# STATISTICAL METHODS IN GENETIC ASSOCIATION STUDIES

**by**

**Rui Xiao**

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2010

Doctoral Committee:

Professor Michael L. Boehnke, Chair
Professor Gonçalo Abecasis
Associate Professor David T. Burke
Assistant Professor Sebastian K. Zöllner
Associate Research Scientist Laura J. Scott

To my grandma, parents, and Yujie

## Acknowledgements

I would have never finished this dissertation without the help of many individuals. Above all, I would like to express my most sincere gratitude to my advisor, Dr. Michael Boehnke, for his tremendous amount of support and guidance throughout this entire process. Mike has been invaluable in helping me develop interesting dissertation topics, improve my writing skills, and prepare for professional presentations, and also in encouraging me to pursue my career. I am particularly grateful to Dr. Laura Scott for sharing her ideas and spending so much time to help me prepare the manuscript of the allelic expression imbalance study. I would also like to thank the other members of my dissertation committee: Dr. Sebastian Zöllner, for thoughtful discussions on the winner's curse paper; Dr. Dave Burke, for helping me understand the biology of the AEI; and Dr. Gonçalo Abecasis, for insightful suggestions on different problems in this dissertation. I would especially like to acknowledge Dr. Peter Song for constructive advice on the statistic methods which contribute a great deal to the AEI study.

I am truly thankful to my friends for their continuing encouragement and support during my long journey of graduate school at Michigan: Wen Ye, Weihua Guan, Jun Ding, Hui Zhang, Xin Gao, Xiaobi Huang, Wei Chen, Matt Zawistowski and Shyam Gopalakrishnan. I also appreciate the support from Peggy White, Anne Jackson, Heather Stringham, Tanya Teslovich and Cristen Willer from our wonderful FUSION group.

Finally, I would like to thank my family for their unwavering support and love throughout my life. Most importantly, I thank my husband Yujie, for making many sacrifices and for helping me in many ways. Without his constant love, support, and understanding, I would have not been able to finish this dissertation.

# TABLE OF CONTENTS

# LIST OF FIGURES

## Figure

# LIST OF TABLES

## Table

# CHAPTER 1

# INTRODUCTION

The past decade has witnessed a tremendous expansion of genetic resources for biomedical research. The wealth of data generated by the Human Genome Project [The International Human Genome Sequencing Consortium, 2001], the International HapMap Project [The International HapMap Consortium, 2007], and the ongoing 1000 Genomes Project [www.1000genomes.org], present great opportunities for statisticians to contribute critical concepts and methods to this field.

Genetic association studies are a powerful tool to detect genetic variants influencing human diseases or traits. High-throughput genotyping technologies make it possible to conduct genomewide association studies (GWAS) of common genetic variation across the entire human genome. In the past three years, there has been a dramatic increase in genetic discoveries involving complex diseases, with hundreds of common genetic variants for more than 80 diseases and traits identified and replicated in GWAS [www.genome.gov/gwastudies]. Analysis and interpretation of GWAS raise interesting statistical challenges, for example, the massive number of statistical tests performed presents a potential for false-positive results, requiring stringent statistical significance levels and replication of findings.

Chapter 2 and 3 of my dissertation are motivated by an important problem encountered in GWAS. Once an associated variant is identified, investigators are often

interested in estimating the genetic effect size of the identified variant. However, estimates of the genetic effect based on the initial GWAS sample(s), in which the significant associations were detected, tend to be upwardly biased as a consequence of "winner's curse" [Lohmueller et al., 2003]. Overestimation of the genetic effect size in initial studies may cause follow-up studies to be underpowered and so to fail.

In Chapter 2, I study the impact of the winner's curse in the context of genetic case-control association studies. I analytically quantify the bias in the estimates of the allele frequency difference and odds ratio, which are used as measures of the strength of the effect in such studies. I show that in realistic situations, these uncorrected estimators can be substantially overestimated, and that the overestimation decreases as power increases. I then propose an ascertainment-corrected maximum likelihood method to reduce the bias of these estimators. I demonstrate that the ascertainment-corrected estimator results in reduced absolute bias compared to the naïve uncorrected estimator when study power is low or moderate (<60%), a range that is typical for most large-scale genetic association studies, and has similar absolute bias when power is higher. I extend these calculations to two-stage association studies, and find that for optimal two-stage designs [Skol et al., 2007], results are similar to those for the corresponding one-stage designs. [Xiao and Boehnke, 2009].

Associations between genotype and disease-related quantitative traits (QTs), such as cholesterol level, body mass index, and systolic and diastolic blood pressure have also been investigated. One rationale behind QT studies is that, because many of the traits examined are closely related to one or more diseases, any identified quantitative trait locus (QTL) may also be a disease predisposing locus. As for case-control association

studies, for QT association studies, investigators usually focus on genetic loci showing significant evidence for SNP-trait association. Again, the estimator of the genetic effect size also tends to overestimate the true effect size as a consequence of the winner's curse.

In Chapter 3, I extend the study of the winner's curse to QT association studies, in which the genetic effect size is parameterized as the slope in a linear regression model. I use analytical calculation to demonstrate that overestimation in the regression slope estimate decreases as power increases. To reduce the ascertainment bias, I propose a three-parameter maximum likelihood method in which the intercept, slope, and error of the linear regression model are estimated. I also simplify this three-parameter likelihood model to a one-parameter model by excluding the nuisance parameters (the regression intercept and the error), since the regression slope is the primary interest of investigators. I show that both likelihood methods reduce the bias when power to detect association is low or moderate, and the one-parameter model generally results in a slope estimator with smaller variance.

Recent development of GWAS has enabled investigation of common variants in most of the human genome, including the non-coding regions which comprise ~98% of the genome [International Human Genome Sequencing Consortium, 2004]. Index SNPs for about 88% of GWAS signals are located in non-coding regions, either in intronic (45%) or intergenic (43%) regions [Hindorff et al., 2009; www.genome.gov/gwastudies]. The relationship between these associated SNPs and the disease is frequently unclear, but regulation of gene expression might be a candidate to account for the connection between these polymorphisms and the disease. Identification of the functional variant(s) and its mechanism of action is often made more difficult by the presence of multiple genes in the

associated region. Testing for association between the associated SNPs and the genes expression levels has the potential to help identify the gene(s) most likely to influence the trait.

Allelic expression imbalance (AEI) between the two alleles of a gene can be used to detect *cis*-acting regulatory SNPs (rSNP) in individuals heterozygous for a transcribed SNP (tSNP). Here the *cis*-regulatory elements are defined as the DNA polymorphisms that reside on the same chromosome as the gene they regulate, and only regulate the allele of the gene on the same chromosome. These elements are often in close proximity to the gene they regulate, but can also be located further away. In contrast, the *trans*-regulatory elements can be located on the same or a different chromosome as the gene they modulate and regulate both alleles of the gene. AEI is measured in individuals heterozygous for the tSNP. An advantage of using AEI is that both alleles are measured within the same environment in each individual, allowing a more direct comparison of *cis*-acting variants. Consequently, AEI will specifically identify the *cis*-regulatory elements, whereas testing for association between total expression level and SNP genotype will identify both *cis* and *trans*-acting elements [Bray et al., 2003; Mahr et al., 2006; Pastinen et al., 2003; Pastinen et al., 2005; Tao et al., 2006]. Methods for AEI analysis vary depending on whether we know phase of the rSNP and tSNP, and linkage disequilibrium (LD) between the rSNP and tSNP if we do not know the phase. Some current AEI analysis methods require phase known data [Serre et al., 2008; Ge et al., 2009], others rely on the LD between the rSNP and tSNP for haplotype reconstruction [Tao et al., 2006; Alachkar et al., 2008].

In Chapter 4, I focus on the situation when LD between the rSNP and tSNP is incomplete (D' < 1) and there is no phase information for the rSNP and tSNP. I propose five tests to detect association between the potential rSNP and AEI, assuming initially that AEI is due to a single rSNP. I show that the type I error rates for all these tests are well controlled, and demonstrate the relative power of the tests strongly depends on the magnitude of the LD between the rSNP and tSNP, and less strongly on the AEI effect size of the rSNP, the number of tSNP heterozygotes, and the allele frequencies of the rSNP and tSNP. I further demonstrate that the impact of a second ungenotyped rSNP on the power of these tests depends on the LD structure of the three SNPs, but almost never invalidates the proposed tests nor substantially changes the rankings of the tests for a given level of LD between the genotyped rSNP and the tSNP. I recommend the use of F test when the rSNP and tSNP are in or near linkage equilibrium (D'~0). When the two SNPs are in LD, in general, the mixture-model based test is most powerful for the intermediate LD levels, and the t test is typically most powerful for high LD.

## References


Alachkar H, Kataki M, Scharre DW, Audrey Papp A, Sadee W. 2008. Allelic mRNA expression of sortilin-1 (*SORL1*) mRNA in Alzheimer's autopsy brain tissues. Neurosci Lett 448(1): 120-124.

Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. 2003. *Cis*-acting variation in the expression of a high proportion of genes in human brain. Hum Genet 113: 149-153.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KCL, Gagné, Dias J, Hoberman R, Montpetit A, Joly M-M, Harvey EJ, Sinnet D, Beaulieu P, Hamon R, Graziani A, Dewar K, Harmsen E, Majewski J, Goring HHH, Naumova AK, Blanchette M, Gunderson KL, Pastinen T. 2009. Global patterns of *cis* variation in human cells revealed by high-density allelic expression analysis. Nat Genet 41: 1216-1222.

Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A catalog of published genome-wide association studies. www.genome.gov/gwastudies.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. [May 27, 2009].

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177-182.

Mahr S, Burmester GR, Hilke D, Göbel U, Grützkau A, Häupl T, Hauschild M, Koczan D, Krenn V, Neidel J, Perka C, Radbruch A, Thiesen HJ, Müller B. 2006. *Cis*- and *trans*-acting gene regulation is associated with osteoarthritis. Am J Hum Genet 78: 793-803.

Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H, Beck A, Verner A, Kingsley J, Harmsen E, Labuda D, Morgan K, Vohl MC, Naumova AK, Sinnett D, Hudson TJ. 2003. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics 16: 184-193.

Pastinen T, Ge B, Hudson TJ. 2006. Influence of human genome polymorphism on gene expression. Hum Mol Genet 15 Spec No 1: R9-16

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan J-B, Hudson TJ. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. PLoS Genet 4: e1000006.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2007. Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31: 776-788.

Tao H, Cox DR, Frazer KA. 2006. Allele-specific *KRT1* expression is a complex trait. PLoS Genet 2: e93.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

The International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-920.

The International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431: 915-916.

Xiao R, Boehnke M. 2009. Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol 33: 453-462.

# CHAPTER 2

# QUANTIFYING AND CORRECTING FOR THE WINNER'S CURSE IN GENETIC ASSOCIATION STUDIES

Genetic association studies are a powerful tool to detect genetic variants that predispose to human disease. Once an associated variant is identified, investigators are also interested in estimating the effect of the identified variant on disease risk. Estimates of the genetic effect based on new association findings tend to be upwardly biased due to a phenomenon known as the "winner's curse". Overestimation of genetic effect size in initial studies may cause follow-up studies to be underpowered and so to fail. In this paper, we quantify the impact of the winner's curse on the allele frequency difference and odds ratio estimators for one- and two-stage case-control association studies. We then propose an ascertainment-corrected maximum likelihood method to reduce the bias of these estimators. We show that overestimation of the genetic effect by the uncorrected estimator decreases as the power of the association study increases and that the ascertainment-corrected method reduces absolute bias and mean square error unless power to detect association is high.

## 2.1 Introduction

Large-scale genetic association studies are now commonly used to localize genetic variants that predispose to a wide range of human diseases. In genetic association studies, once the disease-predisposing variants are identified, it is of interest to estimate the genetic effect of those variants on disease risk. The simplest method of estimating the effect size of the variant is to calculate the difference of the observed risk allele frequency between cases and controls or the corresponding odds ratio. However, these naïve estimators are likely to overestimate the true genetic effect size as a consequence of the "winner's curse" [Lohmueller et al., 2003], a phenomenon first described in the auction theory literature [Bazerman and Samuelson, 1983]. In auctions, participants place bids on an item. Even if the bids are unbiased, the winning bid is likely to overestimate the true item value since it is the highest among all the bids. In genetic association studies, an initial positive finding plays the role of the winning bid, since we generally focus on genetic effect size estimates only for the variants that yield significant evidence for association, resulting in effect size estimates that are upwardly biased. We refer to this bias as 'ascertainment bias' since it is caused by ascertaining only those samples that result in significant association evidence. If the sample size calculation for a subsequent study is based on an overestimated effect size, replication studies are likely to be underpowered and so more likely to fail. A review of association studies [Ioannidis et al., 2001] has described the overestimation in first positive reports, consistent with the winner's curse.

This problem has drawn attention from several investigators in the context of genetic linkage and association studies [Göring et al., 2001; Siegmund, 2002; Allison et

al., 2003; Sun and Bull, 2005; Wu et al., 2006; Garner, 2007; Yu et al., 2007; Zöllner and Pritchard, 2007; Zhong and Prentice, 2008; Ghosh et al., 2008]. Göring et al. [2001] recommended the use of two independent datasets: one for locus mapping, the other for parameter estimation. An obvious disadvantage of this strategy is the power loss due to splitting the sample in two. Sun and Bull [2005] proposed resampling estimators that employ repeated random sample splitting of the data via cross-validation or the bootstrap. Wu et al. [2006] compared their bootstrap estimators for locus-specific quantitative trait linkage analysis, and, in the context of two-stage design, Yu et al. [2007] applied a bootstrap estimator to correct for stage 1 bias and improve sample size estimates for stage 2. Zöllner and Pritchard [2007] used computer simulation to evaluate the magnitude of the winner's curse effect in case-control studies and proposed a maximum likelihood method to correct for it. Their method estimates the frequencies of all genotypes and corresponding penetrance parameters based on a known population prevalence of the disease under different inheritance models. Garner [2007] studied the source of the upward bias in the odds ratio estimate in genome-wide association studies, but did not propose a method to correct for it. Zhong and Prentice [2008] and Ghosh et al. [2008] recently proposed conditional-likelihood-based methods for point and interval estimation of the (logarithm of the) odds ratio in the context of logistic regression analysis of case-control status using genotype categories as a covariate.

In this paper, we take a direct approach to evaluate and correct for the effect of winner's curse in the context of case-control genetic association studies. In contrast to previous simulation-based evaluations, we calculate analytically the impact of the winner's curse on estimates of the allele frequency difference between cases and controls

and the corresponding odds ratios as a function of sample size, allele frequencies, and statistical significance level. We then describe a simple ascertainment-corrected maximum likelihood method to estimate the risk allele frequency difference and odds ratio. Our method is most similar to that of Zöllner and Pritchard [2007], but in contrast to their method, ours estimates directly the allele frequency difference or odds ratio, instead of estimating the penetrance parameters. We compare the performance (bias, standard error, and mean square error (MSE)) of our ascertainment-corrected maximum likelihood estimators (MLEs) to that of the naïve, uncorrected estimators. We extend these calculations to two-stage association studies, in which all markers are genotyped on a set of individuals in Stage 1, and the most promising markers are followed up by genotyping a second set of individuals in Stage 2.

Consistent with Zöllner and Pritchard [2007], we find that (1) the factors that result in overestimation of the allele frequency difference can be summarized by study power, independent of sample size and allele frequency, and that overestimation decreases as power increases; and (2) compared to the uncorrected estimator of the allele frequency difference, the ascertainment-corrected estimator results in reduced absolute bias when study power is low or moderate, and has comparable absolute bias when power is high. Further, we find that (3) for the logarithm of the odds ratio (ln OR), overestimation can again be summarized by study power, independent of sample size and allele frequency, and that overestimation decreases as power increases; (4) compared to the uncorrected estimator, the ascertainment-corrected MLE of the OR generally results in reduced bias and MSE, and (5) for reasonable two-stage designs [Skol et al., 2007], results mirror those for the corresponding one-stage designs. We recommend use of this

ascertainment-corrected maximum likelihood method for estimation of genetic effect size in large-scale genetic association studies.

## 2.2 Methods

### 2.2.1 One-stage design

***Model and assumptions***

We assume independent samples of N cases and N controls genotyped at an autosomal disease locus with alleles D and d. Let p and $p+\delta$ ($\delta \neq 0$) denote the frequency of the risk allele D in controls and cases, respectively. For a complex disease, we expect the genetic effect size to be small, so that Hardy-Weinberg equilibrium predictions provide a good approximation to the genotype frequencies in both controls and cases. Under this assumption, the counts $m_0$ and $m_1$ of the risk allele D in controls and cases follow independent binomial distributions on 2N trials with probabilities of success p and $p+\delta$, respectively.

Let X be the standard Pearson chi-square test statistic for association in a $2 \times 2$ table of allele counts in cases and controls. Under the assumption of Hardy-Weinberg equilibrium, X follows a chi-square distribution with one degree of freedom under the null hypothesis of no association ($\delta = 0$). We claim an association significant if X exceeds the critical value $x_\alpha$ at significance level $\alpha$.

***Uncorrected (naïve) maximum likelihood estimators (MLEs)***

In practice, investigators generally estimate the allele frequency difference between cases and controls by its MLE $\hat{\delta}_{un} = \frac{m_1}{2N} - \frac{m_0}{2N}$, or the corresponding odds ratio by $\hat{OR}_{un} = \frac{m_1(2N - m_0)}{m_0(2N - m_1)}$. We call these uncorrected MLEs "naïve" because they ignore the bias associated with focusing on genetic markers with statistically significant association results.

To model the impact of the winner's curse, we calculate the expected value of the uncorrected MLE $\hat{\delta}_{un}$ of the allele frequency difference δ conditional on obtaining significant evidence for association:

$$E(\hat{\delta}_{un} \mid X > x_\alpha) = \frac{\sum_{(m_0, m_1) \in I} \hat{\delta}_{un} P(m_0, m_1)}{\sum_{(m_0, m_1) \in I} P(m_0, m_1)} \qquad (1)$$

and from it the bias of the estimator as $E(\hat{\delta}_{un} \mid X > x_\alpha) - \delta$, and the proportional bias as $\frac{E(\hat{\delta}_{un} \mid X > x_\alpha) - \delta}{\delta}$. Here, $I = \{(m_0, m_1) : X(m_0, m_1) > x_\alpha\}$ is the set of allele count pairs that result in statistically significant evidence for association and

$$P(m_0, m_1) = \binom{2N}{m_0} p^{m_0} (1-p)^{2N-m_0} \binom{2N}{m_1} (p+\delta)^{m_1} (1-p-\delta)^{2N-m_1} \qquad (2)$$

Note that the denominator in (1) is the power to detect association if we genotype the disease SNP.

The standard error of the uncorrected MLE $\hat{\delta}_{un}$ can be calculated as:

$$SE(\hat{\delta}_{un} \mid X > x_\alpha) = \sqrt{E(\hat{\delta}_{un}{}^2 \mid X > x_\alpha) - \left(E(\hat{\delta}_{un} \mid X > x_\alpha)\right)^2} \qquad (3)$$

where $E(\hat{\delta}_{un}^{\,2} \mid X > x_\alpha)$ may be calculated by replacing $\hat{\delta}_{un}$ by $\hat{\delta}_{un}^{\,2}$ in (1).

We also calculate the absolute bias of $\hat{\delta}_{un}$ as:

$$E(\left|\hat{\delta}_{un} - \delta\right| \mid X > x_\alpha) = \frac{\sum_{(m_0,m_1)\in I} \left|\hat{\delta}_{un} - \delta\right| \times P(m_0, m_1)}{\sum_{(m_0,m_1)\in I} P(m_0, m_1)} \qquad (4)$$

Analogous formulae allow us to calculate the conditional bias, standard error, and absolute bias of the uncorrected MLE of the odds ratio OR, and from the expectation, the proportional bias of the logarithm of the estimator $\frac{E[\ln(\hat{O}R_{un}) \mid X > x_\alpha] - \ln(OR)}{\ln(OR)}$.

***Ascertainment-corrected MLEs***

The naive estimators ignore the fact that we typically are interested in estimates of the allele frequency difference δ and the odds ratio OR only if we have strong evidence for association. To address this, we propose an ascertainment-corrected maximum likelihood method that conditions on obtaining evidence for association. To this end, we calculate the conditional likelihood function

$$L(p, \delta \mid X > x_\alpha) = P(m_0, m_1 \mid X > x_\alpha) = 1\{X > x_\alpha \mid m_0, m_1, N\} \frac{P(m_0, m_1)}{\sum_{(m_0,m_1)\in I} P(m_0, m_1)} \qquad (5)$$

where the indicator function $1\{X > x_\alpha \mid m_0, m_1, N\}$ equals 1 or 0 depending on whether or not $X > x_\alpha$.

We maximize $L(p, \delta \mid X > x_\alpha)$ as a function of p and δ to obtain the ascertainment-corrected MLEs $\hat{p}_{as}$ and $\hat{\delta}_{as}$ by using the Nelder-Mead [1965] simplex method. We calculate the empirical standard errors of these estimators based on 1000

simulation replicates, and the asymptotic-theory standard errors by calculating the observed information matrix (see Appendix) evaluated at the parameter estimates:

$$I(\hat{p}_{as}, \hat{\delta}_{as}) = -\partial^2_{p,\delta} \log L(p, \delta \mid X > x_\alpha)\Big|_{\hat{p}_{as}, \hat{\delta}_{as}} \quad (6)$$

The covariance matrix for $\hat{p}_{as}$ and $\hat{\delta}_{as}$ can be approximated by $I^{-1}(\hat{p}_{as}, \hat{\delta}_{as})$. We take advantage of the invariance property of the MLE to calculate the ascertainment-corrected MLE for the odds ratio, and apply the delta method [Rao, 1965] to obtain its standard error. We calculate the mean square error (MSE) for the estimators by taking the sum of the variance and the squared bias of the estimator.

### 2.2.2 Two-stage design

***Model and assumptions***

We next consider two-stage association studies, in which $N_1$ cases and $N_1$ controls are genotyped for all markers, and only the most promising markers are genotyped in the second stage in an additional $N_2$ cases and $N_2$ controls. Let $p_i$ and $\delta_i$ be the risk allele frequencies in controls and the allele frequency difference between cases and controls in stage i. Given genetic homogeneity between stages 1 and 2, $p_1 = p_2 = p$ and $\delta_1 = \delta_2 = \delta$. At each stage, we calculate the association test statistic using the data only from that stage

$$Z_i = \frac{\hat{p}_{i1} - \hat{p}_{i0}}{\sqrt{\left[\hat{p}_{i0}(1-\hat{p}_{i0}) + \hat{p}_{i1}(1-\hat{p}_{i1})\right]/(2N_i)}} \quad (7)$$

where $\hat{p}_{i0}$ and $\hat{p}_{i1}$ are the naïve MLEs of the risk allele frequencies in controls and cases respectively at stage i, $\hat{p}_{ij} = \frac{m_{ij}}{2N_i} (i = 1, 2; j = 0, 1)$. Under null hypothesis of no disease-marker association ($\delta = 0$), the association test statistic $Z_i$ follows a standard normal distribution with mean 0 and variance 1.

We employ a joint analysis strategy for this two-stage study [Satagopan et al., 2002; Skol et al., 2006] by calculating

$$Z_{12} = \sqrt{\pi_{\text{sample}}} Z_1 + \sqrt{1 - \pi_{\text{sample}}} Z_2 \qquad (8)$$

where $\pi_{\text{sample}} = N_1/(N_1+N_2)$ is the proportion of individuals genotyped in Stage 1. We claim significant association when both $|Z_1|$ and $|Z_{12}|$ exceed the relevant critical values $C_1$ and $C_{12}$ in joint analysis. $C_1$ is calculated so that $P(|Z_1| > C_1) = \pi_{\text{marker}}$, where $\pi_{\text{marker}}$ is the proportion of markers to be genotyped in Stage 2, and $C_{12}$ by finding the threshold so that $P(|Z_1| > C_1, |Z_{12}| > C_{12}) = P(|Z_{12}| > C_{12} \,|\, |Z_1| > C_1) \times P(|Z_1| > C_1)$ results in the desired significance level [Skol et al., 2006].

***Uncorrected (naïve) MLEs***

The uncorrected MLE of the risk allele frequency difference for the two-stage design $\hat{\delta}_{12} = \pi_{sample}\hat{\delta}_1 + (1 - \pi_{sample})\hat{\delta}_2$, where $\hat{\delta}_i = \frac{m_{i1}}{2N_i} - \frac{m_{i0}}{2N_i}$, i = 1, 2. The bias of the uncorrected MLE $\hat{\delta}_{12}$ can be calculated exactly as for one-stage design by formula (1) and similarly the proportional bias. However, exact calculation becomes computationally difficult when $N_1$ or $N_2$ is large, so we simulated n=1000 datasets satisfying $|Z_1| > C_1$ and $|Z_{12}| > C_{12}$ and approximated the expectation and empirical standard error of

$\hat{\delta}_{12}$ by calculating the mean and the standard error of the uncorrected MLE of the n simulated datasets:

$$E(\hat{\delta}_{12} \| |Z_1| > C_1, |Z_{12}| > C_{12}) \approx \bar{\delta}_{12} = \frac{1}{n}\sum_{j=1}^{n}\left(\pi_{sample}\hat{\delta}_{1j} + (1-\pi_{sample})\hat{\delta}_{2j}\right) \quad (9)$$

$$SE(\hat{\delta}_{12} \| |Z_1| > C_1, |Z_{12}| > C_{12}) = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(\hat{\delta}_{12,j} - \bar{\delta}_{12}\right)^2} \quad (10)$$

***Ascertainment-corrected MLEs***

In analogy to the one-stage design, the two-stage ascertainment-corrected likelihood

$$L(p, \delta \,|\, |Z_1| > C_1, |Z_{12}| > C_{12}) = P(m \,|\, |Z_1| > C_1, |Z_{12}| > C_{12})$$

$$= \frac{1\{|Z_1| > C_1, |Z_{12}| > C_{12} \,|\, m, N_1, N_2\} P(m)}{P(|Z_1| > C_1, |Z_{12}| > C_{12})} \quad (11)$$

Here, $m = (m_{10}, m_{11}, m_{20}, m_{21})$ , $1(|Z_1| > C_1, |Z_{12}| > C_{12} \,|\, m, N_1, N_2)$ is an indicator function taking values of 1 or 0 depending on whether or not $|Z_1| > C_1$ and $|Z_{12}| > C_{12}$, and *P(m)* is the product of four binomial probabilities. The denominator of (8) is again the power of the study, and can be evaluated as described by Skol et al. [2006]. We maximize the likelihood (8) to get MLEs of p and δ by using the Nelder-Mead simplex approach, obtain empirical standard errors based on 1000 simulation replicates.

## 2.3 Results

### 2.3.1 One-stage design

***Bias of the uncorrected MLE of the allele frequency difference δ and the odds ratio OR***

For a locus showing association (δ≠0), our analytical calculation demonstrates upward bias in the genetic effect size by the naïve estimator $\hat{\delta}_{un}$ of the allele frequency difference δ (Figure 2.1). This bias is particularly severe when power is low, owing to small sample size N and/or small allele frequency difference δ (Table 2.1, Figure 2.2A). As power approaches one, the bias disappears. Under the null hypothesis (δ = 0), $\hat{\delta}_{un}$ is unbiased, since δ is equally likely to be over- or under-estimated. However, the absolute bias of this uncorrected estimator is extremely high when δ = 0 or when δ is small (Figure 2.1). Due to symmetry, for the rest of the tables or figures, we only provide results for δ > 0 (lnOR > 0).

Given N = 1000 cases and N = 1000 controls, allele frequencies p = .1 and p+δ = .1258 (OR = 1.295), and testing at significance level of $\alpha = 10^{-6}$ (resulting in power = .01), the expected value of the uncorrected estimator of δ is .0524 compared to the true value of .0258, a bias of .0266 and a proportional bias of 103%; similarly, the expected value of the uncorrected OR estimator is 1.699 compared to its true value of 1.295. In this case, a follow-up study designed to have 80% power at significance level α = .05 would include 310 cases and 310 controls, but would have actual power of only 30%.

We found that, for a fixed significance level α, the proportional bias in the uncorrected estimate of δ is solely a function of power, and is otherwise independent of sample size, allele frequency, or genetic model [Zöllner and Pritchard, 2007]. Consistent

with intuition, proportional bias decreases as power increases (Figure 2.2A), since the conditioning event becomes increasingly likely. At significance level $\alpha = 10^{-6}$, the uncorrected estimator of δ gives a proportional bias of ~60% when power is .05 but is nearly unbiased when power is 95%. Interestingly, given fixed power, the proportional bias of the naïve estimator is consistently less when $\alpha = 10^{-6}$ than when $\alpha = 10^{-4}$.

We extended our analytical calculation to the uncorrected MLE of the odds ratio (Table 2.1, Figure 2.1), and observed the same general trend: substantial overestimation of the genetic effect given low to modest power to detect association and no bias given no association or sufficiently strong association. However, the proportional bias of the OR estimator, $\hat{OR}_{un}$, cannot be explained by power alone, but depends on sample size, allele frequency, and genetic model (Figure 2.2B). Interestingly, the proportional bias of the logarithm of the OR estimator, $\log \hat{OR}_{un}$, is a function of power, and follows a very similar pattern as the uncorrected MLE of allele frequency δ.

***Bias of the ascertainment-corrected MLE of δ and OR***

When we correct for ascertainment, the absolute bias of the MLE is substantially reduced (Figure 2.1, Table 2.1), and correction actually results in underestimation unless the genetic effect size is small or power is very low. For example, given N = 1000 cases and N = 1000 controls, allele frequencies p = .1 and p+δ = .1258 (power = .01), and testing at significance level $\alpha = 10^{-6}$, the proportional bias of the corrected MLE of δ is –7%, compared to +103% before correction. In this case, a follow-up study designed to have 80% power at significance level α = .05 would include 1350 cases and 1350 controls and have actual power 85%, whereas 1150 cases and 1150 controls actually

would be sufficient to achieve 80% power. In the absence of association (δ=0), the corrected MLE is again nearly unbiased.

Reduction of the absolute bias is most pronounced when overestimation is most severe, and for fixed significance level α, bias reduction depends solely on study power. The relationship between power and proportional bias of the ascertainment-corrected MLE of δ is summarized in Figure 2.2A. Although the corrected MLE $\hat{\delta}_{as}$ typically underestimates δ by 10-20% over the power range of .001-.95 given testing at significance level of $\alpha = 10^{-6}$, the corrected MLE is considerably less biased than the uncorrected estimator unless power is high (typically > 60%). Even given high power, the magnitude of the bias of the ascertainment-corrected MLE $\hat{\delta}_{as}$ is not much greater than that of the uncorrected MLE $\hat{\delta}_{un}$, and it is of opposite sign. Interestingly, when power greater than .1, the bias in the corrected MLE $\hat{\delta}_{as}$ decreases almost linearly as power increases (Figure 2.2A).

The situation for the odds ratio is similar. With correction, the OR is typically underestimated by 5-10%, and this bias is in general smaller (although of opposite sign) than that for the uncorrected estimator for study powers ranging from .001 to .95 (Table 2.1, Figures 2.1 and 2.2B). Compared to the corrected MLE of δ whose proportional bias can be approximately summarized by power alone, the proportional bias for the corrected OR estimator does depends on sample size and allele frequency (Figure 2.2B), while the proportional bias of the logarithm of the corrected estimator depends essentially on power alone and displays a very similar pattern as that of the corrected estimator for δ (Figure 2.2A). Again, if we focus on the situations in which power < 60%, correction generally results in reduced absolute bias, and in many cases, absolute bias reduction is impressive.

For example, given N = 1000 cases and N = 1000 controls, allele frequencies p = .1 and p+δ = .1258 (OR = 1.295), and testing at significance level $\alpha = 10^{-6}$ (resulting in power = .01), the proportional bias of the corrected MLE of OR is –2%, compared to +31% before correction.

***Standard errors and mean square errors (MSE) of the estimators***

Table 2.2 summarizes the standard errors (SEs) for the MLEs of δ. We observed that the empirical SEs agree well with the asymptotic SEs for the corrected MLE, and both are two to six times greater than the SE of the uncorrected MLE which incorrectly ignores the fact of ascertainment. We also calculated the SE based on a random sample of the same sample size without ascertainment. All calculated SEs demonstrate that the genetic effect size estimates are quite variable in the settings described. The SEs of the corrected MLE are typically 1.5-2 times as large as those for an unascertained independent sample of the same size. This implies that while the ascertained sample is not as informative as a new random sample would be to estimate genetic effect size, the ascertained sample does provide 50-60% of the information in a new random sample, without the extra cost of collecting a new sample. We observed a very similar trend for SEs for the MLE of the odds ratio.

The mean squared error (MSE) provides a measure of estimator quality that takes into account both bias and variance. Figure 2.1 displays the MSE for the naïve and corrected MLEs of δ and lnOR. In general, the naïve estimator has larger MSE than the ascertainment-corrected estimator unless the genetic effect size is sufficiently large to result in high power to detect association. In that case, biases for the two estimators are

similar but the variance of the corrected estimator is larger than that of the naïve estimator (Table 2.2).

### 2.3.2 Two-stage design

For both the allele frequency difference δ and the odds ratio OR, the naïve and ascertainment-corrected MLEs for optimal two-stage designs yield very similar results to those for the one-stage association designs described above (Figure 2.3A). This is hardly surprising, since for optimal two-stage designs, statistical power is very close to that of the corresponding one-stage design in which all markers are genotyped on all samples, and power (approximately) determines proportional bias for δ and lnOR. Even for non-optimal two-stage designs, this continues to be true, except that the proportional bias of both the uncorrected and corrected estimators tends to increase modestly as $\pi_{sample,}$ the fraction of the sample genotyped in Stage 1, increases (Figure 2.3B).

## 2.4 Discussion

In genetic association studies, the genetic effect size for associated markers tends to be overestimated as a consequence of winner's curse. This bias is due to the strong positive correlation between the association test statistic and the estimator of the genetic effect and the focus of investigators on markers that show statistically significant evidence of association. In this paper, we studied the bias of the naïve maximum likelihood estimators for the allele frequency difference and the odds ratio that ignore this ascertainment; these measures are routinely used to estimate the strength of the effect in

genetic association studies. We demonstrated that the proportional bias in the estimators decreases as power increases. Interestingly, at fixed significance level, the proportional biases of the allele frequency difference and the logarithm of odds ratio are functions of power, and otherwise are essentially independent of allele frequency or sample size (see also [Zöllner and Pritchard, 2007]).

We proposed a maximum likelihood method to correct for this ascertainment bias. The ascertainment-corrected MLEs for both the allele frequency difference and the (log) odds ratio are generally less biased than the uncorrected estimators unless study power is moderate to high (>60%). Since large-scale genetic association studies of complex traits typically are underpowered owing to small genetic effect sizes, our method should generally provide a more accurate estimate of genetic effect size in the context of genome-wide association studies and large-scale candidate gene studies. In high power situations, bias for both the naïve and corrected methods are small, so that ascertainment correction again is reasonable. Proportional bias of the corrected and uncorrected estimators for both the allele frequency difference and the odds ratio does show modest dependence on significance level $\alpha$. For example, when significance level $\alpha = 10^{-4}$, biases for all estimators are somewhat increased compared to the case of $\alpha = 10^{-6}$, and the advantage of ascertainment correction is increased slightly.

Zöllner and Pritchard [2007] used simulations to evaluate the impact of the winner's curse effect in genetic association studies and also proposed a maximum likelihood method to correct for it. Their method estimates the frequencies of all genotypes and corresponding penetrance parameters based on a known population prevalence of the disease under different inheritance models. In contrast, our method is

simpler and focuses solely on the parameters of greatest interest: the allele frequency difference and odds ratio. This advantage of our method does require the assumption of Hardy-Weinberg Equilibrium for our case and control samples. Such an assumption is entirely reasonable given the modest locus effect sizes for complex traits, but would not be reasonable in the context of a Mendelian major locus.

Our corrected MLEs for the allele frequency difference and odds ratio generally underestimate the true genetic effects [Zöllner and Pritchard, 2007]. Using computer simulation, we note that the empirical distribution of our corrected MLEs can reasonably be described as a two-component mixture, with one component near zero and the other appearing more nearly normal. Figure 2.4 illustrates this for the ascertainment-corrected estimator of the allele frequency difference. As power increases, the distribution becomes more nearly normal, and the asymptotic unbiasedness of the MLE comes into play.

We investigated the coverage of the asymptotic theory 95% confidence interval for the naïve and ascertainment-corrected MLEs for the allele frequency difference $\delta$. The coverage of the ascertainment-corrected interval ranged from 82-100% for the cases we considered, reflecting the distribution and the bias of the ascertainment-corrected MLE, but still generally better than the coverage for the naïve estimator, which ranged from 0-92%.

Given the usual downward bias of our ascertainment-corrected estimators, one could consider an *ad hoc* bias correction. For the estimators of the allele frequency difference $\delta$ and the log odds ratio lnOR, the downward bias is 5-20% across the situations we considered (control allele frequency .1-.5, allele frequency difference $\delta$=.018-.159 (OR 1.11-2.30), case and control sample sizes 250 to 2,000, and statistical

significance $10^{-4}$ to $10^{-8}$), so that multiplying the resulting estimate by 1.05 – 1.10 would generally reduce absolute bias. However, such an approach is counterproductive when power is very low (<.005). The same criticism holds for taking a (weighted) average of the corrected and uncorrected estimators. More appealing might be to use an alternative estimation approach, and we currently are considering an empirical Bayes method [Carlin and Louis, 2000] that uses information from genome-wide association studies to help define a prior distribution for the genetic effect size.

Realistically, precise and unbiased estimation of genetic effect size will best be obtained by collecting a large sample specifically for this purpose, should resources be available to do so. However, given a sample in which an association is discovered, our ascertainment corrected approach provides more accurate estimation of allele frequency difference and odds ratio than the naïve approach, and permits better design of subsequent replication studies or studies focused on estimating the population effect of the identified variant(s). Standard errors for the ascertainment-corrected MLEs were substantially larger than those for the naïve estimator based on an independent random sample of the same size, correctly reflecting the information loss for estimation based on a sample used for association detection.

In summary, we have presented analytic calculations that quantify the impact of the winner's curse in large-scale genetic association studies, and confirm that in realistic situations, it can result in substantial overestimation of the true genetic effect as measured by the case-control allele frequency difference or the corresponding odds ratio. We propose a maximum likelihood estimator that corrects for the typical focus on statistically significant results, and demonstrate that this estimator results in reduced absolute bias

compared to the naïve uncorrected estimator when study power is low or moderate (<60%), a range that is typical for most large-scale genetic association studies, and similar absolute bias when power is high. Our method does not require specification of a genetic model and is easy to implement. We extended these calculations to two-stage association studies, and found similar results to those for one-stage studies. We recommend the use of this ascertainment-corrected method for estimation of genetic effect size in large-scale genetic association studies.

Software that carries out this analysis for case-controls data is available at http://csg.sph.umich.edu/boehnke/winner.

Table 2.1: Proportional bias (%) of the uncorrected (naïve) and ascertainment-corrected MLEs of the allele frequency difference δ and odds ratio OR. Results are presented only for $\delta > 0$.

| p | N | power | δ | $\frac{\hat{\delta}_{un} - \delta}{\delta}$ | $\frac{\hat{\delta}_{as} - \delta}{\delta}$ | OR | $\frac{\hat{OR}_{un} - OR}{OR}$ | $\frac{\hat{OR}_{as} - OR}{OR}$ |
|---|---|---|---|---|---|---|---|---|
| .1 | 500 | .01 | .0376 | 101.9 | -4.3 | 1.436 | 53.3 | -1.8 |
| | | .10 | .0541 | 47.9 | -12.4 | 1.640 | 27.9 | -5.7 |
| | | .30 | .0665 | 26.2 | -16.1 | 1.798 | 17.6 | -7.2 |
| | | .50 | .0752 | 16.2 | -14.2 | 1.913 | 11.9 | -6.5 |
| | | .80 | .0898 | 6.0 | -8.2 | 2.108 | 5.55 | -1.7 |
| | 1000 | .01 | .0258 | 103.1 | -7.0 | 1.295 | 31.2 | -1.9 |
| | | .10 | .0370 | 48.1 | -15.7 | 1.429 | 19.2 | -4.2 |
| | | .30 | .0453 | 26.5 | -18.3 | 1.530 | 12.3 | -7.1 |
| | | .50 | .0512 | 16.0 | -16.2 | 1.603 | 8.36 | -6.3 |
| | | .80 | .0609 | 6.1 | -9.3 | 1.726 | 3.77 | -4.1 |
| .5 | 500 | .01 | .0576 | 103.0 | -9.0 | 1.260 | 27.2 | -2.1 |
| | | .10 | .0806 | 48.3 | -16.8 | 1.384 | 17.3 | -5.3 |
| | | .30 | .0972 | 26.3 | -18.3 | 1.483 | 11.2 | -6.3 |
| | | .50 | .1086 | 16.2 | -14.3 | 1.555 | 7.72 | -5.8 |
| | | .80 | .1270 | 6.1 | -9.2 | 1.681 | 3.51 | -3.6 |
| | 1000 | .01 | .0405 | 104.0 | -13.1 | 1.176 | 18.5 | -2.0 |
| | | .10 | .0571 | 48.3 | -18.6 | 1.258 | 11.8 | -4.2 |
| | | .30 | .0690 | 26.4 | -18.3 | 1.320 | 7.7 | -4.9 |
| | | .50 | .0772 | 16.2 | -15.0 | 1.365 | 5.4 | -4.5 |
| | | .80 | .0903 | 6.1 | -10.4 | 1.441 | 2.4 | -3.1 |

un: uncorrected; as: ascertainment-corrected

p: disease allele frequency in controls; N: sample size (number of cases and of controls)

Assume testing at significance level $\alpha = 10^{-6}$

Table 2.2: Standard errors (SEs) for the uncorrected (naïve) and ascertainment-corrected MLEs of the allele frequency difference δ and for MLE obtained from an unascertained random sample. Results are presented only for $\delta > 0$.

| p | N | power | OR | δ | SE | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | $\hat{\delta}_{un}$ | $\hat{\delta}^{*}_{as}$ | $\hat{\delta}^{\dagger}_{as}$ | $\hat{\delta}_{rand}$ |
| .1 | 500 | .01 | 1.436 | .0376 | .0049 | .0263 | .0307 | .0142 |
| | | .10 | 1.640 | .0541 | .0064 | .0291 | .0315 | .0150 |
| | | .30 | 1.798 | .0665 | .0080 | .0304 | .0307 | .0148 |
| | | .50 | 1.913 | .0752 | .0094 | .0309 | .0291 | .0153 |
| | | .80 | 2.108 | .0898 | .0120 | .0291 | .0244 | .0154 |
| | 1000 | .01 | 1.295 | .0258 | .0032 | .0179 | .0216 | .0099 |
| | | .10 | 1.429 | .0370 | .0043 | .0200 | .0212 | .0103 |
| | | .30 | 1.530 | .0453 | .0054 | .0216 | .0204 | .0099 |
| | | .50 | 1.603 | .0512 | .0063 | .0218 | .0195 | .0107 |
| | | .80 | 1.726 | .0609 | .0081 | .0195 | .0170 | .0102 |
| .5 | 500 | .01 | 1.260 | .0576 | .0069 | .0392 | .0433 | .0215 |
| | | .10 | 1.384 | .0806 | .0091 | .0442 | .0460 | .0214 |
| | | .30 | 1.483 | .0972 | .0114 | .0447 | .0439 | .0229 |
| | | .50 | 1.555 | .1086 | .0133 | .0445 | .0409 | .0221 |
| | | .80 | 1.681 | .1270 | .0168 | .0411 | .0348 | .0222 |
| | 1000 | .01 | 1.176 | .0405 | .0049 | .0280 | .0319 | .0159 |
| | | .10 | 1.258 | .0571 | .0065 | .0310 | .0320 | .0154 |
| | | .30 | 1.320 | .0690 | .0081 | .0325 | .0305 | .0154 |
| | | .50 | 1.365 | .0772 | .0095 | .0325 | .0286 | .0160 |
| | | .80 | 1.441 | .0903 | .0120 | .0281 | .0247 | .0159 |

un: uncorrected; as: ascertainment-corrected; rand: random sample without ascertainment

*: empirical; †: asymptotic

p: disease allele frequency in controls; N: sample size (number of cases and of controls)

Assume testing at significance level $\alpha = 10^{-6}$

Figure 2.1: Bias, absolute bias, and mean square error (MSE) for allele frequency difference δ and logarithm of odds ratio lnOR with sample size N = 1000 and control allele frequency p = .3. Significance level $\alpha = 10^{-6}$.

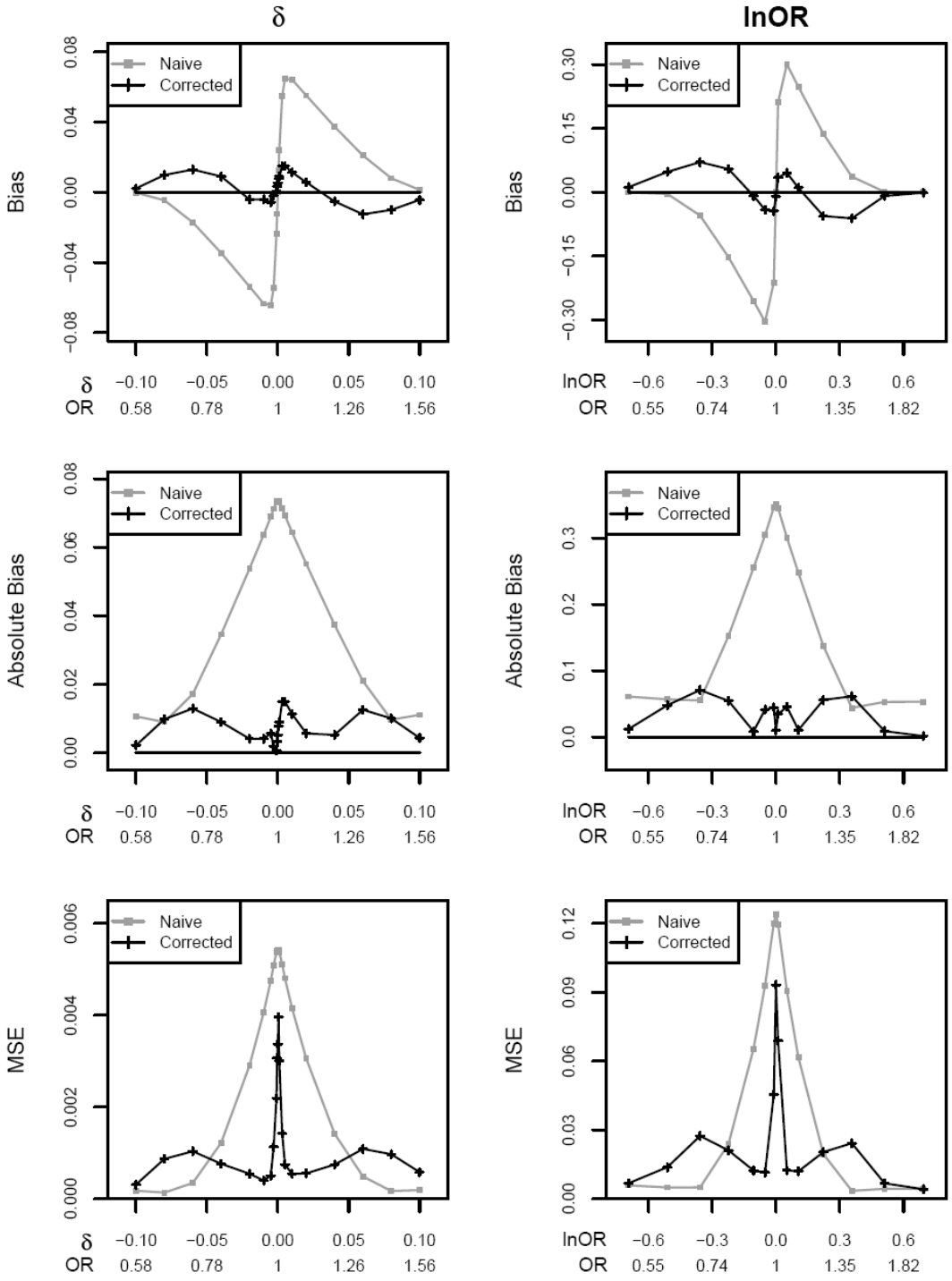
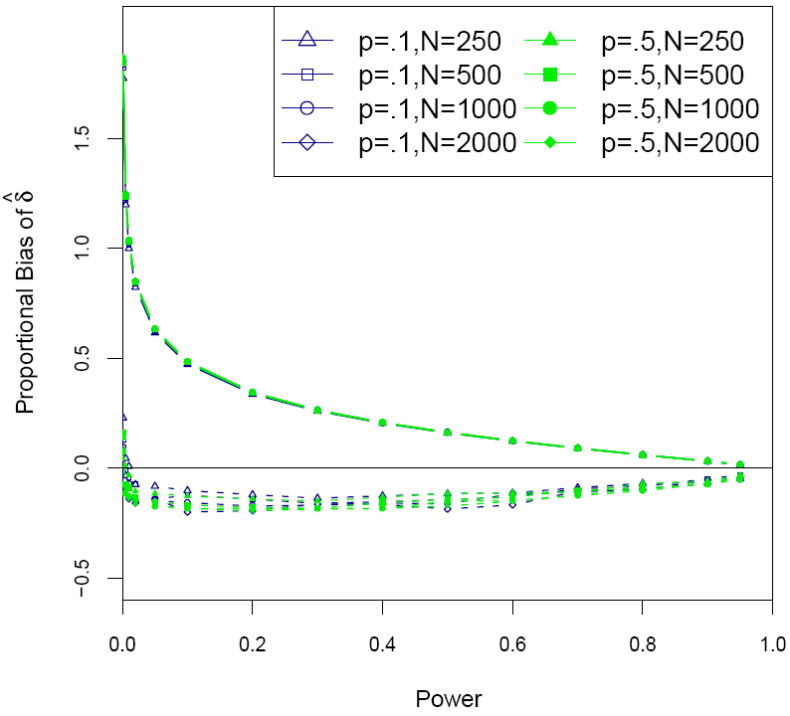


Figure 2.2: Proportional bias versus power for the uncorrected (naïve) (solid lines) and corrected (dashed lines) estimators of the (A) allele frequency difference δ and (B) odds ratio OR. Significance level $\alpha = 10^{-6}$. Results are presented only for $\delta > 0$.

A.




B.

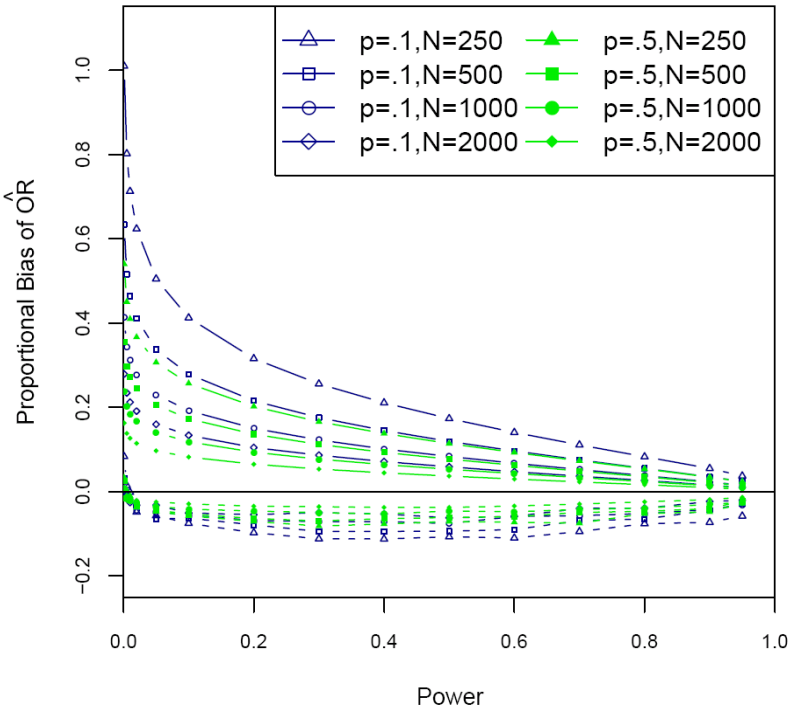
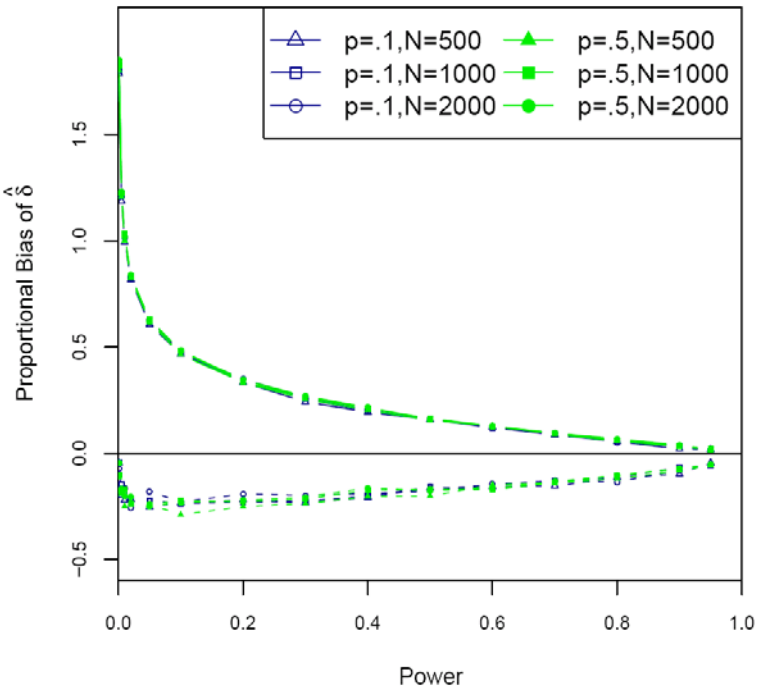


Figure 2.3: Proportional bias versus power for the uncorrected (naïve) (solid lines) and corrected (dashed lines) estimators of the allele frequency difference δ for (A) optimal and (B) non-optimal two-stage designs. Significance level $\alpha = 10^{-6}$. Designs optimal for multiplicative disease model with disease prevalence .10, stage 2 to stage 1 genotype cost ratio 30. For non-optimal designs, $\pi_{marker}$ = 1%, and samples of N=1000 cases and N=1000 controls. Results are presented only for $\delta > 0$.

A.




B.

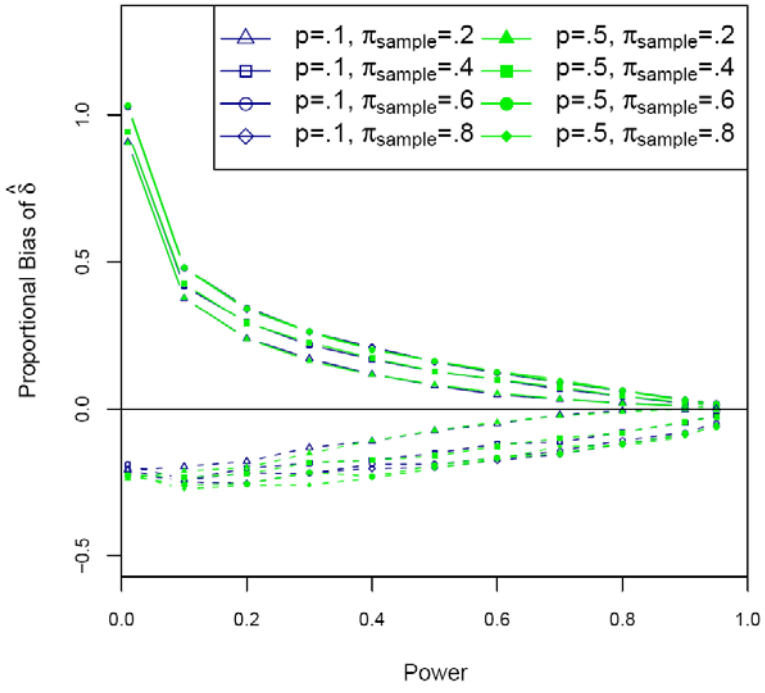
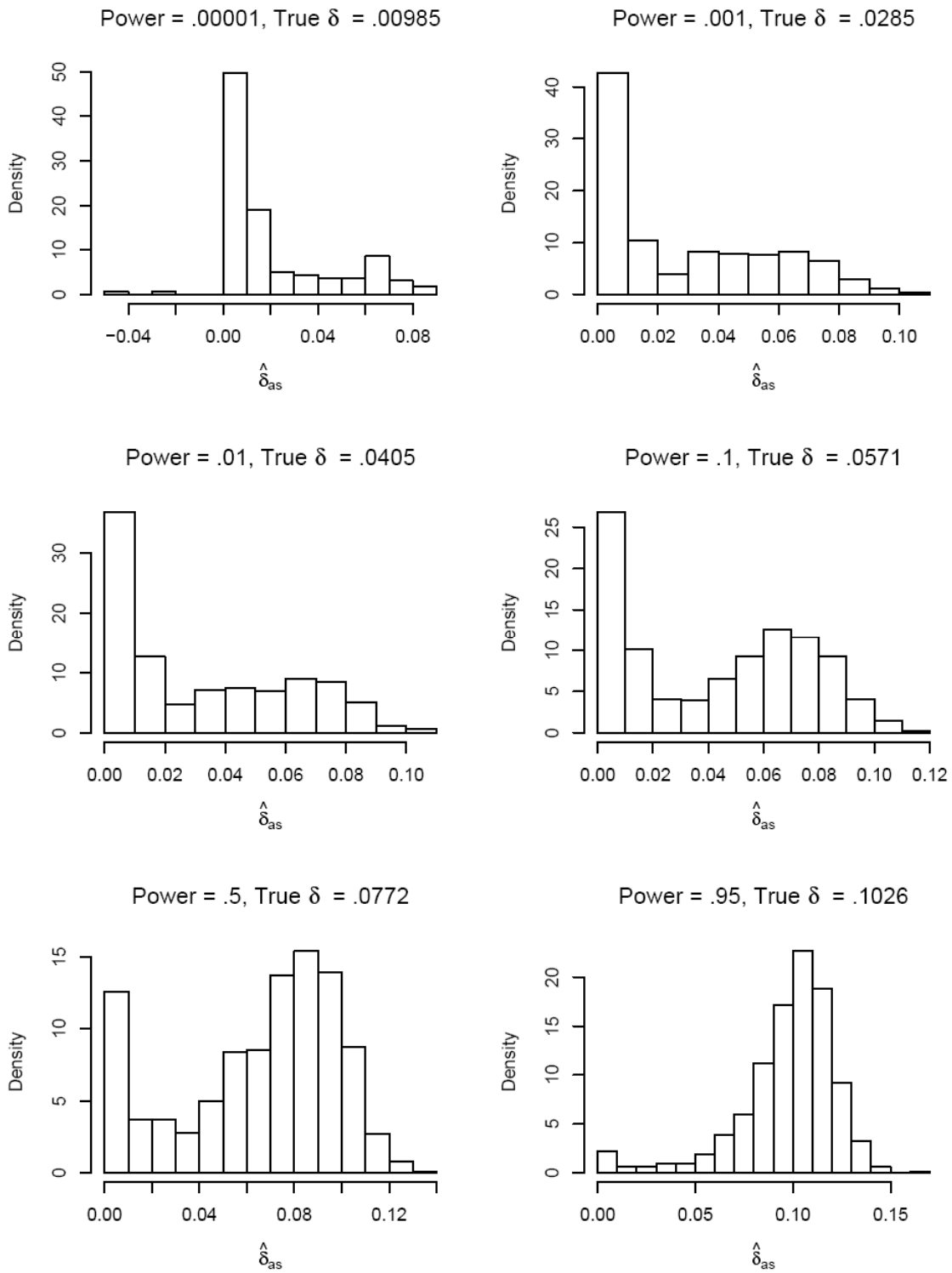


Figure 2.4: Distribution of the ascertainment-corrected MLE of the allele frequency difference δ for different power levels. Results are presented only for $\delta > 0$.




Based on 1000 simulation replicates of N=1000 cases and N=1000 controls, control allele frequency p = .5, and testing at significance level $\alpha = 10^{-6}$.

# Appendix

Calculate the observed information matrix $I$ for one-stage study:

$I(p,\delta)=\begin{pmatrix} a & b \\ b & d \end{pmatrix}$, where

$$a=-\frac{m_0}{p^2}-\frac{2N-m_0}{(1-p)^2}-\frac{m_1}{(p+\delta)^2}-\frac{2N-m_1}{(1-p-\delta)^2}+\frac{A^2}{F^2}+\frac{B}{F}$$

$$b=-\frac{m_1}{(p+\delta)^2}-\frac{2N-m_1}{(1-p-\delta)^2}+\frac{AD}{F^2}+\frac{C}{F}$$

$$d=-\frac{m_1}{(p+\delta)^2}-\frac{2N-m_1}{(1-p-\delta)^2}+\frac{D^2}{F^2}+\frac{E}{F}$$

where $A,\ B,\ C,\ D,\ E$ and $F$ are calculated as follows:

$$A=\sum_{(x_0,x_1)\in I}\left\{\left(\frac{x_0}{p}-\frac{2N-x_0}{1-p}+\frac{x_1}{p+\delta}-\frac{2N-x_1}{1-p-\delta}\right)P(x_0,x_1)\right\}$$

$$B=\sum_{(x_0,x_1)\in I}\left\{\left(\frac{x_0}{p^2}+\frac{2N-x_0}{(1-p)^2}+\frac{x_1}{(p+\delta)^2}+\frac{2N-x_1}{(1-p-\delta)^2}-\left(\frac{x_0}{p}-\frac{2N-x_0}{1-p}+\frac{x_1}{p+\delta}-\frac{2N-x_1}{1-p-\delta}\right)^2\right)P(x_0,x_1)\right\}$$

$$C=\sum_{(x_0,x_1)\in I}\left\{\left(\frac{x_1}{(p+\delta)^2}+\frac{2N-x_1}{(1-p-\delta)^2}-\left(\frac{x_0}{p}-\frac{2N-x_0}{1-p}\right)\times\left(\frac{x_0}{p}-\frac{2N-x_0}{1-p}+\frac{x_1}{p+\delta}-\frac{2N-x_1}{1-p-\delta}\right)\right)P(x_0,x_1)\right\}$$

$$D=\sum_{(x_0,x_1)\in I}\left\{\left(\frac{x_1}{p+\delta}-\frac{2N-x_1}{1-p-\delta}\right)P(x_0,x_1)\right\}$$

$$E=\sum_{(x_0,x_1)\in I}\left\{\left(\frac{x_1}{(p+\delta)^2}+\frac{2N-x_1}{(1-p-\delta)^2}-\left(\frac{x_1}{p+\delta}-\frac{2N-x_1}{1-p-\delta}\right)^2\right)P(x_0,x_1)\right\}$$

$$F=\sum_{(x_0,x_1)\in I}P(x_0,x_1)$$

where $I=\{(x_0,x_1):X(x_0,x_1)>x_\alpha\}$ and $P(x_0,x_1)$ is calculated by formula (2) in the paper.

Our calculation for the asymptotic SE for p and δ was based on the observed information matrix evaluated at $\hat{p}_{as},\hat{\delta}_{as}$.

## References


Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM, Amos CI. 2002. Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. Am J Hum Genet 70: 575-585.

Bazerman, MH, Samuelson WF. 1983. I won the auction but don't want the prize. J Conflict Resolut 27: 618-634.

Carlin BP, Louis TA. 2000. Bayes and empirical Bayes methods for data analysis, 2nd ed. CRC Press, Boca Raton, FL.

Garner C. 2007. Upward bias in odds ratio estimates from genome-wide association studies. Genet Epidemiol 31: 288-295.

Ghosh A, Zou F, Wright FA. 2008. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. Am J Hum Genet 82: 1064-1074.

Göring HHH, Terwilliger JD, Blangero J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69: 1357-1369.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001. Replication validity of genetic association studies. Nat Genet 29: 306-309.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33: 177-182.

Nelder JA, Mead R. 1965. A simplex method for function minimization. Computer J 7: 308-313.

Rao CR. 1965. Linear statistical inference and its applications. New York: John Wiley.

Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. 2002. Two-stage design for gene-disease association studies. Biometrics 58: 163-170.

Siegmund D. 2002. Upward bias in estimation of genetic effects. Am J Hum Genet 71: 1183-1188.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38: 209-213.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2007. Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31: 776-788.

Sun L, Bull SB. 2005. Reduction of selection bias in genome-wide studies by resampling. Genet Epidemiol 28: 352-367.

Wu LY, Sun L, Bull SB. 2006. Locus-specific heritability estimation via bootstrap in linkage scans for quantitative trait loci. Hum Hered 62: 84-96.

Yu K, Chatterjee N, Wheeler W, Li Q, Wang S, Rothman N, Wacholder S. 2007. Flexible design for following up positive findings. Am J Hum Genet 81: 540-551.

Zhong H, Prentice RL. 2008. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics 94(4): 621-634.

Zöllner S, Pritchard JK. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80: 605-615.

# CHAPTER 3

# WINNER'S CURSE IN QUANTITATIVE TRAIT ASSOCIATION STUDIES

Quantitative traits (QT) are an important focus of human genetic studies both because of interest in the traits themselves, and because of their role as risk factors for many human diseases. For large-scale QT association studies including genome-wide association studies (GWAS), investigators usually focus on genetic loci showing significant evidence for SNP-QT association, and consistent with findings for case-control association studies of diseases, QT genetic effect size also tends to be overestimated as a consequence of the winner's curse. In this chapter, I study the impact of the winner's curse on QT association studies, in which the genetic effect size is parameterized as the slope in a linear regression model. Analytical calculation demonstrates that the overestimation in the regression slope estimate decreases as power increases. To reduce the ascertainment bias, I propose a three-parameter maximum likelihood method and further simplify this model to a one-parameter model by excluding nuisance parameters. I show that both methods reduce the bias when power to detect association is low or moderate, and that the one-parameter model generally results in smaller variance in the estimate.

## 3.1 Introduction

In Chapter 2, I quantified the winner's curse effect in one- and two-stage genetic case-control association studies, and described a maximum likelihood method to correct for the ascertainment bias in genetic effect size (parameterized as allele frequency difference and odds ratio) estimation. I showed that the upward bias is particularly severe when study power is low, and that the ascertainment-corrected MLE reduces bias when power is low or moderate, as expected for new gene discoveries in GWAS. In this chapter, I extend winner's curse study to quantitative trait (QT) association studies.

For complex disease genetics research in humans, remarkable progress has been made recently with a number of genome-wide case-control association studies published [Klein et al., 2005; Maraganore et al., 2005; Sladek et al., 2007; Scott et al., 2007; Saxena et al., 2007; Zeggini et al., 2007]. In parallel, there have been increasing efforts to investigate the association between genotype and disease-related QT at population level [Frayling et al., 2007; Saxena et al., 2007; Willer et al., 2008; Levy et al., 2009]. One rationale behind QT studies is that, because the traits examined are in many cases closely related to disease, any quantitative trait locus (QTL) being identified may also act as a disease predisposing locus.

Traditionally, linkage analysis, in which families with multiple affected individuals are scanned, has played an important role in mapping QTL. However, this analysis has the weakness of identifying relatively large chromosomal intervals associated with particular phenotypes. For instance, the initial localization from linkage analysis may define a region of 10-20 Mb. In contrast, association studies are useful only for short-range mapping because association relies on either the presence of linkage

disequilibrium (LD) between the SNP and trait loci, or the SNP being the trait locus itself. Association studies generally have greater power to detect alleles with minor or modest phenotypic effects [Risch and Merikangas, 1996; Sham et al., 2002].

For QT association studies, a commonly used method to detect the SNP-trait association is to regress the observed trait values on a score based on the individual's SNP genotype. The slope of the linear regression is a measure of the strength of the genetic effect. As in disease case-control association studies, for QT association studies, investigators usually focus on genetic loci showing significant evidence for SNP-trait association. As a consequence of the winner's curse described in Chapter 2, the effect size estimator tends to overestimate the true genetic effect size. Several investigators have studied the winner's curse effect in the context of QT linkage analysis [Göring et al., 2001; Siegmund, 2002; Allison et al., 2003; Sun and Bull, 2005; Wu et al., 2006].

In this chapter, I study the winner's curse effect in the context of QT association studies. I analytically quantify the impact of the winner's curse on the estimate of the genetic effect size parameterized as the linear regression slope as a function of sample size, allele frequency, and statistical significance level. I then describe an ascertainment-corrected maximum likelihood method similar to that we derived for case-control disease association studies [Xiao and Boehnke, 2009; Chapter 2] to correct for this bias. I describe both a fully parameterized model in which we estimate the intercept, slope, and error of the linear regression model, and a simplified which focuses only on the regression equation slope parameter resulting in a one-parameter model. I also consider a mean square error (MSE) weighted estimator calculated as the weighted average of the uncorrected and corrected estimators using the MSE as the weight. I compare the

performance (bias, standard error, MSE) of these ascertainment-corrected maximum likelihood estimators (MLEs) and that of the naïve, uncorrected estimators.

As for genetic case-control studies (Chapter 2), I find that (1) the factors that result in overestimation of the regression slope can be summarized by study power alone, independent of sample size and allele frequency, and that overestimation decreases as power increases; (2) compared to the uncorrected estimator of the regression slope, the ascertainment-corrected estimators based on the one- and three-parameter model result in reduced absolute bias when study power is low or moderate, and have comparable absolute bias when power is high; (3) the MSE of the ascertainment-corrected MLE of the regression slope based on the one-parameter model is generally smaller than that of the MLE based on the three-parameter model; and it is also smaller than the uncorrected estimator when power is low or moderate; and (4) the MSE weighted estimator generally improves the ascertainment correction compared to the three- and one-parameter-model based ascertainment-corrected MLEs. I recommend the use of the one-parameter-based ascertainment-corrected maximum likelihood method and the MSE weighted estimator for estimation of genetic effect size in large-scale quantitative trait association studies.

## 3.2 Methods

### 3.2.1 Models and Assumptions

I assume $N$ independent samples genotyped at an autosomal quantitative trait locus (QTL) with alleles A and a. Let p be the frequency for the minor allele a. For individual i, let $X_i$ be the score for allele a, depending on the genetic model we assume. For example, $X_i = 0$ for AA, 1 for Aa, and 2 for aa if we assume an additive model or $X_i =$

0 for AA or Aa, and 1 for aa for a recessive minor allele. Let $y$ be the $N\times1$ vector of trait values for the $N$ samples.

To test for SNP-trait association, I assume the linear regression model: $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $\{\varepsilon_i\}$ are independently and identically distributed (iid) as normal with mean 0 and variance $\sigma^2$. For simplicity in what follows, I assume an additive genetic model and no other covariates in the linear regression model, although these assumptions are easily relaxed.

In a quantitative trait association study, I focus on the slope $\beta_1$ in the linear regression model as a measure of the genetic effect size, and calculate the regression-based t-test statistic $T = \dfrac{|\hat{\beta}_1|}{SE(\hat{\beta}_1)}$ for the null hypothesis of no association ($H_0$: $\beta_1 = 0$). Here, $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$ are the estimated regression slope and its standard error (SE) obtained from the linear regression. I claim the association significant at significance level $\alpha$ when $T > t_{\alpha/2, N-2}$.

### 3.2.2 Uncorrected (naïve) estimators

In practice, investigators often estimate the effect size of the quantitative trait locus, parameterized as the linear regression slope $\beta_1$, using the same data used for the initial association test. I call this uncorrected estimator of $\beta_1$ "naïve" because it ignores the possible bias associated with focusing only on genetic markers with statistically significant association results.

To assess the impact of the winner's curse, I calculate the expected value of this uncorrected estimator $\hat{\beta}_{1,un}$ conditional on obtaining significant evidence for association:

$$E(\hat{\beta}_{1,un} \mid T > t_{\alpha/2,N-2}) = E(\hat{\beta}_{1,un} \mid \frac{|\hat{\beta}_{1,un}|}{SE(\hat{\beta}_{1,un})} > t_{\alpha/2,N-2}) \quad (1)$$

To simplify this calculation, I approximate (1) by assuming $SE(\hat{\beta}_{1,un}) \approx \sigma / \sqrt{S_{xx}}$:

$$E(\hat{\beta}_{1,un} \mid T > t_{\alpha/2,N-2}) \approx E(\hat{\beta}_{1,un} \mid \frac{|\hat{\beta}_{1,un}|}{\sigma / \sqrt{S_{xx}}} > t_{\alpha/2,N-2})$$

$$= \int_{-\infty}^{-a} z \times f(z; \beta_1, \sigma^2 / S_{xx} \mid z > a) dz + \int_{a}^{+\infty} z \times f(z; \beta_1, \sigma^2 / S_{xx} \mid z > a) dz$$

$$= \frac{\beta_1 - \int_{-a}^{a} z \times f(z; \beta_1, \sigma^2 / S_{xx}) dz}{P(T > t_{\alpha/2,N-2})} \quad (2)$$

where $f(z; \beta_1, \sigma^2 / S_{xx})$ is the density function of normal distribution with mean $\beta_1$ and variance $\sigma^2 / S_{xx}$, and $a = t_{\alpha/2,N-2} \sigma / \sqrt{S_{xx}}$. $S_{xx}$ is the sum of squares of the covariate matrix X. Note that the numerator of (2) is the power of the study, which equals 1 minus the cumulative density function (CDF) of the non-central t distribution with degrees of freedom (df) $N - 2$ and non-centrality parameter $ncp = \frac{\beta_1 \sqrt{S_{xx}}}{\sigma}$. The approximation of $SE(\hat{\beta}_{1,un})$ is verified by simulations. From (2), I calculate the bias of the estimator as $E(\hat{\beta}_{1,un} \mid T > t_{\alpha/2,N-2}) - \beta_1$, and the proportional bias as $\frac{E(\hat{\beta}_{1,un} \mid T > t_{\alpha/2,N-2}) - \beta_1}{\beta_1}$.

I also quantify the winner's curse effect in the estimator of the coefficient of determination $R^2 = \frac{SS_{reg}}{SS_{reg} + SS_{res}}$, where $SS_{reg}$ and $SS_{res}$ are the regression and residual sums of squares, respectively. By dividing both the numerator and denominator by $SS_{res}$, $R^2 = \frac{F}{F + N - 2}$, where $F = \frac{SS_{reg} / 1}{SS_{res} / N - 2}$ is the F test statistic for $H_0$: $\beta_1 = 0$. When there are no other covariates in the model, $F = T^2$. I calculate the expected difference in estimates of $R^2$ when taking into account the ascertainment or ignoring this ascertainment as:

$$E\left(R^2 \mid T > t_{\alpha/2,N-2}\right) - E\left(R^2\right) = E\left(\frac{F}{F+N-2} \mid F > F_{\alpha,1,N-2}\right) - E\left(\frac{F}{F+N-2}\right)$$

$$= \frac{\int_{F_{\alpha,1,N-2}}^{+\infty} \frac{x}{x+N-2} g(x)dx}{P(F > F_{\alpha,1,N-2})} - \int_{0}^{+\infty} \frac{x}{x+N-2} g(x)dx \qquad (3)$$

where here $g(x)$ is the density function for F distribution with 1 and $N-2$ degrees of freedom, and $F_{\alpha,1,N-2}$ is the corresponding quantile at significance level $\alpha$. Notice that, were $R^2$ a constant, the expected difference in (3) would be its bias in estimation owing to the winner's curse.

### 3.2.3. Ascertainment-corrected MLEs

#### *Three-parameter model*

The naïve estimator ignores the fact that we typically are interested in estimates of the regression slope $\beta_1$ only if we have strong evidence for SNP-QT association. To

address this, I propose an ascertainment-corrected maximum likelihood method that conditions on obtaining evidence for association. To this end, I calculate the conditional likelihood function

$$L(\beta_0, \beta_1, \sigma \mid y, X, T > t_{\alpha/2, N-2}) = \prod_i f(y; \beta_0 + \beta_1 x_i, \sigma^2 \mid T > t_{\alpha/2, N-2})$$

$$= \frac{1\{T > t_{\alpha/2, N-2}\}}{P(T > t_{\alpha/2, N-2})} \prod_i f(y; \beta_0 + \beta_1 x_i, \sigma^2) \quad (4)$$

where the indicator function $1\{T > t_{\alpha/2, N-2}\}$ equals 1 if $T > t_{\alpha/2, N-2}$ and 0 otherwise, and $f(y; \beta_0 + \beta_1 x_i, \sigma^2)$ is the normal density function with mean $\beta_0 + \beta_1 x_i$ and variance $\sigma^2$.

I maximize the likelihood (4) as a function of $\beta_0$, $\beta_1$ and $\sigma$ to obtain the ascertainment-corrected MLEs $\hat{\beta}_{0,as}, \hat{\beta}_{1,as}$ and $\hat{\sigma}_{as}$ by using the Nelder-Mead [1965] simplex method. I calculate the empirical standard errors of these estimates based on 1000 simulation replicates.

***One-parameter model***

Since the primary interest of investigators is to estimate the regression slope $\beta_1$, $\beta_0$ and $\sigma^2$ are nuisance parameters. I propose a one-parameter model to obtain the corrected estimator for $\beta_1$ only, by assuming that the uncorrected estimator $\hat{\beta}_{1,un}$ obtained from linear regression is a consistent estimator of $\beta_1$, and asymptotically normally distributed when sample size is sufficiently large. Therefore I calculate the conditional likelihood function for $\beta_1$ as:

$$L(\beta_1 \left| \hat{\beta}_{1,un}, T > t_{\alpha/2,N-2} \right.) = f(\hat{\beta}_{1,un}; \beta_1, \hat{var}(\hat{\beta}_{1,un}) \mid T > t_{\alpha/2,N-2})$$

$$\approx \frac{1\{T > t_{\alpha/2,N-2}\} f(\hat{\beta}_{1,un}; \beta_1, \hat{var}(\hat{\beta}_{1,un}))}{P(T > t_{\alpha/2,N-2} \left| \beta_1 \right.)} \quad (5)$$

where $f(\hat{\beta}_{1,un}; \beta_1, \hat{var}(\hat{\beta}_{1,un}))$ is the normal density function with mean $\beta_1$ and variance $\hat{var}(\hat{\beta}_{1,un})$, since $\frac{\hat{var}(\hat{\beta}_{1,un})}{var(\hat{\beta}_{1,un})} \to 1$.

I maximize the likelihood (5) as a function of $\beta_1$ to obtain the ascertainment-corrected MLEs $\tilde{\beta}_{1,as}$, and calculate the empirical standard errors of these estimators based on 1000 simulation replicates.

### 3.2.4. MSE-Weighted MLEs

Following Zhong and Prentice [2008], I also consider a weighted estimator $\hat{\beta}_{1,w}$, calculated as the weighted average of the uncorrected estimator $\hat{\beta}_{1,un}$ and the corrected estimator $\hat{\beta}_{1,cor}$:

$$\hat{\beta}_{1,w} = \hat{K}\hat{\beta}_{1,un} + (1-\hat{K})\hat{\beta}_{1,cor} \quad (6)$$

The weight is defined as $\hat{K} = \frac{\hat{\sigma}^2_{un}}{\hat{\sigma}^2_{un} + \left(\hat{\beta}_{1,un} - \hat{\beta}_{1,cor}\right)^2}$, where the denominator is the estimated mean square error (MSE) and the corrected estimator $\hat{\beta}_{1,cor}$ can be either the three-parameter-model based MLE $\hat{\beta}_{1,as}$ or the one-parameter-model based MLE $\tilde{\beta}_{1,as}$. Zhong and Prentice [2008] showed that in the case-control setting, this estimator

generally results in a smaller bias compared to the naïve and ascertainment-corrected estimators.

### 3.3 Results

#### *Uncorrected Estimators*

For a locus showing significant evidence for QT association, there is clear upward bias in the genetic effect size for the uncorrected estimator $\hat{\beta}_{1,un}$ of $\beta_1$ (Figure 3.1). This bias is particularly severe when power is low, owing to small sample size N and/or small genetic effect size $\beta_1$ (Figure 3.2). As power increases, bias decreases. Under the null hypothesis ($\beta_1 = 0$), $\beta_1$ is equally likely to be over- or under-estimated so that the bias is zero while the absolute bias is large owing to large variance. Due to symmetry, here and for the rest of the tables or figures, I provide results only for $\beta_1 \geq 0$. For example, given an association study with $N$ = 2000 samples, allele frequency p = .3, and testing at significance level of $\alpha = 10^{-6}$ under an additive genetic model, if the true value for $\beta_1$ is 0.1 (power = 5%), the expected value of the uncorrected estimator of $\beta_1$ is .178, resulting in an absolute bias of .078 and a proportional bias of 78%.

Similar to the case-control studies (Chapter 2), I found for a fixed significance level α, the proportional bias in the uncorrected estimator of $\beta_1$ is solely a function of power, independent of sample size, allele frequency, and genetic model (Figure 3.2). As expected, proportional bias decreases as power increases, since the conditioning event becomes increasingly likely as power increases. At significance level $\alpha = 10^{-6}$, the uncorrected estimator of $\beta_1$ gives a proportional bias of 50% when power is 10% but is nearly unbiased when power is 95%.

Interestingly, I found that, at a fixed significance level, the proportional bias in the uncorrected estimator of the coefficient of determination $R^2$ is solely a function of the power as well (Figure 3.3), independent of sample size and allele frequency.

***Corrected ML and MSE-Weighted Estimators***

I found that both three- and one-parameter model-based ascertainment-corrected MLEs for the genetic effect size $\beta_1$ are less biased than the uncorrected estimator when power is low or moderate (< 50%) (Figure 3.2). For example, given N = 2000 samples, allele frequency p = .3, and testing at significance level of $\alpha = 10^{-6}$ under an additive model, if the true value for $\beta_1$ is 0.1 (power = 5%), the proportional bias of the corrected MLE of δ from the three- and one-parameter models are both about –20%, compared to +78% before correction. However, both estimators tend to underestimate the true effect size. As expected, the bias and absolute bias of the MSE-weighted estimator are intermediate between those of the uncorrected estimators and corrected estimators (Figure 3.1).

The three-parameter and one-parameter estimators have very similar performance in bias reduction (Figure 3.1 and 3.2). However, the standard error of the MLE from one-parameter model is smaller compared to that of three-parameter model, and also the uncorrected estimator (Figure 3.1). As with bias and absolute bias, the variance of the MSE-weighted estimator is intermediate between that of the uncorrected and corrected estimators (Figure 3.1).

In the typical range of power for GWAS, I found that the MSE weighted estimator generally results in smaller bias compared to the uncorrected estimator or to the

corrected estimators based on either one- or three- parameter model (Figure 3.3). The improvement of the ascertainment correction is substantial when power is high. For example, at significance level $\alpha = 10^{-6}$, when study power is 80%, the proportional bias is ~10% for the uncorrected estimator of $\beta_1$ , and –18% for the ascertainment corrected estimators from both one- and three- parameter models, but only –5% for the weighted estimator.

### 3.4 Discussion

Similar to disease-marker case-control association studies, in quantitative trait association studies, the genetic effect size for associated markers tends to be overestimated as a consequence of the winner's curse. This is true because the association test statistic is correlated with the estimator of the genetic effect, and since investigators focus primarily on markers that show statistically significant evidence of association. In this chapter, I parameterized the genetic effect size as the slope in a trait-genotype score linear regression, which is often used as a measure to estimate the genetic effect size in QT association studies. I quantified the bias of the naïve estimator that ignore this ascertainment, and showed that the proportional bias in the estimators decreases as power increases. Interestingly, at fixed significance level, the proportional biases of the regression slope and the coefficient of determination are functions solely of power, independent of allele frequency or sample size.

To correct for this ascertainment bias, I proposed a three-parameter-based maximum likelihood method, and then simplified the method to a one-parameter-based model rid of nuisance parameters. The ascertainment-corrected MLEs for the regression

slope obtained from both models are generally less biased than the uncorrected estimators unless study power is moderate to high (>60%). However, both of models tend to overcorrect. Since the uncorrected estimator is expected to be upwardly biased and the corrected estimators tend to underestimate the $\beta_{1,}$ following Zhong & Prentice [2008], I also considered a weighted estimator $\hat{\beta}_{1,w}$ which takes the weighted average of the uncorrected and corrected estimators using the estimated MSE as the weight. Simulations suggest that this MSE weighted estimator generally results in smaller bias compared to the uncorrected estimator or the ascertainment-corrected estimators based on either the one- or the three-parameter model. This weighted estimator improves the ascertainment correction substantially when power is high.

Although the three-parameter and one-parameter model based estimators have very similar performance in bias reduction, the standard error of the MLE from one-parameter model is smaller than that of three-parameter model. This is likely primarily owing to optimizing the likelihood function over a one- versus three-parameter space. In addition, our one-parameter model makes an explicit normality assumption on the slope estimator.

I also considered the application of a Bayesian model to the slope estimate in linear regression. However, given the large sample size typically used in GWAS, a strong prior for $\beta_1$ is needed to influence its estimation. As a consequence, the posterior estimator of $\beta_1$ will highly depend on the prior parameters. The resulting estimate of $\beta_1$ could then be more biased if the prior is mis-specified.

In this study I focused on one-stage designed QT association study, but it is easy to extend this approach to multi-stage designs. In this chapter I only presented results

under additive model, but it is straightforward to generalize the approach to other genetic models by simply reparameterizing the genotype score. Although I investigated the model with only genotypes as the covariate, it is readily to incorporate other covariates such as demographic variables in the model.

In summary, I have presented analytic calculations to quantify the impact of the winner's curse in quantitative trait association studies, and confirm that it can result in substantial overestimation of the true genetic effect parameterized as the linear regression slope when study power is not high. To correct for the ascertainment bias, I propose a fully parameterized maximum likelihood model and also a simplified likelihood model with the nuisance parameters excluded. I demonstrate that the ascertainment-corrected estimators from both the three- and the one-parameter models result in reduced absolute bias compared to the uncorrected estimator when study power is low or moderate (<60%), and similar absolute bias when power is high; but the variance of the one-parameter-model based estimator is generally smaller than that of the three-parameter model. I also consider a MSE weighted estimator and show that it generally results in smaller bias compare to the corrected estimators based on either the one- or the three-parameter model. I recommend the use of the one-parameter maximum likelihood method and the MSE weighted estimator for estimation of genetic effect size in quantitative trait association studies.

Figure 3.1: Bias, absolute bias and mean square error (MSE) of the uncorrected, corrected, and MSE-weighted estimators for $\beta_1$ from three- and one-parameter models with sample size $N$ = 2000 and allele frequency p = .3 under an additive genetic model. Significance level $\alpha = 10^{-6}$.

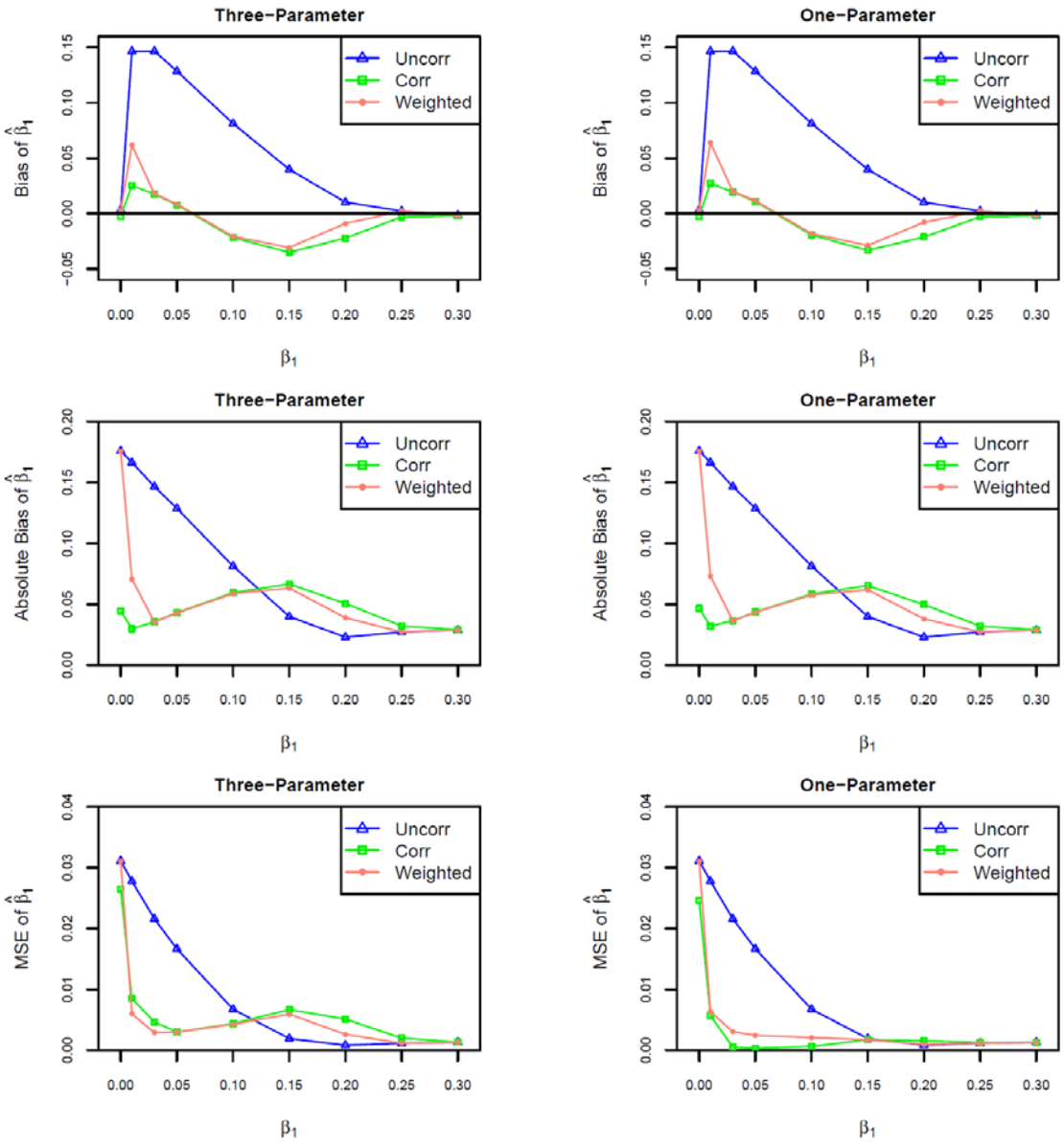


Figure 3.2: Proportional bias of the uncorrected, corrected, and MSE-weighted estimators for $\beta_1$ from three- and one-parameter models. Significance level $\alpha = 10^{-6}$. Results are presented only for $\beta_1 > 0$ under an additive genetic model.

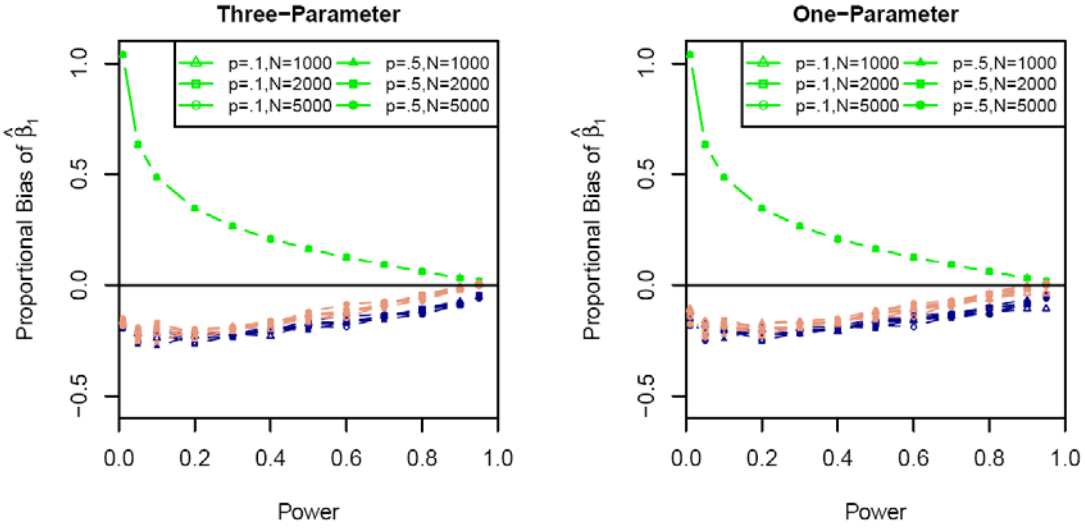
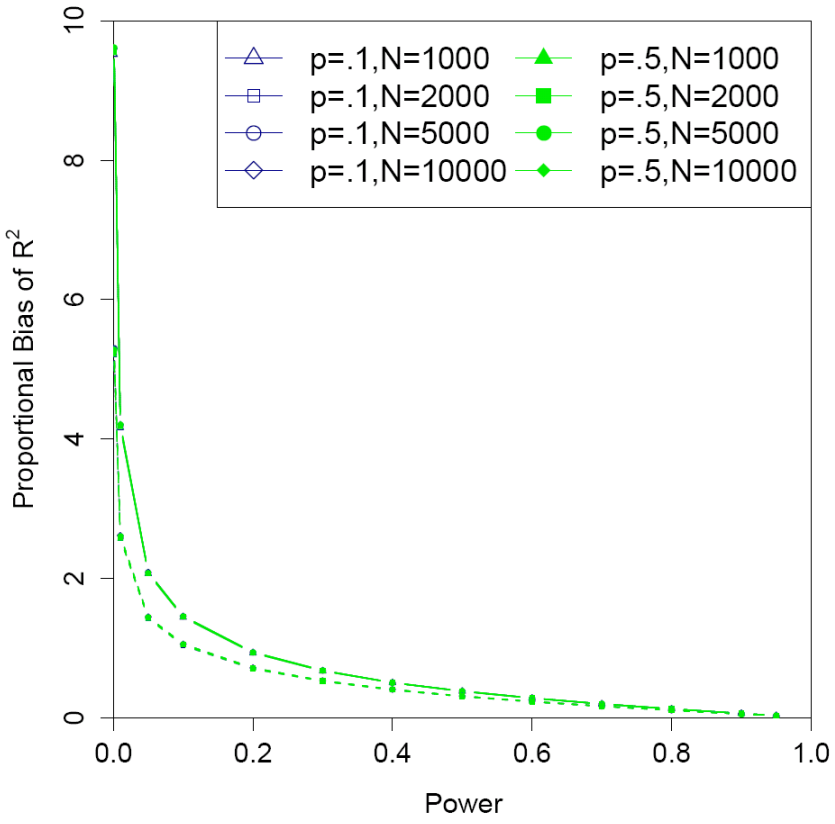


Figure 3.3: Proportional expected difference versus power for the uncorrected estimator of the coefficient of determination $R^2$ under an additive genetic additive model. Significance levels $\alpha = 10^{-4}$ (solid) and $\alpha = 10^{-6}$ (dashed).



## References


Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM, Amos CI. 2002. Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. Am J Hum Genet 70: 575-585.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. 2007. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.

Göring HHH, Terwilliger JD, Blangero J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69: 1357-1369.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-389.

Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, Aulchenko Y, Lumley T, Köttgen A, Vasan RS, Rivadeneira F, Eiriksdottir G, Guo X, Arking DE, Mitchell GF, Mattace-Raso FU, Smith AV, Taylor K, Scharpf RB, Hwang SJ, Sijbrands EJ, Bis J, Harris TB, Ganesh SK, O'Donnell CJ, Hofman A, Rotter JI, Coresh J, Benjamin EJ, Uitterlinden AG, Heiss G, Fox CS, Witteman JC, Boerwinkle E, Wang TJ, Gudnason V, Larson MG, Chakravarti A, Psaty BM, van Duijn CM. 2009. Genome-wide association study of blood pressure and hypertension. Nat Genet 41: 677-687.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PVK, Frazer KA, Cox DR, Ballinger DG. 2005. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77: 685-693.

Nelder JA, Mead R. 1965. A simplex method for function minimization. Computer J 7: 308-313.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273: 1516–1517.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez

JC, Meyer J, Ardlie K, Bengtsson K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331-1336.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341-1345.

Sham, PC, Cherny, SS, Purcell, S, Hewitt, JK. 2000. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. Am J Hum Genet 66: 1616-1630.

Siegmund D. 2002. Upward bias in estimation of genetic effects. Am J Hum Genet 71: 1183-1188.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.

Sun L, Bull SB. 2005. Reduction of selection bias in genome-wide studies by resampling. Genet Epidemiol 28: 352-367.

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.

Wu LY, Sun L, Bull SB. 2006. Locus-specific heritability estimation via bootstrap in linkage scans for quantitative trait loci. Hum Hered 62: 84-96.

Xiao R, Boehnke M. 2009. Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol 33: 453-462.

Zhong H, Prentice RL. 2008. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics 94(4): 621-634.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. 2007. Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. Science 316: 1336-1341.

# CHAPTER 4

# ALLELIC EXPRESSION IMBALANCE TO DETECT THE *CIS*-ACTING REGULATORY SNPS

Testing for association between gene expression and SNPs identified by genomewide association studies (GWAS) can help understand the relationship between these SNPs and the trait of interest and identify the gene(s) most likely to influence the trait. Allelic expression imbalance (AEI) between the two alleles of a gene can be used to detect *cis*-acting regulatory SNPs (rSNP) in individuals heterozygous for a transcribed SNP (tSNP). Appropriate analysis of AEI data depends on the extent of linkage disequilibrium (LD) between the rSNP and tSNP and whether we know linkage phase. In this paper, I propose five tests to detect association between a potential rSNP and AEI when LD between the rSNP and tSNP is incomplete (D' < 1) and linkage phase is unknown. I show that the relative power of the tests depends strongly on the magnitude of the LD between the rSNP and tSNP, and less strongly on the AEI effect size of the rSNP, number of tSNP heterozygotes, and whether the two SNPs have similar allele frequencies. I further demonstrate that the impact of a second ungenotyped rSNP on the relative power of these tests depends on the LD structure of the three SNPs, but almost never invalidates the proposed tests nor substantially changes the rankings of the tests. Based on my results, I recommend the use of F test when the rSNP and tSNP are in or

near linkage equilibrium (D'~0). When the two SNPs are in LD, in general, the mixture-model based test is most powerful for the intermediate LD levels, and the t test is typically most powerful for high LD.

## 4.1 Introduction

Over the past decades, there has been increasing interest in understanding the genetic basis of phenotypic diversity in humans. While earlier studies focused primarily on genetic variants in protein coding regions or regions flanking candidate genes, the recent development of genomewide association studies (GWAS) enables investigation of common variants in most of the human genome, including the non-coding regions that comprise ~98% of the genome [International Human Genome Sequencing Consortium, 2004]. For published GWAS signals, about 88% are located in non-coding regions, either intronic (45%) or intergenic (43%) [Hindorff et al., 2009; www.genome.gov/gwastudies]. The relationship between these associated polymorphisms and the corresponding disease or trait (henceforth trait) are, for the most part, still unknown. Regulation of gene expression is one possible mechanism to account for the connection between these polymorphisms and the trait.

Following a GWAS, a key goal is to identify the functional gene(s) and variant(s) within an identified locus of association. For example, recent GWAS of high-density lipoprotein-cholesterol (HDL-C) identified a region on chromosome 12 which contains four genes: *KCTD10*, *MMAB*, *MVK*, and *UBE3B* [Kathiresan et al., 2009; Willer et al., 2009]. In this 175 kb region of high linkage disequilibrium (LD), multiple common SNPs (minor allele frequency > .05) showed similar strength of association with HDL-C levels.

At least two of the genes, *MMAB* and *MVK*, are functionally relevant to cholesterol biology [Deodato et al., 2006; Goldstein and Brown, 1990]. To identify the gene(s) most likely to influence HDL-C, one approach is to test for association between the associated SNPs and the expression levels of these genes.

Gene expression is affected by both *cis*- and *trans*-regulatory elements, where by *cis*-regulatory elements we mean DNA polymorphisms that reside on the same chromosome as the gene they regulate, and act only on the copy of the gene on the same chromosome. These *cis*-regulatory elements usually are in close proximity to the gene being regulated, but can be far away. Trans-regulatory elements can be located on the same or a different chromosome, but regulate both alleles of the gene.

One commonly used approach to identify potential regulatory SNPs (rSNP) is to test for statistical association between SNP and mRNA transcript levels by regressing the transcript levels on the SNP genotypes [Cheung et al., 2005; Rockman et al., 2006; Pastinen et al., 2006]. An advantage of this approach is that mRNA levels can be measured using high throughput expression arrays, which assay thousands of genes simultaneously. However, the power of this analysis may be reduced by intra-individual differences in a large number of variants involved in the regulation of the gene, and also by any non-genetic differences between samples [Tao et al., 2006]. In addition, this approach does not explicitly distinguish between *cis*- and *trans*-acting regulators [Cheung et al., 2005].

A second approach is to test for association between a potential rSNP and the relative expression level of the two alleles of a gene in individuals heterozygous for a transcribed SNP (tSNP). Polymorphisms regulating gene expression in *cis* will result in

unequal amounts of mRNA from the two alleles of the gene, a phenomenon known as allelic expression imbalance (AEI). An advantage of studying AEI is that the relative mRNA level of one allele with respect to the other can be quantified more accurately than the absolute level, since the amounts of mRNA originating from the two alleles are measured in the same individual. Each allele serves as an internal standard for the other to control for *trans*-acting factors that could affect the expression of both alleles in the same way. Hence, using relative expression levels of the two alleles of the gene in each individual can specifically identify *cis*-acting regulators, in contrast to using total expression level as the outcome variable [Bray et al., 2003; Mahr et al., 2006; Pastinen et al., 2003; Pastinen et al., 2005; Tao et al., 2006]. Because of technical variability in assays for the two alleles, the relative allelic expression needs to be normalized to a reference which has equal amounts of the two corresponding alleles. Genomic DNA (gDNA) of the tSNP heterozygotes is often used for this purpose [Buckland, 2004], and I will do so here.

Some AEI studies have used samples for which phase is known [Serre et al., 2008; Ge et al., 2009]; others have relied on the LD between the rSNP and tSNP and additional genotyped SNPs nearby for haplotype reconstruction [Tao et al., 2006; Alachkar et al., 2008]. Based on the LD structure between the rSNP and tSNP, I consider three scenarios first assuming the AEI is due to a single rSNP (Figure 4.1): (1) $r^2 = 1$; (2) $r^2 < 1$, $D' = 1$; and (3) $D' < 1$, and in this chapter develop tests for the third scenario.

When $r^2 = 1$, all samples heterozygous for the tSNP are also heterozygous for the rSNP, there is only one possible haplogenotype, and we expect to observe only one AEI cluster for rSNP heterozygotes (Figure 4.1A). Here, a two-sample t test comparing mean

AEI of the cDNA to that of the gDNA may be applied for AEI detection, as suggested by Campino et al. [2008]. This approach requires at least one tSNP in the gene to be complete LD ($r^2 = 1$) with the rSNP, a requirement that is often not met.

When $r^2 < 1$ but D' = 1, there is again only one possible haplogenotype for rSNP heterozygotes and we expect to observe one AEI cluster for rSNP heterozygotes and another for rSNP homozygotes (Figure 1B). This allows comparison of mean AEI of the cDNA for rSNP heterozygotes to that of rSNP homozygotes, or alternatively to that of gDNA for all individuals as in the $r^2 = 1$ scenario.

When D' < 1, there are two possible haplogenotypes for rSNP heterozygotes and we expect to observe two AEI clusters for rSNP heterozygotes, and two more for rSNP homozygotes (Figure 1C). Given sampled and genotyped parents or genotype data from the International HapMap Project, the phase of the rSNP and tSNP may be known [Serre et al., 2008; Ge et al., 2009], and Ge et al. [2009] proposed a regression-based test for phase known data. Often, however, phase is unknown and analysis of the AEI data is more challenging in doubly heterozygous individuals. Teare et al. [2006] proposed a four-component mixture model and expectation-maximization (EM) algorithm to analyze AEI data and a likelihood ratio test (LRT) to compare mean AEI in rSNP heterozygotes and homozygotes. However, they did not describe how to assess the significance of the LRT; the usual chi-square distribution cannot be used to due to non-identifiability of parameters in the finite mixture model [Hartigan, 1985].

Here, I describe five alternative statistical procedures to analyze AEI data when D' < 1, including a two-sample t test comparing two means, a one-sided F test comparing two variances, and a mixture-model based test which fits a two-component mixture

model for rSNP heterozygotes. I further propose a test based on the minimum p-value of the t and F tests and a test that combines these two p-values. For the F, t, minimum-p-value, and combined-p-value tests, I use permutation to assess significance while allowing for non-normality or correlated tests, while for the mixture-model based test, I employ the parametric bootstrap.

My simulations demonstrate that type I error rates for all five tests are well controlled, and power of the tests depends on the LD structure (D') between the rSNP and tSNP, whether the two SNPs have similar frequencies, the AEI effect size of the rSNP, and the number of tSNP heterozygotes. The F test is generally the most powerful test when the two SNPs are in linkage equilibrium (LE) or low LD (D' < .2), but has fairly low power when the two SNPs are in high LD (D' > .5). In contrast, the t test generally is the least powerful test when LD is low (D' < .2), but most powerful when LD is high. When LD is intermediate, the mixture-model based test generally has the highest power, slightly higher than the combined-p-value test. I also demonstrate that the presence of a second ungenotyped rSNP generally does not invalidate these tests, but may result in reduced or increased power, depending on the LD structure between the three loci and the direction of effect of the two rSNPs. To detect *cis*-acting regulatory SNPs using AEI, I recommend the use of F test when the rSNP and tSNP are in or near linkage equilibrium (D'~0). When the two SNPs are in LD, in general, the mixture-model based test is most powerful for intermediate LD levels, and the t test is typically most powerful for high LD.

## 4.2 Methods

### 4.2.1 Model and assumptions

I initially assume that the differential expression of a gene is caused in part by a single *cis*-acting regulatory SNP (rSNP) with alleles R and r, with R causing higher expression of the allele on its chromosome compared to r. Allelic expression imbalance is measured in N independent individuals who are heterozygous for a transcribed SNP (tSNP) with alleles T and t. Let $p_{\mathrm{R}}$ and $p_{\mathrm{T}}$ denote the frequencies of R and T. For individual i, let $G_i \in \{\mathrm{RR}, \mathrm{Rr}, \mathrm{rr}\}$ be the genotype of the rSNP and $H_i \in \left\{\frac{\mathrm{RT}}{\mathrm{rt}}, \frac{\mathrm{rT}}{\mathrm{Rt}}, \frac{\mathrm{RT}}{\mathrm{Rt}}, \frac{\mathrm{rT}}{\mathrm{rt}}\right\}$ be the haplogenotype of the rSNP and tSNP.

I define the allelic expression ratio (AER) as the ratio of the allele T transcript level to the allele t transcript level, and use the natural logarithm of this AER normalized by the corresponding ratio in gDNA for the tSNP heterozygotes as the outcome variable

$$y = \ln \mathrm{AER}_{\mathrm{cDNA}} - \mathrm{mean}(\ln {}^{\mathrm{T}}\!/_{\mathrm{t}})_{\mathrm{gDNA\_Tt}} \qquad (1)$$

In what follows, I will refer to this normalized logarithm of AER as lnAER.

Actually, this normalization of lnAER by the gDNA mean does not affect the type I error rate or power of the tests I propose in following sections. However, for purposes of estimating the AEI effect size of the rSNP, normalization by gDNA makes the resulting data easier to interpret since the possible difference of the detection system to quantify the two alleles of the gene is controlled.

Compared to rSNP homozygotes ($h = \frac{\mathrm{RT}}{\mathrm{Rt}}, \frac{\mathrm{rT}}{\mathrm{rt}}$) for which we do not expect to observe AEI, in the presence of AEI, Rr heterozygotes will show an increased T:t expression ratio if $h = \frac{\mathrm{RT}}{\mathrm{rt}}$ and a decreased T:t expression ratio if $h = \frac{\mathrm{rT}}{\mathrm{Rt}}$.

For individual i with haplogenotype $h$, I assume $y_i$ is normally distributed with mean $\mu_h$ and variance $\sigma^2$, where

$$\mu_h = \begin{cases} \mu_0 & \text{for } h = \dfrac{\mathrm{RT}}{\mathrm{Rt}} \text{ or } \dfrac{\mathrm{rT}}{\mathrm{rt}} \\ \mu_0 + \alpha_R & \text{for } h = \dfrac{\mathrm{RT}}{\mathrm{rt}} \\ \mu_0 - \alpha_R & \text{for } h = \dfrac{\mathrm{rT}}{\mathrm{Rt}} \end{cases} \qquad (2)$$

Under the null hypothesis of no AEI effect, $\alpha_R = 0$. I assume that there is no difference in the mean or variance of y between the RR and rr homozygotes [Pastinen et al. 2009].

Our goal is to test for association between AEI and the putative rSNP. In the presence of AEI ( $\alpha_R \neq 0$ ), different haplogenotypes have different mean allelic expression ratios, and I model the data for the rSNP heterozygotes as a mixture of normal distributions for each haplogenotype. I consider a two-sample t test comparing the mean of lnAER between the different rSNP genotype groups, an F test comparing their variances, and a mixture-model based test that explicitly acknowledges the nature of the mixing distribution of the lnAER data. Since the t and F tests have different power characteristics for different levels of LD between the rSNP and tSNP, I also consider a test based on the minimum of the p-values of the t and F tests and another test that combines the two p-values based on Fisher's [1948] method.

### 4.2.2 Two-sample (two-sided) t test and (one-sided) F test

When the rSNP and tSNP are in LD, one of the two RrTt haplogenotypes ($h = \dfrac{\mathrm{RT}}{\mathrm{rt}}, \dfrac{\mathrm{rT}}{\mathrm{Rt}}$) is more common than the other. In the presence of AEI, I expect mean lnAER for Rr heterozygotes to be higher or lower than for the combined RR and rr homozygotes,

depending on which haplogenotype is more common. I propose using a (two-sided) two-sample t test for the hypothesis that mean lnAER of the Rr heterozygotes differs from that of the combined RR and rr homozygotes, allowing for unequal variances between the heterozygous and homozygous groups due to the mixing distribution for the Rr heterozygotes.

When the rSNP and tSNP are in LE or low LD, the two haplogenotypes ($h = \frac{RT}{rt}, \frac{rT}{Rt}$) have similar frequencies. In the presence of AEI, I expect to observe approximately half the Rr heterozygotes to have high lnAER and the remainder to have low lnAER, resulting in an increased variance in lnAER for Rr heterozygotes compared to the combined RR and rr homozygotes. I therefore propose using the F test for equal variances against the one-sided alternative that variance in Rr heterozygotes is greater than in the combined RR and rr homozygotes.

For both these tests, I use permutations to assess significance while accounting for violation of the normality assumption due to the nature of the mixing distribution of the lnAER. I estimate the p-value by the sum of the proportion of permuted data test statistics greater than the observed test statistic plus half the proportion of permuted data test statistics tied with the observed test statistic.

### 4.2.3 Mixture-model based test

Given unknown linkage phase and incomplete LD, the lnAER data follow a mixture distribution. I therefore propose a mixture-model based test which fits a two-component normal mixture model for the rSNP heterozygotes, with likelihood:

$$\mathrm{L} = \prod_{i \in \{G_i = \mathrm{RR},\, \mathrm{rr}\}} f(y_i; \mu_0, \sigma^2) \times$$

$$\prod_{i \in \{G_i = \mathrm{Rr}\}} \left\{ \pi f(y_i; \mu_0 + \alpha_R, \sigma^2) + (1 - \pi) f(y_i; \mu_0 - \alpha_R, \sigma^2) \right\} \qquad (3)$$

Here, $f(\mu, \sigma^2)$ is the density function for normal distributions with mean $\mu$ and variance $\sigma^2$ and $\pi$ is the mixing proportion of the two-component mixture model.

I perform a likelihood ratio test (LRT) for the null hypothesis: $\alpha_R = 0$:

$$\Lambda = -2 \ln \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_1)} \qquad (4)$$

where $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\sigma}^2)$ and $\hat{\theta}_1 = (\hat{\pi}, \hat{\mu}_0, \hat{\alpha}_R, \hat{\sigma}^2)$ are the maximum likelihood estimators (MLEs) under the null and alternative hypotheses, respectively. Since the likelihood cannot be maximized directly, I obtain MLEs by the simplex method [Nelder and Mead, 1965]. To assess significance for $\Lambda$, I apply the parametric bootstrap [McLachlan, 1987], since the chi-square distribution cannot be used to approximate the null distribution of LRT in finite mixture models [Hartigan, 1985]. For each bootstrap, I simulate the lnAER data from the distribution with parameters estimated under null hypothesis, and calculate the LRT statistic based on the bootstrapped data. I estimate the p-value as the proportion of the bootstrap LRT statistics greater than the observed LRT statistic; no ties were observed.

### 4.2.4 Minimum- and combined-p-value tests

As we shall see in the Results, the t test tends to be more powerful when the rSNP and tSNP are in strong LD, while the F test tends to be more powerful given weak LD.

To take advantage of this behavior, I consider two additional tests. The minimum-p-value test

$$T_{min} = \min(P_t, P_F) \qquad (5)$$

selects the minimum of the p-values for the t and F tests ($P_t$, $P_F$). In contrast, the combined-p-value test

$$T_{com} = -2(\ln P_t + \ln P_F) \qquad (6)$$

uses Fisher's (1948) method to meta-analyze the information from the two tests. Both approaches seek to take advantage of whichever test is more powerful. Significance for both test statistics is assessed via permutation as described above and allows me to account for the correlation between the t and F tests.

### 4.2.5 t test when $r^2 = 1$

When the rSNP and tSNP are in complete LD ($r^2 = 1$), I carry out the analysis by a two-sample t test comparing mean lnAER for the rSNP heterozygotes to the corresponding mean for the gDNA Tt heterozygotes.

## 4.3 Simulations

### 4.3.1 One regulatory SNP

I evaluated the performance of the tests to detect association between AEI and the potential rSNP by simulating samples with varying numbers N of Tt heterozygotes, under models with lnAER mean effect $\alpha_R$ and variance $\sigma^2$, allele frequencies $p_R$ and $p_T$, and D' values $0 \leq D'_{RT} < 1$ between the rSNP and tSNP.

For each individual, I simulated haplotype pairs according to the conditional probabilities of the two-locus haplogenotypes assuming ascertainment for Tt heterozygotes. For example,

$$f_{h=\frac{\underline{\mathrm{RT}}}{\underline{\mathrm{Rt}}}} = \frac{p(\underline{\mathrm{RT}},\underline{\mathrm{Rt}})}{p(\mathrm{Tt})} = \frac{w_{\underline{\mathrm{RT}}} w_{\underline{\mathrm{Rt}}}}{p_{\mathrm{T}}(1-p_{\mathrm{T}})} \qquad (7)$$

where $w_l$ is the frequency of haplotype $l \in \{\underline{\mathrm{RT}}, \underline{\mathrm{rT}}, \underline{\mathrm{Rt}}, \underline{\mathrm{rt}}\}$ . I then simulated the corresponding lnAER data from a normal distribution with the appropriate haplogenotype-specific mean described in (2) and variance $\sigma^2$.

When the rSNP and tSNP are in complete LD ($r^2 = 1$), I simulated the gDNA data by assuming the mean and variance of the corresponding logarithm ratio for the gDNA to be the same as those for the rSNP homozygotes cDNA.

### 4.3.2 Two regulatory SNPs

In the preceding work I have assumed a single rSNP. In fact, there could be more than one [see for example Ge et al., 2009]. To assess the impact of a second (ungenotyped) regulatory SNP on the power and relative rankings of the proposed tests, I simulated lnAER data assuming that there is a second *cis*-acting rSNP with alleles $\mathrm{R_U}$ and $\mathrm{r_U}$ influencing allelic expression. $p_{\mathrm{R_U}}$ is the frequency of the allele $\mathrm{R_U}$ causing higher expression for the allele on the same chromosome.

Given two regulatory SNPs $\mathrm{R_G}$ (genotyped) and $\mathrm{R_U}$ (ungenotyped), there are 16 possible haplogenotypes for Tt heterozygotes. Probabilities for these haplogenotypes can be calculated as a function of the pairwise D' values $\mathrm{D'_{R_G R_U}}$, $\mathrm{D'_{R_G T}}$ and $\mathrm{D'_{R_U T}}$, and the third order LD $\mathrm{D_{R_G R_U T}}$ between the three loci. As defined by Bennett [1954]:

$$D_{R_G R_U T} = w_{R_G R_U T} - p_T D_{R_G R_U} - p_{R_U} D_{R_G T} - p_{R_G} D_{R_U T} - p_{R_G} p_{R_U} p_T \qquad (8)$$

where $w_{R_G R_U T}$ is the haplotype frequency, and $D_{R_G R_U}$, $D_{R_G T}$ and $D_{R_U T}$ are the unnormalized pairwise LD for the three loci. The normalized third order LD

$$D'_{R_G R_U T} = \frac{D_{R_G R_U T} - D_{R_G R_U T}(\min)}{D_{R_G R_U T}(\max) - D_{R_G R_U T}(\min)} \qquad (9)$$

[Thomson and Baur, 1984], where $D_{R_G R_U T}(\min)$ and $D_{R_G R_U T}(\max)$ are the lower and upper bounds for $D_{R_G R_U T}$ described in that paper.

I assume that the $R_U$ allele of the ungenotyped rSNP increases mean lnAER by $\alpha_{R_U}$, and that the two regulatory SNPs act additively, resulting in the type of pattern displayed by the balloon plot in Figure 4.2, which shows the expected lnAER data patterns for different levels of LD between the rSNP and tSNP. The diameter of each circle corresponds to the frequency of the haplogenotype(s) to its right while the center of the circle corresponds to mean lnAER in individuals with that halplogenotype(s). For example, lnAER for genotyped rSNP $R_G R_G$ homozygotes may display three clusters, with means $\mu_0 + \alpha_{R_U}$ (corresponding to haplogenotype $k = \frac{R_G R_U T}{R_G r_U t}$), $\mu_0$ ($k = \frac{R_G R_U T}{R_G R_U t}, \frac{R_G r_U T}{R_G r_U t}$), and $\mu_0 - \alpha_{R_U}$ ($k = \frac{R_G r_U T}{R_G R_U t}$). As many as three clusters also may be present for $r_G r_G$ individuals, and six for $R_G r_G$ heterozygotes.

As in the one-rSNP case, for each individual, I simulate haplotype pairs based on probabilities analogous to those in (7), and the corresponding lnAER data with appropriate haplogenotype-specific mean and variance $\sigma^2$.

## 4.4 Results

In this section, I show the type I error rate and power of my five proposed tests for SNP-AEI association given one or two rSNP. My main goals are 1) to illustrate the similarities and differences in relative rankings of power for these tests over a wide range of scenarios and 2) to determine if patterns of the relative power are consistent enough to draw general conclusions about the most powerful test given existing information on the rSNP and tSNP to be tested (allele frequencies and LD between them). To do this I will focus on specific examples from a much larger set of simulations.

### 4.4.1 Single rSNP

***Type I error***

I examined the type I error rates for the F, t, mixture-model based, minimum-p-value, and combined-p-value tests of association between AEI and the potential rSNP using computer simulation. I considered LD levels between the rSNP and tSNP $D'_{RT}$ = .0, .1, …, .9, numbers of tSNP heterozygotes N = 25, 50, 100, and 500, rSNP allele frequencies $p_R$ = .1, .3, .5, .7, and .9, and tSNP frequencies $p_T$ = .1, .3, and .5. I found that in all cases examined, empirical type I error rate estimates based on 10,000 simulation replicates were consistent with nominal significance levels α = .01, .05, and .10. Figure 4.3 shows results for N = 50 and 100, $p_R = p_T$ = .3 for α = .05. For purpose of comparison, I included $D'_{RT}$ = 1 (highlighted by * in Figure 4.3), with type I error rates estimated for a t test comparing the sample mean of lnAER for rSNP heterozygotes to the corresponding sample mean of gDNA when $p_R = p_T$ or to the sample mean of lnAER for the rSNP homozygotes when $p_R \neq p_T$. This is true for all the figures.

***Power***

***Outline*** Next I evaluated the power of the five tests at significance level $\alpha = .05$ as a function of these same values of LD levels $D'_{RT}$, and allele frequencies $p_R$ and $p_T$. For numbers of tSNP heterozygotes N and AEI rSNP effect sizes $\alpha_R$, I considered combinations of N = 50, 100, and 500 and $\alpha_R$ of 0.3 to 1.2, chosen to achieve power in the range of 40 to 90% to yield informative comparisons. For example, when N = 100, I considered $\alpha_R$ = 0.6, 0.8, and 1.0. Throughout this section, I fixed variance $\sigma^2 = 1$.

***Equal allele frequencies: impact of LD*** I first looked at the power of the tests under different levels of $D'_{RT}$ when the two SNPs have equal allele frequencies $p_R = p_T$, holding the other parameters constant. Figure 4.4A displays these results for $p_R = p_T = .3$, N = 100, and $\alpha_R = .8$; I observed nearly identical patterns of the relative rankings of the power for $p_R = p_T = .1$ or .5 for $(N, \alpha_R) = (50, 1.2)$ and (500, 0.3) (see below). My simulations demonstrate that the F test has the highest power among the five tests when the two SNPs are in LE or low LD ($D'_{RT} < .2$), but power decreases dramatically as $D'_{RT}$ increases and is fairly low when $D'_{RT}$ is moderate or high (> .4) (Figure 4.4A). In contrast, the t test is the least powerful test when $D'_{RT}$ is low (< .2), but its power increases rapidly as $D'_{RT}$ increases, and it becomes the most powerful test when $D'_{RT}$ is moderate or high (> .6).

This pattern holds because when rSNP and tSNP are in LE or low LD, on average ~½ the rSNP heterozygotes have high lnAER and the remainder low lnAER, resulting in similar mean but increased variance in lnAER for rSNP heterozygotes compared to rSNP homozygotes (Figure 4.4B). For moderate to high $D'_{RT}$, the majority of rSNP

heterozygotes are expected to have high (when $h = \frac{RT}{rt}$ is the more common haplogenotype) or low (when $h = \frac{rT}{Rt}$ is the more common haplogenotype) lnAER (Figure 4.4B), resulting in a greater difference in mean lnAER between the Rr heterozygotes and the combined RR and rr homozygotes, but similar variances between the two groups.

The mixture-model based test shows a trend similar to that of the t test as $D'_{RT}$ increases, but is substantially more powerful than the t test when $D'_{RT} < .3$, the most powerful test when $.1 < D'_{RT} < .5$, and only slightly less powerful than the t test when $D'_{RT} > .4$ (Figure 4.4A). Here and elsewhere (see Figures 4.5, 4.6, 4.9), power for the minimum-p-value and combined-p-value tests generally are similar to those for the mixture model based-test, but the power for the mixture-model based test usually is greater and never is substantially less.

As for the actual power values, the F test displays a monotonic decreasing trend as $D'_{RT}$ increases, while the other four tests display an increasing-and-then-decreasing trend with power maximized at an intermediate $D'_{RT}$. This is because for high $D'_{RT}$ only a few individuals are rSNP homozygotes, conditional on the fact they are heterozygotes at the tSNP (Figure 4.4B). The small number of rSNP homozygotes results in poor estimation of $\mu_0$ and decreased power for the mixture-model based test, and the unbalance in numbers between the rSNP heterozygous and homozygous groups results in decreased power for the t test and consequently for the minimum- and combined-p-value tests.

***Equal allele frequencies: impact of sample size and effect size*** I next evaluated the power of the tests varying the number of tSNP heterozygotes N and rSNP AEI effect size $\alpha_R$ while continuing to hold $p_R = p_T = .1$, .3, or .5.

Figure 4.5 shows results for $(N, \alpha_R) = (100, 0.8)$, (50, 1.2), and (500, 0.3) when $p_R = p_T = .3$. My simulations show that while power of the tests varies according to scenario, the relative rankings of the tests across different levels of LD remain similar for different $(N, \alpha_R)$ combinations, except that the mixture-model based test is most powerful for a slightly broader range of LD levels when $(N, \alpha_R) = (50, 1.2)$, and a narrower range when $(N, \alpha_R) = (500, 0.3)$. I observed similar results for $p_R = p_T = .1$ or .5, and also for additional combinations of $(N, \alpha_R) = (50, 1.0)$, (100, 0.6) and (100, 0.7) (data not shown).

When $\alpha_R = 1.2$ and N = 50, the mixture-model based test is most powerful for $.1 < D'_{RT} < .6$ (Figure 4.5B) compared to $.1 < D'_{RT} < .5$ when $\alpha_R = 0.8$ and N = 100 (Figure 4.5A). This is because the stronger AEI effect leads to two more separated clusters for rSNP heterozygotes, resulting in a better fit of the mixture model.

When $\alpha_R = 0.3$ and N = 500, the F test is most powerful when $D'_{RT} = 0$, the mixture-model based test is most powerful only for $D'_{RT} = .2$, and the t test is most powerful for a much wider range of LD ($D'_{RT} > .2$) (Figure 4.5C). But in contrast to $\alpha_R = 0.8$ or 1.2, the power of all the tests when $D'_{RT}$ is low or high is much less than the powers for all but the F test when $D'_{RT}$ is moderate.

***Unequal allele frequencies*** I next evaluated the power of the tests varying the rSNP and tSNP allele frequencies $p_R$ and $p_T$. Figure 6 displays results for $p_R = .1$ and .3, $p_T = .3$, N = 100, and $\alpha_R = .8$. While power of the tests varies by scenario, the relative

rankings of the tests are generally consistent across different $p_R$-$p_T$ combinations for fixed $(N, \alpha_R)$ (data not shown). As for $p_R = p_T$, the F test is most powerful when $D'_{RT}$ is low ($D'_{RT} < .2$) and the t test when $D'_{RT}$ is high ($D'_{RT} > .7$), while the mixture-model based test is most powerful in a wider range of $D'_{RT} = .2$ to .7 compared to $D'_{RT} = .2$ to .4 when $p_R = p_T$ (Figure 4.6AB). In addition, when $p_R$ and $p_T$ are sufficiently different, for all tests except the F test, power increases monotonically as $D'_{RT}$ increases, in contrast to an increasing-and-then-decreasing trend when $p_R = p_T$ as described previously.

### 4.4.2 Two rSNPs

***Type I error***

***Outline*** I investigated the impact of a second (ungenotyped) rSNP on the type I error rates of the five tests to detect AEI association with the genotyped putative rSNP (Figures 4.7, 4.8). I assume here that the genotyped putative rSNP has no effect on lnAER ($\alpha_{R_G} = 0$) while the ungenotyped rSNP has mean effect size $\alpha_{R_U} = .8$ and variance $\sigma^2 = 1$. Figure 4.7 and 4.8 display the type I error rates evaluated for different LD structures between the three loci, assuming allele frequencies for the genotyped rSNP $p_{R_G} = .3$, ungenotyped rSNP $p_{R_U} = .1$, and tSNP $p_T = .3$. I also performed simulations for $p_{R_G} = .1$ and .5, and $p_{R_U} = .3$ and .5 (data not shown).

***Ungenotyped rSNP in LE with genotyped putative rSNP and tSNP*** When the ungenotyped rSNP is in LE with both the genotyped putative rSNP and the tSNP ($D'_{R_GR_U} = D'_{R_UT} = 0, D'_{R_GT}$ varies from 0 to 1), empirical type I error rates are consistent with nominal expectation for $\alpha = .05$ (Figure 4.7A), and also for $\alpha = .01$ and .10 (data not

shown). This is because the independent ungenotyped rSNP adds a similar amount of variability to lnAER for genotyped putative rSNP heterozygotes and homozygotes, resulting in a similar lnAER pattern as for the single rSNP case under the null hypothesis (Figure 4.7B).

***Ungenotyped rSNP in LD with genotyped putative rSNP and tSNP*** When the ungenotyped rSNP is in moderate LD with both the genotyped putative rSNP and the tSNP ( $D'_{R_G R_U} = D'_{R_U T} = .5, D'_{R_G T}$ varies in the figure from 0 to 1), I observed both increase and decrease in the type I error rate from nominal expectations (Figure 4.8A), and the increase in type I error rate is more substantial compared to when the ungenotyped rSNP is in stronger LD with the tSNP (Figure 4.8B). In fact, this is not really a type I error, since the genotyped putative rSNP is serving as a proxy for the actual (ungenotyped) rSNP. Interestingly, I observed that the F test is conservative when the genotyped putative rSNP is in moderate or high LD with the tSNP (Figure 4.8A). This is because the variance of lnAER in the $R_G r_G$ heterozygotes could be smaller than that in the combined $R_G R_G$ and $r_G r_G$ homozygotes when $D'_{R_U T}$ is high (balloon plot in Figure 4.8A) due to the presence of the second ungenotyped rSNP, causing the one-side F test to have a type I error rate smaller than nominal expectations.

***Power***

***Outline*** Next I investigated the impact of a second ungenotyped rSNP on the power of the five tests. Here, I assumed the two rSNPs act additively on gene expression and have same effect size with mean $\alpha_{R_G} = \alpha_{R_U} = .8$. I observed similar results for

$(N, \alpha_{R_G} = \alpha_{R_U})$ = (50, 1.2) and (500, 0.3) (data not shown), and report power under the same settings as considered above for type I error.

***Two rSNPs have same regulation direction and are in LE*** I first assumed that the minor alleles of the two rSNPs regulate gene expression in same direction. I found that when the ungenotyped rSNP is in LE with both the genotyped putative rSNP and the tSNP ($D'_{R_G R_U} = D'_{R_U T} = 0$), the relative rankings of the tests are essentially unchanged compared to the single rSNP case, although the power of each test decreases slightly (Figures 4.4A, 4.9A). This is because the presence of the second ungenotyped rSNP increases variation of the lnAER data for all tSNP heterozygotes (Figure 4.9B).

***Two rSNPs have same regulation direction and are in LD*** When the second ungenotyped rSNP is in moderate LD with both the genotyped putative rSNP and the tSNP ($D'_{R_G R_U} = D'_{R_U T} = .5, D'_{R_G T}$ varies in the figure from 0 to 1), I again observed generally unchanged relative rankings of the tests at each D' level between the genotyped putative rSNP and the tSNP (Figure 4.10A). However, the power of the tests increases slightly, due to the LD of the ungenotyped rSNP with the tSNP and the consistent direction of effect on AEI of the two rSNPs (Figure 4.10A). This power increase is more substantial when LD between the ungenotyped rSNP and the tSNP is higher (Figure 4.10B) or, as expected, when the allele frequency of the ungenotyped rSNP $p_{R_U}$ is higher (data not shown).

***Two rSNPs have opposite regulation direction and are in LD*** Next I assumed that the two rSNPs act additively but the minor alleles of the two rSNPs regulate gene expression in opposite directions. When the two rSNPs have similar allele frequencies and are in strong LD, as expected, the power of all tests is much lower compared to the

single rSNP case (data not shown). When the second ungenotyped rSNP is in moderate LD with the genotyped putative rSNP and the tSNP ($D'_{R_G R_U} = D'_{R_U T} = .5$), the power of the F and mixture-model based tests decreases when the genotyped putative rSNP and the tSNP are in low LD ($D'_{R_G T} < .4$), resulting in a wider $D'_{R_G T}$ range ($D'_{R_G T} > 0$) for the t test to be the most powerful test compared to the single rSNP case (Figures 4.4A, 4.11A). Because part of the AEI effect of the two rSNPs cancels, when $D'_{R_G T}$ is low, we expect to see a smaller difference in lnAER variance between the $R_G r_G$ heterozygotes and the combined $R_G R_G$ and $r_G r_G$ homozygotes, and also to see two less separated clusters in the $R_G r_G$ heterozygotes (Figure 4.11B), which are likely to account for the power decrease for the F and mixture-model based tests respectively.

### 4.5 Discussion

Measurement of the relative expression levels of the two alleles of a gene can be used to identify *cis*-acting regulatory SNPs. Recent AEI studies have used samples such as the HapMap CEU with phase known data [Ge et al., 2009], while other studies have estimated haplotypes for the potential rSNPs and tSNP based on the LD between them and additional genotyped SNPs nearby, ignoring the possible errors in haplotype estimation [Tao et al., 2006; Alachkar et al., 2008]. While AEI data may be collected without phase information [for example, Marie Fogarty and Karen Mohlke, personal communication], few studies have tested for AEI when D' < 1 because of the lack of well evaluated methods. It is this situation on which I have focused in this chapter.

I initially proposed a t test, F test, and mixture-model based test of AEI-SNP association. The t test tends to be most powerful when the rSNP and tSNP are in high LD, while the F test tends to be most powerful when the two SNPs are in LE or low LD. The mixture-model based test, which explicitly acknowledges the expected mixed distribution nature of the data, tends to be the most powerful test when the rSNP and tSNP are in intermediate LD, and is not much less powerful than the F test when LD is low but not zero, and the t test when LD is high. To take advantage of the strengths of both the F and t tests under different levels of LD, I also proposed the minimum- and combined-p-value tests which use information from both the t and F tests. My simulations showed that the combined-p-value test is occasionally more powerful, and in most simulations only slightly less powerful than the mixture-model based test when the two SNPs are in moderate LD, suggesting an alternative to the mixture-model based test with the advantage of less computational complexity for investigators uncomfortable with a more complex statistical approach.

For the mixture-model based test, I applied a two-component normal mixture model to the rSNP heterozygotes: $\frac{RT}{rt}$ and $\frac{rT}{Rt}$. Teare et al. [2006] also proposed a mixture-model based method. The four components of their mixture model correspond to the two rSNP heterozygous haplogenotypes $\frac{RT}{rt}$ and $\frac{rT}{Rt}$ and the two rSNP homozygous haplogenotypes $\frac{RT}{Rt}$ and $\frac{rT}{rt}$. Since the same information is available for both models, it is not obvious to us why their results differ.

Teare et al. proposed the use of a likelihood ratio test (LRT) comparing the four-component model to a one-component model given no AEI effect. Within the paper they

did not describe how to assess the significance of the LRT statistic. In further communication they stated that they compare their LRT statistic to a chi-squared distribution on one degree of freedom [Mauro Santibánez-Koref, personal communication]. In fact, the finite mixture model belongs to a non-regular parametric family and most classical asymptotic results do not apply, so that the limiting null distribution of the LRT for homogeneity is complicated and cannot be approximated this way [Hartigan, 1985; Chen and Chen, 2001]. Some investigators have suggested a modified likelihood ratio test which incorporates a penalty term in the likelihood to force the estimates away from the boundary of the parameter space [Chen and Kalbfleisch, 1996; Chen, 1998; Chen et al., 2001]. This method requires specifying a parameter in the penalty term, and the choice of the parameter affects power of the test. I instead chose to use the parametric bootstrap to determine the significance of the LRT, a procedure described originally in this context by McLachlan [1987]. The bootstrap provides an estimate of the null distribution of the LRT based on the distribution parameters estimated from the observed data. McLachlan [1987] has shown that the type I error rate of this method is well controlled, which is also confirmed by my simulations.

I initially assumed a single rSNP model for AEI. In fact, AEI could be due to more than one rSNP [see for example Ge et al., 2009]. To examine the sensitivity of the proposed tests to the presence of >1 rSNP, I studied the impact of an ungenotyped rSNP on the size and power of my tests to detect association between AEI and the genotyped (putative) rSNP. My simulations demonstrated that when the second ungenotyped rSNP is in LE with both the genotyped putative rSNP and the tSNP, the type I error rate of the

tests is well controlled and that the relative rankings of the powers of the various tests are essentially unchanged compared to the single rSNP case.

When the ungenotyped rSNP is in LD with both the genotyped putative rSNP and the tSNP, my simulations demonstrate that the 'false positive' rate of the tests can be high, corresponding to the genotyped putative rSNP serving as a proxy for the ungenotyped rSNP. Thus, when an association between AEI and a potential rSNP is detected, we can at most infer that the AEI is due to the putative rSNP and/or one or more other rSNP(s) in LD with the genotyped putative rSNP.

Given this LD structure, the relative rankings of the tests essentially remain unchanged, while the absolute power of the tests can either decrease or increase depending on the frequencies of the expression-increasing allele of the two rSNPs. When the two rSNPs are in high $r^2$, there will be a substantial increase or decrease of power, depending on whether the rSNP expression-increasing alleles are on the same or opposite chromosomes.

After extensive simulations in which I varied the allele frequencies from .1 to .9, the pairwise LD from 0 to 1, and the third order LD for the three loci from 0 to 1, I did identify two LD scenarios in which the second ungenotyped rSNP leads to inconsistent tests. In the first scenario, pairwise LD values for the three pairs of markers are all (near) zero, but the third-order LD is (near) one. In this case, there are four three-locus haplotypes $\underline{R_G R_U T}$ , $\underline{R_G r_U t}$ , $\underline{r_G R_U t}$ , and $\underline{r_G r_U T}$ , each with probability ~.25, and correspondingly four haplogenotypes $\frac{R_G R_U T}{R_G r_U t}$, $\frac{R_G R_U T}{r_G R_U t}$, $\frac{r_G r_U T}{R_G r_U t}$, and $\frac{r_G r_U T}{r_G R_U t}$ also with probabilities ~.25. If the two rSNPs act additively and have same AEI effect size, the AEI data are expected to be identical in the rSNP heterozygotes and combined homozygotes.

As a consequence, the power of the proposed tests would equivalent to the type 1 error rate regardless of sample size. However, this scenario is unlikely to occur in real data [Nielsen et al., 2004]. I confirmed this through extensive search in the HapMap CEU samples. I examined at all three-SNP combinations in a 200 kb window on chromosome 1 and estimated the three-locus haplotype frequencies for each combination. I did not find a single example even approximating this LD scenario.

The second and more plausible scenario in which an ungenotyped rSNP results in tests with little or no power is when the two rSNPs are in strong LD ($r^2 \sim 1$) and act additively with same effect sizes but in opposite directions on gene expression. In this case, the AEI effects of the two rSNPs cancel.

While AEI has important advantages for studying gene expression, notably to explicitly detect *cis*-acting regulatory SNPs, it also has important limitations. First, individuals used for analysis must be heterozygous for the transcribed SNP, which requires common SNPs in order to obtain a reasonable number of samples. Second, until recently allelic expression has been measured by low throughput procedures such as reverse transcription PCR (RT-PCR) [for example, Heighway et al., 2005]. However, with the development of next-generation sequencing technologies, high throughput allelic expression data has begun to be generated [Ge et al., 2009] and more will be available in the near future, so that we expect this limitation to be addressed.

When the rSNP and tSNP have equal allele frequency and are in high D', my simulations show a decreased power for all the tests I proposed, due to smaller sample size for the rSNP homozygotes. In this situation, it will be useful to incorporate information from genomic DNA (gDNA) for all individuals. In this chapter, I focused on

the situation that the rSNP and tSNP are in incomplete LD (D' < 1). When D' = 1 and $r^2$ < 1, we could choose to compare the sample mean lnAEI of the cDNA for rSNP heterozygotes to that of rSNP homozygotes, or alternatively to that of gDNA for all individuals. For future work, I will consider developing tests which take advantage of the gDNA information and combine with the cDNA for rSNP homozygotes.

My simulations indicate that the power of the most powerful tests maximize at different D' levels between the rSNP and tSNP in the various scenarios considered in my study. This information may prove useful for optimal design of AEI experiments to help choose the tSNP on which to focus when there are multiple tSNPs available in the gene of interest.

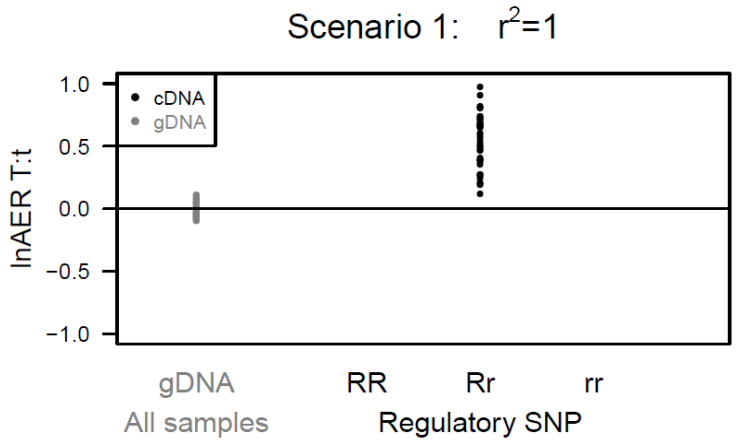
As I proceed to write the manuscript based on this chapter, I plan to develop a test for analyzing AEI data when linkage phase is known, since some of the current AEI studies collected phased data [Serre et al., 2008; Ge et al., 2009]. This test will also be useful for providing an upper bound on the power of tests where phase is unknown and allow comparison of the relative merits of phase known and unknown data. In addition, I plan to study the asymptotic and empirical distributions of the likelihood ratio test (LRT) for the mixture model based test to assess the p-value of the LRT instead of using bootstrap. This has the potential to make computation much faster and consequently make the method more attractive when applied to large scale AEI studies.

In summary, in this chapter I presented five testing procedures for association between AEI and a *cis*-acting rSNP when D' < 1 between the rSNP and a tSNP and there is no phase information. I demonstrated that when AEI is due to a single rSNP, the power of the tests is affected by multiple factors, including the LD between the rSNP and tSNP

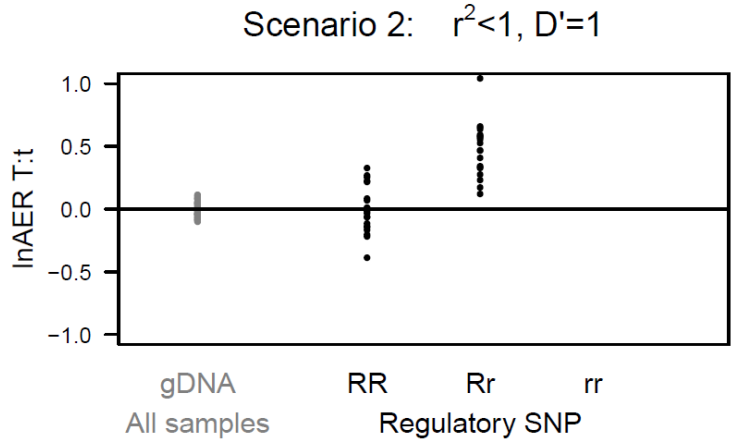
which has strong impact on the relative power, and the allele frequencies of the two SNPs, number of tSNP heterozygotes, and AEI effect size of the rSNP, which have less strong impact on the relative power. I further demonstrated that the presence of a second ungenotyped rSNP may reduce (or increase) statistical power, but does not impact type I error rate, seldom results in inconsistent tests, and tends not to modify the relative ranking of the tests. To maximize power to detect association between AEI and a *cis*-acting regulatory SNP, I recommend the use of the F test when the rSNP and tSNP are in or near linkage equilibrium (D'~0). When the two SNPs are in linkage disequilibrium, in general, the mixture-model based test is most powerful for the intermediate LD levels, and the t test is typically most powerful for high LD.

Figure 4.1: The lnAER data patterns for three different LD structures between the rSNP and tSNP.

A)




B)




C)

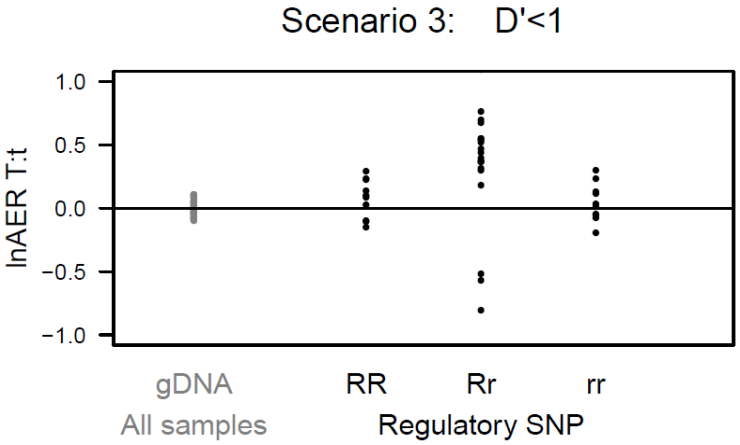


Figure 4.2: Balloon plot for expected lnAER data pattern with presence of a second ungenotyped rSNP. In this example, the two rSNPs and the tSNP are all mutually independent, and have same allele frequency $p_{R_G} = p_{R_U} = p_T = .5$. Assume effect of the genotyped rSNP on lnAER is greater than that of the ungenotyped rSNP ($\alpha_{R_G} > \alpha_{R_U}$) and the two rSNPs act additively. Center and diameter of each circle represent mean lnAER and frequency of the corresponding haplogenotype(s) to its right, respectively.

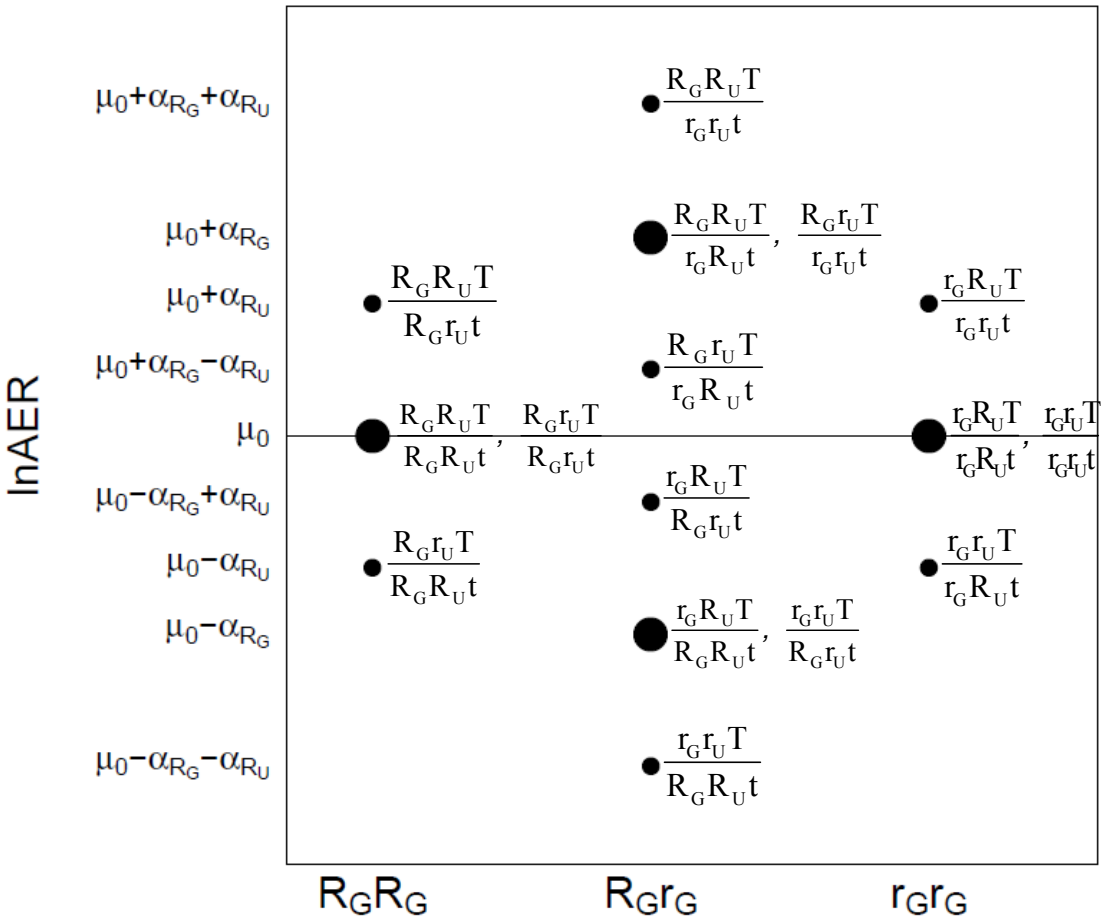
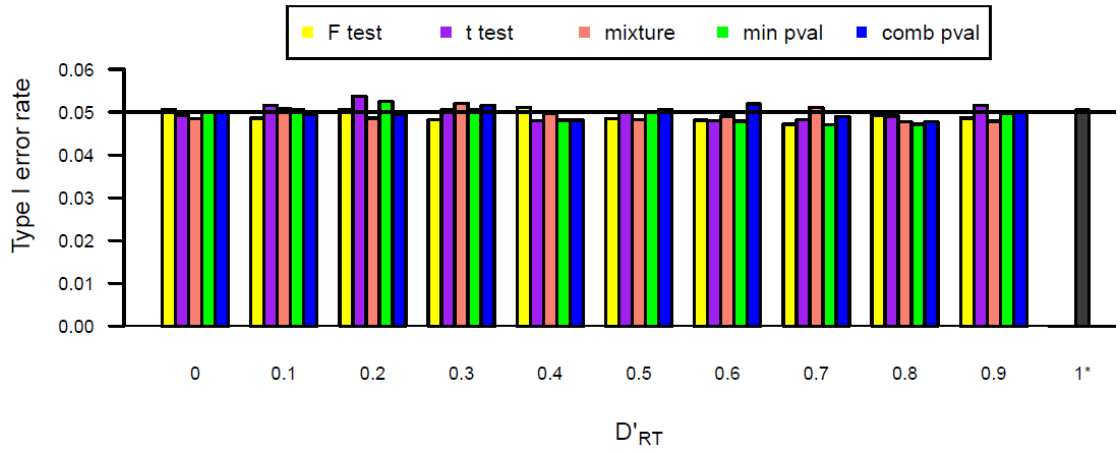
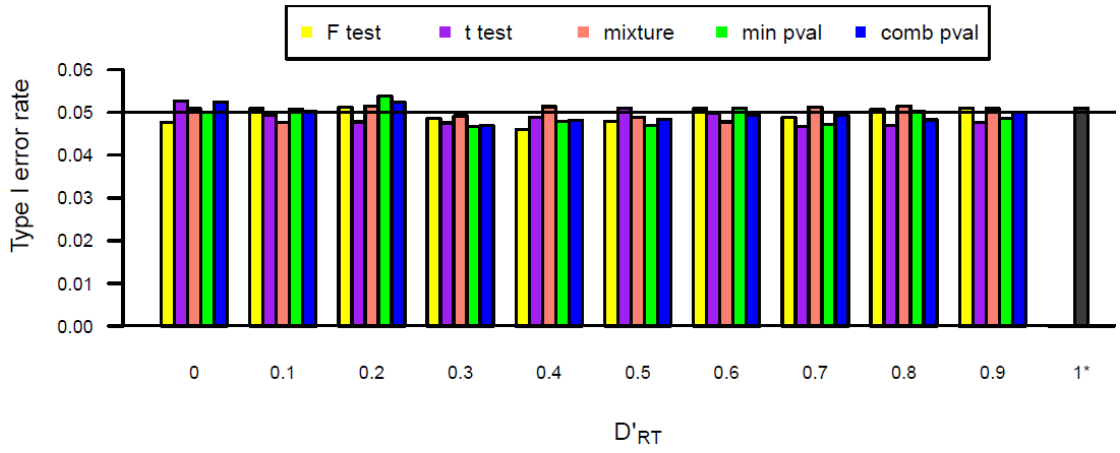


Figure 4.3: Type I error rate at significance level α = .05 for the F, t, mixture-model based, minimum-p-value and combined-p-value tests. A) N = 50 tSNP heterozygotes, frequency of the rSNP expression-increasing allele $p_R$ = .3 and B) N = 100, $p_R$ = .3. The tSNP allele frequency $p_T$ = .3.

A)




B)




P-value test is estimated using 1000 permutations for the F, t, minimum-p-value and combined-p-value, and 1000 bootstraps for the mixture-model based test; type I error rate for each test is calculated based on 10000 simulation replicates.
*: t test to compare the mean of lnAER of the cDNA to that of the corresponding gDNA, assuming that for the gDNA, the mean and variance of logarithm ratio are equal to that of the cDNA for rSNP homozygotes.
Same for Figure 4.4 - 4.11.

Figure 4.4: Impact of LD between the rSNP and tSNP on power at significance level α = .05 for the tests to detect association between AEI and the rSNP. N = 100 tSNP heterozygotes. The frequency of the rSNP expression-increasing allele $p_R$ = .3 and tSNP T allele $p_T$ = .3. Effect size of the rSNP on lnAER $\alpha_R$ = 0.8 with variance $\sigma^2$ = 1. Mean lnAER in rSNP homozygotes $\mu_0$ = 0.

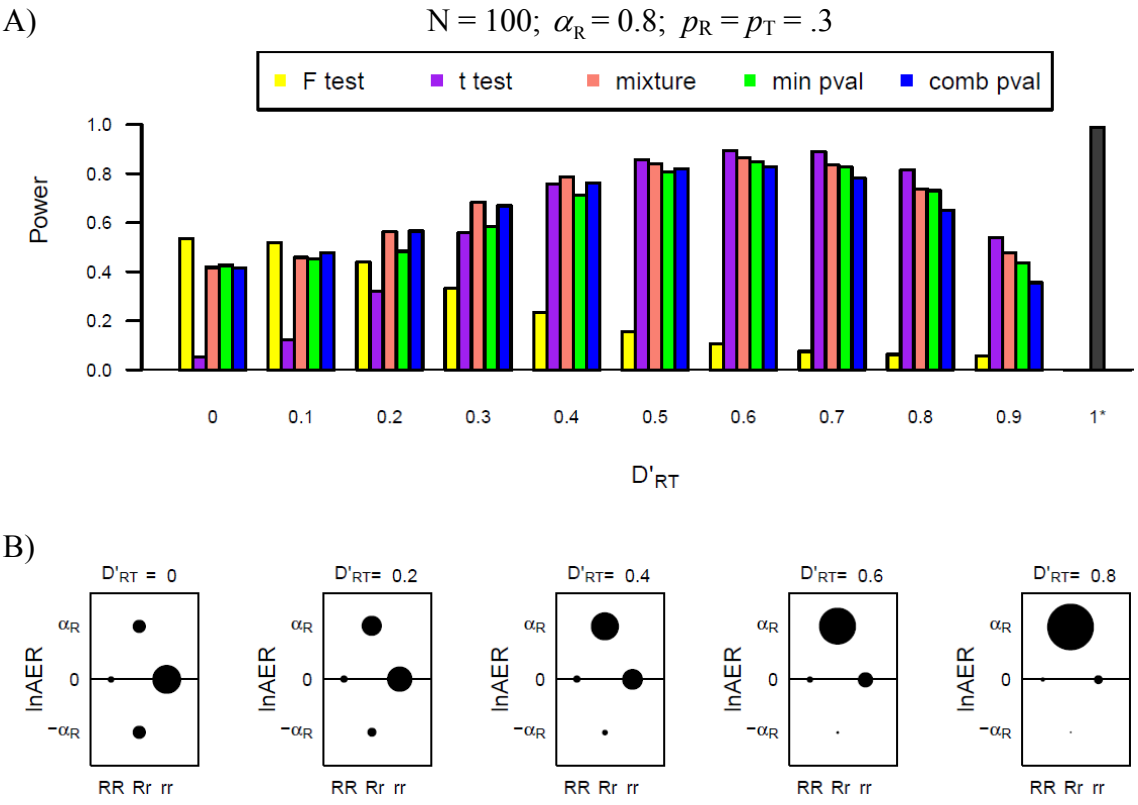
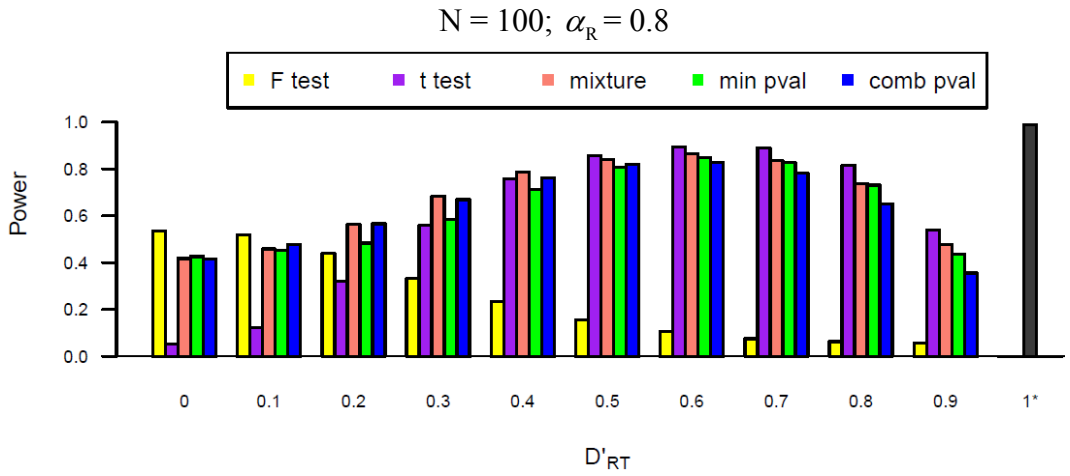
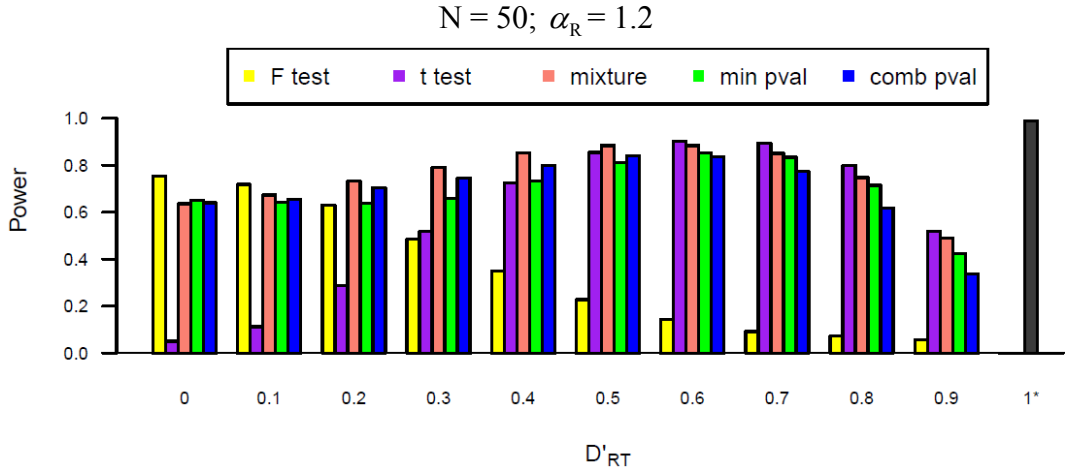


Figure 4.5: Impact of number of tSNP heterozygotes N and AEI effect size $\alpha_R$ of the rSNP on the power of the tests at significance level α = .05. A) N = 100, $\alpha_R$ = 0.8; B) N = 50, $\alpha_R$ = 1.2; and C) N = 500, $\alpha_R$ = 0.3. The frequencies of the rSNP and tSNP are $p_R = p_T$ = .3. The variance of lnAER $\sigma^2$ = 1. Mean lnAER in rSNP homozygotes $\mu_0$ = 0.

A)




B)




C)

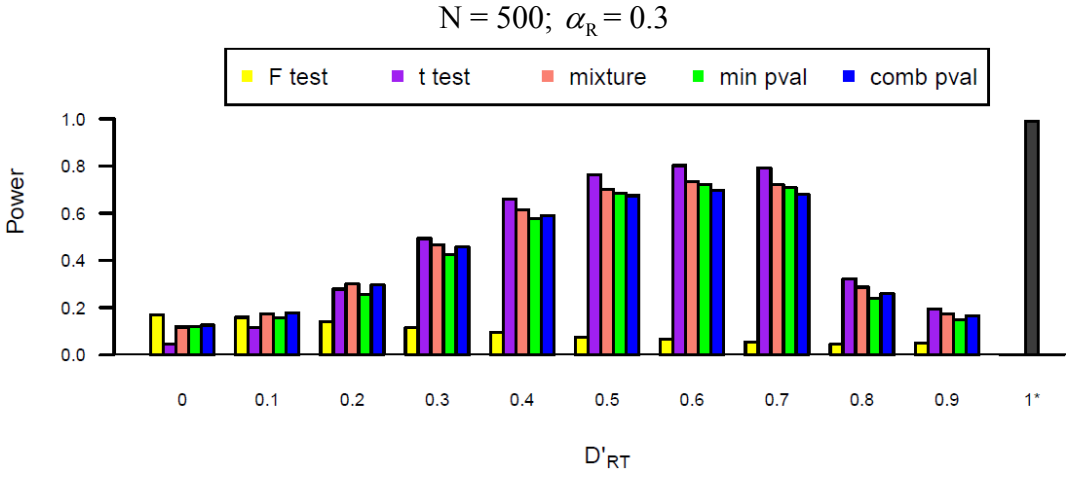


Figure 4.6: Impact of the frequency $p_R$ of the rSNP expression-increasing allele on the power of the tests at significance level α = .05. A) $p_R$ = .3; and B) $p_R$ = .1. N = 100 tSNP heterozygotes with tSNP allele frequency $p_T$ = .3. Effect size of the rSNP on lnAER $\alpha_R$ = .8 with variance $\sigma^2$ = 1. Mean lnAER in rSNP homozygotes $\mu_0$ = 0.

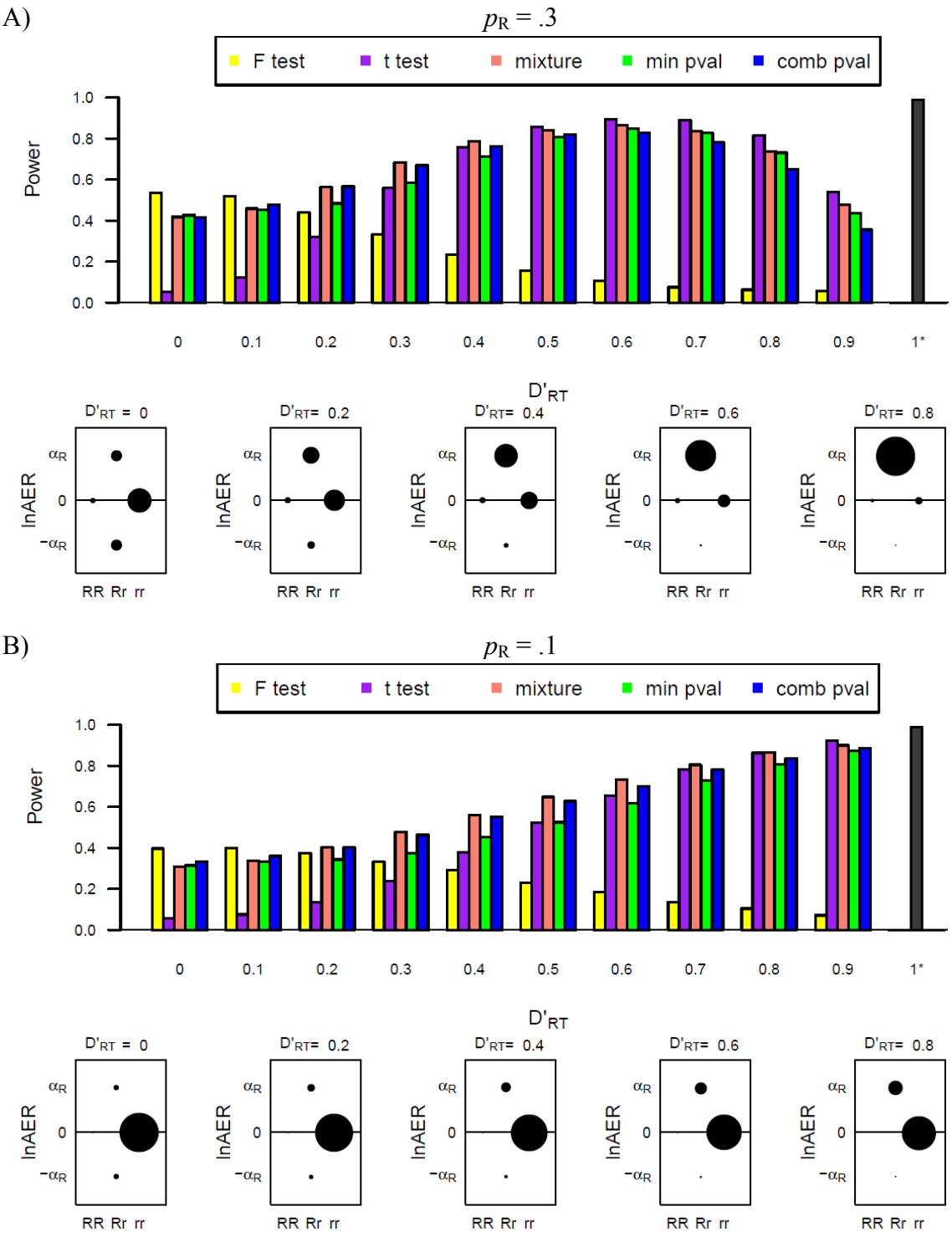
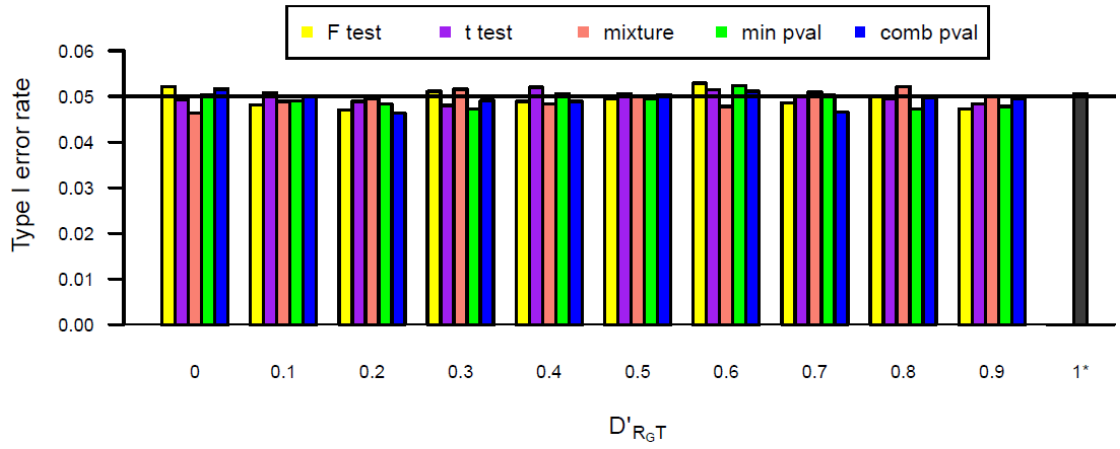
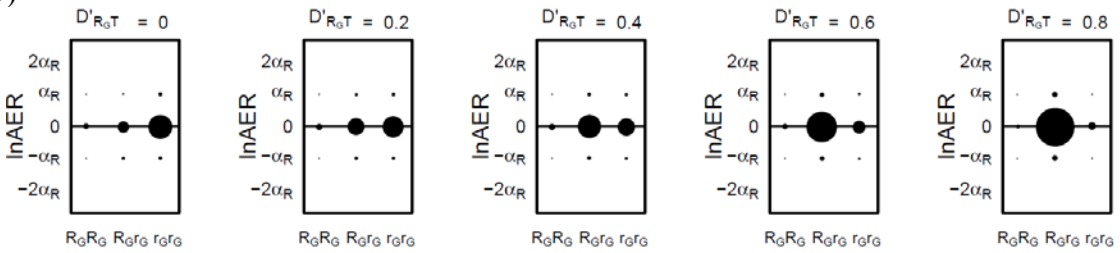


Figure 4.7: When a second ungenotyped rSNP independent from the genotyped putative rSNP and the tSNP ( $D'_{R_G R_U} = D'_{R_U T} = 0$), type I error rate of the tests to detect association between AEI and the genotyped rSNP.

A)




B)

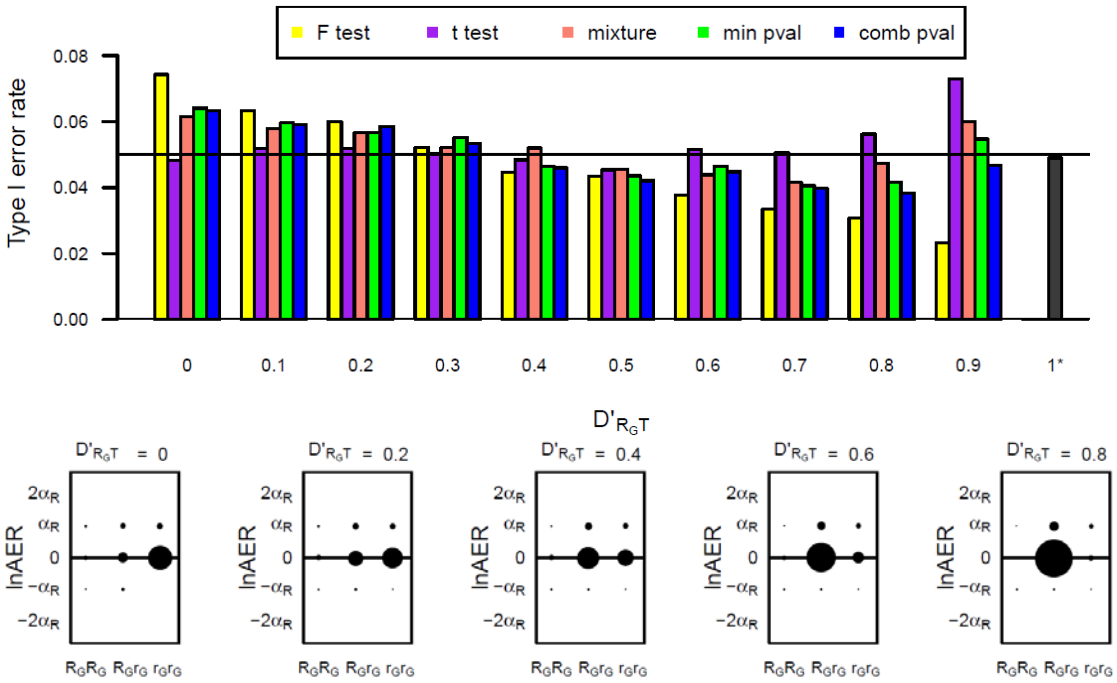



Significance level α = .05. N = 100 tSNP heterozygotes with allele frequency $p_T$ = .3. Allele frequencies for the genotyped rSNP $p_{R_G}$ = .3 and ungenotyped rSNPs $p_{R_U}$ = .1. Effect size of the ungenotyped rSNP on lnAER $\alpha_{R_U} = \alpha_R$ = .8 with variance $\sigma^2$ = 1. Mean lnAER in rSNP homozygotes $\mu_0 = 0$.

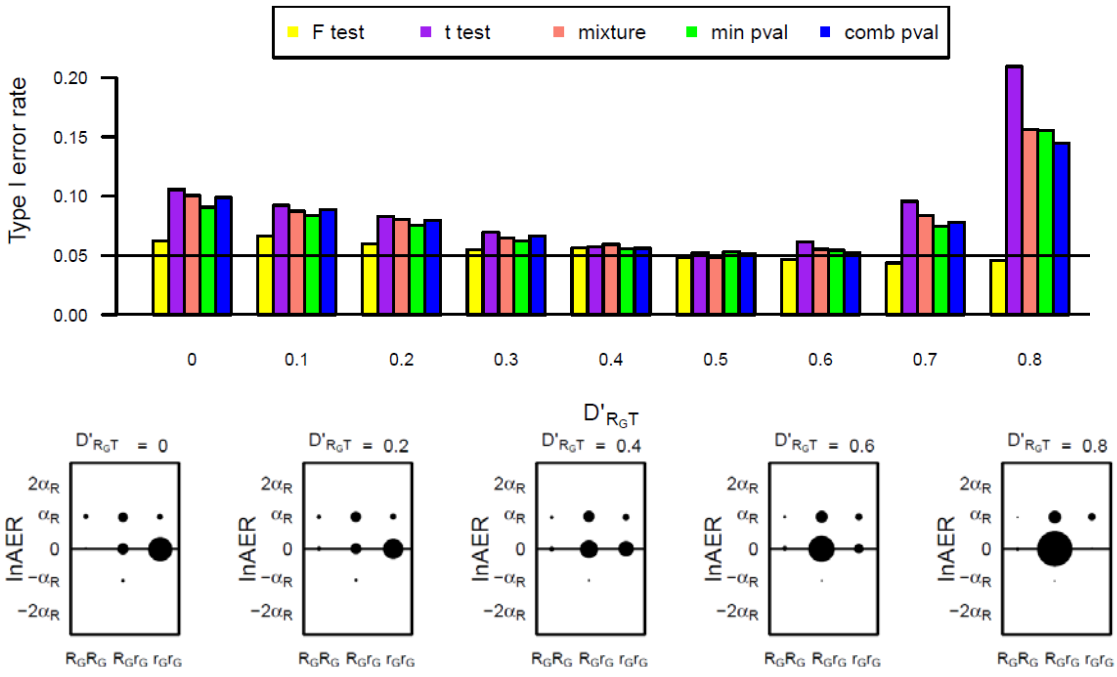
Same for Figure 4.8.

Figure 4.8: When a second ungenotyped rSNP in LD with the genotyped putative rSNP and the tSNP, type I error rate of the tests to detect association between AEI and the genotyped rSNP. A) $D'_{R_GR_U} = D'_{R_UT} = .5$ and B) $D'_{R_GR_U} = .5$, $D'_{R_UT} = 1$.

A)



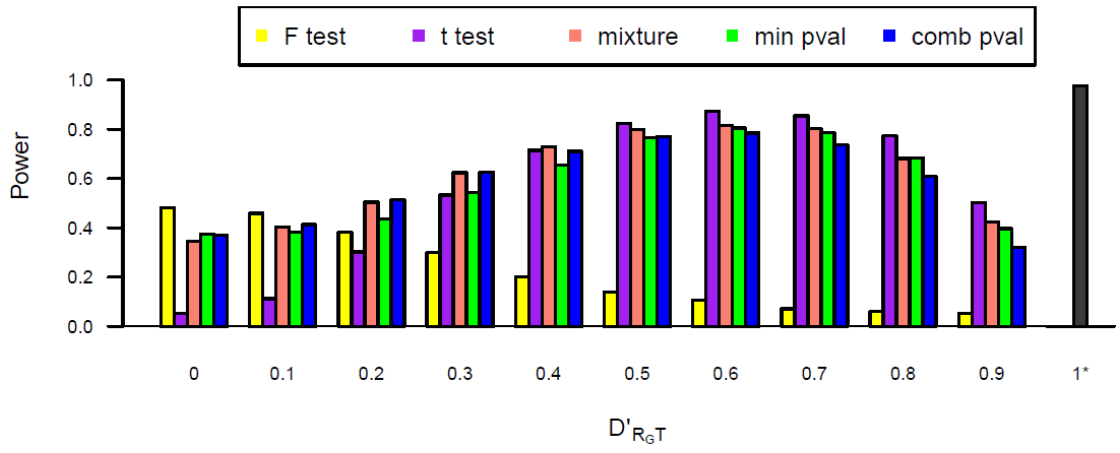

B)*



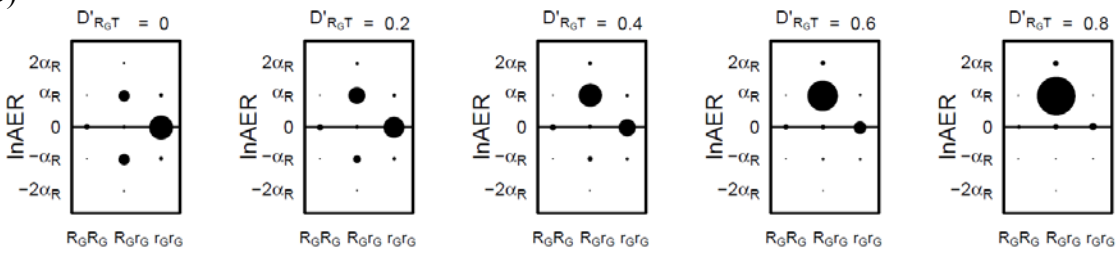

*: $D'_{R_GT}$ maximizes at the level of .8 under the combination of the allele frequencies, pairwise LD and third order LD considered. Same for Figure 4.10B.

Figure 4.9: When a second ungenotyped rSNP independent from the genotyped putative rSNP and the tSNP ( $D'_{R_G R_U} = D'_{R_U T} = 0$), power of the tests to detect association between AEI and the genotyped rSNP.

A)




B)

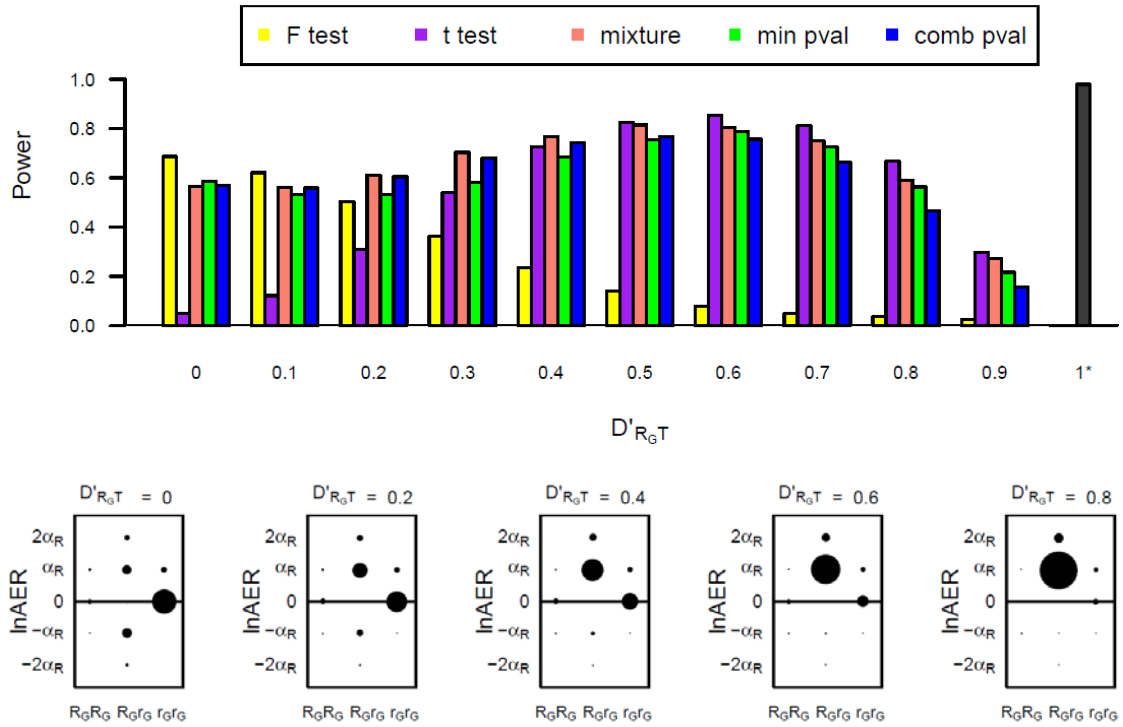



Significance level α = .05. N = 100 tSNP heterozygotes with allele frequency $p_T = .3$. Allele frequencies for the genotyped rSNP $p_{R_G} = .3$ and ungenotyped rSNPs $p_{R_U} = .1$. The two rSNPs have equal effect size on lnAER $\alpha_{R_G} = \alpha_{R_U} = \alpha_R = .8$ with variance $\sigma^2 = 1$, and act additively with same direction on gene expression. Mean lnAER in rSNP homozygotes $\mu_0 = 0$.

Same for Figure 4.10.

Figure 4.10: When a second ungenotyped rSNP in LD with the genotyped putative rSNP and the tSNP, power of the tests to detect association between AEI and the genotyped rSNP. A) $D'_{R_GR_U} = D'_{R_UT} = .5$; and B) $D'_{R_GR_U} = .5$, $D'_{R_UT} = 1$.

A)




B)

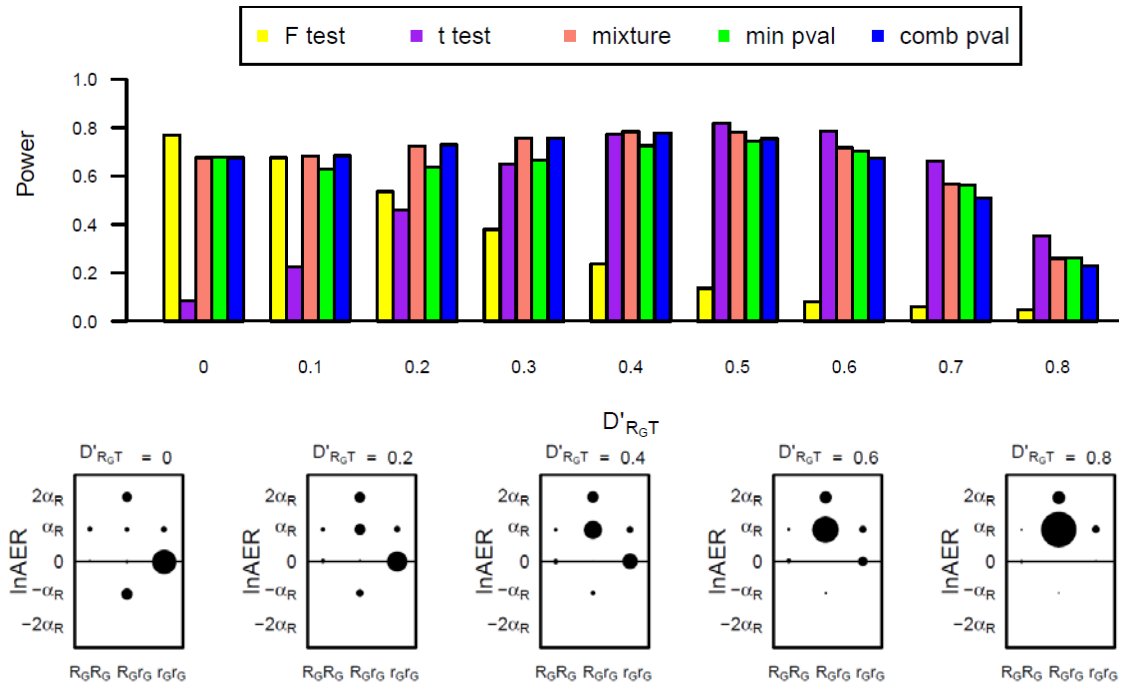
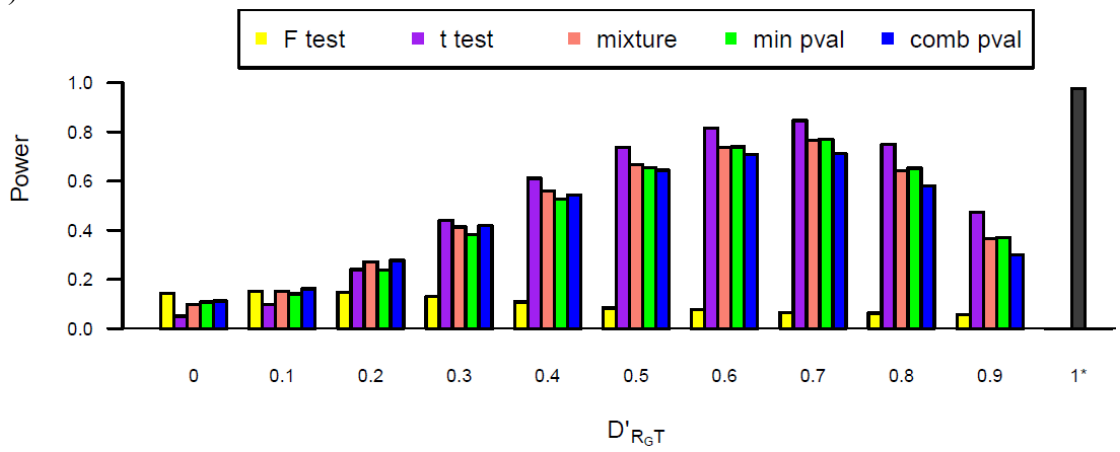
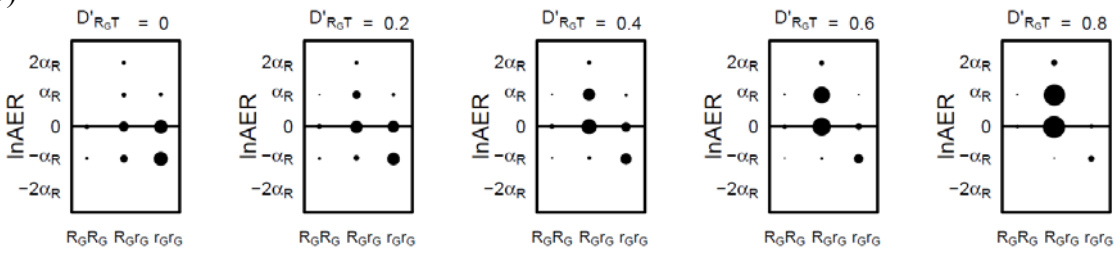


Figure 4.11: When a second ungenotyped rSNP with opposite regulation direction as the genotyped rSNP, power of the tests at significance level α = .05. LD between the ungenotyped rSNP with the genotyped rSNP and the tSNP $D'_{R_G R_U} = D'_{R_U T} = .5$, and third order LD $D'_{R_G R_U T} = 0$. Allele frequencies for the genotyped rSNP and the tSNP $p_{R_G} = p_T = .3$, and ungenotyped rSNP $p_{R_U} = .1$. N = 100 tSNP heterozygotes. The two rSNPs have equal effect size on lnAER $\alpha_{R_G} = \alpha_{R_U} = \alpha_R = .8$ with variance $\sigma^2 = 1$. Mean lnAER in rSNP homozygotes $\mu_0 = 0$.

A)




B)



# References


Alachkar H, Kataki M, Scharre DW, Audrey Papp A, Sadee W. 2008. Allelic mRNA expression of sortilin-1 (*SORL1*) mRNA in Alzheimer's autopsy brain tissues. Neurosci Let 448: 120-124.

Bennett JH. 1954. One the theory of random mating. Ann Eugenics 18: 311-317.

Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. 2003. *Cis*-acting variation in the expression of a high proportion of genes in human brain. Hum Genet 113: 149-153.

Buckland PR. 2004. Allele-specific gene expression differences in humans. Hum Mol Genet 13: R255-R260.

Campino S, Forton J, Raj S, Mohr B, Auburn S, Fry A, Mangano VD, Vandiedonck C, Richardson A, Rockett K, Clark TG, Kwiatkowski DP. 2008. Validating discovered *cis*-acting regulatory genetic variants: application of an allele specific expression approach to HapMap populations. PLoS ONE 3: e4105.

Chen H, Chen J. 2001. Large sample distribution of the likelihood ratio test for normal mixtures. Canad J Statist 29: 201-216.

Chen H, Chen J, Kalbfleisch JD. 2001. A modified likelihood ratio test for homogeneity in finite mixture models. J Roy Statist Soc B63: 19-29.

Chen J. 1998. Penalized likelihood ratio test for finite mixture models with multinomial observations. Canad J Statist 26: 583-599.

Chen J, Kalbfleisch JD. 1996. Penalized minimum-distance estimates in finite mixture models. Canad J Statist 24: 167-175.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365-1369.

Dempster AP, Laird NM, Rubin DB.1977. Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc B 39: 1–38.

Deodato F, Boenzi S, Santorelli FM, Dionisi-Vici C. 2006. Methylmalonic and propionic aciduria. Am J Med Genet C Semin Med Genet142C: 104-112.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KCL, Gagné, Dias J, Hoberman R, Montpetit A, Joly M-M, Harvey EJ, Sinnet D, Beaulieu P, Hamon R, Graziani A, Dewar K, Harmsen E, Majewski J, Goring HHH, Naumova AK, Blanchette M, Gunderson KL, Pastinen T. 2009. Global

patterns of *cis* variation in human cells revealed by high-density allelic expression analysis. Nat Genet 41: 1216-1222.

Goldstein JL, Brown MS. 1990. Regulation of the mevalonate pathway. Nature 343: 425-430.

Hartigan JA. 1985. A failure of likelihood asymptotics for normal mixtures. In: LeCam L, Olshen RA. Proceedings of the Berk Conference in Honor of J. Neyman and J. Kiefer. Vol. 2: 807-810.

Heighway J, Bowers NL, Smith S, Betticher DC, Santibánez Koref MF. 2005. The use of allelic expression differences to ascertain functional polymorphisms acting in *cis*: analysis of *MMP1* transcripts in normal lung tissue. Ann Hum Genet 69: 127-133.

Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A catalog of published genome-wide association studies. www.genome.gov/gwastudies.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106: 9362-9367.

Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T, Voight BF, Bonnycastle LL, Jackson AU, Crawford G, Surti A, Guiducci C, Burtt NP, Parish S, Clarke R, Zelenika D, Kubalanza KA, Morken MA, Scott LJ, Stringham HM, Galan P, Swift AJ, Kuusisto J, Bergman RN, Sundvall J, Laakso M, Ferrucci L, Scheet P, Sanna S, Uda M, Yang Q, Lunetta KL, Dupuis J, de Bakker PI, O'Donnell CJ, Chambers JC, Kooner JS, Hercberg S, Meneton P, Lakatta EG, Scuteri A, Schlessinger D, Tuomilehto J, Collins FS, Groop L, Altshuler D, Collins R, Lathrop GM, Melander O, Salomaa V, Peltonen L, Orho-Melander M, Ordovas JM, Boehnke M, Abecasis GR, Mohlke KL, Cupples LA. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41: 56-65.

Mahr S, Burmester GR, Hilke D, Gobel U, Grutzkau A, Haupl T, Hauschild M, Koczan D, Krenn V. et al. 2006. *Cis*- and *trans*-acting gene regulation is associated with osteoarthritis. Am J Hum Genet 78: 793-803.

McLachlan GJ. 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl Statist 36: 318-324.

Mosteller F, Fisher RA. 1948. Combining independent tests of significance. Am Statist 2: 30-31.

Nelder JA, Mead R. 1965. A simplex method for function minimization. Computer J 7: 308-313.

Nielson DM, Ehm MG, Zaykin DV, Weir BS. 2004. Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics 168: 1029-1040.

Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H, Beck A, Verner A, Kingsley J, Harmsen E, Labuda D, Morgan K, Vohl MC, Naumova AK, Sinnett D, Hudson TJ. 2003. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics 16: 184-193.

Pastinen T, Ge B, Hudson TJ. 2006. Influence of human genome polymorphism on gene expression. Hum Mol Genet 15 Spec No 1: R9-16.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. Nat Rev Genet 7: 862-872.

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan J-B, Hudson TJ. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic *cis*-acting mechanisms regulating gene expression. PLoS Genet 4: e1000006.

Tao H, Cox DR, Frazer KA. 2006. Allele-specific *KRT1* expression is a complex trait. PLoS Genet 2: e93.

Teare MD, Heighway J, Santibánez Koref MF. 2006. An exprectation-maximization algorithm for the analysis of allelic expression imbalance. Am J Hum Genet 79: 539-543.

The International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431: 915-916.

Thomson G, Baur MP. Third order linkage disequilibrium. Tissue Antigens 24: 250-255.

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.

# CHAPTER 5

# CONCLUSION AND FUTURE WORK

For the past three years, the genomewide association studies (GWAS) have rapidly grown in scale and complexity**,** and have provided new insights into complex disease genetics. For loci identified by GWAS, investigators are interested in genetic effect size to help understand the genetic contribution of these loci to the disease risk or trait variation, and moreover, to provide information for designing follow-up studies. Following GWAS, testing for association between gene expression and identified SNPs has the potential to help understand the relationship between these SNPs with the trait, and identify the gene(s) and variants most likely to influence the trait in identified regions that include multiple genes. In this dissertation, I have presented statistical methods to correct for the winner's curse in GWAS and so achieve more accurate estimation of genetic effect size. In addition, I have proposed testing procedures for using the allelic expression imbalance (AEI) to detect *cis*-acting regulatory SNPs.

In Chapter 2 and 3, I studied the impact of winner's curse in the context of genetic case-control and quantitative trait (QT) association studies by analytically quantifying the upward bias for estimators of the genetic effect size. I also proposed ascertainment-corrected maximum likelihood methods to reduce the bias of the estimators.

In Chapter 2, I focused on the estimates of the allele frequency difference and odds ratio (OR) in case-control association studies, measures often used to quantify the strength of the genetic effect in such studies. I showed that in realistic situations, these uncorrected (naïve) estimators can be substantially overestimated, and that the overestimation decreases as power increases. I demonstrated that in the typical power range for most large-scale genetic association studies, the ascertainment-corrected estimators result in reduced absolute bias compared to the naïve uncorrected estimators. I further extended these calculations to two-stage association studies, and found that for optimal two-stage designs [Skol et al., 2007], results are similar to those for the corresponding one-stage designs. [Xiao and Boehnke, 2009]

In Chapter 3, I presented an extension of the winner's curse study in SNP-QT association studies, in which the genetic effect size is parameterized as the linear regression slope. My analytical calculation again demonstrated that the overestimation in the regression slope estimate decreases as power increases. To reduce the ascertainment bias, I proposed a three-parameter maximum likelihood method and also a simplified one-parameter model with the nuisance parameters excluded based on the asymptotic property of the linear regression model. I showed that both likelihood methods reduce the bias when power to detect association is low or moderate, and the one-parameter model generally results in a slope estimator with smaller variance.

I found that, as with other methods [Sun and Bull, 2005; Zöllner and Pritchard, 2007; Zhong and Prentice, 2008], the ascertainment-corrected estimators in both case-control and QT association studies tend to underestimate the effect size, in contrast to the naïve estimators which result in overestimation. Zhong and Prentice [2008] proposed a

mean-square-error (MSE) weighted estimator for the OR in logistic regression model, which is a linear combination of the naïve and corrected estimators. This MSE weighted estimator improves the bias-correction when study power is high. In an effort better to address the overcorrection problem in winner's cures studies, for future work, I plan to use an empirical Bayes method [Carlin and Louis, 2000] utilizing information from GWAS to help define a prior distribution for the genetic effect size.

In Chapter 4, I described testing procedures for using AEI to detect *cis*-acting regulatory SNPs (rSNP), focusing on the situation when the rSNP and a transcribed SNP (tSNP) are in incomplete linkage disequilibrium (LD) and there is no phase information for the two SNPs. I initially assumed that the AEI is due to a single rSNP, and modeled the AEI data as a mixture of normal distributions depending on the haplogenotype of the rSNP and tSNP. I proposed simple t and F tests which ignore the nature of the mixing distribution, and also a mixture-model based test which incorporates this nature. My simulations showed that the type I error rates for all tests are well controlled, and the relative power of the tests depends on the LD between the rSNP and tSNP, allele frequencies of the SNPs, AEI effect size of the rSNP, and number of tSNP heterozygotes. I further investigated how sensitive these tests are to the violation of the single-rSNP assumption, and found that a second ungenotyped rSNP may reduce power of the tests but almost never invalidates the proposed tests nor substantially changes the rankings of the tests for a given level of LD between the genotyped rSNP and the tSNP.

For the mixture-model based test in Chapter 4, I estimated the frequencies of the two rSNP heterozygous haplogenotypes by the mixing proportion $\pi$ of the mixture model. In practice, the two haplogenotype frequencies could be estimated based on additional

genotyped SNPs surrounding the rSNP or the tSNP, or an external source of phased data, for example, the HapMap data. In future work, I will consider an extension of the AEI analysis in a Bayesian framework, which incorporates this prior information on $\pi$ to help infer the haplogenotype frequencies. I also plan to develop methods combining the AEI with the total expression level of the two alleles of the gene to infer jointly the association strength between the rSNP and the candidate gene.

Until recently AEI was measured by low-throughput procedures [e.g., Melani et al., 2007], which has limited the application of this method. With the development of next-generation sequencing technologies and the ongoing 1000 Genomes Project, high throughput allelic expression data has begun to be generated [Ge et al., 2009] and more will be available in the near future. Chapter 4 provides a quantitative framework for using the AEI to identify the potential regulatory variants identified in large scale sequencing studies, and ultimately, may provide new insight into the relationships between gene regulation and disease.

# References

Carlin BP, Louis TA. 2000. Bayes and empirical Bayes methods for data analysis, 2nd ed. CRC Press, Boca Raton, FL.

Milani L, Gupta M, Andersen M, Dhar S, Fryknäs M, Isaksson A, Larsson R, Syvänen AC. 2007. Allelic imbalance in gene expression as a guide to *cis*-acting regulatory single nucleotide polymorphisms in cancer cell. Nucleic Acids Res 35(5): e34.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2007. Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31: 776-788.

Sun L, Bull SB. 2005. Reduction of selection bias in genome-wide studies by resampling. Genet Epidemiol 28: 352-367.

Xiao R, Boehnke M. 2009. Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol 33: 453-462.

Zhong H, Prentice RL. 2008. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics 94(4): 621-634.

Zöllner S, Pritchard JK. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80: 605-615.