Methods for large-scale genetic association studies

by

Karen N. Conneely

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2008

Doctoral Committee:

Professor Michael L. Boehnke, Chair
Associate Professor Gonçalo Abecasis
Associate Professor Kerby A. Shedden
Assistant Professor Sebastian K. Zoellner

To my grandparents,
who carried their genes across great distance,
and to my parents, who gave them to me,
that I might one day treasure them
and marvel at their workings.

ACKNOWLEDGEMENTS

Five short years ago (ok, six-and-a-half long years ago) I couldn't believe my good fortune at being able to start as a graduate student in the statistical genetics group at the University of Michigan. Today, I can't believe my continuing luck as I prepare to start as a professor in the genetics department at Emory University. I suspect that like many temporal variables, luck is autoregressive, and further conjecture that in this case the first stroke of luck which landed me at Michigan can completely explain the second.

Perhaps even harder for me to believe is that the process of writing this dissertation is finally coming to an end. This could not have happened without the help and contributions of many. I am most indebted to my advisor, Dr. Michael Boehnke, who was unfailingly supportive, patient, and generous with both his ideas and his time. His guidance and flexibility gave me the direction and freedom I needed to develop as a researcher, and his example showed me the type of researcher and mentor I want to be. My committee members, Drs. Gonçalo Abecasis, Kerby Shedden, and Sebastian Zoellner, were amazingly constructive in advice and accommodating in schedule. All four committee members as well as Cristen Willer and Laura Scott asked insightful and critical questions that improved many aspects of this work. Andrew Skol, Jeanette Mumford, and Kevin Cheek uncomplainingly answered myriad questions about genetics, statistics, and computing, often late at night. The investigators of the Finland-United States Investigation of NIDDM (FUSION) study generously allowed my use of FUSION data to illustrate my methods. Grant T32 HG00040 from the National Institutes of Health

and a University of Michigan Rackham Predoctoral Fellowship funded my work on this dissertation. And finally, Barry Hall provided not only copious emotional and alimentary support, but a sense of calm that kept me sane through much of this process and him heroically tolerant for the remainder.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

With the ever-increasing availability and decreasing costs of high-throughput SNP genotyping, contemporary genetic association studies now incorporate more information than anyone would have imagined even five years ago. While SNP genotyping was once performed sparingly and painstakingly for highly localized genetic regions harboring candidate genes or strong linkage results, it is now performed automatically, abundantly, and, increasingly, genome-wide.

While statistical geneticists once focused their efforts on elegant methods for extracting the maximum amount of information from sparse genotype data, these efforts have shifted. As genotyping technology soars ahead, the problem of sparse data has been all but eliminated, yet the demand for statistical methods remains strong. Even with the high density of current SNP chips, methods to extrapolate from this information to infer genotypes for even more SNPs are under development (Li et al. 2007, Marchini et al. 2007). On the other hand, the very richness of the available data presents new problems which must be addressed, in particular the still unresolved problem of how to extract the right information and draw meaningful conclusions from the almost overwhelming amount of data.

The problem of multiple testing is well-known and methods for adjustment predate most of us (Bonferroni 1936). Until recently, however, the available methods for multiple-testing adjustment required the assumption of independent tests, an assumption that has become less and less realistic as the density of SNP genotype data has increased. In Chapter 2, I address the problem of adjusting for multiple tests in genetic association studies where substantial correlation between tests is the norm. Recent approaches to this problem have enlisted simulations or permutations of the data to approximate the behavior of correlated test statistics under the null hypothesis of no genetic association. I present a new approach that attains the same accuracy as these methods but requires much less computation and is orders of magnitude faster. Seaman and Müller-Myhsok (2005) have shown that when multiple SNPs are tested for association, the score statistics have a multivariate normal distribution in the common case where association tests are based on generalized linear models. I show that this is also true when testing multiple SNPs for association with multiple traits under multiple genetic models. I derive the appropriate covariance matrix and present a method for computing multiple-testing-adjusted *P*-values directly from the multivariate normal distribution, rather than through simulation. This method achieves the target type I error rate in a variety of simulations, and demonstrates a nearly one-to-one relationship with permutation-based *P*-values computed in the course of a large candidate gene analysis (Gaulton et al. 2007) performed as part of the Finland-United States Investigation of NIDDM Genetics (FUSION) study (Valle et al. 1998).

As technological advances have made it possible to genotype many SNPs, follow-up of interesting results in independent samples has become a more feasible option. It has also become less of an option and more of an obligation, since with the large number of tests in a typical study, even extremely significant results should be treated with caution. Methods for combining results from multiple studies are well-established (for example, Mantel and Haenszel 1959, Mantel 1963), but meta-analyses based on genetic association tests are subject to the same problem described above and will require adjustments for multiple correlated tests.

In Chapter 3, I extend the method presented in Chapter 2 to adjust meta-analyses involving correlated association tests for multiple testing. I present variants of the method to address study design issues present in most meta-analyses, such as data missing-at-random, selection of only the best results for follow-up, or two-stage design, where only results passing a pre-designated significance criteria are followed up in consecutive samples. Simulations based on haplotype data collected as part of a five-sample meta-analysis demonstrate that even when the number of tests is large and correlation between tests is substantial, these methods can provide accurate control of the type I error rate for meta-analyses in a variety of settings.

With the recent strides in genotyping technology, it is likely that the quality of genotype data has continued to improve as well. However, detection of errors in the data remains difficult, especially for case-control studies involving unrelated individuals. The extent to which genotype error influences the power and validity of association tests, and how much error may be tolerated, remain important questions in the context of both case-control and family-based association testing. A related issue is missing genotype data. Since data missing at the individual level are often the result of no-call procedures designed to weed out potential genotype errors, they are unlikely to be missing-at-random, which means that the available non-missing data may not be an accurate representation of the true distribution of genotypes.

In Chapter 4, I address these related issues of genotype error and missing genotype data, with particular attention to the fact that both errors and missing genotypes are likely to occur at differential rates depending on an individual's true genotype. Common problems such as the presence of unknown variants in the primer region have been observed to affect most major genotyping platforms (Koboldt et al. 2006), and may lead to the misclassification of heterozygous genotypes as homozygous genotypes due to failure of one allele to amplify, as well as to the loss of genotypes due to failure of both alleles to amplify. As the least frequent genotypes, minor allele homozygotes may be more likely to be misclassified or classified as no-calls by genotype-calling algorithms, since precision is lower for these genotypes. Using replicate Sequenom data collected as part of the FUSION study and replicate unfiltered HapMap data created as part of the

HapMap quality control exercises (International HapMap Consortium 2005, Online Supplement), I assess the degree of differential rates of genotype error and missing data across a large number of genotyping platforms. I find strong evidence for both differential error and loss of genotypes, and find that the extent and patterns vary considerably across platforms. To investigate the tolerance of case-control and family-based association tests for incomplete and lower-quality data, I perform simulated association tests in which genotype errors and missing genotypes are sampled based on the differential rates of error and missing data observed in initial genotyping attempts in the Sequenom data. I find that for case-control association tests, although the test for equal allele frequencies is quite sensitive to incomplete data, the Cochran-Armitage test for trend (Cochran 1954; Armitage 1955) is remarkably robust to all levels of incomplete data, and the main reason to resolve incomplete data in this case is to avoid power loss. For family-based association tests, however, I find that the transmission/disequilibrium test (TDT) was highly sensitive to data quality, and that the rate of type I error more than doubled when only 5-10% of individuals had missing genotypes. Finally, I assess the effect of the observed distribution of genotype errors and missing data across SNPs on a genome-wide association study, where extremely bad data for a single SNP could potentially alter the conclusions of the entire study. I find that for the allele frequency test and especially the TDT, the expected study-wide false positive rate in a genome-wide study is inflated due to genotype error and missing data. The target type I error rate can be obtained with the TDT if extremely stringent quality control measures are implemented. I conclude by recommending 1) the use of trend tests in place of the allele frequency test, and 2) future work on appropriate levels of quality control for the TDT, or alternatively the use of a model which accounts for differential rates of genotype error and missing data.

CHAPTER 2

SO MANY CORRELATED TESTS, SO LITTLE TIME! RAPID ADJUSTMENT OF *P*-VALUES FOR MULTIPLE CORRELATED TESTS

This chapter has been published with the same title in the American Journal of Human Genetics (2007, v. 81, pp 1158-1168).

Contemporary genetic association studies may test hundreds of thousands of genetic variants for association, often with multiple binary and continuous traits or under more than one model of inheritance. Many of these association tests may be correlated with one another due to linkage disequilibrium (LD) between nearby markers and correlation between traits and models. Permutation tests and simulation-based methods are often employed to adjust groups of correlated tests for multiple testing, since conventional methods such as Bonferroni correction are overly conservative when tests are correlated. We present here a method of computing *P*-values Adjusted for Correlated Tests (P_{ACT}) that attains the accuracy of permutation or simulation-based tests in much less computation time, and show that our method applies to many common association tests based on multiple traits, markers, and genetic models. Simulation demonstrates that P_{ACT} attains the power of permutation testing and provides a valid adjustment for hundreds of correlated association tests. In data analyzed as part of the Finland-United States Investigation of NIDDM Genetics (FUSION) study (Valle et al. 1998), we observe a near one-to-one relationship ($r^2 > .999$) between P_{ACT} and the corresponding permutation-based *P*-values, achieving the same precision as permutation testing but thousands of times faster.

2.1 Introduction

Improvements in genotyping technology and the accompanying reductions in genotyping cost have led to an unprecedented wealth of genetic data to analyze. In genome-wide association (GWA) studies, it has become routine to genotype hundreds of thousands of SNP markers. Even candidate gene studies may now involve hundreds or thousands of SNPs. Studies may test multiple binary and continuous outcome variables for genetic association – for example, one or more diseases and a set of disease-related quantitative traits. It is also possible to test each SNP for association in several ways – for example, by allowing competing models of inheritance when the true model is unknown. The ability to perform so many tests brings with it a greater potential than ever before to identify disease-predisposing variants, but also a new set of issues regarding the most efficient way to use the available information.

An important issue affecting large-scale association analyses is how best to adjust for multiple testing, given the likely correlation between many of the tests. With the density of SNPs in contemporary candidate gene and GWA studies, linkage disequilibrium ensures that there often will be correlation between tests performed on nearby SNPs. Additionally, phenotypic traits collected for a particular study are likely to be correlated, and tests based on different models of inheritance such as the recessive and dominant model will certainly be correlated. A danger of using traditional methods such as Bonferroni correction in this context is that truly interesting findings may be rendered insignificant by an overly severe correction.

For L independent tests with a pre-set significance level, α, approximately αL of the tests will appear significant by chance alone. Without adjustment for multiple testing, the expected Type I error rate for the group of tests (the probability that at least one test is significant given no true association) is $1 - (1 - \alpha)^L \approx \alpha L$, rather than α, the target type I error rate. The best P-values can be adjusted for multiple testing with the Bonferroni procedure, which effectively multiplies the best P-value (P_{min}) by L, or the more precise Šidák procedure, which computes the adjusted P-value as $1 - (1 - P_{min})^L$ and guarantees a type I error rate of α for independent tests (Šidák 1967).

While Bonferroni and Šidák adjustments are valid in the case of independent tests, they tend to be overly conservative in association studies where the tests are correlated. A valid adjustment for multiple testing must account for the correlation between tests. Permutation tests provide a valid adjustment if the data are permuted in a way that simulates the null hypothesis but maintains the original correlation structure. Randomly permuting and re-analyzing the data many times and comparing the permutation-based results to the original results allows estimation of the probability of observing a P-value as small as the original minimum, given the correlation between tests. This solution is attractive due to its simplicity and robustness, and is often considered the gold standard for analysis. However, in the context of large association studies, permutation is likely to require too much computation time, so computationally-efficient alternatives are desirable.

Some proposed alternatives have focused on extending the Bonferroni or Šidák adjustments to account for the correlation between tests. When the L tests are correlated, the true probability of observing a P-value as small as P_{min} is smaller than the Šidák estimate $1 - (1 - P_{min})^L$ because there is less variation between test statistics than if the tests were independent, making extreme test statistics less likely. In effect, it is as though fewer tests were performed; for this reason, several studies suggest replacing L in $1 - (1 - P_{min})^L$ with an estimate of the effective number of independent tests (Cheverud 2001; Nyholt 2004; Li and Ji 2005). However, the suggestion that a single parameter fully captures the correlation structure has been rejected in the majority of cases when tested on SNPs in LD (Dudbridge and Koeleman 2004; Salyakina et al. 2005). Salyakina et al. also found in simulation studies of the method of Nyholt (2004) that the "nominal 5% type I error rate varied from under 3% to over 7%" and that while this approach "may be useful as an exploratory tool, it is not an adequate substitute for permutation tests" (p. 19).

A shortcoming of methods based on an effective number of tests is that they do not account for the distribution of the test statistics. The Šidák-adjusted *P*-value has identical form regardless of distribution, which is appropriate for independent tests; however, the analogous probability for correlated tests depends on the joint distribution of the test statistics, and any valid extension of the Šidák method must take this into account. If the test statistics follow an asymptotic multivariate normal distribution, as is true for many tests, the adjusted *P*-values may be computed as multivariate normal probabilities. This strategy has previously been used in survival analysis (Wei et al. 1989; Wei and Glidden 1997) and clinical trials (James 1991) for ten or fewer correlated tests. More recently, Lin (2005a) and Seaman and Müller-Myhsok (2005) have employed this strategy in the genetics literature to adjust *P*-values from a larger number of tests. In these studies, as in permutation tests, replicates of the test statistics are simulated under the null hypothesis of no association. However, these methods achieve greater speed than permutation tests by simulating the test statistics directly from the asymptotic distribution rather than permuting and re-analyzing the entire dataset in each replicate.

Here we present an alternative method of *P*-value adjustment that attains even greater speed by avoiding the need for simulation altogether. We propose comparing the observed test statistics directly to their asymptotic distribution through numerical integration. We show that for many common association tests the joint distribution of the test statistics is multivariate normal with a simple covariance structure even for association tests involving multiple correlated traits, markers, and genetic models. We demonstrate through simulations and through analysis of data from the Finland-United States Investigation of NIDDM Genetics (FUSION) study (Valle et al. 1998) that this method attains the same accuracy as permutation tests or their simulation-based counterparts and is orders of magnitude faster than these methods.

2.2 Methods

2.2.1 *P*-values Adjusted for Correlated Tests (P_{ACT})

Consider L tests of association with test statistics $T_1, \ldots, T_L$ and P-values $P_1, \ldots, P_L$; denote the ordered P-values $P_{min} \le P_{(2)} \le P_{(3)} \le \ \ldots \ \le P_{(L)}$. It is common to focus interest on the smallest P-values. However, each individual P-value is based on a single hypothesis test that does not account for the fact that L tests were actually performed. The Šidák (1967) P-value,

$$P_{Šidák} = 1 - \left(1 - P_{min}\right)^L , \tag{2.1}$$

estimates the probability of observing at least one P-value $\le p_{min}$ under the null hypothesis for L independent tests. We suggest here an estimator of this probability for correlated tests which we denote P_{ACT} (P-value Adjusted for Correlated Tests). While $P_{Šidák}$ depends only on P_{min}, P_{ACT} is based on the joint distribution of all L statistics $T_1, \ldots, T_L$ and their correlation structure.

As we show in the next section, many common association tests are based on or related to test statistics that are asymptotically distributed as multivariate normal with known covariance matrix. We assume here that the vector of test statistics $\boldsymbol{T} \mathrel{\dot\sim} N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ where $\dot\sim$ denotes asymptotic (large sample) distribution, $\mathbf{0}$ is an L-dimensional vector of zeroes, and $\boldsymbol{\Sigma}$ is an L×L correlation matrix. Then, $P_i = 1 - \Phi\left(T_i\right)$ for one-sided tests and $P_i = 2\left(1 - \Phi\left(\left|T_i\right|\right)\right)$ for two-sided tests, where Φ is the standard normal distribution function.

To adjust the minimum observed *P*-value P_{min} to reflect that *L* correlated tests were performed, we compute the probability of observing at least one *P*-value as small as P_{min} under the null hypothesis of no association, given that $\boldsymbol{T} \sim N(\mathbf{0}, \mathbf{\Sigma})$ when the null hypothesis is true. Denoting this probability P_{ACT}, and letting $Z_1, ..., Z_L$ be random variables from the multivariate normal distribution with covariance matrix $\mathbf{\Sigma}$,

$$P_{ACT} = \begin{cases} 1 - P\left(\max\left(Z_1, ..., Z_L\right) < \Phi^{-1}\left(1 - P_{min}\right)\right) & \text{for one-sided tests} \\ 1 - P\left(\max\left(\left|Z_1\right|, ..., \left|Z_L\right|\right) < \Phi^{-1}\left(1 - \frac{P_{min}}{2}\right)\right) & \text{for two-sided tests,} \end{cases} \tag{2.2}$$

with the obvious generalization to a combination of one and two-sided tests. Figures 1a and 1b illustrate the probabilities for one and two-sided tests when $L = 2$. The elliptical lines represent the contours of the bivariate normal density function. P_{ACT}, is the probability that a random point from this distribution will fall within the shaded area.

Applying the sequentially rejective multiple test procedure of Holm (1979), the ordered *P*-values $P_{min} \leq P_{(2)} \leq P_{(3)} \leq \ ... \ \leq P_{(L)}$ may be adjusted and tested for significance one at a time, starting with P_{min}. We first adjust P_{min} for multiple testing by computing P_{ACT} as in equation (2.2). If $P_{ACT} < \alpha$, the null hypothesis is rejected for the test associated with P_{min} and we proceed to $P_{(2)}$. To adjust $P_{(2)}$ for multiple testing, we can remove the test associated with P_{min} from consideration, since the null hypothesis for this test has been rejected. We can now compute $P_{ACT}^{(2)}$ according to the formula in equation (2.2) but replacing P_{min} with $P_{(2)}$, *L* with $L-1$, and $\mathbf{\Sigma}$ with the covariance matrix between the remaining $L-1$ tests. If $P_{ACT}^{(2)} < \alpha$, we then reject the null hypothesis associated with $p_{(2)}$ and compute $P_{ACT}^{(3)}$ with $P_{(2)}$ removed from consideration, continuing in this fashion until $P_{ACT}^{(k)} \geq \alpha$ for some *k*, at which point we conclude that all remaining tests are insignificant. A good example of this kind of sequential testing in the multivariate normal case can be found in Wei et al. (1989).

2.2.2 Asymptotic Multivariate Normality of Common Association Test Statistics

Adjustment for multiple correlated tests with P_{ACT} requires that test statistics are asymptotically distributed as multivariate normal with known covariance matrix. Seaman and Müller-Myhsok (2005) have shown that for association tests based on *M* markers, one can apply the result that a vector of score statistics has a multivariate normal asymptotic distribution under the null hypothesis (McCullagh and Nelder 1989). We extend this result to include association tests based on correlated traits by deriving the asymptotic distribution for tests of association between *M* markers and *K* binary and continuous outcome variables. We show that this result can also be readily applied when multiple genetic models are tested. Although we focus on score tests, these results also apply to Wald and likelihood ratio tests, since they are asymptotically equivalent to the score test (Cox and Hinkley 1974).

For each individual (i = 1, …, N), let $\mathbf{Y}_i = \begin{bmatrix} Y_{i1} & Y_{i2} & \cdots & Y_{iK} \end{bmatrix}^T$ be a vector of *K* trait variables (where ^{T} indicates transpose) which may include both quantitative traits and binary traits such as disease status. Let $\mathbf{G}_i$ be a genotype vector containing allele counts of 0, 1, or 2 for each of *M* markers, and let $\mathbf{X}_i$ be a covariate vector that contains 1 as the first element and can also include environmental and demographic variables such as age and sex.

Many of the commonly used tests for association between traits and genotype are based on or related to the score statistics from a generalized linear model. Such tests include the simple test of equal allele frequency for cases and controls, the Cochran-Armitage test for trend (Cochran 1954; Armitage 1955) and linear and logistic regression. A key assumption of generalized linear models is that

$$E\left(Y_{ik} \mid \mathbf{X}_i, \mathbf{G}_i\right) = h\left(\eta_{ik}\right),$$

where h is a function and $\eta_{ik} = \mathbf{X}_i^T \alpha_k + \mathbf{G}_i^T \beta_k$, where α_k is a vector of covariate effects that includes an intercept term and β_k is an M-dimensional vector of genetic effects. Under this assumption, a linear combination η_{ik} of genotypes and covariates provides all the information necessary to predict the mean trait value, but the relationship between predicted trait value and η_{ik} may be non-linear. For example, in a trend test or logistic regression model, $h(\eta_{ik}) = \frac{e^{\eta_{ik}}}{1+e^{\eta_{ik}}}$.

If K traits are tested for association with M genotypes, the KM-dimensional vector of score statistics is

$$\mathbf{U}_\beta = \sum_{i=1}^{N} \left(\mathbf{Y}_i - \tilde{\mathbf{Y}}_i\right) \otimes \mathbf{G}_i \, ,$$

where $\tilde{\mathbf{Y}}_i$ is the vector of predicted trait values given covariates, assuming no genetic association, and $\otimes$ represents the Kronecker product. As we show in the appendix, $\mathbf{U}_\beta \dot{\sim} N\left(0, \mathbf{V}_\beta\right)$, where $\mathbf{V}_\beta$ can be estimated as $\Omega \otimes \left[\mathbf{G}\mathbf{G}^T - \mathbf{G}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{G}^T\right]$, the Kronecker product of the sample covariance matrices of traits and genotypes, conditioning on covariates. Here $\mathbf{G} = \left[\mathbf{G}_1 \mid \mathbf{G}_2 \mid \cdots \mid \mathbf{G}_N\right]$ and $\mathbf{X} = \left[\mathbf{X}_1 \mid \mathbf{X}_2 \mid \cdots \mid \mathbf{X}_N\right]$ are matrices of genotypes and covariates and $\Omega = \sum_{i=1}^{N}\left(\mathbf{Y}_i - \tilde{\mathbf{Y}}_i\right)\left(\mathbf{Y}_i - \tilde{\mathbf{Y}}_i\right)^T$ is the trait covariance matrix, conditioned on $\mathbf{X}$.

The P-values from individual association tests are generally based on test statistics that are normalized to have variance one. A vector of L score statistics $\mathbf{U}_\beta$ is easily transformed to a normalized vector of test statistics $\boldsymbol{T}$ by computing each element of $\boldsymbol{T}$ as $T_l = \frac{\mathbf{U}_{\beta,l}}{\sqrt{\mathbf{V}_{\beta,ll}}}$, where $\mathbf{U}_{\beta,l}$ is the lth element of $\mathbf{U}_\beta$ and $\mathbf{V}_{\beta,ll}$ is the lth element along the diagonal of $\mathbf{V}_\beta$ for $l = 1, \ldots, L$; it is also common to work with $T_l^2 = \mathbf{U}_{\beta,l}\mathbf{V}_{\beta,ll}^{-1}\mathbf{U}_{\beta,l}$. It is easy to show that $\boldsymbol{T} \dot{\sim} N\left(0,\mathbf{R}\right)$, where $\mathbf{R}$ is the correlation matrix corresponding to the covariance matrix $\mathbf{V}_\beta$. Using this fact, P_{ACT} can then be computed as in equation (2.2) given only P_{min} and **R. R** in turn can generally be estimated as a simple function of the sample correlation matrices of traits and markers, conditioned on any covariates. Appropriate estimates of $\mathbf{R}$ are shown for a few examples in Table 2.1.

The above model may be trivially extended to include tests based on multiple genetic models. For example, if a marker is tested for association in three ways, assuming an additive, dominant, and recessive model, it can be assigned three elements in $\mathbf{G}_i$, each containing the appropriate genotype code. For instance, the genotype codes for an individual with two copies of the reference allele would be 2, 1, and 1 for the additive, dominant, and recessive model respectively. The score statistics and covariance matrix are then computed as usual.

2.2.3 Computation of P_{ACT}

Computation of p_{ACT} in (2) requires integration of the multivariate normal density function. Although the integral has no closed-form solution, multivariate normal probabilities can be integrated numerically when the covariance matrix is known or can be estimated. Genz (1992; 1993) and Genz and Bretz (2002) have developed a computationally efficient method for numerical integration of the multivariate normal distribution which is available as Fortran code that can handle integrands of up to 1000 dimensions (Genz 2000). This Fortran code has been incorporated into the package ‘mvtnorm’ (Genz et al. 2007) in the R software environment (R Development Core Team 2007), and the latest version of ‘mvtnorm’ (versions ≥ 0.8-0) provides sensible estimates of the multivariate normal integral for up to 1000 dimensions (Genz et al. 2007). We apply Genz’s algorithm as implemented in ‘mvtnorm’ to estimate P_{ACT} for several common association tests. In the interests of computational efficiency, we may choose the requested precision level depending on the magnitude of the *P*-values and the nature of the analysis. For example, one may desire a quick low-precision analysis for exploratory purposes or for clearly non-significant results, but want to devote more computational resources to a high-precision final analysis. Our R code for computation of P_{ACT} is available online at http://csg.sph.umich.edu/boehnke/p_act.php.

2.2.4 Assessment of Type I Error Rate and Power

To estimate the type I error rate and power of adjusting for multiple testing with P_{ACT}, we performed simulations that involved both binary and quantitative traits. In each case, we estimated type I error by simulating 100,000 datasets under the null hypothesis, where trait was assigned at random independent of genotype. Similarly, we estimated power by creating 10,000 replicate datasets where trait was influenced by genotype. For each simulation, we performed the relevant set of association tests and computed three overall *P*-values: P_{ACT} and $P_{Šidák}$, as in equations (2.1) and (2.2) above, and P_{perm}. To calculate P_{perm}, we first created 1000 permutations of the original data by randomly shuffling individual genotype vectors while leaving the trait data and any covariates intact. In this way, the permuted samples simulated the null hypothesis of no association, but maintained the original correlation between genotypes, between traits, and between traits and covariates. We tested each of these 1000 samples for association and estimated P_{perm} as the proportion of samples with a minimum *P*-value as low as that observed in the original data. Although 1000 permutations is much lower than we would use in practice, it is sufficient for estimating type I error and power at the significance level $\alpha = .05$ we chose to use.

Binary trait simulations: We simulated case-control status for 1389 individuals genotyped for 20 *HNF1A* SNPs as part of the Finland-United States Investigation of NIDDM Genetics (FUSION) study of the genetics of type 2 diabetes (Valle et al. 1998).

HNF1A is one of six genes known to be involved in maturity-onset diabetes of the young (Fajans et al. 2001) and was analyzed by FUSION as a potential candidate gene for type 2 diabetes (Bonnycastle et al. 2006). Of the 20 SNPs genotyped for the study, most had been chosen to be non-redundant ($r^2 < .8$) and as Figure 2.2 shows, only moderate LD was present.

For type I error estimation we randomly assigned case-control status in each simulation. For power estimation we chose one of the 20 SNPs as a disease SNP and randomly assigned case-control status according to a multiplicative model of disease risk for each individual, where genotype relative risk (GRR) was chosen to ensure a roughly equal number of cases and controls and a correlation of ~.12 between case-control status and the disease gene. This corresponded to a GRR of 1.2 if the disease SNP was our most common SNP, with a minor allele frequency (MAF) of .48, and a GRR of 1.4 if the disease SNP was our least common SNP (MAF = .04). Individuals missing genotype data for the disease SNP were assigned the mean GRR. To model the common situation in which the genotyped SNPs are proxies for a disease-predisposing variant that was not genotyped, we then omitted the disease SNP from consideration and tested only the remaining 19 SNPs for association when estimating power. For estimation of the type I error rate there was no disease SNP, so in this case we tested all 20 SNPs.

We first tested each of the 19 or 20 SNPs for association with a Cochran-Armitage test for trend, which assumes an additive model of disease risk. In each case, we computed P_{ACT}, $P_{Šidák}$, and P_{perm} to adjust for the 19 or 20 tests. Since 215 individuals were missing data on at least one genotype, we performed each association test using only individuals with data for the SNP being tested, but estimated the covariance matrix using genotype data from all individuals, with missing genotype data for each SNP filled in with the mean allele count for that SNP.

Using the same data, we also tried testing every SNP under the additive, dominant, and recessive models and adjusting for all of the tests with P_{ACT}, $P_{Šidák}$, and

P_{perm}. For SNPs with < 20 minor allele homozygotes, we omitted the relevant dominant or recessive model from analysis. This led to the exclusion of four models, for a total of 56 tests before also removing the disease SNP from consideration.

For the same 1389 genotyped individuals, we simulated 5 correlated binary traits according to a probit model. For each simulation, we first generated 5 equally correlated random variables $Z_{i1}, \ldots, Z_{i5}$ from the multivariate normal distribution for each individual i. For $j = 1, \ldots, 5$, each binary trait Y_{ij} was defined as 1 if $Z_{ij} > 0$, and 0 otherwise. The resulting 5 binary traits were equally correlated with one another, with all pairwise correlations ≈ .7. For power estimation, we allowed one trait to be influenced additively by the disease SNP by defining it to be 1 if $Z_{ij} + \left(G_i - \overline{G}\right)\beta > 0$ and 0 otherwise, where G_i is disease allele count (0, 1, or 2) for individual i and $\overline{G}$ is the mean allele count over all individuals with genotypes for the disease SNP. For individuals missing genotypes for the disease SNP, we set $G_i - \overline{G}$ to zero. We then used Cochran-Armitage trend tests to test each of the 20 SNPs for association with each of the 5 traits, for a total of 100 tests (or 95 when the disease SNP is omitted). We again used P_{ACT}, $P_{Šidák}$, and P_{perm} to adjust for the 95 or 100 tests.

Quantitative trait simulations: We first simulated datasets of 2000 individuals with 10 correlated quantitative traits and genotype data for a single SNP with allele frequency .5. We assigned trait values Y_{ij} according to the linear model $Y_{ij} = \alpha_j X_i + \beta_j G_i + \varepsilon_{ij}$ where G_i is the allele count for individual i, X_i is a covariate generated as a linear function of G_i and a random normal component such that the correlation between X_i and G_i ~ .25, ε_{ij} is a random component, and α_j and β_j are parameters that determine the effect of the covariate and genotype on trait j. For each trait, α_j was drawn from a normal distribution tightly centered around a fixed effect size so that covariates had a similar, though not identical, effect on the ten traits. We set

$\beta_j = 0$ for j = 1,…,10 when computing type I error and $\beta_1 > 0$ and $\beta_j = 0$ for j = 2,…, 10 when computing power. We simulated $\varepsilon_i = \left[\varepsilon_{i1} \quad \varepsilon_{i2} \quad \cdots \quad \varepsilon_{i10}\right]^T$ from the multivariate normal distribution $N\left(0, \mathbf{R_Y}\right)$ with one of the five correlation structures shown in Figure 2.3. For each simulation, we tested the SNP for association with each trait separately with a linear regression of the trait value on allele count and the covariate. We used the results from the 10 tests to compute $P_{Šidák}$, P_{ACT}, and P_{perm}. We performed simulations for lower (.2), higher (.7), and extremely high (.99) values of ρ.

We next randomly drew *HNF1A* genotypes for each individual and simulated ten traits using a similar linear model with no covariates. We tested the traits for association with the 20 *HNF1A* SNPs, for a total of 200 tests. We estimated type I error as in our previous simulations; to estimate power we simulated a model where Y_{i1} is influenced by the least common of the 20 SNPs (MAF = .04). When 200 tests were involved, estimation of P_{perm} was too computationally intensive, so in this case we estimated $P_{Šidák}$ and P_{ACT} only

Finally, we performed both the single-SNP and 20-SNP simulations for a set of 5 binary and 5 continuous traits. We generated 10 multivariate normal random variables according to the model $Z_{ij} = \alpha_j X_i + \beta_j G_i + \varepsilon_{ij}$, with X_i, G_i, ε_{ij}, α_j, and β_j defined as above. We defined the 5 continuous traits as $Y_{ij} = Z_{ij}$ for j = 1, …, 5 and the 5 binary traits by setting $Y_{ij} = 1$ if $Z_{ij} > 1.25$ and 0 otherwise for j = 6, …, 10. Each binary trait had a prevalence of $\sim .1$ and we chose the covariance of ε_i such that all pairwise trait correlations were between .5 and .7. We estimated type I error and power as in previous simulations.

Performance of other methods: We also used the simulations described above to estimate the type I error rate for two methods which estimate an effective number of tests (see Introduction). For the method of Cheverud (2001) and Nyholt (2004), we computed the effective number of tests as $1+\left(L-1\right)\left(1-Var\left(\lambda\right)/L\right)$, where L is the number of tests performed and $Var\left(\lambda\right)$ is the variance of the eigenvalues from the correlation matrix between the tests. For the method of Li and Ji (2005), we computed the effective number of tests as $\sum_{i=1}^{L}\left(I\left(\left|\lambda_i\right|\geq 1\right)+\left(\left|\lambda_i\right|-\left\lfloor\left|\lambda_i\right|\right\rfloor\right)\right)$, where $I\left(\left|\lambda_i\right|\geq 1\right)$ is 1 if the absolute value of the ith eigenvalue $\left|\lambda_i\right| \geq 1$ and 0 otherwise, and $\left\lfloor\left|\lambda_i\right|\right\rfloor$ is the largest integer $\leq\left|\lambda_i\right|$. For each method, we computed a multiple-testing adjusted *P*-value by substituting the effective number of tests for *L* in the Šidák formula. We then estimated the type I error rate as described above.

2.2.5 Comparison between P_{ACT} and P_{perm} in FUSION Data

To assess how closely estimates of P_{ACT} correspond to gold standard estimates based on P_{perm}, we analyzed 3575 SNPs in and near 224 candidate genes which were genotyped on 1161 type 2 diabetes (T2D) cases and 1174 normal glucose-tolerant controls from the FUSION study (Gaulton et al. 2007). We first tested the 3007 SNPs having ≥ 20 individuals in each of the three genotype classes for association with T2D using the additive, dominant, and recessive models and controlling for age category, sex, and birth region as covariates. For each SNP, we estimated both P_{ACT} and P_{perm} to adjust for the three tests, providing 3007 comparisons between P_{ACT} and P_{perm}.

We next tested all 3575 SNPs for association with 18 quantitative T2D-related traits (residualized on age category, sex, and birth region) on the 1174 controls. For each SNP, we estimated both P_{ACT} and P_{perm} to adjust for the 18 correlated tests, providing 3575 comparisons between P_{ACT} and P_{perm}. To provide additional comparisons between P_{ACT} and P_{perm} for highly significant tests, we simulated 9 additional SNPs with minimum *P*-values of 1×10^{-5}, 5×10^{-6}, 2.5×10^{-6}, 1×10^{-6}, 5×10^{-7}, 2.5×10^{-7}, 1×10^{-7}, 5×10^{-8}, and 2.5×10^{-8} and adjusted these minimum *P*-values for multiple testing with P_{ACT} and P_{perm}.

For all comparisons, we computed P_{ACT} at increased precision for more significant SNPs, and under the assumption that covariates were independent of genotype. For P_{perm}, we performed 1,000,000 permutations for the 10 most significant SNPs, 100,000 for the next 190 significant SNPs, and 10,000 for all other SNPs. For the 9 SNPs simulated to be highly significant, we performed 10,000,000 permutations.

2.3 Results

2.3.1 Type I Error Rate and Power for Simulated Data

Table 2.2 presents estimates of type I error rate (first row) and power (subsequent rows) for $P_{Šidák}$, P_{ACT}, and P_{perm} when the 20 *HNF1A* SNPs are tested for disease association. The estimates in the first row (based on 100,000 simulation replicates each) show that both P_{ACT} and P_{perm} have type I error rates ~ .05 and are thus valid in all cases considered: when the 20 SNPs are tested for association with a binary trait under an additive model or under three competing models, or when the SNPs are tested for association with 5 correlated binary traits. Tests based on $P_{Šidák}$ are conservative in each case. A similar pattern was observed for α-levels of .01, .001, and .0001, or when the true model was dominant or recessive (data not shown).

Each of the next four rows of Table 2.2 present power estimates with a different SNP modeled as the disease-predisposing SNP: the most common SNP (MAF=.48), a moderately frequent SNP (MAF=.20), the least common SNP (MAF=.04), and the SNP least well predicted by a linear function of the others. The power estimates (based on 10,000 simulation replicates each) show that tests based on P_{ACT} have near identical power to permutation tests and are consistently more powerful than Šidák (or Bonferroni) adjustment. Results were similar for the other 16 SNPs (data not shown).

Table 2.3 presents estimates of type I error rate and power for tests of association with traits correlated as in Figure 3 with ρ = .7; data are presented for 10 quantitative traits in rows 1-5 and for 5 binary and 5 quantitative traits in row 6. The leftmost panel shows that when a single SNP is tested for association, P_{ACT} and P_{perm} provide valid tests and $P_{Šidák}$ is overly conservative except when traits are independent, as in the first row. The next panel shows the familiar pattern of near identical power for P_{ACT} and P_{perm}, while $P_{Šidák}$ has reduced power in each situation except independence. The two panels on the right show that results are similar even when 20 correlated SNPs are tested for association with 10 correlated traits, for a total of 200 tests. Similar results were also observed for lower levels of correlation (ρ = .2, data not shown) and extremely high levels of correlation (ρ =.99) (data not shown). As expected, the power gains of P_{ACT} and P_{perm} over $P_{Šidák}$ were smaller when ρ = .2 and greater when ρ = .99.

We ran additional simulations testing up to 1000 equicorrelated quantitative traits (ρ = .7) for association with a single SNP and a covariate (data not shown). For 300, 400, and 500 tests, estimated type I error rate was .0121 .0112, and .0102 for $P_{Šidák}$ and .506, .0499, and .0517 for P_{ACT}, suggesting that P_{ACT} can achieve the target type I error rate for several hundred tests, while $P_{Šidák}$ is increasingly conservative. For 600, 750, and 1000 tests, estimated type I error rate was .0102, .0093, and .0086 for $P_{Šidák}$ and .0550, .0593, and .0648 for P_{ACT}, indicating a possible bias or reduction in the precision of P_{ACT} when the number of tests is extremely large.

For the two methods based on the effective number of tests (data not shown), we found that the method of Cheverud (2001) and Nyholt (2004) tended to be overly conservative and the method of Li and Ji (2005) was anti-conservative in all cases except when tests were completely independent. When a binary trait was tested for association with 20 *HNF1A* SNPs, the type I error rates for the two methods were .0389 and .0613 for just the additive model or .0297 and .0667 when three genetic models were tested. When ten traits were tested for association with a single SNP and a covariate, both methods had a type I error rate ~ .05 when traits were independent; for the other trait correlation structures the type I error rate ranged from .0460 to .0504 for the Cheverud/Nyholt method and from .0615 to .0666 for the method of Li and Ji,.

2.3.2 P_{ACT} and P_{perm} in FUSION Data

Figures 2.4 and 2.5 show the relationship between P_{ACT} and P_{perm} in the context of a FUSION study of 3575 SNPs in 224 candidate genes for type 2 diabetes (Gaulton et al. 2007). P_{ACT} and P_{perm} are plotted on a log scale to emphasize the smallest *P*-values (upper right of figure). We obtained the values of P_{ACT} and P_{perm} in Figure 2.4 by testing each SNP for association under the additive, dominant, and recessive models and adjusting the minimum *P*-value from these three tests for multiple testing. We obtained the values of P_{ACT} and P_{perm} in Figure 2.5 by testing each SNP for association with 18 correlated T2D-related traits, and adjusting the minimum *P*-value for each SNP for the 18 tests. Figure 2.5 also includes data for 9 highly significant simulated SNPs, indicated by filled circles. In all cases, P_{ACT} and P_{perm} track each other quite closely, with all points falling very near the identity line ($r^2 > .999$ for both figures).

2.3.3 Computation Speed: Comparison Between Methods

Because the goal of our proposed method is to estimate *P*-values with the same accuracy and precision as permutation tests in less time, we timed computation of *P*-values at a constant level of precision. We compared timings for P_{ACT}, P_{perm}, and one of the simulation-based methods (see Introduction) that has been shown to attain the accuracy of permutation tests – the direct simulation approach (DSA) of Seaman and Müller-Myhsok (2005). We implemented all three methods in R, using the code for the DSA provided on the authors' website. For each method, we measured the time required to compute an adjusted *P*-value for a fixed P_{min} (chosen such that $P_{ACT},$ $P_{DSA},$ or $P_{perm} \approx$

.0001) at a given level of precision (standard error ≤ .00001). Attainment of this level of precision requires ~ 1,000,000 permutations for P_{perm} and ~ 1,000,000 simulations for P_{DSA}. Since the speed of P_{perm} depends on sample size, we present timings for three typical sample sizes. For computational efficiency, we tested for association with a simple Cochran-Armitage test for trend; models requiring additional computation such as logistic or even linear regression would have penalized the permutation tests to a much greater degree. For example, if we had instead tested for association with a logistic regression model of trait on genotype with age and sex as covariates, the timings for P_{ACT} and P_{DSA} would show no noticeable change, but computation of P_{perm} would have taken > 300 times longer.

Table 2.4 compares timings for P_{ACT}, P_{perm}, and P_{DSA} for three representative situations. The first row shows timings when 200 autocorrelated tests are adjusted for multiple testing. This example is meant to approximate the correlation between a series of non-redundant SNPs along a chromosome, since correlation is generally high between neighboring SNPs and decays with distance. In this case, computing P_{ACT} is ~ 60 times faster than P_{DSA} and thousands of times faster than P_{perm}. Similar timings for 20, 40, 60, 80, and 100 autocorrelated tests demonstrate that the computational time required increases approximately linearly in the number of tests for all three estimators (data not shown). We also computed P_{ACT} for even smaller P-values and greater dimension. Adjustment of a minimum P-value of 10^{-8} with P_{ACT} with standard error ≤ 10% of estimate required 11 seconds for 200 autocorrelated tests, 25 seconds for 500 tests, and 70 seconds for 1000 tests. The same computation for only 200 tests would have required > 3 hours for P_{DSA} and 100-800 hours for P_{perm}, depending on sample size.

The second row presents the computational time required to test the 20 *HNF1A* SNPs for association. In this case, P_{ACT} can be computed 60 times faster than P_{DSA} and up to 5000 times faster than P_{perm}. The third row uses the information from the second to consider the prospect of 20,000 independent blocks of 20 SNPs with the correlation structure of *HNF1A*, illustrating what might occur if we tested sets of SNPs from every gene in the human genome. In this situation, permutation testing is essentially infeasible except with massive amounts of parallelization, while the same analysis can be performed with P_{ACT} in a single afternoon.

2.4 Discussion

Permutation testing, when performed appropriately, provides an unbiased test of the null hypothesis and is widely considered the gold standard to which other estimators and tests may be compared. Its main disadvantage is the time and computational resources required to obtain precise *P*-value estimates, so alternative tests that provide similar results with less computational burden can be quite attractive, particularly when a large number of tests is involved, or when data are frequently reanalyzed in light of new samples or genotypes.

While conventional distribution-based statistical tests typically require minimal computational resources, permutation tests are often employed when the asymptotic distribution of the statistic is unknown or difficult to model. However, for many of the tests commonly used in genome-wide association studies, the asymptotic joint distribution of the test statistics is known, making analytical methods possible. As we show above, the asymptotic distribution of test statistics from association tests between correlated traits, markers, and models is often multivariate normal with known covariance matrix. However, the most significant test statistic from a group of multivariate normal test statistics has a distribution function that while known, cannot be computed analytically due to the lack of a closed-form solution to the multivariate normal integral. Lin (2005a) and Seaman and Müller-Myhsok (2005) have suggested simulation-based approaches that can approximate the null distribution of ordered test statistics much more quickly than permutation tests. Our P_{ACT} method relies on numerical integration of the distribution function and can approximate the null distribution much more quickly than permutation or simulation-based approaches.

The data presented here suggest that tests based on P_{ACT} are appropriate substitutes for those based on permutation testing, since P_{ACT} consistently attains essentially identical results to permutation-based P-values both in simulated data and over thousands of association tests performed as a part of a large candidate gene study (Gaulton et al. 2007). While Lin (2005a) and Seaman and Müller-Myhsok (2005) have also demonstrated that their estimators (denoted here P_{Lin} and P_{DSA}) provide valid tests and attain the accuracy of P_{perm}, P_{ACT} demonstrates greater gains in computational efficiency. P_{ACT} is typically thousands of times faster than permutation-based P-values at a given level of precision. This makes P_{ACT} potentially useful in the contexts of both large-scale candidate gene studies, where thousands of tests may be performed, and genome-wide association studies, where millions of tests may be performed. Since the precision of this method can be traded for speed, P_{ACT} can be tailored both to initial exploratory tests where speed is especially important and to more definitive tests where greater precision is needed; it can also be computed at increased precision for more interesting results.

Like any estimator, P_{ACT} is not appropriate for every analysis. It was designed to adjust the minimum P-value and other ordered P-values for a large number of 1-df tests. An advantage of this method is that it allows easy identification of the particular traits, variants, and genetic models associated with the most interesting results. This approach is especially relevant if we are looking for a small number of reasonably large genetic effects. If we instead expect a large number of very small effects, a joint analysis of all associations simultaneously might be more appropriate. Typically these methods are based on multi-degree-of-freedom tests, which are outside the scope of P_{ACT}, but P_{Lin} and P_{DSA} remain useful alternatives to permutation testing in these situations. For example, the DSA software (Seaman and Müller-Myhsok 2005) computes an adjusted P-value for product methods (see Fisher 1932; Zaykin et al. 2002) as well as for the minimum *P*-value.

The validity of P_{ACT} (as well as P_{Lin} and P_{DSA}) depends on knowledge of the correct asymptotic distribution. While many common association test statistics are asymptotically multivariate normal, use of the asymptotic distribution requires reasonably large sample sizes and cell counts and may not be appropriate in all cases – for example, dominant or recessive models with a rare minor allele. The solution we have employed here and elsewhere (Bonnycastle et al. 2006; Willer et al. 2007; Gaulton et al. 2007) is to drop dominant or recessive models with low cell counts from analysis; another solution would be to rely on exact tests such as Fisher's exact test for these models. A related issue is that sample size must be substantially larger than the number of tests for asymptotic properties to hold; however, simulations have shown that P_{Lin} can achieve the target type I error rate when the number of tests far exceeds the sample size (Lin 2005a). For situations where the asymptotic distribution is unknown or the sample size is too small for asymptotic properties to hold, however, permutation testing may be the appropriate choice. The algorithm of Kimmel and Shamir (2006), which relies on importance sampling to sample from the null distribution in a way that mimics permutation testing, can also be computed thousands of times faster than permutation tests and does not require assumptions about the asymptotic distribution. A direct comparison of the asymptotic methods discussed here and this importance sampling method has not been performed but would be of great interest.

The validity of P_{ACT} and P_{DSA} also depends on accurate estimation of the covariance matrix. Improper handling of missing trait or genotype data is one factor that can lead to biased covariance estimates. While it is rare for samples to contain complete genotype and trait data for every individual, only individuals with complete data can be used in computation of sample covariance matrices; otherwise the matrices may not be positive-definite. However, exclusion of individuals with incomplete data may lead to biased estimates of the covariance matrix. Seaman and Müller-Myhsok (2005) suggest performing the entire analysis with missing genotype data imputed, but Lin (2005b) argues that imputation can adversely impact type I error. Lin's estimator is based on individual contributions to the score statistic, and he treats missing data for an individual by setting the individual component of the appropriate score statistic(s) to zero. In the case of P_{ACT}, an analogous approach is to test each trait and marker using only individuals with complete data for that trait and marker, but to estimate the covariance matrix of the tests using the full sample, with missing data for marker m (or trait k) filled in with the mean genotype score for marker m (or the mean value for trait k), conditional on covariates. Although >15% of individuals in our first set of simulations (Table 2.2) were missing data on at least one genotype, P_{ACT} achieved the target type I error when this approach was used.

Valid covariance matrix estimation also depends on how many tests are considered at once. The numerical integration method implemented in the package 'mvtnorm' (Genz et al. 2007) has proved reliable in testing of 750-dimensional integrals (Genz 2007), and we observed that high levels of precision are possible for up to 1000 dimensions. However, even with reliable numerical integration, precision of the covariance estimates may suffer as the ratio between the number of parameters in the covariance matrix and the number of usable samples increases. In our simulations, tests based on P_{ACT} with samples of 2000 were consistently valid for dimension 200, and appeared to be valid in examples with 300–500 tests. However, in the examples we considered with 600–1000 tests, P_{ACT} did not achieve the target type I error rate. Further investigation of the appropriate upper limits on dimension and how they relate to sample size is warranted. Seaman and Müller-Myhsok (2005) treat 0.1 as the upper limit for the ratio of number of tests (L) to sample size (N), which seems an appropriate rule of thumb since the eigenvalues of a sample covariance matrix resemble the eigenvalues of the true matrix quite closely when $L \leq N/10$ (Schäfer and Strimmer 2005). Given conventional sample sizes, large-scale candidate gene studies are quite feasible within such a limit, and we have already used P_{ACT} in several (Bonnycastle et al. 2006; Willer et al. 2007; Gaulton et al. 2007).

With genome-wide association studies becoming a priority, there is also potential for P_{ACT} to be useful on a larger scale. One possible strategy is to break up large analyses into roughly independent blocks of hundreds of tests each (Seaman and Müller-Myhsok 2005). If we then compute P_{ACT} for each group of tests, the Šidák procedure can be used to adjust the most significant values of P_{ACT} for the number of blocks via the sequential Holm (1979) procedure (see Methods). As long as the correlation between the blocks of tests is reasonably low, little power will be sacrificed by approximating in this way since P_{ACT} has accounted for the correlation within the blocks. Use of the P_{ACT} method in such a framework has the potential to facilitate exploration of the genome by highlighting our most significant findings without imposing an overly severe penalty when hundreds, thousands, or millions of association tests are performed.

Appendix

Written in terms of covariate effects(α), the *KM*-dimensional vector of score statistics is

$$\mathbf{U}_{\beta}(\tilde{\alpha},0)=\sum_{i=1}^{N}\left(\mathbf{Y}_i-\tilde{\mathbf{Y}}_i\right)\otimes\mathbf{G}_i=\sum_{i=1}^{N}\left(\mathbf{Y}_i-h\left(\tilde{\boldsymbol{\eta}}_i\right)\right)\otimes\mathbf{G}_i\ ,$$

where $\tilde{\boldsymbol{\eta}}_i$ is the vector $\left[\mathbf{X}_i^{\ T}\tilde{\alpha}_1 \quad \mathbf{X}_i^{\ T}\tilde{\alpha}_2 \quad \cdots \quad \mathbf{X}_i^{\ T}\tilde{\alpha}_K\right]^T$ and $\tilde{\alpha}_k$ is the maximum-likelihood estimate of α_k when β_k is restricted to zero. A first-order Taylor expansion gives us

$$\frac{1}{\sqrt{n}}\mathbf{U}_{\beta}(\tilde{\alpha},0)\approx\frac{1}{\sqrt{n}}\mathbf{U}_{\beta}(\alpha,0)+\frac{\partial}{\partial\alpha}\left(\frac{1}{n}\mathbf{U}_{\beta}(\alpha,0)\right)\sqrt{n}\left(\tilde{\alpha}-\alpha\right)$$

where $\tilde{\alpha}$ and α are the stacked vectors $\left[\tilde{\alpha}_1^{\ T} \quad \tilde{\alpha}_2^{\ T} \quad \cdots \quad \tilde{\alpha}_K^{\ T}\right]^T$ and $\left[\alpha_1^{\ T} \quad \alpha_2^{\ T} \quad \cdots \quad \alpha_K^{\ T}\right]^T$, respectively. The multivariate central limit theorem (Cramér 1946) may be applied to show that $\frac{1}{\sqrt{n}}\mathbf{U}_{\beta}(\alpha,0)\ \dot{\sim}\ N\left(0,Var\left(\sum_{i=1}^{N}\left(\mathbf{Y}_i-h\left(\boldsymbol{\eta}_{\mathbf{i}}\right)\right)\otimes\mathbf{G}_i\right)\right)$ where $\boldsymbol{\eta}_i$ is the vector $\left[\mathbf{X}_i^{\ T}\alpha_1 \quad \mathbf{X}_i^{\ T}\alpha_2 \quad \cdots \quad \mathbf{X}_i^{\ T}\alpha_K\right]^T$. Since under the null hypothesis $\mathbf{Y}_i-h\left(\boldsymbol{\eta}_{\mathbf{i}}\right)$ and $\mathbf{G}$ are independent with mean zero, $Var\left(\sum_{i=1}^{N}\left(\mathbf{Y}_i-h\left(\boldsymbol{\eta}_{\mathbf{i}}\right)\right)\otimes\mathbf{G}_i\right)$ can be estimated efficiently by $\Omega\otimes\mathbf{GG}^T$ where $\Omega=\sum_{i=1}^{N}\left(\mathbf{Y}_i-\tilde{\mathbf{Y}}_i\right)\left(\mathbf{Y}_i-\tilde{\mathbf{Y}}_i\right)^T$. It is also easily shown through Taylor expansion of $\mathbf{U}_{\alpha}(\tilde{\alpha})$ that $\sqrt{n}\left(\tilde{\alpha}-\alpha\right)\dot{\sim}\ N\left(0,Var\left(\left(\mathbf{Y}_i-h\left(\boldsymbol{\eta}_{\mathbf{i}}\right)\right)\otimes\mathbf{X}_i\right)^{-1}\right)$ where $Var\left(\left(\mathbf{Y}_i-h\left(\boldsymbol{\eta}_{\mathbf{i}}\right)\right)\otimes\mathbf{X}_i\right)$ can be estimated by $\Omega^{-1}\otimes\mathbf{XX}^T$. Finally, $\frac{\partial}{\partial\alpha}\left(\frac{1}{n}\mathbf{U}_{\beta}(\alpha,0)\right)=h'\left(\boldsymbol{\eta}_{\mathbf{i}}\right)\otimes\mathbf{GX}^T$, which has sample analogue $\Omega\otimes\mathbf{GX}^T$. Hence,

$$\frac{1}{\sqrt{n}}\mathbf{U}_{\beta}(\tilde{\alpha},0)\ \dot{\sim}\ N\left(0,\mathbf{V}_{\beta}\right),\ \text{ where}$$

$$\begin{aligned}\mathbf{V}_{\beta} &= \Omega \otimes \mathbf{G}\mathbf{G}^{T} - \left(\Omega \otimes \mathbf{G}\mathbf{X}^{T}\right)\left(\Omega \otimes \mathbf{X}\mathbf{X}^{T}\right)^{-1}\left(\Omega \otimes \mathbf{X}\mathbf{G}^{T}\right) \\ &= \Omega \otimes \left[\mathbf{G}\mathbf{G}^{T} - \mathbf{G}\mathbf{X}^{T}\left(\mathbf{X}\mathbf{X}^{T}\right)^{-1}\mathbf{X}\mathbf{G}^{T}\right].\end{aligned}$$

Table 2.1: The covariance matrix of test statistics **R**: three examples

Traits	Markers	**R**
2 traits with correlation ρ	Single SNP	$\mathbf{R}_Y = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$
Single trait	2 SNPs with correlation r	$\mathbf{R}_G = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$
2 traits with correlation ρ	2 SNPs with correlation r	$\mathbf{R}_Y \otimes \mathbf{R}_G = \left[\begin{array}{cc:cc} 1 & \rho & r & r\rho \\ \rho & 1 & r\rho & r \\ \hdashline r & r\rho & 1 & \rho \\ r\rho & r & \rho & 1 \end{array}\right]$

Table 2.2: Type I error rate and power when 20 *HNF1A* SNPs are tested for association with binary traits

				One binary trait tested						5 binary traits tested		
				on additive model			on three models			on additive model		
Disease SNP	MAF	r^2_{total}	r^2_{max}	$P_{Šidák}$	P_{ACT}	P_{perm}	$P_{Šidák}$	P_{ACT}	P_{perm}	$P_{Šidák}$	P_{ACT}	P_{perm}
None (Type I error)	–	–	–	.0301	.0503	.0507	.0247	.0500	.0508	.0259	.0495	.0502
Most common SNP	.48	.88[a]	.78[b]	.899	.927	.925	.859	.911	.910	.806	.857	.859
Moderately frequent SNP	.20	.93	.19	.419	.535	.538	.338	.482	.484	.280	.385	.377
Least common SNP	.04	.91	.79	.878	.916	.915	.811	.874	.874	.686	.772	.773
SNP least predicted by others	.05	.42	.35	.387	.475	.476	.296	.401	.402	.220	.304	.299

[a] r^2_{total} = proportion of variance in disease SNP allele count explained by the other 19 SNPs

[b] r^2_{max} = maximum pairwise r^2 between disease SNP and the other 19 SNPs

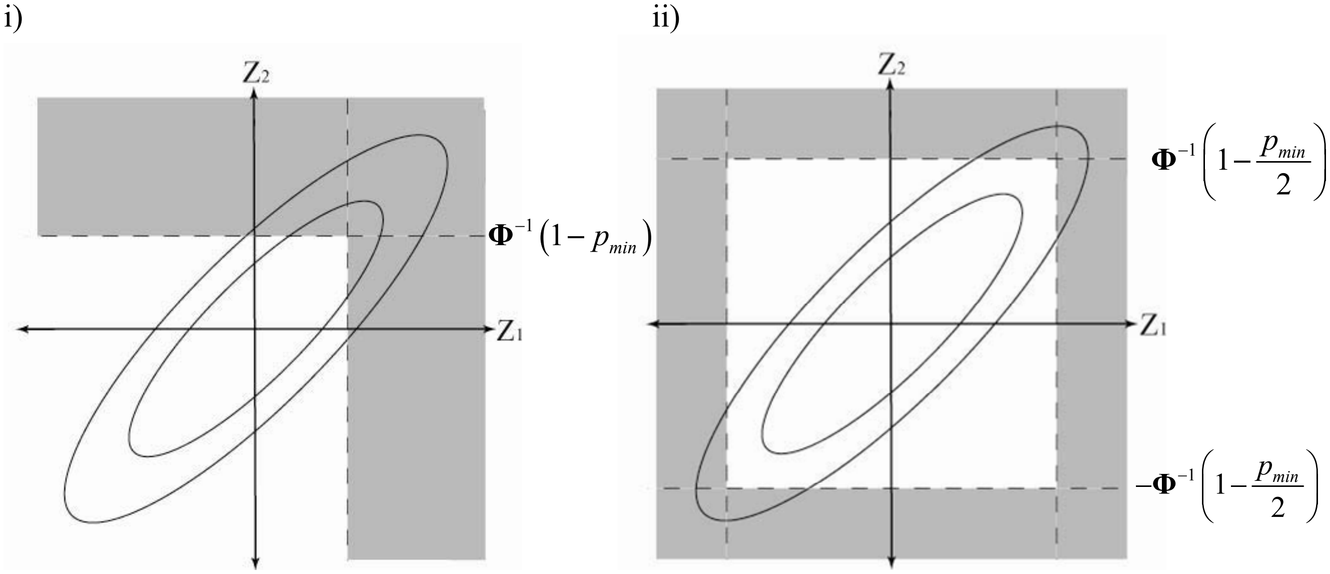
Table 2.3: Type I error rate and power when 10 correlated quantitative traits are tested for association

	Ten traits tested for association with:									
	one SNP and a covariate						20 correlated HNF1A SNPs			
	Type I error rate			Power			Type I error rate		Power	
Trait correlation structure	$P_{\check{S}id\acute{a}k}$	P_{ACT}	P_{perm}	$P_{\check{S}id\acute{a}k}$	P_{ACT}	P_{perm}	$P_{\check{S}id\acute{a}k}$	P_{ACT}	$P_{\check{S}id\acute{a}k}$	P_{ACT}
Independent traits	.0498	.0499	.0496	.819	.819	.816	.0325	.0514	.780	.821
Equicorrelated traits	.0302	.0502	.0503	.826	.880	.878	.0216	.0507	.778	.852
Autocorrelated traits	.0393	.0494	.0495	.820	.842	.839	.0274	.0499	.777	.833
Independent blocks of traits	.0386	.0497	.0501	.824	.850	.848	.0264	.0501	.779	.836
Negatively correlated blocks	.0327	.0496	.0500	.825	.870	.868	.0234	.0503	.779	.846
5 binary, 5 quantitative traits	.0341	.0491	.0488	.825	.864	.860	.0263	.0517	.781	.844

Table 2.4: Computation time required to estimate a *P*-value of .0001 with standard error ≤ .00001

	P_{ACT}	P_{DSA}		P_{perm}	
Correlation structure	(any *N*)	(any *N*)	*N* = 200	*N* = 1000	*N* = 2000
200 autocorrelated SNPs	3.54 s	212 s	1.75 hrs	10.8 hrs	13.9 hrs
HNF1A with 20 SNPs	0.71 s	43.8 s	825 s	2044 s	1 hr
20,000 *HNF1A*s with 20 SNPs each	3.94 hrs	10.1 days	0.52 yrs	1.29 yrs	2.28 yrs

Figure 2.1: Bivariate normal probability represented by P_{ACT} when $L = 2$ for i) one-sided tests, and ii) two-sided tests


i)
ii)
Z2
Z1
$\Phi^{-1}(1 - p_{min})$
Z2
Z1
$\Phi^{-1}\left(1 - \frac{p_{min}}{2}\right)$
$-\Phi^{-1}\left(1 - \frac{p_{min}}{2}\right)$

Note – Elliptical lines represent the contours of a bivariate normal density function with positive correlation. Shaded area represents the space (extending to infinity) over which the probability P_{ACT} is measured.

Figure 2.2: Linkage disequilibrium (r^2) between 20 SNPs from *HNF1A*

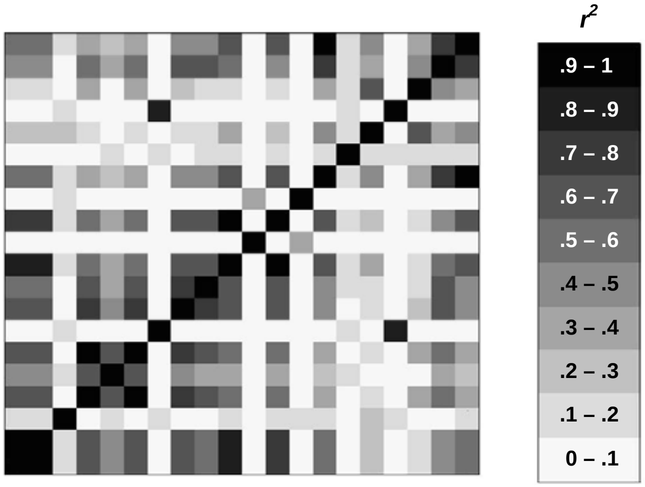


Figure 2.3: Correlation structures used in simulations of 10 correlated traits

Uncorrelated traits

$$\mathbf{R_Y} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Equal correlation between traits

$$\mathbf{R_Y} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

Autocorrelated traits

$$\mathbf{R_Y} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^9 \\ \rho & 1 & \rho & \cdots & \rho^8 \\ \rho^2 & \rho & 1 & \cdots & \rho^7 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^9 & \rho^8 & \rho^7 & \cdots & 1 \end{bmatrix}$$

Independent blocks of correlated traits

$$\mathbf{R_Y} = \left[\begin{array}{cccccc|ccc|c} 1 & \rho & \rho & \rho & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & 1 & \rho & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & \rho & 1 & \rho & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & \rho & \rho & 1 & \rho & 0 & 0 & 0 & 0 \\ \rho & \rho & \rho & \rho & \rho & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & \rho & \rho & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \rho & \rho & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Negatively correlated blocks of correlated traits

$$\mathbf{R_Y} = \left[\begin{array}{cccccc|ccc|c} 1 & \rho & \rho & \rho & \rho & \rho & -\rho & -\rho & -\rho & 0 \\ \rho & 1 & \rho & \rho & \rho & \rho & -\rho & -\rho & -\rho & 0 \\ \rho & \rho & 1 & \rho & \rho & \rho & -\rho & -\rho & -\rho & 0 \\ \rho & \rho & \rho & 1 & \rho & \rho & -\rho & -\rho & -\rho & 0 \\ \rho & \rho & \rho & \rho & 1 & \rho & -\rho & -\rho & -\rho & 0 \\ \rho & \rho & \rho & \rho & \rho & 1 & -\rho & -\rho & -\rho & 0 \\ \hline -\rho & -\rho & -\rho & -\rho & -\rho & -\rho & 1 & \rho & \rho & 0 \\ -\rho & -\rho & -\rho & -\rho & -\rho & -\rho & \rho & 1 & \rho & 0 \\ -\rho & -\rho & -\rho & -\rho & -\rho & -\rho & \rho & \rho & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

Figure 2.4: Estimates of P_{ACT} and P_{perm} for 3007 SNPs tested for disease association under 3 genetic models

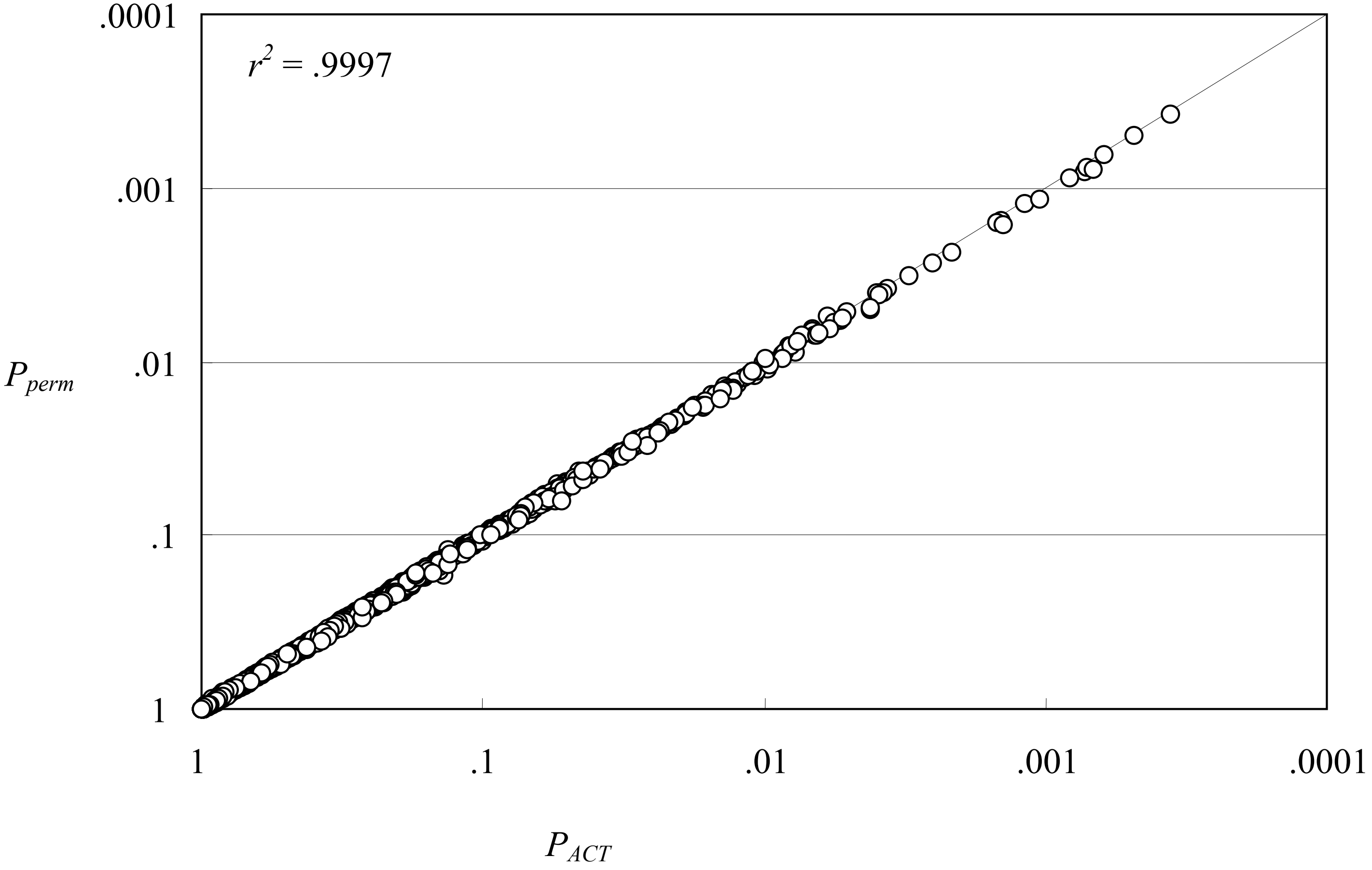
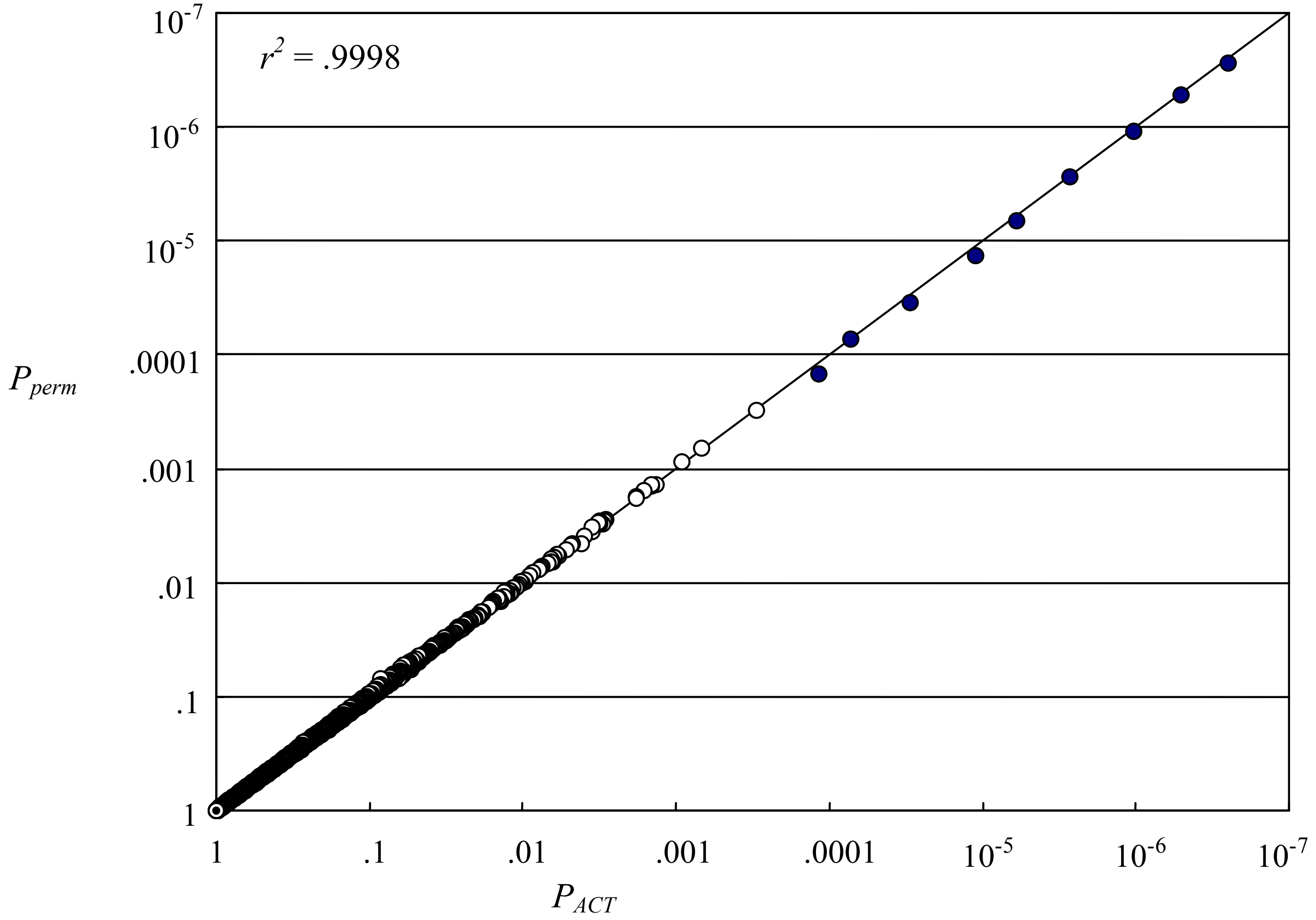


Figure 2.5: Estimates of P_{ACT} and P_{perm} for 3584 SNPs tested for association with 18 quantitative traits



Note – Open circles represent 3575 SNPs genotyped for candidate gene study. Filled circles represent 9 simulated SNPs.

CHAPTER 3

ADJUSTMENT OF META-ANALYSES FOR CORRELATED TESTS

The recent wave of large-scale genetic association studies has led to a host of positive genetic association results. The need for validation through testing in independent samples has in turn led to an increased focus on meta analysis. Meta analyses of genetic association studies based on multiple SNPs and traits are subject to the same multiple testing issues as single-sample studies, but depending on the study design, it is generally more difficult to adjust for these tests. Procedures such as Bonferroni may control the type I error rate, but will generally provide an overly harsh correction given the likely correlation between tests, while permutation testing is often not possible in a meta-analysis framework. We present methods of adjusting for multiple correlated tests for four study designs which are commonly employed in meta analyses of genetic association tests. We show through simulation that these methods accurately control the rate of type I error and achieve improved power over multiple testing adjustments which do not account for correlation.

3.1 Introduction

In Chapter 2 we described P_{ACT}, a multiple-testing adjustment which provides a faster alternative to permutation testing and accounts for the correlation between tests. P_{ACT} can be used to adjust the most significant *P*-values or test statistics from tests of *K* traits for association with *M* genetic variants. We showed that in a generalized linear model framework, the $K \times M$ test statistics had asymptotic distribution $N(0, \mathbf{R})$, where $\mathbf{R}$ is the correlation matrix corresponding to the covariance matrix $\mathbf{V}_{\beta}$, the Kronecker product of the sample covariance matrices of traits and genotypes, conditioned on covariates. Here we show that this result readily extends to meta analyses – an important case where permutation testing is often difficult due to coordination of analyses across centers, and may not be possible when only SNPs passing specific criteria are followed up in additional samples.

For tests of K traits for association with M markers in J independent samples, several common meta test statistics are available. The test statistic of Mantel and Haenszel (1959) is often applied, although an extension to tests of trend (Mantel 1963) is also available and is more general. A trend test assesses the impact of a dose variable with integer levels $0,\ldots,D$ on a binary response variable with the score statistic from a logistic regression (Cochran 1954; Armitage 1955). For $D = 1$ this test is equivalent to a simple chi-square test for independence. Trend test statistics are also commonly used to test SNPs for association with a binary trait under an additive model of association (with $D = 2$ and doses representing allele counts).

Methods based on weighted sums of the test statistics are applicable to a wider variety of models since they require only that the distribution of the summed test statistics is known (which is generally the case when test statistics are χ_1^2 or normally distributed and are summed across independent samples). These methods can be applied to tests with quantitative or binary response variables, and dose variables that are continuous rather than integer-valued; they can also be applied to a variety of more complicated models. Weighted-sum methods are especially useful in genetic studies which may involve quantitative traits, environmental and demographic covariates, and continuous dose variables such as imputed genotype dosage scores (Li et al. 2007).

Since it is common in multi-sample studies for the availability of SNP or trait data to vary across samples, we address this issue in the context of P_{ACT}. Data may be unavailable in certain samples pre-hoc due to study design, constrained resources, or failed assays. This type of unavailability is similar to missingness-at-random, and we show that it is quite easy to deal with in the context of computing meta-statistics and adjusting them for multiple testing with P_{ACT}.

SNP or trait data may also be missing not-at-random for an entire sample or samples. Multi-sample studies often consist of an initial sample, on which many tests may be performed, and one or more follow-up samples, which only test SNPs or traits passing a pre-set significance criterion in the initial study. Skol et al. (2006, 2007) present a weighted-sum meta test statistic that accounts for the conditional selection of SNPs for follow-up samples. Here, we show that P_{ACT} can be used to adjust for multiple correlated tests in this context, as well as in the case where only the best SNP is chosen for follow-up.

We show through simulation that our method provides a valid adjustment for correlated meta-analysis statistics in a number of situations: with Mantel-Haenszel and weighted-sum meta statistics, with binary or continuous traits, and with genotypes missing for an entire sample at random, through follow-up of only the best SNP, or through threshold-based selection of follow-up SNPs.

3.2 Methods

Below, we describe extensions of P_{ACT} that can be can be used for meta-analysis of multiple traits tested for association with multiple SNPs using either the Mantel-Haenszel trend test (which requires traits to be binary and uses a specific model) or a weighted-sum meta statistic (which allows binary or continuous traits and allows a general class of models). We then describe methods for handling the case where not all tests are performed in all samples (either at random or due to selection of specific tests for follow-up.)

3.2.1 Mantel-Haenszel extension to trend test

Given counts of n_{d1} cases and n_{d0} controls at each dose $d = 0,\ldots,D$, and defining marginal counts $n_{d+} = n_{d0} + n_{d1}$, $N_0 = \sum_{d=0}^{D} n_{d0}$, $N_1 = \sum_{d=0}^{D} n_{d1}$, and $N = N_0 + N_1$, the test statistic for a trend test (Cochran 1954; Armitage 1955) is:

$$\frac{\left(\sum_{d=0}^{D} dn_{d1} - E\left(\sum_{d=0}^{D} dn_{d1}\right)\right)^2}{V\left(\sum_{d=0}^{D} dn_{d1}\right)} = \frac{\left(\sum_{d=0}^{D} dn_{d1} - \frac{N_1}{N}\sum_{d=0}^{D} dn_{d+}\right)^2}{\frac{N_1 N_2}{N-1}\left(\frac{1}{N}\sum_{d=0}^{D} d^2 n_{d+} - \left(\frac{1}{N}\sum_{d=0}^{D} dn_{d+}\right)^2\right)} \mathrel{\dot{\sim}} \chi^2_1.$$

The corresponding Mantel-Haenszel meta-statistic for J samples is then

$$\frac{\left(\sum_{j=1}^{J}\sum_{d=0}^{D} dn_{d1,j} - \sum_{j=1}^{J}\frac{N_{1,j}}{N_j}\sum_{d=0}^{D} dn_{d+}\right)^2}{\sum_{j=1}^{J} V\left(\sum_{d=0}^{D} dn_{d1,j}\right)} \mathrel{\dot{\sim}} \chi^2_1,$$

where $n_{d1,j}$, $N_{1,j}$, and N_j are the counts of cases with dose d, total count of cases, and total count respectively for sample j (Mantel 1963).

If we express these counts in terms of Y_i, G_i, and $\mathbf{X}_\mathrm{i}$ where for individual i Y_i is a binary indicator variable equal to 1 for cases and 0 for controls, G_i is the allele count (0, 1, or 2 for an additive model, or 0, 1 for a dominant, recessive, or allele-based model) of a single SNP tested for association, and $\mathbf{X}_i = \left(X_{i1}, X_{i2}, \ldots, X_{iJ}\right)$ is a J-dimensional sample-indicator vector that contains 1 as the jth element and 0 for all other elements if individual i belongs to sample j, then $dn_{d1,j} = \sum_{i=1}^{N} Y_i G_i X_{ij} \cdot I\left(G_i = d\right)$ and $N_{1,j} = \sum_{i=1}^{N} Y_i X_{ij}$.

The Mantel-Haenszel meta test statistic can then be rewritten as

$$\frac{\left(\sum_{j=1}^{J}\sum_{i=1}^{N} Y_i G_i X_{ij} - \sum_{j=1}^{J} N_j \bar{Y}_j \bar{G}_j\right)^2}{\sum_{j=1}^{J} V\left(\sum_{i=1}^{N} Y_i G_i X_{ij}\right)} = \frac{\left(\sum_{i=1}^{N} Y_i G_i - \sum_{j=1}^{J} N_j \bar{Y}_j \bar{G}_j\right)^2}{\sum_{j=1}^{J}\sum_{i=1}^{N} V\left(Y_i\right) V\left(G_i\right) X_{ij}} \mathrel{\dot{\sim}} \chi^2_1.$$

under the null hypothesis that Y_i is independent of G_i, but not $\mathbf{X}_\mathrm{i}$, or equivalently,

$$\frac{\sum_{i=1}^{N} Y_i G_i - \sum_{j=1}^{J} N_j \overline{Y}_j \overline{G}_j}{\sqrt{\sum_{j=1}^{J} N_j {\omega_j}^2 {\sigma_j}^2}} \mathrel{\dot{\sim}} N\left(0,1\right) \tag{3.1}$$

where ${\omega_j}^2 = \overline{Y}_j\left(1-\overline{Y}_j\right)$ and ${\sigma_j}^2 = \frac{1}{N_j}\sum_{i=1}^{N} G_i^2 X_{ij} - {\overline{G}_j}^2$ are the respective trait and genotype variances estimated for sample j.

Both the single-sample and multiple-sample tests for trend shown above are generalized linear models. The single-sample trend test statistic provides a test of the null hypothesis of independence between Y_i and G_i $\left(\beta = 0\right)$ in the logistic model $E\left(Y_i \mid G_i\right) = h\left(\eta_i\right) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ where $\eta_i = \alpha + \beta G_i$ and α and β are scalar parameters. The multiple-sample Mantel-Haenszel statistic provides a test of the null hypothesis that Y_i and G_i are independent conditional on the sample indicator $\mathbf{X}_\mathrm{i}$ by testing $\beta = 0$ using a similar logistic model with fixed effects, where $\eta_i = \alpha^T \mathbf{X}_i + \beta G_i$ and α is now a J-dimensional parameter vector of sample fixed effects. The score statistic from this test is the numerator from equation (3.1):

$$\mathbf{U}_\beta = \sum_{i=1}^{N} Y_i G_i - \sum_{j=1}^{J} N_j \overline{Y}_j \overline{G}_j = \sum_{i=1}^{N}\left(Y_i - \tilde{Y}_i\right) G_i$$

where $\tilde{Y}_i = \overline{Y}_j$ is the predicted value of Y_i conditional on $\mathbf{X}_\mathrm{i}$, and denominator of equation (3.1) provides an estimate of the variance of this statistic $\mathbf{V}_\beta$.

While the above example deals with a single trait tested for association with a single genotype, we have shown in Section 2.2.2 that when multiple traits are tested for association with multiple markers in a single sample, the vector of score statistics

$\mathbf{U}_\beta \dot{\sim} N\left(0, \mathbf{V}_\beta\right)$. Below, we show that this multivariate normality result readily extends to the case of multiple samples in a meta analysis.

For *j* = 1, …, *J* independent samples, let $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, ..., Y_{iK}\right)$ be a vector of *K* binary trait variables for individuals *i* = 1, …, *N*. Let **G**i be an *M*-dimensional genotype vector containing the genotype codes for individual i for each of *M* markers. We assume for now that the same markers are genotyped for all *J* samples, an assumption that we will relax in later subsections. $\mathbf{X}_i = \left(X_{i1}, X_{i2}, ..., X_{iJ}\right)$ is the *J*-dimensional sample-indicator vector defined above. Our single-sample score statistic vector (see Section 2.2.2) is then:

$$\mathbf{U}_{\beta,j} = \sum_{i=1}^{N} X_{ij}\left(\mathbf{Y}_i - \overline{\mathbf{Y}}_j\right) \otimes \mathbf{G}_i \dot{\sim} N\left(0, \mathbf{V}_{\beta,j}\right),$$

where $\mathbf{V}_{\beta,j}$ can be estimated as $N_j \Omega_j \otimes \Sigma_j$, the Kronecker product between the sample covariance matrices of **Y**i and **G**i from sample *j*.

The vector of score statistics for the meta-analysis of all *J* samples may be obtained by summing the single-sample score statistics:

$$\mathbf{U}_\beta = \sum_{j=1}^{J} \mathbf{U}_{\beta,j} = \sum_{i=1}^{N} \mathbf{Y}_i \otimes \mathbf{G}_i - \sum_{j=1}^{J} N_j \overline{\mathbf{Y}}_j \otimes \overline{\mathbf{G}}_j,$$

and because the *J* samples are independent,

$$\mathbf{V}_\beta = \sum_{j=1}^{J} \mathbf{V}_{\beta,j} = \sum_{j=1}^{J} N_j \Omega_j \otimes \Sigma_j.$$

As the sum of *J* multivariate normal vectors, the vector of Mantel-Haenszel test statistics $\mathbf{U}_\beta \dot{\sim} N\left(0, \mathbf{V}_\beta\right)$, so we can compute P_{ACT} for the most extreme Mantel-Haenszel statistic using the correlations between the standardized test statistics $MH_{km} = \frac{\mathbf{U}_{\beta,km}}{\sqrt{\mathbf{V}_{\beta,km,km}}}$:

$$Cor\left(MH_{km},MH_{k'm'}\right)=\frac{\sum_j N_j \omega_{kk'}{}^{(j)}\sigma_{mm'}{}^{(j)}}{\sqrt{\sum_j N_j \omega_{kk}{}^{(j)}\sigma_{mm}{}^{(j)}\sum_j N_j \omega_{k'k'}{}^{(j)}\sigma_{m'm'}{}^{(j)}}}.$$

For an association test of a single binary trait with M markers, this reduces to

$$Cor\left(MH_{m},MH_{m'}\right)=\frac{\sum_j N_j \bar{Y}_j\left(1-\bar{Y}_j\right)\sigma_{mm'}{}^{(j)}}{\sqrt{\sum_j N_j \bar{Y}_j\left(1-\bar{Y}_j\right)\sigma_{mm}{}^{(j)}\sum_j N_j \bar{Y}_j\left(1-\bar{Y}_j\right)\sigma_{m'm'}{}^{(j)}}}.$$

3.2.2 Weighted-sum statistics

In this section, we show how P_{ACT} may used to adjust meta-statistics created as weighted sums of normally distributed test statistics from J independent samples. These statistics do not assume a specific model, so we address the common situation where all statistics are asymptotically equivalent to score statistics from GLMs – a general case which covers most types of normally distributed statistics estimated in genetic studies. This implies that if L tests are performed in each sample j = 1, …, J, the L-dimensional vector of standardized score statistics for sample j has a multivariate normal distribution:

$$\mathbf{T}^{(j)}=\left(\frac{\mathbf{U}_{\beta,1}{}^{(j)}}{\sqrt{\mathbf{V}_{\beta,11}{}^{(j)}}},\ldots,\frac{\mathbf{U}_{\beta,L}{}^{(j)}}{\sqrt{\mathbf{V}_{\beta,LL}{}^{(j)}}}\right)\dot{\sim} N\left(0,\mathbf{R}^{(j)}\right)$$

where $\mathbf{R}^{(j)}$, the covariance matrix of the test statistics, can be estimated as the sample correlation matrix between the traits and/or genotypes being tested (see Section 2.2.2). It is common to create a meta test statistic based on a weighted sum of normally distributed test statistics across samples. For example, a weighted sum of the standardized test statistics for test l across J samples (WS_l) would have the form $WS_l=\sum_j w_{jl}T_l^{(j)}$. Since the vector of weighted-sum meta statistics for the L tests $\mathbf{WS}=\left(WS_1,WS_2,\ldots,WS_L\right)$ is a weighted sum of J multivariate normal vectors, it will itself be distributed multivariate normally. The independence of the *J* samples guarantees that $Var\left(WS_l\right)=\sum_j w_{jl}^2$ for l = 1,…,*L*, and it is easy to show that the correlation between the meta statistics for tests *k* and *l* is $\sum_j w_{jk}w_{jl}\mathbf{R}_{kl}^{(j)}$. The square root of the sample proportion $\sqrt{\frac{N_j}{N}}$ is often used to weight standardized normal test statistics in which case $Var\left(WS_l\right)=1$ and the correlation between test statistics *k* and *l* is $\frac{1}{N}\sum_j N_j\mathbf{R}_{kl}$.

3.2.3 Missing data and incomplete genotyping across samples

To estimate P_{ACT} in the presence of missing data at the individual level, we have suggested in Section 2.4 that each association test can be performed using all available observations, and setting the individual component of the score statistic for missing observations to zero for purposes of variance estimation. In the context of genetic association studies, this would involve estimating variance on the full dataset with missing values for trait *k* set to the sample mean trait value $\overline{Y}_k$ and missing values for marker *m* set to the sample genotype mean $\overline{G}_m$.

A similar approach can be applied in a meta-analysis framework, where the complete set of traits or markers is not necessarily available in every sample. Assuming that the availability of traits or markers in certain samples is independent of results in other samples (an assumption which will be relaxed in the next subsection), unavailable traits or markers can simply be treated as missing data. If trait *k* is not analyzed in sample *j*, then the score statistics and variance can be computed with Y_{ik} set to zero for all individuals in sample *j*. Similarly, if marker *m* is not analyzed in sample *j*, then $\mathbf{U}_\beta$ and $\mathbf{V}_\beta$ can be calculated with G_{im} set to zero for all individuals in sample *j*, which will effectively set to zero the *m*th element of the score statistic $\mathbf{U}_{\beta,m}^{(j)}$ and all covariances and variances involving marker *m* in sample *j*. This allows computation of the meta test statistics and covariance estimates based only on samples with data for the relevant traits and markers while ensuring a positive-definite covariance matrix.

To demonstrate with a simple example, we consider the case where a single binary trait is tested for association with SNP A and SNP B in one sample and only with SNP A in a second sample (where SNP B can be assumed to be lost due to a random reason such as a failed assay). Assuming that sample 1 consists of individuals $i = 1, \ldots, N_1$ and sample 2 consists of individuals $i = N_1 + 1, \ldots, N$, the vector of score statistics for the Mantel-Haenszel tests will be:

$$\mathbf{U}_\beta = \begin{bmatrix} \sum_{i=1}^{N} Y_i G_{iA} - N_1 \overline{Y}_1 \overline{G}_A^{(1)} - N_2 \overline{Y}_2 \overline{G}_A^{(2)} \\ \sum_{i=1}^{N_1} Y_i G_{iB} - N_1 \overline{Y}_1 \overline{G}_B^{(1)} \end{bmatrix}.$$

The variance of the score statistic can be estimated as:

$$\mathbf{V}_\beta = \sum_{j=1}^{J} \overline{Y}_j \left(1-\overline{Y}_j\right)\Sigma_j = N_1 \overline{Y}_1 \left(1-\overline{Y}_1\right)\begin{bmatrix} \sigma_{AA}^{(1)} & \sigma_{AB}^{(1)} \\ \sigma_{AB}^{(1)} & \sigma_{BB}^{(1)} \end{bmatrix} + N_2 \overline{Y}_2 \left(1-\overline{Y}_2\right)\begin{bmatrix} \sigma_{AA}^{(2)} & 0 \\ 0 & 0 \end{bmatrix},$$

so the vector of standardized Mantel-Haenszel test statistics will be

$$\mathbf{MH} = \begin{bmatrix} MH_A \\ MH_B \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{i=1}^{N} Y_i G_{iA} - N_1 \overline{Y}_1 \overline{G}_A^{(1)} - N_2 \overline{Y}_2 \overline{G}_A^{(2)}}{\sqrt{N_1 \overline{Y}_1 \left(1-\overline{Y}_1\right)\sigma_{AA}^{(1)} + N_2 \overline{Y}_2 \left(1-\overline{Y}_2\right)\sigma_{AA}^{(2)}}} \\ \dfrac{\sum_{i=1}^{N_1} Y_i G_{iB} - N_1 \overline{Y}_1 \overline{G}_B^{(1)}}{\sqrt{N_1 \overline{Y}_1 \left(1-\overline{Y}_1\right)\sigma_{BB}^{(1)}}} \end{bmatrix}$$

Both the numerator and denominator of MH_A contain an additional component due to the N_2 additional individuals tested on SNP A. We can then adjust the two statistics for multiple testing using P_{ACT} and the fact that the correlation between MH_A and MH_B is

$$\frac{\sigma_{AB}{}^{(1)}}{\sqrt{\left(\sigma_{AA}{}^{(1)}+\frac{N_2\bar{Y}_2\left(1-\bar{Y}_2\right)}{N_1\bar{Y}_1\left(1-\bar{Y}_1\right)}\sigma_{AA}{}^{(2)}\right)\sigma_{BB}{}^{(1)}}}.$$

The correlation between the meta statistics is smaller than $\sigma_{AB}{}^{(1)}/\sqrt{\sigma_{AA}{}^{(1)}\sigma_{BB}{}^{(1)}}$, the correlation between the trend test statistics from sample 1, since the component of MH_A based on individuals $i = N_1 + 1, \ldots, N$ is independent of MH_B. The lower correlation will cause the P_{ACT} correction to be somewhat more severe, which is appropriate given that additional testing was performed.

Missing data and incomplete genotyping across samples can be handled similarly in meta statistics created as weighted sums of standard normal test statistics. If the sample proportion square roots $\sqrt{\frac{N_j}{N}}$ are used as weights, then we can simply establish different weights for each test $l = 1,\ldots,L$. Define the weight for test l as $\sqrt{\frac{N_{l,j}}{N_l}}$, where $N_{l,j}$ represents the number of non-missing observations for test l in sample j (and will be zero if a variable needed for test l is missing sample-wide) and $N_l = \sum_j N_{l,j}$. Using these weights, the weighted-sum meta statistic can then be computed as described in section 3.2.2.

3.2.4 Replication samples

The previous section addressed the case where traits or genotypes are unavailable in certain samples for reasons independent of any observed results. In the example presented, the decision to type only SNP A in the second sample was considered to have occurred independently of any observed results in the first sample, perhaps due to a failed assay. This is a common scenario in multi-sample studies, and fortunately it does not create much of a problem in the meta-analysis.

Another common scenario in meta-analyses involves the genotyping of many markers in an initial study, followed by the genotyping of only a subset of markers in follow-up studies. We show below that P_{ACT} can be used to control the study-wide type I error even when tests are correlated in two common scenarios: the case where the best result from a group of tests is followed up in additional samples, and the case where all tests exceeding a pre-set significance level are selected for follow-up.

Best result selected for follow-up – While not necessarily a best practice, it is a common practice to follow up automatically the most significant result in an initial study for further testing. If the best result is not especially strong, this can lead to wasted resources and publication bias. However, at least the latter problem can be avoided since the type I error of the resulting meta-analysis can be easily controlled with P_{ACT}.

If L tests are performed in an initial sample, and only the best result is followed up in $J - 1$ additional samples, it is important to adjust the initial best result for multiple testing before its inclusion in a meta statistic, but the remaining $J - 1$ samples are independent and do not require adjustment for multiple testing. In this situation, P_{ACT} can be computed to adjust the best result from the initial sample for multiple testing. If we then compute a Z-score from P_{ACT} as though it were a standard normal quantile:

$$Z_{ACT} = \pm\Phi^{-1}\left(1 - \frac{P_{ACT}}{2}\right),$$

where the sign matches that of the initial test statistic, we can then estimate the weighted-sum statistic as usual, but with Z_{ACT} in place of $T^{(1)}$:

$$WS = w_1 Z_{ACT} + \sum_{j=2}^{J} w_j T^{(j)}. \qquad (3.2)$$

Since P_{ACT} has a U(0,1) distribution under the null hypothesis of no association, $Z_{ACT} \sim$ $N(0,1)$ under the null hypothesis, so the type I error rate will be accurately controlled by this adjustment. Appendix A shows that under the alternative hypothesis, Z_{ACT} is generally not normally distributed. If the best result in sample 1 is the result of a true association, 1) the expected value of Z_{ACT} will be larger than the actual effect size, since this result is only observed conditional on being larger than all other test statistics, and 2) the variance of Z_{ACT} will be < 1. Both 1) and 2) will be increasingly true for larger *L*. As we show in Appendix A, the sample proportion square roots $\sqrt{\frac{N_j}{N}}$ are not necessarily the power-maximizing weights in this situation, and there is no closed form solution for the optimal weights. A recursive solution can be obtained, but will depend on the effect size (which is unknown) and desired critical value, as well as several integrals which depend on *L* and require numerical rather than analytical computation. Furthermore, the analysis in Appendix A assumes independent tests; this problem will naturally become even more complicated in the presence of correlation. A simpler solution would be to employ a more appropriate and powerful study design (see *Two-stage design* below). Hence, we do not belabor this point but instead investigate the efficacy of this method using the sample proportion square roots $\sqrt{\frac{N_j}{N}}$ as our weights – a strategy which is less powerful than the hypothetical alternative strategy involving optimal weights, but should lead to unbiased type I error rates.

Two-stage design – A generally more appropriate study design involves selecting tests for follow-up based on pre-set criteria. In a genetic association context, two-stage design involves the selection of all markers with test statistics exceeding a carefully chosen cutoff for follow-up analysis. Skol et al. (2006, 2007) show that if $z_1 \sim N\left(0,1\right)$ is a test statistic observed in an initial study and only SNPs for which $\left|z_1\right| > T_1$ are tested in a replication study, the conditional probability of the weighted meta-statistic $z_{\text{joint}} = \sqrt{\frac{N_1}{N}} z_1 + \sqrt{\frac{N_2}{N}} z_2$ reaching significance, $P_{\text{joint}} = P\left(\left|z_{\text{joint}}\right| > T_{\text{joint}} \,|\, \left|z_1\right| > T_1\right)$, can be obtained through integration of the conditional normal CDF. The overall *P*-value for the joint analysis is then $P_1 P_{\text{joint}}$, where $P_1 = P\left(\left|z_1\right| > T_1\right)$ and P_1 and P_{joint} are both computed under the null hypothesis.

A similar method can be applied to a meta-analysis involving correlated tests, and P_{ACT} can be used conditionally in this context to adjust for the correlation between tests while taking the conditional selection of tests into account. The appropriate adjusted *P*-value for the best observed meta statistic is the joint probability that under the null hypothesis, at least one of the *L* tests 1) passes the pre-set cutoff in the initial sample, and 2) is at least as extreme as the best observed meta statistic T_{joint}:

$$P_{ACT-2s} = P\left(\left|z_{l,1}\right| > T_1, \left|z_{l,\text{joint}}\right| > T_{\text{joint}} \text{ for some } l \in \{1,2,...,L\}\right) \tag{3.3}$$

Defining the *L* initial test statistics as $z_{1,1}, z_{2,1}, ..., z_{L,1}$ and the *L* joint test statistics (which are only observed if the corresponding initial statistic passes the cutoff T_1) as $z_{1,\text{joint}}, z_{2,\text{joint}}, ..., z_{L,\text{joint}}$, the set of 2*L* initial and joint statistics has a multivariate normal distribution with covariance matrix

$$\left[\begin{array}{c|c} \mathbf{R}^{(1)} & w_1\mathbf{R}^{(1)} \\ \hline w_1\mathbf{R}^{(1)} & \sum w_j{}^2\mathbf{R}^{(j)} \end{array}\right]$$

where $\mathbf{R}^{(j)}$ and w_{j} are the sample correlation matrices and weights for sample *j* as defined in section 3.2.2. $P_{ACT\text{-}2s}$ (equation 3.3) can then be computed as a sum of multivariate normal probabilities or a much less computationally intensive approximation which yields near-identical results (see Appendix B).

3.2.5 Simulations

To assess the validity and power of the Mantel-Haenszel (MH) and weighted-sum (WS) meta statistics, we simulated haplotype association tests in an initial sample and four replication samples. In each simulation, we randomly drew 6-SNP haplotypes from a larger sample of individual haplotypes. To simulate realistic variation between samples, we drew haplotypes from 5 different samples collected as part of a case-control meta-analysis; this allowed the correlation between haplotypes, and hence between haplotype association tests, to vary between samples as it would in a typical meta-analysis. Haplotypes were inferred in the original data using MACH 1.0 (Li et al. 2007). One of the five samples contained data on only 5 of the 6 SNPs; for this sample, we also used MACH to impute data for this SNP based on haplotype frequencies from the HapMap (CEU sample). For the initial sample, which contained data on 2632 haplotypes, we randomly drew 2000 haplotypes to represent 1000 individuals. In the follow-up samples, which included 1404, 1810, 1818, and 2206 haplotypes, respectively, we drew haplotypes for 500, 650, 750, and 900 individuals. We simulated a single binary trait, as described below.

In the simulated initial sample, we tested each common haplotype involving between one and six SNPs for association with a Cochran-Armitage test for trend, for a total of 170 unique tests. 169 of these tested haplotypes were present and polymorphic in the four follow-up samples. We tested these 169 haplotypes for association with the binary trait in the four follow-up samples and used P_{ACT} to adjust the most significant *P*-values in each simulated meta-analysis.

We assessed type I error by performing 10,000 simulations where the binary trait tested for association was assigned independently of genotype, and we compared the 10,000 observed *P*-values to the expected sample quantiles. For assessment of power, we created 1000 simulation replicates where the binary trait was influenced by a single haplotype. To do this, we simulated the binary trait y_{ij} for individual *i* in sample *j* such that $y_{ij} = 1$ if $Z_{ij} + \beta_j \left(G_{im}{}^{(j)} - \overline{G}_m{}^{(j)} \right) > 0$ and 0 otherwise, where Z_{ij} is a normal random variable, $G_{im}{}^{(j)}$ is the number of copies of haplotype m possessed by individual i, with mean $\bar{G}_m^{(j)}$, and β_j is an effect size drawn from the uniform distribution for each sample, to allow for heterogeneity of effects between samples.

We computed type I error and power for four different study designs. In the first, all 169 haplotypes were tested on all 5 samples. We computed Mantel-Haenszel and weighted-sum meta statistics for each of the 169 haplotypes, and adjusted the most significant meta statistic for association using P_{ACT} as described in section 3.2.2.

For the second study design, we assumed that SNPs in the four follow-up samples were unavailable for random reasons, such as failed assays. We allowed each SNP to be missing in each follow-up sample with probability 0.1. A missing SNP meant that all haplotype combinations containing this SNP were missing and thus unavailable for testing. On average, samples were missing 0, 1, 2, 3, 4, or 5 SNPs 53%, 35%, 10%, 1.4%, 0.1%, and 0.002% of the time. We then computed the meta statistics and P_{ACT} assuming data were missing-at-random, as described in section 3.2.3.

For the third study design, we assumed that only the best haplotype from the first sample was followed up in subsequent samples. We computed P_{ACT} to adjust the best P-value from the first sample for 169 correlated tests, and computed the weighted sum of the $N(0,1)$ test statistics from the 5 samples as in equation (3.2), using $Z_{ACT} = \pm\Phi^{-1}\left(1-\frac{P_{ACT}}{2}\right)$ as the statistic for sample 1.

For the fourth study design, we assumed a two-stage design where only haplotypes with individual association test P-values < .1 were re-tested in the four follow-up samples. In practice the number of haplotypes passing the criteria for further testing in each simulation under the null hypothesis ranged from 0 (in 20% of simulations) to 108, with a median of 10. For the haplotypes which were tested in all samples, we computed meta test statistics as described in section 3.2.2 and adjusted the most extreme meta test statistic for multiple correlated tests and conditional haplotype selection with P_{ACT} as in equation (3.3), using the approximation described in Appendix B.

3.3 Results

3.3.1 Type I error rate for different study designs

We computed meta test statistics for 169 haplotypes tested for association in 10,000 simulated 5-sample meta-analyses. For each simulation, we adjusted the best meta *P*-value in each simulation for multiple testing with P_{ACT} as described in Section 3.2.2. Adjusted *P*-values for Mantel-Haenszel tests and weighted-sum test statistics are plotted against their theoretical quantiles in Figure 3.1 (i). Values of P_{ACT} follow the identity line quite closely for the entire range of *P*-values, indicating that the appropriate type I error rate is maintained at all levels of significance. The distributions of unadjusted *P*-values (the best meta *P*-value in each simulation before adjustment for multiple testing) and Šidák-adjusted *P*-values are also plotted, demonstrating that relying on unadjusted *P*-values would lead to greatly inflated rates of type I error (for example, .62 at an α-level of .05), while Šidák-adjusted *P*-values provide conservative tests, with a type I error rate of .01 at an α-level of .05.

We next treated the five samples as an initial sample and four follow-up samples, and we allowed the six SNPs underlying the haplotypes to be missing-at-random (MAR) in the follow-up samples, where each SNP in each sample had a .1 probability of being MAR and a missing SNP meant that all haplotype combinations involving alleles of that SNP were missing as well. We computed the meta statistics and P_{ACT} for both Mantel-Haenszel and weighted-sum test statistics as described in Section 3.2.3, and plotted P_{ACT} against its quantiles in Figure 3.1 (ii). Again, P_{ACT} demonstrates a near one-to-one relationship with its quantiles, indicating that the appropriate type I error rate can be achieved for any α-level.

Returning to the complete data case (ie, no SNPs missing-at-random), we considered two cases where only key results are selected for follow-up. In the first case, we allowed only the haplotype combination showing the strongest association in the first sample to be followed up in the next four samples. We computed P_{ACT} for each simulated meta-analysis as in equation (3.2) and plotted the values against their quantiles in Figure 3.2 (i). As above, P_{ACT} tracked the quantiles very closely both when Mantel-Haenszel test statistics and weighted-sum statistics were used, indicating that the correct type I error rate is achieved for all α-levels.

Finally, we performed the meta-analysis with a two-stage design, where all haplotypes with association *P*-values < .1 in the initial sample were tested in the four follow-up samples. We computed P_{ACT} for each simulation as in equation (3.3) and plotted the values against their quantiles in Figure 3.2 (ii). For simulations where no haplotype had a *P*-value < .1, which was the case in 21.5% of simulations, no meta-analysis was performed, so the meta *P*-value in these cases is set to 1. For simulations where a meta-analysis was performed, P_{ACT} once again tracks its quantiles quite closely, indicating that it achieves the correct type I error rate for all reasonable α-levels.

3.3.2 Power for different study designs

A comparison of the power of the methods described above is presented in Figure 3.3 for 1000 simulated meta-analyses based on (i) Mantel Haenszel test statistics and (ii) weighted-sum test statistics. For comparison purposes, the leftmost pair of bars represents power for the single-sample analysis. All four meta-analysis study designs considered here demonstrate a clear gain in power over the single-sample analysis, with the complete data meta-analysis (all SNPs genotyped and all haplotypes tested) showing the greatest gains. Due to substantial correlation between tests, which is typical in haplotype analysis, adjustment for multiple testing with P_{ACT} leads to large gains in power over Bonferroni or Šidák adjustment, with gains of nearly 80% for the scenario where only the best haplotype combination is followed up.

Although the sample proportion weights were not necessarily the optimal weights for the study design which follows up only the best result, our implementation of P_{ACT} for this study design did show substantial gains in power over the alternatives of Šidák correction or P_{ACT} applied to just the first sample (which is equivalent to assigning a weight of 1 to the first sample and zero to all subsequent samples.) To investigate the effect of increasing or decreasing w_1, the weight placed on the adjusted test statistic from the initial sample, we re-ran our power simulations with w_1 increased or decreased by .02, .05, or .10. In each case, we adjusted the weights for other samples proportionally so that the squared weights continue to sum to one. In each set of simulations, we observed that there was a cutoff, which we denote α_{switch}, such that for critical values < α_{switch}, greater power was achieved with decreased values of w_1, while for critical values > α_{switch}, greater power was achieved with increased values of w_1. This seems surprising, but is consistent with the relationship between the critical value and optimal weights implied by the analysis in Appendix A. The level of α_{switch} varies depending on L and on whether we simulated correlated or independent tests, but it often fell within the range of reasonable critical values. For example, for 10 independent tests and the same effect sizes as in Figure 3.3, α_{switch}, ≈ .01, implying that if the target type I error rate is in the .01 – .05 range, it is more powerful to increase w_1, but if the desired type I error rate < .01, then it is more powerful to decrease w_1. For the same effect size, α_{switch} increases with L and decreases if tests are correlated.

Next, to assess whether application of P_{ACT} to these four study designs alters the rank ordering of significant results, we compare estimates of P_{ACT} from the weighted-sum meta analysis shown in Figure 3.3 (ii), to the estimates of the Šidák P-values. Comparisons are plotted in Figure 3.4. For the full meta-analysis case where every test is performed in every sample (Figure 3.4 i), the relationship is one-to-one. Although the

Šidák test is clearly more conservative in the presence of correlation, the rank ordering of significant *P*-values remains the same whether P_{ACT} or a Šidák adjustment is used. When data are missing-at-random (ii), the pattern is similar but the relationship is no longer one-to-one due to the noise generated by the missing data. When only the best result is followed up (Figure 3.4 iii), there are some simulations where P_{ACT} is significant even though the Šidák-adjusted *P*-value is 1. In this case, the Šidák adjustment does more poorly than usual because it is essentially adjusting the meta-statistic for the number of tests performed in the first sample, even though only a single test is performed in the remaining samples. For two-stage analysis, the rank ordering of P_{ACT} and the Šidák-adjusted *P*-values is again the same. Similar results were observed for the Mantel-Haenszel statistics (data not shown).

3.4 Discussion

We have presented a new set of tools for adjustment of multiple correlated association tests in meta-analyses for a variety of common study designs and for any number of independent samples. In simulations of a large number of highly correlated haplotype association tests, our methods attained the appropriate type I error rates for a range of study designs and were substantially more powerful than Šidák adjustments, which do not account for correlation between tests.

For the four study designs we discussed, three have pre-hoc designs and are generally considered good research practice (two-stage design and the full meta-analysis with and without missingness-at-random), while the one post-hoc study design – automatic follow-up of the most significant result – is inefficient and likely to lead to wasted resources. It is also potentially biased if the multiple tests performed on the initial sample are not properly adjusted for. Since this study design is more likely accident than design, arrived at for a variety of reasons including human nature, we felt it was still worthwhile to provide an adjustment that would allow studies falling into this trap to be dug back out, at least partially. Controlling the type I error rate will prevent false positives and further devotion of resources to unpromising results; however, it cannot compensate for the decreased power of this study design compared to alternatives such as a two-stage design. Nonetheless, this study design did surprisingly well in terms of power under the simulated situations we observed, especially considering that the weights used (square root of sample proportion) were not necessarily power-maximizing for our meta test statistic. Recursive computation of the optimal weights would require assumption of an effect size, as well as specification of a particular critical value to be used and other computations that depend on the number of tests and the extent of correlation between the tests. Whether this could be done presents an interesting question for future research. Another relevant question is whether such an endeavor would even be worthwhile when better study designs requiring fewer assumptions are available. As replication of significant results becomes more and more routine, meta-analyses are often incorporated into the initial planning of studies, thus averting the pitfalls of post-hoc design.

In conclusion, well-designed meta-analyses have become increasingly common given the wealth of available data and the drive to combine information as a means of affirming valid results and ruling out spurious ones. Given the current emphasis placed on replication of important association results in independent samples, we feel that these methods are timely and have the potential to be useful in a variety of settings.

Appendix A: Distribution of Z_{ACT} under the alternative hypothesis for independent test statistics

Consider a test statistic $Z_1 \sim N(a,1)$ and L - 1 independent test statistics $Z_2, \ldots, Z_{\mathrm{L}} \sim N(0,1)$. Define $Z_{|\max|}$ as the test statistic with the largest absolute value. Since the tests are independent, P_{ACT} will be asymptotically equivalent to the Šidák-adjusted P-value:

$$P_{ACT} \approx 1-\left(2\Phi\left(\left|Z_{|\max|}\right|\right)-1\right)^{L}$$

and we can define

$$Z_{ACT} = sign\left(Z_{|\max|}\right)\Phi^{-1}\left(1-\frac{P_{ACT}}{2}\right).$$

It can then be shown through the appropriate transformation that

$$f\left(Z_{ACT} \mid Z_{|\max|} = Z_1\right) = \frac{\phi\left(Z_1-a\right)}{L\int\limits_0^\infty\left[\phi\left(x-a\right)+\phi\left(x-a\right)\right]\left(2\Phi\left(x\right)-1\right)^{L-1}dx}\frac{\phi\left(Z_{ACT}\right)}{\phi\left(Z_1\right)},$$

which is $N(0,1)$ under the null hypothesis that all L test statistics have mean 0, since $a = 0$ in this case. Under the alternative hypothesis, $a \neq 0$, so this distribution is not normal, but the mean and variance of Z_{ACT} can be obtained through a Taylor series expansion:

$$E\left(Z_{ACT} \mid Z_{|\max|} = Z_1\right) = \beta\left(L\right)a$$

$$\text{where } \beta\left(L\right) = 2\int\limits_0^\infty x\Phi^{-1}\left(\frac{1+\sqrt[L]{2\Phi\left(x\right)-1}}{2}\right)\phi\left(x\right)dx;\quad \beta\left(L\right)>1,\ \beta'\left(L\right)>0$$

$$Var\left(Z_{ACT} \mid Z_{|\max|} = Z_1\right) = 1-\delta\left(L\right)a^2 \text{ where } \delta\left(L\right)>0,\ \delta'\left(L\right)>0,$$

where $\delta\left(L\right)$ is the sum of $\beta\left(L\right)^2$ and two additional integrals which depend on L and lack closed-form solutions.

Now consider the case where $Z_1, \ldots, Z_L$ and $Y_1, \ldots, Y_L$ are realizations of the same L independent test statistics tested on two different samples of n_1 and n_2 individuals:

$$Z_1 \sim N\left(\sqrt{n_1}\varepsilon,1\right),\ Z_2,\ldots,Z_L \sim N\left(0,1\right)$$
$$Y_1 \sim N\left(\sqrt{n_2}\varepsilon,1\right),\ Y_2,\ldots,Y_L \sim N\left(0,1\right)$$

If we only performed one test in each sample, the weighted sum statistic $\lambda Z_l + \sqrt{1-\lambda^2}Y_l$ would be distributed $N\left(\left(\lambda\sqrt{n_1}+\sqrt{1-\lambda^2}\sqrt{n_2}\right)\varepsilon,1\right)$for *l* = 1 or *N*(0,1) for *l* = 2, …, *L*. We can then maximumize power while maintaining the appropriate type I error rate by choosing $\lambda = \sqrt{\dfrac{n_1}{n_1+n_2}}$.

If instead we perform all *L* tests in the first sample, and then follow up the test with the most extreme test statistic, the simple weighted-sum statistic $\lambda Z_l + \sqrt{1-\lambda^2}Y_l$ (where $Z_l = Z_{|\max|}$) is not *N*(0,1) under the null hypothesis. The adjusted statistic $\lambda Z_{ACT} + \sqrt{1-\lambda^2}Y_l \sim N\left(0,1\right)$ under the null hypothesis, allowing control of the type I error rate. Under the alternative hypothesis, the adjusted statistic has mean $\left(\lambda\beta\sqrt{n_1}+\sqrt{1-\lambda^2}\sqrt{n_2}\right)\varepsilon$ and variance $1-\delta a^2\lambda^2$. The inflated mean alone would imply an increase in the optimal weight for sample 1: $\lambda = \sqrt{\dfrac{\beta n_1}{\beta n_1 + n_2}}$. However, due to the presence of λ in the variance, there is no longer a closed-form solution for the optimal λ. Optimal weights can be obtained through recursion, but they will depend on the unknown quantity *ε*, as well as *β*, *δ*, n_1, n_2, and the desired significance level *α*.

Appendix B: Approximation for $P_{ACT\text{-}2s}$

The probability $P_{ACT-2s} = P\left(\left|z_{l,1}\right| > T_1, \left|z_{l,\text{joint}}\right| > T_{\text{joint}} \text{ for some } l \in \{1,2,...,L\}\right)$ from equation (3.3) may be computed as 1 minus the piecewise sum of the probabilities of all possible events where there is no $l \in \{1,2,...,L\}$ for which both $\left|z_{l,1}\right| > T_1$ and $\left|z_{l,\text{joint}}\right| > T_{\text{joint}}$. This method requires the computation of 3^L separate probabilities, and hence is feasible only for small *L*. However, if the correlation between tests in the initial sample $\mathbf{R}^{(1)}$ is similar to $\sum w_j^{\,2}\mathbf{R}^{(j)}$, the weighted sums of all correlations, a good approximation to (3.3) is available.

To perform this approximation, we adjust the minimum *P*-value P_{min} from the 2-stage meta-analysis in two steps. We first adjust for the 2-stage test by computing the probability $P' = P\left(\left|z_1\right| > T_1, \left|z_{\text{joint}}\right| > \left|T_{\text{joint}}\right|\right)$ that a single test passes both the initial cutoff in the first sample and attains the magnitude of the best test statistic observed in the combined samples, T_{joint}. This probability can be easily computed as the sum of four probabilities using the fact that the joint distribution of z_1 and z_{joint} is bivariate normal with correlation w_1. We can then convert this probability to an adjusted test statistic $T' = \left(1 - P'/2\right)$. For the second step, we adjust T' for the *L* tests that were performed by computing P_{ACT} assuming *L* tests with correlation matrix $\mathbf{R}^{(1)}$.

To test the performance of the approximation, we computed P_{ACT-2s} using both the exact method and the approximation described above for 1000 simulations of a reduced-dimension version of the analysis presented in Figure 3.2 (ii), where only 8 of the 169 haplotypes were included. As Figure 3.A.1 shows, we obtained near-identical results using the two methods, demonstrating that the approximation is very close in a situation with heterogeneous samples and high correlation between tests.

Figure 3.A.1 Comparison of $P_{ACT\text{-}2s}$ estimated for 8 correlated tests as either a piecewise sum of probabilities or a faster 2-step approximation

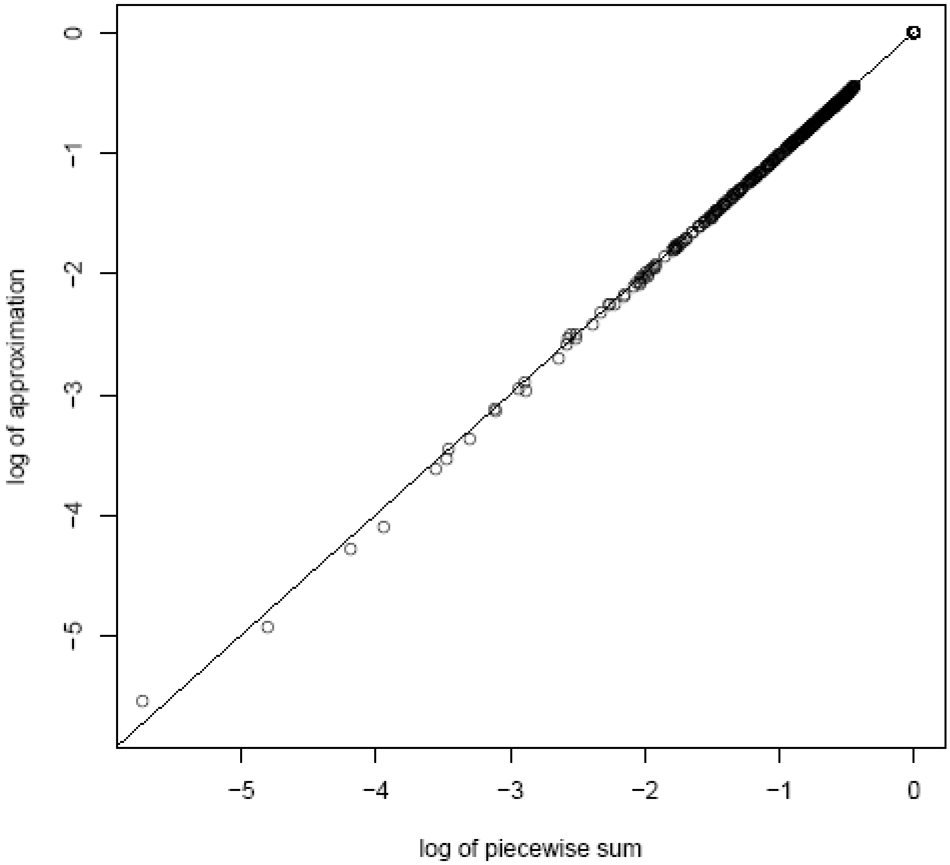
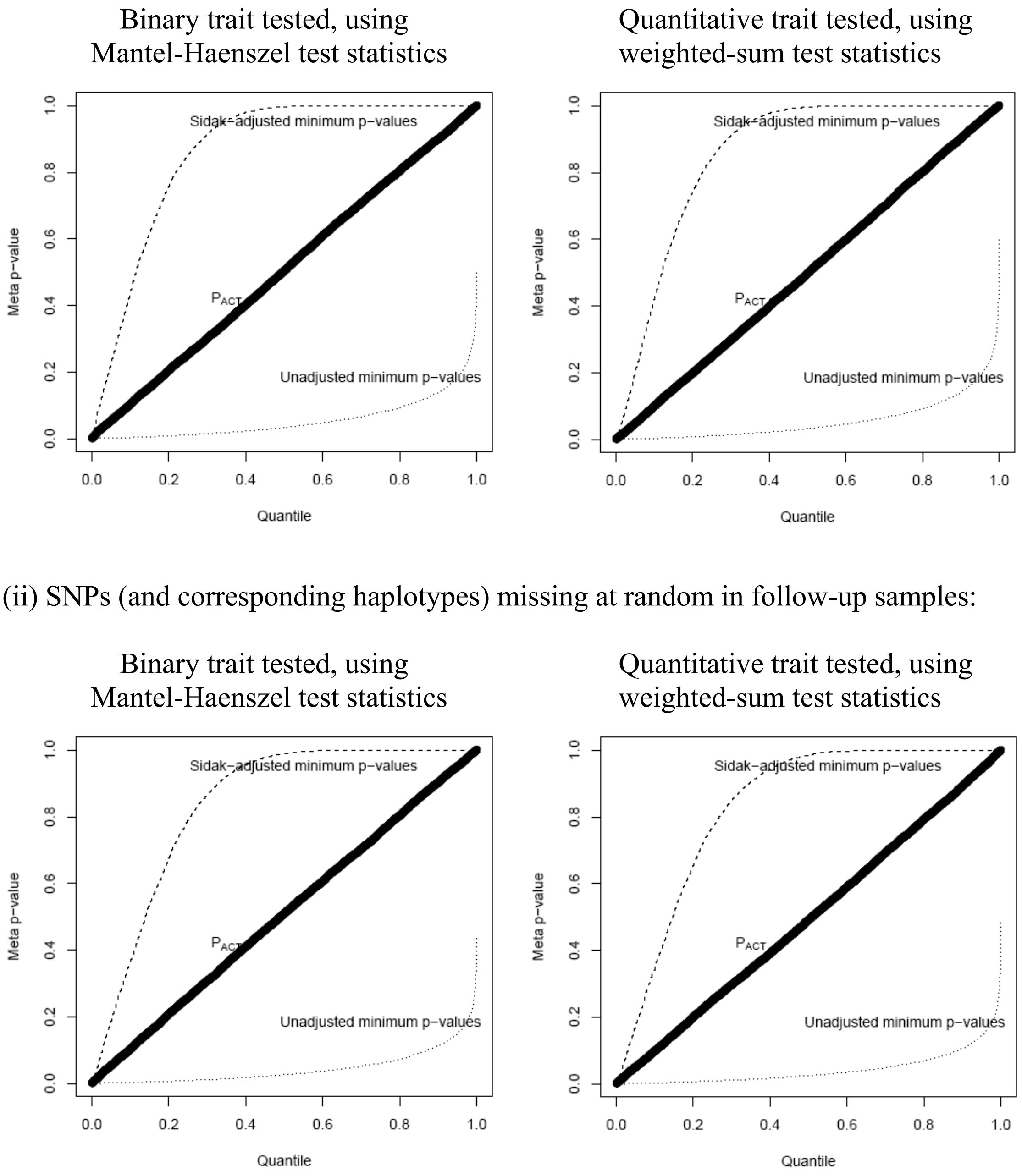


Figure 3.1: Quantiles of most significant *P*-values from 10,000 simulated meta-analyses. In each simulation, 169 haplotypes were tested in 5 samples and P_{ACT} was computed from a sample-size-weighted meta test statistic.



(i) All 169 haplotypes tested in initial sample and four follow-up samples using:

Binary trait tested, using Mantel-Haenszel test statistics

Quantitative trait tested, using weighted-sum test statistics

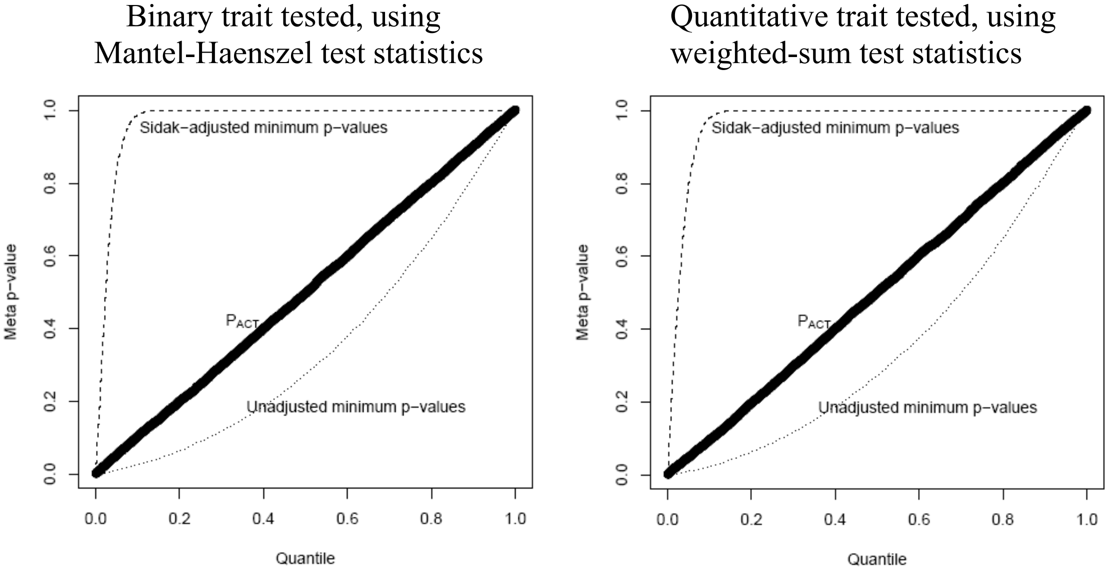
(ii) SNPs (and corresponding haplotypes) missing at random in follow-up samples:

Binary trait tested, using Mantel-Haenszel test statistics

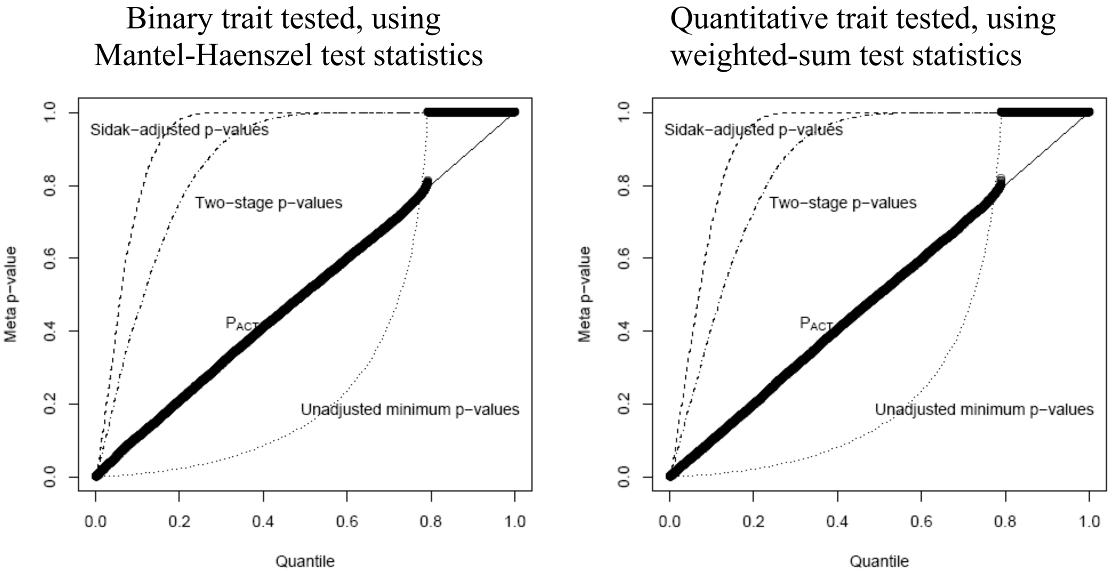
Quantitative trait tested, using weighted-sum test statistics

Figure 3.2: Quantiles of most significant P-values from 10,000 simulated meta-analyses. In each simulation, 169 haplotypes were tested in initial sample and selected haplotypes followed up in 4 additional samples.

(i) Best haplotype in initial sample followed up, P_{ACT} computed as in equation (3.2).

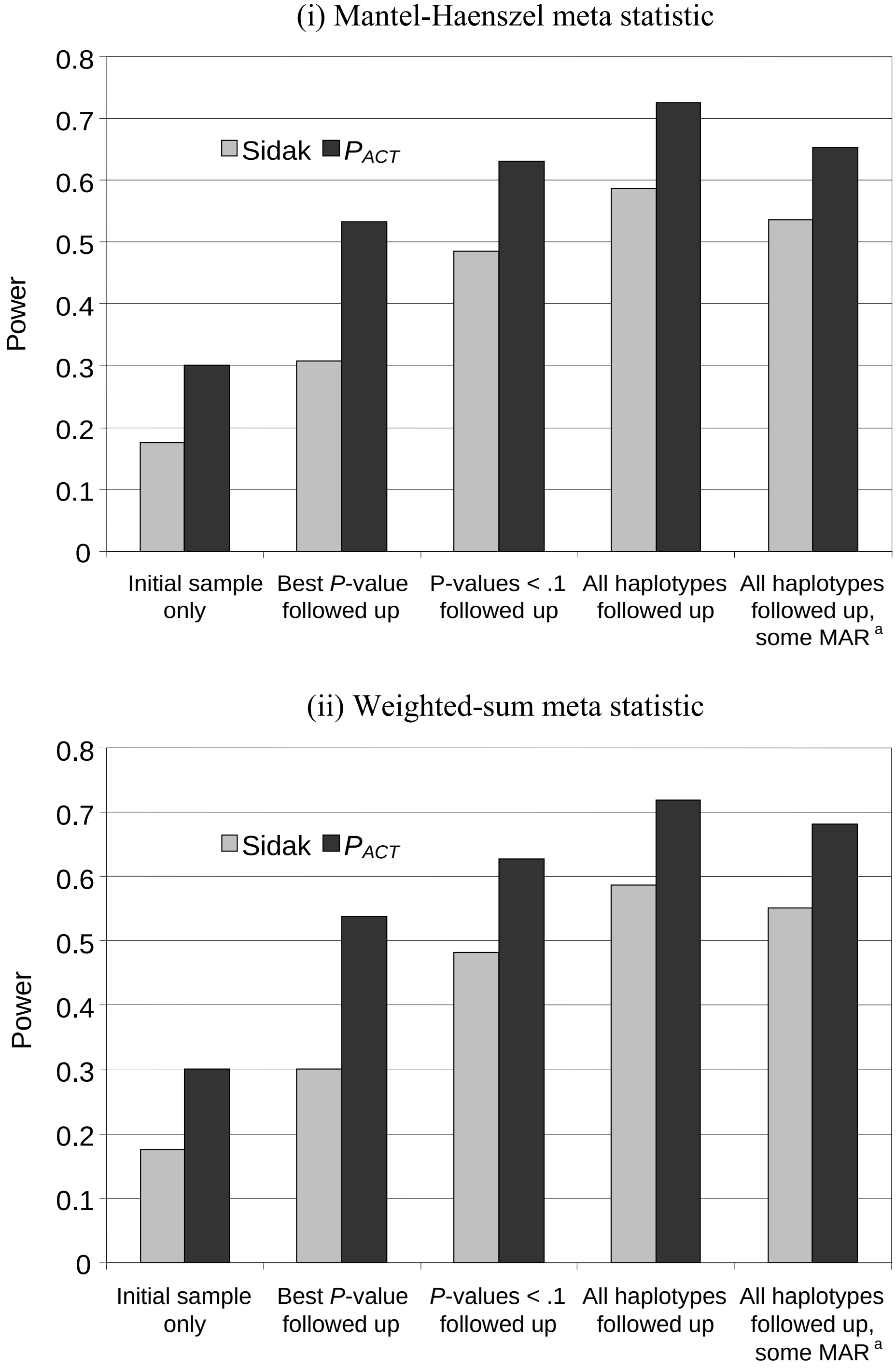


(ii) P-values < .1 in initial sample followed up, P_{ACT} computed as in equation (3.3).



Note – Meta P-values of 1 indicate that none of the 169 SNPs passed the cutoff in the initial sample, and hence a meta-analysis was not performed.

Figure 3.3: Power to detect a heterogeneous genetic effect in an initial sample and four follow-up samples.



(i) Mantel-Haenszel meta statistic

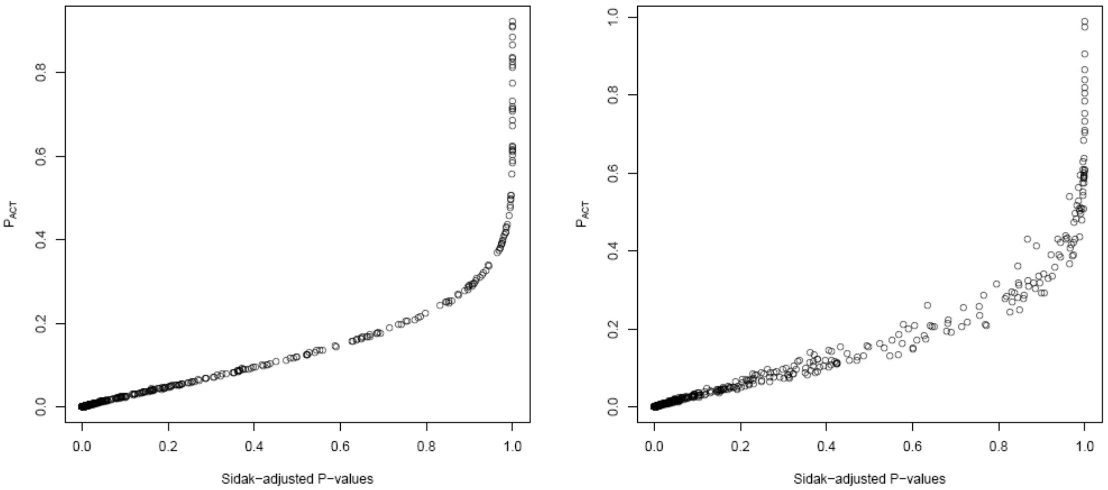
(ii) Weighted-sum meta statistic

Note – [a] MAR indicates that SNPs (and corresponding haplotypes) were missing-at-random in the follow-up samples, as described in section 3.2.5

Figure 3.4: Joint distribution of P_{ACT} and Šidák *P*-values from weighted-sum meta-analyses

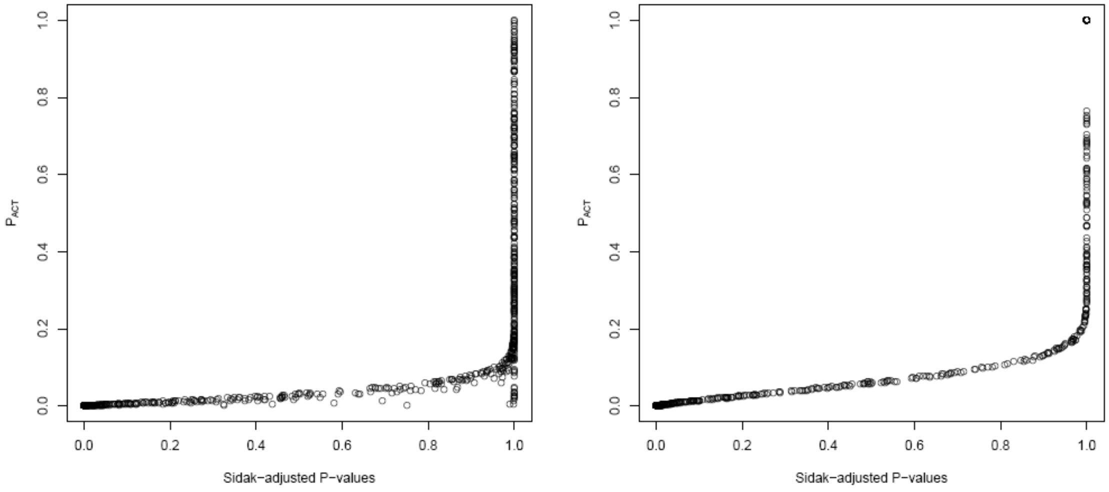
All 169 haplotypes tested in initial samples and four follow-up samples:

i) No missing data ii) Data missing-at-random



Only select results followed up in all samples:

iii) Single best result followed up iv) All *P*-values < .1 followed up



CHAPTER 4

EVIDENCE AND IMPACT OF DIFFERENTIAL RATES OF GENOTYPE ERROR AND MISSING GENOTYPE DATA IN SNP ASSOCIATION STUDIES

Many common methods of SNP genotyping are more prone to certain types of errors than others. For example, on many genotyping platforms heterozygous genotypes are at greater risk for being incorrectly assigned as another genotype or dropped due to ambiguity. Using replicate genotype data from a variety of genotyping platforms, we investigate the extent to which rates of genotype error and missing data differ depending on true genotypes. We find that rates do differ significantly across genotypes, and that depending on the genotyping platform, either heterozygotes or minor-allele homozygotes have the highest rates of genotype error and missingness. We use simulation to investigate the impact of the observed distributions of errors and missing data on three common tests of association. We find that genome-wide analyses based on allele frequency tests and transmission/disequilibrium tests may have inflated study-wide false positive rates, while trend-based tests of association are generally robust to the presence of differentially missing and erroneous genotypes. We conclude by reinforcing the recommendation of Sasieni (1997) that trend tests should be used as a more robust alternative to the allele-frequency test, and recommending stringent quality control or direct modeling of genotype error and missingness for transmission/disequilibrium tests.

4.1 Introduction

Recent advances in SNP genotyping have resulted in remarkable improvements in genotyping throughput, cost, completeness, and accuracy. Nonetheless, genotype error and missing data remain a concern even in high throughput labs, and especially in other labs.

Genotyping error in SNPs may occur due to human error, unanticipated reactions with the primers, weak PCR amplification, or poor clustering. The first three of these factors may also result in genotypes classified as missing due to lack of a discernible signal. Missing genotype data may also be created deliberately to ensure data quality, often through an error detection process that removes inconsistent genotypes or a no-call procedure that removes ambiguous genotypes. For data on related individuals, errors may be detected through checks for Mendelian inconsistency; however, Mendelian checks do not catch most errors in SNPs (Gordon et al. 1999) and are not possible in the context of unrelated cases and controls. For data on unrelated individuals, error detection at the individual level is generally not possible; instead, a no-call procedure is used to remove experimentally ambiguous genotypes, and the quality of the entire plate of genotypes is assessed through quality control (QC) filters such as checks for data completeness, Hardy-Weinberg equilibrium (HWE), and concordance between quality control duplicates.

Much of the ambiguity that leads to genotype error and missing genotypes is due to low signal resulting from weak amplification or a failed primer reaction. Many common SNP genotyping platforms -- including Sequenom, Affymetrix, Illumina, Perlegen, MIP (Molecular Inversion Probe), and pyrosequencing -- use a procedure that measures a separate signal for each allele and calls genotypes based on signal intensities. Homozygotes genotyped on these platforms generally have a strong signal for one allele and no signal for the other, while heterozygotes have weaker signals for both alleles. The weaker heterozygote signals are more likely to blend into the noise, especially if they are further weakened by inefficient amplification. Signal ambiguity can lead to an incorrect genotype call or a no-call, the latter resulting in missing data. If one allele signal is too weak to observe, the heterozygote will be mistyped as a homozygote; if both are too weak, it will be classified as missing. Homozygotes are at lower risk for this type of error and missingness because of their generally stronger signal intensity.

A common source of signal ambiguity is the failure to amplify one or both alleles due to an unrecognized variant in the primer region. Koboldt et al. (2006) examined the primer sequences used for SNPs genotyped in Phase I of the HapMap (International HapMap Consortium, 2005) for recently discovered variants, and found that the presence of SNPs in the primer region was predictive of increased genotype error (specifically heterozygotes scored as homozygotes) and no-call rates. The pervasiveness of the SNP-in-primer problem varied with genotyping platform due to differences in average primer lengths across platforms.

Ambiguity is also present in translation of the raw signal data into called genotypes. SNP genotypes are commonly called by algorithms which cluster individuals into three genotype categories based on signal intensity scores. Extremely poor clustering that affects an entire plate of genotypes is often detectable through HWE testing and other quality control measures. Misclassification of genotypes at the individual level is less detectable and is most likely to occur when the dispersion of intensity scores within a cluster is high. Similarly, genotypes belonging to highly variable clusters are more likely to appear ambiguous in terms of cluster membership and to be classified as no-calls. Since clusters with the fewest data points will tend to have the highest variance, minor allele homozygotes may be at the greatest risk of being classified as no-calls or misclassified into the incorrect cluster, especially for rare minor-allele SNPs.

It is well-known that heterozygotes are often at greater risk of being mistyped (Cutler et al. 2001, Mitchell et al. 2003) and missing (Hirschhorn and Daly 2005, Hao and Cawley 2007) for the reasons described above. However, in the literature on SNP genotyping error, studies allowing for differential rates of error have been the exception rather than the rule.

Studies of the effects of SNP genotyping error on case-control association tests have typically assumed error models that do not allow for differential rates of error by genotype. For example, Gordon and Ott (2001) have shown that for a random-allele model of genotyping error, misclassification of genotypes reduces power of the test for equal allele frequency between cases and controls but does not affect the rate of type I error. However, this random-allele model, which assumes that genotyping error occurs independently for each allele, is a unique case in that it guarantees that Hardy-Weinberg equilibrium is maintained, even for very high rates of error. Sasieni (1997) has shown that the allele frequency test is sensitive to HWE violations, so the conclusions of Gordon and Ott may not generalize to other models of genotype error.

Ahn et al. (2006) considered a model of differential rates of genotype error in an investigation of how different types of genotype error affect the power of the Cochran-Armitage test for trend (Cochran 1954; Armitage 1955). They found that the errors leading to the greatest loss of power are those involving misclassification of the more common homozygote. They addressed the power of the test rather than its validity because the trend test provides an unbiased test in the presence of genotype misclassification if 1) the null hypothesis of no genetic association is true and 2) the probabilities of misclassification are identical for cases and controls (Brenner 1992). The Cochran-Armitage test for trend is also robust to violations of HWE (Sasieni 1997). Hence, the influence of genotype error on trend tests is limited to a decrease in the power to detect a genetic effect (attenuation bias).

The transmission/disequilibrium test (TDT), however, can exhibit substantial bias in the presence of differential rates of genotyping error. Mitchell et al. (2003) have shown that several models of undetected genotyping error can lead to apparent over-transmission of the common allele, including both the random-allele model discussed above and a model where all error is due to miscalling heterozygotes as homozygotes. However, they noted that the latter model required higher error rates to produce the same magnitude of distortion as the random-allele model.

A related issue is the sensitivity of association tests to missing data and in particular to rates of missingness which vary by genotype. This topic has not been addressed until recently, probably because the simplest scenario -- random loss of genotypes -- is merely a problem of reduced power due to a diminished sample size. However, differential rates of missing genotypes can lead to problems much like those associated with differential rates of genotype error.

For the TDT, Hirschhorn and Daly (2005) found that transmission distortion similar to that observed by Mitchell et al. (2003) could occur even in the absence of genotyping error if genotypes were missing at different rates. In particular, the proportion of false positives vastly exceeded the desired type I error rate when heterozygous genotypes were missing at a greater rate than homozygous genotypes. They noted that this was also the case for other scenarios involving differential rates of missingness, such as when the rate of missingness was largest for minor-allele homozygotes. In a recent study addressing the impact of differential dropout rates between heterozygotes and homozygotes, Hao and Cawley (2007) observed a similar bias in the expected transmission ratio which increased with the ratio of rates of missingness between heterozygotes and homozygotes.

Hao and Cawley also investigated the impact of differential rates of missing genotypes on a trend test for association and found that the odds ratio remained unbiased even when the rate of missingness was ten times greater for heterozygotes than homozygotes. This is consistent with Sasieni's (1997) endorsement of the trend test as a robust test for association compared to the commonly used test for equal allele frequency, which is sensitive to departures from HWE.

Allele frequency tests and trend tests are asymptotically equivalent when HWE holds. However, trend tests are based on genotype frequencies while the allele frequency test is based on an allele frequency estimate which requires the assumption of HWE. This distinction is particularly relevant to the problem of differential rates of genotype error and missing genotypes, since these differential rates lead to distortion in genotype frequencies and hence in HWE. Nevertheless, previous studies on the effects of differential rates of missing genotypes and genotype error on the validity and power of association tests have emphasized tests for trend or the TDT, but not the allele frequency test. In section 4.2.1, we investigate the properties of the allele frequency test under a simple model where data are missing only for heterozygotes. We demonstrate theoretically that differential rates of missing genotypes can invalidate the allele frequency test, although noticeable bias generally requires substantial levels of error. We present similar results for a simple model of differential genotype error.

To assess the extent to which differential rates of genotype error and missingness are realistic in practice, in section 4.2.3 we measure and compare rates of genotype error and missing data among heterozygotes and both major and minor-allele homozygotes for a variety of genotyping platforms and allele frequencies. To do this, we take advantage of two existing datasets where replicate genotyping has been performed: a set of 1388 384-well plates of genotypes obtained using the Sequenom platform for the FUSION study (Valle et al. 1998), and a set of >200,000 plates of 30 trios each genotyped by three or more centers as part of the HapMap (International HapMap Consortium 2005, 2007). The FUSION data include the final resolved genotype calls, which can be treated as the consensus or true genotypes. For the HapMap data, we develop an EM algorithm which infers true genotypes from the replicate genotype data, accounting for the relatedness of individuals. In both FUSION and HapMap data, we find clear evidence of differential rates of genotype error and missingness; in particular, we find that rates of genotype error and missing genotypes are higher for heterozygotes and minor allele homozygotes than for major-allele homozygotes. In the HapMap data, we observe substantial variation in rates of genotype error and missingness between platforms.

We next explore the impact of data completeness on the validity and power of association studies, given the types of genotype error and differential rates of missing data that may occur on incomplete plates of genotypes. Many studies impose a set of plate-wide QC criteria that includes removing or re-genotyping plates with low genotype success rates. Since substantial resources may be spent re-genotyping incomplete plates, we investigate the efficacy of this endeavor in section 4.2.4. We use simulations to assess type I error rates and power in an association-testing framework for a range of genotype completeness, given the rates of error and missingness by genotype observed on initiall plates of FUSION genotypes. The impact of incompleteness on validity and power varies depending on the type of analysis desired. For allele frequency tests, even a small proportion of missing genotypes is sufficient to distort HWE and yield anti-conservative tests. However, trend tests attain the correct type I error rate even when up to 80% of genotype data are missing due to no-calls, although they do lose power in this situation due to the 80% reduction in sample size. For TDT analyses, however, even relatively low rates of missing genotypes (.05–.10) cause the rate of type I error to more than double in our simulations.

Finally, we investigate the distribution of genotype error and missingness rate across plates of genotypes, and find that 70% of the genotype errors in the Sequenom (HapMap) data are concentrated in <1% (~2%) of the plates. This is worrisome because in a large-scale association study, interest is typically focused on the SNP or SNPs showing the strongest evidence of association; hence, systematic but undetected error for a single SNP could impact the conclusions of an entire study. We perform simulations to investigate the impact of the observed distribution of genotype error and missingness on the study-wide false positive rate of a genome-wide.association study. We observe inflated rates of type I error when the allele frequency test is used to test for case-control association, particularly for platforms exhibiting greater rates of genotype error. The removal of plates failing HWE checks does not eliminate this bias; however, the Cochran-Armitage test for trend achieves the correct type I error rate for all platforms. For a TDT-based genome-wide association study, the observed distributions of error and missingness lead to study-wide false positive results in every simulation. However, the target type I error can be attained with the imposition of strict plate-wide QC measures. We conclude that the non-uniform distribution of genotype errors and missing data across both genotypes and plates has the potential to invalidate association studies, particularly the allele frequency test and the TDT, so additional precautions should be taken when using these methods.

4.2 Methods

4.2.1 Impact of differential rates of missing genotypes and genotype error on type I error rate on the test of equal allele frequencies

Assume that true genotypes AA, AB, and BB are in HWE with allele frequencies p and q and true genotype frequencies p^2, $2pq$, and q^2. In this case, allele frequency can be estimated from observed genotype frequencies:

$$\hat{p} = \hat{p}_{AA} + \frac{1}{2}\hat{p}_{AB}.$$

This estimate of p is unbiased under HWE but is generally biased when HWE is violated, since HWE is required for the allele counts to follow a binomial distribution (Sasieni 1997). If heterozygotes are missing at rate m and no other genotypes are missing or mistyped, it is easy to show that our estimate of p will be biased:

$$E\left(\hat{p}\right) = p\left(\frac{1-mq}{1-2mpq}\right).$$

Figure 4.1 i) shows this bias due to preferential missingness of heterozygotes (or of homozygotes, although this is atypical) for a range of allele frequencies. While small in absolute terms, this bias can be a large portion of the estimated p for low allele frequencies.

Define N as the number of individuals for whom genotypes were attempted and n as the number of genotypes observed (that is, N minus the number of missing heterozygotes). The usual variance estimate $\hat{p}(1-\hat{p})/2n$ is not appropriate if a portion of heterozygotes are missing, since the allele counts do not follow a binomial distribution in this case (Sasieni 1997). We use the delta method to approximate the true variance:

$$Var(\hat{p}) \approx \frac{pq\left(1-m+2mpq\right)}{2N(1-2mpq)^3}.$$

The expectation of the conventional estimate is approximately

$$E\left(\frac{\hat{p}(1-\hat{p})}{2n}\right) \approx \frac{pq\left(1-m+m^2pq\right)}{2N(1-2mpq)^3},$$

so the approximate ratio of true to estimated variance can be expressed as

$$k = \frac{1-m+2mpq}{1-m+m^2pq}.$$

Note that the ratio is independent of sample size, and that $k = 1$ when $m = 0$. Because the incorrect variance is used when calculating the *t*-statistic, the usual allele frequency test statistic $t \xrightarrow{d} \sqrt{k}Z \sim N(0,k)$, and the resulting test will be biased with expected type I error

$$P(\left|\sqrt{k}Z\right| > 1.96) = 2\Phi\left(-\frac{1.96}{\sqrt{k}}\right).$$

Figure 4.1 ii) shows the expected rates of type I error for the allele frequency test when there is a proportional loss of heterozygotes or homozygotes. The allele frequency test will be anti-conservative when heterozygotes are more likely to be missing than homozygotes, and overly conservative if homozygotes are more likely to be missing.

A similar derivation can be performed to allow for differential rates of genotype error. Figure 4.1 iii) shows the expected rates of type I error when heterozygotes are erroneously genotyped as homozygotes, and vice versa. Since the allele frequency test is dependent only on allele counts, the scenario where heterozygotes are misclassified as the major and minor allele homozygote in equal numbers leads to a valid test. However, this case seems implausible given the likely sources of the error, since it would imply that the two alleles each failed to amplify with equal probability. In general the allele frequency test will be biased when genotype error rate varies by true genotype. If heterozygotes have a larger (smaller) probability of error than homozygotes, the allele frequency test becomes anti-conservative (overly conservative), especially for SNPs with low MAF.

4.2.2 Samples

To estimate empirical rates of error and missingness by genotype, we take advantage of replicate genotype data from two distinct datasets. Our first sample contains replicate mass spectrometry genotyping data for 499 SNPs genotyped by the Sequenom Biomass system for the FUSION study (Valle et al. 1998) which were genotyped more than once and ultimately resolved with confidence. Our second sample consists of data from the January 2007 HapMap release, which included files containing HapMap genotype data in its raw form, prior to the application of QC filters

For both samples, we restricted our analysis to plates which were valid according to a set of lenient criteria. We did this to remove plates that would normally be discarded due to obvious plate-wide errors, since our main interest was in estimating the amount of individual genotype error and missingness on seemingly valid plates. For the analysis described in section 4.2.3, we considered plates valid if

1.) ≥ 80% of genotypes were called (non-missing)

2.) the genotypes on the plate passed HWE ($P > 1\times10^{-5}$)

3.) the plate included at least 2 heterozygotes and 2 major-allele homozygotes

4.) plate allele frequency differed from the consensus allele frequency by < .20

5.) there was not an obvious allele switch

We defined an allele switch as a plate with > 80% of homozygotes recorded as the opposite homozygote as the consensus. For Sequenom genotypes, we defined the consensus allele frequency for each SNP as the allele frequency computed based on the final genotypes. For the HapMap genotypes, we defined the consensus allele frequency as the average allele frequency across all QC positive plates for each SNP, and used a majority rule to identify plates where allele labels had been switched.

For the Sequenom data passing these criteria, we defined our sample to include the first valid genotyping attempt for each plate, as well as the final resolved genotypes. Our final sample included sequential data on 1388 384-well plates that were genotyped more than once and ultimately resolved with confidence, either by the Sequenom software or manually, for a total of 485,667 final genotypes.

For the unfiltered HapMap data passing the above criteria, our analysis sample consisted of 121,713 SNPs that had been genotyped on one or both of the CEU or YRI samples at three or more center/platform combinations. Including SNPs genotyped on both samples, our analysis sample included data on 105,686 SNPs genotyped in replicate for the 30 YRI trios, and 105,673 for the 30 CEU trios, for a total of 211,359 × 90 distinct genotypes attempted. With three or more centers submitting genotypes for each of these SNPs, data were available for a total of 664,369 × 90 submitted genotypes, Replicate genotyping of the same SNP by multiple centers occurred both inadvertently and as a part of the HapMap quality assessment exercises (International Hapmap Consortium 2005, Online Supplement). Tables 4.1 and 4.2 show the counts of plates submitted in replicate by center, platform, and the number of replicates for the CEU and YRI samples. Genotyping was performed by ten centers on eight different platforms, for a total of 14 different center/platform combinations for CEU samples and 15 center/platform combinations for YRI samples.

In section 4.2.4, we will relax the first of the above sample-definition criteria so we can assess the tolerance of association testing methods to lower-quality and less-complete genotypes. For the GWA simulation analysis in 4.2.5, we will apply all of the above criteria and will further restrict our sample of replicate HapMap genotypes to include only plates of genotypes which passed the HapMap QC filters (International

HapMap Consortium, 2005). Application of these additional criteria is equivalent to removing plates with

1.) ≥ 2 Mendelian inconsistencies

2.) ≥ 2 of 5 duplicate genotypes inconsistent with original genotype

3.) HWE P-value $\leq .001$

as well as plates that had been flagged as QC failures by the submitting centers.

4.2.3 Estimation of rates of genotype error and missingness

For each SNP, we labeled the major allele (defined as the allele observed most frequently over all valid plates for that SNP) as 'A' and the minor allele as 'B'. The conditional probabilities for the four possible genotype calls (AA, AB, BB, no call) given the three possible true genotypes can be expressed as P(observe genotype y | true genotype is x) = e_{xy} and written as a 3 x 4 matrix

$$E = \begin{bmatrix} e_{11} & e_{12} & e_{13} & e_{14} \\ e_{21} & e_{22} & e_{23} & e_{24} \\ e_{31} & e_{32} & e_{33} & e_{34} \end{bmatrix}$$

where 1, 2, 3, and 4 are shorthand for AA, AB, BB, and no call, and each row sums to 1.

Replicate Sequenom genotypes: Treating the final genotype as the true genotype, we computed e_{xy} for $x = 1, 2, 3$, $y = 1, 2, 3, 4$ as the proportion of all x genotypes which had been initially called as y, where initial call was defined as the genotype from the first valid plate. e_{14}, e_{24}, and e_{34} are the respective rates of missingness for AA, AB, and BB genotypes. We computed the genotype error rate among called genotypes as $(e_{12} + e_{13})/(e_{11} + e_{12} + e_{13})$.when the true genotype was AA, $(e_{21} + e_{23})/(e_{21} + e_{22} + e_{23})$ when the true genotype was AB, and $(e_{31} + e_{32})/(e_{31} + e_{32} + e_{33})$ when the true genotype was BB.

Unfiltered replicate HapMap genotypes: Since the HapMap replicate data did not include attempts to resolve discrepant genotypes, we used the known parent-child relationships and the repeated independent observations of genotypes for each SNP to compute the posterior probabilities that each underlying true genotype was AA, AB, or BB. To ensure that we considered only independent attempts at genotyping each SNP, we omitted any replicate plates having the same SNP/center/platform combination by randomly selecting a single plate for inclusion in our analysis.

We applied the Expectation Maximization algorithm to estimate a separate conditional error matrix E^j for each center/platform combination $j = 1,\ldots,J$. We initialized p, the vector of SNP allele frequencies, to the average allele frequency based on all called genotypes submitted on all included plates. We also initialized $E^j = E_{(0)}$ for all centers, where $E_{(0)}$ contains naïve moment-based estimates of the conditional probabilities based on two plates for each SNP.

Expectation step: Posterior probabilities for trios and individuals:
For each trio $i = 1,\ldots,30$ and each SNP $s = 1,\ldots,S$, we define the genotypes submitted by center j for the mother, father, and child as $G_{si}{}^{j} = \left\{ M_{si}{}^{j}, F_{si}{}^{j}, C_{si}{}^{j} \right\}$. For each SNP, we estimate the posterior probability of each possible value for the unobserved true genotypes $G_{si} = \left\{ M_{si}, F_{si}, C_{si} \right\}$ conditional on the observed genotypes as the product of the Mendelian probability of observing G_{si} and the conditional probability of all submitted genotypes conditional on G_{si}

$$
\begin{aligned}
P\left(G_{si} \mid G_{si}{}^{1},\ldots,G_{si}{}^{J_i}\right) &= \frac{P\left(G_{si}\right)\prod_{j} P\left(G_{si}{}^{j} \mid G_{si}\right)}{\sum_{g_{trio}}\left(P\left(g_{trio}\right)\prod_{j} P\left(G_{si}{}^{j} \mid g_{trio}\right)\right)} \\
&= \frac{P\left(M_{si},F_{si},C_{si};p_s\right)\prod_{j} E^{j}_{M_{si}M_{si}{}^{j}} E^{j}_{F_{si}F_{si}{}^{j}} E^{j}_{C_{si}C_{si}{}^{j}}}{\sum_{g_M,g_F,g_C}\left(P\left(g_M,g_F,g_C;p_s\right)\prod_{j} E^{j}_{g_M M_{si}{}^{j}} E^{j}_{g_F F_{si}{}^{j}} E^{j}_{g_C C_{si}{}^{j}}\right)}
\end{aligned}
$$

$P\left(M_{si}, F_{si}, C_{si}; p_s\right)$ is simply the Mendelian probability of genotypes M_{si}, F_{si}, C_{si} given allele frequency p_s; hence, any Mendelian-inconsistent true genotypes will automatically have a posterior probability of zero.

Posterior probabilities for individuals can be obtained by summing the trio posterior over the relevant trio members. For instance, the posterior probability that the mother's genotype is AA can be computed as

$$Post_{M,1} = P\left(M_{si} = 1 \mid G_{si}^{\,1}, \ldots, G_{si}^{\,J_i}\right) = \sum_{c,f} P\left(M_{si} = 1, F_{si} = f, C_{si} = c \mid G_{si}^{\,1}, \ldots, G_{si}^{\,J_i}\right)$$

Maximization step: conditional probabilities of erroneous genotypes and no-calls

For each center, we compute updated estimates of each element of E^j as

$$e_{xy}^{j} = \frac{\sum_s \sum_i \left(Post_{M_{si},x} \cdot I_{M_{si}^{\,j}=y} + Post_{F_{si},x} \cdot I_{F_{si}^{\,j}=y} + Post_{C_{si},x} \cdot I_{C_{si}^{\,j}=y}\right)}{\sum_s \sum_i \left(Post_{M_{si},x} + Post_{F_{si},x} + Post_{C_{si},x}\right)}$$

where $I_{M_{si}^{\,j}=y}$ is an indicator that the reported genotype $M_{si}^{\,j} = y$. The numerator is the predicted number of genotypes submitted as *y* by center *j* which are truly *x*. This expected count is computed by taking a weighted count of genotypes reported as *y* over all SNPs and individuals, where the weights are the posterior probability that the true genotype is *x*. Similarly, the denominator is the expected count of all genotypes submitted by center *j* which are truly *x*, which we compute by summing over all SNPs and individuals the posterior probability that the true genotype is *x*.

Our criterion for convergence of the EM algorithm was a difference of $< 10^{-10}$ in the L2-norm of the 3 × 4 × *J* array $[E^1, \ldots, E^J]$ over two successive iterations. We used this algorithm to estimate E^j for the 15 center/platform combinations described above based on all submitted genotypes in the combined CEU and YRI samples. We allowed SNPs that were genotyped in both the CEU and YRI samples to have different allele frequencies (and hence, different Mendelian probabilities) in each sample. To estimate E^j separately for five allele frequency categories (.01–.05, .05–.10, .10–.20, .20–.35, and

.35–.50), we performed the analysis using only SNPs with consensus allele frequencies in the appropriate category. We computed rates of genotype error for called genotypes based on the estimate of E^j as described for the replicate Sequenom data.

The posterior probabilities for individual genotypes can also be used to establish a set of consensus genotypes for the HapMap samples. All genotypes had a posterior probability > .5 for one of the three possible genotype, and 99.95% of genotypes had a posterior probability > .9999, so consensus genotypes could be determined with little ambiguity.

4.2.4 Estimation of type I error rate and power by plate completeness

Because substantial resources may be spent re-genotyping incompletely genotyped plates, we were interested in what kinds of bias and power loss could be expected if data from incomplete plates were used in analyses, given that incomplete plates are likely affected by differential rates of genotype errors and missingness, and what kinds of gains were associated with obtaining complete data.

To simulate realistic incomplete data, we grouped the replicate Sequenom data by completeness of initial plate to form eight sets of plates with missing genotype rates of < 2%, 2–5%, 5–10%, 10–20%, 20–30%, 30–40%, 40–50%, 50–60%, and 60–80%. Since the focus of this analysis was incomplete plates, we excluded plates which were initially complete and we included plates which had been excluded from the analysis in section 4.2.3 due to having no-call rates >20%. We performed 1,000,000 simulations for each of the eight categories. In each simulation, we randomly drew a plate from the appropriate set of plates. We used the final genotype frequencies for each plate to draw complete-data genotypes for a simulated sample of 500 cases and 500 controls. We then used the distribution of initial versus final genotype calls observed on each plate to form conditional probabilities of being observed as AA, AB, BB, or no-call given an individual's complete-data genotype. We used these conditional probabilities to draw initial genotypes for each of the cases and controls, conditional on their assigned true genotypes.

For each simulated sample of 500 cases and 500 controls, we performed allele frequency tests and trend tests using both the complete data and the initial, incomplete data. For each of the eight levels of plate completeness, we computed type I error rate as the proportion of 1,000,000 simulations with *P*-values < .05.

To estimate power, we re-performed the above analysis, but assigned complete-data genotypes for cases with probabilities P(AA) - ε, P(AB), and P(BB) + ε, where ε was chosen such that association tests had power ~ .8 when performed on complete data. In this case, we estimated power as the proportion of 1,000,000 simulations with *P*-values < .05.

To assess the impact of realistic incomplete data on the TDT, we also created 1000 simulations of 1000 trios, and performed the TDT on the subset of trios which were informative. In each simulation, we randomly drew a plate from the appropriate set of Sequenom plates, and simulated parental genotypes for the 1000 trios based on final genotype frequencies on the plate. Child genotypes were simulated directly from true parent genotypes assuming Mendelian inheritance. We then simulated observed genotypes for all individuals based the conditional probabilities of genotype error and missingness inferred from the distribution of initial versus final genotype calls on the Sequenom plate. We performed the TDT on all informative trios based on the complete data. We performed it again using the incomplete data, after first removing all Mendelian-inconsistent trios and incompletely genotyped trios. Type I error rate was estimated for each level of incompleteness for both initial and final genotypes as the proportion of simulations with *P*-values < .05.

4.2.5 Estimation of type I error rate for genome-wide association testing

To assess the impact of the observed patterns of genotype error and missingness on genome-wide association (GWA) studies, we simulated genotype data for ~300,000 SNPs. To allow the rates of error and missingness to vary realistically across SNPs, we randomly drew a plate-wide estimate of the error probability matrix E for each simulated SNP. The plate-wide estimates of E were based on observed rates of error and missingness from plates of HapMap SNPs passing QC. We used the plate-wide probabilities of error and missingness to randomly induce missing genotypes and errors into our simulated genotypes, and then tested all SNPs for association.

Construction of plate-wide estimates of E: We computed separate estimates of E for 642,879 plates of attempted genotypes in the HapMap CEU and YRI data which had consensus MAF > .01 and passed both our QC measures (see section 4.2.2) and the set of QC measures defined by the HapMap. Each plate contains 90 genotypes from either the CEU or YRI sample for a particular SNP ($s = 1, \ldots, S$) and center ($j = 1, \ldots, 15$). A simple estimate of the error probability matrix for each plate,

$$E^{s,j} = \begin{bmatrix} e_{11}{}^{s,j} & e_{12}{}^{s,j} & e_{13}{}^{s,j} & e_{14}{}^{s,j} \\ e_{21}{}^{s,j} & e_{22}{}^{s,j} & e_{23}{}^{s,j} & e_{24}{}^{s,j} \\ e_{31}{}^{s,j} & e_{32}{}^{s,j} & e_{33}{}^{s,j} & e_{34}{}^{s,j} \end{bmatrix}$$

can be constructed by comparing observed genotypes to the consensus genotypes inferred from the posterior probabilities computed by our EM algorithm. For example, we can compute the conditional probabilities $e_{31}{}^{s,j}$, $e_{32}{}^{s,j}$, $e_{33}{}^{s,j}$, and $e_{34}{}^{s,j}$ as the proportion of BB consensus genotypes called as AA, AB, BB, or missing. A problem with this simple estimate is that with 90 individuals, SNPs with MAF < .2 will generally have fewer than 10 BB genotypes, so often only very crude estimates of $e_{31}{}^{s,j}$, $e_{32}{}^{s,j}$, $e_{33}{}^{s,j}$, and $e_{34}{}^{s,j}$ will result.

To smooth these crude estimates, we fit a logit model to the data that allows the rate of each type of error to vary across plates as a log-linear function of true MAF (p_s) and includes random effects for each plate and for each individual. We fit a separate model for each true genotype. For example, for the $n_{BB}{}^{s}$ individuals with consensus genotype BB for SNP/plate s, the number of missing genotypes has a $\mathrm{Bin}\left(n_{BB}{}^{s}, e_{34}{}^{s,j}\right)$ distribution, where $e_{34}{}^{s,j}$ is the rate of missingness for BB genotypes on the plate for SNP s submitted by center j. For all combinations of SNPs $s = 1, \ldots, S$ and individuals $i = 1, \ldots, 90$ where the consensus genotype is BB, we fit the model

$$e_{34}{}^{s,i} = \frac{\exp\left(\mu + \beta p_s + \varepsilon_s + \delta_i\right)}{1 + \exp\left(\mu + \beta p_s + \varepsilon_s + \delta_i\right)}$$

where $\varepsilon_s \sim N\left(0, \sigma_\varepsilon{}^2\right)$ and $\delta_i \sim N\left(0, \sigma_\delta{}^2\right)$ are plate-specific and individual specific random effects. The ^{j} superscript is omitted in the model above because we fit the model separately for each center. For each plate, we then obtain $e_{34}{}^{s,j}$ as the average of the predicted rates $e_{34}{}^{s,i}$ over the $n_{BB}{}^{s}$ individuals. Although MAF is included in the model as a log-linear covariate (p_s), we also fit the model separately for four categories of consensus MAF (.01–.10, .10–.20, .20–.35, and .35–.50) to allow all parameters to vary flexibly with allele frequency.

We fit similar models to estimate plate-specific rates corresponding to each of the conditional probabilities in E^j. We fit a separate set of such models to estimate E^j for $j = 1, \ldots, 15$, which allows us to estimate and store a separate set of plate-wide error probabilities for each of the 15 center/platform combinations.

Simulation of GWA samples: We computed genotype frequencies for 310,151 autosomal SNPs genotyped on >2000 cases and controls from the FUSION study (Scott et al. 2007) on the Illumina HumanHap300 BeadChip. Based on these genotype frequencies, we simulated 310,151 independent genotypes (“true genotypes”) under the null hypothesis of no association, either for 1) 10,000 replicate samples of 1000 cases and 1000 controls, or 2) 1000 replicate samples of 1000 trios.

To draw a stratified sample of 310,151 of the 642,879 HapMap-based plate-wide estimates of the conditional error matrix *E*, we first sorted the plate-wide estimates into bins according to the consensus HapMap MAF (89 categories representing minor allele proportions ranging from 2/180 to 90/180) and the center/platform which provided the raw genotypes. For 15 center-platform combinations × 310,151 SNPs, we randomly drew a plate-wide probability matrix $E^{s,j}$ from the bin corresponding to center *j* and to the allele frequency of the simulated GWA SNP. For HapMap SNPs with no BB consensus genotypes, the plate-wide estimates were missing estimates of $e_{31}{}^{s,j}$, $e_{32}{}^{s,j}$, $e_{33}{}^{s,j}$, and $e_{34}{}^{s,j}$. When such a plate-wide estimate was drawn, we re-drew estimates of e_{31}, e_{32}, e_{33}, and e_{34} from another plate, but kept the original estimates of the other conditional probabilities in $E^{s,j}$. For MAFs < .045, BB consensus genotypes were rare, so we combined the eight bins corresponding to MAF < .045 when re-drawing estimates of e_{31}, e_{32}, e_{33}, and e_{34}.

We next used a multinomial distribution based on the assigned plate-wide probability matrix $E^{s,j}$ to randomly induce genotype errors and missing genotypes into each simulated GWA SNP. Through this process, we obtained 16 sets of GWA data to analyze in each replicate: one set of simulated "true" genotypes and 15 sets of "observed genotypes", each corresponding one of the 15 center/platform combinations.

Estimation of GWA study-wide type I error rates based on case-control association: We tested each SNP in each of the 16 samples for association with case-control status using either the allele frequency test or the Cochran-Armitage test for trend (Cochran 1954; Armitage 1955), and used the Šidák method (Šidák 1967) to adjust the *P*-value of the most significant SNP genome-wide for the number of SNPs tested. We repeated this analysis on 10,000 replicate GWA samples, each time recording the adjusted minimum *P*-value for each of the 16 analysis samples. To estimate the study-wide type I error rate, we computed the proportion of replicates where the Šidák-adjusted *P*-value < .05. We performed this analysis both with and without a HWE check which restricted the set of analysis SNPs to include only SNPs passing HWE ($P_{HWE} > .001$) on the control sample.

Estimation of GWA study-wide type I error rates based on the TDT: In each analysis sample, we computed the TDT statistic for every SNP with at least 100 informative trios. The number of informative trios available for each TDT varied with the MAF of the SNP being tested, as well as the loss of trios due to Mendelian inconsistencies or missing data. For each of the 16 analysis samples, we computed the Šidák-adjusted *P*-value for the most significant TDT statistic genome-wide. To assess the impact of the distribution of genotype error and missing data on smaller-scale analyses, we also computed Šidák-adjusted *P*-values for the largest TDT statistic out of 1000, 100, or 10 randomly selected SNPs. For comparison purposes, we also recorded the *P*-value of a single SNP, selected at random. We repeated this analysis on 1000 replicates of each of the 16 GWA samples. We performed this analysis both with and without the application of QC filters. When QC filters were applied, we restricted the set of analysis SNPs in each sample to those with ≤ 2 Mendelian-inconsistent trios and at least 95% completeness. For comparison purposes, we also tried further restricting the set of SNPs to those with 99% completeness and no Mendelian inconsistencies.

4.3 Results

4.3.1 Replicate Sequenom data

In a sample of 485,667 genotypes on 499 SNPs from the Sequenom Biomass genotyping system in the FUSION study that were ultimately resolved with confidence, models of identical genotype error rates and identical missingness rates across genotypes were strongly rejected (p<<.0001). The rate of missing genotypes was highest for heterozygous genotypes: 3.4% of heterozygotes were initially no-calls, compared to 2.1%

of AA or major allele homozygotes and 2.3% of BB or minor allele homozygotes. Figure 4.2 shows the initial rates of missingness by final genotype both overall and broken down by allele frequency categories. In all categories, heterozygous genotypes are the most likely to initially be missing. BB homozygotes are generally slightly more likely than AA homozygotes to be missing, especially for lower allele frequencies.

In contrast, the rate of genotype error among called genotypes was the highest among minor allele homozygotes. 0.52% of BB homozygotes were initially misclassified as an incorrect genotype, which was equally likely to be AA or AB. 0.37% of heterozygotes were initially called as homozygotes, while only 0.20% of AA homozygotes were initially misclassified as another genotype. Figure 4.3 shows rates of genotype error by final genotype overall and broken down by allele frequency. Genotypes homozygous for the minor allele have the highest rates of genotype error at all allele frequencies, followed by heterozygous genotypes.

Genotype errors and missing genotypes were not evenly distributed across plates. 62% of the 1388 plates passing our criteria for inclusion were initially complete, and 86% of no-call genotypes were concentrated in just 20% of the 1388 plates. Similarly, 96% of plates had no initial genotype error, and 70% of genotype errors occurred on the 8 most error-prone plates.

Because scarce resources are often spent resolving bad or incomplete plates, we investigated the potential gains in power and validity of association tests associated with these efforts. Figure 4.4 i) compares expected type I error rates for case-control association tests based on simulated samples with incomplete versus complete genotype data. The initial incomplete data reflect the differential rates of genotype error and missingness observed in our Sequenom data, as described in section 4.2.4. Type I error rate for the allele frequency test is represented by the solid bars, and it is clear that the allele frequency test is biased in favor of rejection when the data are not fully resolved.

However, trend tests, represented by shaded bars, are robust to the variations in data quality and provide unbiased tests for all levels of genotype completeness.

Since trend tests can generally be used in place of the allele frequency test, valid case-control SNP association tests may be achieved even with low-quality initial data, although this does not generalize to more complicated association tests, such as those involving haplotypes. In the context of case-control association tests, however, power loss may be the main reason to worry about the quality and completeness of genotypes. Figure 4.4 ii) compares expected power for the simulated initial and resolved genotype data. The power of both allele frequency and trend tests clearly decreases as the no-call rate increases. However, the power losses are mild for plates with >90% complete data.

The TDT is much more sensitive to low data quality than case-control association tests. Figure 4.5 shows the expected type I error rates when the TDT is performed on initial versus complete genotype data. Even when the data are 90 – 95% complete, the observed type I error rate can be more than double the nominal type I error rate. It appears that requirements for genotyping completeness should be much more stringent in the case of the TDT than for case-control association tests unless the error can be appropriately modeled in the analysis.

4.3.2 HapMap data

For the 664,369 × 90 replicate genotypes submitted by the HapMap centers, respective rates of missingness for inferred true AA, AB, and BB genotypes were 0.7%, 1.1%, and 1.2%. Figure 4.6 i) shows the rates of missingness for all three genotypes overall and broken down by allele frequency. In contrast to the Sequenom data, the BB homozygotes have the highest rates of missing genotypes for all allele frequencies.

Figure 4.6 ii) presents rates of genotype error estimated using the HapMap data. As with the Sequenom data, BB homozygotes are the most likely to be misclassified overall (0.36%) and for all allele frequency categories, followed closely by heterozygotes, which are erroneously genotyped 0.33% of the time. Major allele (AA) homozygotes are misclassified as another genotype only 0.10% of the time.

Figure 4.7 presents rates of genotype error and missing data by inferred true genotype, broken down by platform and by the center submitting the genotype. Rates of genotype error vary significantly with genotype for every platform and center (*P* << .0001), and missing data rates vary significantly with genotype for every center (*P* << .0001) except for the Sanger Center (*P* = .045). However, the patterns of differential genotype error and missingness differ across platforms. For plates assayed using the Perlegen and Sequenom platforms, heterozygous genotypes are the most likely to be misclassified or missing. For Perlegen genotypes, this pattern is consistent with the known sensitivity of the amplicon long-range PCR structure of the Perlegen design to undetected variants in the primer region (International Hapmap Consortium 2007, Online Supplement). For other platforms, BB homozygotes are more likely than other genotypes to be missing (Affymetrix and MIP) or misclassified as an incorrect genotype (GoldenGate, Affymetrix, and FP-TDI). These patterns are similar when broken down by allele frequency (data not shown), although for low MAFs, BB genotypes are the most likely to be lost to genotype error or missing data on all platforms.

The above analysis is based on 664,369 plates of genotypes which passed the set of criteria described in section 4.2.2. We chose somewhat lenient criteria in order to assess the types of errors that may occur in raw genotype data. However, the more stringent QC measures defined for the HapMap data (International HapMap Consortium 2005) are probably closer to the types of filters applied to genotype data in large-scale association studies. Strengthening our criteria to include the HapMap QC filters reduces our analysis set to 642,879 plates of genotypes. We applied our EM algorithm to estimate rates of error and missingness for this reduced set of plates. Compared to the above analysis, we observed 1) very similar rates of missingness, and 2) reduced rates of genotype error rates for all platforms and centers, especially the Sanger Center. Although error rates were lower across the board, the overall pattern of differential error rates by genotype and platform was similar to the pattern in Figure 4.7 (data not shown).

As with the Sequenom data, no-call genotypes and genotype errors were not distributed evenly across plates. For the 642,879 plates passing the HapMap QC filters, 87% of no-call genotypes were concentrated on just 20% of plates, and 70% of genotype errors occurred on just 2.1% of plates. To assess the potential impact of this uneven distribution of error and missingness on large-scale association studies, we induced errors and missing data in simulated genome-wide SNP data based on probabilities sampled from the 642,879 plates of genotypes passing the HapMap QC filters.

Figure 4.8 shows the expected study-wide type I error rate for a genome-wide case-control association study given the distribution of genotype errors and missing data observed for each center and platform. For studies employing allele frequency tests (solid-colored bars), the study-wide type I error rate is somewhat inflated and can be as high as .08 depending on the genotyping center. The bias is most pronounced for the genotyping centers and platforms which show the highest incidence of genotype error in Figure 4.7 ii). Restricting the analysis to SNPs which pass HWE (P_{HWE} >.001) in the control sample only slightly mitigates this bias. Even in the absence of errors or missing data (leftmost bars), the allele frequency test has a type I error rate of .059, with a 95% confidence interval (.054, .063) that does not include .05. As discussed above, the allele frequency test is not robust to departures from HWE. It is likely that even with our reasonably stringent HWE check, one or more of the ~300,000 SNPs will have sufficient HWE departure to inflate the type I error rate of the allele frequency test. This appears to occur even in the absence of genotype error and missing data, and to be exacerbated by differential rates of error and missingness.

The estimated type I error rate of .0480 for the trend test in the absence of error and missing data (leftmost shaded bars) does include .05 in its 95% confidence interval (.0459, .0501). When error and missing data are induced in the data, the trend test achieves type I error rates within or slightly below this confidence interval, falling below the target rate of .05 for all centers and platforms. Since the trend test attains the appropriate rate of type I error, we also investigated the impact of the observed distribution of genotype error and missing data on the power of the trend test in GWA (data not shown). Power for all centers and platforms was slightly diminished compared to power in the absence of error and missing data. Depending on the genotyping center, power ranged from 87% to 100% of power in the absence of error and missing data.

We next investigated the impact of the observed distributions of error and missing data on large-scale TDT studies. Figure 4.9 shows the study-wide type I error rate for TDT-based association studies involving 1, 10, 100, 1000, or ~300,000 independent SNPs, either i) with no QC filters applied, or ii) with a reasonable QC filter (95% complete data, ≤2 Mendelian inconsistencies) which on average led to the removal of 12% of SNPs. The pre-QC results in Figure 4.9 i) are striking. A false positive result is observed for the GWA for every center, in every simulation. For comparison purposes, the leftmost bars show that the appropriate study-wide rate of type I error is attained when the "true" genotypes are tested, so the 100% false positive rate was induced by our addition of genotype errors and missing data. While it is possible to attain reasonable rates of type I error for studies of 1, 10, or 100 SNPs, depending on the genotyping platform, even studies of 1000 or more SNPs are wildly anti-conservative in this scenario.

In contrast, Figure 4.9 ii) shows the study-wide rates of type I error attained once a QC filter has been applied. After ~12% of SNPs with < 95% completeness and/or >2 Mendelian inconsistencies were removed from consideration, studies of 1 – 1000 SNPs are now unbiased. The GWA studies are still somewhat anti-conservative, depending on the center and platform, but they are much less biased than the pre-QC studies. By applying an even stricter QC filter requiring >99% completeness and 0 Mendelian inconsistencies, we were able to achieve the target rate of type I error for the GWA studies as well (data not shown), although application of this filter led to the unacceptable removal of an average of 37% of SNPs.

4.4 Discussion

Our results demonstrate that the rates of genotype error and missing genotype data vary considerably across genotypes, SNPs, genotyping platforms, and genotyping centers. Differential rates of error and missingness across genotypes can lead to bias in the allele frequency test due to distortion of HWE, or bias in the TDT due to distortion in the transmission ratio. Given the wide variation in levels of error and missingness across SNPs, the tendency towards distortion can become amplified in large-scale studies, where even a small number of undetected bad SNPs have the potential to severely impact the outcome of the study.

However, although our results show that this scenario is possible, they also show that it is not inevitable. In the case of the TDT, stringent QC filters provided one possible solution. Excluding SNPs with even a single Mendelian inconsistency is conservative, but may be warranted given the sensitivity of the TDT to genotype error and the likelihood that Mendelian-detectable errors are accompanied by non-detectable errors. 99% complete genotype data is also a strict requirement, but as Figures 4.5 and 4.9 ii) demonstrate, the TDT is quite sensitive to data completeness, especially given non-differential rates of missing data. Another possible solution with the TDT might be to model the differential rates of error and missingness directly, perhaps with an algorithm which incorporates the information from the “non-informative” trios, including trios discarded due to missing genotypes or Mendelian inconsistencies.

. In the case of the allele frequency test, this kind of solution is not necessary since the trend test is already available as an improved version of the test. When testing for case-control association with trend tests, the main consequence of incomplete or erroneous genotype data will be loss of power due to reduced sample size and attenuation bias.

It is clear from our analysis and from previous work that the impact of genotype error and missing data on association tests varies substantially depending on the type of association test. This has important implications for study design, since a QC filter that is appropriate for one type of association test may be too weak or too stringent for another. A good example of this is the analysis on the impact of plate completeness presented in sections 4.2.4 and 4.3.1. For the TDT, genotyping completeness > 98% was necessary to achieve the correct type I error rate. Given the preferential missingness of heterozygotes in the underlying genotype data, this makes sense, since removal of even a few heterozygous genotypes can bias transmission in favor of the major allele. In contrast, the Cochran-Armitage test for trend achieved the correct type I error rate for all levels of completeness, and only suffered minor losses of power due to reduced sample size. In cases like this, it almost certainly makes sense to allocate genotyping resources differently depending on the goal of the study. For a TDT-based analysis, it would probably be wasteful not to re-genotype incomplete plates, since a meaningful analysis would not be possible without the complete data. For case-control association, some leniency in terms of required completeness might be warranted, depending on the trade-offs in cost between resolving incomplete plates and genotyping additional variants. If the only goal of the analysis is to perform case-control association testing on SNPs with a trend test, efficient use of resources might involve accepting less-than-complete plates of genotypes, and instead diverting the resources towards the genotyping of additional variants or samples.

Table 4.1: Counts of plates submitted in replicate for CEU sample

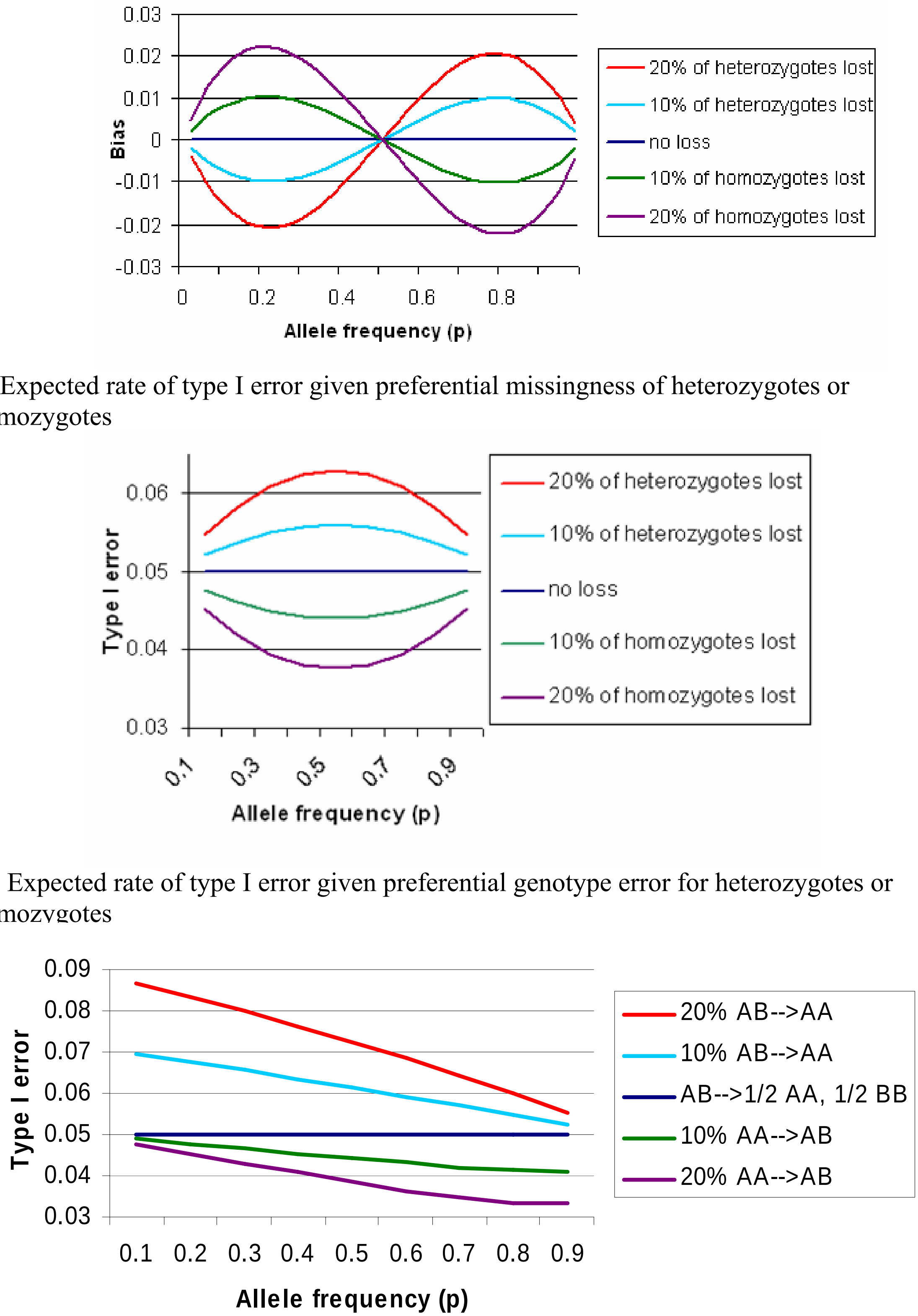
			Number of times plate is replicated by any platform and center								
Platform	Center	# plates	3	4	5	6	7	8	9	10	11
Perlegen	Perlegen	48,815	41,907	6450	339	46	16	25	27	4	1
BeadChip	Illumina	92,858	82,649	9130	789	87	45	100	45	12	1
BeadArray	Illumina	25,192	20,309	3322	504	243	464	256	79	14	1
	Sanger	26,096	21,286	3900	700	92	27	49	33	9	0
	McGill	15,449	12,230	1925	200	258	481	261	79	14	1
	Broad	5276	4692	499	40	14	4	12	12	2	1
	CHMC	5532	4872	418	47	45	46	68	30	6	0
Sequenom	Broad	4178	2246	647	325	219	427	231	69	13	1
	CHMC	3078	1879	152	100	192	436	236	69	13	1
Affymetrix	Affymetrix	71,322	62,892	7528	615	84	63	92	36	11	1
MIP	Affymetrix	2218	1775	377	61	3	1	1	0	0	0
	BCM	9316	5869	1888	471	263	470	261	79	14	1
Invader	RIKEN	21,877	17,821	2441	516	281	470	255	78	14	1
FP-TDI	UCSF-WU	2257	1084	151	73	165	445	249	75	14	1
Total number of plates:		333,464	93,837×3	9707×4	956×5	332×6	485×7	262×8	79×9	14×10	1×11

Table 4.2: Counts of plates submitted in replicate for YRI sample

			# of times plate is replicated by any platform/center				
Platform	Center	# plates	3	4	5	6	7
Perlegen	Perlegen	44,993	38,440	5940	555	52	6
BeadChip	Illumina	89,296	77,807	10,330	1047	104	8
BeadArray	Illumina	23,217	19,019	3618	528	49	3
	Sanger	25,418	20,497	4060	739	114	8
	McGill	13,439	11,316	1907	186	27	3
	Broad	4,620	4157	430	25	8	0
	CHMC	5,901	5233	604	49	13	2
Sequenom	Broad	2,797	1934	560	234	63	6
	CHMC	1,661	1385	228	43	5	0
Affymetrix	Broad	22,103	17,602	3935	506	57	3
	Affymetrix	67,466	58,315	8235	812	97	7
MIP	Affymetrix	2,108	1705	333	61	8	1
	BCM	7,989	5681	1858	399	48	3
Invader	RIKEN	18,950	15,819	2585	467	73	6
FP-TDI	UCSF-WU	947	801	125	19	2	0
Total number of plates:		330,905	93,237×3	11,187×4	1134×5	120×6	8×7

Figure 4.1: Expected biases in the allele frequency test due to differential rates of missingness and genotype error

i) Bias in allele frequency estimate given preferential missingness of heterozygotes or homozygotes



ii) Expected rate of type I error given preferential missingness of heterozygotes or homozygotes

iii) Expected rate of type I error given preferential genotype error for heterozygotes or homozygotes

Figure 4.2: Rates of missingness in replicate Sequenom data by allele frequency and true genotype

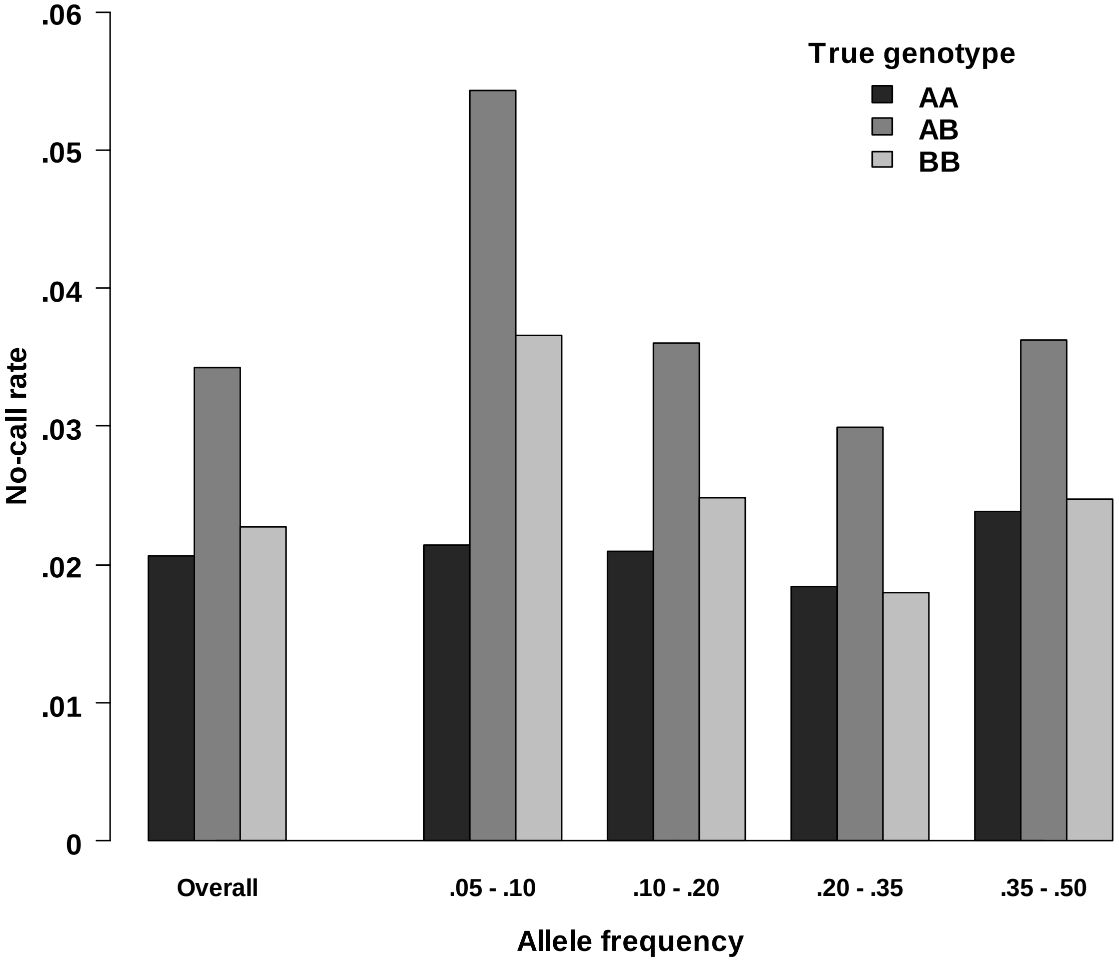


Figure 4.3: Genotype error rates in replicate Sequenom data by allele frequency and true genotype

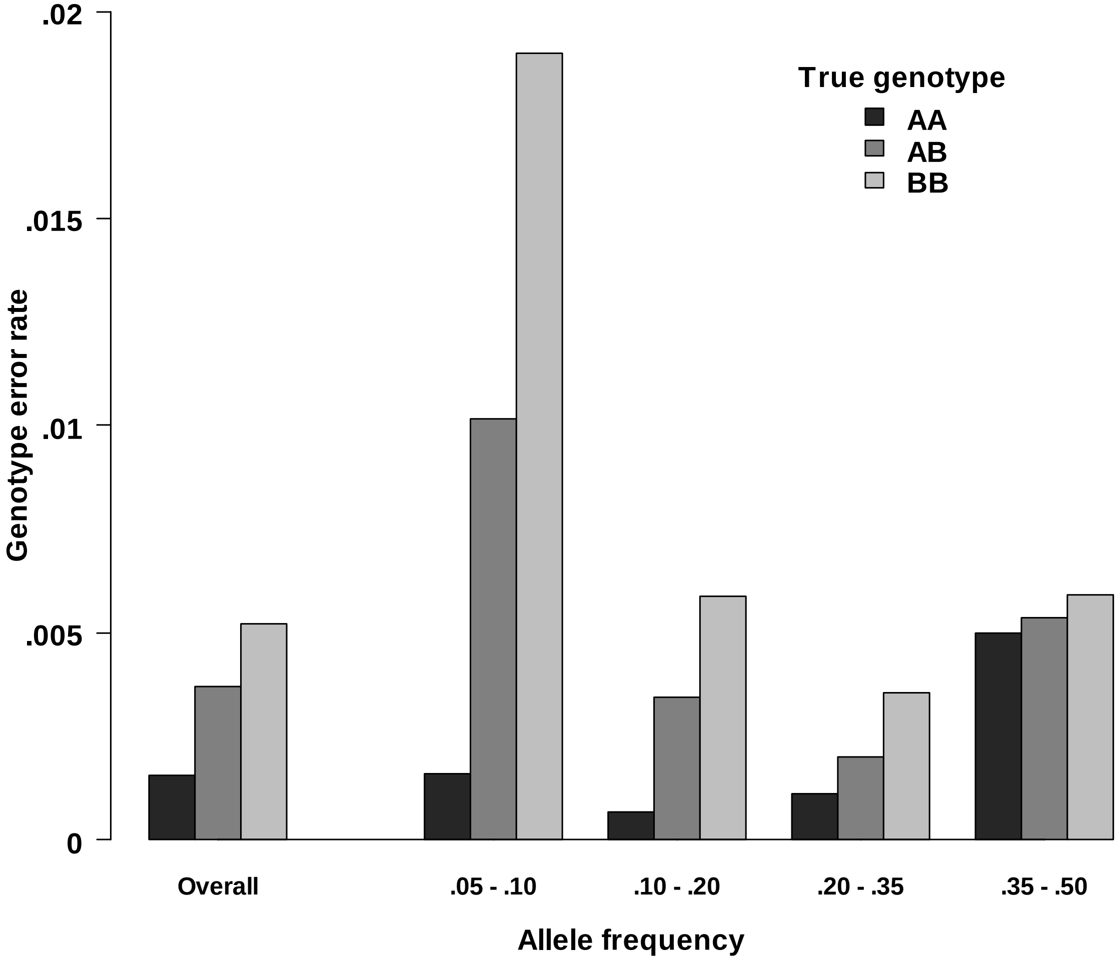
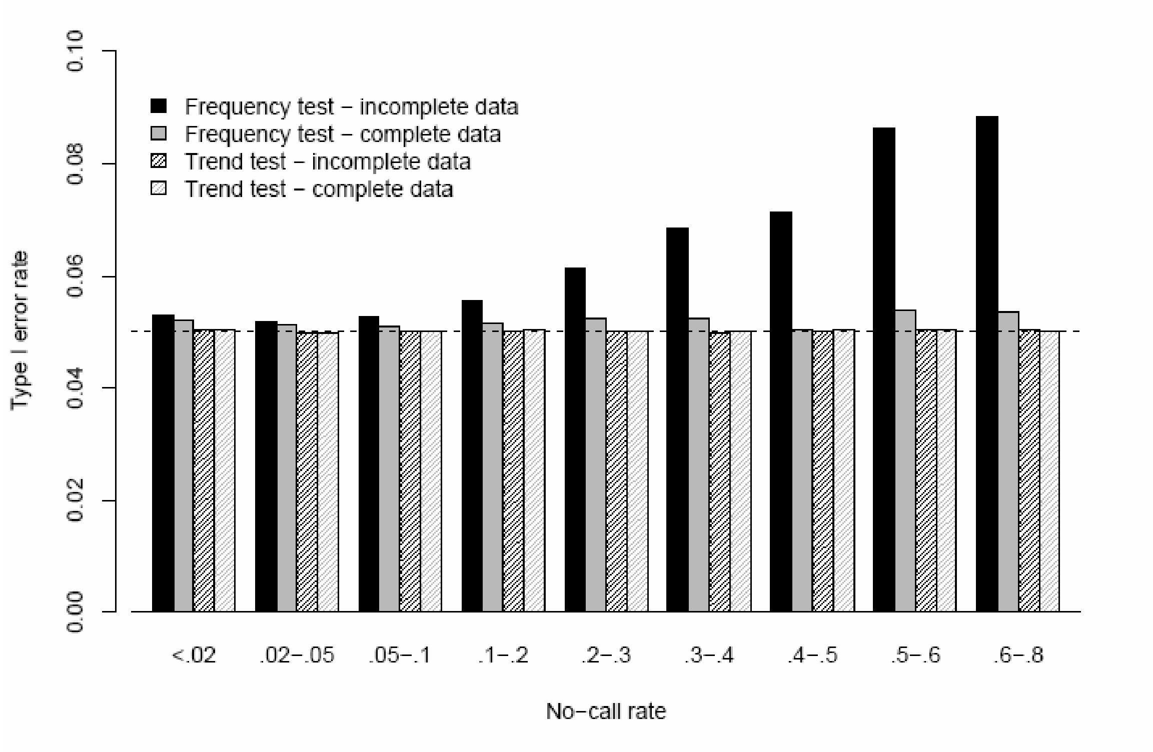


Figure 4.4: Expected i) type I error rate and ii) power for allele frequency test and trend test by proportion of missing genotype data, estimated using replicate Sequenom data

i)



ii)

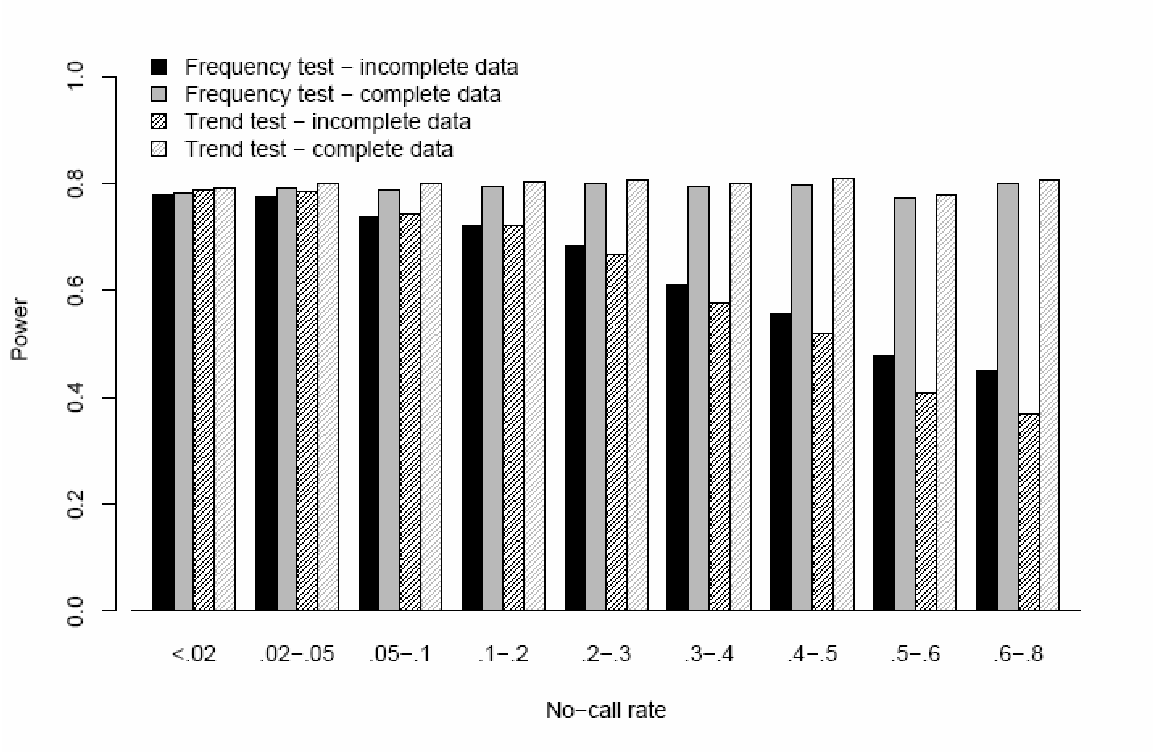


Figure 4.5: Expected type I error rate for TDT by proportion of missing genotype data, estimated using replicate Sequenom data

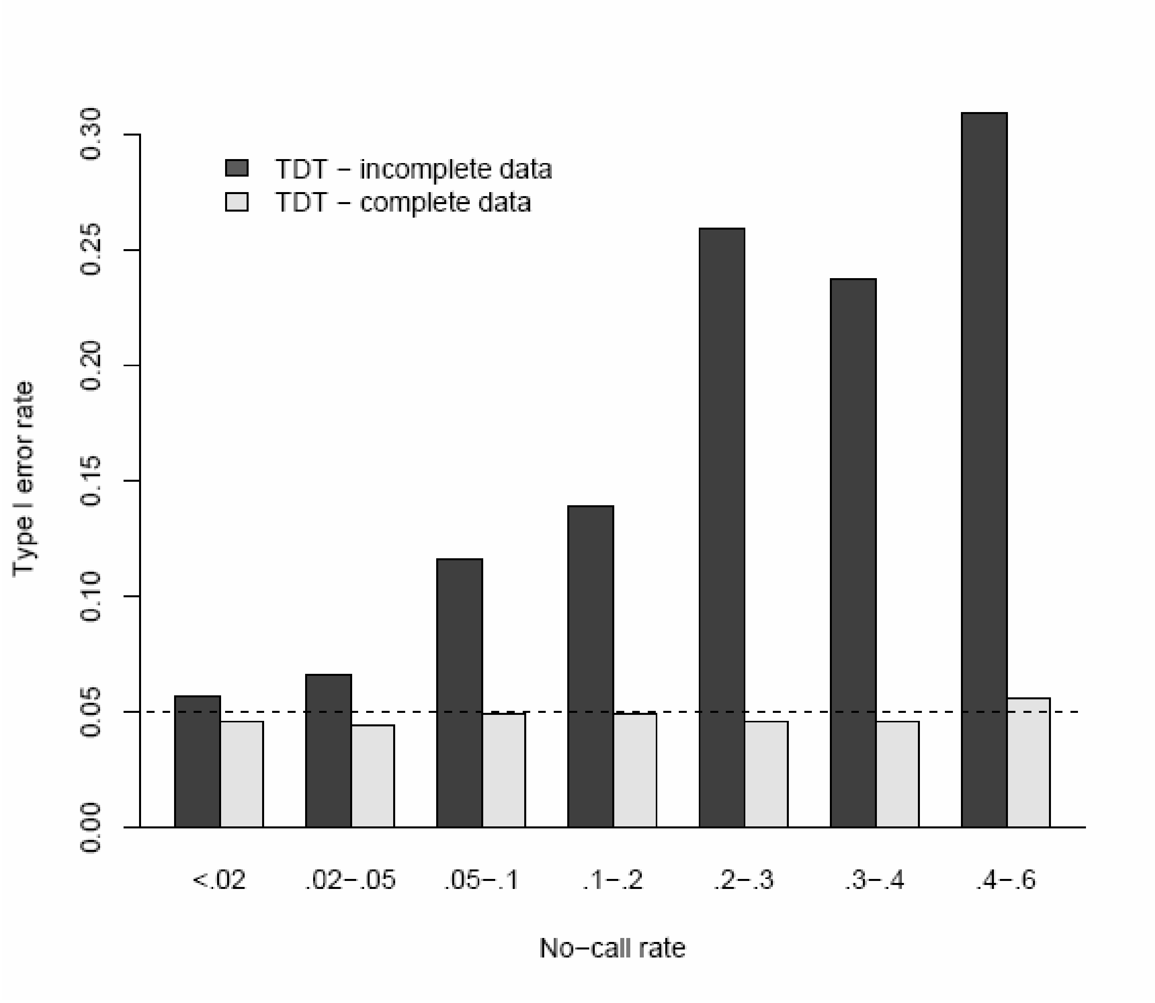
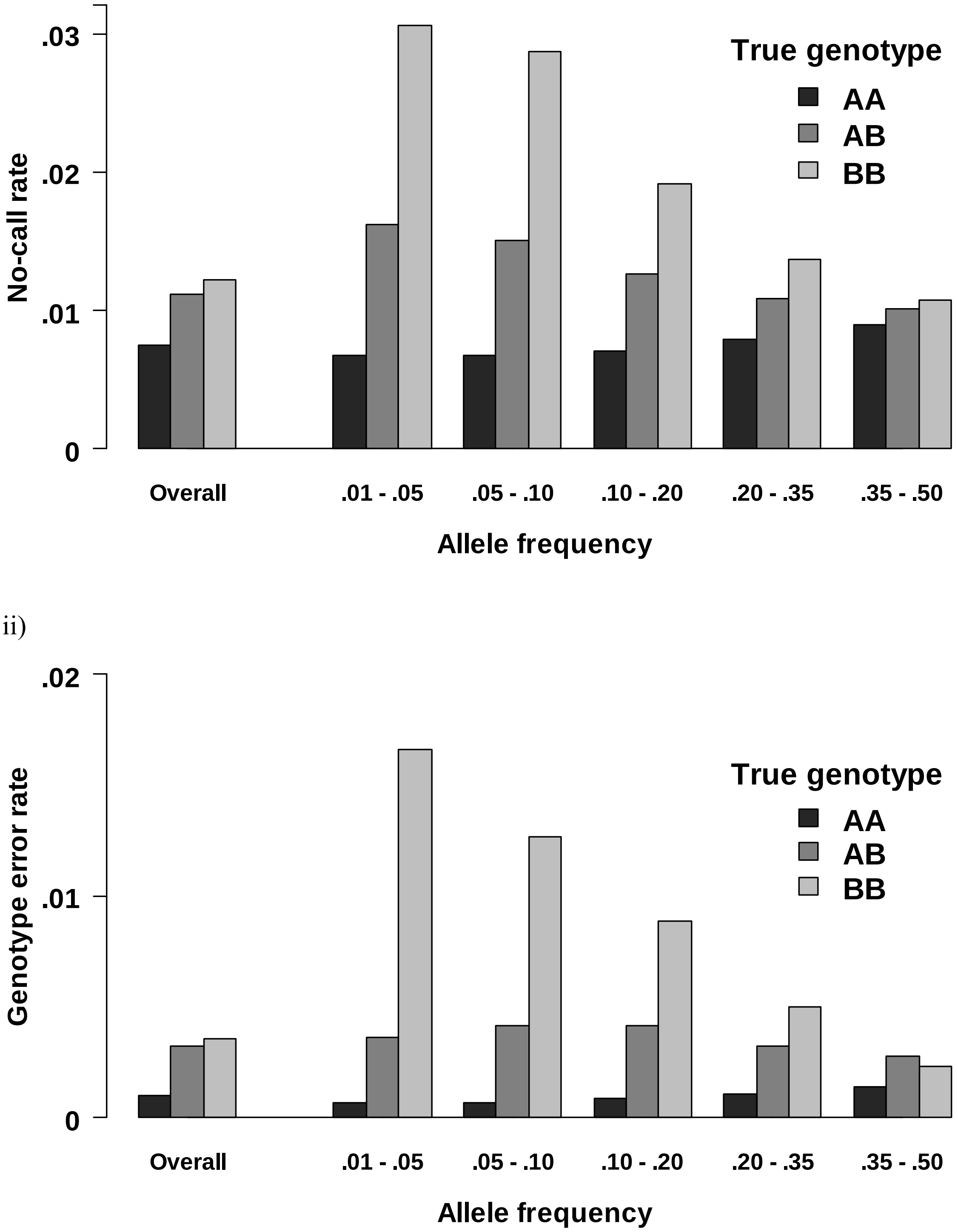


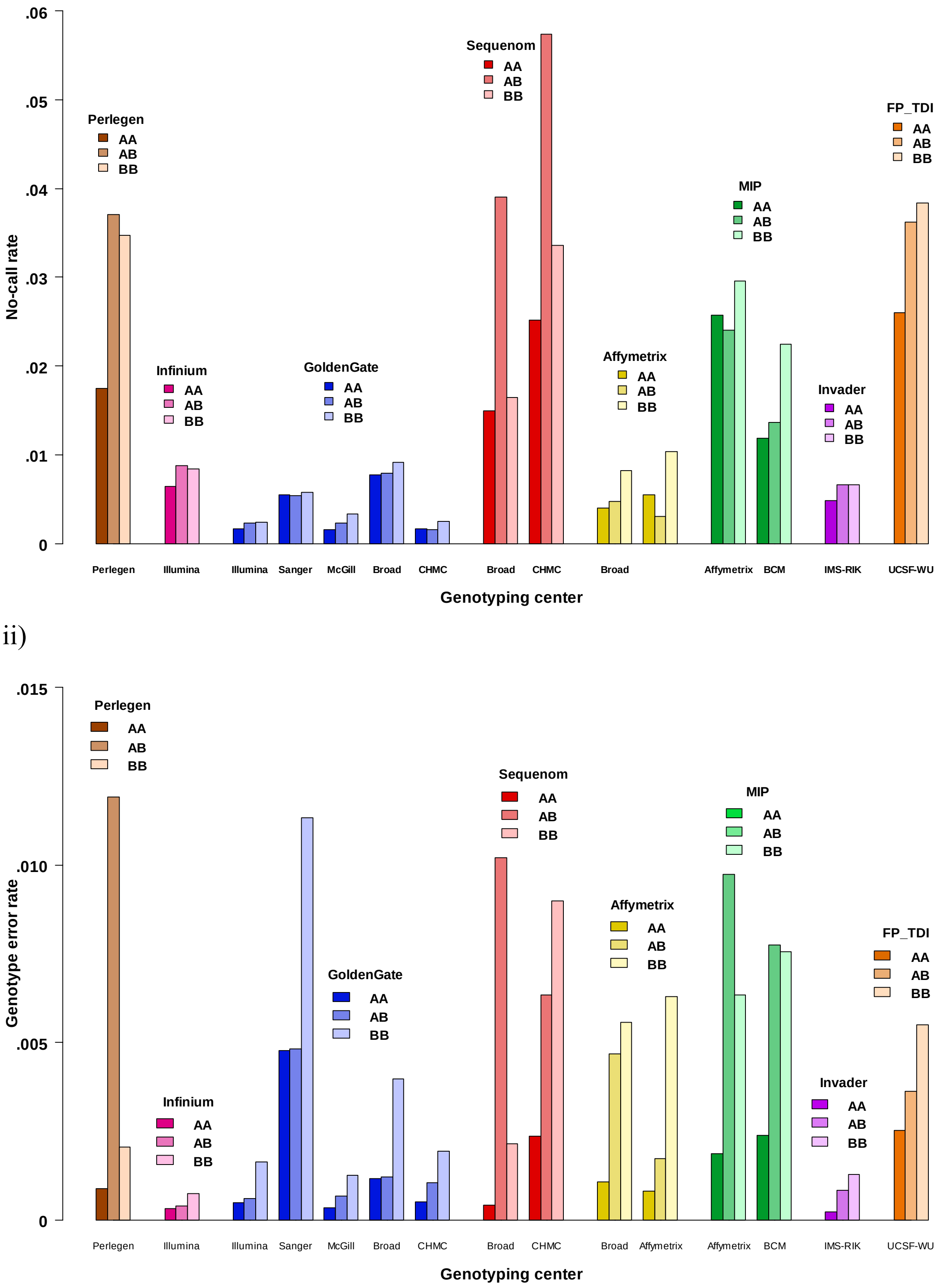
Figure 4.6: Rates of i) missing genotype data and ii) genotype error in replicate Hapmap data by allele frequency and true genotype



i)

ii)

Figure 4.7: Rates of i) missing genotype data and ii) genotype error in replicate Hapmap data by true genotype and genotyping platform



i)

ii)

Figure 4.8: Expected study-wide type I error rate for genome-wide case-control association study, given distributions of genotyping error and missing genotypes

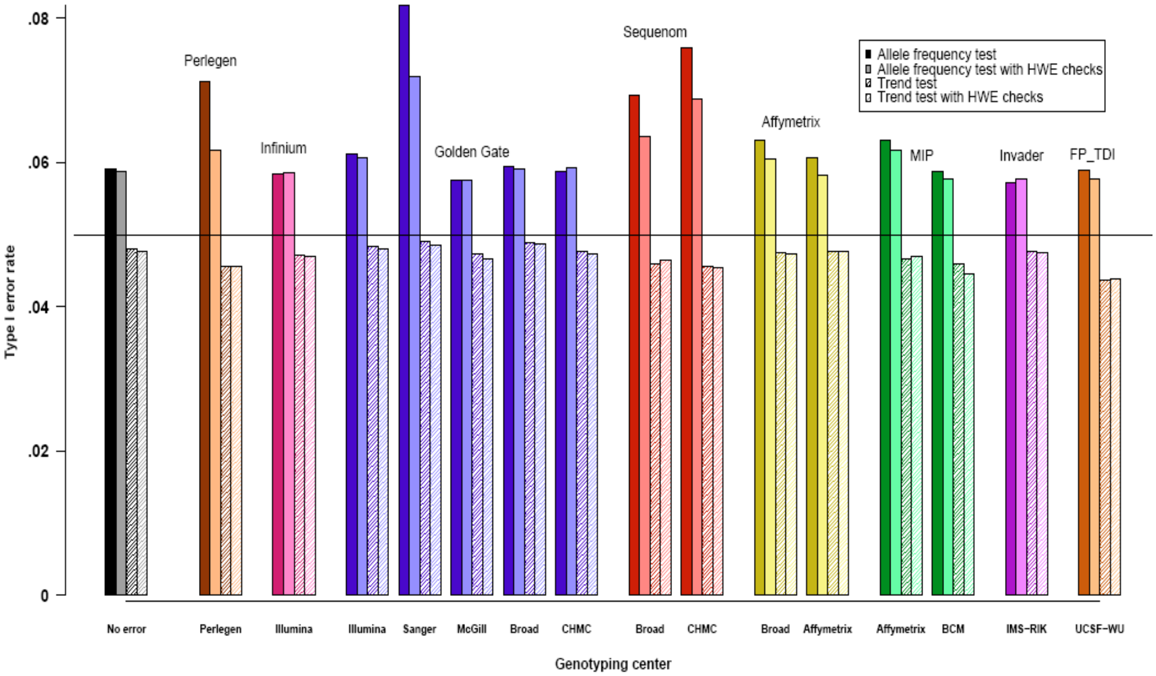
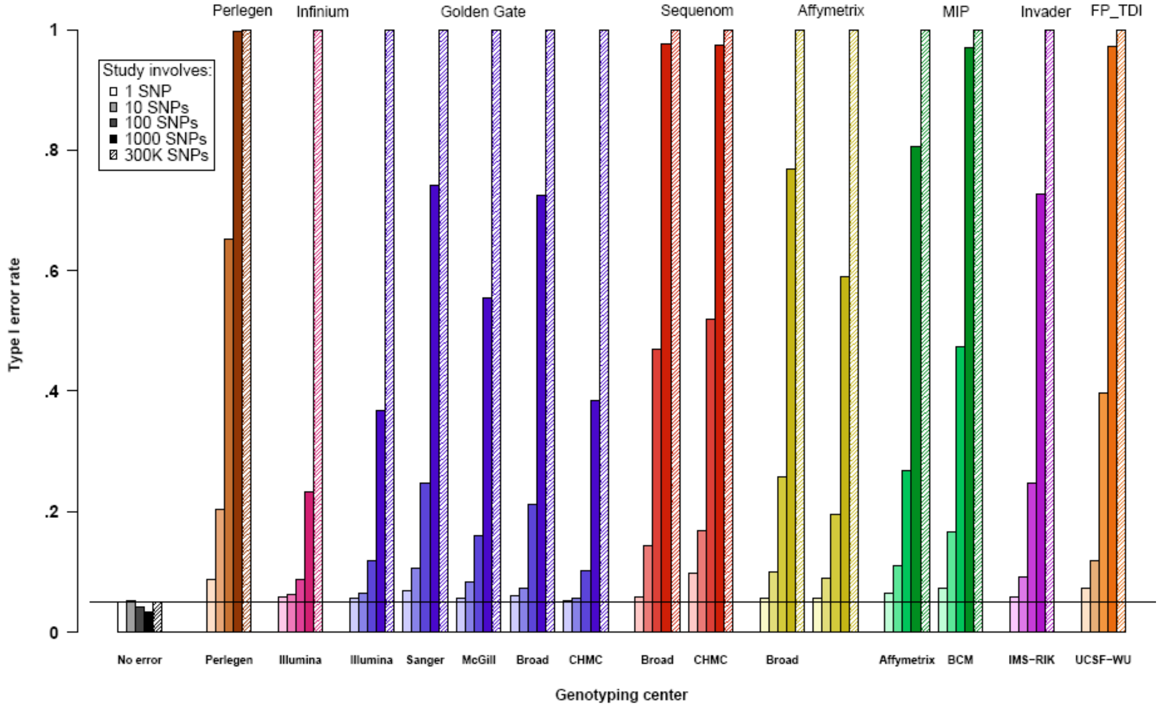
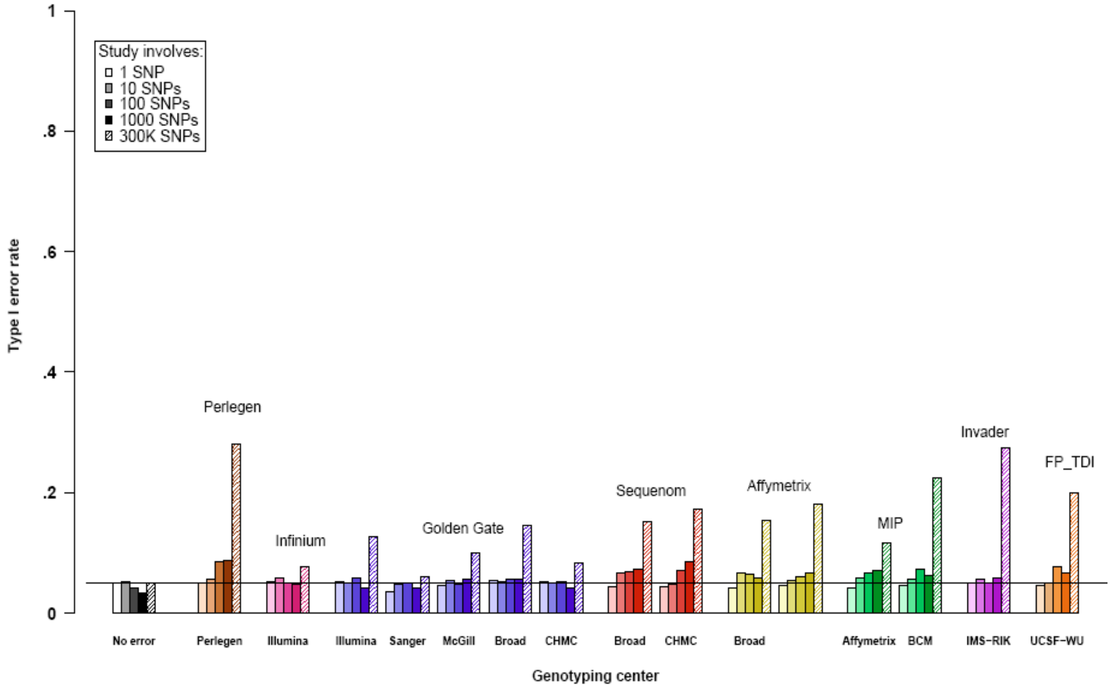


Figure 4.9: Expected study-wide type I error rate for TDT, given distributions of genotyping error and missing genotypes

i) All SNPs with ≥100 informative trios; no other QC filters applied



ii) All SNPs w/ ≥100 informative trios, 95% completeness, ≤2 Mendelian inconsistencies



CHAPTER 5

CONCLUSION

As our ability to obtain high quality genotype data continues to grow, so does our interest in large-scale genetic studies and related statistical issues. Adjusting for multiple tests that are correlated becomes more and more important as we continue to genotype SNPs in greater numbers and density. Large-scale meta-analyses are becoming increasingly common, both because they are now more feasible and because there is a greater volume of positive results to validate. Although the quality of genotyping has improved and will continue to improve, it is still a concern, and considerable resources are still devoted to ensuring genotype quality in most large-scale studies. In this dissertation, I have addressed these issues as follows:

In Chapter 2, I presented a method for dealing with multiple correlated tests in genetic association studies. P_{ACT} (P-value adjusted for correlated tests) can be computed for the minimum P-value or P-values from up to 500 correlated tests which may involve multiple SNPs, traits, and models. We have published this method (Conneely and Boehnke 2007), and have made publicly available our software for computing P_{ACT}. Our method has been applied to adjust P-values based on multiple SNPs, traits, and models in several large candidate gene studies (Bonnycastle et al. 2006; Willer et al. 2007; Gaulton et al. 2007). Based on emails we have received in response to our paper and software, there also appears to be interest in applying this method to more general situations, including microarray-based and protein expression analyses. This is entirely appropriate, since P_{ACT} should be applicable to a wide class of generalized linear models involving single-degree-of-freedom (single-df) tests between multiple correlated explanatory and outcome variables.

Two of the main limitations to P_{ACT} are 1) that it cannot be reliably applied to adjust for more than ~500 tests at once, and 2) that its applicability is limited to single-df tests. These limitations both present directions for future research. Given the current focus on studies of genome-wide association which may involve hundreds of thousands of SNPs, there is high demand for an approach to adjust for genome-wide multiplicity. This could be accomplished with a method that groups SNPs into roughly independent blocks which can then be adjusted with P_{ACT}; it could also be accomplished by addressing genome-wide multiplicity as a whole, perhaps through Monte Carlo methods such as that of Lin (2005a). Application of P_{ACT} to multiple-degree-of-freedom tests would also be a useful extension given current interest in tests involving multiple SNPs and interactions between SNPs. Efforts to derive the appropriate correlation matrices for these types of tests have been fruitless to-date due to the mathematical complexity of the relevant multivariate order statistic distributions. This interesting but possibly unsolvable problem currently occupies a back burner, but may rear its head again one day.

In Chapter 3, I presented an extension to P_{ACT} that adjusts the results of meta-analyses for multiple correlated tests which may be based on multiple SNPs, traits, and models. I discuss how to apply P_{ACT} under four common study designs, including full meta analyses where every test is performed in every sample except where missing-at-random, and follow-up studies where only results passing specific criteria are followed up. My simulation results for adjustment of meta test statistics with P_{ACT} demonstrate accurate control of type I error rates and improved power over adjustment methods which do not account for correlation. My simulations are based on 169 single-degree-of-freedom haplotype association tests based on haplotypes involving the same 6 SNPs. Given the high levels of correlation between sets of overlapping haplotypes, this is another situation where P_{ACT} can be useful. I am currently applying P_{ACT} in a meta-analysis involving highly correlated haplotypes, and am planning to make available software for adjusting meta-analyses with P_{ACT} for use in the ongoing FUSION study (Valle et al. 1998).

A limitation in one of the study designs presented in Chapter 3 is that there is no closed form expression for the power-maximizing sample weights for this particular design. This is probably not a major limitation, since the study design in question (automatic follow-up of the strongest result) is a post-hoc design that should probably not be used at all given that better alternatives such as two-stage designs are available. Also, I was able to attain the appropriate type I error rate and much-improved power over simple Šidák adjustment by applying my method with simple population-based sample weights. Hence, while post-hoc study designs should generally be avoided, P_{ACT} can still be used to avoid potential biases resulting from such a design while maintaining adequate power. Estimation of the optimal weights for this estimator does pose an interesting question for future research, though it is probably not a crucial question given the availability of more powerful and efficient study designs.

In Chapter 4, I investigated the extent to which rates of SNP genotype errors and missing genotype data vary depending on an individual's true genotype, and the potential impact of differential rates on several types of association studies. For two datasets where SNP genotyping had been performed in replicate, I observed that depending on the genotyping platform, either heterozygotes or minor-allele homozygotes were generally at greater risk of being mistyped or missing. It was also clear that these errors and no-calls were not distributed uniformly across plates of genotypes, but were concentrated on a few especially problematic plates. Simulations showed that inclusion of low-quality plates did lead to inflated rates of type I error for the allele frequency test of association and especially the TDT, while the trend test remained robust to low-quality plates. It was also apparent that the standard practice of removing low-quality plates was effective in avoiding bias, although the appropriate cutoff for complete plates is probably higher for the TDT than for other tests. A similar pattern was observed for genome-wide false positive rates, when SNP genotypes had rates of error and missingness drawn from the observed distribution of plate-wide error and no-call rates. A mild anti-conservative bias was observed in the false positive rate for the allele frequency test even in the absence of error and missingness; this is due to the assumption of Hardy-Weinberg equilibrium inherent in the allele frequency test. With genome-wide data, minor deviations from Hardy-Weinberg equilibrium can be expected unless we remove all SNPs with even mild evidence of Hardy-Weinberg failure from consideration. We suggest instead simply using the Cochran-Armitage test for trend (Cochran 1954; Armitage 1955), since it tests the same hypothesis as the allele frequency test but is robust to imperfect data. For the TDT, a very strong anti-conservative bias was observed. This bias could be controlled to some extent through use of standard quality control measures such as removal of SNPs with multiple Mendelian inconsistencies or high levels of incompleteness. The target type I error rate could be achieved with the use of very stringent measures, but these measures led to the removal of ~37% of SNPs. In future work, I plan to investigate whether an optimal combination of quality control measures exists which can achieve the target type I error rate with the removal of fewer SNPs. I am also working on an EM algorithm which uses the information from the incomplete and Mendelian-inconsistent trios to directly model the differential rates of genotype error and missing data. Given the lack of robustness of the TDT to imperfect data and the observed differential rates of genotype error and missingness, this EM-based approach has the potential to be quite useful.

REFERENCES

Ahn K, Haynes C, Kim W, St. Fleur R, Gordon D, Finch S (2006) The effects of SNP genotyping erros on the power of the Cochran-Armitage linear trend test for case/control association studies. Ann Hum Genet 71:249-261

Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11:375-386

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilit `a. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3-62

Bonnycastle LL, Willer CJ, Conneely KN, Jackson AU, Burrill CP, Watanabe RM, Chines PS, Narisu N, Scott LJ, Enloe ST, et al. (2006) Common variants in maturity-onset diabetes of the young genes contribute to risk of type 2 diabetes in Finns. Diabetes 55:2534-2540

Brenner H (1992) Notes on the assessment of trend in the presence of nondifferential exposure misclassification. Epidemiology 3:420-427

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52-58

Cochran WG (1954) Some methods for strengthening the common χ^2 tests. Biometrics 10:417-451

Conneely KN, Boehnke M (2007) So many correlated tests, so little time! Rapid adjustment of *P*-values for multiple correlated tests. Am J Hum Genet 81:1158-1168

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A (2001) High-throughput variation detection and genotyping using microarrays. Genome Res 11:1913-25

Dudbridge F, Koeleman BP (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet 75:424-435

Fajans SS, Bell GI, Polonsky KS (2001) Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. N Engl J Med 345:971–980

Fisher RA (1932) Statistical methods for research workers. Oliver and Boyd, London

Gaulton KJ, Willer CJ, Li Y, Scott LJ, Conneely KN, Jackson AU, Duren WL, Chines PS, Narisu N, Bonnycastle L et al. (2007) Comprehensive association study of type 2 diabetes and related quantitative traits with 222 candidate genes. Diabetes (submitted)

Genz A (1992) Numerical computation of multivariate normal probabilities. J Comput Graph Stat 1:141-149

Genz A (1993) Comparison of methods for the computation of multivariate normal probabilities. Comput Sci Stat 25:400-405

Genz, A (2000) MVTDST: a set of Fortran subroutines, with sample driver program, for the numerical computation of multivariate t integrals, with maximum dimension 100 (increased to 1000 – 7/07); http://www.math.wsu.edu/faculty/genz/homepage)

Genz A (2007) personal communication

Genz A, Bretz F (2002) Comparison of methods for the computation of multivariate t-probabilities. J Comput Graph Stat 11:950-971

Genz A, Bretz F, Hothorn T (2007) mvtnorm: Multivariate normal and t distribution. R package version 0.8-0

Gordon D, Heath SC, Ott J (1999) True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. Hum Hered 49:65-70

Gordon D, Ott J (2001) Assessment and management of SNP genotype errors in genetic association analysis. Pac Symp Biocomput: 18-29

Hao K, Cawley S (2007) Differential dropout among SNP genotypes and impacts on association tests

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95-108

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65-70

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-320

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

James S (1991) Approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. Stat Med 10:1123-1135

Kimmel G, Shamir R (2006) A fast method for computing high-significance disease association in large population-based studies. Am J Hum Gen 79:481-492

Koboldt DC, Miller RD, Kwok PY (2007) Distribution of human SNPs and its effect on high-throughput genotyping. Hum Mutat 27:249-254

Li J, Ji L (2005) Adjusting multiple testing in multilocus analysis using the eigenvalues of a correlation matrix. Heredity 95:221-227

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2007) Markov model for rapid haplotyping and genotype imputation in genome wide studies. Nat Genet (submitted)

Lin DY (2005a) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21:781-787

Lin DY (2005b) On rapid simulation of P values in association studies. Am J Hum Genet 77:513-514

Mantel N (1963) Chi-square tests with 1 degree of freedom – extensions of the Mantel-Haenszel procedure. J Am Stat Assoc 58:690-700

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22:719-748

Marchini J, Howie B, Myers S, McVean G, Donneely P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906-913

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd ed. Chapman and Hall, London

Mitchell AA, Cutler DJ, Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet 72:598-610

Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765-769

R Development Core Team (2007) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org)

Salyakina D, Seaman SR, Browning BL, Dudbridge F, Müller-Myhsok B (2005) Evaluation of Nyholt's procedure for multiple testing correction. Hum Hered 60:19-25

Sasieni PD (1997) From genotypes to genes: Doubling the sample size. Biometrics 53: 1253-1261

Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance-matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4:Article 32

Seaman SR, Müller-Myhsok B (2005) Rapid simulation of p values for product methods and multiple-testing adjustment in association studies. Am J Hum Genet 76:399-408

Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62:626-633

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

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, et al. (1998) Mapping genes for NIDDM: Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Diabetes Care 21: 949-958

Wei LJ, Lin DY, Weissfeld L (1989) Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Am Stat Assoc 84:1065-1073

Wei LJ, Glidden DV (1997) An overview of statistical methods for multiple failure time data in clinical trials. Stat Med 16:833-839

Willer CJ, Bonnycastle LL, Conneely KN, Duren WL, Jackson AU, Scott LJ, Narisu N, Chines PS, Skol A, Stringham HM, et al. (2007) Screening of 134 single nucleotide polymorphisms (SNPs) previously associated with type 2 diabetes replicates association with 12 SNPs in nine genes. Diabetes 56:256-264

Zaykin DV, Zhivotovskty LA, Westfall PH, Weir BS (2002) Truncated product method for combining P-values. Genet Epidemiol 22:170-185