. Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. Hypertension. 2000; 36(4):477–483. [PubMed: 11040222]
23. Shi G, Rao DC. Ignoring temporal trends in genetic effects substantially reduces power of quantitative trait linkage analysis. Genet.Epidemiol. 2008; 32(1):61–72. [PubMed: 17703462]
24. de Andrade M, Gueguen R, Visvikis S, Sass C, Siest G, Amos CI. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet.Epidemiol. 2002; 22(3):221–232. [PubMed: 11921082]
25. de Andrade M, Olswold C. Comparison of longitudinal variance components and regression-based approaches for linkage detection on chromosome 17 for systolic blood pressure. BMC.Genet. 2003; 4(Suppl 1):S17. [PubMed: 14975085]
26. Zhang H, Zhong X. Linkage analysis of longitudinal data and design consideration. BMC.Genet. 2006; 7:37. [PubMed: 16768806]
27. Wang Y, Huang C, Fang Y, Yang Q, Li R. Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing. Applied Statistics. 2012; 61:1–23. [PubMed: 22581986]