

**TOPICS IN STATISTICAL METHODS  
FOR HUMAN GENE MAPPING**

by

Chia-Ling Kuo

MS, Biostatistics, National Taiwan University, Taipei, Taiwan, 2003

BBA, Statistics, National Chengchi University, Taipei, Taiwan, 2001

Submitted to the Graduate Faculty of  
Graduate School of Public Health in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

**Chia-Ling Kuo**

It was defended on

**April 16, 2009**

and approved by

Dissertation Advisor: **Eleanor Feingold**, Ph.D., Professor, Depts. of Human Genetics and  
Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: **Daniel E. Weeks**, Ph.D., Professor, Depts. of Human Genetics and  
Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: **Michael M. Barmada**, Ph.D., Associate Professor, Department of  
Human Genetics, Graduate School of Public Health, University of Pittsburgh

Committee Member: **George C. Tseng**, Ph.D., Associate Professor, Depts. of Biostatistics  
and Human Genetics, Graduate School of Public Health, University of Pittsburgh

Copyright © by Chia-Ling Kuo

2010

**TOPICS IN STATISTICAL METHODS FOR HUMAN GENE MAPPING**

Chia-Ling Kuo, PhD

University of Pittsburgh, 2010

Statistical approaches used for gene mapping can be divided into two types: linkage and association analysis. This dissertation work addresses statistical methods in both areas.

In the area of linkage analysis, I consider the problem of QTL (Quantitative Trait Locus) linkage analysis. Linkage analysis requires family data, and if the families are selected according to phenotype or if the trait of interest has a non-Gaussian distribution, standard analysis methods may be inappropriate. The score statistic, derived by taking the first derivative of the likelihood with respect to the linkage parameter, maintains the power of likelihood-based methods and with the use of an empirical variance estimator is robust against non-normal traits and selected samples. I investigate a number of empirical variance estimators that can be used for general pedigrees and evaluate the effects of different variance estimators and trait parameter estimates on the power of the score statistic.

In the area of association analysis, I consider the question of what is the best model for a simple genome-scan analysis of a case-control study. In a case-control genome-wide association study, hundreds of thousands of SNPs are genotyped and statistical analysis usually starts with 1 or 2 df chi-squared test or logistic regression model. Power comparisons among subsets of these methods have been done but none of these papers have comprehensively tackled the question of which method is best for univariate scanning in a genome scan. I compare different test

procedures and regression models for case-control studies starting from single-locus analysis followed by scanning with covariates and then genome-wide analysis. Based on the simulation results, I offer guidelines for choosing robust test procedures or regression models for testing the genetic effect.

The methods proposed here can be used to improve the efficiency of gene mapping studies. This will lead to quicker and more reliable discoveries of genetic risk factors for many different diseases with great public health importance, which should in turn lead to improved prevention and treatment strategies.

## TABLE OF CONTENTS

<b>PREFACE.....</b>	<b>XV</b>
<b>1.0 INTRODUCTION.....</b>	<b>1</b>
<b>1.1 OVERVIEW .....</b>	<b>1</b>
<b>1.2 PRINCIPLES OF LINKAGE ANALYSIS .....</b>	<b>2</b>
<b>1.3 LINKAGE ANALYSIS METHODS FOR BINARY TRAITS .....</b>	<b>3</b>
<b>1.4 LINKAGE ANALYSIS METHODS FOR QUANTITATIVE TRAITS.....</b>	<b>5</b>
<b>1.5 PRINCIPLES OF ASSOCIATION ANALYSIS.....</b>	<b>6</b>
<b>1.6 ASSOCIATION TESTS FOR CASE-CONTROL STUDIES.....</b>	<b>7</b>
<b>1.7 FAMILY-BASED ASSOCIATION METHODS.....</b>	<b>8</b>
<b>1.8 NOTES ON DIFFERENT METHODS.....</b>	<b>10</b>
<b>1.9 WHAT IS THIS DISSERTATION ABOUT?.....</b>	<b>11</b>
<b>2.0 ROBUST S CORE STATISTICS F OR Q TL L INKAGE ANALYSIS US ING EXTENDED PEDIGREES .....</b>	<b>12</b>
<b>2.1 ABSTRACT .....</b>	<b>12</b>
<b>2.2 INTRODUCTION .....</b>	<b>12</b>
<b>2.3 MATERIALS AND METHODS.....</b>	<b>15</b>
<b>2.3.1 Review: CT and MERLIN Type Score Tests.....</b>	<b>15</b>
<b>2.3.2 Newly Proposed Score Tests .....</b>	<b>20</b>
<b>2.4 SIMULATION.....</b>	<b>26</b>

2.4.1	Robust score tests .....	26
2.4.2	Sensitivity analysis.....	32
2.5	RESULTS.....	34
2.5.1	Robust score tests .....	34
2.5.2	Sensitivity analysis.....	39
2.6	DISCUSSION.....	43
3.0	WHAT'S THE BEST STATISTIC FOR A SIMPLE TEST OF GENETIC ASSOCIATION IN A CASE-CONTROL STUDY? .....	45
3.1	ABSTRACT .....	45
3.2	INTRODUCTION.....	46
3.3	MATERIALS AND METHODS.....	47
3.3.1	TEST PROCEDURES .....	47
3.3.2	PREVIOUSLY ESTABLISHED RESULTS .....	50
3.3.3	METHODS FOR POWER CALCULATION.....	51
3.4	SIMULATION METHODS .....	52
3.4.1	SINGLE-LOCUS ANALYSIS.....	52
3.4.2	SCAN WITH COVARIATES .....	53
3.4.3	GENOME-WIDE ANALYSIS .....	53
3.5	RESULTS.....	54
3.5.1	SINGLE-LOCUS ANALYSIS.....	54
3.5.2	SCAN WITH COVARIATES .....	57
3.5.3	GENOME-WIDE ANALYSIS .....	59
3.6	DISCUSSION.....	62

3.6.1	SINGLE-LOCUS ANALYSIS.....	62
3.6.2	SCAN WITH COVARIATES .....	62
3.6.3	GENOME-WIDE ANALYSIS .....	63
3.6.4	OVERALL CONCLUSIONS.....	64
4.0	DISCUSSION .....	65
4.1	MY CONTRIBUTIONS .....	65
4.2	PROPOSED FUTURE WORK.....	67
4.3	OPEN PROBLEMS IN GENE MAPPING.....	69
4.3.1	Population stratification .....	69
4.3.2	Multiple testing .....	70
4.3.3	Meta analysis.....	70
4.3.4	Multilocus analysis .....	71
4.3.5	Modeling strategy .....	71
4.3.6	Rare variant analysis.....	72
4.3.7	Summary .....	73
APPENDIX A	SUPPLEMENTAL MATERIALS FOR CHAPTER 2.....	75
A.1	MEAN, VARIANCE, AND COVARIANCE OF ESTIMATED IBD SHARING OF TWO SIMILAR PEDIGREE TYPES .....	75
A.2	MATHEMATICAL INSIGHT ON DROPPING PARENTAL PHENOTYPES FROM SIBSHIP DATA IN THE CALCULATION OF SCORE TEST STATISTIC .....	80
A.3	POWER SIMULATION RESULTS .....	82
A.4	ANALYTICAL COMPARISON OF FS CORE.CT' <sub>ALL-TC</sub> AND SCORE.NULL.CT' <sub>ONE</sub> .....	95

<b>APPENDIX B SUPPLEMENTAL MATERIALS FOR CHAPTER 3 .....</b>	<b>98</b>
<b>B.1 ANALYTICAL COMPARISON OF 2×2 ALLELE AND 2×3 TREND .....</b>	<b>98</b>
<b>B.2 TABLES AND FIGURES.....</b>	<b>100</b>
<b>B.3 GENETIC MODEL OF A MARKER IN LD WITH A DISEASE LOCUS.....</b>	<b>107</b>
<b>BIBLIOGRAPHY .....</b>	<b>109</b>

## LIST OF TABLES

Table 2-1. Score tests considered in our work .....	21
Table 2-2. Trait models .....	27
Table 2-3. Pair types produced by a pedigree of each type .....	30
Table 2-4. Sample sizes .....	30
Table 2-5. Population trait parameters .....	31
Table 2-6. Phenotypic correlations for the normal trait models .....	33
Table 2-7. Phenotypic correlations for the moderately non-normal trait models .....	33
Table 2-8. Phenotypic correlations for the extremely non-normal trait models .....	34
Table 2-9. Number of replicates with a negative variance for MERLIN and $CT'_{FAM}$ type score tests given the 4G in power simulation .....	35
Table 3-1. $2 \times 3$ genotype-based table .....	47
Table 3-2. $2 \times 2$ allele-based table .....	47
Table 3-3. $2 \times 2$ genotype-based table combing the rarer homozygote class with the heterozygote class .....	48
Table 3-4. Genetic models .....	52
Table 3-5. Genome-wide simulation results for the sample size 500 .....	60
Table 3-6. Genome-wide simulation results for the sample size 1500 .....	61

Table A1-1. Joint distribution of estimated IBD sharing for the three sibpairs.....	76
Table A1-2. Marginal distribution of estimated IBD sharing for a single sibpair, e.g. 3-4.....	77
Table A1-3. Joint distribution of estimated IBD sharing for two sibpairs, e.g. 3-4 and 4-5 .....	77
Table A1-4. Joint distribution of estimated IBD sharing for the three siblings given the genotypes of the three siblings.....	79
Table A1-5. Variances and covariances of estimated IBD sharing for the sibpairs in the pedigrees 1-2-3-4-5 and 3-4-5.....	80
Table B2-1. Penetrances and sample sizes for interaction studies.....	100
Table B2-2. Power of single test procedure under each genetic model used for genome-wide simulations with sample size 500 (1500) and significance level 0.05 (0.0001) .....	101
Table B2-3. Power simulation results for the marker of intermediate genetic effect .....	101
Table B2-4. Penetrances of a marker in LD with a disease locus with complete genetic effect	102

## LIST OF FIGURES

Figure 2-1. Pedigree diagrams .....	29
Figure 2-2. 4G power simulation results.....	37
Figure 2-3. 4G null simulation results .....	38
Figure 2-4. 4G power simulation results using model-based corr. estimated using PC pairs .....	41
Figure 2-5. 4G power simulation results using model-based corr. estimated using SB pairs .....	42
Figure 3-1. Single-locus power simulation results for the minor allele frequencies: A. 0.05, B. 0.1, and C. 0.3. M1-M7 are the models as defined in Table 3-4. X-axis is power and y-axis is statistic name.....	56
Figure 3-2. Additive marker power simulation results for the three fitted logistic regression models: $G$ , $G+E$ , and $G+E+G\times E$ at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene $\times$ exposure interaction in which there is only a genetic effect in the exposed group, and D. gene $\times$ exposure interaction with effects in both groups. ....	58
Figure A3-1. 4SIBS power simulation results .....	83
Figure A3-2. HS power simulation results .....	84
Figure A3-3. 3G power simulation results.....	85
Figure A3-4. 2+4SIBS power simulation results .....	86
Figure A3-5. 4SIBS+3G power simulation results.....	87

Figure A3-6. 4SIBS+3G+HP power simulation results.....	88
Figure A3-7. 4SIBS power simulation results using model-based corr. estimated by PC pairs...	89
Figure A3-8. 4SIBS power simulation results using model-based corr. estimated by SB pairs...	90
Figure A3-9. HS power simulation results using model-based corr. estimated by PC pairs .....	91
Figure A3-10. HS power simulation results using model-based corr. estimated by SB pairs .....	92
Figure A3-11. 3G power simulation results using model-based corr. estimated by PC pairs .....	93
Figure A3-12. 3G power simulation results using model-based corr. estimated by SB pairs .....	94
Figure B2-1. Recessive marker power simulation results for the three fitted logistic regression models: $G$ , $G+E$ , and $G+E+G\times E$ at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene $\times$ exposure interaction in which there is only a genetic effect in the exposed group, and D. gene $\times$ exposure interaction with effects in both groups. ....	103
Figure B2-2. Dominant marker power simulation results for the three fitted logistic regression models: $G$ , $G+E$ , and $G+E+G\times E$ at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene $\times$ exposure interaction in which there is only a genetic effect in the exposed group, and D. gene $\times$ exposure interaction with effects in both groups. ....	104
Figure B2-3. Over-dominant marker power simulation results for the three fitted logistic regression models: $G$ , $G+E$ , and $G+E+G\times E$ at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene $\times$ exposure interaction in which there is only a genetic effect in the exposed group, and D. gene $\times$ exposure interaction with effects in both groups. ....	105

Figure B2-4. Under-dominant marker power simulation results for the three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G\times E$  at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene $\times$ exposure interaction in which there is only a genetic effect in the exposed group, and D. gene $\times$ exposure interaction with effects in both groups. .... 106

## PREFACE

In the past five years, I have been most fortunate to have Dr. Eleanor Feingold as my dissertation advisor. I thank her for guiding me with patience, sending me to conferences and short courses, and giving me opportunities to review papers and books. I enjoyed the meetings with her and was encouraged by her comments and suggestions. Besides professional knowledge, she taught me the characters of how to be a good scientist. She has made a significant impact on me not only as an advisor but also as a mentor during these years.

Being Graduate Student Researcher (GSR) under the mentorship of Dr. Daniel E. Weeks has been another fruitful experience for me. I like the flexibility he gave me to think independently about research questions and experiment on my ideas. I was encouraged to speak out my thoughts and opinions and have learned from his insightful comments and feedback. He has been providing me useful tools and resources to accomplish our research goals and giving me advice by telling great stories from his life experience. It has been fun and I have enjoyed working with Dr. Daniel E. Weeks.

I acknowledge Dr. Daniel E. Weeks, Dr. Michael M. Barmada, and Dr. George C. Tseng for serving as my proposal and dissertation committee members. Their input has greatly improved this dissertation. I am grateful to my co-workers Samsiddhi and Nandita for useful discussions and technical assistance and also to Ryan for maintaining the server Gattaca on which I ran all the simulations for my dissertation. I thank Johanna for the training before I started working with Dr. Daniel E. Weeks; John and Candy for helping me improve my presentation skills; the friends I made in Pittsburgh (Statistical Genetics group, Department of Biostatistics, and Badminton Club) and friends back in Taiwan. All my dissertation work is dedicated to my family. I appreciate the respect and faith they have in me. Their love and support are my greatest inspiration.

## 1.0 INTRODUCTION

### 1.1 OVERVIEW

Gene mapping is the process of localizing a gene that affects disease risk. Once a gene is known and characterized, it can be sequenced and the function of its molecular products can be analyzed. Statistical methods used for gene mapping can be grouped into two categories, linkage and association analysis.

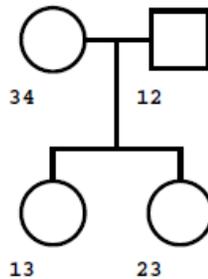
Linkage analysis requires family data to study allele transmission from one generation to the next, but association analysis, associating genes with one or multiple traits, can be done with unrelated samples and/or family data. In both analyses, genetic markers are typically SNPs (Single Nucleotide Polymorphism) and the phenotype can be binary (e.g. affection status), continuous (e.g. blood pressure), censored (e.g. survival time), and so on. Generally, rare variants with a moderate to strong effect can be detected using linkage while common but weaker variants are best detected by association analysis [Manolio, et al. 2009]. Linkage analysis has been successfully used to unravel the genetic basis of Mendelian disorders but the majority of diseases don't follow Mendelian inheritance patterns (complex diseases). People pursue GWAS assume that a complex disease is controlled by a number of variants, each with a high frequency but a low penetrance (common disease-common variant hypothesis). Application of high-throughput genotyping technology (e.g. completion of the Human Genome Project in 2003 and the International HapMap Project in 2005) and the illustration by Risch and Merikangas [1996]<sup>\*</sup> have motivated the recent wave of genome-wide association studies (GWAS). Achievements of GWAS, such as gene discovery of type 2 diabetes and age-related macular degeneration, are prominent but it is clear that much more than that remains to be done and I here leave statistical problems induced by GWAS for the final chapter, Discussion.

\*A common polymorphism tightly linked to a multiplicative disease gene with a relative risk less than 4 is much less detectable by linkage analysis than by association analysis provided that a dense map is available.

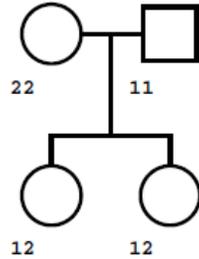
## 1.2 PRINCIPLES OF LINKAGE ANALYSIS

Linkage analysis measures the deviation from independent assortment of a disease and a genetic marker, but must be tested within families because causal haplotypes might vary across families. If a marker and trait locus are on different chromosomes, by Mendel's law of independent assortment, their alleles are transmitted to offspring at random. Suppose an individual has the genotype  $D_1D_2$  at the trait locus and  $M_1M_2$  at the marker. When the two loci are on different chromosomes, his or her offspring might inherit the gametes of  $D_1M_1$ ,  $D_1M_2$ ,  $D_2M_1$ , or  $D_2M_2$  with an equal probability of 0.25, i.e., the frequency of recombination between the marker and trait locus is expected to be 0.5. In contrast, if the two loci are linked, physical proximity may suppress the frequency of crossovers and gives a recombination frequency of less than 0.5.

Linkage analysis approaches either estimate the recombination frequency or indirectly study the correlation between the phenotype and genetic similarity shared by relative pairs. It is justified by the fact that a relative pair phenotypically similar should be genetically similar if the trait is controlled by genes. The genetic similarity is commonly measured by the number of alleles shared identical by descent (IBD) (number of the same ancestral alleles). Another similar measure is the proportion of alleles shared identical by descent (IBD sharing), which is simply the IBD divided by 2 (number of alleles at the locus). Consider a sibship of size 2 and the genotypes of father, mother, and their two children being 12, 34, 13, and 23.



It is apparent that the allele 3 in both children is from the mother and the father passed allele 1 to child 1 and allele 2 to child 2. By definition, the sibling pair shares one common ancestral allele so has IBD equal to 1 (IBD sharing = 0.5).



If the genotypes are replaced by 11, 22, 12, and 12, the real IBD then can be 0, 1, or 2 with the probabilities of 0.25, 0.5, and 0.25 [Ziegler and König 2006]. In such a case, the estimated IBD calculated by  $0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1$  (IBD sharing = 0.5) is used as a measure of genetic similarity for further statistical analysis.

### 1.3 LINKAGE ANALYSIS METHODS FOR BINARY TRAITS

I here classify linkage analysis methods for binary traits into two categories, model-based and model-free approaches. Model-based approaches require information on disease inheritance including allele frequencies and penetrances, but model-free approaches don't. Overall, model-based approaches are inferior to model-free approaches in computational efficiency and robustness to genetic effects but usually have higher power if the parameters of disease inheritance are correctly specified. The most typical model-based approach is the LOD (logarithm (base 10) of odds) score likelihood method [Morton 1955]. It evaluates the difference between the likelihoods of pedigree data at the recombination frequencies of 0.5 and the maximum likelihood estimate  $\hat{\theta}$ . Define the LOD score test as

$$Z = \log_{10} \frac{L(0.5)}{L(\hat{\theta})}.$$

By sequential probability ratio test [Wald 1947],  $Z \geq 3$  corresponds to a p-value of 0.0001 and linkage is unlikely if  $Z \leq -2$ . The likelihood for a nuclear family as

$$L(\theta) = P(g_F)P(g_M) \prod_{\text{offspring}} P(g_O|g_F, g_M, \theta),$$

where  $g_F$ ,  $g_M$ , and  $g_O$  are the genotypes of father, mother, and their children and  $\theta$  is the recombination frequency. In the situation of unknown phases and missing genotypes, the likelihood is given by

$$L(\theta) = \sum_{h_F} P(h_F) \sum_{h_M} P(h_M) \prod_{\text{offspring}} \sum_{h_O} P(h_O|h_F, h_M, \theta),$$

where  $h_F$ ,  $h_M$ , and  $h_O$  are the genotypes with the phase of father, mother, and their children. Assume that the marker and trait locus are in HWE,  $P(h_i)$  is a function of allele frequencies and  $P(h_O|h_F, h_M, \theta)$  is defined by the recombination frequency.

Model-free methods for binary traits are usually allele sharing methods and initially designed for affected sibpairs. The test statistic can be defined by

$$Z = \frac{\sum_{i=1}^m w_i Z_i}{\sqrt{\sum_{i=1}^m w_i^2}}$$

Each  $Z_i$  is defined as

$$Z_i = \frac{S_i - \mu_i}{\sigma_i},$$

where  $w_i$  is the weight assigned to the  $Z_i$  contributed by the  $i_{\text{th}}$  family,  $S_i$  is a function of IBD sharing among affected family members,  $\mu_i$  is the expectation and  $\sigma_i^2$  is the variance of  $S_i$  under the null hypothesis of no linkage. For example, let  $S_i$  be the IBD for the  $i_{\text{th}}$  affected sibpair and let each have an equal weight, i.e.  $w_1 = w_2 = \dots = w_m$ . The resulting  $Z$  is exactly the mean test

statistic [Blackwelder and Elston 1985] which compares the mean IBD to its expectation calculated under the null hypothesis. When both parents are unaffected or their phenotypes are missing, the mean test is equivalent to the LOD score test assuming a fully recessive model [Knapp, et al. 1994b]. It is locally most powerful regardless of mode of inheritance and is uniformly most powerful for multiplicative and fully recessive models [Knapp, et al. 1994a].

Be cautious that allele sharing methods using the theoretical variance perform conservatively if the marker is not fully polymorphic and this can be simply overcome by an empirical variance or simulation-based p-value. Another important issue for allele-sharing methods or even methods that model the IBD sharing is that the way they deal with uninformative pairs (assign them the null expected IBD sharing) might lead to an underpowered test [Schork and Greenwood 2004]. Strategies suggested by Schork and Greenwood [2004] to overcome this inherent problem are using weighting procedures, modified (and more appropriate) mixture models, and/or removing uninformative relative pairs from linkage analysis.

#### **1.4 LINKAGE ANALYSIS METHODS FOR QUANTITATIVE TRAITS**

In the development of linkage analysis methods for quantitative traits, the two earliest methods are Haseman and Elston (HE) regression [Haseman and Elston 1972] and the variance components (VC) method [Almasy and Blangero 1998; Amos 1994]. The HE method regresses the squared trait difference on the IBD sharing and tests linkage by the null hypothesis that the regression coefficient is zero against the alternative that it is negative. The expectation of the squared trait difference given the IBD sharing at the marker is a linear function of IBD sharing and the probability of one allele shared identical by descent in which the coefficient of IBD sharing is zero if and only if the recombination frequency is 0.5 or the additive trait variance explained by the trait locus is zero. After Wright [1997] pointed out that the trait difference cannot capture all the trait information without considering the trait sum, a number of HE revisited methods were proposed either using a more informative regression outcome (highly correlated with the IBD sharing) or efficiently combining the coefficient estimates for the squared trait difference and squared mean-corrected trait sum.

The HE revisited methods can only deal with nuclear families but the VC method doesn't have this constraint. The VC method models the trait values within a family by a multivariate normal distribution and assesses linkage by variance components using likelihood ratio test. Sensitivity of the VC method to the normal assumption however has motivated research in developing methods robust to different trait distributions and sampling schemes, for example, score statistics.

The score statistic for linkage analysis is the first derivative of the VC likelihood with respect to the additive trait variance evaluated under the null hypothesis of no linkage [Wang 2005]. Its variance derived conditioning on trait values generally has no closed form and can be estimated by different methods. Empirical variance estimators can maintain validity and power of the score statistic when the normal assumption doesn't hold. They however can be computed in many different ways and it is of practical interest to understand which gives robust power for different trait distributions and sampling schemes. Score statistic calculation requires population parameters including the mean and variance of the trait distribution and phenotypic correlations for different relative pair types. Misspecified parameters might have a great effect on the power of the score statistic. Power comparison of score tests (and other tests) thus should account for the effect of parameter misspecification.

A variety of variance estimators for the score statistic using sibpairs were investigated by Szatkiewicz, et al. [2003] and T. Cuenco, et al. [2003] and the studies later were extended for sibships by Bhattacharjee, et al. [2008] and for extended pedigrees by Dupuis, et al. [2009]. Compared to the study of Dupuis, et al. [2009], our work detailed in the later chapter of "Robust Score Statistics for QTL Linkage Analysis Using Extended Pedigrees" has surveyed a more complete collection of variance estimators under a broader range of realistic scenarios. For a more complete review of QTL linkage analysis methods, see the papers of Feingold [2001] and Feingold [2002].

## **1.5 PRINCIPLES OF ASSOCIATION ANALYSIS**

A marker is associated with a disease if it is correlated with affection status or other phenotypic measure. The trait locus might be untyped but its association with the disease can be captured by

its neighboring markers if they are in strong LD (Linkage Disequilibrium)\* with the trait locus. Both population and family data can be used to test for association. Family data, depending on how they are used, can test for population association or transmission distortion. For example, population association is tested if we compare the allele frequencies in affected and unaffected individuals across the families with correction for familial relatedness. In contrast, if we track how parental alleles are transmitted to offspring and compare the transmitted allele frequency of one allele to the others, we then are testing for transmission distortion. By conditioning on parental alleles, the idea used for testing transmission distortion guarantees the robustness against population stratification so avoids an inflated type I error. Testing for population association either using population samples or family data on the other hand might incur spurious association due to population stratification if cases and controls are sampled from two distinct ethnic populations with different allele frequencies. Remedies for solving this problem aimed at case-control data include Pritchard and Rosenberg [1999], Pritchard et al. [2000], Devlin and Roeder [1999], and Price, et al. [2006]. In the following, I first review association methods for case-control studies and then family-based association methods intended for testing transmission distortion hereafter simply referred to as family-based association methods.

\*Linkage disequilibrium: Two loci are in linkage disequilibrium if their haplotype frequencies deviate from the product of individual allele frequencies.

## 1.6 ASSOCIATION TESTS FOR CASE-CONTROL STUDIES

In a case-control study, the data for a SNP without consideration of environmental factors can be summarized in a  $2 \times 2$  table of disease status by allele or in a  $2 \times 3$  table of disease status by genotype. A SNP is associated with the disease if alleles or genotypes are not independent of disease status. 1 or 2 df independence chi-squared test and Cochran-Armitage trend test [Armitage 1955] are appropriate to analyze this type of data and genotype-based tests essentially are equivalent to logistic regression models with genotype main effect only. The logistic regression model however has greater flexibility than the chi-squared test in the use of control variables. It should be noted that 1df allele-based test requires that both case and control

populations be in HWE (Hardy-Weinberg Equilibrium)\* and is asymptotically equivalent to the trend test [Guedj, et al. 2008; Knapp 2008]. The control population is usually assumed to be in HWE but the case population might deviate from HWE if there is non-additive genetic effect [Guedj, et al. 2008]. Moreover, HWE in overall population is sufficient but not necessary to have HWE in case and control populations. Among these test procedures, practical questions of interest here include “What is the best test procedure in a single-locus analysis?”, “What is the best fitted logistic regression for detecting the genetic effect?” and “Are the inferences from a single locus analysis still valid in a genome scan?”. All of these issues will be addressed in the chapter titled “What’s the best statistic for a simple test of genetic association in a case-control study?” [Kuo and Feingold 2010].

\*Hardy-Weinberg disequilibrium: A biallelic locus is in HWE if the frequencies of the genotypes  $A_1A_1$ ,  $A_1A_2$ ,  $A_2A_2$  equal  $p^2$ ,  $2pq$ , and  $q^2$  where  $p$  and  $q$  are the allele frequencies of  $A_1$  and  $A_2$ .

## 1.7 FAMILY-BASED ASSOCIATION METHODS

Family-based association methods are typically TDT-based (Transmission Disequilibrium Test) tests. The original TDT [Spielman, et al. 1993], designed for binary traits, tests for linkage in the presence of association using trio data. It uses parental alleles not transmitted to offspring as controls and transmitted ones as cases and tests if a specific allele is more likely to be transmitted than the others. By the rule, families without heterozygous parents are not used so the power is attenuated but the idea of “pseudo” cases and controls makes the TDT robust to population stratification. As the method was first introduced, it didn’t consider the circumstance of missing parental genotypes. It had potential to extend for a multi-allelic marker, a quantitative trait, complex pedigrees (multiple offspring affected or unaffected and general pedigrees), multiple traits, and haplotypes but required further investigation. Since the 1990s, consequent TDT-based methods have been developed for specific issues [Laird and Lange 2008; Zhao 2000] but the most generalized one is probably the FBAT (Family-Based Association Test) [Rabinowitz and Laird 2000].

The FBAT is a score test derived from a multinomial likelihood conditioning on parental genotypes, trait values, and Mendel's first law. Its test statistic is the covariance between the phenotype (binary or quantitative) and a coding function of transmitted alleles. From a biallelic marker to a multi-allelic marker, the FBAT simply implements the idea of Spielman and Ewens [1998] by a  $p$ -dimensional vector in which each element represents a dichotomous coding for a specific transmitted allele. With regard to the extension for multiple offspring, as the null hypothesis is 1) no association and no linkage or 2) association but no linkage between the marker and trait locus, different offspring can be treated as if they are from separate families. To optimize the power, the contribution to the test statistic by offspring with different disease status might be assigned different weights. If the null hypothesis is 3) linkage but no association, allele transmission depends on the recombination frequency. To construct a valid test statistic, one can calculate its variance conditioning on the IBD sharing (greatly reduce the sample size, not recommended), by modeling the recombination frequency, or simply using an empirical variance estimator [Lake, et al. 2000]. Among the three null hypotheses, 1) is commonly tested in a genome-wide association study, 2) is used in conventional linkage analysis, and 3) is appropriate for a candidate gene study which might be a follow-up of previous linkage studies. For more details on how the FBAT has been generalized, see the review paper of Laird and Lange [2008] and the papers it refers to.

Another popular family-based association method for quantitative traits is the QTDT (Quantitative Transmission Disequilibrium Test) [Abecasis, et al. 2000]. Similar to the traditional VC method, it assumes that the trait values within a family are distributed as a multivariate normal distribution with a covariance matrix specified by variance components and estimated IBD sharing. Unlike the VC method, it doesn't use constant genotypic values at the QTL but estimate the allelic effect by incorporating a covariate of marker genotype. This implies that the success of association analysis relies on strong linkage disequilibrium (LD) between markers and trait loci. To avoid spurious association due to population stratification, it considers the extreme case that each family is drawn from a different stratum and decomposes the association into between- and within-family effects [Fulker, et al. 1999]. When population substructure is present, association testing based on within-effect maintains its validity but that based on between-effect doesn't. By manipulating the constraints on variance components and within-effect under the null and alternative hypotheses, the QTDT offers appreciable flexibility

for testing linkage and/or association. Moreover, permutation tests are used in case the normal assumption is not met. A practical question here is whether it is desirable to explore population stratification ahead of association analysis. The decomposition turns to be unnecessary if population stratification is not found. Binary trait data through the use of a threshold liability model can be analyzed in the framework of the VC method [Williams, et al. 1999] and QTDT. The FBAT using a quantitative trait (quantitative FBAT) in fact is the score test derived from the QTDT likelihood assuming no phenotypic correlation within a family [Lange, et al. 2002]. Power comparison of quantitative FBAT and QTDT was done by [Lange, et al. 2002] but that of binary FBAT and QTDT on continuous liabilities remains open.

## 1.8 NOTES ON DIFFERENT METHODS

It should be noted that the two most popular family-based association methods, FBAT and QTDT, are technically evolved from the TDT and the VC method. From the viewpoint of statistical theory, the VC method and QTDT are doing likelihood ratio tests. TDT and FBAT on the other hand are essentially score tests. It is well known that likelihood-based methods make assumptions about a model and rely on that for their performance. In general, score statistics conditioning on partial observed information are computationally more efficient and the use of empirical variance estimator makes them less sensitive to the distributional assumption; likelihood based methods are more flexible in conducting hypotheses as demonstrated by QTDT; score statistics require nuisance parameters be specified but likelihood based methods estimate them instead (ascertained samples cannot be used to estimate population parameters however).

One general and important issue that deserves serious attention is how to appropriately handle uninformative relative pairs. Traditional allele sharing methods have been long criticized for this and Schork and Greenwood [2004] provided us a simple coin example to demonstrate what potential biases might be. Removing uninformative relative pairs as they have suggested is probably an option to overcome this problem. “Uninformative” might be defined differently for different test procedures. Basically, uninformative relative pairs contribute the same amount as expected under the null hypothesis to the test statistic. Relative pairs that completely have no variation in the measure of interest undoubtedly are uninformative. Otherwise, observations

identical to the expected value might be driven by scientific facts or by systematic errors or biases such as an uninformative marker. Unless we can differentiate one from the other, we will not be able to appropriately handle uninformative relative pairs.

## 1.9 WHAT IS THIS DISSERTATION ABOUT?

This dissertation consists of four chapters: in the present chapter, “Overview” I have simply reviewed a few methods that play a role in the history of linkage and association analysis. The purpose is not to be complete and detailed but to introduce the ideas behind and reveal the questions of interest. As noted, score statistics in linkage analysis and association tests for case-control studies will be expanded in chapters 2 and 3. In chapter 2, “Robust Score Statistics for QTL Linkage Analysis Using Extended Pedigrees” the topic is a straightforward extension of Bhattacharjee, et al. [2008]. Rather than sibships, we study the score tests derived from those previously considered in Bhattacharjee, et al. [2008] using general pedigrees and compare their performance by simulating various scenarios of trait distribution, sampling scheme, and distribution of pedigree type. In chapter 3, “What’s the best simple genetic association test in a case-control study?”, suppose that the “simple” strategy, simple models for individual SNPs followed by fancy ones for a small subset, might work if the sample size is sufficiently large. To provide a guideline for choosing an appropriate test procedure or logistic regression model for genome-wide association analysis, we perform a comprehensive comparison study of chi-squared tests and logistic regression models from three aspects, single-locus analysis, scan with covariates, and genome-wide analysis. In chapter 4, “Discussion”, first I will summarize my research contributions and propose future work; then address open problems of GWAS and lead a discussion on strategies for gene mapping.

## **2.0 ROBUST SCORE STATISTICS FOR QTL LINKAGE ANALYSIS USING EXTENDED PEDIGREES**

### **2.1 ABSTRACT**

Score statistics for quantitative trait locus (QTL) linkage analysis have been proposed by many authors as an alternative to variance components (VC) and/or Haseman-Elston (HE) type methods because they have high power and can be made robust to selected samples and/or non-normal traits. But most literature exploring the properties of these statistics has focused on nuclear families. There are a number of computational complexities involved in implementing the score statistics for extended pedigrees, primarily having to do with computation of the statistic variance. In our work, we propose several different practical methods for computing this variance in general pedigrees, some of which are based only on relative pairs and some of which require working with the overall pedigree structure, which is computationally more difficult. We evaluate the performance of these different score tests using various trait distributions, ascertainment schemes, and pedigree types.

### **2.2 INTRODUCTION**

Currently available methods for QTL linkage analysis using general pedigrees include the variance components (VC) method [e.g., Amos 1994; Almasy and Blangero 1998] and score statistics such as the reverse regression method [Sham, et al. 2002] and the score statistics proposed by Dupuis, et al. [2009]. The VC method has been commonly used for mapping quantitative trait loci; however, is quite sensitive to the assumption of normality that the trait values of family members follow a multivariate normal distribution [Allison, et al. 1999]. Score

statistics are locally most powerful, efficient to compute, and can be made robust to the normality assumption through the use of an empirical variance estimate. The reverse regression method has been a significant improvement to the VC method in terms of comparable power as the normality assumption holds and robustness against non-normal traits and selected samples but requires intensive computation for the analysis of huge pedigrees. Dupuis, et al. [2009] evaluated few score statistics using population samples but their statistical properties might not hold for selected samples. The VC method and score statistics in fact can be unified in the framework of generalized estimating equations (GEEs) [Chen, et al. 2004; Chen, et al. 2005]. GEE-based score statistics incorporating the higher-moments of the trait distribution (skewness and kurtosis), referred to as higher-moment (HM) score statistics, are believed to improve the robustness against non-normal traits and/or selected samples. [Bhattacharjee, et al. 2008; Chen, et al. 2005].

The score statistic is derived by taking the first derivative of the VC likelihood with respect to the additive trait variance evaluated under the null hypothesis of no additive genetic effect [e.g., Wang, 2005]. To be clear, let the score test be the score statistic (numerator with expectation equal to zero) divided by its standard deviation (denominator). The numerator is essentially the same for population samples and for pedigrees ascertained based on phenotypes [Peng and Siegmund 2006; Wang 2005]. The denominator conditioning on trait values and on identical by descent (IBD) information can be made partially or fully empirical to enhance the robustness of the score test. To ensure the validity of a score test, empirical estimators of the denominator are preferred over null theoretical ones in case the normality assumption is violated. Additionally, score tests using the denominator conditioning on IBD information might be invalid for non-normal traits and/or selected samples [e.g., Bhattacharjee, et al. 2008; Dupuis, et al. 2009] and in this study we consider the denominator conditioning on trait values only.

We tackle the question of what score tests for general pedigrees have robust power for a variety of trait distributions, sampling schemes, and distributions of pedigree types. We define a pedigree type as a fixed pedigree structure for the members with a non-missing genotype and phenotype. Previous studies including T. Cuenco, et al. [2003], Szatkiewicz, et al. [2003], and Bhattacharjee, et al. [2008] touched on this issue for nuclear families. Dupuis, et al. [2009] considered extended pedigrees while focused on population samples and a limited number of variance estimators of the score statistic. To improve the robustness of a score test, we especially

consider many different partially or fully empirical denominators estimating the variances and covariances of estimated IBD sharing by grouping the same type of relative pairs of different extents. As the number of each type of relative pair, e.g. parent-child and sibling pairs, is not sufficiently large, we expect that their corresponding score tests might perform differently.

Several segregation/nuisance parameters are required for score statistic calculation. These include the mean and variance of the trait distribution, and phenotypic correlations for different relative pair types. The segregation parameters are theoretically orthogonal to the linkage parameter. It may reduce power if these are incorrectly specified. Bhattacharjee, et al. [2008] comprehensively studied the effect of parameter misspecification using sibships. Based on those simulation results, misspecified mean and correlation(s) can have a substantial effect on power and HM score statistics are fairly sensitive to the skewness and kurtosis. As the number of correlation parameters rapidly increases with pedigree size and complexity, there is a great chance to misspecify any of them. Those for more distant pairs might play a less important role in the power of score statistic, however. Which relative pair correlation parameters deserve most attention requires further investigation. We here, as MERLIN-REGRESS [Abecasis, et al. 2002] does, consider the correlations derived from a simple trait model.

In this study, we first review the score statistic and its variance estimators conditioning on trait values previously considered and newly proposed. We focus on the score tests empirically estimating their denominators not using pedigree type information. We compare their performance with those using pedigree type information and using other variance and covariance estimators. We evaluate the effect of ignoring pedigree types and our goal ultimately is to find score statistics that are robust to different trait distributions, sampling schemes, and distributions of pedigree types. Among the parameters required for the calculation of score statistics, we are interested in the feasibility of reducing the number of correlation parameters by assuming a simple trait model. We compare the score tests using model-based correlations and true correlations with power.

## 2.3 MATERIALS AND METHODS

### 2.3.1 Review: CT and MERLIN Type Score Tests

As aforementioned, a score test is essentially a  $Z$  test defined as the score statistic divided by its standard error, hereafter referred to as *numerator* and *denominator* respectively. Before introducing more variants of the denominator considered in our work, we review the score statistic and a few variance estimators suggested for sibships in Bhattacharjee, et al. [2008]. Assume that there are  $m_k$  pedigrees of type  $k$ ,  $k = 1, 2, \dots, K$ , and the total number of pedigrees is  $m = \sum m_k$ . The numerator  $S$  [see, e.g., Wang 2005] under the null hypothesis of no additive genetic effect for a QTL can be written as

$$S = \sum_{i=1}^m S_i = \sum_{i=1}^m v_i' \text{vec}(\Pi_i - 2\Phi_i),$$

where  $v_i$  is the transformed phenotype vector for the  $i$ th pedigree defined as

$$\text{vec} \left[ (\Sigma_i)^{-1} (y_i - \mu \mathbf{1}_{n_i}) (y_i - \mu \mathbf{1}_{n_i})' (\Sigma_i)^{-1} - (\Sigma_i)^{-1} \right],$$

and  $\text{vec}$  is an operator that vectorizes the upper-diagonal elements of a matrix in a row-wise order.  $y_i$  is the vector of trait values of family  $i$ .  $\mu$  is the trait mean and  $\mathbf{1}_{n_i}$  is a 1-vector (all elements equal 1) of family size  $n_i$ .  $\Sigma_i$  is the covariance matrix specified by the trait variance and phenotypic correlations for different relative pair types.  $\Pi_i$  and  $\Phi_i$  are the matrix of estimated IBD sharing and kinship coefficient. The variance of the score statistic conditioning on trait values is given by

$$\text{Var}(S) = \sum_{k=1}^K \sum_{i=1}^{m_k} v_{ki}' \text{Var}(\text{vec}(\Pi_{ki})) v_{ki},$$

where  $Var(vec(\Pi_{ki}))$  is the covariance matrix of estimated IBD sharing depending on the pedigree type. We can also write  $S$  and  $Var(S)$  in the framework of GEE [see, e.g., Chen, et al. 2005]. Define

$$D_i = [0 \quad 0 \quad vec(\Pi_i - 2\Phi_i)'],$$

$$U_i = \left[ (y_i - \mu 1_{n_i})' \quad \left[ (y_i - \mu 1_{n_i})^2 - \sigma^2 1_{n_i} \right]' \quad vec \left\{ (y_i - \mu 1_{n_i})(y_i - \mu 1_{n_i})' - \Sigma_i \right\}' \right],$$

$$G_i = \begin{bmatrix} \Sigma_i & 0 & 0 \\ 0 & [2(\Sigma_{irs})^2] & [2\Sigma_{irv}\Sigma_{irw}] \\ 0 & [2\Sigma_{its}\Sigma_{ius}] & [\Sigma_{itv}\Sigma_{iuw} + \Sigma_{itw}\Sigma_{iuv}] \end{bmatrix},$$

for  $1 \leq r, s \leq n_i, 1 \leq t < u \leq n_i$ , and  $1 \leq v < w \leq n_i$ .  $G_i$  is the null Gaussian working covariance matrix of  $U_i$ .  $[2(\Sigma_{irs})^2]$  represents a matrix consisting of the elements  $2(\Sigma_{irs})^2$  where  $\Sigma_{irs}$  is the element in the  $r$ th row and  $s$ th column of  $\Sigma_i$ . GEE-based  $S$  and  $Var(S)$  then can be written as

$$S_{GEE} = \sum_{i=1}^m S_i = \sum_{i=1}^m D_i' G_i^{-1} U_i,$$

$$Var(S_{GEE}) = \sum_{k=1}^K \sum_{i=1}^{m_k} U_{ki}' G_{ki}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & Var(vec(\Pi_{ki})) \end{bmatrix} G_{ki}^{-1} U_{ki}.$$

The vector  $v_i$  consists of the last  $C_2^{n_i}$  elements of  $G_i^{-1} U_i$ . Let  $h_i$  be the vector with the last  $C_2^{n_i}$  elements of  $M_i^{-1} U_i$  where  $M_i$  is the working covariance matrix of higher-moment defined as

$$M_i = \begin{bmatrix} \Sigma_i & \hat{\gamma}_3 \sigma^3 I_{n_i \times n_i} & 0 \\ \hat{\gamma}_3 \sigma^3 I_{n_i \times n_i} & [2(\Sigma_{irs})^2] + \hat{\gamma}_4 \sigma^4 I_{n_i \times n_i} & [2\Sigma_{irv}\Sigma_{irw}] \\ 0 & [2\Sigma_{its}\Sigma_{ius}] & [\Sigma_{itv}\Sigma_{iuw} + \Sigma_{itw}\Sigma_{iuv}] \end{bmatrix},$$

for  $1 \leq r, s \leq n_i, 1 \leq t < u \leq n_i$ , and  $1 \leq v < w \leq n_i$ . In  $M_i, \hat{\gamma}_3$  and  $\hat{\gamma}_4$  are empirical estimates for skewness and kurtosis of the trait distribution. The higher-moment score test replaces  $v_i$  by  $h_i$  in  $S$  and  $Var(S)$ . Its numerator and denominator are denoted by  $S_{HM}$  and the square root of  $Var(S_{HM})$  respectively,

$$S_{HM} = \sum_{i=1}^m S_i = \sum_{i=1}^m h_i' vec(\Pi_i - 2\Phi_i),$$

$$Var(S_{HM}) = \sum_{k=1}^K \sum_{i=1}^{m_k} h_{ki}' Var(vec(\Pi_{ki})) h_{ki}.$$

MERLIN and CT (Conditioning on Trait values) score tests have the same numerator (lower-moment or higher-moment) and use the denominator conditioning on trait values, but estimate the variances and covariances of estimated IBD sharing in  $Var(vec(\Pi_{ki}))$  differently. MERLIN score tests (essentially equivalent to the reverse regression method [Sham, et al. 2002]) estimate them by imputed variances and covariances (prior covariance (variance) minus posterior covariance (variance) given the marker data of family members). CT score tests estimate them by partially or fully empirical variances and covariances.

Let SCORE.MERLIN and HM.MERLIN denote the lower-moment and higher-moment MERLIN score test respectively. Both estimate  $Var(vec(\Pi_{ki}))$  by

$$\hat{\Sigma}_{ki}^{MERLIN} = Var(vec(\tilde{\Pi}_{ki})) - Var(vec(\tilde{\Pi}_{ki})|M_{ki}),$$

where  $M_{ki}$  is the marker data, and  $\tilde{\Pi}_{ki}$  is the matrix of expected IBD sharing. The idea behind comes from

$$\begin{aligned} Var(vec(\tilde{\Pi}_{ki})) &= Var[E(vec(\tilde{\Pi}_{ki})|M_{ki})] + E[Var(vec(\tilde{\Pi}_{ki})|M_{ki})] \\ &= Var(vec(\Pi_{ki})) + Var(vec(\tilde{\Pi}_{ki})|M_{ki}). \end{aligned}$$

MERLIN score tests occasionally fail because a negative statistic variance is estimated. They don't require pedigree type information; additional grouping by pedigree type might speed up statistical computation but is unnecessary.

Let SCORE.CT and SCORE.NULL.CT denote the lower-moment CT score tests and HM.CT denote the higher-moment CT score test. SCORE.NULL.CT estimates  $Var(vec(\Pi_{ki}))$  by

$$\hat{\Sigma}_{ki}^{\text{NULL.CT}} = \frac{1}{m_k} \sum_{i=1}^{m_k} vec(\Pi_{ki} - 2\Phi_k) vec(\Pi_{ki} - 2\Phi_k)',$$

where  $E(\Pi_{ki}) = 2\Phi_{ki} = 2\Phi_k$ . SCORE.CT and HM.CT estimate  $Var(vec(\Pi_{ki}))$  by

$$\hat{\Sigma}_{ki}^{\text{CT}} = \frac{1}{m_k - 1} \sum_{i=1}^{m_k} vec(\Pi_{ki} - \bar{\Pi}_k) vec(\Pi_{ki} - \bar{\Pi}_k)',$$

where  $\bar{\Pi}_k = \frac{1}{m_k} \sum_{i=1}^{m_k} \Pi_{ki}$ . SCORE.CT and HM.CT center the estimated IBD sharing at the sample average while SCORE.NULL.CT does it at the expected value given the relationship. In theory, SCORE.CT is expected to be more powerful than SCORE.NULL.CT whenever  $E(\Pi_{ki}) \neq 2\Phi_k$ .

SCORE.CT is a test taking advantage of the knowledge that 1) the variances and covariances of estimated IBD sharing for the relative pairs of the same type might vary with pedigree types, and 2) covariances of estimated IBD sharing between relative pairs within families are dependent on the persons involved. For example, pairs of 4-9 and 8-9 (numbers are person ids) and pairs of 4-9 and 7-11 are both avuncular-cousin pairs but their covariances of IBD sharing differ because the 4-9 and 8-9 pairs are more correlated than 4-9 and 7-11 pairs because of person 9. SCORE.CT estimates the variances and covariances of estimated IBD sharing simply using a fixed pair in the pedigrees of the same type. For example, take a sibship of size three (trisib) and denote the person ids of the three siblings by age in any trisibs are 3, 4, and 5. SCORE.CT uses all the 3-4 sibpairs (all the 3-4 and 4-5 sibpairs) in trisibs to estimate the variance (covariance) of estimated IBD sharing for the 3-4 sibpair (3-4 and 4-5 sibpairs) in this trisib.

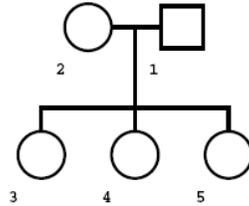
Given a pedigree type, we expect more or less power gain for SCORE.CT depending on whether it estimates the variances and covariances of estimated IBD sharing by grouping pairs of the same type, or by using one fixed pair in the pedigrees of the same type. For a sibship, given the parental genotypes, the variance of estimated IBD sharing for any sibpair is 0.5 and the covariance (of estimated IBD sharing) for any two sibpairs is fixed at  $0.25r(1-r)(r^2-r+1)$  where  $r$  is the minor allele frequency of the marker (see Appendix A.1). Sibships are more common than other pedigree types, and generally we have observed that the majority of covariance terms are close to zero. Both motivate us to propose additional CT type score tests that group more pairs of the same type(s) to estimate the variances and covariances of estimated IBD sharing.

As applied to general pedigrees, SCORE.CT might end up with a huge number of pedigree types and the number of pedigrees of each type might be insufficient to give a good estimate of the variances and covariances of the estimated IBD sharing. Other definitions of pedigree type might be appropriate in terms of power of a score test but require further investigation. A compromise approach to avoid using pedigree types is estimating the variances and covariances of estimated IBD sharing by relative type not using pedigree type information. Relative pairs in a genetically less informative pedigree are expected to have a smaller variance of estimated IBD sharing but this may not hurt the efficiency of a score statistic much if we estimate the IBD sharing based on the joint IBD distribution for each pedigree with neither parental genotypes missing and/or integrate the IBD information from neighbor markers (multipoint IBD sharing) as implemented in many algorithms for the estimation of IBD sharing like MERLIN [Abecasis, et al. 2002].

Based on our analytic illustrations in Appendix A.1, we can show that the variance of estimated IBD sharing for each sibpair and covariance between sibpairs given both parents genotyped are quite close to those without parental genotypes if the IBD sharing is estimated based on the joint IBD distribution given the genotypes of both parents and the three siblings. We thus expect that the effect of ignoring pedigree type for the pedigrees of similar types is minor. In the following, we introduce the newly proposed score statistics including one using pedigree type information and many not using pedigree type information.

### 2.3.2 Newly Proposed Score Tests

Overall, the score tests considered in our work fall into three classes, 1) MERLIN type; 2) CT type using pedigree type information, 3) CT type not using pedigree type information.



Suppose that we are estimating the variance of estimated IBD sharing for the sibpair 3-4. MERLIN score tests estimate it using the imputed variance. Among the CT type using pedigree type information (1<sup>st</sup> column of Table 2-1), the CT score tests (SCORE.CT, SCORE.NULL.CT, and HM.CT) estimate it using the 3-4 sibpair in each trisib and the CT<sub>ALL</sub> (SCORE.CT<sub>ALL</sub>, SCORE.NULL.CT<sub>ALL</sub>, and HM.CT<sub>ALL</sub>) score tests estimate it using *all* the sibpairs in trisibs. Among the CT type score tests not using pedigree type information (2<sup>nd</sup> column of Table 2-1), CT<sub>ONE</sub>' (SCORE.NULL.CT<sub>ONE</sub>') estimates it using the 3-4 pair in this trisib, the CT<sub>FAM</sub>' (SCORE.CT<sub>FAM</sub>', SCORE.NULL.CT<sub>FAM</sub>', and HM.CT<sub>FAM</sub>') score tests estimate it using all sibpairs in this trisib, and CT<sub>ALL</sub>' (SCORE.CT<sub>ALL</sub>', SCORE.NULL.CT<sub>ALL</sub>', HM.CT<sub>ALL</sub>', SCORE.CT<sub>ALL-TC</sub>', and HM.CT<sub>ALL-TC</sub>') score tests estimate it using *all* the sibpairs across all pedigrees. All the score tests considered in our work are summarized in Table 2-1.

We compare the classes 2) and 3) for evaluating the effect of ignoring pedigree types but ultimately intend to find score tests robust to different trait distributions, sampling schemes, and distributions of pedigree types. We have reviewed the MERLIN and CT score tests and now would propose CT<sub>ALL</sub> in the class 2), and CT<sub>ALL</sub>', CT<sub>FAM</sub>', and CT<sub>ONE</sub>' in the class 3). The prime notation indicates that the score test doesn't use pedigree type information and the subscript reveals the extent of relative pairs used in the calculation of the variances and covariances of estimated IBD sharing.

**Table 2-1. Score tests considered in our work**

CT type using pedigree type information		CT type not using pedigree type information		MERLIN type
CT	SCORE.CT SCORE.NULL.CT HM.CT	CT' <sub>ONE</sub>	SCORE.NULL.CT' <sub>ONE</sub>	SCORE.MERLIN HM.MERLIN
CT <sub>ALL</sub>	SCORE.CT <sub>ALL</sub> SCORE.NULL.CT <sub>ALL</sub> HM.CT <sub>ALL</sub>	CT' <sub>FAM</sub>	SCORE.CT' <sub>FAM</sub> SCORE.NULL.CT' <sub>FAM</sub> HM.CT' <sub>FAM</sub>	
		CT' <sub>ALL</sub>	SCORE.CT' <sub>ALL</sub> SCORE.NULL.CT' <sub>ALL</sub> HM.CT' <sub>ALL</sub> SCORE.CT' <sub>ALL-TC</sub> HM.CT' <sub>ALL-TC</sub>	

Note: **NULL** - variance and covariance of estimated IBD sharing centered at the expected value instead of sample average; **HM** - **H**igher-**M**oment transformed phenotypes incorporating skewness and kurtosis of the trait distribution; **CT** - using one fixed pair in the pedigrees of the same type; **CT<sub>ALL</sub>** - using *all* the same type of pairs in the same type of pedigrees; **CT'<sub>ONE</sub>** - using a single pair for the variance and two pairs for the covariance; **CT'<sub>FAM</sub>** - using the pairs of the same type in the current pedigree; **CT'<sub>ALL</sub>** - using the pairs of the same type across all pedigrees.

For simplicity in defining the score tests not using pedigree type information, we rewrite the score statistic and its variance in terms of pairs. Denote by  $S_{pairs}$  and  $Var(S_{pairs})$  the numerator and the denominator conditioning on trait values.

$$S_{pairs} = \sum_{i=1}^m \sum_{j=1}^{n_i} v_{ij} (\hat{\pi}_{ij} - 2\phi_{ij}),$$

$$Var(S_{pairs}) = \sum_{i=1}^m \left\{ \sum_{j=1}^{n_i} v_{ij}^2 Var(\hat{\pi}_{ij}) + \sum_{p \neq q} v_{ip} v_{iq} Cov(\hat{\pi}_{ip}, \hat{\pi}_{iq}) \right\},$$

where  $v_{ij}$ ,  $\hat{\pi}_{ij}$ , and  $\phi_{ij}$  are the transformed phenotype (the  $j$ th element in  $v_i$ ), estimated IBD sharing, and kinship coefficient for the  $j^{\text{th}}$  pair in the  $i$ th pedigree. The higher-moment score statistic and its variance are denoted by  $S_{HM.pairs}$  and  $Var(S_{HM.pairs})$ , and are created by replacing  $v$  by  $h$  in  $S_{pairs}$  and  $Var(S_{pairs})$ .

Let  $p_i$  be the pedigree type of the  $i$ th pedigree,  $r_j$  be the relative type of the  $j$ th pair,  $\bar{\pi}_{r_j}$  be the average of estimated IBD sharing for the relative type  $r_j$  calculated using all the pairs of that type

across the pedigrees, and  $n_{r_j}^{FAM_g}$  be the number of  $r_j$  type pairs in the  $g$ th pedigree. Newly proposed score tests using pedigree type information  $CT_{ALL}$  estimate  $Var(\hat{\pi}_{ij})$  and  $Cov(\hat{\pi}_{ip}, \hat{\pi}_{iq})$  using *all* the same type of relative pairs in the same type of pedigrees.

○ SCORE.NULL.CT<sub>ALL</sub>

$$\widehat{Var}_{NULL.CT_{ALL}}(\hat{\pi}_{ij} | p_i = k) = \frac{1}{\sum_{g: p_g=k} n_{r_j}^{FAM_g}} \sum_{g: p_g=k} \sum_{t: r_t=r_j} (\hat{\pi}_{gt} - 2\phi_{gt})^2$$

$$\begin{aligned} \widehat{Cov}_{NULL.CT_{ALL}}(\hat{\pi}_{ip}, \hat{\pi}_{iq} | p_i = k) \\ = \frac{1}{\sum_{g: p_g=k} n_{r_p}^{FAM_g} n_{r_q}^{FAM_g}} \sum_{g: p_g=k} \sum_{u,v: r_u=r_p, r_v=r_q} (\hat{\pi}_{gu} - 2\phi_{gu})(\hat{\pi}_{gv} - 2\phi_{gv}) \end{aligned}$$

○ SCORE.CT<sub>ALL</sub> and HM.CT<sub>ALL</sub>

$$\widehat{Cov}_{CT_{ALL}}(\hat{\pi}_{ip}, \hat{\pi}_{iq} | p_i = k) = \begin{cases} \frac{1}{\sum_{g: p_g=k} n_{r_p}^{FAM_g} n_{r_q}^{FAM_g} - df} \sum_{g: p_g=k} \sum_{u,v: r_u=r_p, r_v=r_q} (\hat{\pi}_{gu} - \bar{\pi}_{r_p})(\hat{\pi}_{gv} - \bar{\pi}_{r_q}), \\ \quad \text{if } \sum_{g: p_g=k} n_{r_p}^{FAM_g} n_{r_q}^{FAM_g} - df > 0 \\ \widehat{Cov}_{NULL.CT_{ALL}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}), \text{ if } \sum_{g: p_g=k} n_{r_p}^{FAM_g} n_{r_q}^{FAM_g} - df \leq 0 \end{cases}$$

where  $df = 1$ , if  $r_p = r_q$ , and  $df = 2$ , if  $r_p \neq r_q$ .

For the CT type score tests not using pedigree type information, we consider the score tests that estimate  $Var(\hat{\pi}_{ij})$  using a single pair and  $Cov(\hat{\pi}_{ip}, \hat{\pi}_{iq})$  using two pairs, using the pairs of the same type in that pedigree, and using all the pairs of the same type across all pedigrees.

1) using a single pair for the variance and two pairs for the covariance of estimated IBD sharing

○ SCORE.NULL.CT'ONE

$$\widehat{Var}_{\text{NULL.CT'ONE}}(\hat{\pi}_{ij}) = (\hat{\pi}_{ij} - 2\phi_{ij})^2$$

$$\widehat{Cov}_{\text{NULL.CT'ONE}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}) = (\hat{\pi}_{ip} - 2\phi_{ip})(\hat{\pi}_{iq} - 2\phi_{iq})$$

Both variance and covariance estimators were considered in Dupuis, et al. [2009] denoted by *Empirical*.

2) using the pairs of the same type in that pedigree

○ SCORE.NULL.CT'FAM

$$\widehat{Var}_{\text{NULL.CT'FAM}}(\hat{\pi}_{ij}) = \frac{1}{n_{r_j}^{FAM_i}} \sum_{t \in i: r_t=r_j} (\hat{\pi}_{it} - 2\phi_{it})^2$$

$$\widehat{Cov}_{\text{NULL.CT'FAM}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}) = \frac{1}{n_{r_p}^{FAM_i} n_{r_q}^{FAM_i}} \sum_{u,v \in i: r_u=r_p, r_v=r_q} (\hat{\pi}_{iu} - 2\phi_{iu})(\hat{\pi}_{iv} - 2\phi_{iv})$$

where  $t \in i$  denotes the  $t$  pair belongs to the  $i$  pedigree and similarly  $u, v \in i$  denotes the  $u$  and  $v$  pairs belong to the  $i$  pedigree.

○ SCORE.CT'FAM and HM.CT'FAM

$$\widehat{Var}_{\text{CT'FAM}}(\pi_{ij}) = \begin{cases} \frac{1}{n_{r_j}^{FAM_i} - 1} \sum_{t \in i: r_t=r_j} (\hat{\pi}_{it} - \bar{\pi}_{r_j})^2, & \text{if } n_{r_j}^{FAM_i} - 1 > 0 \\ \widehat{Var}_{\text{NULL.CT'FAM}}(\hat{\pi}_{ij}), & \text{if } n_{r_j}^{FAM_i} - 1 \leq 0 \end{cases}$$

$$\widehat{Cov}_{CT'_{FAM}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}) = \begin{cases} \frac{1}{n_{r_p}^{FAM_i} n_{r_q}^{FAM_i} - df} \sum_{u,v \in i: r_u=r_p, r_v=r_q} (\hat{\pi}_{iu} - \bar{\pi}_{r_p})(\hat{\pi}_{iv} - \bar{\pi}_{r_q}), \\ \text{if } n_{r_p}^{FAM_i} n_{r_q}^{FAM_i} - df > 0 \\ \widehat{Cov}_{NULL.CT'_{FAM}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}), \text{ if } n_{r_p}^{FAM_i} n_{r_q}^{FAM_i} - df \leq 0 \end{cases}$$

where  $df = 1$ , if  $r_p = r_q$ , and  $df = 2$ , if  $r_p \neq r_q$ .

Note that the variances and covariances of estimated IBD sharing estimated in this way might lead to a negative variance of the score statistic which is probably due to a negative definite covariance matrix of estimated IBD sharing.

3) using the pairs of the same type across the pedigrees

○ SCORE.NULL.CT'\_{ALL}

$$\widehat{Var}_{NULL.CT'_{ALL}}(\hat{\pi}_{ij}) = \frac{1}{\sum_g n_{r_j}^{FAM_g}} \sum_g \sum_{t: r_t=r_j} (\hat{\pi}_{gt} - 2\phi_{gt})^2$$

$$\widehat{Cov}_{NULL.CT'_{ALL}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}) = \frac{1}{\sum_g n_{r_p}^{FAM_g} n_{r_q}^{FAM_g}} \sum_g \sum_{u,v: r_u=r_p, r_v=r_q} (\hat{\pi}_{gu} - 2\phi_{gu})(\hat{\pi}_{gv} - 2\phi_{gv})$$

○ SCORE.CT'\_{ALL} and HM.CT'\_{ALL}

$$\widehat{Var}_{CT'_{ALL}}(\hat{\pi}_{ij}) = \begin{cases} \frac{1}{\sum_g n_{r_j}^{FAM_g} - 1} \sum_g \sum_{t: r_t=r_j} (\hat{\pi}_{gt} - \bar{\pi}_{r_j})^2, \text{ if } \sum_g n_{r_j}^{FAM_g} - 1 > 0 \\ \widehat{Var}_{NULL.CT'_{ALL}}(\hat{\pi}_{ij}), \text{ if } \sum_g n_{r_j}^{FAM_g} - 1 \leq 0 \end{cases}$$

$$\widehat{Cov}_{CT'_{ALL}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}) = \begin{cases} \frac{1}{\sum_g n_{r_p}^{FAM_g} n_{r_q}^{FAM_g} - df} \sum_g \sum_{u,v: r_u=r_p, r_v=r_q} (\hat{\pi}_{gu} - \bar{\pi}_{r_p})(\hat{\pi}_{gv} - \bar{\pi}_{r_q}), & \text{if } \sum_g n_{r_p}^{FAM_g} n_{r_q}^{FAM_g} - df > 0 \\ \widehat{Cov}_{NULL.CT'_{ALL}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}), & \text{if } \sum_g n_{r_p}^{FAM_g} n_{r_q}^{FAM_g} - df \leq 0 \end{cases}$$

where  $df = 1$ , if  $r_p = r_q$ , and  $df = 2$ , if  $r_p \neq r_q$ .

If all pedigrees are of a single type, SCORE.CT'\_{ALL} is exactly equivalent to SCORE.CT\_{ALL} and so are their null and higher-moment versions.

- SCORE.CT'\_{ALL-TC} and HM.CT'\_{ALL-TC} (TC: Theoretical Correlation)

$$\widehat{Var}_{CT'_{ALL-TC}}(\hat{\pi}_{ij}) = \widehat{Var}_{CT'_{ALL}}(\hat{\pi}_{ij})$$

$$\widehat{Cov}_{CT'_{ALL-TC}}(\hat{\pi}_{ip}, \hat{\pi}_{iq}) = Cor(\pi_{ip}, \pi_{iq}) \sqrt{\widehat{Var}_{CT'_{ALL}}(\hat{\pi}_{ip}) \widehat{Var}_{CT'_{ALL}}(\hat{\pi}_{iq})}$$

where  $Cor(\pi_{ip}, \pi_{iq})$  is the theoretical correlation between the IBD sharing of  $\pi_{ip}$  and  $\pi_{iq}$  assuming a completely polymorphic marker. *Estimated Variances* in Dupuis, et al. [2009] is similar to  $\widehat{Var}_{CT'_{ALL-TC}}(\hat{\pi}_{ij})$  or  $\widehat{Cov}_{CT'_{ALL-TC}}(\hat{\pi}_{ip}, \hat{\pi}_{iq})$  but estimates the variance of estimated IBD sharing using  $\widehat{Var}_{NULL.CT'_{ALL}}(\hat{\pi}_{ij})$  instead of  $\widehat{Var}_{CT'_{ALL}}(\hat{\pi}_{ij})$ .

It should be noted that the variances and covariances of estimated IBD sharing in SCORE.CT\_{ALL}, SCORE.CT'\_{FAM}, SCORE.CT'\_{ALL}, SCORE.CT'\_{ALL-TC}, and their higher moment versions center the estimated IBD sharing at the sample average which is calculated using *all* the pairs of each type across the pedigrees.

The pairs used to estimate a variance or covariance of estimated IBD sharing theoretically should have the same variance or covariance. Among the CT type score tests (see Table 2-1), this is true for the CT only. Basically, CT type using pedigree type information increases homogeneity among pairs used in the estimate as compared to tests that don't use pedigree type. The tradeoff is that tests that use pedigree type will base the variance/covariance estimates on smaller sample sizes.

We considered 10 relative-pair types: parent-child (PC), sibling (SB), grandparent-grandchild (GG), avuncular (AV), cousin (CS), half-sibling (HS), half-avuncular (HA), half-cousin (HC), unrelated (UR), and distant (DS). Any relative pairs which don't belong to the first nine types are counted in DS. The variances and covariances of estimated IBD sharing for distant pairs in non-NULL score tests are estimated as in NULL score tests since distant pairs are in fact different types of relative pairs and their estimated IBD sharing centered at the *overall* sample average would result in an inflated variance estimate. Parent-child and unrelated pairs are zeroed out because the IBD sharing is fixed at 0.5 and 0 respectively. This applies to both numerator and denominator. For the sibship data, disregarding the data from parent-child and unrelated pairs is mathematically equivalent to including them but assuming that the phenotypic correlations for parent-child and unrelated pairs are zero (Appendix A.2). This implies that uninformative pairs can contribute to the score test statistic if their phenotypic correlations can be correctly specified.

## 2.4 SIMULATION

### 2.4.1 Robust score tests

We evaluated the performance of score tests in terms of power and type I error. We considered different trait distributions, sampling schemes, and distributions of pedigree types. We used a marker with eight equiprecurrent alleles in linkage with a biallelic QTL with a minor allele frequency of 0.1 or 0.5. Real IBD data would be multipoint, and much more informative than a single SNP. We did a back-of-the-envelope calculation and decided that typical multipoint

information from a microsatellite scan was about equivalent to single marker with eight equiprequent alleles. Assume that the marker and QTL are in Hardy-Weinberg Equilibrium (HWE) and their alleles are transmitted to offspring with a recombination frequency 0.5 under the null and 0 under the alternative hypothesis.

For population (POP) samples and a normal trait denoted by  $x$ , the trait values of family members were simulated from a multivariate normal distribution with a mean vector dependent on the QTL genotypes and the covariance matrix with each element given by

$$[P(\text{IBD} = 2) + 0.5P(\text{IBD} = 1)]\sigma_{pa}^2 + P(\text{IBD} = 2)\sigma_{pd}^2 + r_e\sigma_e^2$$

where  $\sigma_{pa}^2$  and  $\sigma_{pd}^2$  are the additive and dominant polygenic variances,  $\sigma_e^2$  is the environmental variance, and  $r_e$  is the pairwise environmental correlation. Assume that  $\sigma_{pa}^2$  and  $\sigma_{pd}^2$  are zero, and we increase phenotypic covariances by environmental correlation: 1 for self-self, 0.15 for parent-child, 0.25 for sibling, and 0.05 for others.  $\sigma_e^2$  and the trait mean for each genotype vary with trait models and we considered five trait models with a heritability of 0.15 representing different genetic effects (Table 2-2).

**Table 2-2. Trait models**

Model	1	2	3	4	5
Inheritance Type	Additive	Dominant	Recessive	Additive	Dominant
Heritability	0.15	0.15	0.15	0.15	0.15
Minor allele freq.	0.1	0.1	0.1	0.5	0.5
Trait mean	-1, 0, 1	0, 1, 1	0, 0, 1	-1, 0, 1	0, 1, 1
Environmental SD	1.010	0.934	0.237	1.683	1.031

To make non-normal traits, we transformed the trait values  $x$  using the functions of  $x|x|$  (models 1'-5') and  $x^3$  (models 1''-5''). To make selected samples, we considered **SINGLE** proband sampling (SINGLE) and **Extreme Concordant** sampling (EC). A nuclear (extended) family is ascertained according to SINGLE if it has at least one sibling (member) whose trait value is in the top 10% of the trait distribution and is ascertained according to EC if it has at least two siblings (members) whose trait values are in the top 10% of the trait distribution. Technically, we first found the top and bottom 10% quantiles by simulating 10,000 trait values

from the population. We then kept screening pedigrees following the sampling schemes until we collected sufficient samples for a reasonable power.

We considered four single pedigree types and three mixed combinations of pedigree types. Each pedigree type is shown in Figure 2-1. For the single pedigree type studies, we considered sibships of size 4 (4SIBS), nuclear families with half-sibling (HS), 3-generation (3G), and 4-generation (4G) pedigrees. For the mixed pedigree types, we considered 50% : 50% of 2SIBS and 4SIBS (2+4SIBS), 75% : 25% of 4SIBS and 3G (4SIBS+3G), and 4SIBS+3G plus few specific “*huge pedigrees*” (4SIBS+3G+HP). Relative-pair types produced by each type of pedigree are summarized in Table 2-3. Sample sizes by distribution of pedigree types for each sampling scheme and trait distribution are given in Table 2-4.

We assumed that all the parameters required for score statistics calculation are available as given in Table 2-5. Most of them were empirically estimated using simulated population samples except the mean, variance, and correlations for models 1-5 were theoretically derived. 4G is the only single pedigree type having distant pairs which are essentially great-grandparent-great-grandchild pairs and we simply picked a *fixed* pair of that type in terms of person ids per pedigree to calculate the Pearson correlation coefficient. For the distant pairs in the mixed pedigree type 4SIBS+3G+HP, instead of using a *fixed* pair, we randomly selected a distant pair from each pedigree. We considered the significance level 0.01 in all our statistical tests and simulated 1,000 and 10,000 replicates for power and type I error analyses, respectively.

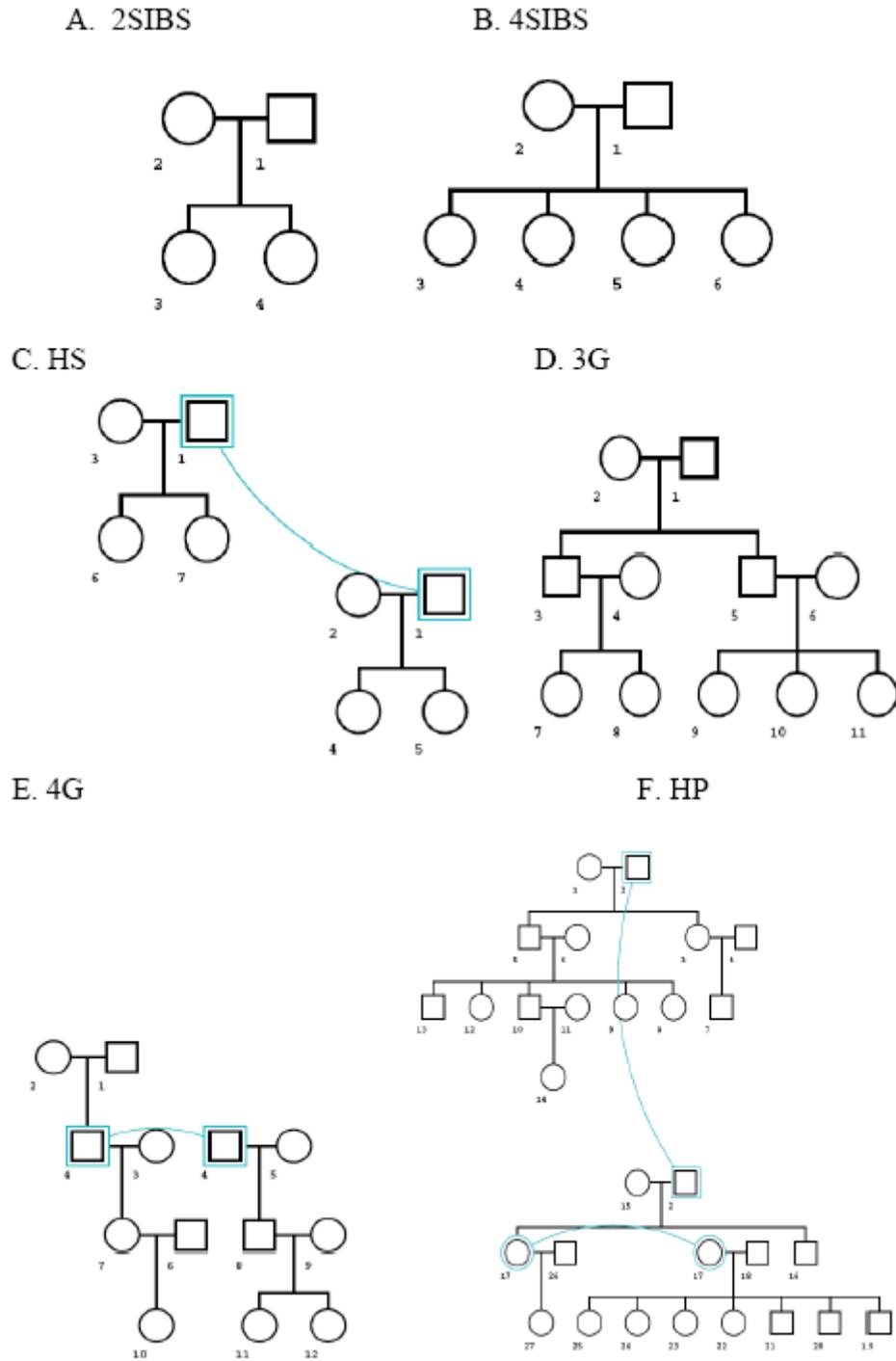


Figure 2-1. Pedigree diagrams

**Table 2-3. Pair types produced by a pedigree of each type**

Relative pair type	2SIB	4SIB	HS	3G	4G	HP
PC	4	8	8	14	12	38
SB	1	6	2	5	1	33
GG	0	0	0	10	10	30
AV	0	0	0	5	0	18
CS	0	0	0	6	0	5
HS	0	0	4	0	1	11
HA	0	0	0	0	3	28
HC	0	0	0	0	2	48
DS	0	0	0	0	6	14
UR	1	1	7	15	31	126

**Table 2-4. Sample sizes**

Sampling scheme	Trait	4SIBS	HS	3G	4G	2+4SIBS	4SIBS+3G	4SIBS+3G+HP
POP	x	450	900	300	400	400+400	300+100	240+80+10
	x x	750	1200	500	500	700+700	360+120	300+100+10
	x <sup>3</sup>	1000	1500	1000	1000	800+800	600+200	600+200+10
SEL	x	100	250	100	200	75+75	75+25	60+20+5
	x x	200	350	200	250	175+175	90+30	75+25+5
	x <sup>3</sup>	300	400	300	350	250+250	150+50	135+45+5
EC	x	50	150	75	150	50+50	60+20	60+20+5
	x x	100	150	150	250	75+75	60+20	60+20+5
	x <sup>3</sup>	200	200	200	350	125+125	75+25	75+25+5

**Table 2-5. Population trait parameters**

<b>Model</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
Mean	-0.800	0.190	0.010	0.000	0.750
Variance	1.200	1.026	0.066	3.333	1,250
Skewness	0.110	0.092	0.572	0.000	-0.067
Kurtosis	0.058	0.011	2.138	-0.023	-0.015
Cor <sub>PC</sub>	0.203	0.199	0.141	0.203	0.178
Cor <sub>SB</sub>	0.288	0.286	0.257	0.288	0.275
Cor <sub>GG</sub>	0.080	0.078	0.049	0.080	0.068
Cor <sub>AV</sub>	0.080	0.078	0.049	0.080	0.068
Cor <sub>CS</sub>	0.061	0.060	0.046	0.061	0.055
Cor <sub>HS</sub>	0.061	0.060	0.046	0.061	0.055
Cor <sub>HA</sub>	0.080	0.078	0.049	0.080	0.068
Cor <sub>HC</sub>	0.061	0.060	0.046	0.061	0.068
Cor <sub>DS</sub>	0.052	0.051	0.044	0.052	0.049
Cor <sub>UR</sub>	0.043	0.043	0.043	0.043	0.043
<b>Model</b>	<b>1'</b>	<b>2'</b>	<b>3'</b>	<b>4'</b>	<b>5'</b>
Mean	-1.486	0.340	0.011	-0.012	1.418
Variance	6.722	3.365	0.023	33.000	6.907
Skewness	-1.656	1.235	5.563	-0.013	1.724
Kurtosis	6.168	9.053	61.481	8.352	6.560
Cor <sub>PC</sub>	0.163	0.169	0.105	0.165	0.141
Cor <sub>SB</sub>	0.238	0.250	0.227	0.251	0.231
Cor <sub>GG</sub>	0.064	0.071	0.048	0.070	0.055
Cor <sub>AV</sub>	0.058	0.081	0.039	0.063	0.060
Cor <sub>CS</sub>	0.049	0.055	0.023	0.047	0.043
Cor <sub>HS</sub>	0.054	0.053	0.025	0.050	0.045
Cor <sub>HA</sub>	0.060	0.070	0.046	0.068	0.050
Cor <sub>HC</sub>	0.049	0.046	0.031	0.052	0.043
Cor <sub>DS</sub>	0.043	0.043	0.024	0.045	0.043
Cor <sub>UR</sub>	0.039	0.039	0.022	0.036	0.038
<b>Model</b>	<b>1''</b>	<b>2''</b>	<b>3''</b>	<b>4''</b>	<b>5''</b>
Mean	-3.240	0.687	0.012	0.038	3.150
Variance	56.890	18.797	0.023	541.864	59.052
Skewness	-3.762	3.412	12.950	-0.028	3.968
Kurtosis	27.527	46.171	234.652	39.290	30.243
Cor <sub>PC</sub>	0.110	0.131	0.062	0.122	0.106
Cor <sub>SB</sub>	0.186	0.193	0.239	0.174	0.172
Cor <sub>GG</sub>	0.040	0.051	0.038	0.051	0.035
Cor <sub>AV</sub>	0.040	0.055	0.037	0.045	0.047
Cor <sub>CS</sub>	0.034	0.037	0.021	0.038	0.029
Cor <sub>HS</sub>	0.040	0.041	0.027	0.038	0.032
Cor <sub>HA</sub>	0.041	0.051	0.036	0.046	0.036
Cor <sub>HC</sub>	0.038	0.037	0.026	0.034	0.027
Cor <sub>DS</sub>	0.032	0.030	0.013	0.032	0.026
Cor <sub>UR</sub>	0.027	0.021	0.009	0.034	0.023

## 2.4.2 Sensitivity analysis

Score tests applied for general pedigrees require specifying many correlation parameters. To reduce the number of parameters, it is practically preferred to use model-based correlations and simply estimate a single heritability parameter. To evaluate the effect of using model-based correlations on the power of score tests, we here introduce a simple trait model and based on that derive phenotypic correlations for sensitivity analysis.

Suppose a quantitative trait is completely controlled by a locus obeying HWE. Let  $y_{ij}$  be the trait value and  $x_{ij}$  be the genotype of the  $j$ th individual in the  $i$ th pedigree. Consider the trait model

$$y_{ij} = \mu + g_{ij} + u_i + \varepsilon_{ij}$$

where  $\mu$  is the grand mean,  $g_{ij}$  is the genotypic value depending on  $x_{ij}$ ,  $u_i$  is the environmental effect shared by family members, and  $\varepsilon_{ij}$  is the unshared environmental effect. Without loss of generality, let  $E(g_{ij}) = E(u_i) = E(\varepsilon_{ij}) = 0$ .  $\sigma^2 = \sigma_g^2 + \sigma_u^2 + \sigma_\varepsilon^2$  is the total trait variance where  $\sigma_g^2$  is the genotypic variance due to the QTL,  $\sigma_u^2$  is the shared environmental variance, and  $\sigma_\varepsilon^2$  is the unshared environmental variance. Partition the  $\sigma_g^2$  into an additive variance  $\sigma_a^2$  and a dominance variance  $\sigma_d^2$ . Rewrite  $\sigma^2$  as  $\sigma_a^2 + \sigma_d^2 + \sigma_u^2 + \sigma_\varepsilon^2$ . The covariance of  $y_{ij}$  and  $y_{ik}$  is given by

$$Cov(y_{ij}, y_{ik}) = 2\phi_{jk}\sigma_a^2 + P(\text{IBD} = 2)\sigma_d^2 + \sigma_u^2, j \neq k.$$

Assume that the dominance effect is minor and ignorable, i.e.  $\sigma_d^2 = 0$ , and there are no environmental correlations between family members. The correlation of  $y_{ij}$  and  $y_{ik}$ ,  $Cor(y_{ij}, y_{ik})$ , is simplified to  $2\phi_{jk}h$  (referred to as model-based correlation), where  $h$  is the heritability defined as  $\sigma_a^2/\sigma^2$  which can be estimated using parent-child pairs (sibling pairs) by  $2\widehat{Cor}_{PC}$  ( $\widehat{Cor}_{SB}$ ). We assume that an accurate correlation estimate for parent-child pairs (sibling pairs) is readily available; here we simply used the true correlation. For distant pairs, we used an empirical correlation estimated using population samples. Based on our limited simulation results, the score tests are not sensitive to the correlation for distant pairs. True and model-based correlations for different relative pair types are given in Tables 2-6-2-8. Eventually, we

compared the power of different score tests using true correlations and model-based correlations with the use of the same datasets.

**Table 2-6. Phenotypic correlations for the normal trait models**

Model	Type <sup>a</sup>	Cor <sub>UR</sub>	Cor <sub>PC</sub>	Cor <sub>SB</sub>	Cor <sub>GG</sub>	Cor <sub>AV</sub>	Cor <sub>CS</sub>	Cor <sub>HS</sub>	Cor <sub>HA</sub>	Cor <sub>HC</sub>	Cor <sub>DS</sub>
1	TRUE	0.043	0.203	0.288	0.080	0.080	0.061	0.061	0.080	0.061	0.052
	PC	0.000	0.203	0.203	0.101	0.101	0.051	0.061	0.101	0.051	0.025
	SB	0.000	0.288	0.288	0.144	0.144	0.072	0.061	0.144	0.072	0.036
2	TRUE	0.043	0.199	0.286	0.078	0.078	0.060	0.060	0.078	0.060	0.051
	PC	0.000	0.199	0.199	0.099	0.099	0.050	0.060	0.099	0.050	0.025
	SB	0.000	0.286	0.286	0.143	0.143	0.071	0.060	0.143	0.071	0.036
3	TRUE	0.043	0.141	0.257	0.049	0.049	0.046	0.046	0.049	0.046	0.044
	PC	0.000	0.141	0.141	0.071	0.071	0.035	0.046	0.071	0.035	0.018
	SB	0.000	0.257	0.257	0.128	0.128	0.064	0.046	0.128	0.064	0.032
4	TRUE	0.043	0.203	0.288	0.080	0.080	0.061	0.061	0.080	0.061	0.052
	PC	0.000	0.203	0.203	0.101	0.101	0.051	0.061	0.101	0.051	0.025
	SB	0.000	0.288	0.288	0.144	0.144	0.072	0.061	0.144	0.072	0.036
5	TRUE	0.043	0.178	0.275	0.068	0.068	0.055	0.055	0.068	0.068	0.049
	PC	0.000	0.178	0.178	0.089	0.089	0.044	0.055	0.089	0.044	0.022
	SB	0.000	0.275	0.275	0.138	0.138	0.069	0.055	0.138	0.069	0.034

a: TRUE - true correlation; PC – model-based correlation estimated using parent-child pairs; SB – model-based correlations estimated using sibling pairs.

**Table 2-7. Phenotypic correlations for the moderately non-normal trait models**

Model	Type <sup>a</sup>	Cor <sub>UR</sub>	Cor <sub>PC</sub>	Cor <sub>SB</sub>	Cor <sub>GG</sub>	Cor <sub>AV</sub>	Cor <sub>CS</sub>	Cor <sub>HS</sub>	Cor <sub>HA</sub>	Cor <sub>HC</sub>	Cor <sub>DS</sub>
1	TRUE	0.039	0.163	0.238	0.064	0.058	0.049	0.054	0.060	0.049	0.043
	PC	0.000	0.163	0.163	0.082	0.082	0.041	0.054	0.082	0.041	0.020
	SB	0.000	0.238	0.238	0.119	0.119	0.059	0.054	0.119	0.059	0.030
2	TRUE	0.039	0.169	0.250	0.071	0.081	0.055	0.053	0.070	0.046	0.043
	PC	0.000	0.169	0.169	0.085	0.085	0.042	0.053	0.085	0.042	0.021
	SB	0.000	0.250	0.250	0.125	0.125	0.063	0.053	0.125	0.063	0.031
3	TRUE	0.022	0.105	0.227	0.048	0.039	0.023	0.025	0.046	0.031	0.024
	PC	0.000	0.105	0.105	0.053	0.053	0.026	0.025	0.053	0.026	0.013
	SB	0.000	0.227	0.227	0.114	0.114	0.057	0.025	0.114	0.057	0.028
4	TRUE	0.036	0.165	0.251	0.070	0.063	0.047	0.050	0.068	0.052	0.045
	PC	0.000	0.165	0.165	0.082	0.082	0.041	0.050	0.082	0.041	0.021
	SB	0.000	0.251	0.251	0.125	0.125	0.063	0.050	0.125	0.063	0.031
5	TRUE	0.038	0.141	0.231	0.055	0.060	0.043	0.045	0.050	0.043	0.043
	PC	0.000	0.141	0.141	0.071	0.071	0.035	0.045	0.071	0.035	0.018
	SB	0.000	0.231	0.231	0.055	0.116	0.058	0.045	0.116	0.058	0.029

a: TRUE - true correlation; PC – model-based correlation estimated using parent-child pairs; SB – model-based correlations estimated using sibling pairs.

**Table 2-8. Phenotypic correlations for the extremely non-normal trait models**

Model	Type <sup>a</sup>	Cor <sub>IR</sub>	Cor <sub>PC</sub>	Cor <sub>SB</sub>	Cor <sub>GG</sub>	Cor <sub>AV</sub>	Cor <sub>CS</sub>	Cor <sub>HS</sub>	Cor <sub>HA</sub>	Cor <sub>HC</sub>	Cor <sub>DS</sub>
1	TRUE	0.039	0.163	0.238	0.064	0.058	0.049	0.054	0.060	0.049	0.043
	PC	0.000	0.163	0.163	0.082	0.082	0.041	0.054	0.082	0.041	0.020
	SB	0.000	0.238	0.238	0.119	0.119	0.059	0.054	0.119	0.059	0.030
2	TRUE	0.039	0.169	0.250	0.071	0.081	0.055	0.053	0.070	0.046	0.043
	PC	0.000	0.169	0.169	0.085	0.085	0.042	0.053	0.085	0.042	0.021
	SB	0.000	0.250	0.250	0.125	0.125	0.063	0.053	0.125	0.063	0.031
3	TRUE	0.022	0.105	0.227	0.048	0.039	0.023	0.025	0.046	0.031	0.024
	PC	0.000	0.105	0.105	0.053	0.053	0.026	0.025	0.053	0.026	0.013
	SB	0.000	0.227	0.227	0.114	0.114	0.057	0.025	0.114	0.057	0.028
4	TRUE	0.036	0.165	0.251	0.070	0.063	0.047	0.050	0.068	0.052	0.045
	PC	0.000	0.165	0.165	0.082	0.082	0.041	0.050	0.082	0.041	0.021
	SB	0.000	0.251	0.251	0.125	0.125	0.063	0.050	0.125	0.063	0.031
5	TRUE	0.038	0.141	0.231	0.055	0.060	0.043	0.045	0.050	0.043	0.043
	PC	0.000	0.141	0.141	0.071	0.071	0.035	0.045	0.071	0.035	0.018
	SB	0.000	0.231	0.231	0.055	0.116	0.058	0.045	0.116	0.058	0.029

a: TRUE - true correlation; PC – model-based correlation estimated using parent-child pairs; SB – model-based correlations estimated using sibling pairs.

## 2.5 RESULTS

### 2.5.1 Robust score tests

For the lower-moment (LM) score tests, the score tests ignoring pedigree types are as powerful as those using pedigree type information. SCORE.CT<sub>ALL</sub> and SCORE.CT'<sub>ALL</sub> are mathematically equivalent for the single pedigree types. Both perform similarly for the mixed pedigree types of 2+4SIBS, 4SIBS+3G, and 4SIBS+3G+HP. NULL and non-NULL score tests have similar power except CT'<sub>FAM</sub> (Figures A3-4-A3-6). SCORE.NULL.CT'<sub>ONE</sub> and SCORE.CT'<sub>FAM</sub> are overly conservative and generally less powerful. Overall, SCORE.MERLIN has the best power, followed by SCORE.CT and SCORE.CT'<sub>ALL-TC</sub>, and then SCORE.NULL.CT'<sub>FAM</sub> and SCORE.CT'<sub>ALL</sub> (SCORE.CT<sub>ALL</sub>). The power difference among them is usually minor but SCORE.NULL.CT'<sub>FAM</sub> and SCORE.CT'<sub>ALL</sub> (SCORE.CT<sub>ALL</sub>) are significantly less powerful for the 4G (Figure 2-2).

SCORE.MERLIN tends to have a type I error rate to be right of the 95% confidence interval of the presumed value (0.008, 0.012), i.e., it has an inflated type I error. In contrast, SCORE.NULL.CT'<sub>FAM</sub> and SCORE.CT'<sub>ALL</sub> tend to have it left to the left of the confidence interval (e.g. Figure 2-3), i.e., have conservative type I error. SCORE.CT has higher power than

SCORE.CT<sub>ALL</sub>, which implies that when the sample size for each pedigree type is sufficiently large to estimate the variances and covariances of estimated IBD sharing for each individual pair, we can gain extra power by accounting for the variation between pairs of the same type. SCORE.CT'<sub>ALL-TC</sub> has higher power than SCORE.CT'<sub>ALL</sub>, which implies that the score test using theoretical correlations between IBD sharing is probably more robust to different pedigree types.

SCORE.MERLIN (HM.MERLIN) and SCORE.CT'<sub>FAM</sub> (HM.CT'<sub>FAM</sub>) might have a negative variance of the score statistic and fail to return the statistic and p-value. The negative variances we have observed appeared across the pedigree types for SCORE.MERLIN and/or HM.MERLIN, but for SCORE.CT'<sub>FAM</sub> and/or HM.CT'<sub>FAM</sub> we saw negative variances only when the 4G was considered. They usually came with non-normal traits and frequently occurred for the recessive models (models 3' and 3''). We reported the number of replicates with a negative variance for the score tests of both types (e.g. Table 2-9 for 4G power simulation) but didn't count them in the calculation of type I error rate and power.

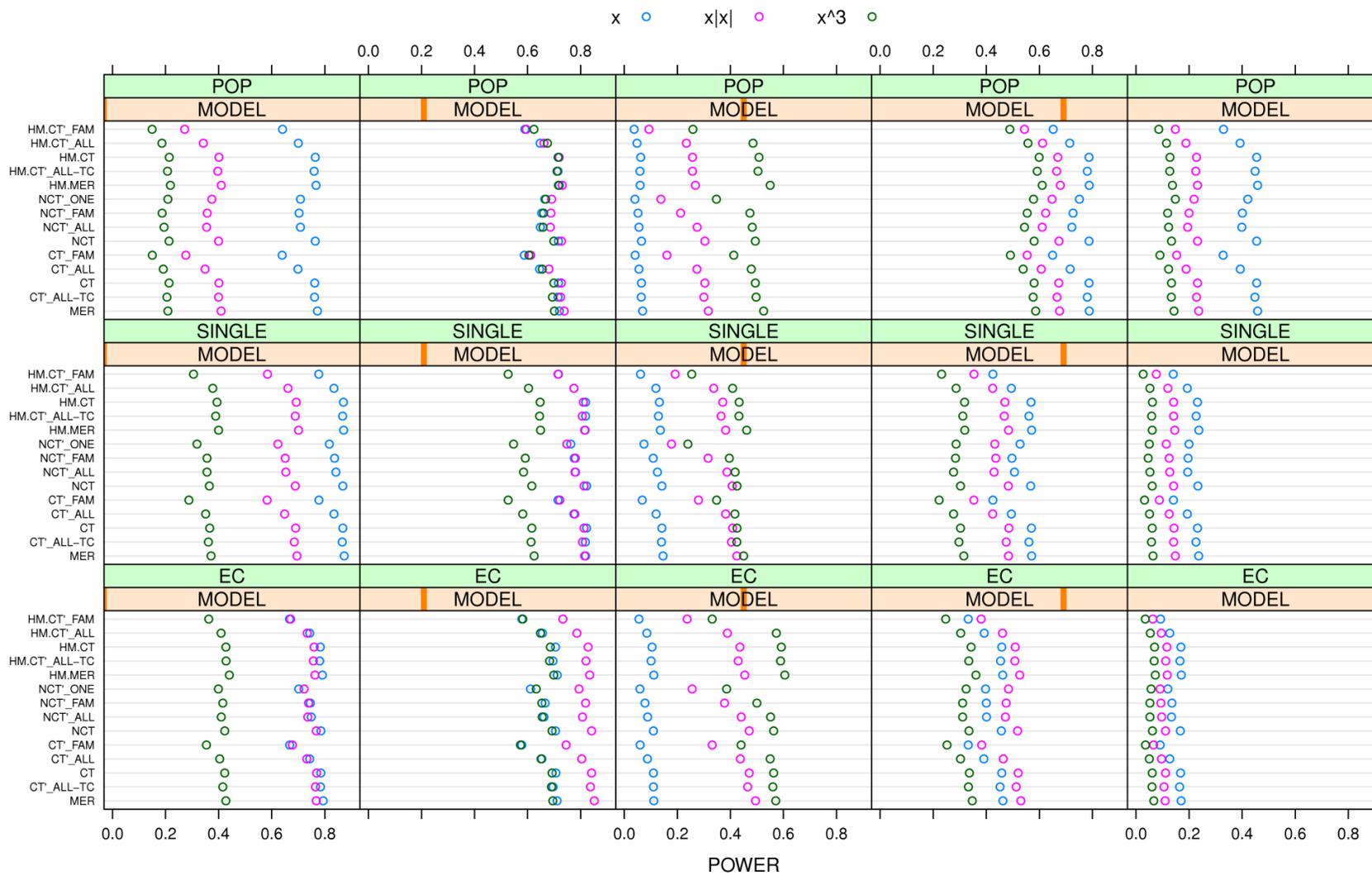
**Table 2-9. Number of replicates with a negative variance for MERLIN and CT'<sub>FAM</sub> type score tests given the 4G in power simulation**

POP	Trait model									
	1'	1''	2'	2''	3	3'	3''	4''	5''	
MERLIN	0	0	0	0	0	1	1	1	0	
CT' <sub>FAM</sub>	0	0	0	0	0	3	3	0	0	
HM.MERLIN	0	0	1	2	0	2	13	2	0	
HM.CT' <sub>FAM</sub>	0	0	0	1	0	4	1	0	0	
SEL	1'	1''	2'	2''	3	3'	3''	4''	5''	
MERLIN	0	0	0	2	0	2	9	0	0	
CT' <sub>FAM</sub>	0	0	0	0	0	3	2	1	2	
HM.MERLIN	0	0	0	2	0	2	20	0	0	
HM.CT' <sub>FAM</sub>	0	0	0	1	0	3	1	0	2	
EC	1'	1''	2'	2''	3	3'	3''	4''	5''	
MERLIN	0	0	0	1	0	2	3	0	0	
CT' <sub>FAM</sub>	1	0	0	0	1	2	2	0	0	
HM.MERLIN	1	0	0	2	0	4	8	1	0	
HM.CT' <sub>FAM</sub>	0	0	0	0	1	4	1	0	0	

Note: 1,000 replicates were used for power simulations.

Additionally, we dropped SCORE.CT'<sub>ALL-TC</sub> (HM.CT'<sub>ALL-TC</sub>) and SCORE.MERLIN (HM.MERLIN) for the 4SIBS+3G+HP due to the intensive computation necessary for the joint IBD distribution which is required by SCORE.MERLIN (HM.MERLIN) and we used it for

SCORE.CT<sub>ALL-CT</sub> (HM.CT<sub>ALL-CT</sub>) by using uninformative genotypes to obtain theoretical correlations between IBD sharing.



**Figure 2-2. 4G power simulation results**

Note: CT\*\_ALL-TC is SCORE.CT\*\_ALL-TC, NCT\*\_ALL is SCORE.NULL.CT\*\_ALL and so on. Power for SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL were not presented because they are exactly equivalent to SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL for single pedigree types.



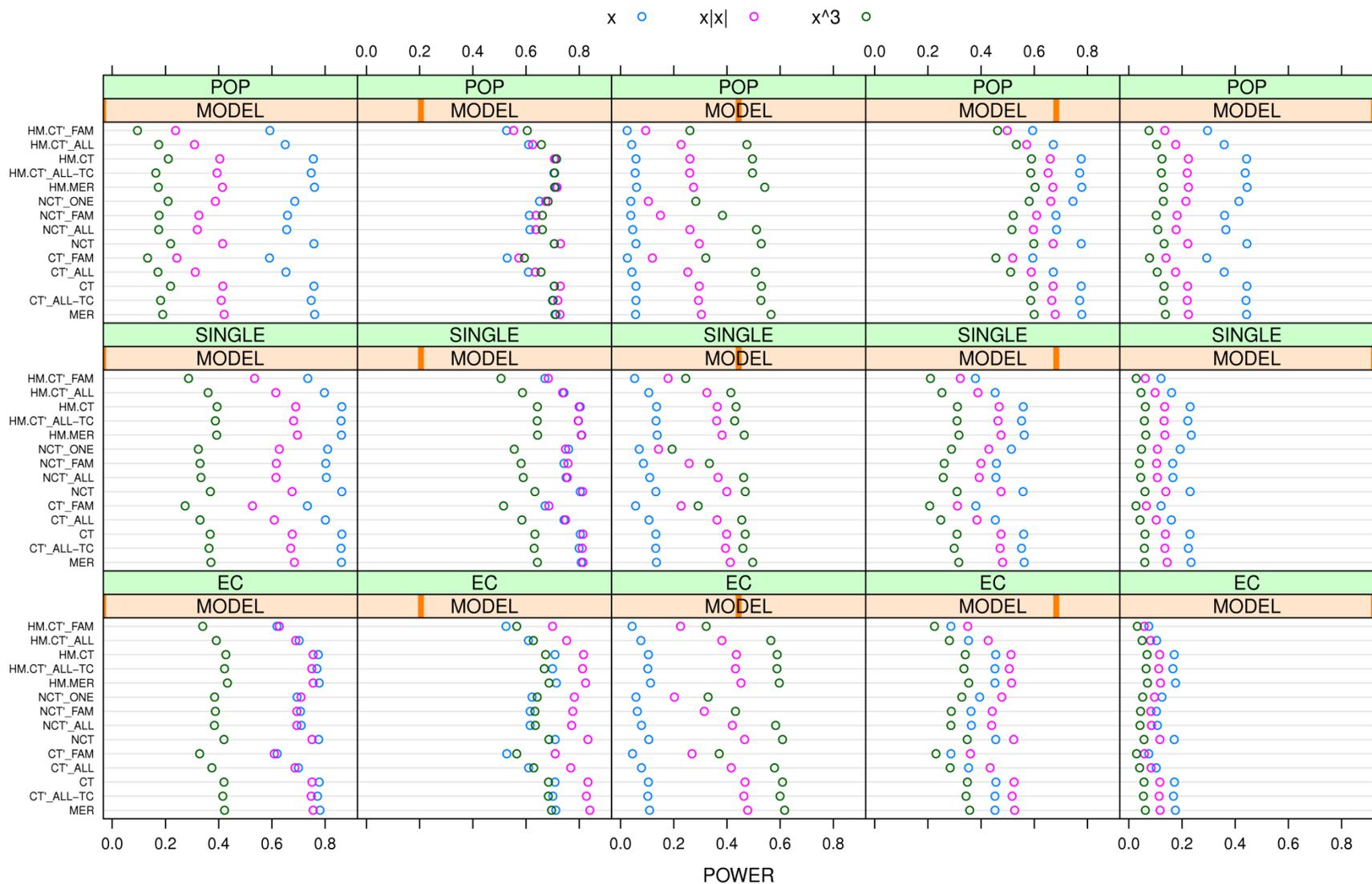
For the higher-moment (HM) score tests, similarly,  $HM.CT'_{FAM}$  always has lower power than other HM score tests.  $HM.CT'_{ALL}$  and  $HM.CT'_{ALL}$  are mathematically equivalent for the single pedigree types and our simulation results showed that they perform similarly for the mixed pedigree types.  $HM.CT'_{ALL}$  can have a significant power drop for the 4G.  $HM.CT'_{FAM}$ , like its LM version, has a conservative type I error rate and low power. HM score tests compared to LM score tests don't seem to have much power advantage. Both perform similarly for the normal trait,  $x$ . LM score tests are preferred over HM score tests for  $x|x|$  but the superiority is reversed for  $x^3$ . Since we have observed appreciable power gain for using LM score tests for  $x|x|$  (e.g. model 3'/POP/HS: ~15%; model 3'/POP, SEL, EC/3G: ~10%; model 3'/POP/4SIBS+3G: ~10%; model 3'/POP/ 4SIBS+3G+HP: ~10%), and for using HM score tests for  $x^3$  (e.g. model 1''/SEL/2+4SIBS: ~10%; model 2''/EC/2+4SIBS: ~10%) (Figures A3-1-A3-6), we recommend LM score tests for normal or moderately non-normal traits but HM score tests for extremely non-normal traits if skewness and kurtosis of the trait distribution are available.

## 2.5.2 Sensitivity analysis

Relative performance between the score tests using model-based correlations remains as that using true correlations. MERLIN, CT, and  $CT'_{ALL-TC}$  score tests using model-based correlations maintain a substantial amount of power while  $CT'_{ALL}$ ,  $CT'_{FAM}$ , and  $CT'_{ONE}$  might have a significant power drop (e.g. see Figures 2-4, 2-5 and Figures A3-7-A3-12). As for the comparison of LM and HM score tests, LM score tests are generally preferred over HM score tests if using model-based correlations estimated by parent-child pairs; with the use of model-based correlations estimated by sibling pairs, LM score tests have similar power to HM score tests for normal and moderately non-normal traits but slightly lower power than HM score tests for extremely non-normal traits. Overall, the score tests using model-based correlations estimated by parent-child pairs are slightly more powerful than those using model-based correlations estimated by sibling pairs and this implies that misspecifying the correlation for parent-child pairs or sibling pairs seems to have a similar effect on the power of score tests if the effect of other correlations is minor. For the trait models we have considered, we presume that sibling pairs have a higher phenotypic correlation than that of parent-child pairs (Table 2-5).

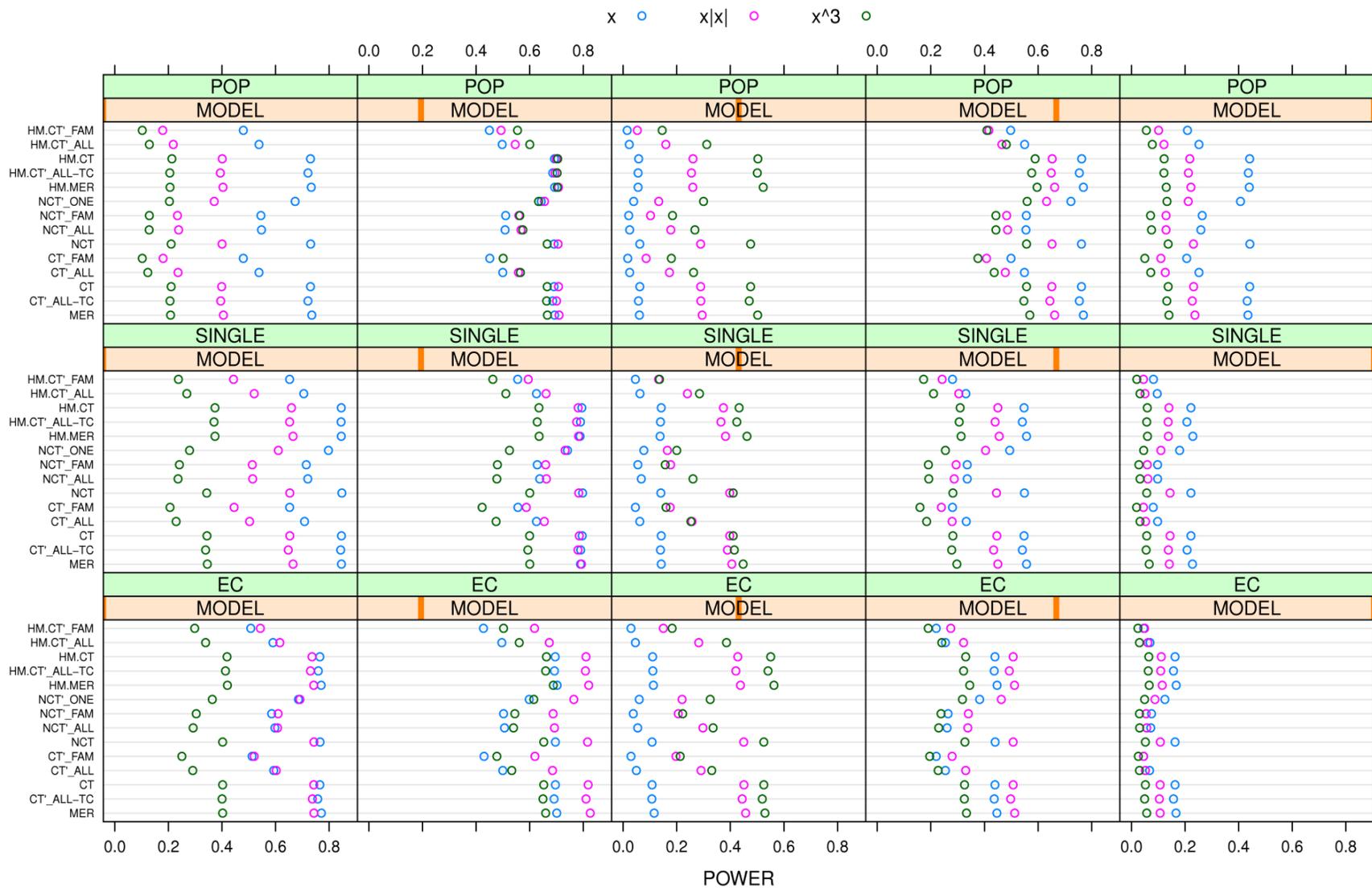
Model-based correlations estimated by parent-child pairs are derived based on a substantially over-estimated heritability including a true correlation for parent-child pairs and an under-estimated correlation for sibling pairs (Tables 2-6-2-8). Similarly, model-based correlations estimated by sibling pairs again are derived based on a substantially over-estimated heritability including a true correlation for sibling pairs and an over-estimated correlation for parent-child pairs (Tables 2-6-2-8).

Overall, SCORE.MERLIN, SCORE.CT and SCORE.CT'<sub>ALL-TC</sub> have robust power over the scenarios we have considered and the robustness remains for each using model-based correlations. HM score tests using model-based correlations compared to their lower-moment versions turn to be less powerful or have trivial power advantage for extremely non-normal traits but require specifying two more parameters so become less useful.



**Figure 2-4. 4G power simulation results using model-based correlations estimated using PC pairs**

Note: CT\_ALL-TC is SCORE.CT<sub>ALL-TC</sub>, NCT'\_ALL is SCORE.NULL.CT<sub>ALL</sub> and so on. Power for SCORE.CT<sub>ALL</sub>, SCORE.NULL.CT<sub>ALL</sub>, and HM.CT<sub>ALL</sub> were not presented because they are exactly equivalent to SCORE.CT'<sub>ALL</sub>, SCORE.NULL.CT'<sub>ALL</sub>, and HM.CT'<sub>ALL</sub> for single pedigree types.



**Figure 2-5. 4G power simulation results using model-based correlations estimated using SB pairs**

Note: CT\_ALL-TC is SCORE.CT<sub>ALL-TC</sub>, NCT'\_ALL is SCORE.NULL.CT<sub>ALL</sub> and so on. Power for SCORE.CT<sub>ALL</sub>, SCORE.NULL.CT<sub>ALL</sub>, and HM.CT<sub>ALL</sub> were not presented because they are exactly equivalent to SCORE.CT'<sub>ALL</sub>, SCORE.NULL.CT'<sub>ALL</sub>, and HM.CT'<sub>ALL</sub> for single pedigree types.

## 2.6 DISCUSSION

MERLIN, CT, and  $CT'_{ALL-TC}$  score tests have robust power for different pedigree types against non-normal traits and selected samples. They especially outperform other score tests for the single pedigree type 4G. 4G relative to other pedigree types has more levels but fewer sibling nodes. It gives pairs of various types, but GG pairs appear to dominate (1 SB, 10 GG, 1 HS, 3 HA, and 2 HC; see Table 2-3). By definition,  $CT_{ALL}$ ,  $CT'_{FAM}$ , and  $CT'_{ALL}$  estimate the covariance of estimated IBD sharing between a GG pair and a SB pair using 10 pairs of GG and SB produced by 10 GG pairs and 1 SB pair in a 4G pedigree. It is probably that the reduced variation of estimated IBD sharing in SB due to extremely unbalanced matching leads to biased covariance estimates and then decreases the power. The CT score tests (SCORE.CT, SCORE.NULL.CT, and HM.CT) don't have this balance problem; they use one pair of GG and SB in every 4G pedigree but requires pedigree type information, which is practically intractable. Rather than an empirical covariance,  $CT'_{ALL-TC}$  score tests estimate a covariance by two empirical variances and one theoretical correlation between IBD sharing. The most computationally demanding part for MERLIN and  $CT'_{ALL-TC}$  is to calculate the joint IBD distribution. MERLIN score tests require both prior and posterior joint IBD distributions to calculate the variances and covariances of estimated IBD sharing.  $CT_{ALL-TC}$  score tests use the prior one only to obtain theoretical correlations between IBD sharing, which alternatively can be estimated by simulation, but both probably have similar computational complexities.

Overall, unlike CT,  $CT'_{ALL-TC}$  score tests are not using pedigree type information and are computationally more efficient than MERLIN mainly for large pedigrees. The  $CT'_{ALL-TC}$  score test using model-based correlations can maintain a substantial amount of power and SCORE. $CT'_{ALL-TC}$  always has similar power to HM. $CT'_{ALL-TC}$ . We thus recommend SCORE. $CT'_{ALL-TC}$  using model-based correlations for general use. Other methods for heritability estimation or a more realistic trait model to derive model-based correlations might improve the power. The method adopted by us is easy to implement and guarantees appreciable power.

Our conclusions on the performance of the score tests don't completely agree with those by Dupuis, et al. [2009]. SCORE.NULL.CT'<sub>ONE</sub> and SCORE.CT'<sub>ALL-TC</sub> are essentially the score tests using *Empirical* and *Estimated Variances* in Dupuis, et al. [2009]. SCORE.CT'<sub>ALL-TC</sub> and the score test using *Estimated Variances* have slightly different definitions but are anticipated to perform similarly for modest genetic effects. In Dupuis, et al. [2009], it was assumed that both phenotypes and genotypes of grandparents are missing and their main conclusion is that the score tests using *Empirical* and *Estimated Variances* have conservative type I error rates and similar power. In our study, we used the data of grandparents and our simulation results showed that SCORE.CT'<sub>ALL-TC</sub> consistently has a reasonable type I error rate and higher power than SCORE.NULL.CT'<sub>ONE</sub> especially for normal or moderately non-normal traits and recessive trait models (e.g., Figure A3-3 for the 3G which is similar to the pedigree type considered in Dupuis, et al. 2009). The genotypes of grandparents can improve the power of score tests but don't affect the relative performance among the score tests. We further explored SCORE.NULL.CT'<sub>ONE</sub> and SCORE.CT'<sub>ALL-TC</sub> (Appendix A.4). Our analytical work relying on a multivariate normal assumption can show that both of them are asymptotically equivalent under the null hypothesis and when the alternative hypothesis is true SCORE.CT'<sub>ALL-TC</sub> is usually more powerful than SCORE.NULL.CT'<sub>ONE</sub>.

A wide variety of pedigree types might be seen in reality and the majority of them might have not been considered in our simulations. Generally speaking, in terms of pedigree diagrams, pedigrees with a *long* shape can create plenty of relative-pair types. When the number of pedigrees of this type is not sufficiently large, score tests using pedigree type information and some of those not using pedigree type information might be less powerful for insufficient pairs of each type, e.g. CT'<sub>ONE</sub> and CT'<sub>FAM</sub>, or *unbalanced* information carried by different pair types (CT'<sub>ALL</sub> and CT'<sub>FAM</sub>) in the estimation of variance and covariance of estimated IBD sharing. Pedigrees with a *wide* shape tend to produce few pair types with numerous pairs in each. CT'<sub>FAM</sub> score tests compared to other score tests benefit more from this pedigree type. CT score tests in contrast benefit less from this pedigree type. Other score tests are expected to perform similarly but CT'<sub>ONE</sub> probably requires a larger sample size. Be cautious that the power of the score statistic depends on how different pedigree types are mixed, although the power might be dominated by few pedigree types.

### **3.0 WHAT'S THE BEST STATISTIC FOR A SIMPLE TEST OF GENETIC ASSOCIATION IN A CASE-CONTROL STUDY?**

Our association work has been published in *Genetic Epidemiology*, volume 34, issue 3, pages 246-253 [Kuo and Feingold 2010]. Chapter 3 is almost the paper except that I moved the supplemental material to appendices and renumbered tables and figures. Additionally, I fixed few errors I found in the references. The journal grants the authors rights to include the article in part or in full in any publication provided that such re-used material constitutes less than half of the total material in such publication.

#### **3.1 ABSTRACT**

Genome-wide genetic association studies typically start with univariate statistical tests of each marker. In principle, this single-SNP scanning is statistically straightforward – the testing is done with standard methods (e.g. chi-squared tests, regression) that have been well-studied for decades. However, a number of different tests and testing procedures can be used. In a case-control study, one can use a 1 df allele-based test, a 1 df or 2 df genotype-based test, or a compound procedure that combines two or more of these statistics. Additionally, most of the tests can be performed with or without covariates included in the model. While there are a number of statistical papers that make power comparisons among subsets of these methods, none has comprehensively tackled the question of which of the methods in common use is best suited to univariate scanning in a genome-wide association study. In this paper, we consider a wide variety of realistic test procedures, and first compare the power of the different procedures to detect a single locus under different genetic models. We then address the question of whether or when it is a good idea to include covariates in the analysis. We conclude that the most

commonly-used approach to handling covariates - modeling covariate main effects but not interactions - is almost never a good idea. Finally, we consider the performance of the statistics in a genome scan context.

## 3.2 INTRODUCTION

Large-scale genetic association studies test for association between genotype and phenotype for as many as one million SNPs at a time. Analysis in these studies usually starts with univariate statistical tests of correlation between the phenotype and each individual marker genotype. In principle, this scanning is statistically straightforward - the testing is done with standard methods (e.g. chi-squared tests, regression) that have been well-studied for decades. However, a number of different tests can be used. For a case-control study, most people perform a genotype-based chi-squared test, but this can be a 1 or 2 df test on the  $2 \times 3$  table (3 genotypes), or a 1 df test that combines the heterozygote class with the rarer homozygote class. It is also common to perform scans with two or more of these statistics and consider the largest statistics from both lists. Such an approach should really be considered as a distinct test procedure, with the multiple testing issues properly accounted for. Some studies use logistic regression instead of chi-squared tests so that covariates can be incorporated into the initial scan. This presents essentially the same options as the chi-squared test for modeling the genotype using 1 or 2 degrees of freedom, but in addition the model can involve interactions between the genotypes and the covariates as well as main effects of the covariates if desired.

Surprisingly, we have been able to find relatively little literature comparing the power of different test procedures, and we have observed a huge variety of procedures used in real application. Several papers (see further discussion below) have compared 1 df genotype-based tests, but these typically have not considered the 2 df test. In this paper, we review the options for a simple genetic association test in a case-control study, summarize previous results, and then examine several unanswered questions about what the best association tests are. Specifically, we consider three issues. First, among tests that do not consider covariates, which has the most robust power to detect association between a trait and a single marker? Second, if we have important covariates but are primarily interested in testing the genetic effect, what logistic

regression model is best? We compare models with the genetic effect only, with an environmental covariate included, and with interaction. Finally, we simulate a simple genome-wide scenario and ask whether our single-locus conclusions are still appropriate in a genome scan context.

### 3.3 MATERIALS AND METHODS

#### 3.3.1 TEST PROCEDURES

In a case-control genetic association study, the genotypic data for each biallelic marker can be summarized as Table 3-1, which presents the numbers of cases and of controls with 0, 1, and 2 copies of the minor allele. This can be transformed to a  $2 \times 2$  allele-based table as Table 3-2. Or, to avoid the problem of sparse cells, one might combine the heterozygote class with the rarer homozygote class as Table 3-3.

**Table 3-1.  $2 \times 3$  genotype-based table**

	Number of minor alleles			
	0	1	2	Total
Case	$r_0$	$r_1$	$r_2$	$R$
Control	$s_0$	$s_1$	$s_2$	$S$
Total	$n_0$	$n_1$	$n_2$	$N$

**Table 3-2.  $2 \times 2$  allele-based table**

	Major allele	Minor allele	Total
Case	$2r_0 + r_1$	$2r_2 + r_1$	$2R$
Control	$2s_0 + s_1$	$2s_2 + s_1$	$2S$
Total	$2n_0 + n_1$	$2n_2 + n_1$	$2N$

**Table 3-3. 2×2 genotype-based table combing the rarer homozygote class with the heterozygote class**

	Without minor allele	With minor allele	Total
Case	$r_0$	$r_2 + r_1$	$R$
Control	$s_0$	$s_2 + s_1$	$S$
Total	$n_0$	$n_2 + n_1$	$N$

Based on these tables, we consider several different statistical tests. We refer to the chi-squared test of independence on Table 3-1 as the 2×3 two df test. We refer to the chi-squared test of independence on Table 3-2 as the 2×2 allele test. We refer to the chi-squared test of independence on Table 3-3 as the 2×2 geno test. Finally, we refer to the linear trend test with score vector (0, 1, 2) on Table 3-1 as the 2×3 trend test. Note that the 2×2 geno test can also be considered a trend test with score vector (0, 1, 1). These test statistics can be written as follows, using the notation presented in the tables.

$$\chi_{2 \times 3 \text{ two df}}^2 = \sum_{i=0}^2 \left( \frac{Nr_i - n_i R}{n_i R} + \frac{Ns_i - n_i S}{n_i S} \right) \stackrel{H_0}{\sim} \chi_2^2$$

$$\chi_{2 \times 2 \text{ allele}}^2 = \frac{2N \left[ (2r_0 + r_1)(2s_2 + s_1) - (2r_2 + r_1)(2s_0 + s_1) \right]^2}{(2R)(2S)(2n_0 + n_1)(2n_2 + n_1)} \stackrel{H_0}{\sim} \chi_1^2$$

$$\chi_{2 \times 2 \text{ geno}}^2 = \frac{N \left[ r_0 (s_1 + s_2) - s_0 (r_1 + r_2) \right]^2}{RSn_0 (n_1 + n_2)} \stackrel{H_0}{\sim} \chi_1^2$$

$$\chi_{2 \times 3 \text{ trend}}^2 = \frac{N \left[ N(r_1 + 2r_2) - R(n_1 + 2n_2) \right]^2}{RS \left[ N(n_1 + 4n_2) - (n_1 + 2n_2)^2 \right]} \stackrel{H_0}{\sim} \chi_1^2$$

It is also possible to construct a “recessive” test using the trend test with score vector (0, 0, 1). We denote this as REC. We don’t consider it as an independent test procedure, but we do incorporate it into the compound statistics which combine two or more statistics. We consider three compound statistics, as defined as follows.

min 4p = minimum p-value of 2×3 trend, 2×2 geno, REC, and 2×3 two df

min 3p = minimum p-value of 2×3 trend, 2×2 geno, and REC

min 2p = minimum p-value of 2×2 geno and 2×3 two df

The statistic min 4p is provided by PLINK [Purcell, et al. 2007] which is a popular software package for genome-wide association analysis. The statistic min 3p is suggested by Freidlin, et al. [2002] for general use in the framework of trend tests, and is often referred to as  $Z_{MAX}$ . Since the asymptotic distributions of the compound statistics are not easily available, we calculate p-values by simulation.

We alter some of these test statistics slightly in order to mimic their real use in a genome-scan context. In a genome scan, loci with rare alleles will cause problems for the REC and 2×3 two df statistics. In our simulations, when  $n_2$  is less than 10 for any given dataset, we apply the 2×2 geno test instead of the 2×3 two df test and skip REC in the compound tests. Thus, for example, our version of min 2p is precisely defined as

min 2p = minimum p-value of 2×2 geno and 2×3 two df, when  $n_2 \geq 10$   
p-value of 2×2 geno, when  $n_2 < 10$

In addition to the tests defined above, association between phenotype and genotype can be modeled using logistic regression with dummy coding for the three genotypes. Any covariates suspected of affecting the disease can be included in order to improve the accuracy of the model, but in terms of hypothesis testing, the addition of covariates is punished by the loss of degrees of freedom. We consider three different logistic regression models for testing association. For the sake of simplicity, we consider only a single binary covariate. Let  $G$  be genotype coded as 0, 1, and 2, and  $E$  be the binary exposure covariate coded as 1 for the exposed and 0 for the non-exposed. We then consider three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G \times E$ . We test the genetic effect in the  $G$  model using the likelihood ratio test to compare with the null model. For the  $G+E$  and  $G+E+G \times E$  models, we test for genetic effect by comparing with the  $E$  model. Note that this is a 2 df test for the interaction model - we believe that this is the most appropriate test to perform when this model is being used primarily to test for genetic effect. It is probably most common in genetic association studies to use the  $G+E$  model, but we consider the

question of when that model has an advantage over the  $G$  model (again, for the specific purpose of testing genetic effect), and when the interaction model might have an advantage over both.

### 3.3.2 PREVIOUSLY ESTABLISHED RESULTS

Summarizing previous power comparison results for single-locus tests [Freidlin, et al. 2002; Guedj, et al. 2007; Li, et al. 2009b; Ohashi, et al. 2001; Sasieni 1997; Zheng, et al. 2003; Zheng and Ng 2008], it is fairly clear that for additive models the preferred test is  $2 \times 3$  trend or  $2 \times 2$  allele. For recessive models, the best choice is REC or  $2 \times 3$  two df. For dominant models the best choice is  $2 \times 2$  geno or  $2 \times 3$  two df. The  $Z_{MAX}$  test has robust power for additive, dominant, and recessive models. Previous studies have not considered over-dominant or under-dominant models, but it is fairly clear that the  $2 \times 3$  two df test would be best for these. What is not clear is which test(s) are most powerful for intermediate models, and which have robust power over the widest variety of models. Although a number of authors have suggested that compound statistics such as  $Z_{MAX}$  might meet this need, there have not been systematic comparisons among different statistics of that type.

There is a much more limited amount of previous literature comparing methods for incorporating covariates. Kraft, et al. [2007] and Selinger-Leneman, et al. [2003] compared the models  $G$  and  $G+E+G \times E$  and both concluded that  $G+E+G \times E$  has much higher power than  $G$  when the genetic effect expresses only in the presence of the exposure, and that the power gain is decreased with the exposure frequency, so that the simple model  $G$  might be appropriate for a common “exposure” such as smoking or sex. Neither of these studies considered the model  $G+E$ , although this is probably the most commonly-used model in genetic association studies (certainly more common than an interaction model in genome-wide studies). Moreover, both of these studies only considered the situation in which the interaction model fit to the data is actually the correct model, a scenario that is unlikely to be the case in a real dataset.

There has also been very limited literature looking at how these issues play out in a genome-scan setting. Some studies have looked at how many markers are required for a follow-up study [Gail, et al. 2008; Zaykin and Zhivotovsky 2005], but we focus on comparing the different test procedures. We use the success metric “detection rate” which we present as  $A(B)$  where  $A$  is the average number of true positive loci that are detected (averaged over many

simulation replicates) referenced by the total number,  $B$ , of truly associated loci. It is similar to the detection probability (DP) used by Gail et al. [2008]. We consider top-10, top-20, and top-30 gene lists, chosen by ranking SNPs by p-value for single test procedures and by minimum p-value for compound procedures, though it would also be possible to use other ranking criteria. For example, the distribution of  $Z_{MAX}$  under the null depends on the correlation between statistics in terms of allele frequencies so the  $Z_{MAX}$  of different SNPs might not be comparable. Ranking SNPs by the p-value, however, requires more computational effort [Li, et al. 2008b]. Li et al. [2008a] compared the  $Z_{MAX}$  and its p-value in the ranking of true associations and found that both have similar performance; thus, recommended the  $Z_{MAX}$  for simplicity of computation. Zaykin, et al. [2008] proposed to use the p-value of a case-control association test for rare associated SNPs and the allele frequency difference between cases and controls for common associated SNPs.

### **3.3.3 METHODS FOR POWER CALCULATION**

Comparing power of 1 df and 2 df statistics using asymptotic approximations is methodologically challenging, because the differences in power can be smaller than the errors in the asymptotic approximations. Briefly, a number of approximate power functions for independence chi-squared tests and trend tests have been proposed in the literature [Bukhszar and van den Oord 2006; Chapman and Nam 1968; Ferguson 1996; Jackson, et al. 2002; Slager and Schaid 2001]. Power comparison in principle can be done by plugging alternative parameters into approximate power functions but most of them require that each expected cell count is greater than 5, which might not be true in reality. Moreover, a number of the statistics that we want to consider, such as the compound statistics, do not have asymptotic power functions. All of the power comparisons in this paper are thus based on the gold standard of simulations with a large number of replicates, avoiding the pitfalls of analytical comparisons by the less elegant but more accurate application of substantial computational firepower.

### 3.4 SIMULATION METHODS

#### 3.4.1 SINGLE-LOCUS ANALYSIS

We first performed simulations to answer the question of what statistic is best for single-locus testing without consideration of covariates. We start with the following model for a single marker in LD with the disease locus. Suppose the marker has minor allele  $A$  and major allele  $a$ , occurring with frequencies  $p$  and  $q$ . Assume that the overall population is in HWE. Given the disease status, the numbers of each genotype in cases and in controls are trinomially distributed with equal sample sizes and the probabilities as following:

$$\left\{ \begin{array}{l} P(aa | case) = \frac{q^2 f_0}{K} \\ P(Aa | case) = \frac{2pqf_1}{K} \\ P(AA | case) = \frac{p^2 f_2}{K} \end{array} \right. \quad \left\{ \begin{array}{l} P(aa | control) = \frac{q^2(1-f_0)}{1-K} \\ P(Aa | control) = \frac{2pq(1-f_1)}{1-K} \\ P(AA | control) = \frac{p^2(1-f_2)}{1-K} \end{array} \right.$$

where  $f_0$ ,  $f_1$ , and  $f_2$  are the penetrances for  $aa$ ,  $Aa$ , and  $AA$ , and  $K$  is the disease prevalence defined as  $q^2 f_0 + 2pqf_1 + p^2 f_2$ . Under the null,  $f_0 = f_1 = f_2 = K$  and the probabilities used to simulate  $aa$ ,  $Aa$ , and  $AA$  for cases and for controls are  $q^2$ ,  $2pq$ , and  $p^2$ . We considered three minor allele frequencies, 0.05, 0.1, and 0.3, and seven locus effects as defined in Table 3-4 to total 21 genetic models.

**Table 3-4. Genetic models**

Sig.	$p^a$	Model	$f_0^b$	$f_1^c$	$f_2^d$
0.05	0.05	M1. add	0.01	0.015	0.02
		M2. rec	0.01	0.01	0.015
0.001	0.1	M3. dom	0.01	0.015	0.015
		M4. over-dom	0.01	0.02	0.015
0.0001	0.3	M5. over-dom	0.01	0.02	0.01
		M6. under-dom	0.015	0.01	0.02
		M7. under-dom	0.02	0.01	0.02

<sup>a</sup>Minor allele frequency. <sup>b, c, d</sup>Penetrances for the genotypes with 0, 1, and 2 copies of minor alleles. Each model has a disease prevalence between 0.01 and 0.02. We considered the significance levels 0.05, 0.001, and 0.0001. The sample sizes were chosen to achieve power around 0.8 and the number of replicates used for the null and the alternative simulation is 100,000.

### 3.4.2 SCAN WITH COVARIATES

Our second set of simulations was aimed at comparing the power of the  $G$ ,  $G+E$ , and  $G+E+G\times E$  models. We considered “exposure” frequencies of 0.15, 0.3, 0.5, 0.7, and 0.85 and assumed that genotype and exposure are independent. We simulated data under four models: genetic effect only, genetic and exposure main effects only, gene $\times$ exposure interaction in which there is only a genetic effect in the “exposed” group, and gene $\times$ exposure interaction with effects in both groups. The penetrances for all models are given in Table B2-1. We used an allele frequency of 0.3 for all models and simulated 1,000 replicates for each power simulation. Sample sizes differed for each model, and were chosen to achieve power around 0.8 at the exposure frequency 0.5 so that different methods could be easily compared (Table B2-1).

### 3.4.3 GENOME-WIDE ANALYSIS

The question of interest in our genome-wide simulations was whether our conclusions based on single-marker are still valid when we consider how genome-scans are performed in practice. In the genome scan context, our outcome measure is no longer power to detect a single locus. We consider instead the more realistic success metric of the expected number of true positive loci that are detected, in other words, that appear on a “most significant” gene list. We consider “top- $k$ ” lists for  $k = 10, 20,$  and  $30$ .

Our genome-wide simulations assume that all markers are independent and in HWE. This is a simple model, but it is sufficient to ask basic questions about how statistics compare. For example, Zaykin and Zhivotovsky [2005] showed that realistic linkage disequilibrium structures have little effect on the ranks of true positive loci. We simulated 100,000 markers not associated with the disease, with minor allele frequencies chosen from the uniform distribution on (0.05, 0.5). We simulated ten “alternative hypothesis” markers, with the minor allele frequency 0.1, in LD with disease loci. Of these, six were additive, two dominant, and two recessive. We did our simulations for two sample sizes, 500 and 1500, to mimic both underpowered and adequately powered studies. To assure that the 10 markers were approximately equally likely to be detected, we chose genetic models that had approximately equal power based on single-marker

simulations (Table B2-2). We let the penetrance for an individual without any risk alleles be 0.01 and the effects of the 10 loci be additive, as expressed by

$$penetrance = 0.01 + 0.005 (ADD_i + REC_j + DOM_k), i = 1, \dots, 6, j, k = 1, 2$$

$$\text{where } ADD_i = \begin{cases} 0, & \text{if } aa \\ 1, & \text{if } Aa \\ 2, & \text{if } AA \end{cases}, REC_j = \begin{cases} 0, & \text{if } aa \\ 0, & \text{if } Aa \\ 6, & \text{if } AA \end{cases}, DOM_k = \begin{cases} 0, & \text{if } aa \\ 1, & \text{if } Aa \\ 1, & \text{if } AA \end{cases}$$

This yields a disease prevalence of 0.0185. We performed 100 replicates of our genome-wide simulation.

## 3.5 RESULTS

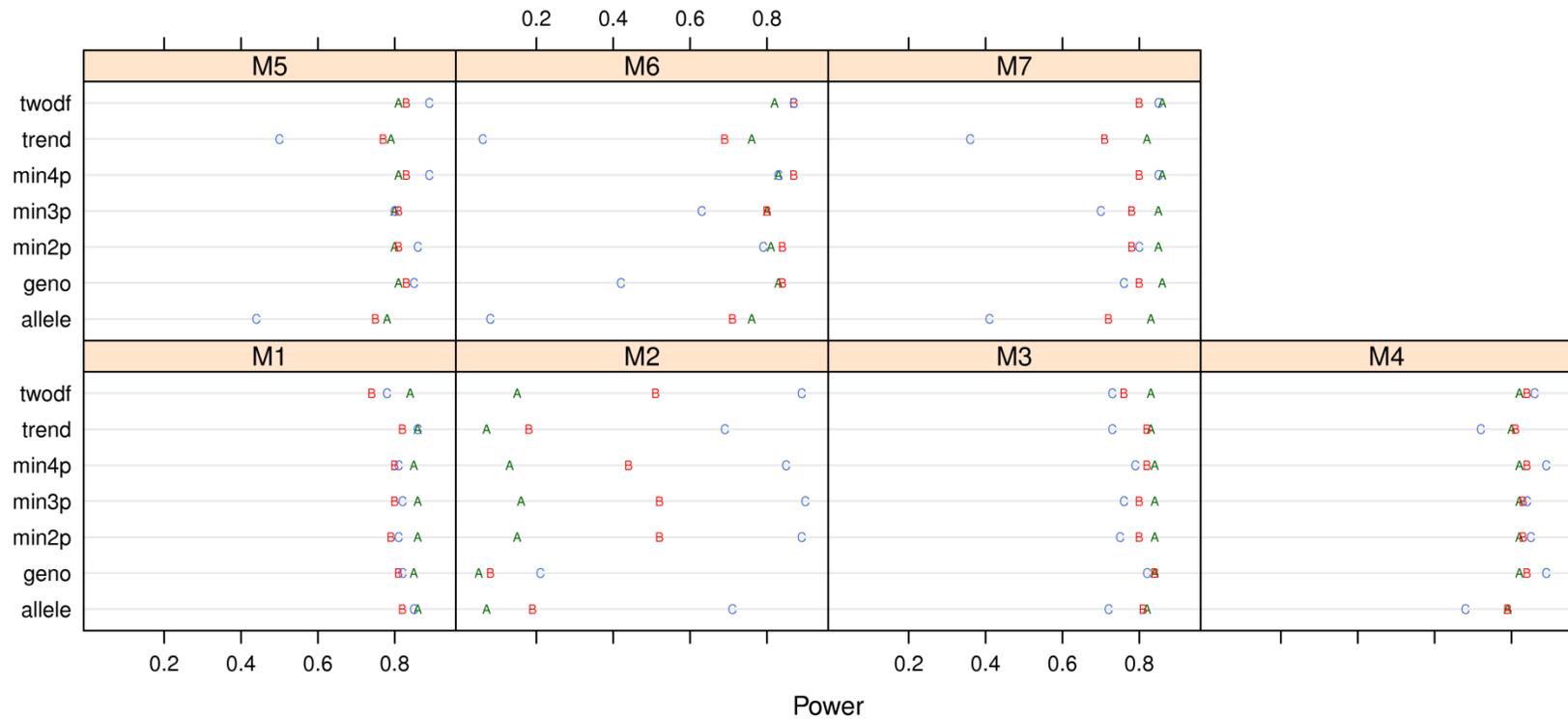
### 3.5.1 SINGLE-LOCUS ANALYSIS

In single-locus analysis, we considered the allele-based test ( $2 \times 2$  allele), three genotype-based tests ( $2 \times 2$  geno,  $2 \times 3$  trend,  $2 \times 3$  two df), and three compound statistics (min 4p, min 3p, min 2p) that combine two or more single statistics. Our goals were to find the most robust test procedure over a wide variety of genetic models and to understand the performance of different test procedures under different genetic models. Results for the significance level 0.05 are provided as an example - other significance levels gave consistent conclusions.

In the null simulation, the nominal type I error of 5% was completely within the 95% confidence intervals for all scenarios (results not shown). In other words, the type I error is correct in all scenarios. For the power simulation, the estimated power for different minor allele frequencies has been summarized in Figure 3-1. When the allele frequency  $p = 0.05$  (“A” in Figure 3-1), the test procedures perform similarly for all models. When  $p = 0.1$  (“B” in Figure 3-1),  $2 \times 3$  two df and the compound statistics have much higher power than the other statistics for detecting a recessive locus. When  $p = 0.3$  (“C” in Figure 3-1), there are obvious power

differences among the test procedures.  $2 \times 3$  two df has much higher power for the over-dominant and the under-dominant models, as well as for the recessive.

Overall,  $2 \times 2$  geno is optimal for the dominant but has poor power for the recessive.  $2 \times 2$  allele and  $2 \times 3$  trend are asymptotically equivalent and are best for the additive but generally less powerful for the others (see analytical analysis for  $2 \times 2$  allele and  $2 \times 3$  trend in Appendix B.1).  $2 \times 3$  two df is the only single test procedure that performs well for the recessive. It also has much higher power for the under-dominant and the over-dominant, and is nearly optimal for the additive and the dominant. The compound statistics clearly have the most robust power, although  $\min 3p$  has much less power for the under-dominant and the over-dominant models because it does not incorporate  $2 \times 3$  two df. We conclude that  $2 \times 3$  two df, and the compound statistic  $\min 4p$ , have the most robust power over a wide variety of models. In the discussion we touch on the issue of whether robustness over *all* models is in fact desirable, however.

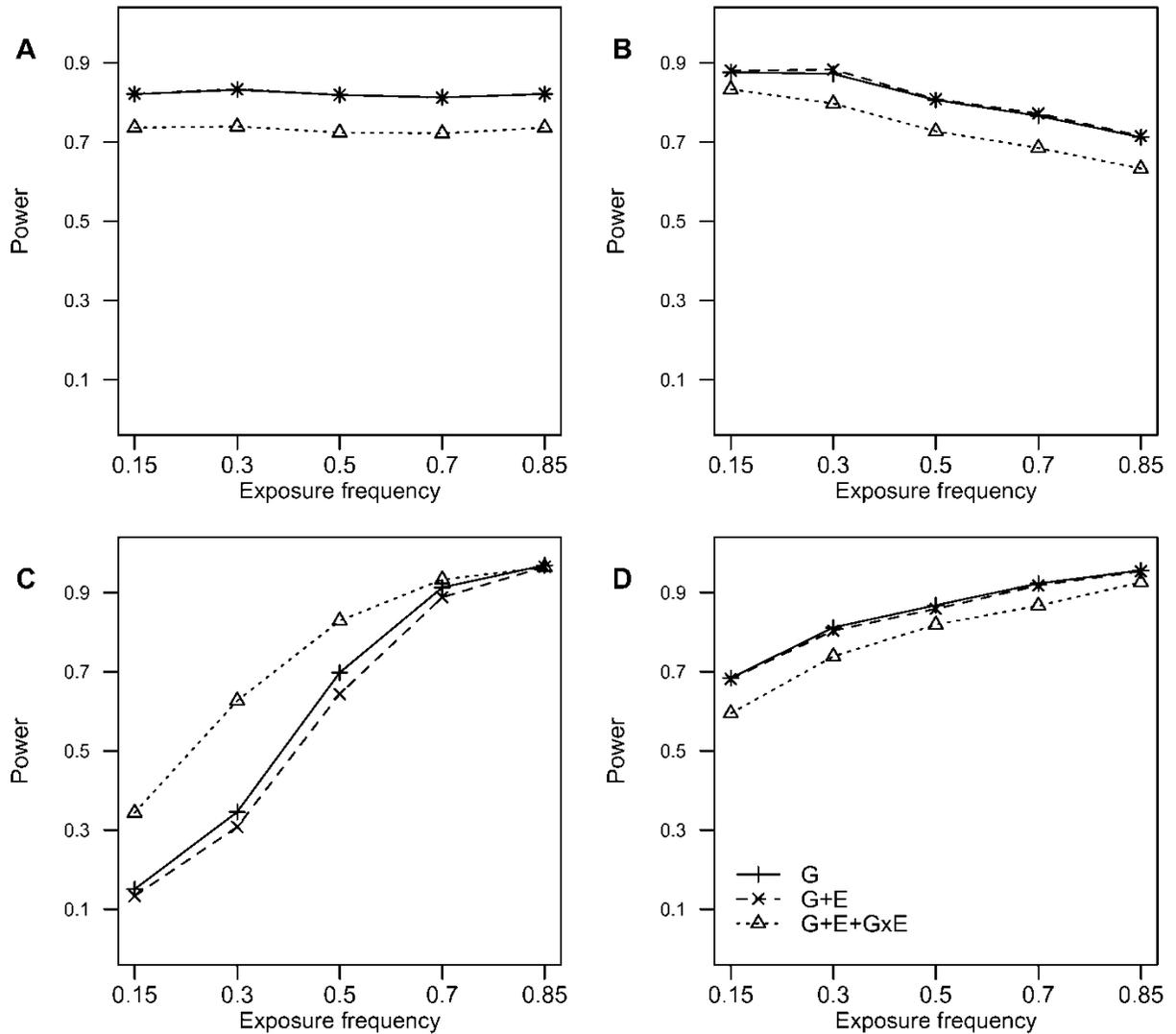


**Figure 3-1. Single-locus power simulation results for the minor allele frequencies: A. 0.05, B. 0.1, and C. 0.3. M1-M7 are the models as defined in Table 3-4. X-axis is power and y-axis is statistic name.**

### 3.5.2 SCAN WITH COVARIATES

We studied three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G\times E$ , with the goal of asking which model is most efficient for detecting a genetic effect. We simulated data under the models described above, which included additive, recessive, dominant, over-dominant, and under-dominant models, each with genetic effect only, genetic and exposure effects, and gene $\times$ exposure interaction. In all fitted models, the genotype  $G$  was coded as 0, 1, and 2, as would most likely be done in a real analysis in which the true genetic models are unknown.

Results for the additive effect are summarized in Figure 3-2; the others are presented in Figures B2-1-B2-4. We also repeated the simulations for the minor allele frequency of 0.1 and obtained essentially the same results. When there is no exposure effect or no interaction between the genotype and the exposure, the fitted logistic regression model  $G$  is as powerful as  $G+E$ . Both consistently have 8-10% higher power than  $G+E+G\times E$  (Figure 3-2 A and B). When there is interaction and the genetic effect is expressed in both exposed and non-exposed groups (Figure 3-2 D),  $G$  still has the highest power. This is probably because the marginal genetic effect evaluated under the 1df chi-square is larger than the genetic effect in either group evaluated under the 2df chi-square. When there is interaction and the genetic effect is expressed only in the exposed group (Figure 3-2 C),  $G+E+G\times E$  has higher power than  $G$  for most exposure frequencies. In a short,  $G$  is the best model for detecting genetic effect except if there is quite strong interaction, in which case the interaction model is best. The  $G+E$  model, which is probably the most frequently used in real studies, was *never* the best choice in our simulations.



**Figure 3-2. Additive marker power simulation results for the three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G \times E$  at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene  $\times$  exposure interaction in which there is only a genetic effect in the exposed group, and D. gene  $\times$  exposure interaction with effects in both groups.**

### 3.5.3 GENOME-WIDE ANALYSIS

Single-locus simulation results answer the question of what the best test procedure is for a single-locus analysis. But in the context of a genome scan, there are two major things that are different from our set-up above. First, instead of a single locus, we envision a moderate number of true positive loci, with a variety of genetic models. Second, because we are changing our success metric in the genome-scan context, it is not guaranteed that our best test procedure chosen by the single-test power metric will still be the best. We considered 10 truly associated and 100,000 non-associated mutually independent markers and were interested in the question of which test procedure could help identify most true positive loci on average.

Table 3-5 gives results for the “underpowered” study with sample size 500, in which detection rates for all statistics are low. When we look at only a top-10 or top-20 list of results, detection rates appear to be driven entirely by whether the “best” locus is picked up in a given replicate, and thus the best statistics are those with highest power for specific models. The compound statistics give up some power to detect that best locus in order to gain a robustness that turns out not to be useful on average when the overall power is so low. But when we examine the top-30 list (which is somewhat similar to increasing power), the compound statistics have the highest power. Table 3-6 gives results for the study with sample size 1500. In this case, overall detection rates are much improved, and the compound statistics consistently have the highest detection rates. We did not perform simulations with very large sample sizes and high detection rates, but based on our results we would expect the following. For the loci of major effect, detection rates would be very high and the choice of statistic would not matter much. For loci of more moderate effect, detection rates would be in the range of what we have tested in our simulation studies, so similar results would hold.

**Table 3-5. Genome-wide simulation results for the sample size 500**

List	Test	Detection rate				# of replicates in 100			
		ADD <sup>a</sup>	DOM <sup>b</sup>	REC <sup>c</sup>	ALL <sup>d</sup>	0 <sup>e</sup>	1 <sup>f</sup>	2 <sup>g</sup>	3 <sup>h</sup>
Top-10	2×2 geno	0.04	0.05	0.00(2)	0.09 (10)	91	9	0	0
	2×3 two	0.02	0.04	0.03(2)	0.09 (10)	92	7	1	0
	2×2 allele	0.06	0.03	0.01(2)	0.10 (10)	91	8	1	0
	2×3 trend	0.06	0.03	0.00(2)	0.09 (10)	91	9	0	0
	min 4p	0.03	0.01	0.01	0.05 (10)	95	5	0	0
	min 3p	0.04	0.02	0.02	0.08 (10)	92	8	0	0
	min 2p	0.03	0.02	0.00	0.05 (10)	95	5	0	0
Top-20	2×2 geno	0.09	0.11	0.00	0.20 (10)	81	18	1	0
	2×3 two	0.05	0.08	0.04	0.17 (10)	84	15	1	0
	2×2 allele	0.11	0.06	0.01	0.18 (10)	85	13	1	1
	2×3 trend	0.10	0.09	0.01	0.20 (10)	82	17	1	0
	min 4p	0.06	0.05	0.03	0.14 (10)	87	12	1	0
	min 3p	0.07	0.05	0.03	0.15 (10)	86	13	1	0
	min 2p	0.07	0.10	0.03	0.20 (10)	82	17	1	0
Top-30	2×2 geno	0.10	0.13	0.00	0.23 (10)	79	19	2	0
	2×3 two	0.07	0.10	0.04	0.21 (10)	80	19	1	0
	2×2 allele	0.13	0.10	0.01	0.24 (10)	79	19	1	1
	2×3 trend	0.13	0.10	0.01	0.24 (10)	79	19	1	1
	min 4p	0.10	0.10	0.06	0.26 (10)	76	23	1	0
	min 3p	0.10	0.10	0.06	0.26 (10)	76	23	1	0
	min 2p	0.10	0.12	0.04	0.26 (10)	76	23	1	0

<sup>a, b, c, d</sup>Detection rate (DR) of additive, dominant, recessive, and overall true positive loci where DR = average number of detected true positive loci (total number of true positive loci).

<sup>e, f, g, h</sup>Number of replicates detecting 0, 1, 2, 3 true positive loci.

**Table 3-6. Genome-wide simulation results for the sample size 1500**

List	Test	Detection rate				# of replicates in 100			
		ADD <sup>a</sup>	DOM <sup>b</sup>	REC <sup>c</sup>	ALL <sup>d</sup>	0 <sup>e</sup>	1-2 <sup>f</sup>	3-4 <sup>g</sup>	5-6 <sup>h</sup>
Top-10	2×2 geno	0.97 (6)	0.27 (2)	0.00 (2)	1.24	29	61	10	0
	2×3 two	0.72 (6)	0.15 (2)	0.41 (2)	1.28	26	64	10	0
	2×2 allele	1.11 (6)	0.19 (2)	0.05 (2)	1.35	24	63	13	0
	2×3 trend	1.14 (6)	0.20 (2)	0.05 (2)	1.39	25	60	15	0
	min 4p	0.85 (6)	0.17 (2)	0.45 (2)	1.47	23	60	17	0
	min 3p	0.89 (6)	0.18 (2)	0.46 (2)	1.53	23	57	20	0
	min 2p	0.88 (6)	0.20 (2)	0.32 (2)	1.40	26	60	14	0
Top-20	2×2 geno	1.27 (6)	0.37 (2)	0.00 (2)	1.64	22	53	25	0
	2×3 two	0.94 (6)	0.26 (2)	0.55 (2)	1.75	17	56	27	0
	2×2 allele	1.39 (6)	0.27 (2)	0.10 (2)	1.76	20	51	27	2
	2×3 trend	1.40 (6)	0.29 (2)	0.06 (2)	1.75	19	52	29	0
	min 4p	1.13 (6)	0.26 (2)	0.58 (2)	1.97	17	53	25	5
	min 3p	1.15 (6)	0.27 (2)	0.58 (2)	2.00	17	50	28	5
	min 2p	1.18 (6)	0.33 (2)	0.46 (2)	1.97	19	46	33	2
Top-30	2×2 geno	1.48 (6)	0.45 (2)	0.00 (2)	1.93	15	52	32	1
	2×3 two	1.12 (6)	0.30 (2)	0.60 (2)	2.02	13	54	25	8
	2×2 allele	1.65 (6)	0.36 (2)	0.11 (2)	2.12	16	40	40	4
	2×3 trend	1.66 (6)	0.36 (2)	0.07 (2)	2.09	15	43	39	3
	min 4p	1.27 (6)	0.32 (2)	0.62 (2)	2.21	15	43	34	8
	min 3p	1.31 (6)	0.35 (2)	0.64 (2)	2.30	12	46	33	9
	min 2p	1.28 (6)	0.38 (2)	0.54 (2)	2.20	14	46	34	6

<sup>a, b, c, d</sup>Detection rate (DR) of additive, dominant, recessive, and overall true positive loci where DR = average number of detected true positive loci (total number of true positive loci).

<sup>e, f, g, h</sup>Number of replicates detecting 0, 1-2, 3-4, 5-6 true positive loci.

We must note that our results are very dependent on the mixture of models for the true loci that we chose, as well as on the number of truly associated loci. Had we chosen, for example, additive models for all or most of the truly associated loci, then the statistics that are powerful for additive models (e.g. 2×3 trend) would have also been most powerful in our genome scan results. On the other hand, if there was only one truly associated locus (or only one strong enough to detect), then our scan might fail entirely unless we used one of the statistics that has robust power. We expand on these points in the discussion.

## 3.6 DISCUSSION

### 3.6.1 SINGLE-LOCUS ANALYSIS

Based on our simulation results, we have shown that the  $2 \times 3$  two df test is the best single test procedure for the recessive and nearly optimal for the additive and the dominant. The compound statistics are even more robust, although  $\min 3p$  ( $Z_{\text{MAX}}$ ) doesn't incorporate  $2 \times 3$  two df and thus loses much power for the over-dominant and the under-dominant models.  $\min 2p$  has slightly lower power for the recessive model, but outperforms  $\min 3p$  and  $\min 4p$  for the over-dominant and the under-dominant models. All of this is quite consistent with previous results in the literature (summarized in our Materials and Methods), except that previous studies did not consider over-dominant and under-dominant models, nor did they compare different compound statistics.

Our work does not, unfortunately, provide a simple answer to the question of what statistic is best, because that depends on one's *a priori* beliefs about what genetic models are likely. For example, if one believes that over-dominant and under-dominant models are exceedingly rare, then it is unnecessary to choose a statistic that has robust power for detecting them, and so  $\min 3p$  is probably the best choice. In the discussion below of our genome-scan results, we return to this issue, and present an argument that robust power for *recessive* models may not be critical, in which case the  $2 \times 3$  trend test might be best. One practical note is that if one desires to compute p-values out to extremely high levels of significance, the necessary simulations for the  $\min 3p$  and  $\min 4p$  statistics might be quite time-consuming.

### 3.6.2 SCAN WITH COVARIATES

Our results showed that the  $G$  model is most powerful for detecting a genetic effect whether or not there is an environmental effect in most cases. The exception is when there is a very strong interaction, in which case the  $G+E+G \times E$  model is best. Most interestingly, we showed that the  $G+E$  model used most commonly in genetic epidemiology studies is never best, although it is a bit better than the  $G$  alone model if there is strong interaction. This is probably because the genetic effect detected by  $G$  is as strong as that detected by  $G+E$ , if it is assumed that  $G$  and  $E$

are independent. Thus the choice of analysis methods once again requires a personal decision about what underlying genetic models one thinks are most likely. If there is a prior reason to believe that strong interaction is a possibility, then the interaction model makes sense. But otherwise, we would suggest that the  $G$  model is the most robust. That is, logistic regression need not be used at all, and the tests described in the previous section will suffice. We also point out that in our simulations the multiplicative interaction model we fit was at least close to correct. With a continuous covariate it might be much less likely that the chosen interaction model would be correct, which would argue for just using the model with  $G$  only even if interaction is suspected. An additional argument for the simple  $G$  model is that if the environmental factor is mis-modeled (e.g. linear when it should be non-linear), this could result in false positives for an interaction effect. Further studies with continuous covariates and various interaction models would be necessary to explore these issues further, however. Note also that we assumed in all cases that the genetic and environmental factors were independent. If this assumption were to be violated (by virtue of population stratification or true genetic effect), the results might be very different. Of course, covariates should be modeled at a later stage of analysis [Chatterjee, et al. 2006] in order to understand the biology, in order to build predictive models, etc., but our results suggest that at the genome-scanning stage, it is counterproductive in most cases to include them.

### **3.6.3 GENOME-WIDE ANALYSIS**

Our genome scan results are relatively consistent with our single-locus results - they show that for an adequately powered study the compound statistics are best, because they have the most robust power under a variety of genetic models. But once again it is necessary to consider the question of which models one thinks are most likely. One subtle but critical issue is that in the genome-scan context we expect that the true functional loci are unlikely to be a part of our SNP panel. Rather, we expect to detect loci that are in linkage disequilibrium (LD) with the functional loci. Even if the functional locus is purely recessive or purely dominant, the marker locus near it may not be (though an additive functional locus will always give an additive marker locus). In Appendix B.3, we show some example models. The larger the difference in allele frequencies between the trait and marker loci, the more additive the marker locus looks, even if the trait locus

is dominant or recessive. Thus it could be argued that truly recessive and dominant loci are actually quite rare in genome-scanning. If this is true, then the best statistic might be the one with the most power for intermediate models (partly dominant to partly recessive). We investigated this question by considering the performance of different test procedures under the intermediate genetic models listed in Table B2-3. The recessive, additive, and dominant models are in the top, middle, and bottom rows of the table, respectively. Note that the  $2 \times 3$  two df test is only optimal for the bottom two (most recessive) models. If we believe that those models are very unlikely (in a marker, as opposed to in a true functional locus), then all of the robustness properties of that test discussed previously in the paper are irrelevant, and the  $2 \times 3$  trend is probably the best test.

### 3.6.4 OVERALL CONCLUSIONS

We would not dream of telling any individual investigator what to believe about which genetic models are most likely, but we have presented results that can be used by anyone to choose the statistic that is most sensible given whatever his or her beliefs are. We feel that strong arguments can be made for the  $2 \times 3$  trend test, based on our LD discussion immediately above, for the min  $3p$  ( $Z_{MAX}$ ) test if one is also interested in robustness for recessive models, and for the min  $4p$  if one also wants robustness for over-dominant and under-dominant models. With regard to modeling covariates, our results argue strongly against the current practice of controlling for covariates when genome-scanning – we argue that a simple  $G$ -only model is to be preferred under almost any circumstances.

## 4.0 DISCUSSION

### 4.1 MY CONTRIBUTIONS

I summarize my research contributions from three aspects, methodology development in linkage and association analysis and real data analysis on age-related macular degeneration (AMD). Regarding methodology development, in linkage analysis, we have been devoted to making a robust score test through the use of empirical variance estimators for different trait distributions, sampling schemes, and distributions of pedigree types. On the basis of our previous work as described in the paper of Bhattacharjee, et al. [2008], we were able to narrow down the variance estimators which might be robust for general pedigrees and started considering additional ones without using pedigree type information, in order to reduce computational burden. We assumed that from sibships to general pedigrees the effect of parameter misspecification is similar. In the analysis of general pedigrees, the number of correlation parameters dramatically increases with pedigree size and complexity. This motivated us to evaluate the strategy of specifying them by model-based correlations. It turns out that score tests using pedigree type information or not perform similarly and score tests using model-based correlations compared to those using true phenotypic correlations reduce minor power only. For the variance estimators we have considered, their favored distributions of pedigree types are predictable based on our simulation results; overall, the most robust score test is  $\text{SCORE.CT}'_{\text{ALL-TC}}$  which estimates the covariance of estimated IBD sharing using two empirical variance estimates and one theoretical correlation.

In association analysis, we have provided a guideline for choosing an appropriate test procedure or regression model mainly for a genome scan depending on one's prior on the distribution of genetic effects (numbers of truly associated loci of different genetic effects, additive, dominant, recessive, etc.). Although numerous fancy methods for association analysis have been proposed, they have not been applied for GWAS as commonly as simple ones,

probably due to popularity and/or computational complexity. When fitting a model, people seem to be used to incorporating an environmental factor that has a significant effect on the outcome but we have shown that the model with main effect only is sufficient for the detection of genetic effect if the genotype is independent of the environmental factor. In practice, the most popular test procedure used in a GWAS is probably the trend test which is justified by our exploration of the genetic effects of markers in LD with a trait locus and test procedures robust for detecting the association between the trait locus and the disease. Based on our results, the compound statistics are the best choices for those who are concerned about the robustness of a test procedure for different genetic effects. In single-locus analysis, they have correct type I error and nearly optimal power. In genome-wide analysis, they surprisingly are not greatly affected by extreme p-values occurring by chance and may detect a variety of genetic effects.

I have been assisting with the genetic study of AMD for my last-year GSR (Graduate Student Researcher) work under the advisement of Dr. Daniel E. Weeks. AMD is the major cause of irreversible vision loss in the elderly population of United States. About 1.75 million Americans have advanced symptoms of AMD and this number will increase to almost 3 million by 2020 due to the rapid aging of the U.S. population. It has been shown that genetic factors contribute significantly to the development of AMD and the five variants identified are claimed to explain about 50% of heritability [Manolio, et al. 2009]. Among the five variants, the one in the region of ARMS2/HTRA1 genes was first found by our group [Jakobsdottir, et al. 2005]. Our aim, currently, is to assess additional candidate genes that are implicated in AMD risk. Preliminary evaluation showed that the five susceptibility loci highly associated with AMD don't guarantee good classification of disease status [Jakobsdottir, et al. 2009] and novel genes of modest effect may provide new insights into functional analysis and an appropriate prediction model. Based on previous linkage and association results, we first screened candidate genes and then associated the SNPs in the regions of interest with disease status. We assumed that undiscovered genes have a small or moderate effect. To enhance statistical power, we integrate our results from different cohorts and the results of external GWAS.

## 4.2 PROPOSED FUTURE WORK

Our future directions are derived from the project of “What’s the best statistic for a simple test of genetic association in a case-control study?”. In the study, we assumed that the genotype and binary exposure are independent. This might hold for some but not all casual variants of a complex disease. Motivated by two realistic examples of gene-environment interaction (see below), we plan to answer the question of what the best (logistic) regression model is for testing the genetic effect when the genotype is not independent of the exposure.

Example 1: Obesity is probably causal to diabetes. In addition to its causal variants (set Y), diabetes may also be affected by the genes of obesity (set X).

$$\begin{aligned} X &\rightarrow \text{obesity} \rightarrow \text{diabetes} \\ Y &\rightarrow \text{diabetes} \end{aligned}$$

If we are interested in detecting the gene set of Y, what is the best (logistic) regression model?

Example 2: Gestational age (used to measure prematurity) and birth-weight are highly correlated but might be controlled by different sets of genes.

$$\begin{aligned} X &\rightarrow \text{prematurity} \\ Y &\rightarrow \text{birth-weight} \end{aligned}$$

If our goal is to find the genes of birth-weight, what (logistic) regression model is the best choice?

Power comparison of different (logistic) regression models will be done by simulation. Following the relationships as described in the two examples, we will start with the simple case of one gene in X and one gene in Y, followed by multiple genes in each as well as hundreds of thousands of loci not associated with the trait. By fitting a (logistic) regression model, we will relate each individual locus to the trait with adjustment of the exposure or not using unrelated samples. Next, I describe our simulation plan for the simple case that both X and Y are of size 1.

Assume that genes (markers)  $x$  and  $y$  are in HWE. Denote the genotypes of  $x$  and  $y$  by  $g_x$  and  $g_y$  corresponding to the genotypic values,  $\mu_x$  and  $\mu_y$ . In example 1, let  $Y_{ob}$  be the phenotype of obesity and  $Y_{lat}$  be a latent variable of diabetes. Given the  $g_x$ , the  $y_{ob}$  is simulated from a normal mixture with the mean  $\mu_x$  and the variance  $\sigma^2$

$$y_{ob} \sim N(\mu_x, \sigma^2),$$

and  $y_{lat}$  is simulated from a normal distribution with the mean defined by a function of  $g_y$  and  $y_{ob}$  and the variance  $\sigma^2$ ,

$$y_{lat} \sim N(f(g_y, y_{ob}), \sigma^2).$$

An individual is diagnosed with diabetes if  $y_{lat}$  is equal or greater than a threshold, and otherwise is treated as a control. Let  $Y$  be affection status and  $\alpha$  be the threshold

$$y = \begin{cases} 1, & \text{if } y_{lat} \geq \alpha \\ 0, & \text{if } y_{lat} < \alpha \end{cases}$$

In example 2, let  $Y_{g-age}$  be the gestational age and  $Y_{bw}$  be the birth-weight.  $(y_{g-age}, y_{bw})$  are simulated from a bivariate normal with the mean vector  $(\mu_x, \mu_y)$  and the covariance matrix

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

where  $\sigma_{11}$  and  $\sigma_{22}$  are the variances of  $Y_{g-age}$  and  $Y_{bw}$  and  $\sigma_{12} = \sigma_{21}$  is the covariance between  $Y_{g-age}$  and  $Y_{bw}$ .

### 4.3 OPEN PROBLEMS IN GENE MAPPING

I will now more broadly discuss open problems still challenging the field of gene mapping. Many of them are related to GWAS, since the design has been a powerful tool for the investigation of complex diseases. Hundreds of genes have been identified by GWAS but for most of complex diseases, they can only account for a small portion of heritability. A vast amount of information produced by GWAS might not have been completely used and other sources or strategies of gene mapping may complement genome-wide association approaches to discover genes responsible for complex traits.

#### 4.3.1 Population stratification

GWAS are typically case-control designs. Case-control association tests are in principle more powerful than family-based association tests but might lead to spurious association due to population stratification. Population genetic structure usually cannot be fully specified by ancestry information because of vague definitions of ancestry groups and/or self-report biases. Thomas [2004] summarized principles of the methods proposed to address this issue: With a panel of markers unlinked to associated loci, 1) use an overdispersion model to determine a test statistic's appropriate empirical distribution [Devlin and Roeder 1999], 2) evaluate whether stratification exists [Pritchard and Rosenberg 1999], or 3) use a latent-class model to distinguish homogeneous subpopulations [Pritchard, et al. 2000a; Pritchard, et al. 2000b]. 1) adjusts a single parameter of the test statistic distribution using the median of null marker statistics and makes a strong assumption that adjustment is the same throughout the genome. 2) is the exploration prior to 3). 3) uses an association test that takes population stratification into account but requires specifying the number of subpopulations and may not handle admixture well. Currently, the most popular methods to adjust for population stratification are those based on principal components [e.g. Price, et al. 2006]. Adjustment by unnecessary principal components however would lead to significant power loss [Yu, et al. 2008]. To best select principal components for the correction of population stratification, alternatives to a fixed number of top-ranked principal components include the approach according to the Tracy-Wisdom test [Patterson, et al. 2006] and the test

procedure [Li, et al. 2009a] based on the distance-based regression model originally proposed by McArdle and Anderson [2001].

### 4.3.2 Multiple testing

Millions of SNPs initially are screened by univariate association tests and their resulting statistics must be adjusted for multiple testing to control the error rate like family-wise error rate (FWER) or false discovery rate (FDR) (FWER: probability of *any* false positive; FDR: expected proportion of false positives among all the rejections). The number of tests can exceed the number of SNPs if multiple phenotypes are involved or epistasis is of interest or each SNP is analyzed using multiple analytical methods or with consideration of more than one inheritance model.

The Bonferroni correction, most typical to adjust for FWER, is generally too conservative especially when the number of tests is large. Methods that control FDR first introduced by Benjamini and Hochberg [1995] are gaining popularity for researchers to identify and then follow up a set of candidate genes. Parametric FDR methods assume that p-values under the null hypothesis are distributed as uniform  $[0, 1]$  and require alternative distribution of test statistic be specified. Alternatively, one can derive empirical p-values by permutation testing but computational time is a challenge and it may be tricky to determine what to maintain and what to permute.

### 4.3.3 Meta analysis

Thousands of samples must be collected to ensure adequate power. Ideally, people share data to facilitate integration of existing and future data sets. Fixed and random effect models are commonly used to combine similar results across studies. If one concludes that the genetic effects of different studies are homogeneous and their estimates vary simply due to sampling errors, it is appropriate to pool information together using a fixed effect model. On the other hand, if one judges that the variation in genetic effect estimates is attributable to within and between-study variation, a random effect model can be used to account for the two sources of variation. Missing or untyped genotypes can be imputed based on the LD information using the

HapMap as a reference dataset. Bear in mind that between-study heterogeneity must be carefully investigated to select an appropriate method for meta-analysis and to detect or correct systematic errors and biases.

#### **4.3.4 Multilocus analysis**

For the joint detection of multiple loci by association analysis, it has been debated whether haplotype analysis should be preferred over unphased multilocus analysis. Clark [2004] outlined three primary reasons to consider haplotype analysis, 1) sequences of functional genes (protein coding genes) are in haplotypes; 2) genetic variations in population are structured into haplotypes and are likely to be transmitted as a unit; 3) haplotypes serve to reduce dimensionality and so may increase power. Haplotypes however require resolution of gametic phase and this usually must be inferred statistically. Clayton, et al. [2004] further argued that haplotype analysis is generally less powerful than unphased multilocus analysis for detecting “indirect” associations of haplotypes with functional loci and so with the trait. In reality, joint association analysis of multiple loci has difficulty being applied for a GWAS not only for high computational demand but also for stringent multiple testing adjustment. An exhaustive search for pairwise gene-gene interaction recently becomes feasible. With the increase of fitted models, statistics suffer a more severe penalty for multiple testing than that on the statistics of single-locus analysis. Even truly associated loci with modest effect after adjustment might be filtered out.

#### **4.3.5 Modeling strategy**

For the concerns about computational intensity and over-penalty of multiple testing, multi-stage analysis and simple models for individual SNPs followed by fancy ones for a small subset are strategies for GWAS. For example, a two-stage testing procedure partitions the information into two orthogonal components and reduces the penalty for multiple testing by adjusting a subset of markers in the second stage. Simple models are not expected to reflect real biological mechanism. In preliminary analysis, we don't intend to build up a model for prediction but a model for testing. Once susceptibility loci are identified, we can proceed with more complicated modeling

to understand how genes work. This supports the strategy of simple models for individual SNPs followed by fancy ones for a small subset.

Another strategy I would like to point out about modeling is measuring association using a model computationally tractable and technically flexible. More specifically, it doesn't matter to use either genotype or phenotype as the response variable. To avoid complex adjustment for the sampling scheme, it makes intuitive sense to condition on phenotypes and let the genotype be random. This may simplify the effort of model fitting and also provide flexibility to account for missing genotypes and phenotypes. Successful examples of this spirit include the reverse regression method [Sham, et al. 2002] for linkage analysis and the MQLS [Thornton and McPeck 2007] for association analysis. A flexible method according to the development of popular methods such as the FBAT [Laird and Lange 2008] and MQLS [Thornton and McPeck 2010; Zheng and McPeck 2007] should be adjustable for population stratification, multiple loci, environmental covariates, general pedigrees etc.

#### **4.3.6 Rare variant analysis**

A variety of methods have been proposed for the selection of tag SNPs but most of them require the LD information provided by HapMap. Ideally, tag SNPs are in strong LD with causal variants; association of causal variants with the trait can be indirectly measured from tag SNPs. The HapMap project however emphasized common SNPs and their association with rare causal variants is weaker than rare SNPs. Rare causal variants thus are unlikely to be detected by GWAS. In post-GWAS era, people turn to pursue the hypothesis of common disease-rare variant.

The latest DNA sequencing technologies have enabled association analysis of rare variants. Due to the high cost of sequencing, a common strategy is to re-sequence regions of candidate genes. With the launch of 1,000 genome project, at least 1,000 genomes of 10 ethnic backgrounds will be sequenced and one of the project's goals is to identify all the variants of the samples. I will assume that the goal can be reached, although the sample size about 100 for each ethnic group is probably too small and rare variants can be identified by chance [Li and Leal 2009]. What we can gain from the map includes designing appropriate chips for the study of rare variants and using the LD information for selecting tag SNPs, imputing missing genotypes, and so on.

Rare variant detection requires a huge number of population samples which might be practically infeasible. Ascertained families are more informative than unrelated samples for the aggregation of rare variants. In addition to case-control data collected for GWAS, it is worth extra effort to enroll families of the cases. To optimize statistical power, this motivates development of the methods that can make use of family and case-control data and if needed, adjust for ascertainment bias.

Single-locus or multilocus association approaches for common variants are probably straightforward to extend for rare variants but might lose a substantial amount of power. Given a region of chromosome (single SNP is a special case), a mixture of rare and other variants associated with the outcome of interest essentially is a realization of sparse data tested for association. It is known that the main statistical problem of sparse data is inaccurate approximation of null distribution of the test statistic. Traditionally, we use an exact or permuted p-value instead of a theoretical one or collapse data across less frequent observations or turn to multivariate analysis, e.g. haplotype analysis. Similar ideas should work for association testing on rare variants and the methods recently proposed are the best examples: Combine Multivariate and Collapsing (CMC) method by Li and Leal [2008], the method based on accumulations of rare variants by Morris and Zeggini [2010], and the haplotype-based method by Zhu et. al [2010]. I thus would suggest exploring association of rare variants with a disease on the basis of literature for handling sparse data.

#### **4.3.7 Summary**

I have been focusing on statistical issues around GWAS. I assessed the GWAS starting from its typical study design followed by its unprecedentedly large-scale and then its connection to the hypotheses of common disease-common variant and then common disease-rare variant. Beyond the polymorphism of DNA, one can do gene mapping by studying structure variation of DNA (Copy Number Variation (CNV) analysis) or other molecular levels like RNA (microarray analysis). Complexity of disease etiology due to locus heterogeneity, phenocopy, gene-gene or gene-environment interaction must be accounted for by statistical modeling. Some people argue that linear models are insufficient to map genotype to phenotype and that the computational

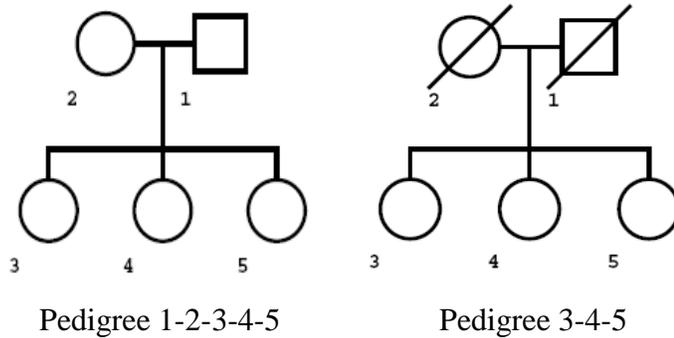
methods of data mining and machine learning let the data tell what the model is instead of fitting a “good” model and thus are more appropriate to address this complexity [Moore, et al. 2010].

Statistical genetics has co-evolved with genotyping technologies. Current genotyping technologies generate a massive amount of data but the use of such data requires better tools and strategies or we may face “an embarrassment of riches” [Stein and Elston 2009].

**APPENDIX A SUPPLEMENTAL MATERIALS FOR CHAPTER 2**

**A.1 MEAN, VARIANCE, AND COVARIANCE OF ESTIMATED IBD SHARING OF TWO SIMILAR PEDIGREE TYPES**

Consider a sibship of size 3 with parents and a similar pedigree with parental genotypes missing. If the effect of ignoring pedigree types on the power of a score test is minor, we expect that the mean, variance, and covariance of estimated IBD sharing for the sibpairs 3-4, 3-5, and 4-5 will be similar in the two pedigree types.



For a SNP with a minor allele frequency of  $r$  ( $s = 1 - r$ ), let  $g_3$ ,  $g_4$ , and  $g_5$  be the genotypes for the siblings 3, 4, and 5. The joint distribution of estimated IBD sharing for the three sibpairs given the genotypes of both parents and the three siblings is given in Table A1-1. Based on Table A1-1, the marginal distribution of estimated IBD sharing for any single sibpair saying 3-4 is given in Table A1-2, and it is straightforward to show that the expectation and variance of estimated IBD sharing are 0.5 and  $0.25r(1 - r)(r^2 - r + 1)$ .

**Table A1-1. Joint distribution of estimated IBD sharing for the three sibpairs given the genotypes of both parents and the three siblings**

Mating	$g_3$	$g_4$	$g_5$	freq.	$\hat{\pi}_{34}$	$\hat{\pi}_{45}$	$\hat{\pi}_{35}$
11×11	11	11	11	$r^4$	0.5	0.5	0.5
11×12	11	11	11	$0.5 r^3 s$	0.75	0.75	0.75
	11	11	12	$0.5 r^3 s$	0.75	0.25	0.25
	11	12	11	$0.5 r^3 s$	0.25	0.25	0.75
	11	12	12	$0.5 r^3 s$	0.25	0.75	0.25
	12	11	11	$0.5 r^3 s$	0.25	0.75	0.25
	12	11	12	$0.5 r^3 s$	0.25	0.25	0.75
	12	12	11	$0.5 r^3 s$	0.75	0.25	0.25
	12	12	12	$0.5 r^3 s$	0.75	0.75	0.75
11×22	12	12	12	$2 r^2 s^2$	0.5	0.5	0.5
22×22	22	22	22	$s^4$	0.5	0.5	0.5
22×12	12	12	12	$0.5 r s^3$	0.75	0.75	0.75
	12	12	22	$0.5 r s^3$	0.75	0.25	0.25
	12	22	12	$0.5 r s^3$	0.25	0.25	0.75
	12	22	22	$0.5 r s^3$	0.25	0.75	0.25
	22	12	12	$0.5 r s^3$	0.25	0.75	0.25
	22	12	22	$0.5 r s^3$	0.25	0.25	0.75
	22	22	12	$0.5 r s^3$	0.75	0.25	0.25
	22	22	22	$0.5 r s^3$	0.75	0.75	0.75
12×12	11	11	11	$0.0625 r^2 s^2$	1	1	1
	11	11	12	$0.125 r^2 s^2$	1	0.5	0.5
	11	11	22	$0.0625 r^2 s^2$	1	0	0
	11	22	11	$0.0625 r^2 s^2$	0	0	1
	11	22	12	$0.125 r^2 s^2$	0	0.5	0.5
	11	22	22	$0.0625 r^2 s^2$	0	1	0
	11	12	11	$0.125 r^2 s^2$	0.5	0.5	1
	11	12	12	$0.25 r^2 s^2$	0.5	0.5	0.5
	11	12	22	$0.125 r^2 s^2$	0.5	0.5	0
	22	11	11	$0.0625 r^2 s^2$	0	1	0
	22	11	12	$0.125 r^2 s^2$	0	0.5	0.5
	22	11	22	$0.0625 r^2 s^2$	0	0	1
	22	22	11	$0.0625 r^2 s^2$	1	0	0
	22	22	12	$0.125 r^2 s^2$	1	0.5	0.5
	22	22	22	$0.0625 r^2 s^2$	1	1	1
	22	12	11	$0.125 r^2 s^2$	0.5	0.5	0
	22	12	12	$0.25 r^2 s^2$	0.5	0.5	0.5
	22	12	22	$0.125 r^2 s^2$	0.5	0.5	1
	12	11	11	$0.125 r^2 s^2$	0.5	1	0.5
	12	11	12	$0.25 r^2 s^2$	0.5	0.5	0.5
	12	11	22	$0.125 r^2 s^2$	0.5	0	0.5
	12	22	11	$0.125 r^2 s^2$	0.5	0	0.5
	12	22	12	$0.25 r^2 s^2$	0.5	0.5	0.5
	12	22	22	$0.125 r^2 s^2$	0.5	1	0.5
12	12	11	$0.25 r^2 s^2$	0.5	0.5	0.5	
12	12	12	$0.5 r^2 s^2$	0.5	0.5	0.5	
12	12	22	$0.25 r^2 s^2$	0.5	0.5	0.5	

**Table A1-2. Marginal distribution of estimated IBD sharing for a single sibpair, e.g. 3-4**

$\hat{\pi}_{34}$	0	0.25	0.5	0.75	1
Frequency	$0.5r^2s^2$	$2rs(r^2 + s^2)$	$r^4 + 5r^2s^2$	$2rs(r^2 + s^2)$	$0.5r^2s^2$

$$E(\hat{\pi}_{34}) = 0.25 \times 2rs(r^2 + s^2) + 0.5 \times (r^4 + 5r^2s^2 + s^4) + 0.75 \times 2rs(r^2 + s^2) + 0.5r^2s^2 = 0.5$$

$$Var(\hat{\pi}_{34}) = 0.25^2 \times 2rs(r^2 + s^2) + 0.5^2 \times (r^4 + 5r^2s^2 + s^4) + 0.75^2 \times 2rs(r^2 + s^2) + 0.5r^2s^2$$

$$- 0.5^2 = 0.25r(1-r)(r^2 - r + 1)$$

Similarly, the joint distribution of estimated IBD sharing for any two sibpairs saying 3-4 and 4-5 is given in Table A1-3. By the formula of covariance,  $Cov(X, Y) = E(XY) - E(X)E(Y)$ , we have showed that the covariance of  $\hat{\pi}_{34}$  and  $\hat{\pi}_{45}$  is zero, and it is obvious from Table A1-3 that  $P(\hat{\pi}_{34}, \hat{\pi}_{45}) \neq P(\hat{\pi}_{34})P(\hat{\pi}_{45})$ .  $\hat{\pi}_{34}$  and  $\hat{\pi}_{45}$  thus are uncorrelated but not independent.

**Table A1-3. Joint distribution of estimated IBD sharing for two sibpairs, e.g. 3-4 and 4-5**

$\hat{\pi}_{34} \setminus \hat{\pi}_{45}$	0	0.25	0.5	0.75	1	Total
0	$\frac{1}{8}r^2s^2$	0	$0.25r^2s^2$	0	$\frac{1}{8}r^2s^2$	$\frac{1}{2}r^2s^2$
0.25	0	$r^3s + rs^3$	0	$r^3s + rs^3$	0	$2rs(r^2 + s^2)$
0.5	$\frac{1}{4}r^2s^2$	0	$r^4 + \frac{9}{2}r^2s^2 + s^4$	0	$\frac{1}{4}r^2s^2$	$r^4 + 5r^2s^2 + s^4$
0.75	0	$r^3s + rs^3$	0	$r^3s + rs^3$	0	$2rs(r^2 + s^2)$
1	$\frac{1}{8}r^2s^2$	0	$0.25r^2s^2$	0	$\frac{1}{8}r^2s^2$	$\frac{1}{2}r^2s^2$
Total	$\frac{1}{2}r^2s^2$	$2rs(r^2 + s^2)$	$r^4 + 5r^2s^2 + s^4$	$2rs(r^2 + s^2)$	$\frac{1}{2}r^2s^2$	1

$$E(\hat{\pi}_{34}\hat{\pi}_{45}) = \frac{1}{16}(r^3s + rs^3) + \frac{3}{16}(r^3s + rs^3) + \frac{1}{4}\left(r^4 + s^4 + \frac{9}{2}r^2s^2\right) + \frac{1}{8}(r^2s^2)$$

$$+ \frac{3}{16}(r^3s + rs^3) + \frac{9}{16}(r^3s + rs^3) + \frac{1}{8}(r^2s^2) + \frac{1}{8}(r^2s^2) = \frac{1}{4}$$

$$\Rightarrow Cov(\hat{\pi}_{34}\hat{\pi}_{45}) = E(\hat{\pi}_{34}\hat{\pi}_{45}) - E(\hat{\pi}_{34})E(\hat{\pi}_{45}) = \frac{1}{4} - \frac{1}{2} \times \frac{1}{2} = 0$$

For the pedigrees in which parental genotypes are unavailable, we can calculate mean and variance of estimated IBD sharing based on weighted averages from Table A1-1. For example, if

$g_3, g_4,$  and  $g_5$  are all 11, according to Table A1-1, this configuration can be from the mating type 11×11 or 11×12 or 12×12 with probability  $r^4, 0.5r^3s,$  and  $r^2s^2/16,$  respectively. The weighted estimated IBD sharing assigned to 3-4, 4-5, and 3-5 sibpairs is then given by

$$\frac{1}{w_1 + w_2 + w_3} [w_1(0.5,0.5,0.5) + w_2(0.75,0.75,0.75) + w_3(1,1,1)]$$

where  $w_1 = r^4, w_2 = 0.5r^3s, w_3 = \frac{1}{16} r^2s^2.$  The joint distribution of estimated IBD sharing for the three siblings given the genotypes of the three siblings only is given in Table A1-4. The mean of estimated IBD sharing for each sibpair is 0.5. The variances and covariances for different sibpairs are equivalent but mathematically not in a simple form (not shown).

In Table A1-5, we compare the difference in variance and covariance between the pedigrees 1-2-3-4-5 and 3-4-5 by considering a number of minor allele frequencies. The absolute difference in variance or covariance increases with the minor allele frequency. The variance for the sibpairs in 3-4-5 is smaller than that for the sibpairs in 1-2-3-4-5. The covariance between the sibpairs in 3-4-5 tends to be negative but the covariance between the sibpairs in 1-2-3-4-5 is proved to be zero. All the differences look quite small and we expect that they are approaching zero with the consideration of multipoint IBD sharing or a more polymorphic marker.

**Table A1-4. Joint distribution of estimated IBD sharing for the three siblings given the genotypes of the three siblings**

$g_3$	$g_4$	$g_5$	frequency	$(\hat{\pi}_{34}, \hat{\pi}_{45}, \hat{\pi}_{35})$
11	11	11	$r^4 + 0.5 r^3 s + 0.0625 r^2 s^2$	$(r^2 s^2 + 8 r^3 s + r^4)^{-1} (r^2 s^2 + 6 r^3 s + 4 r^4) (1, 1, 1)$
11	11	12	$0.5 r^3 s + 0.125 r^2 s^2$	$(2 r^2 s^2 + 8 r^3 s)^{-1} (2 r^2 s^2 + 6 r^3 s, r^2 s^2 + 2 r^3 s, r^2 s^2 + 2 r^3 s)$
11	11	22	$0.0625 r^2 s^2$	$(1, 0, 0)$
11	12	11	$0.5 r^3 s + 0.125 r^2 s^2$	$(2 r^2 s^2 + 8 r^3 s)^{-1} (r^2 s^2 + 2 r^3 s, r^2 s^2 + 2 r^3 s, 2 r^2 s^2 + 6 r^3 s)$
11	12	12	$0.5 r^3 s + 0.25 r^2 s^2$	$(2 r^2 s^2 + 4 r^3 s)^{-1} (r^2 s^2 + r^3 s, r^2 s^2 + 3 r^3 s, r^2 s^2 + r^3 s)$
11	12	22	$0.125 r^2 s^2$	$(0.5, 0.5, 0)$
11	22	11	$0.0625 r^2 s^2$	$(0, 0, 1)$
11	22	12	$0.125 r^2 s^2$	$(0, 0.5, 0.5)$
11	22	22	$0.0625 r^2 s^2$	$(0, 1, 0)$
12	11	11	$0.5 r^3 s + 0.125 r^2 s^2$	$(2 r^2 s^2 + 8 r^3 s)^{-1} (r^2 s^2 + 2 r^3 s, 2 r^2 s^2 + 6 r^3 s, r^2 s^2 + 2 r^3 s)$
12	11	12	$0.5 r^3 s + 0.25 r^2 s^2$	$(2 r^2 s^2 + 4 r^3 s)^{-1} (r^2 s^2 + r^3 s, r^2 s^2 + r^3 s, r^2 s^2 + 3 r^3 s)$
12	11	22	$0.125 r^2 s^2$	$(0.5, 0, 0.5)$
12	12	11	$0.5 r^3 s + 0.25 r^2 s^2$	$(2 r^2 s^2 + 4 r^3 s)^{-1} (r^2 s^2 + 3 r^3 s, r^2 s^2 + r^3 s, r^2 s^2 + r^3 s)$
12	12	12	$0.5 r^3 s + 2 r^2 s^2 + 0.5 r s^3$	$(4 r s^3 + 20 r^2 s^2 + 4 r^3 s)^{-1} (3 r s^3 + 10 r^2 s^2 + 3 r^3 s) (1, 1, 1)$
12	12	22	$0.5 r s^3 + 0.25 r^2 s^2$	$(2 r^2 s^2 + 4 r s^3)^{-1} (r^2 s^2 + 3 r s^3, r^2 s^2 + r s^3, r^2 s^2 + r s^3)$
12	22	11	$0.125 r^2 s^2$	$(0.5, 0, 0.5)$
12	22	12	$0.5 r s^3 + 0.25 r^2 s^2$	$(2 r^2 s^2 + 4 r s^3)^{-1} (r^2 s^2 + r s^3, r^2 s^2 + r s^3, r^2 s^2 + 3 r s^3)$
12	22	22	$0.5 r s^3 + 0.0125 r^2 s^2$	$(2 r^2 s^2 + 8 r s^3)^{-1} (r^2 s^2 + 2 r s^3, 2 r^2 s^2 + 6 r s^3, r^2 s^2 + 2 r s^3)$
22	11	11	$0.0625 r^2 s^2$	$(0, 1, 0)$
22	11	12	$0.125 r^2 s^2$	$(0, 0.5, 0.5)$
22	11	22	$0.0625 r^2 s^2$	$(0, 0, 1)$
22	12	11	$0.125 r^2 s^2$	$(0.5, 0.5, 0)$
22	12	12	$0.5 r s^3 + 0.25 r^2 s^2$	$(2 r^2 s^2 + 4 r s^3)^{-1} (r^2 s^2 + r s^3, r^2 s^2 + 3 r s^3, r^2 s^2 + r s^3)$
22	12	22	$0.5 r s^3 + 0.125 r^2 s^2$	$(2 r^2 s^2 + 8 r s^3)^{-1} (r^2 s^2 + 2 r s^3, r^2 s^2 + 2 r s^3, 2 r^2 s^2 + 6 r s^3)$
22	22	11	$0.0625 r^2 s^2$	$(1, 0, 0)$
22	22	12	$0.5 r s^3 + 0.125 r^2 s^2$	$(2 r^2 s^2 + 8 r s^3)^{-1} (2 r^2 s^2 + 6 r s^3, r^2 s^2 + 2 r s^3, r^2 s^2 + 2 r s^3)$
22	22	22	$s^4 + 0.5 r s^3 + 0.125 r^2 s^2$	$(s^4 + 8 r s^3 + r^2 s^2)^{-1} (8 s^4 + 6 r s^3 + r^2 s^2) (1, 1, 1)$

**Table A1-5. Variances and covariances of estimated IBD sharing for the sibpairs in the pedigrees 1-2-3-4-5 and 3-4-5**

MAF <sup>a</sup>	Variance			Covariance		
	1-2-3-4-5	3-4-5	DIFF <sup>b</sup>	1-2-3-4-5	3-4-5	DIFF <sup>c</sup>
0.01	0.0025	0.0021	0.0004	0	-0.0003	0.0003
0.05	0.0113	0.0095	0.0018	0	-0.0016	0.0016
0.1	0.0205	0.0167	0.0038	0	-0.0032	0.0032
0.2	0.0336	0.0260	0.0076	0	-0.0056	0.0056
0.3	0.0415	0.0310	0.0105	0	-0.0070	0.0070
0.4	0.0456	0.0333	0.0123	0	-0.0076	0.0076
0.5	0.0469	0.0339	0.0130	0	-0.0077	0.0077

a: Minor Allele Frequency; b: column 2-column3; c: column 5-column6.

## A.2 MATHEMATICAL INSIGHT ON DROPPING PARENTAL PHENOTYPES FROM SIBSHIP DATA IN THE CALCULATION OF SCORE TEST STATISTIC

*The score test statistic for a sibship without parental phenotypes is equivalent to that for the sibship including the parental phenotypes if it is assumed that the phenotypic correlations for the parent-child and unrelated pairs are 0.*

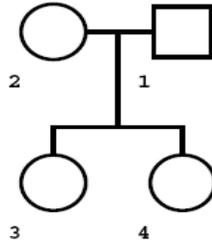
**Proof:** Let the score statistic and its variance contributed by the  $i$ th pedigree be  $S_i$  and  $Var(S_i)$  where  $v_{ij}$  is the transformed phenotype,  $\hat{\pi}_{ij}$  is the estimated IBD sharing, and  $\phi_{ij}$  is the kinship coefficient for the  $j$ th pair in the  $i$ th pedigree.

$$S_i = \sum_{j=1}^{n_i} v_{ij} (\hat{\pi}_{ij} - 2\phi_{ij})$$

$$Var(S_i) = \sum_{j=1}^{n_i} v_{ij}^2 Var(\hat{\pi}_{ij}) + \sum_{p \neq q} v_{ip} v_{iq} Cov(\hat{\pi}_{ip}, \hat{\pi}_{iq})$$

$$= \sum_{j=1}^{n_i} v_{ij}^2 Var(\hat{\pi}_{ij}), \text{ if sibship.}$$

Possible relative-pair types generated by a sibship include unrelated, parent-child, and sibling pairs. The estimated IBD sharing for unrelated and parent-child pairs are fixed at their expected values so the effective terms (not zeroed out) in  $S_i$  and  $Var(S_i)$  are only those for sibling pairs. Without loss of generality, we consider one sibship of size two as follows and denote by pedigree 1-2-3-4 the pedigree with non-missing genotypes and phenotypes and denote by pedigree 3-4 the pedigree with missing parental phenotypes.



For simplicity, we now define  $\hat{\pi}_{34}$  to be the estimated IBD sharing for the sibpair 3-4 and  $E(\hat{\pi}_{34}) = 2\phi_{34}$ . The score statistic and variance of the score statistic for the pedigrees 1-2-3-4 and 3-4 are represented by

$$S = v_{34}(\hat{\pi}_{34} - 2\phi_{34}),$$

$$Var(S) = v_{34}^2 Var(\hat{\pi}_{34}).$$

The two pedigrees are expected to have the same  $\hat{\pi}_{34}$  and  $Var(\hat{\pi}_{34})$ . To prove that the pedigree 3-4 has the same score test statistic ( $S$  and  $Var(S)$ ) as the pedigree 1-2-3-4 assuming that the phenotypic correlations for parent-child and unrelated pairs are zero is simply to show that the sibpair 3-4 in each pedigree has the same transformed phenotype  $v_{34}$ .

For the pedigree of 1-2-3-4, let  $Y_1 = (y_1 \ y_2 \ y_3 \ y_4)'$  be the mean-centered trait vector and  $\Sigma_1$  be the covariance matrix

$$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & c \\ 0 & 0 & c & 1 \end{bmatrix}.$$

where  $\sigma^2$  is the total trait variance and the phenotypic correlations for parent-child and unrelated pairs are assumed to be zero. For the pedigree of 3-4, let  $Y_2 = (y_3 \ y_4)'$  be the mean-centered trait vector and  $\Sigma_2$  be the covariance matrix

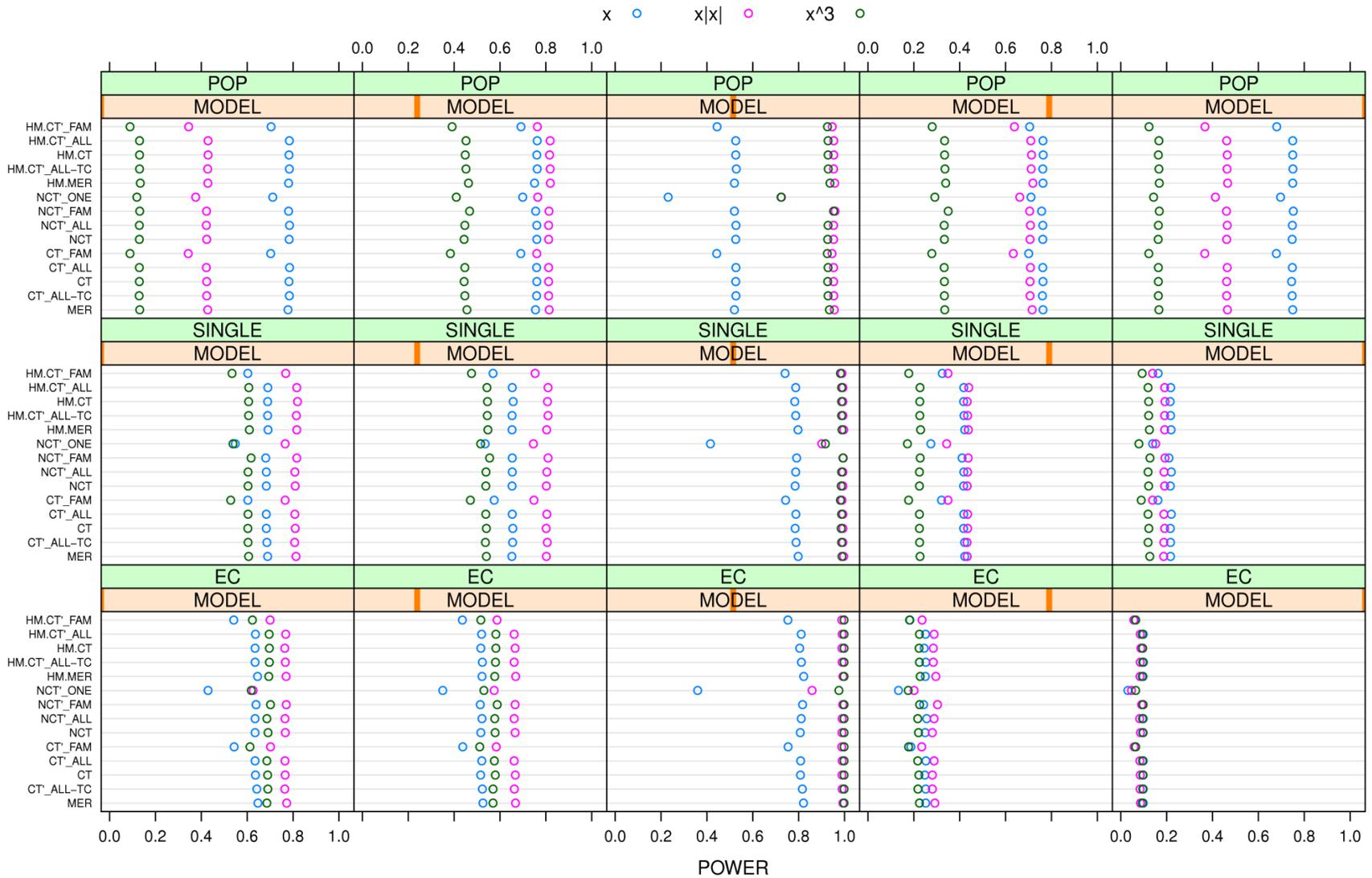
$$\sigma^2 \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix}.$$

Through basic matrix operations, it can be proved that the transformed phenotype  $v_{34}$  derived by  $vec[\Sigma_1^{-1}Y_1Y_1'\Sigma_1^{-1} - \Sigma_1^{-1}]$  or  $vec[\Sigma_2^{-1}Y_2Y_2'\Sigma_2^{-1} - \Sigma_2^{-1}]$  is equivalent to

$$\frac{y_3y_4 + c^2y_3y_4 - c^3\sigma^2 - c(y_3^2 + y_4^2 - \sigma^2)}{(-1 + c^2)\sigma^4}.$$

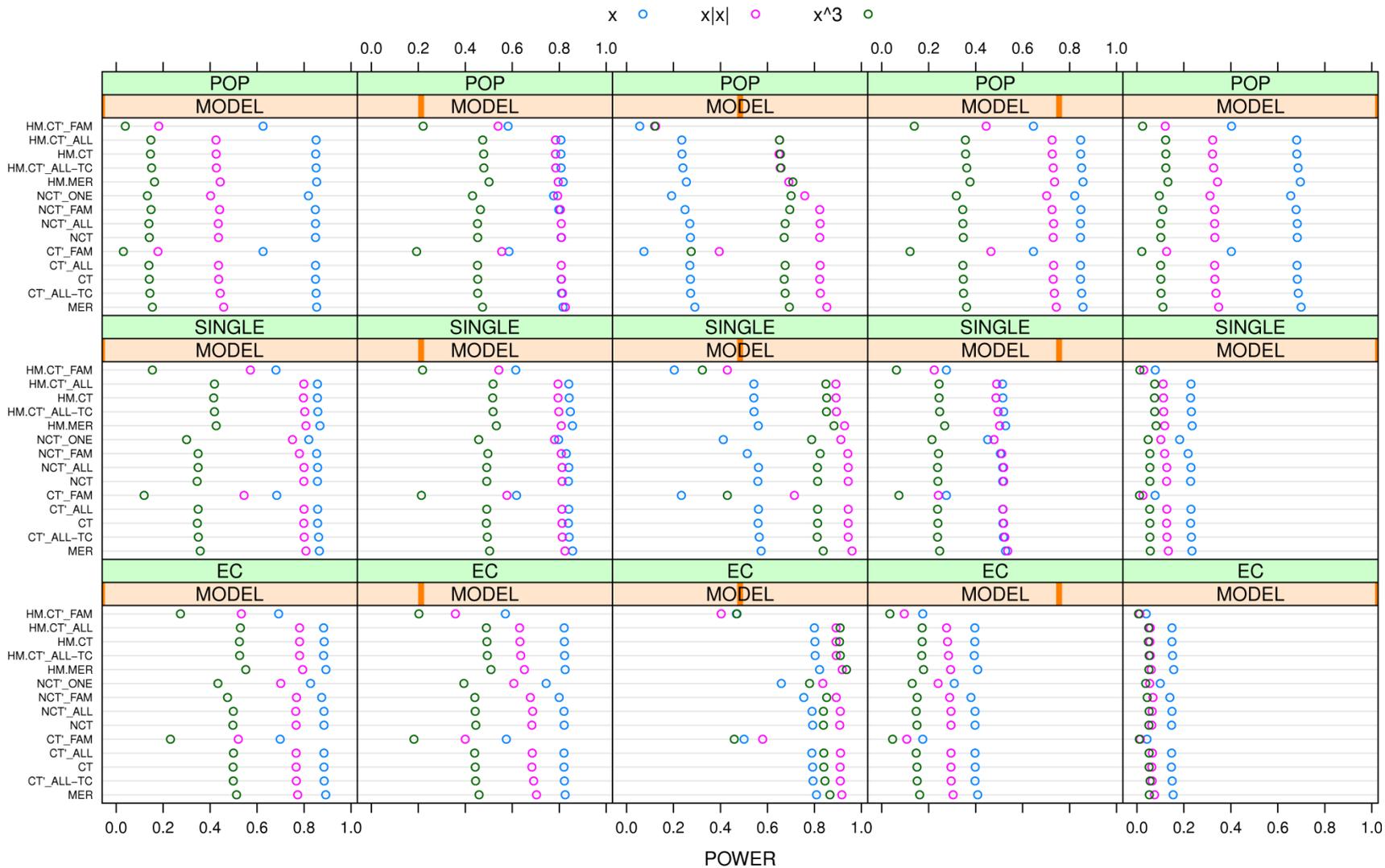
Thus, the transformed phenotype for the sibpair 3-4 in the pedigrees 1-2-3-4 and 3-4 is the same and so one sibship without parental phenotypes has the same score test statistic as the sibship with parental data (genotypes and phenotypes), assuming that both phenotypic correlations for parent-child and unrelated pairs are zero.

### A.3 POWER SIMULATION RESULTS



**Figure A3-1. 4SIBS power simulation results**

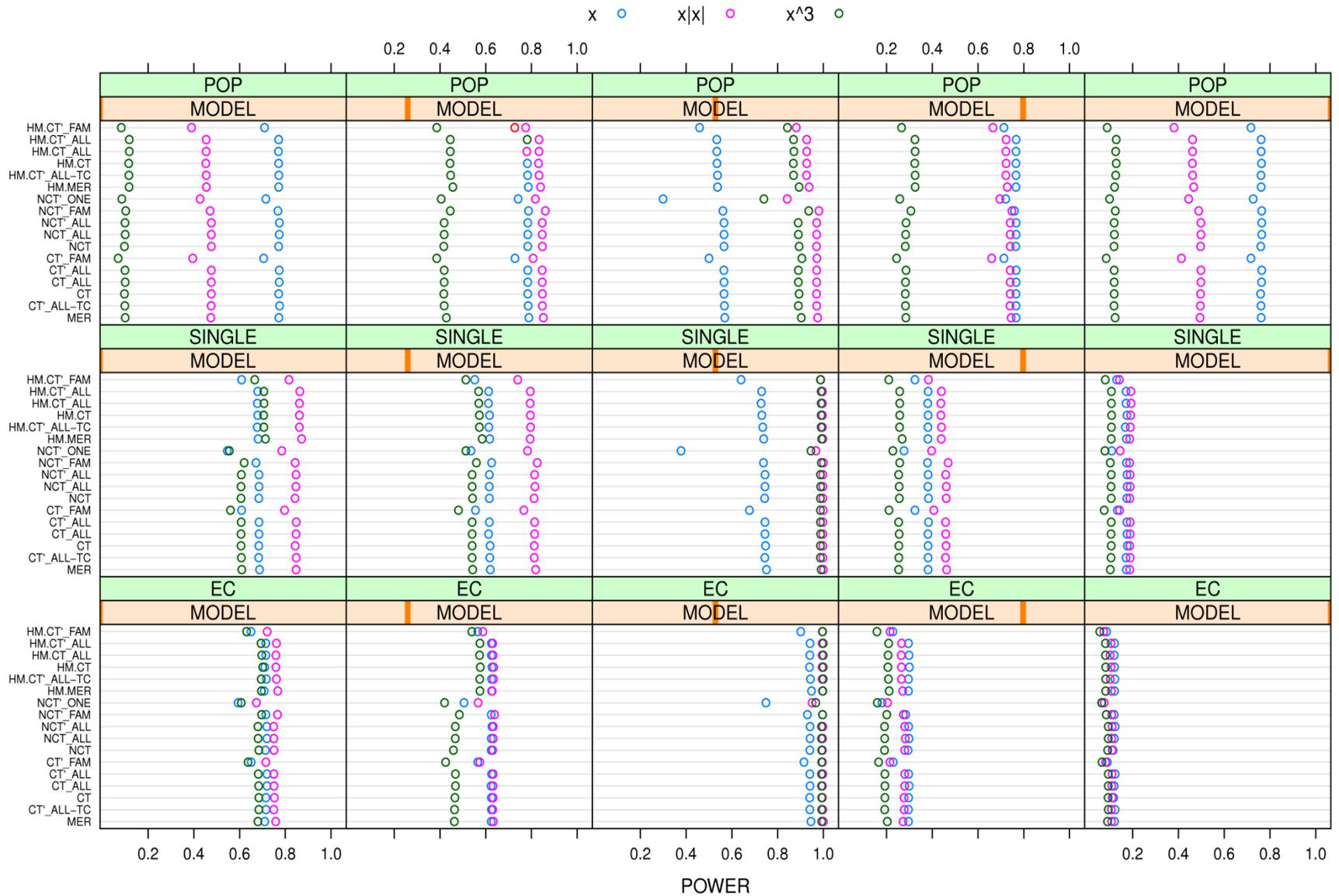
Note: CT\*\_ALL-TC is SCORE.CT\*\_ALL-TC, NCT\*\_ALL is SCORE.NULL.CT\*\_ALL and so on. Power for SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL were not presented because they are exactly equivalent to SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL for single pedigree types.



**Figure A3-2. HS power simulation results**

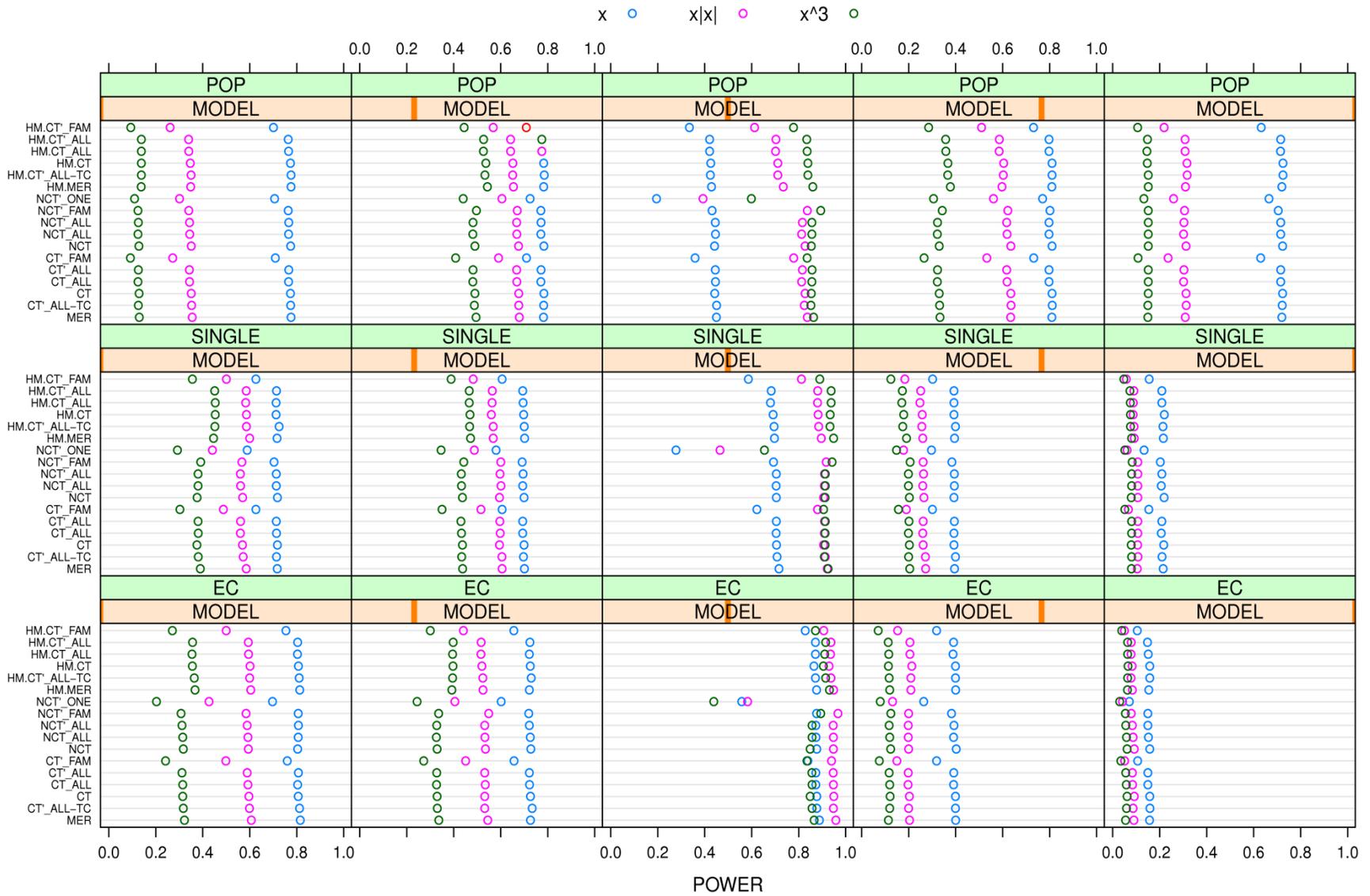
Note: CT\*\_ALL-TC is SCORE.CT\*\_ALL-TC, NCT\*\_ALL is SCORE.NULL.CT\*\_ALL and so on. Power for SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL were not presented because they are exactly equivalent to SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL for single pedigree types.





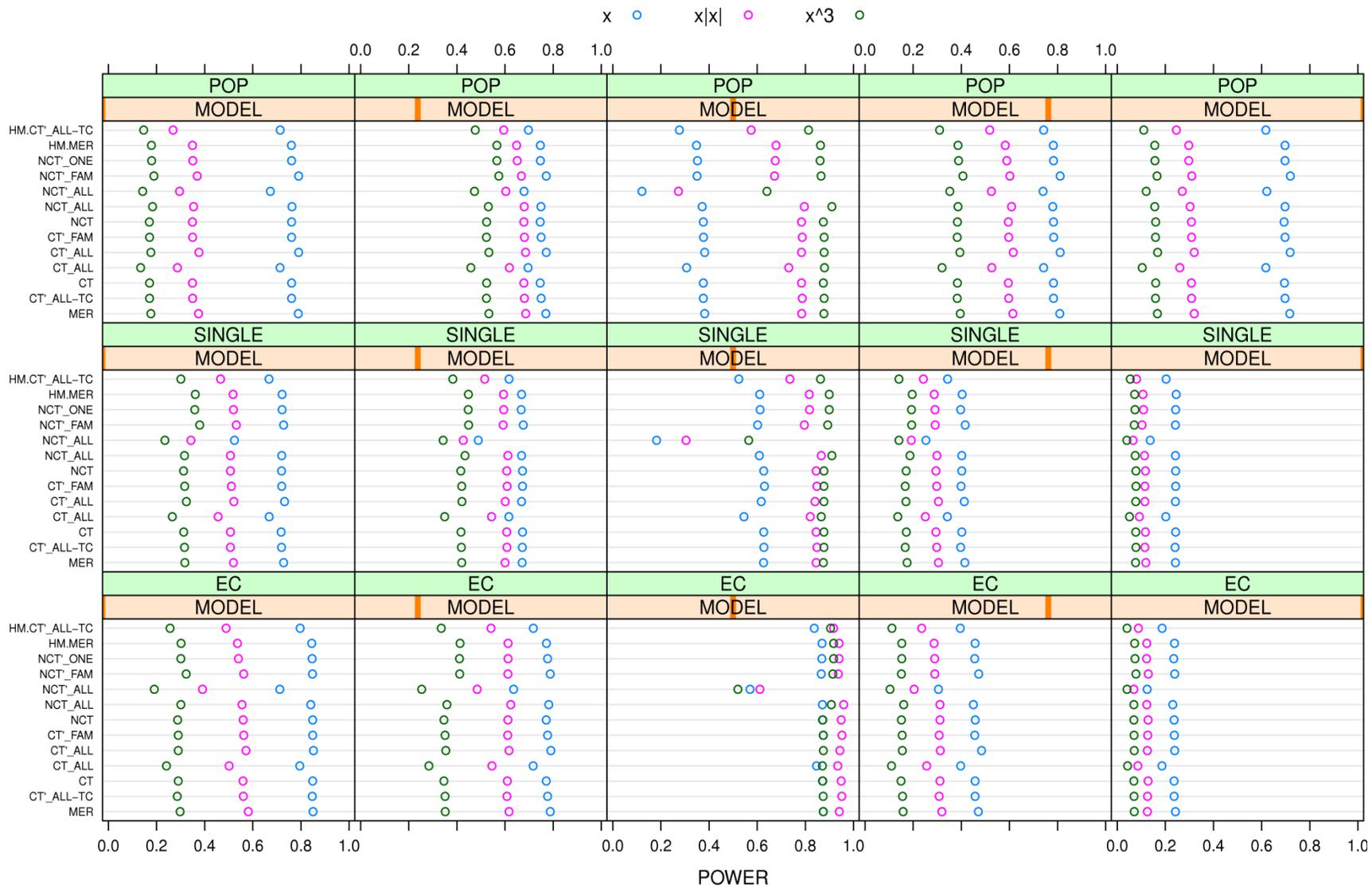
**Figure A3-4. 2+4SIBS power simulation results**

Note: CT'\_ALL-TC is SCORE.CT'<sub>ALL-TC</sub>, NCT'\_ALL is SCORE.NULL.CT'<sub>ALL</sub> and so on.



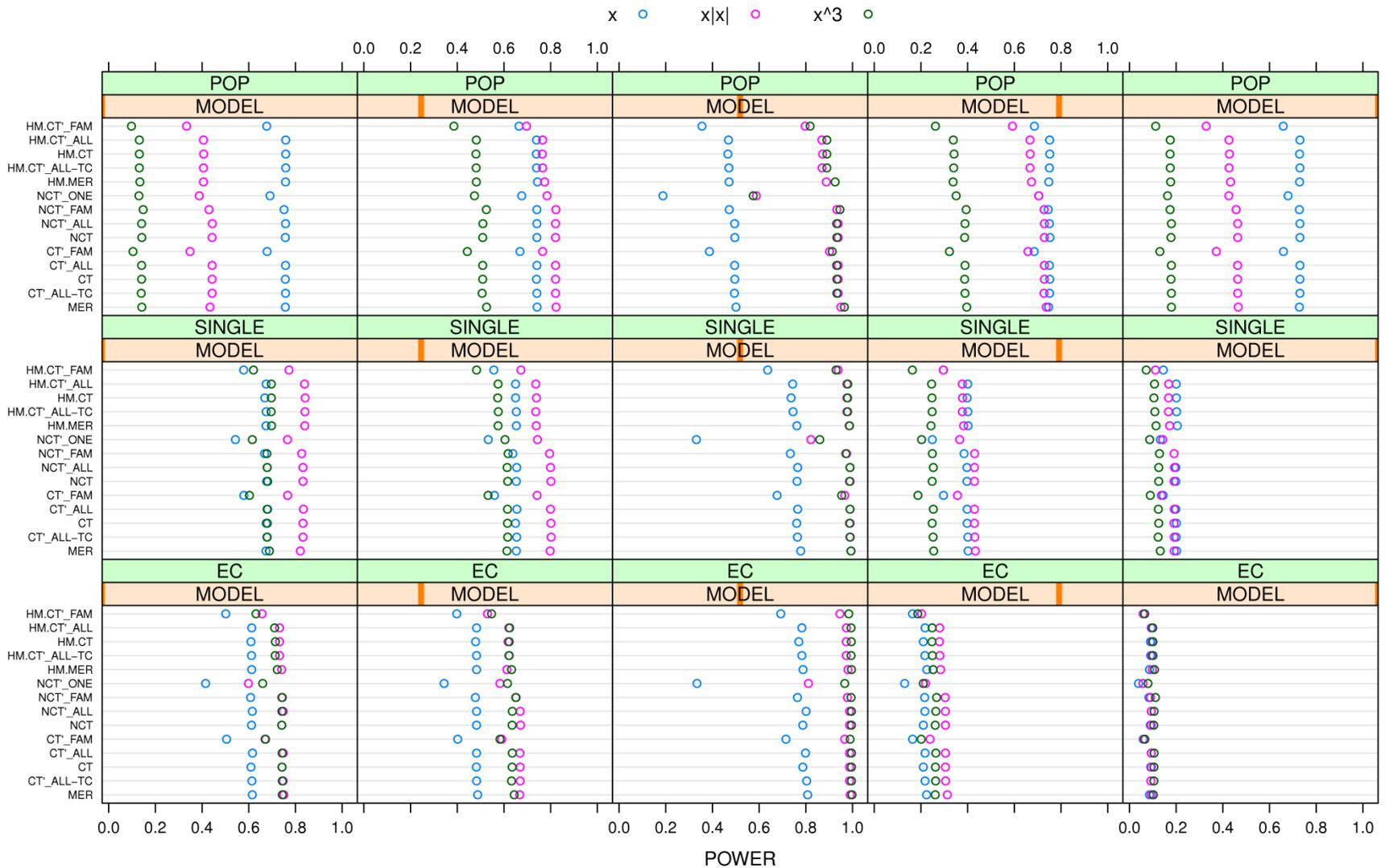
**Figure A3-5. 4SIBS+3G power simulation results**

Note: CT'\_ALL-TC is SCORE.CT'<sub>ALL-TC</sub>; NCT'\_ALL is SCORE.NULL.CT'<sub>ALL</sub> and so on.



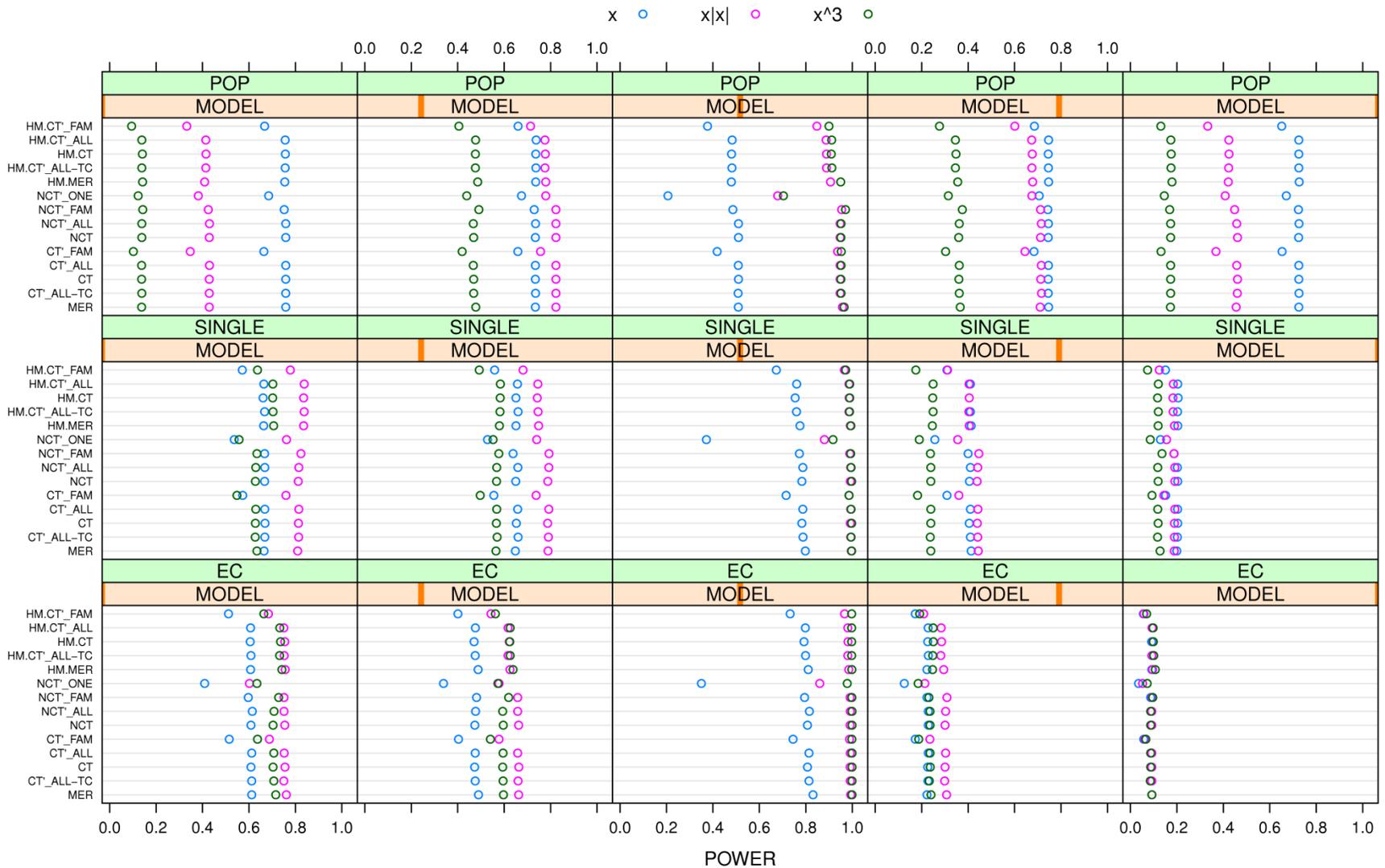
**Figure A3-6. 4SIBS+3G+HP power simulation results**

Note: CT'\_ALL-TC is SCORE.CT'<sub>ALL-TC</sub>; NCT'\_ALL is SCORE.NULL.CT'<sub>ALL</sub> and so on.



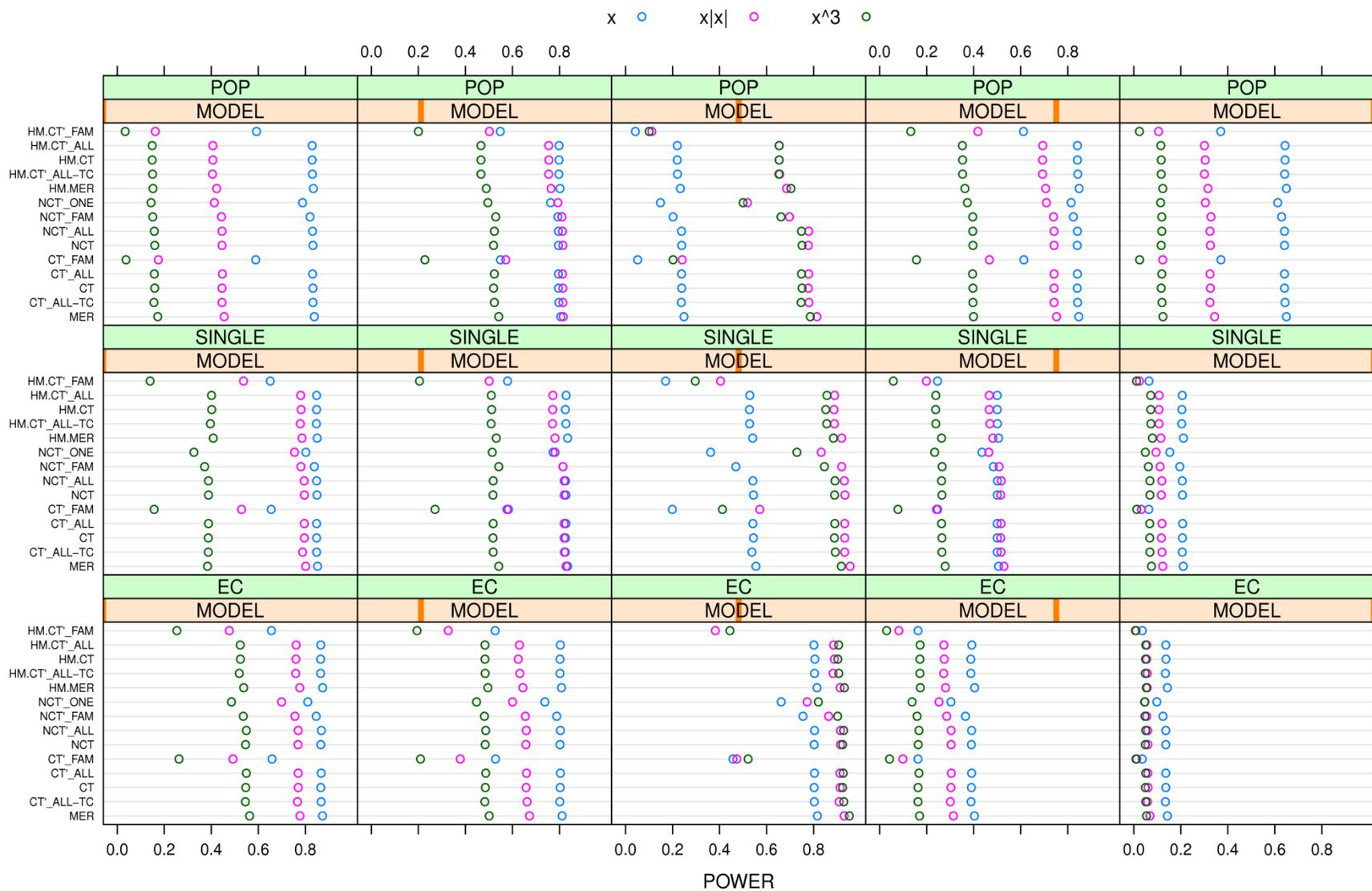
**Figure A3-7. 4SIBS power simulation results using model-based correlations estimated by PC pairs**

Note: CT'\_ALL-TC is SCORE.CT'\_ALL-TC, NCT'\_ALL is SCORE.NULL.CT'\_ALL and so on. Power for SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL were not presented because they are exactly equivalent to SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL for single pedigree types.



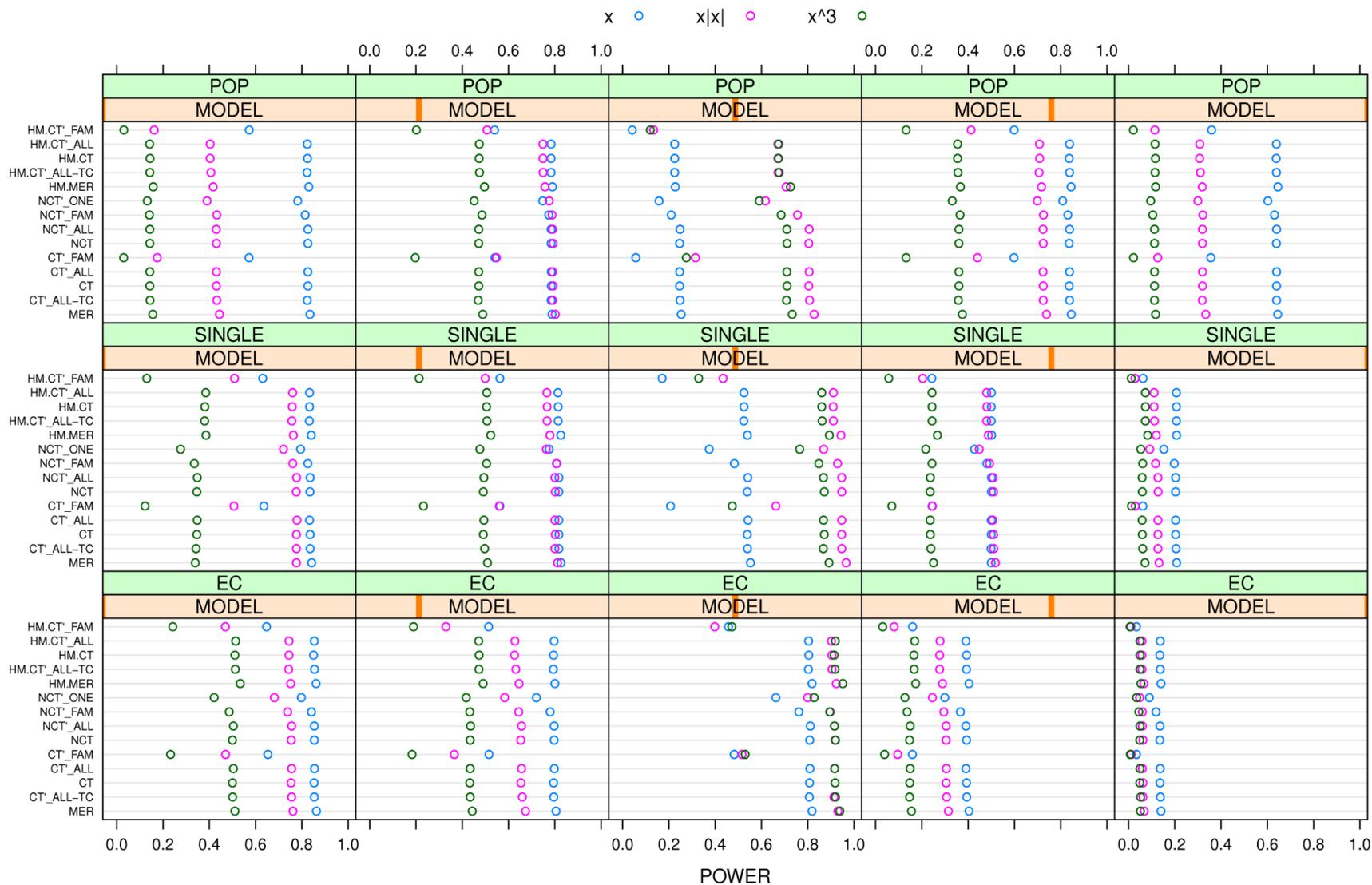
**Figure A3-8. 4SIBS power simulation results using model-based correlations estimated by SB pairs**

Note: CT\*\_ALL-TC is SCORE.CT\*\_ALL-TC, NCT\*\_ALL is SCORE.NULL.CT\*\_ALL and so on. Power for SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL were not presented because they are exactly equivalent to SCORE.CT\*\_ALL, SCORE.NULL.CT\*\_ALL, and HM.CT\*\_ALL for single pedigree types.



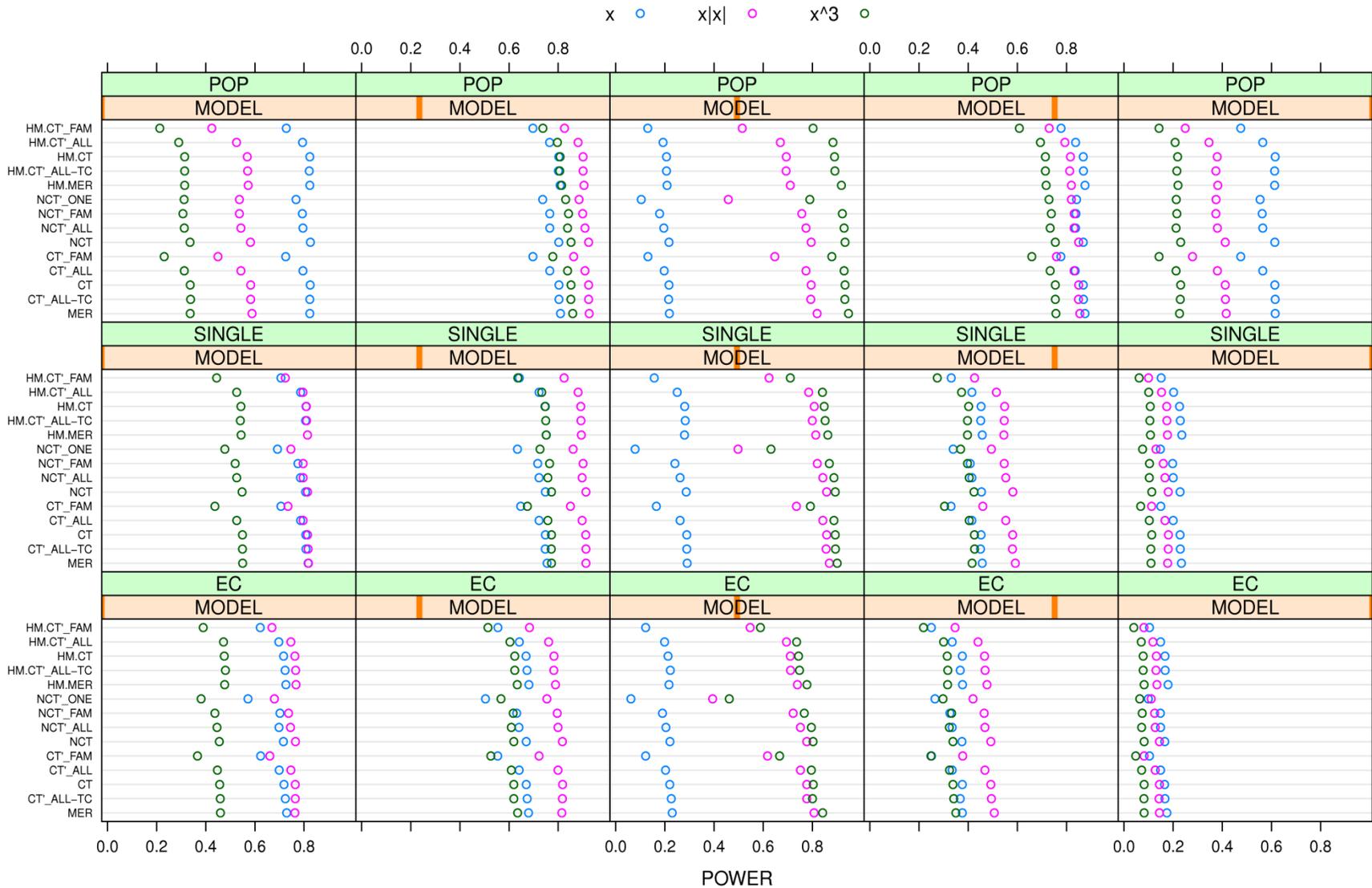
**Figure A3-9. HS power simulation results using model-based correlations estimated by PC pairs**

Note: CT'\_ALL-TC is SCORE.CT'\_ALL-TC, NCT'\_ALL is SCORE.NULL.CT'\_ALL and so on. Power for SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL were not presented because they are exactly equivalent to SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL for single pedigree types.



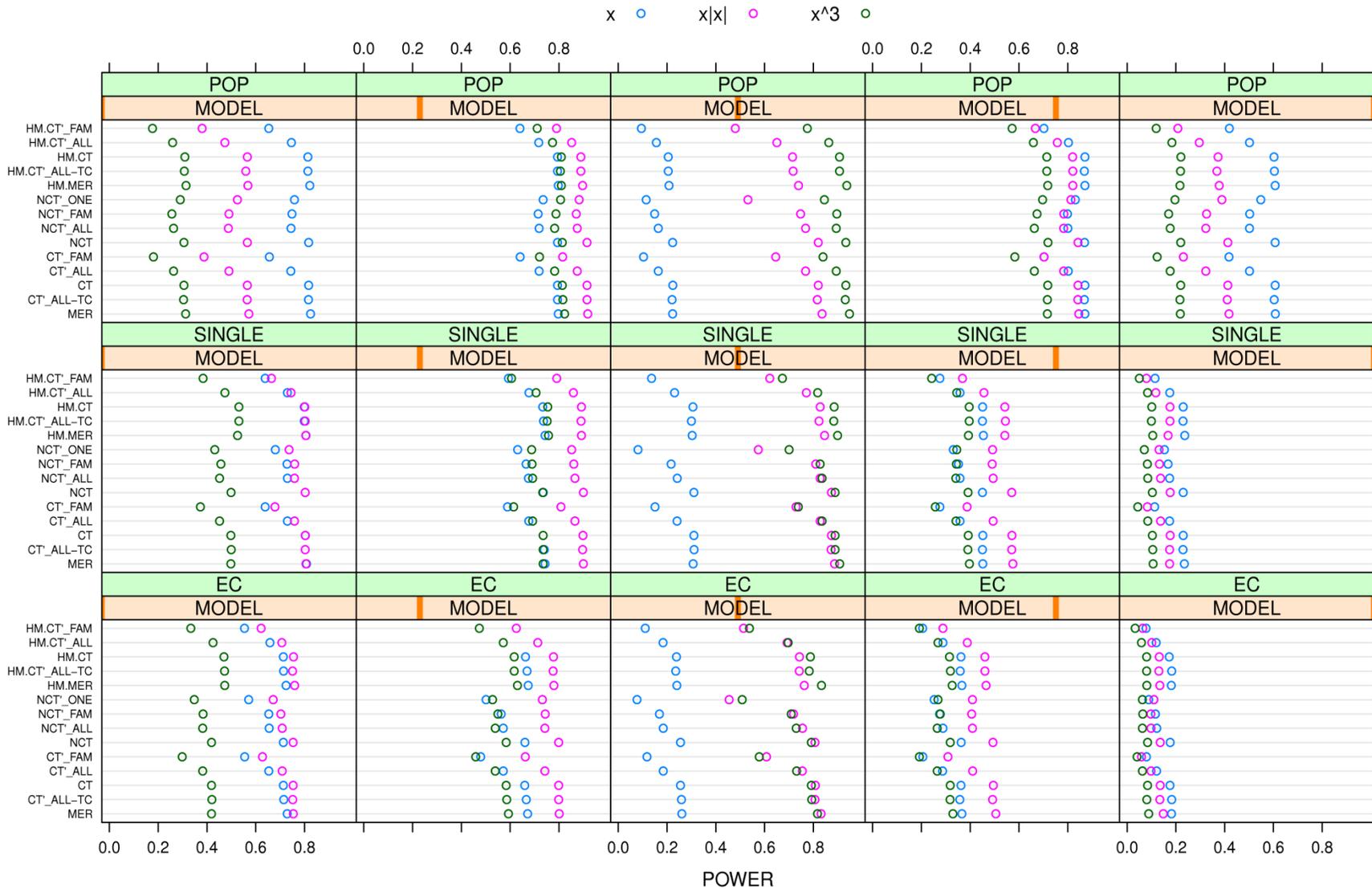
**Figure A3-10. HS power simulation results using model-based correlations estimated by SB pairs**

Note: CT'\_ALL-TC is SCORE.CT'\_ALL-TC, NCT'\_ALL is SCORE.NULL.CT'\_ALL and so on. Power for SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL were not presented because they are exactly equivalent to SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL for single pedigree types.



**Figure A3-11. 3G power simulation results using model-based correlations estimated by PC pairs**

Note: CT'\_ALL-TC is SCORE.CT'\_ALL-TC, NCT'\_ALL is SCORE.NULL.CT'\_ALL and so on. Power for SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL were not presented because they are exactly equivalent to SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL for single pedigree types.



**Figure A3-12. 3G power simulation results using model-based correlations estimated by SB pairs**

Note: CT'\_ALL-TC is SCORE.CT'\_ALL-TC, NCT'\_ALL is SCORE.NULL.CT'\_ALL and so on. Power for SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL were not presented because they are exactly equivalent to SCORE.CT'\_ALL, SCORE.NULL.CT'\_ALL, and HM.CT'\_ALL for single pedigree types.

#### A.4 ANALYTICAL COMPARISON OF SCORE.CT'\_{ALL-TC} AND

#### SCORE.NULL.CT'\_{ONE}

Let  $\hat{\pi}_{ij}$  be the mean-corrected estimated IBD sharing and  $v_{ij}$  be the transformed phenotype (see section 2.3.1) of the  $j$ th pair in the  $i$ th pedigree. The score statistic variance of SCORE.NULL.CT'\_{ONE} is given by

$$S_{\text{NULL.CT}'_{\text{ONE}}}^2 = \sum_i \left\{ \sum_p \sum_q v_{ip} v_{iq} \hat{\pi}_{ip} \hat{\pi}_{iq} \right\}$$

The score statistic variance of SCORE.CT'\_{ALL-TC} is similar to that based on *Estimated Variances* [Dupuis, et al. 2009] which in turn is asymptotically equivalent to the score statistic variance of SCORE.NULL.CT'\_{ALL}.

$$\begin{aligned} S_{\text{CT}'_{\text{ALL-TC}}}^2 &\approx S_{\text{Estimated}}^2 \\ &= \sum_i \left\{ \sum_p v_{ip}^2 \widehat{\text{Var}}(\hat{\pi}_{ip}) + \sum_{p \neq q} v_{ip} v_{iq} \text{Cor}(\pi_{ip}, \pi_{iq}) \sqrt{\widehat{\text{Var}}(\hat{\pi}_{ip}) \widehat{\text{Var}}(\hat{\pi}_{iq})} \right\} \\ &\approx \sum_i \left\{ \sum_p v_{ip}^2 \left( \sum_r \frac{\hat{\pi}_r^2}{n_r} \right) + \sum_{p \neq q} v_{ip} v_{iq} \left( \frac{\sum_r \sum_s \hat{\pi}_r \hat{\pi}_s}{n_{rs}} \right) \right\} \\ &= \sum_i \left\{ \sum_p \sum_q v_{ip} v_{iq} \overline{\pi_r \pi_s} \right\} = S_{\text{NULL.CT}'_{\text{ALL}}}^2 \end{aligned}$$

where the pairs  $r$  and  $p$  are of the same type and so are the pairs  $q$  and  $s$ ;  $n_r = \sum_r 1$  and  $n_{rs} = \sum_r \sum_s 1$ . The difference of  $S_{\text{NULL.CT}'_{\text{ONE}}}^2$  and  $S_{\text{NULL.CT}'_{\text{ALL-TC}}}^2$  is given by

$$\begin{aligned}
S_{\text{NULL.CT'ONE}}^2 - S_{\text{CT'ALL-TC}}^2 &\approx \sum_i \left\{ \sum_p \sum_q v_{ip} v_{iq} (\hat{\pi}_{ip} \hat{\pi}_{iq} - \overline{\pi_r \pi_s}) \right\} \\
&= \sum_i \left\{ \sum_p \sum_q (v_{ip} v_{iq} - \overline{v_r v_s} + \overline{v_r v_s}) (\hat{\pi}_{ip} \hat{\pi}_{iq} - \overline{\pi_r \pi_s}) \right\} \\
&= \sum_i \left\{ \sum_p \sum_q (v_{ip} v_{iq} - \overline{v_r v_s}) (\hat{\pi}_{ip} \hat{\pi}_{iq} - \overline{\pi_r \pi_s}) \right\}
\end{aligned}$$

Its expectation in fact is the covariance between  $v_{ip} v_{iq}$  and  $\hat{\pi}_{ip} \hat{\pi}_{iq}$

$$E \left( S_{\text{NULL.CT'ONE}}^2 - S_{\text{CT'ALL-TC}}^2 \right) = \text{Cov}(v_{ip} v_{iq}, \hat{\pi}_{ip} \hat{\pi}_{iq})$$

Assume that  $Y = (v_p, v_q, \hat{\pi}_p, \hat{\pi}_q)$  is distributed as a multivariate normal distribution with the mean vector  $(0, 0, 0, 0)$  and the covariance matrix

$$\Omega_Y = \sigma^2 \begin{bmatrix} 1 & r_e & \rho_1 & \rho \\ r_e & 1 & \rho & \rho_2 \\ \rho_1 & \rho & 1 & \tau \\ \rho & \rho_2 & \tau & 1 \end{bmatrix}$$

By Isserlis' theorem [Isserlis 1918],

$$E(X_i X_j X_k X_n) = \sigma_{ij} \sigma_{kn} + \sigma_{ik} \sigma_{jn} + \sigma_{in} \sigma_{jk}$$

where  $\sigma_{ij}$  denotes  $\text{Cov}(X_i, X_j)$ . We can calculate the  $\text{Cov}(v_p v_q, \hat{\pi}_p \hat{\pi}_q)$  by

$$\begin{aligned}
\text{Cov}(v_p v_q, \hat{\pi}_p \hat{\pi}_q) &= E(v_p v_q \hat{\pi}_p \hat{\pi}_q) - E(v_p v_q) E(\hat{\pi}_p \hat{\pi}_q) \\
&= \sigma^2 [(r_e \tau + \rho_1 \rho_2 + \rho^2) - r_e \tau] = \sigma^2 (\rho_1 \rho_2 + \rho^2)
\end{aligned}$$

Under the null hypothesis of no linkage, the correlation of the transformed phenotype and estimated IBD sharing within or between relative pairs within families is expected to be zero, i.e.  $\rho_1 = \rho_2 = \rho = 0$ . When that condition holds,  $Cov(v_p v_q, \hat{\pi}_p \hat{\pi}_q)$  is equal to zero and this implies that the two tests are asymptotically equivalent. Under the alternative hypothesis, SCORE.CT'\_{ALL-TC} is more powerful than SCORE.NULL.CT'\_{ONE} if  $\rho_1 \rho_2 + \rho^2 > 0$ , which is the case since  $\rho$  should be around zero and  $\rho_1$  and  $\rho_2$  are anticipated to be away from zero and in the same direction (evidence of linkage).

## APPENDIX B SUPPLEMENTAL MATERIALS FOR CHAPTER 3

### B.1 ANALYTICAL COMPARISON OF 2×2 ALLELE AND 2×3 TREND

We have taken the expected ratio derived by Sasieni [1997]

$$\frac{\chi_{2 \times 2 \text{ allele}}^2}{\chi_{2 \times 3 \text{ trend}}^2} = 1 + \frac{4n_0n_2 - n_1^2}{(n_1 + 2n_2)(n_1 + 2n_0)}$$

for analyzing the type I error and the power of 2×2 allele and 2×3 trend. Under the null, the expected ratio is supposed to be close to 1 if 2×3 trend is a standard and 2×2 allele has correct type I error. Under the alternative, the expected ratio is >1 (<1) if 2×2 allele (2×3 trend) has higher power. The expected ratio based on Taylor expansion is the sum of a number of terms in the form of  $E(r_0^{a_0} r_1^{a_1} r_2^{a_2})$  or  $E(s_0^{b_0} s_1^{b_1} s_2^{b_2})$  where  $a_i, b_j = 0, 1, 2, 3, 4, i, j = 0, 1, 2$ . Each can be obtained through the operation on factorial moments.

Let  $X = 4n_0n_2 - n_1^2$  and  $Y = (n_1 + 2n_2)(n_1 + 2n_0)$ . The first and the second-order Taylor expansion of the expectation of  $X$  over  $Y$  are approximated by

- First-Order Taylor Expansion

$$E\left(\frac{X}{Y}\right) \approx \frac{E(X)}{E(Y)}$$

- Second-Order Taylor Expansion [Mood, et al. 1974]

$$E\left(\frac{X}{Y}\right) \approx \frac{E(X)}{E(Y)} - \frac{1}{E(Y)^2} \text{cov}(X, Y) + \frac{E(X)}{E(Y)^3} \text{Var}(Y)$$

$$\begin{aligned}
&= \frac{E(X)}{E(Y)} - \frac{1}{E(Y)^2} [E(XY) - E(X)E(Y)] + \frac{E(X)}{E(Y)^3} [E(Y^2) - E(Y)^2] \\
&= \frac{E(X)}{E(Y)} - \frac{E(XY)}{E(Y)^2} + \frac{E(X)E(Y^2)}{E(Y)^3}
\end{aligned}$$

where  $E(X)$ ,  $E(Y)$ ,  $E(XY)$ , and  $E(Y^2)$  are functions of  $E(r_0^{a_0} r_1^{a_1} r_2^{a_2})$  and  $E(s_0^{b_0} s_1^{b_1} s_2^{b_2})$ .

$$\begin{aligned}
(r_0, r_1, r_2) &\sim \text{multinomial}(R; p_{10}, p_{11}, p_{12}), \\
(p_{10}, p_{11}, p_{12}) &= \left( \frac{q^2 f_0}{K}, \frac{2pqf_1}{K}, \frac{p^2 f_2}{K} \right); \\
(s_0, s_1, s_2) &\sim \text{multinomial}(S; p_{00}, p_{01}, p_{02}) \\
(p_{00}, p_{01}, p_{02}) &= \left( \frac{q^2(1-f_0)}{1-K}, \frac{2pq(1-f_1)}{1-K}, \frac{p^2(1-f_2)}{1-K} \right) \text{ and } \sum_{j=0}^2 p_{ij} = 1, i = 0, 1.
\end{aligned}$$

The generalized factorial moment [Mosimann 1962] is defined as

$$\begin{aligned}
E(r_0^{(a_0)} r_1^{(a_1)} r_2^{(a_2)}) &= R^{\sum_{i=0}^2 a_i} p_{10}^{a_0} p_{11}^{a_1} p_{12}^{a_2} \\
\text{where } r^{(a)} &= r(r-1)\mathbf{L} \dots (r-a+1) \text{ and } R^{\sum_{i=0}^2 a_i} = R(R-1)\mathbf{L} \dots \left( R - \sum_{i=0}^2 a_i + 1 \right).
\end{aligned}$$

Expectations in the form of  $E(r_0^{a_0} r_1^{a_1} r_2^{a_2})$  or  $E(s_0^{b_0} s_1^{b_1} s_2^{b_2})$  taking  $E(r_1^2)$  for example thus can be calculated by

$$E(r_1^2) = E(r_1^{(2)}) + E(r_1^{(1)}) = R(R-1)p_{11}^2 + Rp_{11} = Rp_{11}[(R-1)p_{11} + 1].$$

Our simulation-based results basically follow the prediction of expected ratio and the first-order expansion in fact works as well as the second-order for giving close expected ratios.

## B.2 TABLES AND FIGURES

**Table B2-1. Penetrances and sample sizes for interaction studies**

Locus effect	Genetic model <sup>a</sup>	Sample size <sup>b</sup>	Exposed			Non-exposed		
			$f_0$	$f_1$	$f_2$	$f_0$	$f_1$	$f_2$
add	1	300	0.01	0.015	0.02	0.01	0.015	0.02
	2	400	0.015	0.02	0.025	0.01	0.015	0.02
	3	600	0.01	0.015	0.02	0.01	0.01	0.01
	4	150	0.01	0.02	0.03	0.01	0.015	0.02
rec	1	2000	0.01	0.01	0.015	0.01	0.01	0.015
	2	3000	0.015	0.015	0.02	0.01	0.01	0.015
	3	5000	0.01	0.01	0.015	0.01	0.01	0.01
	4	1000	0.01	0.01	0.02	0.01	0.01	0.015
dom	1	500	0.01	0.015	0.015	0.01	0.015	0.015
	2	700	0.015	0.02	0.02	0.01	0.015	0.015
	3	1200	0.01	0.015	0.015	0.01	0.01	0.01
	4	300	0.01	0.02	0.02	0.01	0.015	0.015
over-dom	1	400	0.01	0.02	0.01	0.01	0.02	0.01
	2	600	0.015	0.025	0.015	0.01	0.02	0.01
	3	1200	0.01	0.02	0.01	0.01	0.01	0.01
	4	600	0.01	0.02	0.01	0.01	0.015	0.01
under-dom	1	600	0.02	0.01	0.02	0.02	0.01	0.02
	2	800	0.025	0.015	0.025	0.02	0.01	0.02
	3	1500	0.02	0.01	0.02	0.01	0.01	0.01
	4	600	0.02	0.01	0.02	0.02	0.015	0.02

<sup>a</sup>1. genetic effect only; 2. genetic and exposure main effects only; 3. gene×exposure interaction in which there is only a genetic effect in the exposed group, and 4. gene×exposure interaction with effects in both groups.

<sup>b</sup>Sample size required for each genetic model and locus effect to achieve the power 0.8 at the exposure frequency 0.5.

**Table B2-2. Power of single test procedure under each genetic model used for genome-wide simulations with sample size 500 (1500) and significance level 0.05 (0.0001)**

	$f_0$	$f_1$	$f_2$	2×2 geno	2×3 two df	2×2 allele	2×3 trend
ADD	0.01	0.015	0.02	0.805 (0.838)	0.737 (0.778)	0.814 (0.852)	<b>0.816 (0.854)</b>
REC	0.01	0.01	0.04	0.156 (0.011)	<b>0.800 (0.875)</b>	0.468 (0.266)	0.427 (0.199)
DOM	0.01	0.015	0.015	<b>0.767 (0.781)</b>	0.697 (0.686)	0.740 (0.726)	0.746 (0.740)

The highest power in each row appears in bold. Those in the parenthesis are for sample size 1500 and significance level 0.0001.

**Table B2-3. Power simulation results for the marker of intermediate genetic effect**

$p^a$	$R=S^b$	$f_0^c$	$f_1^d$	$f_2^e$	2×2 geno	2×3 two df	2×2 allele	2×3 trend
0.3	150	0.01	0.02	0.02	<b>0.85</b>	0.77	0.73	0.76
0.3	200	0.01	0.019	0.02	<b>0.90</b>	0.84	0.83	0.84
0.3	200	0.01	0.018	0.02	<b>0.85</b>	0.78	0.80	0.81
0.3	250	0.01	0.017	0.02	<b>0.88</b>	0.82	0.85	0.86
0.3	250	0.01	0.016	0.02	0.82	0.76	0.82	<b>0.83</b>
0.3	300	0.01	0.015	0.02	0.82	0.78	<b>0.85</b>	<b>0.85</b>
0.3	400	0.01	0.014	0.02	0.83	0.85	<b>0.91</b>	<b>0.91</b>
0.3	400	0.01	0.013	0.02	0.71	0.81	<b>0.87</b>	<b>0.87</b>
0.3	400	0.01	0.012	0.02	0.55	0.78	<b>0.83</b>	0.82
0.3	450	0.01	0.011	0.02	0.41	<b>0.84</b>	0.81	0.80
0.3	450	0.01	0.010	0.02	0.23	<b>0.88</b>	0.74	0.72

<sup>a</sup>Minor allele frequency.

<sup>b</sup>Sample size for cases and for controls.

<sup>c, d, e</sup>Penetrances for the genotypes with 0, 1, and 2 copies of minor alleles.

The highest power in each row appears in bold.

**Table B2-4. Penetrances of a marker in LD with a disease locus with complete genetic effect**

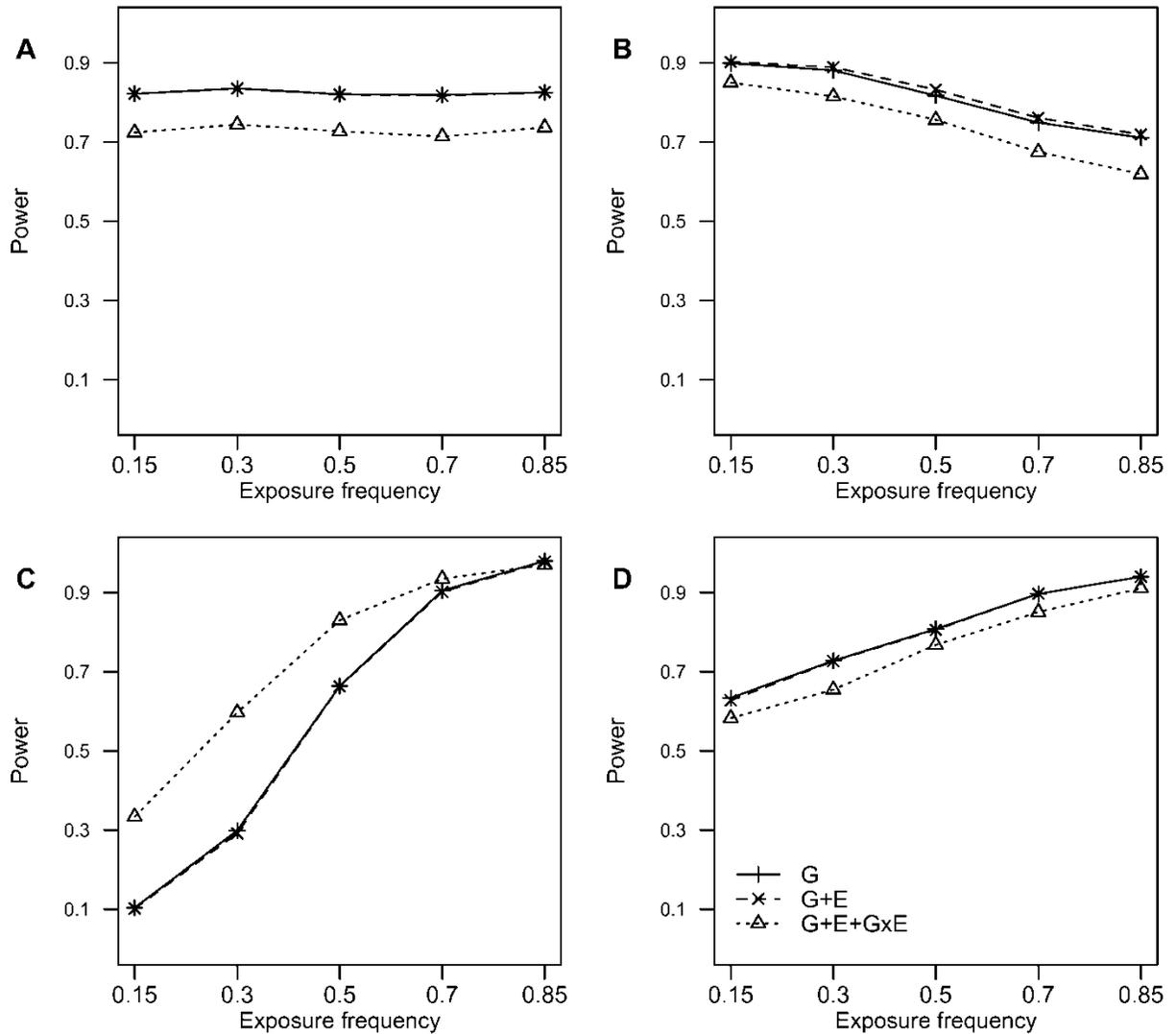
$R^{2a}$	$P(M_1D_1)^b$	$P(D_1)^c$	$P(M_1)^d$	Disease locus <sup>e</sup>	$f_{M_2M_2}^f$	$f_{M_1M_2}^g$	$f_{M_1M_1}^h$
1	0.2	0.2	0.2	Additive	0.01	0.015	0.02
				Dominant	0.01	0.015	0.015
				Recessive	0.01	0.01	0.015
0.583	0.2	0.2	0.3	Additive	0.01	0.0133	0.0167
				Dominant	0.01	0.0133	0.0144
				Recessive	0.01	0.0100	0.0122
0.375	0.2	0.2	0.4	Additive	0.01	0.0125	0.0150
				Dominant	0.01	0.0125	0.0138
				Recessive	0.01	0.0100	0.0113
0.473	0.15	0.2	0.2	Additive	0.0106	0.0141	0.0175
				Dominant	0.0106	0.0138	0.0147
				Recessive	0.0100	0.0102	0.0128
0.241	0.15	0.2	0.3	Additive	0.0107	0.0129	0.0150
				Dominant	0.0107	0.0127	0.0138
				Recessive	0.0100	0.0102	0.0113
0.128	0.15	0.2	0.4	Additive	0.0108	0.0123	0.0137
				Dominant	0.0108	0.0121	0.0130
				Recessive	0.0100	0.0102	0.0107
0.141	0.1	0.2	0.2	Additive	0.0112	0.0131	0.0150
				Dominant	0.0112	0.0128	0.0128
				Recessive	0.0101	0.0103	0.0113
0.048	0.1	0.2	0.3	Additive	0.0129	0.0131	0.0133
				Dominant	0.0124	0.0126	0.0128
				Recessive	0.0104	0.0105	0.0106
0.010	0.1	0.2	0.4	Additive	0.0117	0.0121	0.0125
				Dominant	0.0115	0.0119	0.0122
				Recessive	0.0101	0.0102	0.0103

<sup>a</sup>LD measure as defined in Appendix B.3.

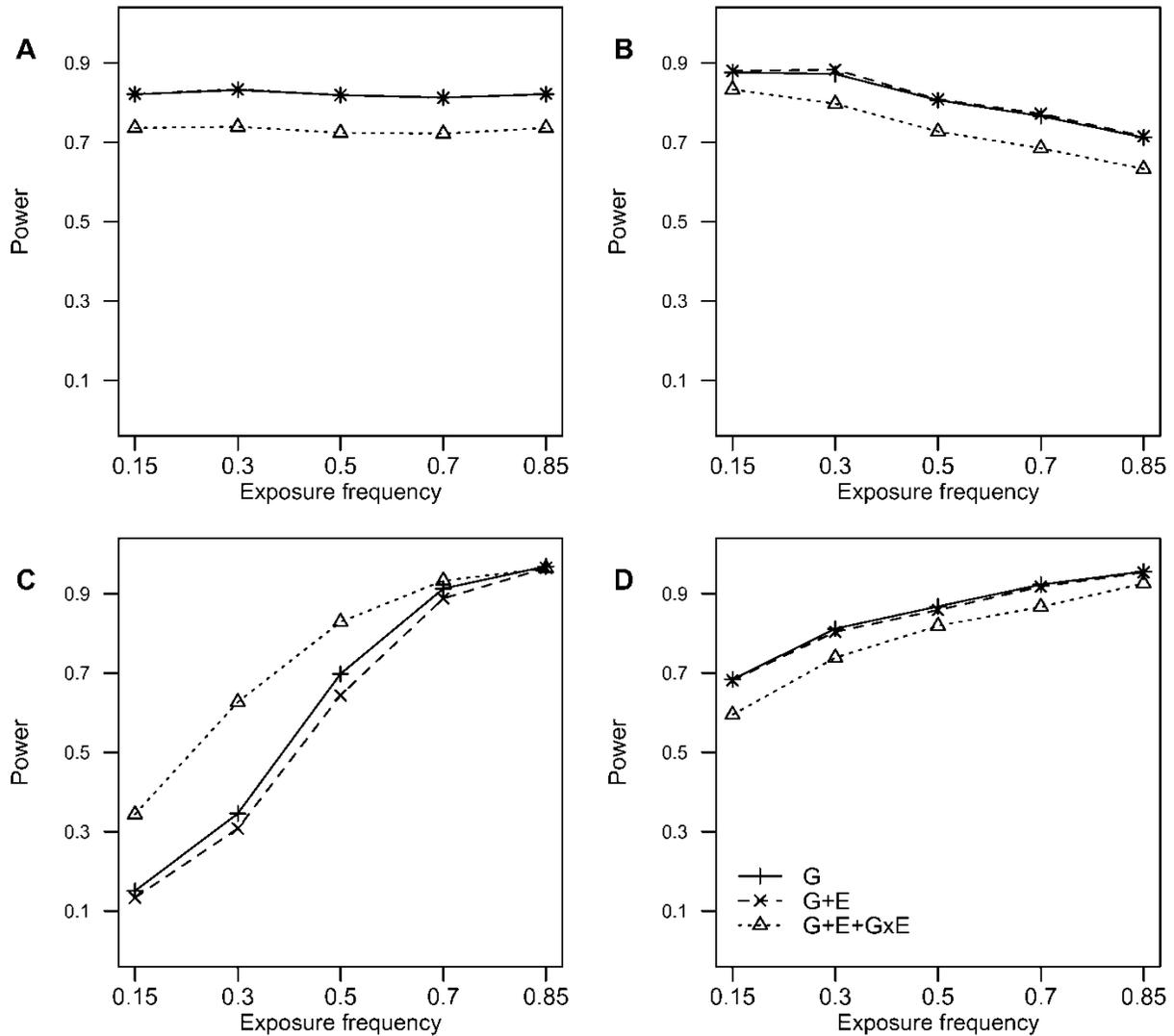
<sup>b, c, d</sup>Frequencies of  $M_1D_1$ ,  $M_1$ , and  $D_1$  where  $M_1(D_1)$  is the minor allele of the marker (disease locus).

<sup>e</sup>Genetic effect of the disease locus (see penetrances in Appendix B.3).

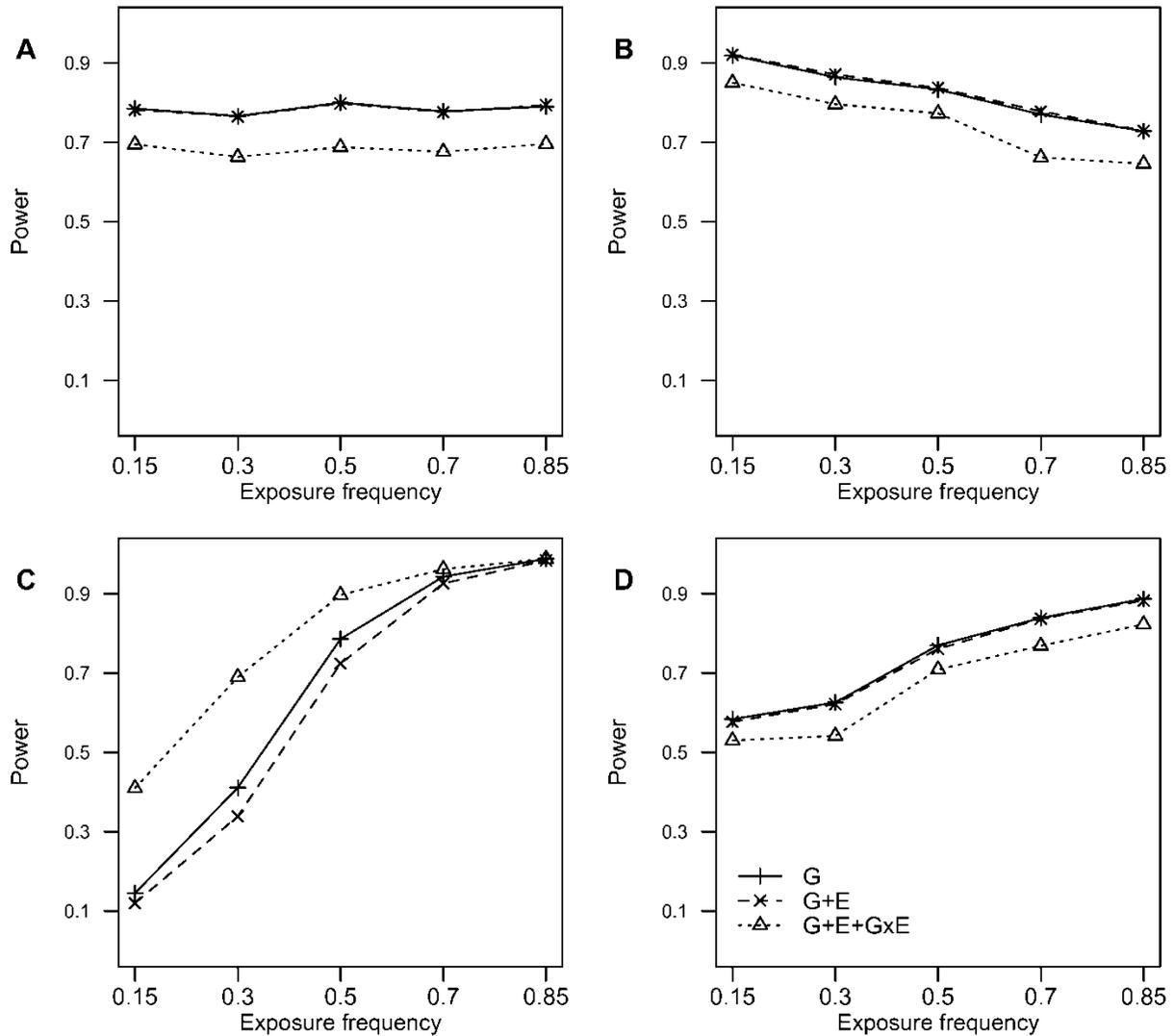
<sup>f, g, h</sup>Marker penetrances for the genotypes with 0, 1, and 2 copies of minor alleles respectively.



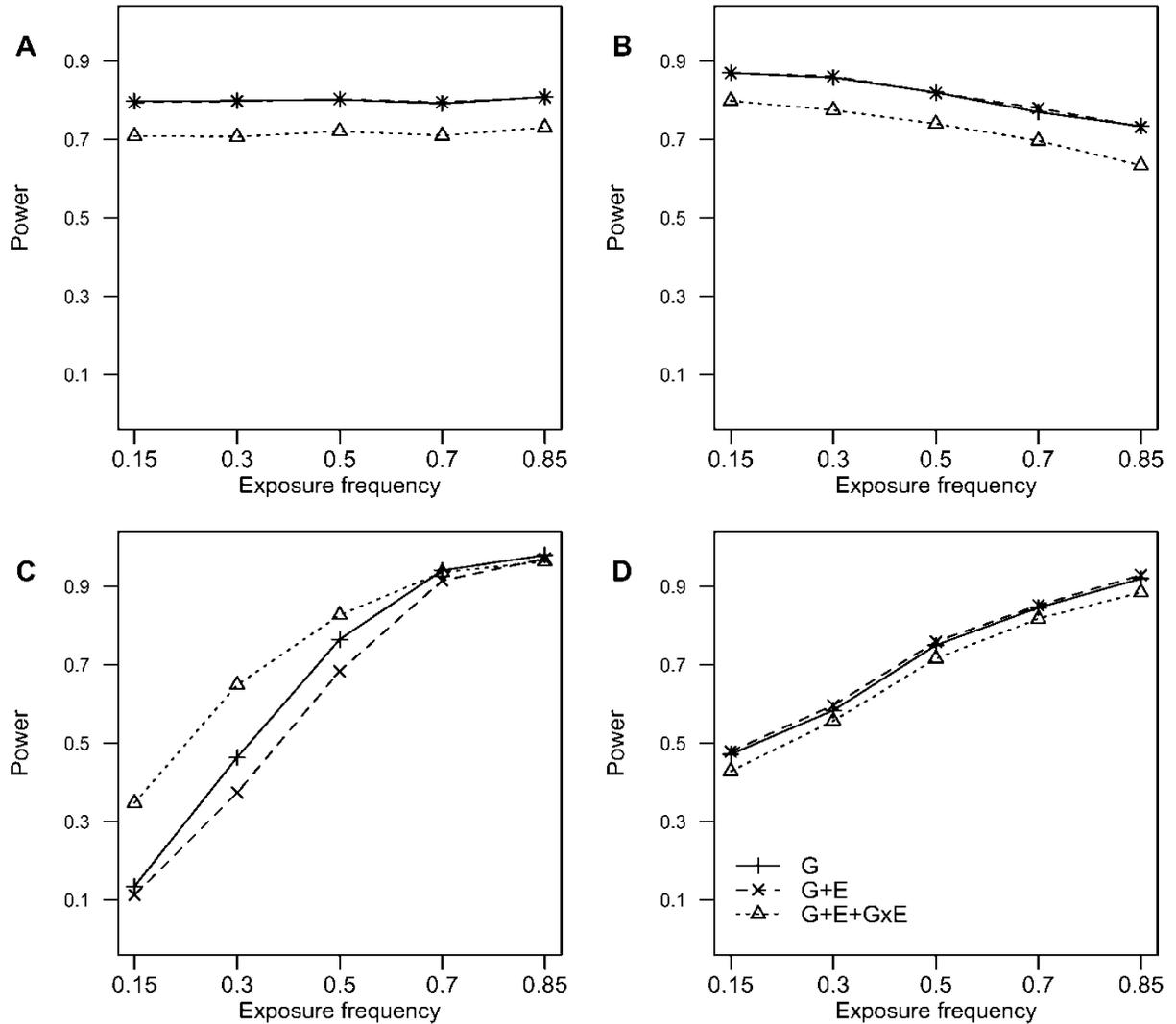
**Figure B2-1. Recessive marker power simulation results for the three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G \times E$  at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene  $\times$  exposure interaction in which there is only a genetic effect in the exposed group, and D. gene  $\times$  exposure interaction with effects in both groups.**



**Figure B2-2. Dominant marker power simulation results for the three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G \times E$  at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene  $\times$  exposure interaction in which there is only a genetic effect in the exposed group, and D. gene  $\times$  exposure interaction with effects in both groups.**



**Figure B2-3. Over-dominant marker power simulation results for the three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G\times E$  at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene $\times$ exposure interaction in which there is only a genetic effect in the exposed group, and D. gene $\times$ exposure interaction with effects in both groups.**



**Figure B2-4. Under-dominant marker power simulation results for the three fitted logistic regression models:  $G$ ,  $G+E$ , and  $G+E+G\times E$  at the exposure frequencies 0.15, 0.3, 0.5, 0.7, and 0.85 given the genotypic data simulated from the models, A. genetic effect only; B. genetic and exposure main effects only; C. gene $\times$ exposure interaction in which there is only a genetic effect in the exposed group, and D. gene $\times$ exposure interaction with effects in both groups.**

### B.3 GENETIC MODEL OF A MARKER IN LD WITH A DISEASE LOCUS

Since the disease loci might not be genotyped, markers near a completely recessive disease locus (for example) might have an intermediate genetic effect dependent on the LD and allele frequencies. We explored this effect as a function of marker penetrances, LD, and allele frequencies. Assume that both marker and disease locus are in HWE. The marker (disease locus) has minor allele  $M_1$  ( $D_1$ ) and major allele  $M_2$  ( $D_2$ ). Let  $f_{M_2M_2}$ ,  $f_{M_1M_2}$ , and  $f_{M_1M_1}$  ( $f_{D_2D_2}$ ,  $f_{D_1D_2}$ , and  $f_{D_1D_1}$ ) be the penetrances of marker (disease locus) for the genotypes with 0, 1, and 2 copies of minor alleles. One of linkage disequilibrium measures in common use, the correlation coefficient ( $R^2$ ), is calculated as

$$R^2 = \frac{D}{P(M_1)P(M_2)P(D_1)P(D_2)} \text{ where } D = P(M_1D_1) - P(M_1)P(D_1)$$

Given a completely additive ( $f_{D_2D_2}=0.01$ ,  $f_{D_1D_2}=0.015$ ,  $f_{D_1D_1}=0.02$ ), dominant ( $f_{D_2D_2}=0.01$ ,  $f_{D_1D_2}=0.015$ ,  $f_{D_1D_1}=0.015$ ), or recessive disease locus ( $f_{D_2D_2}=0.01$ ,  $f_{D_1D_2}=0.01$ ,  $f_{D_1D_1}=0.015$ ) with a minor allele frequency 0.2, the corresponding marker penetrances for each scenario are presented in Supplemental Table IV. With a decrease in  $R^2$  or an increase of allele frequency difference between the marker and the disease locus, markers in LD with a completely dominant or recessive disease locus become more additive. Markers in LD with a completely additive marker however remain additive regardless of  $R^2$  and allele frequencies. This can be simply proved as follows. To be general, denote the allele and haplotype frequencies of the marker and the disease locus as shown in the table below.

	$D_1$	$D_2$	
$M_1$	$x_1$	$y_1$	$r$
$M_2$	$x_2$	$y_2$	$s$
	$p$	$q$	

Marker penetrances for the genotypes with 0, 1, and 2 copies of minor alleles would be

$$f_{M_2M_2} = \frac{1}{s^2} (x_2^2 f_{D_1D_1} + 2x_2y_2 f_{D_1D_2} + y_2^2 f_{D_2D_2})$$

$$f_{M_1M_1} = \frac{1}{r^2} (x_1^2 f_{D_1D_1} + 2x_1y_1 f_{D_1D_2} + y_1^2 f_{D_2D_2})$$

$$f_{M_1M_2} = \frac{1}{rs} \{x_1x_2 f_{D_1D_1} + (x_1y_2 + x_2y_1) f_{D_1D_2} + y_1y_2 f_{D_2D_2}\}$$

For additive genetic effects, without loss of generality, let  $f_{D_2D_2} = 0$ ,  $f_{D_1D_2} = 1$ , and  $f_{D_1D_1} = 2$ . The three marker penetrances then are simplified to

$$f_{M_1M_1} = \frac{2x_1}{r}, \quad f_{M_1M_2} = \frac{2(sx_1 + rx_2)}{rs}, \quad f_{M_2M_2} = \frac{2x_2}{s}.$$

Then

$$f_{M_1M_1} + f_{M_2M_2} = \frac{2x_1}{r} + \frac{2x_2}{s} = \frac{2(sx_1 + rx_2)}{rs} = f_{M_1M_2}.$$

So the marker still has a completely additive genetic effect.

## BIBLIOGRAPHY

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30(1):97-101.
- Abecasis GR, Cookson WO, Cardon LR. 2000. Pedigree tests of transmission disequilibrium. *Eur J Hum Genet* 8(7):545-51.
- Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am J Hum Genet* 65(2):531-44.
- Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62(5):1198-211.
- Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 54(3):535-43.
- Armitage P. 1955. Tests for linear trends in proportions and frequencies. *Biometrics* 11:375-386.
- Bhattacharjee S, Kuo CL, Mukhopadhyay N, Brock GN, Weeks DE, Feingold E. 2008. Robust score statistics for QTL linkage analysis. *Am J Hum Genet* 82(3):567-82.
- Blackwelder WC, Elston RC. 1985. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* 2(1):85-97.
- Bukszar J, van den Oord EJ. 2006. Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for Pearson's statistic. *Biometrics* 62(4):1132-7.
- Chapman DG, Nam JM. 1968. Asymptotic power of chi square tests for linear trends in proportions. *Biometrics* 24(2):315-27.
- Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. 2006. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet* 79(6):1002-16.

- Chen WM, Broman KW, Liang KY. 2004. Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. *Genet Epidemiol* 26(4):265-72.
- Chen WM, Broman KW, Liang KY. 2005. Power and robustness of linkage tests for quantitative traits in general pedigrees. *Genet Epidemiol* 28(1):11-23.
- Clark AG. 2004. The role of haplotypes in candidate gene studies. *Genet Epidemiol* 27(4):321-33.
- Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 27(4):415-28.
- Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55(4):997-1004.
- Dupuis J, Shi J, Manning AK, Benjamin EJ, Meigs JB, Cupples LA, Siegmund D. 2009. Mapping quantitative traits in unselected families: algorithms and examples. *Genet Epidemiol* 33(7):617-27.
- Feingold E. 2001. Methods for linkage analysis of quantitative trait loci in humans. *Theor Popul Biol* 60(3):167-80.
- Feingold E. 2002. Regression-based quantitative-trait-locus mapping in the 21st century. *Am J Hum Genet* 71(2):217-22.
- Ferguson TS. 1996. A course in large sample theory. London ; New York: Chapman & Hall.
- Freidlin B, Zheng G, Li Z, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 53(3):146-52.
- Fulker DW, Cherny SS, Sham PC, Hewitt JK. 1999. Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* 64(1):259-67.
- Gail MH, Pfeiffer RM, Wheeler W, Pee D. 2008. Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics* 9(2):201-15.
- Guedj M, Della-Chiesa E, Picard F, Nuel G. 2007. Computing power in case-control association studies through the use of quadratic approximations: application to meta-statistics. *Ann Hum Genet* 71(Pt 2):262-70.
- Guedj M, Nuel G, Prum B. 2008. A note on allelic tests in case-control association studies. *Ann Hum Genet* 72(Pt 3):407-9.

- Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2(1):3-19.
- Isserlis L. 1918. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika* 12:134-139.
- Jackson MR, Genin E, Knapp M, Escary JL. 2002. Accurate power approximations for chi<sup>2</sup>-tests in case-control association studies of complex disease genes. *Ann Hum Genet* 66(Pt 4):307-21.
- Jakobsdottir J, Conley YP, Weeks DE, Mah TS, Ferrell RE, Gorin MB. 2005. Susceptibility genes for age-related maculopathy on chromosome 10q26. *Am J Hum Genet* 77(3):389-407.
- Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. 2009. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet* 5(2):e1000337.
- Knapp M. 2008. On the asymptotic equivalence of allelic and trend statistic under Hardy-Weinberg equilibrium. *Ann Hum Genet* 72(Pt 5):589.
- Knapp M, Seuchter SA, Baur MP. 1994a. Linkage analysis in nuclear families. 1: Optimality criteria for affected sib-pair tests. *Hum Hered* 44(1):37-43.
- Knapp M, Seuchter SA, Baur MP. 1994b. Linkage analysis in nuclear families. 2: Relationship between affected sib-pair tests and lod score analysis. *Hum Hered* 44(1):44-51.
- Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* 63(2):111-9.
- Kuo CL, Feingold E. 2010. What's the best statistic for a simple test of genetic association in a case-control study? *Genet Epidemiol* 34(3):246-53.
- Laird NM, Lange C. 2008. Family-based methods for linkage and association analysis. *Adv Genet* 60:219-52.
- Lake SL, Blacker D, Laird NM. 2000. Family-based tests of association in the presence of linkage. *Am J Hum Genet* 67(6):1515-25.
- Lange C, DeMeo DL, Laird NM. 2002. Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet* 71(6):1330-41.
- Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83(3):311-21.

- Li B, Leal SM. 2009. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet* 5(5):e1000481.
- Li Q, Wacholder S, Hunter DJ, Hoover RN, Chanock S, Thomas G, Yu K. 2009a. Genetic background comparison using distance-based regression, with applications in population stratification evaluation and adjustment. *Genet Epidemiol* 33(5):432-41.
- Li Q, Yu K, Li Z, Zheng G. 2008a. MAX-rank: a simple and robust genome-wide scan for case-control association studies. *Hum Genet* 123(6):617-23.
- Li Q, Zheng G, Li Z, Yu K. 2008b. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. *Ann Hum Genet* 72(Pt 3):397-406.
- Li Q, Zheng G, Liang X, Yu K. 2009b. Robust tests for single-marker analysis in case-control genetic association studies. *Ann Hum Genet* 73(2):245-52.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747-53.
- McArdle BH AJ. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82:290-297.
- Mood AM, Graybill FA, Boes DC. 1974. *Introduction to the theory of statistics*. New York,: McGraw-Hill.
- Moore JH, Asselbergs FW, Williams SM. 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26(4):445-55.
- Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34(2):188-93.
- Morton NE. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet* 7(3):277-318.
- Mosimann J. 1962. On the compound multinomial distribution, the multivariate beta distribution, and correlations among proportions. *Biometrika* 49:65-82.
- Ohashi J, Yamamoto S, Tsuchiya N, Hatta Y, Komata T, Matsushita M, Tokunaga K. 2001. Comparison of statistical power between 2 \* 2 allele frequency and allele positivity tables in case-control studies of complex disease genes. *Ann Hum Genet* 65(Pt 2):197-206.
- Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.

- Peng J, Siegmund D. 2006. QTL mapping under ascertainment. *Ann Hum Genet* 70(Pt 6):867-81.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904-9.
- Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65(1):220-8.
- Pritchard JK, Stephens M, Donnelly P. 2000a. Inference of population structure using multilocus genotype data. *Genetics* 155(2):945-59.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000b. Association mapping in structured populations. *Am J Hum Genet* 67(1):170-81.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559-75.
- Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 50(4):211-23.
- Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273(5281):1516-7.
- Sasieni PD. 1997. From genotypes to genes: doubling the sample size. *Biometrics* 53(4):1253-61.
- Schork NJ, Greenwood TA. 2004. Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. *Am J Hum Genet* 74(2):306-16.
- Selinger-Leneman H, Genin E, Norris JM, Khlat M. 2003. Does accounting for gene-environment (GxE) interaction increase the power to detect the effect of a gene in a multifactorial disease? *Genet Epidemiol* 24(3):200-7.
- Sham PC, Purcell S, Cherny SS, Abecasis GR. 2002. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet* 71(2):238-53.
- Slager SL, Schaid DJ. 2001. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum Hered* 52(3):149-53.

- Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 62(2):450-8.
- Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52(3):506-16.
- Stein CM, Elston RC. 2009. Finding genes underlying human disease. *Clin Genet* 75(2):101-6.
- Szatkiewicz JP, K TC, Feingold E. 2003. Recent advances in human quantitative-trait-locus mapping: comparison of methods for discordant sibling pairs. *Am J Hum Genet* 73(4):874-85.
- T. Cuenco K, Szatkiewicz JP, Feingold E. 2003. Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. *Am J Hum Genet* 73(4):863-73.
- Thomas DC. 2004. *Statistical methods in genetic epidemiology*. New York: Oxford University Press.
- Thornton T, McPeck MS. 2007. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet* 81(2):321-37.
- Thornton T, McPeck MS. 2010. ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. *Am J Hum Genet* 86(2):172-184.
- Wald A. 1947. *Sequential analysis*. New York: Dover Publications.
- Wang K. 2005. A likelihood approach for quantitative-trait-locus mapping with selected pedigrees. *Biometrics* 61(2):465-73.
- Williams JT, Van Eerdewegh P, Almasy L, Blangero J. 1999. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. *Am J Hum Genet* 65(4):1134-47.
- Wright FA. 1997. The phenotypic difference discards sib-pair QTL linkage information. *Am J Hum Genet* 60(3):740-2.
- Yu K, Wang Z, Li Q, Wacholder S, Hunter DJ, Hoover RN, Chanock S, Thomas G. 2008. Population substructure and control selection in genome-wide association studies. *PLoS One* 3(7):e2551.

- Zaykin DV, Zhivotovsky LA. 2005. Ranks of genuine associations in whole-genome scans. *Genetics* 171(2):813-23.
- Zaykin DV, Zhivotovsky LA, Czika W, Shao S, Wolfinger RD. 2008. P-values may not provide optimal ranks of true associations in whole-genome scan. American Society of Human Genetics Meeting.
- Zhao H. 2000. Family-based association studies. *Stat Methods Med Res* 9(6):563-87.
- Zheng G, Freidlin B, Li Z, Gastwirth J. 2003. Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical Journal*(45):335-348.
- Zheng G, Ng HK. 2008. Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics* 9(3):391-9.
- Zheng M, McPeck MS. 2007. Multipoint linkage-disequilibrium mapping with haplotype-block structure. *Am J Hum Genet* 80(1):112-25.
- Zhu X, Feng T, Li Y, Lu Q, Elston RC. 2010. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol* 34(2):171-87.
- Ziegler A, König IR. 2006. A statistical approach to genetic epidemiology : concepts and applications. Weinheim: Wiley-VCH.