BMC Bioinformatics

Methodology article

Genotyping and inflated type I error rate in genome-wide association case/control studies

Joshua N Sampson* and Hongyu Zhao

Address: Department of Epidemiology and Public Health, Yale University School of Medicine, New haven, CT, USA

Email: Joshua N Sampson* - joshua.sampson@yale.edu; Hongyu Zhao - hongyu.zhao@yale.edu

* Corresponding author

Published: 23 February 2009

BMC Bioinformatics 2009, **10**:68 doi:10.1186/1471-2105-10-68

Received: 12 December 2008

Accepted: 23 February 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/68

Abstract

Background: One common goal of a case/control genome wide association study (GWAS) is to find SNPs associated with a disease. Traditionally, the first step in such studies is to assign a genotype to each SNP in each subject, based on a statistic summarizing fluorescence measurements. When the distributions of the summary statistics are not well separated by genotype, the act of genotype assignment can lead to more potential problems than acknowledged by the literature.

Results: Specifically, we show that the proportions of each called genotype need not equal the true proportions in the population, even as the number of subjects grows infinitely large. The called genotypes for two subjects need not be independent, even when their true genotypes are independent. Consequently, p-values from tests of association can be anti-conservative, even when the distributions of the summary statistic for the cases and controls are identical. To address these problems, we propose two new tests designed to reduce the inflation in the type I error rate caused by these problems. The first algorithm, logiCALL, measures call quality by fully exploring the likelihood profile of intensity measurements, and the second algorithm avoids genotyping by using a likelihood ratio statistic.

Conclusion: Genotyping can introduce avoidable false positives in GWAS.

Background

One common goal of a genome wide association (GWAS) study is to search the entire genome for single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) associated with a disease or some other phenotype. In this article, we focus our analysis on SNPs. The two possible alleles at a SNP are arbitrarily labeled A and B, and association is often tested by measuring and comparing the frequencies of the genotypes AA, AB, and BB, in case and control groups. As technology currently allows close to one million SNPs to be examined simultaneously, there is a need for fast, automated methods to test for association. As only a small minority of SNPs are expected to be associated with the disease, even a modest false positive rate could bury true associations beneath those occurring by chance.

Using Affymetrix 500 k GeneChips as an example, each of the 500,000+ SNPs is represented by a series of probes on a pair of arrays. Each probe is an oligonucleotide designed to bind to either the A or B allele. A subject's fluorescently labeled DNA is allowed to hybridize with these probes, and then a spectrometer measures the relative fluorescent levels between the A and B probes. Each genotyping algo-

rithm (see *Methods*) summarizes the fluorescence information, or the likelihood a subject has allele A at that SNP, by its own statistic. In any population, these statistics usually cluster into three groups, corresponding to the three genotypes. Studies have noted that 1) The mean and variance of these clusters, or the shape of their distribution in general, varies by SNP and 2) For a single SNP, because of differences in processing or duration of storage, the shape of the statistic's distribution can differ between the case and control groups [1-3].

The current test of association requires *calling*, or assigning a genotype, to each SNP for each study subject and then comparing the called genotypes between the case and control groups. The majority of SNPs are easy to call and any of the available methods will call them correctly. Unfortunately, there is a difficult minority that cannot be easily clustered into three distinct groups. Because there can be as many as 500,000 SNPs, this minority can greatly inflate the type I error rate and cause the large, characteristic, deviations from the x = y line in the qq-plots of test statistics. Most studies assume these consequences are the unavoidable results of population substructure and poor data. In this paper, we dispel the myth that these are the sole issues. In fact, the inflated type I error rate and general misbehavior of the test statistic may also result from the act of genotype assignment and a poor choice of statistical methodology.

The goal of this manuscript is two-fold. First, our primary goal is to show that genotyping overlapping clusters can lead to potential problems that we have yet to see fully acknowledged in the literature. The proportions of each called genotype need not equal their true proportions in the population, *even as the number of subjects grows infinitely large*. As we compare genotype calls, p-values from tests of association will be anti-conservative when the distribution of the summary statistic differs between cases and controls. Moreover, the called genotypes for two subjects need not be independent, *even when their true genotypes are independent*. Therefore, p-values from tests of association can be anti-conservative, *even when the distributions of the summary statistic for the cases and controls are identical*, a fact we believe has yet to be fully demonstrated. Although previous studies have examined the effects of genotyping error on tests of association [4-6], studies has neither fully explored the effects caused by case/control differences in distributions nor dependence of error. Second, we discuss two new tests that can circumvent these potential problems. One test compares calls made from a genotyping algorithm designed to minimize the type I error. The second test compares the fluorescence distributions, instead of the called genotypes. We start the *Methods section* by discussing currently available methods for genotyping SNPs and testing for association. Concurrently, we introduce logiCALL, our new genotyping algorithm, and the likelihood ratio-based test of association. In the *Results and Discussion section*, we start by showing that the proportion of genotypes called AA, AB, and BB need not converge to the true population proportions. Then we discuss how the called genotypes can be dependent. We conclude the section by comparing the proposed tests of association with the current standards through both simulation studies and real data analysis. Then a short *Conclusions section* summarizes the key points.

Methods

Calling Genotypes

There is currently a wide variety of programs available for genotyping SNPs. The most popular supporting Affymetrix are RLMM [7], BRLMM [8], CRLMM [9], CHIAMO [10], SNiPer-HD [11], and MAMS [12]. The most popular program for Illumina is their own proprietary software, BeadStudio, but other methods have been recently suggested by Moorhead et. al. [3], Teo et. al. [13], and Dunning et. al. [14] (Table 1). To introduce these methods, we start by defining notation. Let there be n subjects. Let $G_{ij} \in \{AA, AB, BB\}$ be the true genotype of SNP j in subject i, $1 \le i \le n$. For Affymetrix chips, assume there are n_p probe quartets representing each SNP on an array, and let $\vec{I}^{ijk} \equiv \{I^{ijk}_{PM_A}, I^{ijk}_{MM_A}, I^{ijk}_{PM_B}, I^{ijk}_{MM_B}\}$ be the normalized probe intensities for subject i, SNP j, and probe k. Here, the subscripts PM_A and MM_A signify the perfect match and mismatch probes for allele A. PM_B and MM_B are similarly defined for allele B. The log transformed intensities will be $\vec{Y}^{ijk} = log_2(\vec{I}^{ijk})$. To maintain notational consistency, for Illumina chips, we denote the BeadStudio intensity values of the two SNP alleles by $\{I^{ij}_{PM_A}, I^{ij}_{PM_B}\}$. As we will only discuss a single SNP for the majority of the paper, we will omit the superscript "j" from all future notation.

With the exception of dynamic modeling (DM) [15], all calling algorithms share the same general form, and we exploit this form to summarize their key features. The process of transforming raw signal into genotypes can be divided into four steps: 1) Normalize the intensity values; 2) Describe the normalized intensity values by a single, possibly multivariate, summary statistic; 3) Estimate the mean and variance of the summary statistic for the three possible genotypes, AA, AB, and BB; and 4) Compare the value of the statistic from a subject to the group parameters found in the third step to make the call. The universal first step, normalizing the intensity values, is tangential to our discussion here. The second step is to choose a statistic, S_i that Summarizes the intensities. For example,

Table 1: Programs available for genotyping SNPs.

Name	Summary Statistic	MM[1]	Data[2]	Data[3]	Notes
RLMM	$\{\Theta_A, \Theta_B\}$	No	T	M	
BRLMM	$\frac{1}{asinh(4)} asinh\left(\frac{4(I_A - I_B)}{I_A - I_B}\right)$	No	E-U	M	Assumes genotypes in "training" data are known. "Training" data only uses high quality SNPs. Incorporates info from other SNPs as a *B*ayesian Prior.
CRLMM	$\{\Theta_{A+} - \Theta_{B+}, \Theta_{A-} - \Theta_{B-}\}$	No	T	L	*C*orrects for the effect of total intensity level and probe length on $\{S_{ij}\}$ through more complex method, and allows corrections to vary by array.
CHIAMO	$\left\{\frac{1}{np}\sum_{k=1}^{n_p} Y_{Ak}, \frac{1}{np}\sum_{k=1}^{n_p} Y_{Ak}\right\}$	Yes	E-L, T*	W	CHIAMO is a Bayesian hierarchical mixture model and is greatly simplified by this brief summary
SNiPer-HD	$\{R_1, ..., R_{n_p}\}$, where $R_k = (I^{ijk}_{PM_A})/(I^{ijk}_{PM_A} + I^{ijk}_{PM_B})$	No	E-U	W	Assumes genotypes in "training" data are unknown and requires the EM algorithm. "Training" data should only use high quality SNPs.
Moorhead	$\frac{1}{sinh(2)} sinh\left(\frac{I_{PM_A} + I_{PM_B}}{I_{PM_B}}\right)$	N/A	E-U	W	Originally for MIP, but applicable to Affymetrix. Plagnol demonstrated how to link genotype probabilities between cases and controls.
logiCALL	$\left\{\frac{1}{sinh(2)} sinh\left(\frac{I_{PM_A} + I_{PM_B}}{I_{PM_B}}\right), I_{PM_A} + I_{PM_B}\right\}$	No	E-L	W-F	Designed to lower false positive rate and assigns calls based on cumulative distribution, not density functions.

[1]Indicates use of mismatched probes
[2]Parameters were estimated by *E*xperimental or *T*raining data. For experimental data, under the null, cases and control genotype proportions could be *L*inked or *U*nlinked. * indicates optional.
[3]Distance can be *M*ahalanobis, *W* eighted Likelihood, or unweighted *L*ikelihood

RLMM models the probe intensities as $Y^{ik}_{PM_A} = \zeta^i_A + \beta^k_A + \varepsilon^{ik}_A$ and $Y^{ik}_{PM_B} = \zeta^i_B + \beta^k_B + \varepsilon^{ik}_B$. Then, $S_i \equiv \{\zeta^i_A, \zeta^i_B\}$. In the updated version, CRLMM models sense(+) and antisense(-) probes separately, resulting in a 4D statistic, $S_i \equiv \{\zeta^i_{A+}, \zeta^i_{A-}, \zeta^i_{B+}, \zeta^i_{B-}\}$ Each method assumes that the distribution, $\phi^M(S_i|\theta^q)$, of their statistic in a given population q is a Mixture of multivariate normal distributions where θ, to be defined below, are the parameters characterizing the distribution in population q. When it is clear we are discussing only a single population, we will omit the superscript q. Although problems can arise when the distribution is not a true mixture of normals, those complications are beyond the scope of this paper [16]. For completeness, we point out that a minority of programs, including CRLMM, allow these parameters to vary within a group (e.g. to be subject/array specific). Ignoring that some methods allow for an additional null distribution, the general form is

$$\begin{aligned}\phi^M(S_i \mid \theta) = \; & p_{AA}\phi(S_i \mid \vec{\mu}_{AA}, \Sigma_{AA}) + \\ & p_{AB}\phi(S_i \mid \vec{\mu}_{AB}, \Sigma_{AB}) + \\ & p_{BB}\phi(S_i \mid \vec{\mu}_{BB}, \Sigma_{BB})\end{aligned} \quad (1)$$

where $\psi(\cdot)$ is the multivariate normal density. The three mean vectors, $\mu \equiv \{\vec{\mu}_{AA}, \vec{\mu}_{AB}, \vec{\mu}_{BB}\}$, variance matrices, $\Sigma \equiv \{\Sigma_{AA}, \Sigma_{AB}, \Sigma_{BB}\}$, and probabilities, $p \equiv \{p_{AA}, p_{AB}, p_{BB}\}$, correspond to the three possible genotypes, AA, AB, and BB. Define $\theta \equiv \{\vec{\mu}, \Sigma, \vec{p}\}$ and later, we let $\Phi(\cdot)$ and $\Phi^M(\cdot)$ be the cumulative distribution for a normal variable and a mixture of normals. The third step is to estimate θ. Some algorithms, such as RLMM, use a training data set, where the genotypes are known. Other algorithms, such as SNiPer-HD, use no training data, and find the best estimates that describe their experimental values. The fourth step is to assign a genotype, $\hat{G}_{ij}$, to SNP j in subject i. Often, for a given value of S_i, the assigned genotype maximizes a similarity function: $\hat{G}_i = argmax_g D(g, S_i|\theta)$. The similarity function is usually a modified version of one of

the following three quantities: Mahalanobis distance: $-(S_i - \vec{\mu}_g)\Sigma_g^{-1}(S_i - \vec{\mu}_g)^t$, Unweighted Likelihood: $\phi(S_i \mid \vec{\mu}_g, \Sigma_g)$, or the Weighted Likelihood: $p_g\phi(S_i \mid \vec{\mu}_g, \Sigma_g)$. The similarity function is also modified to ensure monotonicity of assignment. When we let $S_i = \{M_i, A_i\}$, we force $D(AB, S_i|\theta) = D(BB, S_i|\theta) = 0$ if $M_i \le \mu_{AA}$, $D(BB, S_i|\theta) = 0$ if $\mu_{AA} \le M_i \le \mu_{AB}$, $D(AA, S_i|\theta) = 0$ if $\mu_{AB} \le M_i \le \cdot\ \mu_{BB}$, and $D(AB, S_i|\theta) = D(AA, S_i|\theta) = 0$ if $M_i \ge \mu_{BB}$, where μ_g, in this case, is the mean of M_i when $G_i = g$. This modification is standard in calling algorithms. As we do not know the true value of the parameters in experiments, we replace $D(g, S_i|\theta)$ by $D(g, S_i|\hat{\theta})$. A subject's SNP may not be called, or assigned a missing value, if the difference or ratio between $D(\hat{G}_i, S_i|\hat{\theta})$ and $D(g_{2i}, S_i|\hat{\theta})$ is not large enough, where $g_{2i}$is the genotype with the second largest value of $D(g, S_i|\hat{\theta})$. A SNP may be omitted from further study if too many values were set to missing. Table 1 describes the details of the four steps for popular methods. For many purposes or to understand the details of the method, especially in handling rare alleles, this table will seem an oversimplification. For our purposes here, it highlights the features of interest.

Tests of Association

The current tests of association start by calling genotypes for a given SNP j in a group of subjects with the disease and in a group of controls. They then compare the resulting proportions, $\hat{P}^A \equiv \{\hat{P}^A_{AA}, \hat{P}^A_{AB}, \hat{P}^A_{BB}\}$ and $\hat{P}^U \equiv \{\hat{P}^U_{AA}, \hat{P}^U_{AB}, \hat{P}^U_{BB}\}$, from these Affected and Unaffected groups using either a Cochran-Armitage test or logistic regression. Here $\hat{p}^q_g \equiv \frac{1}{nq}\sum_{i:Q_{i=q}} 1(\hat{G}_i = g)$, where the indicator function is defined by 1(x) = 1 if x is true, 0 otherwise, $Q_i \in \{A, U\}$ is the disease status for individual i, and n_q is the number of subjects with disease status q. In this manuscript, any 'p-value' from a genotype-based association test will be calculated using ANOVA on the logistic regression model with Q_i and genotype (unordered grouping) as the dependent and independent variables.

Standard tests tend to err anti-conservatively as we will discuss below. We will propose four alterations that can reduce type I error rate, with only a minimal decrease in power. These are the four differences that separate logiCALL from standard methods. The first is based on the observation, which is discussed later, that the likelihood profile of $\phi^M(S_i|\theta)$ will have multiple local maxima near the overall maximum. When estimating θ, the EM algorithm converges to multiple solutions. For many of those solutions, the resulting parameter set, $\hat{\theta}_{lm}$, satisfies $\prod_i \phi^M(S_i \mid \theta) \le \prod_i \phi^M(S_i \mid \theta_{lm}) + \tau n$. For each parameter set satisfying this inequality, we will make a new group of genotype assignments, $\{G_i^{lm}\}$. If more than 10% of such assignments disagree with $\hat{G}_i$ (τ = 0.06), we label that subject's call as questionable. We also continue the practice of marking calls with small values of $D(\hat{G}_i, S_i|\hat{\theta})/D(g_{2i}, S_i|\hat{\theta})$ as questionable. The second alteration is that we do not discard questionable calls, an act which can create false positives. Instead, we assign questionable S_i so the proportions of genotypes in the cases and controls are as similar as possible, which is defined as minimizing $\sum_g |p_g^U - p_g^A|$, with the restriction that the final call for subject i must be either $\hat{G}_i$ or g_{2i}. Here, we let g_{2i} be the genotype which is either the runner-up in terms of distance or the most common genotype among the dissenting calls, depending on why the genotype was labeled as questionable. The third alteration, which is already incorporated into other programs is to perform the EM algorithm under the null hypothesis that the genotype proportions in the two populations are identical [3]. The fourth is the use of a weighted Mahalonobis distance, which is defined later. Given these changes, logiCALL then compares the estimated genotype proportions in cases and controls using logistic regression. Note that none of the changes affect calls for the vast majority of SNPs.

We also introduce a completely new method for testing association based on a likelihood ratio statistic. For our method, steps 1 and 2, normalization and choice of summary statistic, can mimic any of the previously described methods. As our real data to be analyzed was collected on Illumina chips, we choose the statistic from Moorhead, et al. [3], $S = \frac{1}{sinh(2)} sinh(\frac{I_{PM_A} + I_{PM_B}}{I_{PM_B}})$ for exposition. We then assume that S_i follows the mixture model described by equation (1), but allow the parameters to differ by disease status:

$$\phi^M_{AU}(S_i \mid \theta) \equiv \begin{cases} \phi^M(S_i \mid \theta^A), Q_i = A \\ \phi^M(S_i \mid \theta^U), Q_i = U \end{cases} \quad (2)$$

and $\theta = \{\theta^A, \theta^U\}$, where

$$\theta^A = \left\{ p^A_{AA}, \mu^A_{AA}, \sigma^{2A}_{AA}, p^A_{AB}, \mu^A_{AB}, \sigma^{2A}_{AB}, p^A_{BB}, \mu^A_{BB}, \sigma^{2A}_{BB} \right\}$$
$$\theta^U = \left\{ p^U_{AA}, \mu^U_{AA}, \sigma^{2U}_{AA}, p^U_{AB}, \mu^U_{AB}, \sigma^{2U}_{AB}, p^U_{BB}, \mu^U_{BB}, \sigma^{2U}_{BB} \right\}$$

Although θ contains 18 parameters, it has only 16 degrees of freedom (df) because $p^q_{AA} + p^q_{AB} + p^q_{BB} = 1$ for $q \in \{A, U\}$. Our new test will reject the null hypothesis of no association, when $LR(\vec{S})$, the likelihood ratio, is large, where

$$LR(\vec{S}) = \left(\frac{max_{\theta \in \Omega} \prod \phi^M_{AU}(S_i|\theta)}{max_{\theta_R \in \Omega_R} \prod \phi^M_{AU}(S_i|\theta_R)}\right) \quad (3)$$

Clearly, the restricted parameter space, $\{\Omega_R : p^A_{AA}, p^U_{AA}, p^A_{AB} = p^U_{AB}, p^A_{BB} = p^U_{BB}\}$ is a subset of the unrestricted parameter space, Ω. In an ideal scenario, the distribution of 2log(LR) would converge to a chi-squared distribution, $F_{\chi^2_2}$, with 2 degrees of freedom. Therefore, the 'p-value' from a likelihood ratio-based test will be calculated as 1 - $F_{\chi^2_2}(2log(LR))$.

Data Source

To demonstrate the problems of genotyping and compare the genotype- and likelihood ratio-based tests of association, we use three types of data. First, for discussion, we may assume a hypothetical study measuring a one dimensional summary statistic, S_i, for a SNP *j* with only two possible genotypes, $G_i \in \{0, 1\}$. Furthermore, to show problems can exist even under the best conditions, where model and truth coincide, we assume that S_i follows a normal distribution given Q_i and G_{ij}, and that the full distribution can be described by

$$\phi^M(S_i \mid \theta) = p_0\phi(S_i \mid \mu_0, \sigma^2_0) + p_1\phi(S_i \mid \mu_1, \sigma^2_1).$$

We compare our two new tests of association to a standard method using simulated data. The standard method mimics the general Bead-Studio approach by a) fitting parameters with the EM algorithm; b) calling genotypes based on the Mahalanobis Distance; c) removing all calls where $D(\hat{G}_i, S_i|\hat{\theta})/D(g_2, S_i|\hat{\theta}) > 0.5$; and d) comparing the two sets of resulting estimates, $\{P^A_{AA}, P^A_{AB}, P^A_{BB}\}$ and $\{P^U_{AA}, P^U_{AB}, P^U_{BB}\}$. We generated 10 simulated datasets, containing 1000 subjects (500 cases, 500 controls) and 303,100 SNPs for each of 18 scenarios. For each gene *j* and each subject *i*, we generated a 2D summary statistic (M_{ji}, A_{ji}). The distribution of M_{ji} depended on genotype. If $G_{ji} = AA$, then $M_{ji} \sim 2X - 1$, and if $G_{ji} = BB$, then $M_{ji} \sim 1 - 2X$, where $X \sim beta(\alpha = 3, \beta = 30)$. If $G_{ji} = AB$, then M_{ji} also followed a beta distribution, but the parameters varied by SNP, disease status, and scenario. For all SNPs and all subjects, $A_{ji} \sim N(10, 1.5)$. We generated three types of SNPs, background, shifted, and influential. First, 300,000 background SNPs were generated and included in all 10 × 18 = 180 data sets. For each SNP, a single minor allele frequency was generated from a uniform(0.2, 0.4) distribution and genotype probabilities ($\vec{p}^U = \vec{p}^A$) were generated assuming Hardy-Weinberg Equilibrium. Here, $E[M_{ji}|G_{ji} = AB] = 0$ and was independent of disease status. These SNPs, which formed three distinct clusters, can be easily identified and represent a well-behaved group. For 3,000 shifted SNPs, MAF ~*uniform*(0.2, 0.4) and genotype probabilities ($\vec{p}^U = \vec{p}^A$) were generated assuming Hardy-Weinberg Equilibrium. Here, $E[M_{ji}|G_{ji} = AB] \in \{-0.739 + 0.2, -0.739 + 0.3 - 0.739 + 0.5\}$, where we note $E[M_{ji}|G_{ji} = AA] = -0.739$ and $E[M_{ji}|G_{ji} = AB, Q_i = A] - E[M_{ji}|G_{ji} = AB, Q_i = U] \in \{0, 0.2\}$. This group represents difficult to call SNPs. For 100 influential SNPs, MAF ~*uniform*(0.2, 0.4) and genotype probabilities ($\vec{p}^U \neq \vec{p}^A$) were chosen so that, under a disease prevalence of 0.01 and a model of additive effects, the genotype relative risk for subjects homogeneous for the minor allele, $P(Q_i = A|BB)/P(Q_i = A|AA) \in \{1.5, 2.0, 2.5\}$. Combing the degree of shift for poor quality SNPs and the effect size of truly associated SNPs, we have a total of 18 scenarios used in our simulation.

The next set of data is from a recent GWAS of Inflammatory Bowel Disease (IBD) that compared 983 subjects with IBD to 1004 subjects without the disease. Using Illumina microarrays, 308,330 SNPs on the autosomal chromosomes were tested. Jewish and non-Jewish cohorts, approximately equal in size, were analyzed separately, a practice continued here. Details have been previously published [17,18]. Because the overwhelming majority of the SNPs are easy to genotype, as any of the summary statistics neatly divide into three clusters, we chose a 3137 *difficult* SNP subset where at least two clusters overlap

(definition below). Because association was tested separately in Jewish and non-Jewish cohorts, a total of 3137 × 2 = 6274 tests were possible. To demonstrate called genotype dependency, we bootstrapped 40 samples, ignoring case/control status, of 500 subjects for each SNP. A sample would be discarded, and replaced by another random selection, if one or both of the groups were lacking an AA (S_i<-0.7) or a BB (S_i > 0.7) genotype.

We defined *difficult* SNPs as follows. For SNP *j*, we first estimated the density, $\hat{f}(M_{j.})$ nonparametrically using the R function 'density(adjust = 0.3)'. In theory, $\hat{f}(M_{j.})$ is a mixture of three normal distributions, corresponding to the three genotypes. If the SNP is well-behaved, then the three underlying densities will not overlap, and the empirical density $\hat{f}(M_{j.})$ will attain minima near 0 in the valleys between μ_{jAA} and μ_{jAB} and between μ_{jAB} and μ_{jBB}. If either of these minima exceeded 0.2, then at least two of underlying densities overlapped, and that SNP was defined as *difficult*. To speed the process, we found that approximating the center of the peaks (i.e. μ_{jAA}, μ_{jAB}, and μ_{jBB}) by the median values of $M_{j.}$ in the three windows, $\{M_{j.} \le -0.6; -0.3 \le M_{j.} \le 0.3, M_{j.} \ge 0.5\}$, worked well.

Results and discussion

Two Genotype Example: Parameters

We choose to use the hypothetical, two-genotype, study, to highlight that the estimated parameters can be inconsistent even in the simplest scenario, where the summary statistic is distributed normally and our fitted model is correct. When dealing with only a single population, we define the parameter p_0 (p_1) to be the probability that a subject's *true* genotype is 0 (1).

$$p_0 \equiv P(G_i = 0) \text{ and } p_1 \equiv P(G_i = 1) \quad (4)$$

We define the parameter $p_0^{n^*}(p_1^{n^*})$ to be the probability that a subject's *called* genotype is 0 (1).

$$p_0^{n^*} \equiv P(G_i = 0 \mid n, \theta) \text{ and } p_1^{n^*} \equiv P(G_i = 1 \mid n, \theta) \quad (5)$$

Probabilities of called genotypes implicitly depend on the number of subjects in the sample. We then define the parameter c^* to be the point which is equidistant to the two genotype groups when θ is known.

Therefore, $\mu_0 \le c^* \le \mu_1$ is defined to be a solution to equation 6

$$D(G_i = 0, S_i = c^* \mid \theta) = D(G_i = 1, S_i = c^* \mid \theta) \quad (6)$$

For the remainder of the paper, we shall assume such a c^* exists. This assumption is safe in practice as genes with extremely rare minor alleles are discarded. When D is the Mahalanobis distance, c^* is a solution to

$$\frac{(c^* - \mu_0)^2}{\sigma_0^2} = \frac{(c^* - \mu_1)^2}{\sigma_1^2} \quad (7)$$

We define the final parameter, $p_0^*(p_1^*)$, to be the probability that a subject's S_i value is less than c^* given their true genotype is 0 (1).

$$p_0^* \equiv P(S_i < c^* \mid G_i = 0, \theta)$$
$$p_1^* \equiv P(S_i < c^* \mid G_i = 1, \theta) \quad (8)$$

Our first goal in this *Results section* is to show that

$$p_0^* \neq p_0 \quad (9)$$

may be true when two clusters, $\{S_i : G_i = 0\}$ and $\{S_i : G_i = 1\}$ overlap. We will refer to $p_0^* - p_0$ as asymptotic bias, or bias, and we note that it depends on D, p_0, and the magnitude of the overlap. Here, we also define C, our estimate for c^*, to be the solution to the equation

$$D(G_i = 0, S_i = C \mid \theta) = D(G_i = 1, S_i = C \mid \theta) \quad (10)$$

such that $\hat{\mu}_0 \le C \le \hat{\mu}_1$. If no such C exists, to be consistent with monotonicity of assignment, we let $C = \hat{\mu}_g$ where $D(G_i = g, S_i = C\hat{\theta})$ is the smaller of the two measures when $\hat{\mu}_0 \le S_i \le \hat{\mu}_1$. The variable C is the *cut-point* or threshold value of *S* which separates 0 and 1 calls. Therefore, by monotonicity of assignment, $S_i < C \Rightarrow \hat{G}_i = 0$ and $S_i > C \Rightarrow \hat{G}_i = 1$. If $S_i = C$, we assign the genotype randomly. In the specific example of the Mahalanobis distance, C is usually the solution to

$$\frac{(C - \mu_0^{MLE})^2}{\sigma_0^{2MLE}} = \frac{(C - \mu_1^{MLE})^2}{\sigma_1^{2MLE}} \quad (11)$$

Therefore, by convergence of the MLE, we know that

$$lim_{n \to \infty} C \to_p c^* \quad (12)$$

which will useful for the next section.

Two Genotype Example: $p_0^* \neq p_0$

We start by assigning genotypes according to their Mahalanobis distance, as done in BRLMM. Recall, we assign

subject i to genotype 0 if $S_i < C$, and to genotype 1 otherwise. Therefore, the probability, P_M^n, that a subject with genotype 1 is *m*isclassified as genotype 0 will be $P_M^n = P(S_i < C|n, \theta, G_i = 1)$, and in the limit, we know $P_M^n \to \Phi(c^* \mid \mu_1, \sigma_1^2) \equiv P_M^*$. Now, it's easy to show that P_M^* must also be the limiting probability that a subject with genotype 0 is assigned as genotype 1.

$$\begin{aligned} P(S_i > c^* \mid G_i = 0) &= 1 - \Phi(c^* \mid \mu_0, \sigma_0^2) \\ &= 1 - \Phi(\frac{c^* - \mu_0}{\sigma_0^2} \mid 0,1) = 1 - \Phi(\frac{\mu_1 - c^*}{\sigma_1^2} \mid 0,1) \\ &= \Phi(\frac{c^* - \mu_1}{\sigma_1^2} \mid 0,1) = \Phi(c^* \mid \mu_1, \sigma_1^2) \equiv P_M^* \end{aligned} \tag{13}$$

Therefore, given the genotype, the limiting conditional probabilities, $P(\hat{G}_i = 1 \cdot \hat{G}_i \neq G_i | G_i = 0)$ and $P(\hat{G}_i = 0 \cdot \hat{G}_i \neq G_i | G_i = 1)$ are equal. If $p_0 > p_1$, then the unconditional probabilities cannot be equal, specifically

$$\begin{aligned} &P(G_i = 1 \cap G_i \neq G_i) \\ &\quad = P(G_i = 1 \cap G_i \neq G_i \mid G_i = 0)p_0 \\ &\quad > P(G_i = 0 \cap G_i \neq G_i \mid G_i = 1)p_1 \\ &\quad = P(G_i = 0 \cap G_i \neq G_i) \end{aligned} \tag{14}$$

Obviously the opposite inequality holds if $p_0 < p_1$. Therefore, with the Mahalanobis distance, the bias will be

$$p_0^* - p_0 = P_M^* p_1 + (1 - P_M^*)p_0 - p_0 = P_M^*(p_1 - p_0) \tag{15}$$

Clearly, when $p_0 = p_1 = 0.5$, p_0^* for all values of P_M^*. However, when $p_0 \neq p_1$, the bias is a non-zero function P_M^*, and therefore depends on the parameter group, $\{\mu_1 - \mu_0, \sigma_1^2 / \sigma_0^2\}$ (Figure 1). As shown by Figure 1, the bias can be quite large when either P_M^* and/or $|p_0 - 0.5|$ is large.

Next, assume that genotype assignments are based on a likelihood, or weighted likelihood measure. For a value $\omega \in c(0, 1)$, the probability that $\phi_g(S_i)$ exceeds ω, or $2\Phi_g(\phi_g^{-1}(\omega)) - 1$, changes with σ_g^2, where $\phi_g^{-1}(\omega)$ returns a value greater than μ_g. Therefore, $\phi_0(s) = \phi_1(s)$ does not imply anything about the relationship between $\Phi_0(s) = \Phi_1(s)$. We illustrate the potential for bias by a simple example where $\mu_0 = 0$, $\sigma_0^2 = 1$, and $\mu_1 = 1$. Let $p_0 = p_1 = 0.5$. Then, for any value of $\sigma_1^2 \neq 1, 1 - \Phi(\phi_0^{-1}(\omega)) \neq \Phi(\phi_1^{-1}(\omega))$. Equivalently, when two normal densities intersect at the threshold value, the probability of misclassifying a genotype 0 subject will not equal the probability of misclassifying a genotype 1 subject, and therefore $lim_{n\to\infty}\hat{p}_0 \neq 0.5 = p_0$. For a given threshold, c^*, we can define the bias by $p_0^* - p_0 = p_0\Phi_0(c^*) + p_1\Phi_1(c^*) - p_0 = p_0(\Phi_0(c^*) - 1) + (1 - p_0)\Phi_1(c^*)$. Unlike the previous example, with the Mahalanobis distance, Figure 2 shows that describing the bias as a function of σ_2^2 and p_0 can be difficult. In current tests of associa-

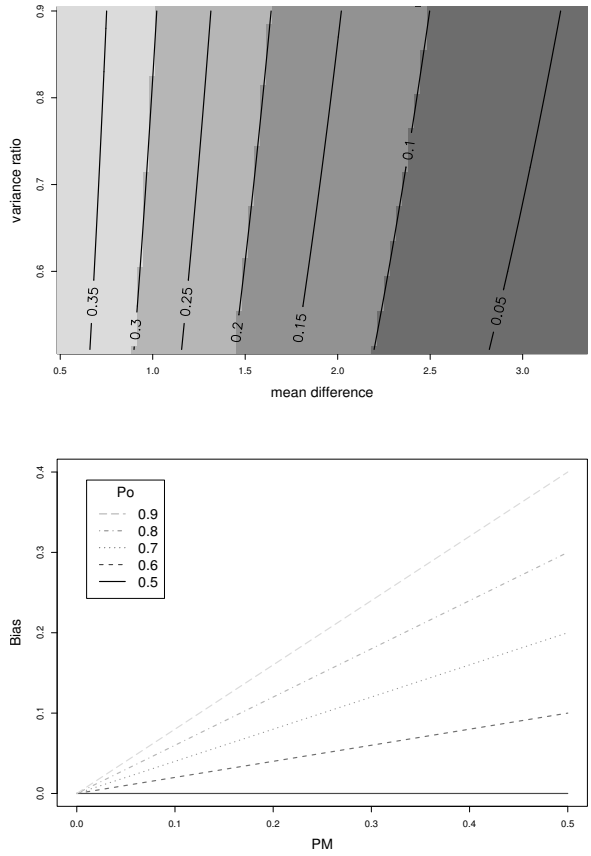

Figure 1

(P_M and Bias) a) P_M depends on the ratio, $\sigma_1^2 : \sigma_0^2$ (y-axis) and on the difference, μ_1 - μ_0(x-axis) b) The bias, p_0^* - p_0 (y-axis) depends on P_M (x-axis), and is shown for different values of p_0.

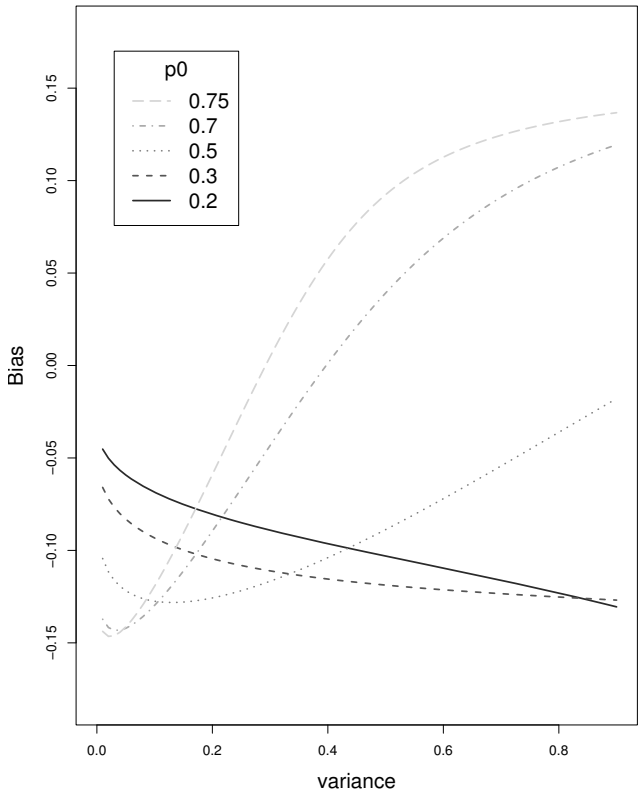

Figure 2

(P_0 and Bias) The bias of the estimate, $p_0^* = lim_{n\to\infty}\sum_{i=1}^{n} 1(G^i = 0)/n$, depends on p_0, μ_1, μ_2, σ_0^2 and σ_1^2. After fixing, $\mu_0 = 0$, $\sigma_0^2 = 1$, and $\mu_1 = 1$, we plot the bias, p_0^* - p_0 (y-axis), against σ_1^2 (x-axis), for different values of p_0.

tion, this inequality shows that it possible for $p_0^{n*A} \neq p_0^{n*U}$ even when $p_0^A = p_0^U$, which as we will see, will lead to an inflated type I error and too many low p-values. Start by noting that GWAS test a surrogate hypothesis, $H_0^* : p_0^{n*A} = p_0^{n*U}$, not the true hypothesis of interest, $H_0 : p_0^A = p_0^U$. Because p_0^{n*A} need not equal p_0^{n*U} when the distributions for cases and controls differ, H_0^*, the tested hypothesis, can be false even when H_0 is true. Let T^* be a standard test statistic for H_0^*, which is believed to have the following property, $P(T^* > t_\alpha^* \mid H_0^*) = \alpha$. Let us make the reasonable assumption that the difference between $P(T^* > t_\alpha^* \mid H_0, H_0^*)$ and $P(T^* > t_\alpha^* \mid H_0^*)$ is small, or, in words, when H_0^* is known to be true, the validity of H_0 has little effect on the distribution of T^*. Then,

$$\begin{aligned} P(T^* > t_\alpha^* \mid H_0) = \; & P(T^* > t_\alpha^* \mid H_0, H_0^*)P(H_0^* \mid H_0) \\ & + P(T^* > t_\alpha^* \mid H_0, H_1^*)P(H_1^* \mid H_0) \\ & \approx \alpha P(H_0^* \mid H_0) + \beta(1 - P(H_0^* \mid H_0)) \\ & > \alpha \end{aligned} \tag{16}$$

Here β is a measure of the power to reject H_0^* and we assume $\beta > \alpha$. Therefore, the current method of rejecting H_0 whenever $T^* > t_\alpha^*$ is actually anti-conservative if the stated p-value is α

Two Genotype Example: Inconsistency

As with any GWAS experiment, we can estimate p_0 and p_1 by

$$\hat{p}_0^n \equiv \frac{1}{n}\sum_i 1(\hat{G}_i = 0) \text{ and } \hat{p}_1^n \equiv \frac{1}{n}\sum_i 1(\hat{G}_i = 1) \tag{17}$$

For presentation, we will omit the superscript n, writing $\hat{p}_0^n$ and $\hat{p}_1^n$ as $\hat{p}_0$ and $\hat{p}_1$. As $\hat{G}_i$, from equation 5, is equivalent to $E\,[P(S_i < C|n, \theta)]$, we know that

$$lim_{n\to\infty} p_0^{n^*} \to p_0^* \text{ and } lim_{n\to\infty} p_1^{n^*} \to p_1^* \tag{18}$$

Therefore, by the convergence of $p_0^{n^*}$ and $p_1^{n^*}$ to constants and the convergence of the MLE, we have

$$lim_{n\to\infty} \hat{p}_0 \to_p p_0^* \text{ and } lim_{n\to\infty} \hat{p}_1 \to_p p_1^* \tag{19}$$

Having just discussed cases where equation (9) holds, our standard estimates of p_0 and p_1 are not consistent. Specifically,

$$lim_{n\to\infty} \hat{p}_0 \not\to_p p_0 \text{ and } lim_{n\to\infty} \hat{p}_1 \not\to_p p_1 \tag{20}$$

for these scenarios.

Return of Consistency: Modifying the Mahalanobis Distance

The Mahalanobis Distance, $D(g, s_i|\theta)$, measures the conditional probability of getting a value as extreme as s_i given genotype g. Therefore, we could achieve the same results using

$$D^{\dagger}(g, s_i \mid \theta) \equiv P(\frac{(S_i-\mu_g)^2}{\sigma_g^2} > \frac{(S_i-\mu_g)^2}{\sigma_g^2} \mid \mu_g, \sigma_g^2) \tag{21}$$

where $g \in \{0, 1\}$ and S_i is again presumed to be normally distributed. As we saw, the current estimators suffer because they don't account for the genotype probabilities. Borrowing Bayesian terminology, we simply need to weight our distance measure by the prior probability of a subject having each genotype. Therefore, returning to step 4 of our genotyping process, we now define $\hat{G}_i = g_{max}$ where g_{max} maximizes the function $p_g D_g^{\dagger}(s)$, a weighted version of the Mahalanobis distance. Let $c^{\dagger}$ be the point such that $p_0 D_0^{\dagger}(c^{\dagger}) = p_1 D_1^{\dagger}(c^{\dagger})$. Then we are guaranteed consistency as

$$\begin{aligned} p_0^{*} &= p_0 P(S_i < c \mid G_i = 0) + p_1 P(S_i < c \mid G_i = 1) \\ &= p_0(1 - 0.5 D_0(c\ )) + 0.5 p_1 D_1^{\dagger}(c^{\dagger}) \\ &= p_0 \end{aligned} \tag{22}$$

With this change, our estimate $\hat{p}_0 \equiv \frac{1}{n}\sum_i 1(\hat{G}_i = 0)$ would now be consistent.

Dependence/Correlation of Called Genotypes

If we knew $\{G_1,..., G_n\}$, the true genotypes for a group of n subjects at SNP j, as opposed to only knowing their called genotypes, $\{G_1,..., G_n\}$, it would easy to construct a test of H_0: $H_0 : p_0^A = p_0^U$ with a specified α level. Reject H_0 if $T(\tilde{p}_0^A, \tilde{p}_0^u) > t_\alpha$. Here, $\tilde{p}_g^q \equiv \frac{1}{n_q}\sum_{i:Q_i=q} G_i$, t_α is the 1-α percentile of a χ_2^2 distribution, and

$$T(\tilde{p}_0^A, \tilde{p}_0^U \equiv (\sqrt{n}\,\frac{\tilde{p}_0^A}{\sqrt{\tilde{p}_0^A - (\tilde{p}_0^A)^2}} - \sqrt{n}\,\frac{\tilde{p}_0^U}{\sqrt{\tilde{p}_0^U - (\tilde{p}_0^U)^2}})^2 \tag{23}$$

The central limit theorem allows us to be confident that we have an α-level test because $\sqrt{n}(\tilde{p}_g^q - p_g^q) \approx N(0, p_g^q(1-p_g^q))$ when $\{G_1,..., G_n$ is a vector of independent Bernoulli random variables. Again, we have returned to the two genotype scenario to simplify our discussion.

By statements about the perceived α-levels of $T(p_0^A, p_0^U)$, $\{G_1,..., G_n\}$ are often implicitly treated as the true genotypes and are assumed to be a vector of independent Bernoulli random variables. The truth, however, is that $\hat{G}_{i_1}$ is not independent of $\hat{G}_{i_2}$. Specifically if C, the threshold for calls, is relatively small, then both $P(\hat{G}_{i_1} = 1|C)$ and $P(\hat{G}_{i_2} = 1|C)$ are relatively large. Using this common dependence on C, it is simple to show that $\hat{G}_{i_1}$ and $\hat{G}_{i_2}$ are positively correlated.

$$\begin{aligned} &p(G_{i_1} = G_{i_2}) \\ &\quad = \int_{C=-\infty}^{\infty} P(G_{i_1 j} = G_{i_2 j} \mid C) f(C) dC \\ &\quad = \int_{C=-\infty}^{\infty} [(p_0^{n^*}(C))^2 + (p_1^{n^*}(C))^2] f(C) dC \\ &\quad = \int_{C=-\infty}^{\infty} (p_0^{n^*}(C))^2 f(C) dC \\ &\quad + \int_{C=-\infty}^{\infty} (p_1^{n^*}(C))^2 f(C) dC \\ &\quad \geq (\int_{C=-\infty}^{\infty} (p_0^{n^*}(C)) f(C) dC)^2 \\ &\quad + (\int_{C=-\infty}^{\infty} (p_1^{n^*}(C)) f(C) dC)^2 \\ &\quad = (p_0^{n^*})^2 + (p_1^{n^*})^2 \end{aligned} \tag{24}$$

This proof, which uses Jensen's inequality, clearly shows that the two variables, $\hat{G}_{i_1}$ and $\hat{G}_{i_2}$, are not independent as that would have implied

$P(G_{i_1} = G_{i_2}) = P(G_{i_1} = 0, G_{i_2} = 0) + P(G_{i_1} = 1, G_{i_2} = 1) = (p_0^{n^*})^2 + (p_1^{n^*})^2$. The consequence of this dependence is that $\sqrt{n}(p_g^q - p_g^q)$ does not follow a $N(0, p_g^q(1-p_g^q))$ distribution, and in turn, that $T(p_0^A, p_0^U)$ neither follows a χ_2^2 distribution nor has $P(T(p_0^A, p_0^U) > t_\alpha) = \alpha$.

We next examine the behavior of $\hat{p}_0^A, \hat{p}_0^U$, and $T(p_0^A, p_0^U)$. First, as is common, the dependence increases the variance of these estimates. For any population,

$$var(\sqrt{n}\,\frac{\sum_{i=1}^{n} 1(G_i=0)}{n}) = nE[var(\frac{\sum_{i=1}^{n} 1(G_i=0)}{n} \mid C)] + nvar(E[\frac{\sum_{i=1}^{n} 1(G_i=0)}{n} \mid C]) \quad (25)$$

The first term, $nE[var(\frac{\sum_{i=1}^{n} 1(G_i=0)}{n} \mid C)]$, represents the uncertainty in the true number of subjects with genotype 0, and can be roughly approximated by $\hat{p}_0 - (\hat{p}_0)^2$. The second term, $nvar(E[\frac{\sum_{i=1}^{n} 1(G_i=0)}{n} \mid C])$ reflects that there will be a subpopulation, of a random size $\sim n|\Phi^M(C) - \Phi^M(E[C])|$, that is assigned the 'non-ideal' genotype, where a call is labeled 'non-ideal' if it would have been different had the threshold been $E[C]$. We can approximate this second term, the overall increase in the variance, by $nvar(\Phi^M(C))$ if $P(\hat{G}_i = 0|C) \approx \Phi^M(C)$ and the $cor(\hat{G}_{i_1}, \hat{G}_{i_2} \mid C) \approx 0$. In experiments, we can estimate $nvar(\Phi^M(C))$ by bootstrapping samples of C or $\hat{\Phi}^M(C)$.

To focus on the distributions instead of just the variances, we decompose $\sqrt{n}(p_0 - p_0^{n^*})$ as

$$\begin{aligned}&\sqrt{n}(p_0 - p_0^{n^*}) = \\ &\sqrt{n}(\Phi^M(C) - E[\Phi^M(C)]) = \\ &\sqrt{n}(\Phi^M(C) - \Phi^M(C)) + \Phi^M(C) - \Phi^M(E[C])) + \\ &(\Phi^M(E[C]) - E[\Phi^M(C)]))\end{aligned} \quad (26)$$

Note that $\hat{p}_0 \equiv \hat{\Phi}^M(C)$ and $p_0^{n^*} \equiv E[\Phi^M(C)]$. The appropriate multiple of the first term, $n(\hat{\Phi}^M(C) - \Phi^M(C))$, should be well approximated by X-E [C], where $X \sim$binomial(n, $E[C]$). From our own experience, we have seen that the third term, $(\Phi^M(E[C]) - E\hat{\Phi}^M(C)])$, is a constant close to 0. The second term, $(\Phi^M(C) - \Phi^M(E[C]))$ is the variable which causes deviation from normality. Again, we could approximate the distribution of this term by bootstrapping C.

Next, we offer an example to demonstrate the effect of dependence among called genotypes. Specifically, using the IBD data, we show that a $N(0, E[p_{AB}](1 - E[p_{AB}]))$ is a poor approximation for the distribution of $\sqrt{n}(p_{AB} - E[p_{AB}])$. Because we will use real data, we have purposely chosen to discuss $\sqrt{n}(p_{AB} - E[p_{AB}])/\sqrt{E[p_{AB}] - (E[p_{AB}])^2}$ instead of T. There are many reasons that T may not follow a χ_2^2 distribution, including p_g^* being a poor estimate of p_g, the distributions of S_i being far from normal and population substructure. However, for large n, the only reason that $\sqrt{n}(p_{AB} - E[p_{AB}])/\sqrt{E[p_{AB}] - (E[p_{AB}])^2}$ will not be approximately normal is if $\{G_1, ..., G_n\}$ are not independent. Also, we chose to focus on the AB genotype as this is certain to be one of the genotypes with an overlapping cluster.

For each of our 3137 SNPs in the IBD data, we bootstrap 40 samples of 500 subjects, and calculate 40 values of $\sqrt{n}(p_{AB} - E[p_{AB}])/\sqrt{E[p_{AB}] - (E[p_{AB}])^2}$, where E[$p_{AB}$] is estimated by $\frac{1}{40}\sum_{k=1}^{40}\hat{p}_{AB}$. The qq-plot in Figure 3 compares these 40 × 3137 values with a N(0,1). The distribution is far from normal, which implies that $\{G_1, ..., G_n\}$ are dependent. Some SNPs were more likely to contribute skewed values than others, but the top and bottom 100 values (200 total) are from 64 different SNPs, indicating that no one SNP, or small number of SNPs, is responsible for the deviation in the qq plot. In contrast, the qq-plot from well-behaved SNPs, where the contributions to the density of S from the three genotypes were separated, was the expected straight line, showing that it was not the normal approximation skewing the results (Figure 3). Because the magnitude of the observed values were larger than predicted by theory, the practical implication is that tests based on the statistic, T, estimated by $F_{\chi_2^2}(T)$, will be anti-conservative(i.e. too many significant p-values after adjusting for multiple testing) under the null hypothesis. Here we also note that if S_i were truly normal, the impact of the dependency would be much less. For more details on the origin of dependency, please see Table 1.

Comparing Tests of Association

Table 2 shows the results from the simulations, and lists the percentage of the influential SNPs that were ranked among the top 100 most significant SNPs, where the rank-

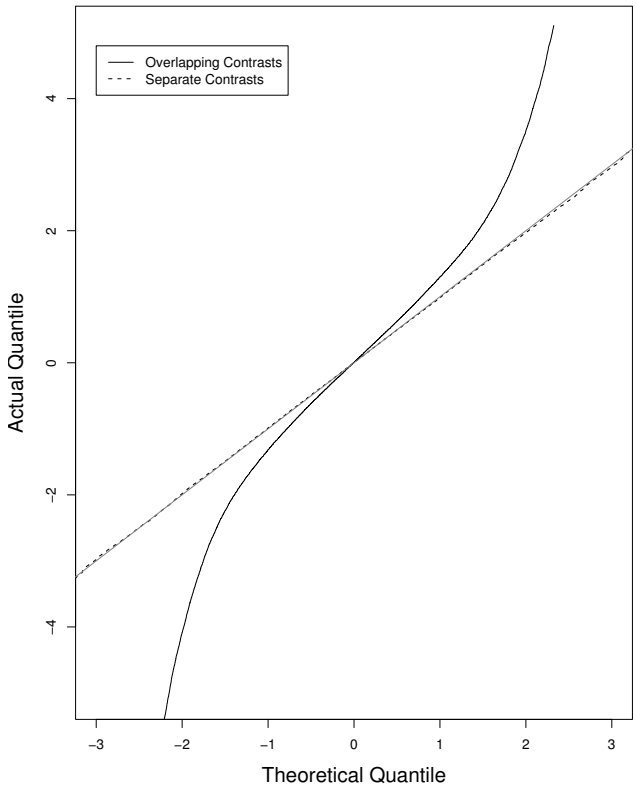

Figure 3

(Dependency of Calls) The density of $\sqrt{n}(p_{AB} - E[p_{AB}])/(Ep_{AB} - E[p_{AB}^2])$ is compared to a N(0,1) density for 500 subject samples in a quantile-quantile plot. The deviation from the Y = X line indicates that $n\hat{p}_{AB}$ is not distributed as a binomial(n, p_{AB}) variable.

ings were determined by either LogiCALL, a likelihood ratio test, or a standard test, similar to Bead-Studio. As there were 100 influential SNPs in each simulation, an ideal scenario would have 100% of the influential SNPs in the top 100 SNPs. As expected, the percentages increase as the relative risk, comparing the two homogeneous genotypes, increases from 1.5 to 2.5. So long as the densities of $\phi(M_{ji}|G_{ji} = AA)$ and $\phi(M_{ji}|G_{ji} = AB)$ were distinct, which was the case when the distance between μ_{AA} and $(\mu_{AB}^A + \mu_{AB}^U)/2$ exceeded 0.5 (and was less than 1.239), all three tests performed equally well. As the shifts were decreased in simulations, many of these shifted SNPs started to appear in the top 100 most significant SNPs, when ranked by a standard test. Furthermore, the loss in power was exaggerated when the amount of overlap differed between cases and controls, $\phi(M_{ji}|G_{ji} = AB, Q_i = A) \neq \phi(M_{ij}|G_{ij} = AB, Q_i = U)$. In the most extreme case, when $E[M_{ji}|G_{ji} = AB] = -0.539$ and $E[M_{ji}|G_{ji} = AB, Q_i = A] - E[M_{ji}|G_{ji} = AB, Q_i = U] = 0.2$, the standard test only detected about half as many influential genes as it did when there were no shifted genes. In contrast, LogiCALL almost never ranked any of the shifted SNPs in the top 100. However, had these shifted SNPs been influential, LogiCALL would have had less power to detect them. The performance of the likelihood ratio was in between the two other tests, but performed nearly as well as LogiCALL when $\phi(M_{ji}|G_{ji} = AB, Q_i = A) = \phi(M_{ij}|G_{ij} = AB, Q_i = U)$.

In GWAS, each marker is tested for association with the disease. Here, we compare four methods for testing the 3137 chosen SNPs in the IBD study. In the first method, we let Bead-Studio call the genotypes and perform its

Table 2: The percentage of influential genes among the top 100 most significant SNPs, as ranked by LogiCALL, a likelihood ratio test, and a standard test.

		RR = 1.5			RR = 2.0			RR = 2.5		
Shift	Difference	LogiCALL	LR	Standard	LogiCALL	LR	Standard	LogiCALL	LR	Standard
0.5	0	4.4	4.6	4.5	41.8	41.8	41.3	78.8	79.2	79.3
	0.2	4.4	4.6	4.5	42.0	41.8	41.3	78.8	79.3	79.4
0.3	0	4.4	4.6	4.6	42.0	41.9	41.2	78.8	79.2	79.4
	0.2	4.4	1.9	0.2	41.6	29.5	14.4	78.8	67.3	42.7
0.2	0	4.4	4.4	2.1	42.0	41.4	25.9	78.8	78.9	47.1
	0.2	3.7	0.1	0.0	38.9	3.4	0.0	75.3	21.3	0.0

Simulated data sets are full described in the *Methods Section*. The shift is the distance between the and μ_{AA} and $(\mu_{AB}^A + \mu_{AB}^U)/2$. The Difference is the distance between μ_{AB}^A and μ_{AB}^U. RR is the genotype relative risk for subjects homogeneous for the minor allele.

standard Cochran-Armitage test. Default settings were used to assign calls as missing and remove poor quality SNPs. In the second method, we call the SNPs using logiCALL and test for association using logistic regression, although using Cochran-Armitage would not change our conclusions. In the third method, we calculate the Likelihood Ratio Statistic comparing $\hat{\theta}^j$ and $\hat{\theta}_R^j$ from equation (3) using all SNPs, and in the fourth method, we calculate the LR statistic using only those SNPs where $\hat{\theta}^j$ and $\hat{\theta}_R^j$ yield identical calls, and Hardy-Weinberg equilibrium, for controls alone, is not violated at a statistical significance level of $< 10^{-16}$. For each method, we calculated the proportion of 'p-values' that were less than 0.005, 0.001, and 0.0005 (Table 3).

When genotyping all SNPs and all subjects, the proportion of Bead-Studio 'p-values' below the three thresholds far exceeded 0.005, 0.001, and 0.0005. Even after removing low quality SNPs and allowing missing calls, the proportion of 'p-values' below the three thresholds were 0.015, 0.005, and 0.004. In contrast, LogiCALL eliminated nearly all false positives. The proportion below the three thresholds were 0.004, 0.002, and 0.002. If the majority of these SNPs are presumed to be null associations, logiCALL appears to be the superior method, so long as the power loss is minimal. Assigning conservative 'p-values' to problematic SNPs is nearly equivalent to removing them. However, because of the tendency for there to be multiple, nearly equivalent, maximum likelihood estimates, the relative distances to the AA, AB, and BB genotypes using the single set of maximum likelihood estimates may not be adequate in identifying questionable calls. Therefore, logiCALL gains an advantage by combining two methods for identifying suspect calls. Additionally, it avoids false positives caused by differential bias, where the proportion of missing calls differs between cases and controls. This new method simplifies testing by requiring no preprocessing and testing all SNPs.

Table 3: The percentage of 'p-values' less than traditional α-levels (0.005,0.001,0.005) are listed for four tests of association.

Method	n	p < 0.005	p < 0.001	p < 0.0005
BeadStudio	3487	0.015	0.005	0.004
logiCALL	5533	0.004	0.002	0.002
LR	5533	0.087	0.061	0.052
LR-(same)	3014	0.006	0.002	0.001

1) BeadStudio (default setting for missing assignements and omitting SNPs).
2) logiCALL.
3) Likelihood Ratio (LR) using all SNP.
4) Likelihood ratio (LR-same) using only SNP where calls were the same for restricted and unrestricted parameter sets.

The power loss from logiCALL depends on the quality of the data. When the statistic for a SNP cleanly separates into an AA, AB, and BB group, there is no power loss. In our IBD example, the 'p-values' reported for rs2066843 and rs2076756, the two SNPs that are believed to be truly associated with IBD were similar for the two methods, Bead-Studio (2.9×10^{-9} and 5.1×10^{-10}) and logiCALL (1.5×10^{-8} and 1.6×10^{-9}). Among those subjects meeting the 96% Bead-Studio call rate, logiCALL found no questionable calls.

The likelihood ratio method had mixed results. Clearly, when the distribution of the summary statistic is a mixture of normals, the estimated genotype proportions are asymptotically unbiased. Unfortunately, this method still resulted in an increased number of false positives. However, if we removed those SNPs where at least one call changed when switching from $\hat{\theta}$ to $\hat{\theta}_R$, the false positive rate decreased to the expected level. In theory, all calls should be identical and only the resulting likelihoods should differ. When assigning genotypes, the cost, in likelihood, incurred from forcing the vectors of genotype proportions to be equal should be far less than the cost of switching calls. When the reverse is true, and calls switch, the statistic for the three groups cannot be well separated, and the p-value is suspect.

Conclusion

In genome-wide association tests, under the null hypothesis, the test statistic rarely follows the expected chi-squared distribution. This deviation tends to result in an excess of false positives. Unfortunately, the investigation into the origin of this deviation has yet to be completed. The problems associated with poor signal quality and population substructure have been thoroughly explored. However, the overlap of fluorescent signals has only been identified as a serious problem, and has yet to be fully explained. In this paper, we have provided two reasons, parameter inconsistency and called genotype dependency, that help explain how overlap causes this deviant behavior. Furthermore, we propose two methods, logiCALL and a method based on the likelihood ratio statistic that better handle the problems of inconsistency and dependency. These methods will perform similarly to the common, genotype-based, test statistics for the well-behaved SNPs and appear to create fewer false positives for the difficult-to-call SNPs. We have also identified a new characteristic of some false positives, that the call differs when using $\hat{\theta}_R^j$ vs. $\hat{\theta}^j$, that will help distinguish

which low p-values represent significant disease/marker association.

We have demonstrated that increasing sample size alone will not eliminate type I error, as genotyping, in its current form, leads to inconsistent estimates of population parameters. To alleviate this inconsistency, the distance measure used for assignment would need to be switched to $p_g d_g^{\dagger}(s)$, defined in the *Results and Discussion Section*. Moreover, we have proven that the called genotypes can be dependent under certain conditions, and that tests based on called genotypes need to account for the increased variance caused by dependence. Finally, we illustrated that the likelihood profile of the data can be relatively flat near the MLEs. Therefore, judging the quality of calls from distances based only on the MLE may not provide an adequate means to identify questionable calls. Hence logiCALL, which looks at all locally-maximal likelihood estimates of the parameters can reduce the type I error rate. Testing association by the likelihood ratio statistic is another promising method for addressing the problems associated with overlapping signals.

Availability and requirements

Computer programs are available on author's website, http://bioinformatics.med.yale.edu/group/josh/index.html.

Authors' contributions

JS identified the original problem. Both JS and HZ developed the problem. JS drafted the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Judy Cho for generously sharing her IBD data. This project was supported by Dr. Zhao's NIH grant GM 50507 and Dr. Sampson's Ruth L Kirschstein post-doctoral fellowship.

References

1. Plagnol V, Cooper J, Todd J, Clayton D: **A method to Address Differential Bias in Genotyping in Large-Scale Association Studies.** *PLoS Genet* 2007, **3(5):**759-767.
2. Clayton D, Walker N, Smyth D, Pask R, Cooper J, Maier L, Smink L, Lam A, Ovington N, Stevens H, Nutlad S, Howson J, Faham M, Moorhead M, Jones H, Falkowski M, Hardenbol P, Willis T, Todd J: **Population Structure, Differential Bias, and Genomic Control in a Large-Scale, Case-Control Association Study.** *Nature Genetics* 2005, **37:**1243-1246.
3. Moorhead M, Hardenbol P, Siddiqui F, Falkowski M, Bruckner C, Ireland J, Jones H, Jain M, Willis T, Faham M: **Optimal Genotype Determination in Highly Multiplexed SNP Data.** *Eur J Hum Genet* 2006, **14:**207-15.
4. Miller M, Schwander K, Rao D: **Genotyping Errors and Their Impact on Genetic Analysis.** *Advances in Genetics* 2008, **60:**141-152.
5. Sobel E, Papp J, K L: **Detection and Integration of Genotyping Errors in Statistical Genetics.** *American Journal of Human Genetics* 2002, **70:**496-508.
6. Rice K, Holmans P: **Allowing for Genotyping Error in Analysis of Unmatched Case-Control Studies.** *Annals of Human Genetics* 2003, **67:**165-174.
7. Rabbee N, Speed TP: **A genotype calling algorithm for affymetrix SNP arrays.** *Bioinformatics* 2006, **22:**7-12.
8. Affymetrix: **BRLMM: an improved Genotype Calling Method for the GeneChip Human Mapping 500 K Array Set.** [http:www.affymetrix.com/support/technical/whitepaperbrlmwhitepaper.pdf].
9. Carvalho B, Bengtsson H, Speed TP, Irizarry RA: **Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.** *Biostatistics* 2007, **8(2):**485-499.
10. Consortium TWTCC: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
11. Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA: **SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays.** *Bioinformatics* 2007, **23:**57-63.
12. Xiao Y, Segal MR, Yang Y, Yeh RF: **A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays.** *Bioinformatics* 2007, **23(12):**1459-1467.
13. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, Clark TG: **A genotype calling algorithm for the Illumina BeadArray platform.** *Bioinformatics* 2007, **23(20):**2741-2746.
14. Dunning MJ, Smith ML, Ritchie ME, Tavare S: **beadarray: R classes and methods for Illumina bead-based data.** *Bioinformatics* 2007, **23(16):**2183-2184.
15. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen Mm, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S: **Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays.** *Bioinformatics* 2005, **21(9):**1958-1963.
16. Murphy EA, Bolling DR: **Testing of single locus hypotheses where there is incomplete separation of the phenotypes.** *American Journal of Human Genetics* 1967, **19:**322-334.
17. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH: **A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene.** *Science* 2006, **314(5804):**1461-1463.
18. Rioux JD, Xavier R, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffths A, Targan S, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, Rotter J, Duerr RH, Cho JH, Daly MJ, Brant SR: **Genome-Wide Association Study Identifies New Susceptibility Loci for Crohn Disease and Implicates Autophagy in Disease Pathogenesis.** *Nature Genetics* 2007, **39:**596-604.