1-29-2009

ASSOCIATON TESTS THAT ACCOMMODATE GENOTYPING ERRORS

Ingo Ruczinski
Johns Hopkins University, Bloombeg School of Public Health, Department of Biostatistics, ingo@jhu.edu

Qing Li
Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics

Benilton Carvalho
Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics

M. Daniele Fallin
Johns Hopkins University, Bloomberg School of Public Health, Department of Epidemiology

Rafael A. Irizarry
Johns Hopkins University, Bloomberg School of Public Health, Department of Biostatistics

See next page for additional authors

Authors
Ingo Ruczinski, Qing Li, Benilton Carvalho, M. Daniele Fallin, Rafael A. Irizarry, and Thomas A. Louis

Association Tests that Accommodate Genotyping Errors

Ingo Ruczinski[1,*], Qing Li[1], Benilton Carvalho[1], M Daniele Fallin[2], Rafael A Irizarry[1], Thomas A Louis[1,*]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

[2]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

* Author to whom correspondence should be addressed (ingo@jhu.edu or tlouis@jhsph.edu)

Abstract

High-throughput SNP arrays provide estimates of genotypes for up to one million loci, often used in genome-wide association studies. While these estimates are typically very accurate, genotyping errors do occur, which can influence in particular the most extreme test statistics and p-values. Estimates for the genotype uncertainties are also available, although typically ignored. In this manuscript, we develop a framework to incorporate these genotype uncertainties in case-control studies for any genetic model. We verify that using the assumption of a "local alternative" in the score test is very reasonable for effect sizes typically seen in SNP association studies, and show that the power of the score test is simply a function of the correlation of the genotype probabilities with the true genotypes. We demonstrate that the power to detect a true association can be substantially increased for difficult to call genotypes, resulting in improved inference in association studies.

Keywords:

Association studies, genotypes, genotype uncertainty, score tests, single nucleotide polymorphisms.

1 Introduction

The technological advancements of high-throughput arrays have revolutionized many aspects in the fields of statistical genetics and genomics. Some of the latest platforms, such as the Affymetrix Genome-Wide Human SNP Array 6.0, allow for the simultaneous assessment of almost one million genotypes at polymorphic loci (single nucleotide polymorphisms, SNPs) in human DNA. The estimates of these genotypes, however, are subject to error. The frequency of these genotyping errors can depend on various factors, such as the sample quality and the chemistry of the SNP probe. It has long been recognized that statistical algorithms play a crucial role for the accuracy and precision of the genotype calls. In particular for the Affymetrix platforms, originally described as a high-throughput assay for calling genotypes at about ten thousand SNPs (Kennedy et al. (2003)), an extensive list of proposed algorithms has been published (DM, Di et al. (2005); RLMM, Rabbee and Speed (2006); BRLMM, Affymetrix (2006); CRLMM, Carvalho et al. (2007); SNiPer-HD, Hua et al. (2007); BirdSeed, Korn et al. (2008)). Similar to gene expression technologies, pre-processing of probe-level data is a crucial consideration to account for biases induced for example by fragment-length and sequence effects, possibly introduced by the polymerase chain reaction (Carvalho et al. (2007)). Evaluating and comparing the overall performances of the genotyping algorithms as well as the assessment of remaining batch and plate effects in those algorithms is a very active research area (Lin et al. (2008); Hong et al. (2008); Nishida et al. (2008)).

Genotyping errors can lead to a variety of undesirable consequences. These include incorrect estimates of parameters such as genetic distance (Goldstein et al. (1997)) and linkage disequilibrium between loci (Akey et al. (2001)), and possibly even more worrisome, incorrect inference in linkage and association studies (Buetow (1991); Abecasis et al. (2001); Kang et al. (2004d); Hao and Cawley (2007)), as well as a general loss of power in hypothesis tests (Gordon et al. (1999, 2002, 2003); Rice and Holmans (2003); Kang et al. (2004c,b)). Acknowledging the existence of such genotyping errors and accounting for the genotype uncertainties accordingly can be essential, and has been proposed for many settings. This includes haplotype estimation (Kang et al. (2004a); Zhu et al. (2007)) and linkage studies (Gordon et al. (2001, 2004)), but most prominently, this pertains to SNP association studies. Gordon and Ott (2001) propose a "reduced penetrance model method" that allows for errors in a case-control design, deriving a contingency table based test statistic. Rice and Holmans (2003) propose association tests for

settings when the overall genotyping error rate is known, for example from an external or pilot study, and investigate the effects of the genotyping error rate on the type I error rate, the power, and the bias in the odds ratio estimates. Hao and Wang (2004) employ weighted contingency table and likelihood ratio tests to incorporate genotyping error rates, finding that the weighted contingency table test improves the power to detect true associations more than the likelihood ratio test, while the latter is more useful when the decrease in bias in the parameter estimates is of interest.

Population based case-control designs, considered in the above approaches, are typically more powerful in detecting associations compared to family-based designs in the absence of population stratification. If population stratification is a strong concern however, these family-based designs often do become appealing. Genotyping errors must be specifically addressed in family based designs though, since these can result in Mendelian inconsistencies in the data. Morris and Kaplan (2004) propose likelihood ratio tests, applicable when testing for associations using case-parent designs, which incorporate additional model parameters that account for genotyping errors. The EM algorithm is employed in the calculation of the test statistic, and thus, no prior specification of the mechanism underlying the genotyping errors is required. As a consequence, the proposed test is applicable for case-parent based association tests even in the presence of Mendelian inconsistencies. Cheng and Lin (2007) offer an alternative approach, addressing population stratification and genotyping errors simultaneously.

Instead of considering the overall error rate in association tests, many genotype calling algorithms such as BRLMM (Affymetrix (2006)), CRLMM (Carvalho et al. (2007)) and BirdSeed (Korn et al. (2008)) also generate a measure of genotype uncertainty together with the estimate of the genotype, or actual genotype probabilities, which allow for association tests taking individual, SNP-specific uncertainties into account. By default, several of these algorithms do not produce a genotype call if the uncertainty is larger than a pre-set threshold. Thus, the issue of genotype uncertainty can also be seen as a missing data problem: a more stringent threshold on the genotype quality produces less genotyping errors but more missing data, and vice versa. For example, Hao and Cawley (2007) investigate power and odds ratios in population based case-control and family based association tests from a missing data perspective, allowing for potentially different rates of missing data among homozygote and heterozygote genotypes. The latter is particularly important, as the data typically are not missing at random. Heterozygote genotypes, on average, are harder to call than homozygote genotypes (see Table 1 in Affymetrix (2006)

for a particularly clear example). However, if the genotype estimates and their respective uncertainties are quantified correctly, then they contain all the information available, and from a statistical perspective, should be taken into account accordingly in the tests of association.

In particular, score tests represent a correct and efficient framework to approach such data. Plagnol et al. (2007) propose score tests that incorporate "fuzzy" genotype calls, based on an additive model under a logistic link. The main motivation in this approach was to address and properly account for potential differences in genotype estimates due to differential data (DNA) sources, avoiding unnecessarily high rates of false positives. The developed methods however are also perfectly applicable to "fuzzy" genotype calls stemming from genotype calling algorithms such as CRLMM. An almost identical approach is suggested by Marchini et al. (2007), addressing in particular genotype uncertainty for unobserved, imputed SNPs (i. e. SNPs not represented by probes on the genotyping platform).

In this manuscript, we develop a framework for score tests based on arbitrary genetic models. We verify the validity of the test, illustrate that the assumption of a "local alternative" is very reasonable for effect sizes typically seen in SNP association studies, and show that the power of the score test is simply a function of the correlation of the "fuzzy" genotype call with the true genotype. We demonstrate that the power to detect a true association can be substantially increased for difficult to call genotypes.

2 Methods and Data

We derive the test statistic and its distribution under the null and alternative hypothesis for an arbitrary genetic model. Let Y_i be the binary disease status indicator for individual $i \in \{1, \ldots, n\}$, and denote the outcome vector for all n subjects by $\mathbf{Y} = (Y_1, \ldots, Y_n)$. We assume that all SNPs are bi-allelic, and code the SNPs as the number of variant alleles. Let g_i be the actual genotype for subject i, with $g_i \in \{0, 1, 2\}$. We denote the vector of genotypes as $\mathbf{g} = (g_1, \ldots, g_n)$, and let G_i be the random variable taking on values $g_i = 0, 1,$ or 2. To make the analogy with a trend test for dose effects, we let d_g be the score assigned to genotype g. Without loss of generality, we set $d_0 = 0$, and thus denote $d = (d_1, d_2)$. This encodes the genetic model assumed in the association study, and includes the most commonly employed genetic models, for example the dominant model $d = (1, 1)$, the recessive model

$d = (0, 1)$, and the additive model $d = (1, 2)$. We derive the methods in this general framework, and illustrate our findings using an additive model in the association test, also referred to as trend test.

Let π_g be the probability of disease given genotype g, e. g., $\pi_g = Pr(Y = 1 \mid g) = H(\mu + \theta d_g)$. The function H can for example encode a logistic model. We denote the derivative of H with respect to its argument as h, e. g. $H' = h$. To quantify the effects of using different types of genotype information available (for example, genotype probabilities versus genotype calls), we use the terms *true* and *working* genotype probabilities. Let t_{ij} be the true genotype probabilities for subject i, e. g., $Pr(g_i = j)$ with $j \in \{1, 2, 3\}$. Obviously $0 \leq t_{ij} \leq 1$, and $t_{i0} = 1 - t_{i1} - t_{i2}$, and thus, we summarize the information in the true genotype probabilities as $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2)$, where $\mathbf{t}_j = (t_{1j}, \ldots, t_{nj})$. Note that we also set $d_0 = 0$. We further denote a genotype "average" as $\bar{t}_j = \sum_i t_{ij}/n = t_{+j}/n$. We use the exact same terminology for the working genotype probabilities $w_{ij} = Pr(g_i = j)$, namely $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)$, with $\mathbf{w}_j = (w_{1j}, \ldots, w_{nj})$, and $\bar{w}_j = \sum_i w_{ij}/n = w_{+j}/n$.

We want to test the hypothesis of no association between genotype and disease, i. e. $H_0 \colon \pi_0 = \pi_1 = \pi_2 = \pi$ (or equivalently, $\theta = 0$) using a score test that accounts for not knowing $\mathbf{g}$. The test is constructed using the $\mathbf{w}$ as the working probabilities, and we then evaluate the sample properties of the test under the true probabilities $\mathbf{t}$. In general, when dealing with incomplete data, the score test or any other likelihood-based inference needs to use the conditional probabilities $Pr(G_i = g_i \mid Y_i, \mathbf{w}, \mu, \theta)$. However, under H_0 the Y_i are un-informative, and hence we only need to use the $\mathbf{w}$ or $\mathbf{t}$ with no conditioning on the outcome variable.

It is straightforward to show that under the null hypothesis of no genotype/phenotype association, the genotype predictive probabilities in the score test produced by the incomplete data likelihood do not depend on the phenotype. Therefore the trend test statistic based on the $\mathbf{w}$ is a one degree of freedom test with statistic:

$$Z(\mathbf{w}, d) \quad = \quad \frac{\sum_i (w_{i1} d_1 + w_{i2} d_2)(Y_i - \hat{\pi})}{\sqrt{n \times \mathrm{Var}(\mathbf{w_1} d_1 + \mathbf{w_2} d_2) \times \hat{\pi}(1 - \hat{\pi})}} \tag{1}$$

with $\hat{\pi} = \bar{Y} = \frac{1}{n} \sum_i Y_i$ and $\mathrm{Var}(\mathbf{w_1} d_1 + \mathbf{w_2} d_2) = d_1^2 \times \mathrm{Var}(\mathbf{w}_1) + d_2^2 \times \mathrm{Var}(\mathbf{w}_2) + 2 d_1 d_2 \times \mathrm{Cov}(\mathbf{w}_1, \mathbf{w}_2)$.

In a score test, the variance term measures the statistical information, and larger is better. Further

analysis of the variance term above shows the loss of information from not knowing the g_i:

$$\begin{aligned} & n \times \mathsf{Var}(\mathbf{w_1}d_1 + \mathbf{w_2}d_2) \\ = & \sum_{j=1}^{2} d_j^2 \left\{ \bar{w}_j(1-\bar{w}_j) - \sum_i w_{ij}(1-w_{ij}) \right\} + n \times 2d_1d_2 \times \mathsf{Cov}(\mathbf{w}_1, \mathbf{w}_2) \\ = & \sum_{j=1}^{2} d_j^2 \bar{w}_j(1-\bar{w}_j) - n2d_1d_2\bar{w}_1\bar{w}_2 - \left[\sum_{j=1}^{2} d_j^2 \sum_i w_{ij}(1-w_{ij}) - 2d_1d_2 \sum_i w_{i1}w_{i2} \right]. \end{aligned} \quad (2)$$

The first term in equation (2) is the variance when the g_i are known, e.g., $w_{ij} = 0$ or 1, $w_{i1}w_{i2} = 0$ and $Pr(G = g) = \bar{w}_g$. The term in the square brackets is non-negative and accounts for the loss of information associated with not knowing the g_i. This decomposition is a standard *Variance = E(conditional variance) + Var(conditional expectation)*, where the conditioning is on i. Note that if $\mathbf{w} \equiv \bar{w}$ the variance is 0, and there is no information. If $w_i \in \{0, 1\}$, there is no penalty. However, also note that $w_i \in \{0, 1\}$ does not imply validity in general. In particular, if there are doubts about the validity of the working genotype probabilities in general (i. e. , if methods are employed that generate poor $\mathbf{w}$'s), it may be better to use the actual genotype calls than relying on the fine details of the $\mathbf{w}$'s.

Under H_0, $Z(\mathbf{w}, d) \sim N(0, 1)$, the null distribution is correct irrespective of the working probabilities $\mathbf{w}$. To calculate the distribution of the test statistic under the alternative, note that when the $\mathbf{t}$ are the true calling probabilities (they could be the actual g_i) with the test statistic computed under the working probabilities $\mathbf{w}$ and for general π_g, the expectation of the the numerator in equation (1) is

$$\begin{aligned} & E\left[\sum_i (w_{i1}d_1 + w_{i2}d_2)(Y_i - \hat{\pi}) \right] \\ = & \; n(\pi_1 - \pi_0)[d_1 \times \mathsf{Cov}(\mathbf{w}_1, \mathbf{t}_1) + d_2 \times \mathsf{Cov}(\mathbf{w}_2, \mathbf{t}_1)] + \\ & \; n(\pi_2 - \pi_0)[d_1 \times \mathsf{Cov}(\mathbf{w}_1, \mathbf{t}_2) + d_2 \times \mathsf{Cov}(\mathbf{w}_2, \mathbf{t}_2)] \end{aligned} \quad (3)$$

and

$$E(\hat{\pi}) = \pi(\mathbf{t}) = (1 - \bar{t}_1 - \bar{t}_2)\pi_0 + \bar{t}_1\pi_1 + \bar{t}_2\pi_2. \quad (4)$$

In our mathematical analysis we consider only local alternatives (small θ), and so we have

$$\begin{aligned}
\pi_g - \pi_0 &\doteq \theta d_g h(\mu), \\
E\left[\sum_i (w_{i1}d_1 + w_{i2}d_2)(Y_i - \hat{\pi})\right] &\doteq n\theta h(\mu) \times [d_1^2 \times \mathsf{Cov}(\mathbf{w}_1, \mathbf{t}_1) + d_1 d_2 \times \mathsf{Cov}(\mathbf{w}_2, \mathbf{t}_1) \\
&\quad + d_1 d_2 \times \mathsf{Cov}(\mathbf{w}_1, \mathbf{t}_2) + d_2^2 \times \mathsf{Cov}(\mathbf{w}_2, \mathbf{t}_2)] \\
&= n\theta h(\mu) \times \mathsf{Cov}(\mathbf{w}_1 d_1 + \mathbf{w}_2 d_2, \mathbf{t}_1 d_1 + \mathbf{t}_2 d_2), \\
\pi(t) &\doteq H(\mu) + \theta h(\mu)[d_1 \bar{t}_1 + d_2 \bar{t}_2]. \qquad (5)
\end{aligned}$$

Combining this with the denominator of equation (1) and using $\mu = H^{-1}(\pi(\mathbf{t}))$, we obtain

$$\begin{aligned}
& E(Z(\mathbf{w}, d) \mid \mu, \theta, d, \mathbf{w}, \mathbf{t}) \\
= \;& \theta \times \sqrt{\frac{n}{\pi(\mathbf{t})(1 - \pi(\mathbf{t}))}} \times h(\mu) \times \mathsf{Corr}\left(\{\mathbf{w}_1 d_1 + \mathbf{w}_2 d_2\}, \{\mathbf{t}_1 d_1 + \mathbf{t}_2 d_2\}\right) \times \sqrt{\mathsf{Var}(\mathbf{t}_1 d_1 + \mathbf{t}_2 d_2)} \\
=: \;& m(\mu, \theta, d, \mathbf{w}, \mathbf{t}) \qquad (6)
\end{aligned}$$

Now, represent a small departure from H_0 by writing $\theta = \theta_0/\sqrt{n}$. From equation (5) we obtain $\pi(\mathbf{t}) = H(\mu) + O(1/\sqrt{n})$, and $\pi(\mathbf{t})$ in equation (4) can then be replaced by $\hat{\pi}$. The local, non-null variance of Z is equal to 1. The logistic model produces $h(\mu) = H(\mu)(1 - H(\mu)) = \pi(1 - \pi)$, and using $\hat{\pi} = \bar{Y}$ we obtain

$$Z \sim N(m(\mu, \theta_0, d, \mathbf{w}, \mathbf{t}), 1) \qquad (7)$$

with

$$\begin{aligned}
& m(\mu, \theta_0, d, \mathbf{w}, \mathbf{t}) \\
= \;& \theta_0 \times \sqrt{\hat{\pi}(1 - \hat{\pi})} \times \mathsf{Corr}\left(\{\mathbf{w}_1 d_1 + \mathbf{w}_2 d_2\}, \{\mathbf{t}_1 d_1 + \mathbf{t}_2 d_2\}\right) \times \sqrt{\mathsf{Var}(\mathbf{t}_1 d_1 + \mathbf{t}_2 d_2)}. \qquad (8)
\end{aligned}$$

Squaring the Z-statistic produces a one degree of freedom, chi-square statistic with non-centrality $\lambda = m^2(\mu, \theta_0, d, \mathbf{w}, \mathbf{t})/2$.

Note that if $\mathbf{w} = \mathbf{t}$, then the correlation is equal to 1 and the score test is fully efficient (within the context of available information). Note from equation (2) that the first term in the variance is the variance associated with knowing the g_i (e.g., the true genotype) and the second term is the loss of information associated with not knowing the genotype, even using valid $\mathbf{t}$. Also, if the $\mathbf{w}$ are chosen very poorly ($\mathbf{w} \neq \mathbf{t}$), it is possible that Corr $(\{\mathbf{w}_1 d_1 + \mathbf{w}_2 d_2\}, \{\mathbf{t}_1 d_1 + \mathbf{t}_2 d_2\}) = 0$, i. e. a non-null situation looks like H_0, and if they are chosen very, very poorly it is possible that Corr $(\{\mathbf{w}_1 d_1 + \mathbf{w}_2 d_2\}, \{\mathbf{t}_1 d_1 + \mathbf{t}_2 d_2\}) < 0$, converting a risk genotype to a protective one. It is hard to imagine this ever happening in practice though, except through a systematic mistake in the genotype annotation. More realistically, the working probabilities $\mathbf{w}$ will have a high, positive correlation with $\mathbf{t}$ for most SNPs.

We demonstrate the potential benefits of our proposed approach via simulation studies using genomic data from the HapMap project (http://www.hapmap.org/). Launched in 2002, the International HapMap Project periodically released data on selected SNPs in 269 individuals from four different populations: 30 parent-child trios from Ibadan, Nigeria (YRI), 30 trios of U.S. residents of northern and western European ancestry (CEU), 44 unrelated individuals from Tokyo, Japan (JPT) and 45 unrelated Han Chinese individuals from Beijing, China (CHB). The genotype information used for the simulations in this manuscripts were selected from these 269 HapMap phase 2 individuals (International HapMap Consortium (2007)), specifically, the HapMap Public Release #22 mapped to NCBI build 36. To illustrate our points a subset of all of the known SNPs is certainly sufficient; the ones used in this manuscript are the SNPs from the HIND sub-array of the Affymetrix 100K chip. We only focus on SNPs with complete genotype information for all 269 subjects, and for which at least ten of these subjects have at least one minor allele. This yields a total of 32,443 SNPs.

Since the HapMap genotypes have been verified across laboratories and platforms, they are considered to be the true genotypes, and serve as the gold standard in the following comparisons. The data derived using high-throughput arrays on the other hand are subject to experimental error and biases, and thus, the resulting genotypes can be considered as estimates only. In this manuscript, we use two different approaches for genotype estimation: the default algorithm recommended by the manufacturer for 100K Affymetrix SNP chip arrays (BRLMM, Affymetrix (2006)), and an alternative approach (CRLMM, Carvalho et al. (2007)). We emphasize that the point of the current research is not to demonstrate the superiority of one algorithm over the other, but solely to demonstrate the effect of genotype uncertainty

in SNP association studies.

As mentioned previously, the genotype estimates are derived from raw intensity patterns, and many algorithms (such as BRLMM) do not produce a genotype call if the uncertainty is beyond a certain threshold (which can be varied, up to the point were a genotype call is forced for every SNP in the sample). The CRLMM method on the other hand produces a likelihood derived from both sense and anti-sense strands for each possible genotype for each locus in a sample, and thus, "fuzzy" genotypes such as the triple of posterior genotype probabilities can be calculated. In the reported research, we use a flat prior, which is equivalent to calculating the normalized likelihoods. No statement of optimality for this prior is made or implied. Certainly, in a homogeneous population with genotypes in Hardy-Weinberg equilibrium (and possibly known minor allele frequencies) more informative priors could be used, but again, the point of the reported research is not to delineate the most powerful genotyping method.

For our comparison we use four sets of genotype calls: the BRLMM genotype calls with missing data when the uncertainty exceeded the default threshold, the BRLMM genotype calls where an estimate was forced (and thus, no missing data are permitted), the CRLMM fuzzy genotype calls, and the called CRLMM genotypes, i. e. the genotype chosen via the posterior mode. In the language of the previous paragraphs, for the comparisons of efficiency we can use the HapMap gold standards as the true genotype "probabilities" and the various estimates as working genotype "probabilities", but also use the CRLMM posterior probabilities as the true genotype probabilities and the posterior modes as the working genotype probabilities.

3 Results

The accuracy with which SNPs are genotyped depends on various factors, including the quality of the DNA used in the sample, and artifacts introduced in the amplification of the DNA. However, besides those array-wide effects, there are also SNP specific effects due to different chemical properties of the probes. This can be seen for example when the raw fluorescent intensities from the sense and anti-sense strands of the DNA are plotted for a collection of individuals (Figure 1). For the majority of SNPs the three genotype clusters separate well (such as the ones of rs1641760), however, for an appreciable

proportion of SNPs this is not the case (such as the SNPs denoted by rs1665933, rs1678775, and rs1659131). Thus, these SNPs are much more prone to genotyping error in genomic assays. Note that the exact magnitude of separation for each SNP is still subject to systematic effects such as a laboratory effect, however, this does not affect the overall qualitative nature (easy/difficult) of the SNP.

The uncertainty for a genotype call obviously depends on the separation of these clusters. Further, for a newly typed sample, the location of the new spot (derived from the fluorescent pattern) within a cluster also determines the genotype uncertainty (center of the cluster versus edge), as does the overall quality of the sample. All of these pieces of information are included in the derivation of the genotype probabilities in CRLMM (Carvalho et al. (2007)). Most genotyping algorithms do not return genotype probabilities but make a genotype call and quantify the uncertainty associated with the call. Typically, if the uncertainty exceeds a certain threshold, the genotype will be recorded as missing.

[FIGURE 1 ABOUT HERE]

If all of the available information is used and fuzzy genotype calls such as the CRLMM posterior probabilities are employed in the score test (equation (1)), all samples contribute to the test as there are no missing data. Thus, it is not too surprising that by properly taking the genotype uncertainty into account one can improve the power to detect associations, particularly for hard to call SNPs. As an illustration, we used the HapMap genotype data for SNP rs1678775 (second panel in Figure 1) from the 269 samples, and simulated an artificial case status via a logistic additive model with intercept $\mu = -1$ (corresponding to a 27% probability of disease in non-carriers), and slope parameters $\theta = 0$ (under the null) and $\theta = 0.41$ (under the alternative), corresponding to odds ratios of 1.0 and 1.5 respectively. Using 10,000 replications under each scenario, we compare the theoretical and empirical cumulative distribution functions (CDFs) for the score test statistics using the true genotypes, and the complete BRLMM calls (i. e. , no missing data) and CRLMM probabilities (Figure 2). We note that the empirical CDFs are indeed very close to the theoretical CDFs, in particular under the null. In this example, using the CRLMM genotype probabilities results in a test with more power under the alternative than using the BRLMM genotype calls, in fact, its cumulative distribution function is almost identical with the one derived from the true genotypes. We note that the oscillation in some of the curves is due to the

discrete nature of the data, and that the differences in magnitude of the oscillations in the lines for the true genotypes and the BRLMM calls are a feature of this particular sample.

[FIGURE 2 ABOUT HERE]

In general, the power to detect associations can differ substantially depending on which approach is used in the score test. The difference in power is virtually nil for easy to call genotypes such as SNP rs1641760 (Figure 3, right column). For difficult SNPs such as rs1665933, rs1678775, and rs1659131 (Figure 3, columns 1-3) the difference in power between the three approaches can be substantial (rs1665933), substantial for one particular approach (rs1678775), or relatively small (rs1659131). As we expected, the true genotype always yields the largest power, and using the fuzzy CRLMM genotypes typically yields improved power compared to the genotype calls from BRLMM.

[FIGURE 3 ABOUT HERE]

The efficiency of the score test is a function of the correlation of the true genotypes and the working genotypes (equation (8)), and thus, the observed differences can simply be explained by the respective correlations (Table 1). The SNPs in this illustration were deliberately chosen such that the fuzzy CRLMM genotypes and the BRLMM genotype calls were weakly correlated to the true genotypes (rs1665933), one was somewhat strongly correlated with the true genotypes but the other one was not (rs1678775), both are somewhat strongly correlated with the true genotypes (rs1659131), and both are perfectly or almost perfectly correlated with the true genotypes (rs1641760).

[TABLE 1 ABOUT HERE]

Assuming we have a homogeneous population and genotype data for a disease risk affecting SNP in Hardy-Weinberg equilibrium (the HapMap data are not in Hardy-Weinberg equilibrium, as the samples come from four diverse populations), then, for a fixed sample size, the power to detect the association with the response depends on the minor allele frequency of the SNP, the effect size, and the correlation of the working genotype (employed in the score test) with the true genotype. We find that, in general,

the power diminishes rapidly with decreasing correlation (Figure 4). For example, assuming a minor allele frequency of 0.5 and an odds ratio of 1.75 in a sample of 269 subjects, the power to detect the association decreases from 92% to 74% if the correlation between the working genotype and the true genotype decreases from 1.0 to 0.75.

[FIGURE 4 ABOUT HERE]

To demonstrate how this change in power can affect inferences in association studies, we carried out a simulation study. We defined a pool of easy to call SNPs and a pool of difficult to call SNPs. The difficulty of a SNP was defined as the BRLMM average confidence score across all 269 subjects. We selected 100 easy to call genotypes (BRLMM average confidence scores between 0.005 and 0.009) and 100 difficult to call genotypes (BRLMM average confidence scores between 0.17 and 0.37). This resulted in two sets of SNPs that differed substantially in the correlations of the estimated and the true genotypes (Table 2).

[TABLE 2 ABOUT HERE]

In each replication of the simulation study (separately for the easy and the difficult SNPs), we randomly picked a SNP from the respective pool to carry the signal. Using the true genotypes for the chosen SNP, we simulated case-control outcomes using a logistic additive model with intercept $\mu = -1$ (corresponding to a 27% probability of disease in non-carries). The slopes were chosen to correspond to odds ratios between 1 (the null) and 2 (strong signal), in increments of 0.1 units. Using 10,000 replications under each scenario, we calculated the score test z-statistics for all 32,443 SNPs, using the true genotypes, the fuzzy CRLMM genotype calls, the genotype calls derived as the modes from the CRLMM posterior probabilities, all BRLMM genotypes regardless of the individual confidence scores, and the BRLMM genotypes after dropping the uncertain genotypes. In each replication, we recorded the rank of the causal SNP among all 32,443 SNPs. As expected, when comparing the 5 different flavors of genotypes, we observe no differences in the ranks of the causal SNPs when the SNP is selected from the pool of easy to call SNPs (Table 3, Figure 5). In essence, the genotype calls from BRLMM and the true genotypes are virtually identical, and the CRLMM fuzzy genotype calls (i.e. the posterior probabilities) are very close to binary in favor of the true genotype.

[TABLE 3 ABOUT HERE]

[FIGURE 5 ABOUT HERE]

Matters are different, of course, for the difficult to call SNPs. Naturally, the best results are achieved when the true genotype is used. The genotype calling methods yield different results in particular for the intermediate effect sizes, where real improvements can be achieved. For small effect sizes the signal is close to the null, and the truly associated SNP rarely floats to the top, regardless of the genotyping method. When the effect size is large, the truly associated SNP almost always floats to the top, regardless of the genotyping method. In the mid range, the improved genotype calling translates into improved detection of the association (the CRLMM genotypes had a higher correlation to the true genotypes than the BRLMM genotypes, Table 2). Further improvement (of a lesser magnitude) can be achieved using the entire information available in the fuzzy CRLMM genotypes, compared to the CRLMM posterior mode genotypes. The same pattern can be seen in the actual z-statistics (Table 4). We further note that the variability in the z-statistics is close to unity, in particular for the small and intermediate effect sizes that we would expect in real association studies, lending credibility to the assumption of a local alternative in the score test. However, the variability in the z-statistics does increase with the effect size. If this is a concern in the actual analysis, the departure from unity can be addressed by using a second-order expansion of the variance in the derivation of the test statistic. The local non-null variance used here is just for convenience in the evaluation and comparison of the different approaches.

[TABLE 4 ABOUT HERE]

When comparing the correlations of the true genotypes with the CRLMM fuzzy genotype probabilities (across the 269 samples) to the correlations of the true genotypes with the BRLMM genotype calls forcing complete data for all 32,443 SNPs, the histogram of the ratios shows a clear enrichment in the right tail (Figure 6, upper panel), quantifying the efficiency gain in using CRLMM fuzzy calls versus BRLMM forced genotype calls. A ratio larger than one for any particular SNP indicates more power in the score test. A ratio less than one can arise if there is some uncertainty in the genotype estimates, but BRLMM gets all genotype calls correct. However, the clear excess of ratios larger than one shows the merit of using the CRLMM continuous genotype probabilities.

For 662 SNPs, the correlation of the true genotypes with the CRLMM fuzzy genotype probabilities exceeds the one of the true genotypes with the forced BRLMM calls by more than 5%, while the reverse is only true for 56 SNPs. It should be emphasized that this enrichment is not the sole reason to use fuzzy genotype calls: even when the above mentioned ratios are less than one, using those genotype probabilities is the correct method to quantify the uncertainty in the data, and to properly use all available information in the association tests. We further note that it is also beneficial not to force the genotype calls in CRLMM, as a histogram of the correlation with the true genotype shows an enrichment in favor of the fuzzy genotype calls (Figure 6, lower panel). While the magintude of this effect is considerably smaller in this setting relative to the one in the previous comparison, there is still a clear enrichment in the right tail. For example, for 258 SNPs, the correlation of the true genotypes with the CRLMM fuzzy genotypes exceeds the one of the true genotypes with the forced CRLMM calls by more than 1%, while the reverse is only true for 3 SNPs. In summary, using the fuzzy genotypes is the correct way for testing associations as it properly quantifies all available information, and also tends to increase the power to detect associations.

[FIGURE 6 ABOUT HERE]

4 Discussion

We demonstrate that using the available knowledge about the uncertainty in the genotype estimates can lead to improved inference in association studies. We propose an easy to implement score test that utilizes all the information available, and show that improved inference can be achieved using genotype probabilities instead of genotype calls, since the latter often contain missing data. We use data from Affymetrix SNP chips to illustrate our findings, since methods to quantify genotype uncertainty have been predominantly developed for arrays from this manufacturer. If the genotype uncertainties can be properly quantified, the score test can certainly be used for association tests using genomic data from other platforms such as Illumina arrays as well.

Scalability and memory use have to be taken into account for any approach that utilizes high throughput genomic data. In this paper, we provide a proof of principle, using 30,000+ SNPs. Some of the newer

platforms such as the Affymetrix Genome-Wide Human SNP Array 6.0 contain probes for almost one million SNPs. The size of one such CEL file that contains the raw fluorescent intensities is about 65Mb, and given that a typical association study uses data from several thousand people, this poses an enormous computational challenge. When actual genotype calls are used to test for associations instead of genotype probabilities, much more efficient memory handling can be employed, as a single genotype can be stored using only 2 bits. The most prominent example of such an efficient data handling approach is PLINK (Purcell et al. (2007)), defined as a "free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner" (http://pngu.mgh.harvard.edu/∼purcell/plink/). Employing the here introduced score test for one million SNPs would be somewhat of a challenge with regards to CPU time, and a very serious challenge with regards to memory use. However, this is not necessary, as the genotypes for the vast majority of SNPs are easy to call, yielding in essence identical inference for the different approaches considered in this manuscript. Since the genotype uncertainties are available from the genotype processing algorithms, we can readily identify the hard to call SNPs, and only run the analysis for these selected SNPs. This strategy will not change the inference for the majority of the SNPs, and the difference in strength of association assessed for some of the difficult to call SNPs might not be very substantial either (we also note that when using the score test in association studies, the same pre-cautions with regards to population stratification have to be employed, such as potential adjustments via principal components or genomic control).

However, it is important to keep in mind that (unfortunately) the pre-dominant analytic approach for genome-wide association studies is to calculate a p-value for each SNP, and report only those SNPs as associated whose p-values withstand a Bonferroni-type correction (completely ignoring type II errors in the process). Using the score test, the inference for a non-trivial number of SNPs will change, and it is quite feasible that some of the SNPs previously declared non-significant will make the list of SNPs declared significant. The reverse can also be true, though we have demonstrated that the score test tends to increase the power to detect associations for some of the difficult to call SNPs. Further, many SNPs are typically not even considered in the actual analysis: if genotype calls are made but the uncertainty is beyond a pre-defined threshold, the data are reported as missing. If a certain number of genotypes are missing for any particular SNP, the SNP is usually removed from the analysis altogether.

In some instances, SNPs are even removed before the data are seen. Some SNPs are notoriously hard to genotype, and at times, lists of well known trouble makers are assembled, to be removed a priori from the analysis. However, when the genotype uncertainty is properly quantified we still have valid information, and using the genotype probabilities is the correct way for testing associations, while taking advantage of all information available. Moreover, this approach also avoids potential biases induced by the informative missingness process.

The point of this paper is not to demonstrate the superiority of one genotyping algorithm over an other, but solely to demonstrate the effect of genotype uncertainty in SNP association studies. Comparing genotype algorithms in terms of accuracy and their capabilities to deal with artifacts such as batch or plate effects is a very important issue, but is not considered here. We refer the interested reader to the existing literature regarding the comparison of the performances of established genotyping algorithms (e. g. Lin et al. (2008); Hong et al. (2008); Nishida et al. (2008)), and also note that an "alternative allele calling" working group has been established within the Genetic Association Information Network (GAIN, http://www.genome.gov/19518664) that carries out these comparisons. We also note that most genotype calling algorithms do make use of (at least some of) the information available in the HapMap data to train parameters etc., and thus, a completely fair comparison of the algorithms using the HapMap data alone would not have been possible in the first place.

Similar comparisons as the ones stated above will be crucial in the near future when association studies using copy number variants will hit the main stream (Korn et al. (2008); McCarroll (2008)). There is evidence that the largest effects of inter laboratory variability might manifest itself in the total fluorescent intensities (Carvalho et al. (2007)), which is the basis for gene copy number estimation. It seems likely that obtaining accurate and precise estimates for gene copy numbers and their uncertainties will proof much more challenging than the respective method developments for genotypes, making valid and efficient statistical approaches imperative.

Acknowledgments

Support was provided by NIH grants R01 DK061662 from the National Institute of Diabetes, Digestive and Kidney Diseases, R01 HL090577 from the National Heart, Lung, and Blood Institute, R01 GM083084, and a CTSA grant to the Johns Hopkins Medical Institutions.

References

Abecasis GR, Cherny SS, Cardon LR (2001) The impact of genotyping error on family-based analysis of quantitative traits. Eur J Hum Genet 9(2):130–134

Affymetrix (2006) Brlmm: An improved genotype calling method for the genechip human mapping 500k array set. Tech. rep., Affymetrix

Akey JM, Zhang K, Xiong M, Doris P, Jin L (2001) The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. Am J Hum Genet 68(6):1447–1456

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet 49(5):985–994

Carvalho B, Bengtsson H, Speed TP, Irizarry RA (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide snp array data. Biostatistics 8(2):485–499

Cheng KF, Lin WJ (2007) Simultaneously correcting for population stratification and for genotyping error in case-control association studies. Am J Hum Genet 81(4):726–743

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, mei Shen M, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S (2005) Dynamic model based algorithms for screening and genotyping over 100 k snps on oligonucleotide microarrays. Bioinformatics 21(9):1958–1963

Goldstein DR, Zhao H, Speed TP (1997) The effects of genotyping errors and interference on estimation of genetic distance. Hum Hered 47(2):86–100

Gordon D, Finch SJ, Nothnagel M, Ott J (2002) Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 54(1):22–33

Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. Eur J Hum Genet 12(9):752–761

Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 69(2):371–380

Gordon D, Levenstien MA, Finch SJ, Ott J (2003) Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case-control association studies. Pac Symp Biocomput pp. 490–501

Gordon D, Matise TC, Heath SC, Ott J (1999) Power loss for multiallelic transmission/disequilibrium test when errors introduced: Gaw11 simulated data. Genet Epidemiol 17 Suppl 1:S587–S592

Gordon D, Ott J (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput pp. 18–29

Hao K, Cawley S (2007) Differential dropout among snp genotypes and impacts on association tests. Hum Hered 63(3-4):219–228

Hao K, Wang X (2004) Incorporating individual error rate into association test of unmatched case-control design. Hum Hered 58(3-4):154–163

Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, Xu J, Chen JJ, Han T, Kaput J, Fuscoe JC, Tong W (2008) Assessing batch effects of genotype calling algorithm brlmm for the affymetrix genechip human mapping 500 k array set using 270 hapmap samples. BMC Bioinformatics 9 Suppl 9:S17

Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA (2007) Sniper-hd: improved genotype calling accuracy by an expectation-maximization algorithm for high-density snp arrays. Bioinformatics 23(1):57–63

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million snps. Nature 449(7164):851–861

Kang H, Qin ZS, Niu T, Liu JS (2004a) Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. Am J Hum Genet 74(3):495–510

Kang SJ, Finch SJ, Haynes C, Gordon D (2004b) Quantifying the percent increase in minimum sample size for snp genotyping errors in genetic model-based association studies. Hum Hered 58(3-4):139–144

Kang SJ, Gordon D, Brown AM, Ott J, Finch SJ (2004c) Tradeoff between no-call reduction in genotyping error rate and loss of sample size for genetic case/control association studies. Pac Symp Biocomput pp. 116–127

Kang SJ, Gordon D, Finch SJ (2004d) What snp genotyping errors are most costly for genetic association studies? Genet Epidemiol 26(2):132–141

Kennedy GC, Matsuzaki H, Dong S, min Liu W, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SPA, Jones KW (2003) Large-scale genotyping of complex dna. Nat Biotechnol 21(10):1233–1237

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D (2008) Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs. Nat Genet 40(10):1253–1260

Lin S, Carvalho B, Cutler D, Arking D, Chakravarti A, Irizarry R (2008) Validation and extension of an empirical bayes method for snp calling on affymetrix microarrays. Genome Biol 9(4):R63

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–913

McCarroll SA (2008) Extending genome-wide association studies to copy-number variation. Hum Mol Genet 17(R2):R135–R142

Morris RW, Kaplan NL (2004) Testing for association with a case-parents design in the presence of genotyping errors. Genet Epidemiol 26(2):142–154

Nishida N, Koike A, Tajima A, Ogasawara Y, Ishibashi Y, Uehara Y, Inoue I, Tokunaga K (2008) Evaluating the performance of affymetrix snp array 6.0 platform with 400 japanese individuals. BMC Genomics 9:431

Plagnol V, Cooper JD, Todd JA, Clayton DG (2007) A method to address differential bias in genotyping in large-scale association studies. PLoS Genet 3(5):e74

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix snp arrays. Bioinformatics 22(1):7–12

Rice KM, Holmans P (2003) Allowing for genotyping error in analysis of unmatched case-control studies. Ann Hum Genet 67(Pt 2):165–174

Zhu WS, Fung WK, Guo J (2007) Incorporating genotyping uncertainty in haplotype frequency estimation in pedigree studies. Hum Hered 64(3):172–181

SNP	AA	AB	BB	cor(g,t)	cor(g,w)	cor(w,t)	$\frac{\text{cor}(g,t)}{\text{cor}(g,w)}$	$\log(\cdot)$
rs1665933	2	28	239	0.72	0.49	0.72	1.47	0.38
rs1678775	204	57	8	0.94	0.64	0.68	1.47	0.38
rs1659131	2	38	229	0.91	0.87	0.95	1.05	0.04
rs1641760	133	111	25	1.00	1.00	1.00	1.00	0.00

Table 1: The true genotype distributions in the 269 HapMap samples for the four selected SNPs. Since the subjects came from four distinct populations, we do not expect the SNPs to be in Hardy-Weinberg equilibrium, and thus do not cite a minor allele frequency. For the genotype calls derived using BRLMM (w) and CRLMM (t), we show the pairwise correlations with each other and the true genotype (g), and the ratios and the log ratios of the correlations with the true genotype. Here, the BRLMM genotypes were called regardless of confidence threshold, and thus, do not contain missing data.

	Pool of 100 difficult SNPs			Pool 100 easy SNPs		
	mean	median	iqr	mean	median	iqr
Missing BRLMM genotypes (%)	14.05	13.38	(11.15, 16.36)	0.00	0.00	(0.00, 0.00)
Genotype confidence Score	0.21	0.21	(0.19, 0.22)	0.01	0.01	(0.01, 0.01)
$cor(gt)$	0.92	0.97	(0.93, 1.00)	1.00	1.00	(1.00, 1.00)
$cor(gw)$	0.78	0.81	(0.72, 0.86)	1.00	1.00	(1.00, 1.00)
$cor(wt)$	0.81	0.83	(0.76, 0.89)	1.00	1.00	(1.00, 1.00)
$\log \frac{cor(gt)}{cor(gw)}$	0.17	0.17	(0.10, 0.25)	0.00	0.00	(0.00, 0.00)

Table 2: Summary statistics shown as mean, median, and interquartile range, derived from the pools of SNPs (100 easy and 100 difficult to call) used in in the simulation study. The difficulty to call a SNP was defined by the average BRLMM confidence score across the 269 subjects. The true genotype is denoted by g, and the BRLMM and CRLMM genotype estimates are denoted by w and t, respectively. For the calculation of the correlations, all BRLMM genotypes were used. In parts of the simulation study, BRLMM genotypes were also recorded as missing if the confidence score exceeded the default threshold set by the algorithm.

	True genotype		CRLMM-1		CRLMM-2		BRLMM-1		BRLMM-2	
OR	med	iqr	med	iqr	med	iqr	med	iqr	med	iqr
Difficult SNPs										
1.0	50.8	(25.1, 76.0)	50.5	(24.9, 76.2)	50.3	(25.2, 76.3)	51.1	(25.3, 75.9)	51.3	(24.9, 75.9)
1.1	64.0	(38.5, 85.5)	63.3	(37.5, 85.0)	63.0	(37.2, 84.9)	62.1	(35.7, 84.0)	61.6	(35.3, 83.6)
1.2	77.3	(52.7, 92.3)	76.0	(49.7, 91.9)	75.8	(49.5, 91.8)	72.7	(46.4, 89.8)	72.7	(46.9, 90.2)
1.3	85.8	(63.1, 96.2)	83.9	(60.9, 95.5)	83.7	(60.5, 95.5)	80.1	(55.1, 93.8)	80.0	(56.3, 93.9)
1.4	91.7	(73.8, 98.1)	90.3	(71.2, 97.8)	90.1	(70.7, 97.8)	86.2	(65.0, 96.4)	86.6	(65.6, 96.6)
1.5	95.2	(82.3, 99.2)	94.2	(79.2, 98.9)	94.0	(78.7, 98.9)	90.9	(72.5, 98.0)	91.1	(73.4, 98.1)
1.6	97.3	(87.7, 99.6)	96.4	(85.0, 99.5)	96.3	(84.5, 99.4)	93.4	(78.2, 98.9)	93.8	(78.9, 98.9)
1.7	98.5	(92.0, 99.8)	98.0	(89.3, 99.8)	97.9	(89.0, 99.7)	95.9	(83.2, 99.4)	96.2	(83.9, 99.4)
1.8	99.3	(94.8, 99.9)	98.9	(92.8, 99.9)	98.9	(92.6, 99.9)	97.3	(87.3, 99.7)	97.6	(88.4, 99.7)
1.9	99.6	(96.6, 100.0)	99.4	(95.0, 100.0)	99.4	(94.7, 100.0)	98.2	(90.3, 99.8)	98.4	(90.7, 99.8)
2.0	99.8	(97.8, 100.0)	99.7	(96.5, 100.0)	99.6	(96.2, 100.0)	98.8	(92.0, 99.9)	98.9	(93.0, 99.9)
Easy SNPs										
1.0	49.8	(25.6, 74.0)	49.8	(25.6, 74.0)	49.7	(25.6, 74.0)	49.8	(25.5, 74.0)	49.8	(25.6, 74.0)
1.1	67.8	(41.0, 87.0)	67.7	(41.0, 87.0)	67.8	(41.0, 87.0)	67.8	(41.0, 87.0)	67.8	(41.0, 87.0)
1.2	81.7	(58.2, 94.4)	81.8	(58.3, 94.4)	81.8	(58.3, 94.4)	81.8	(58.3, 94.4)	81.8	(58.3, 94.4)
1.3	90.1	(71.4, 97.5)	90.1	(71.3, 97.5)	90.1	(71.3, 97.5)	90.0	(71.4, 97.5)	90.0	(71.4, 97.5)
1.4	95.4	(82.8, 99.1)	95.3	(82.9, 99.1)	95.3	(82.9, 99.1)	95.3	(82.8, 99.1)	95.4	(82.8, 99.1)
1.5	97.8	(90.1, 99.7)	97.8	(90.1, 99.7)	97.8	(90.1, 99.7)	97.8	(90.1, 99.7)	97.8	(90.1, 99.7)
1.6	99.1	(94.5, 99.9)	99.1	(94.5, 99.9)	99.1	(94.5, 99.9)	99.1	(94.5, 99.9)	99.1	(94.5, 99.9)
1.7	99.6	(96.9, 100.0)	99.6	(96.9, 100.0)	99.6	(96.9, 100.0)	99.6	(96.9, 100.0)	99.6	(96.9, 100.0)
1.8	99.8	(98.5, 100.0)	99.8	(98.5, 100.0)	99.8	(98.5, 100.0)	99.8	(98.5, 100.0)	99.8	(98.5, 100.0)
1.9	99.9	(99.1, 100.0)	99.9	(99.1, 100.0)	99.9	(99.1, 100.0)	99.9	(99.1, 100.0)	99.9	(99.1, 100.0)
2.0	100.0	(99.5, 100.0)	100.0	(99.5, 100.0)	100.0	(99.5, 100.0)	100.0	(99.5, 100.0)	100.0	(99.5, 100.0)

Table 3: The median and inter-quartile range for the association percentile rank of the sole causal SNP among all other 32,442 non-associated SNPs (in truth), separate for easy and difficult to call SNPs, and different degrees of associations (odds ratios, OR). The strongest statistical association measured in each iteration of the simulation corresponds to the percentile rank 100. The results are shown for the cases when the true genotypes are used to calculate the statistical association, the fuzzy CRMM genotype calls (CRLMM-1), genotype calls derived as the modes from the CRLMM posterior probabilities (CRLMM-2), all BRLMM genotypes regardless of confidence scores (BRLMM-1), and the BRLMM genotypes after dropping the uncertain genotypes (BRLMM-2).

	True genotype		CRLMM-1		CRLMM-2		BRLMM-1		BRLMM-2	
OR	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Difficult SNPs										
1.0	0.01	(1.017)	0.01	(1.018)	0.01	(1.017)	0.01	(1.020)	0.01	(1.017)
1.1	0.38	(1.020)	0.36	(1.018)	0.35	(1.017)	0.31	(1.021)	0.30	(1.015)
1.2	0.75	(1.031)	0.71	(1.039)	0.70	(1.039)	0.60	(1.035)	0.61	(1.033)
1.3	1.07	(1.082)	1.00	(1.082)	0.99	(1.083)	0.86	(1.073)	0.86	(1.070)
1.4	1.40	(1.104)	1.31	(1.113)	1.30	(1.116)	1.12	(1.093)	1.13	(1.090)
1.5	1.70	(1.143)	1.60	(1.160)	1.58	(1.161)	1.37	(1.131)	1.39	(1.131)
1.6	1.97	(1.169)	1.85	(1.187)	1.83	(1.190)	1.58	(1.168)	1.60	(1.154)
1.7	2.25	(1.194)	2.10	(1.230)	2.08	(1.232)	1.80	(1.197)	1.83	(1.180)
1.8	2.51	(1.238)	2.35	(1.278)	2.33	(1.281)	2.01	(1.242)	2.04	(1.214)
1.9	2.74	(1.252)	2.58	(1.303)	2.55	(1.310)	2.19	(1.260)	2.22	(1.237)
2.0	2.95	(1.294)	2.78	(1.354)	2.75	(1.357)	2.37	(1.332)	2.41	(1.298)
Easy SNPs										
1.0	0.00	(0.987)	0.00	(0.988)	0.00	(0.988)	0.00	(0.988)	0.00	(0.988)
1.1	0.45	(1.015)	0.45	(1.015)	0.45	(1.015)	0.45	(1.015)	0.45	(1.015)
1.2	0.91	(1.048)	0.91	(1.048)	0.91	(1.049)	0.91	(1.048)	0.91	(1.048)
1.3	1.30	(1.064)	1.30	(1.063)	1.30	(1.063)	1.30	(1.063)	1.30	(1.063)
1.4	1.70	(1.113)	1.70	(1.113)	1.70	(1.113)	1.70	(1.113)	1.70	(1.113)
1.5	2.08	(1.135)	2.08	(1.134)	2.08	(1.134)	2.08	(1.135)	2.08	(1.135)
1.6	2.42	(1.161)	2.42	(1.161)	2.42	(1.161)	2.42	(1.161)	2.42	(1.161)
1.7	2.72	(1.217)	2.72	(1.217)	2.72	(1.217)	2.72	(1.217)	2.72	(1.217)
1.8	3.02	(1.231)	3.02	(1.231)	3.02	(1.231)	3.02	(1.232)	3.02	(1.232)
1.9	3.30	(1.263)	3.30	(1.263)	3.30	(1.263)	3.30	(1.264)	3.30	(1.264)
2.0	3.56	(1.280)	3.55	(1.281)	3.55	(1.281)	3.55	(1.281)	3.55	(1.281)

Table 4: The sample mean and sample standard deviation of the z-scores of the causal SNP used in the simulation, separate for the easy and difficult to call SNPs, and different degrees of association (odds ratios, OR). The notation for the genotype methods employed is the same as in Table 3.

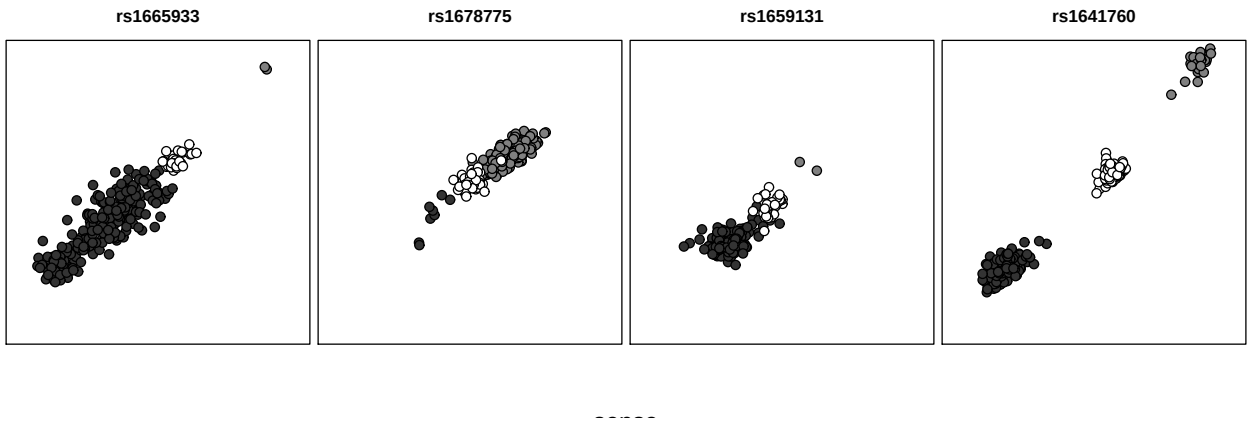

Figure 1: The raw fluorescence intensities from the sense (x-axis) and anti-sense (y-axis) strands for the 269 HapMap phase II individuals, shown for 4 selected SNPs (rs numbers shown in the title of the panels). The true genotypes for each individual are shown as dark grey (AA), white (AB), and light grey (BB) dots in the respective panels. The three genotype clusters in rs1641760 separate well, and thus, most subjects will be correctly genotyped at this variant in genomic assays. The separation of the genotype clusters for the other three SNPs is less clear, and thus, future genotype estimates at these loci are much more prone to error.

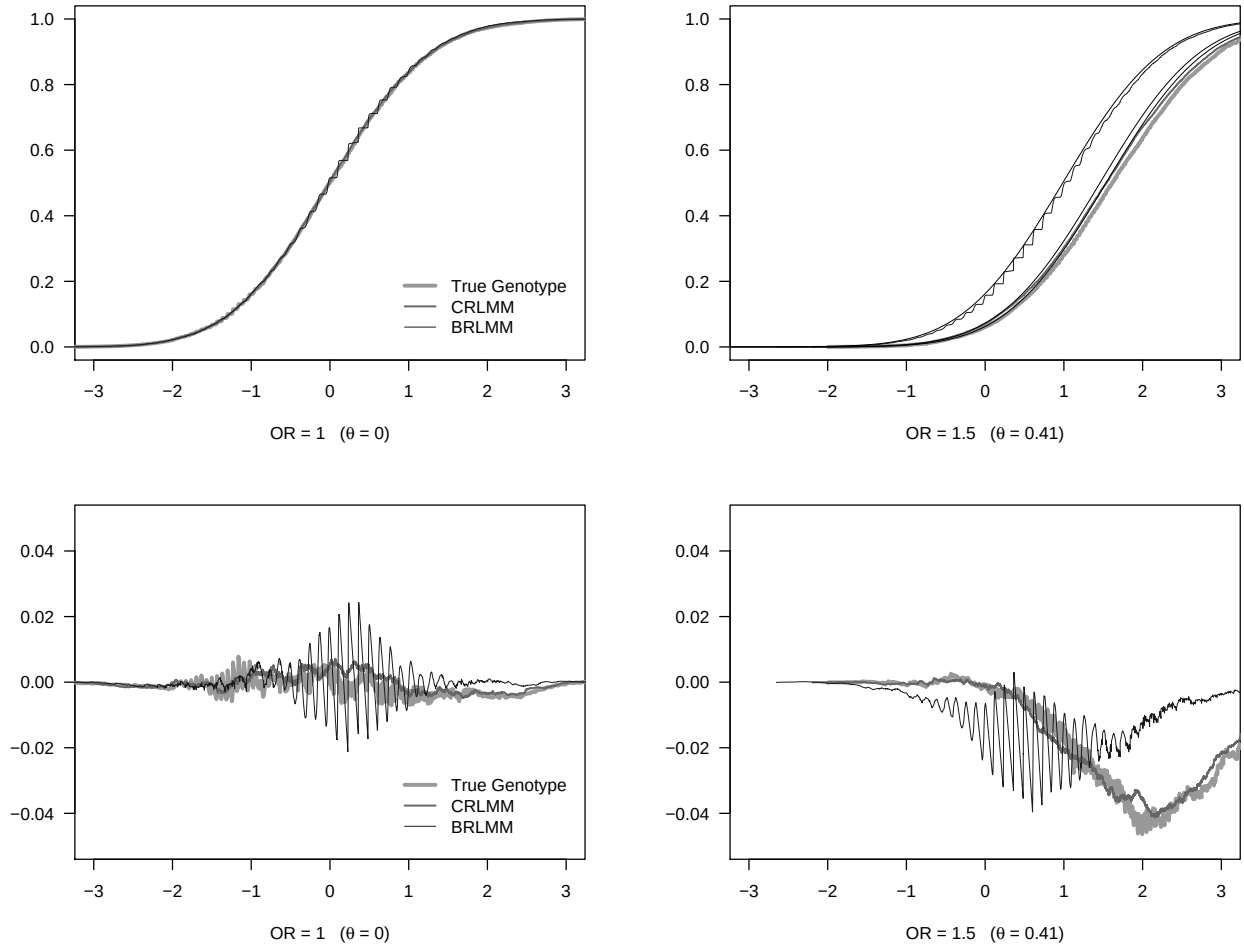

Figure 2: Simulated distributions for the score test statistics for SNP rs1678775, using the true genotypes (thick and light line), and the output from the genotype calling algorithms BRLMM (thin and dark line) and CRLMM (intermediate thickness and color). The upper panels show the empirical cumulative distribution functions, with the theoretical cumulative distribution functions underlying as thin smooth curves. The lower panels show the differences between the empirical and theoretical cumulative distribution functions. The left column shows the results under the null (OR $= 1, \theta = 0$), the right column under an alternative (OR $= 1.5, \theta = 0.41$). In general, the theoretical cumulative distribution functions match the respective empirical cumulative distribution functions well, lending credibility to the validity of the test. The strong oscillation in some of the curves is due to the discrete nature of the data. Note that the curve for the true genotypes is also oscillating, albeit with a much smaller magnitude, a feature of this particular sample.

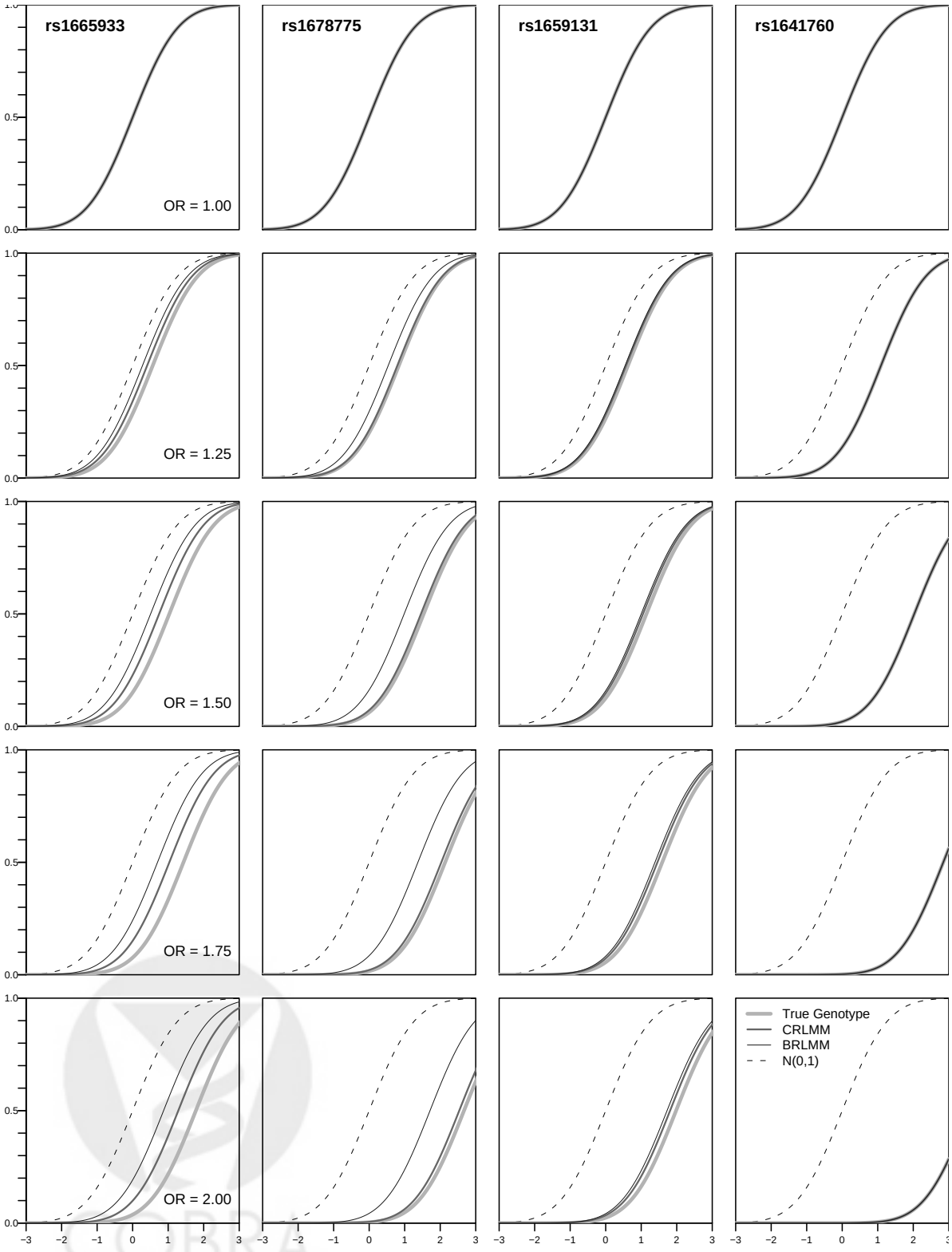

Figure 3: The theoretical cumulative distribution functions for the score test statistics, for the four selected SNPs (columns 1-4). Shown are results under the null (row 1, OR=1), and 4 different alternatives (rows 2-5). In each panel, the distributions for the true genotype, and the genotypes estimated by BRLMM and CRLMM are compared. This allows for a comparison of the loss of power due to genotyping uncertainties. Also indicated is the Normal (0,1) null distribution as a dashed line.

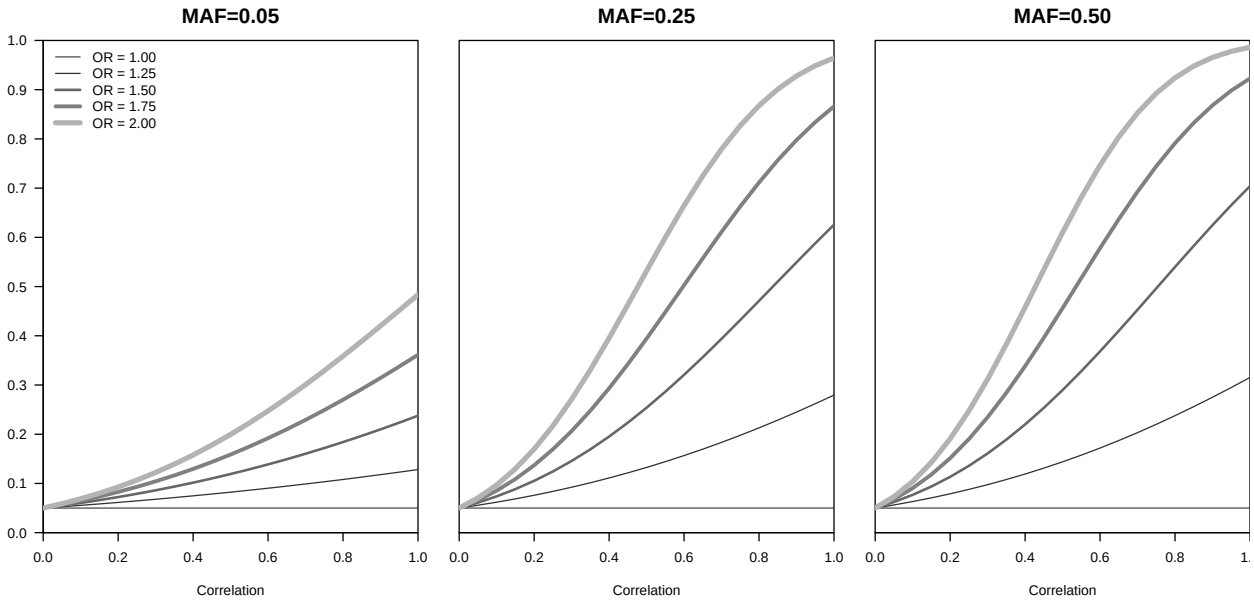

Figure 4: The power to detect an association as a function of the correlation between the true genotype and the output from a genotype calling algorithm, shown for three minor allele frequencies (panels 1-3), and various effect sizes (lines within a panel), assuming a sample size of 269. In all instances, the power diminishes rapidly with decreasing correlation.

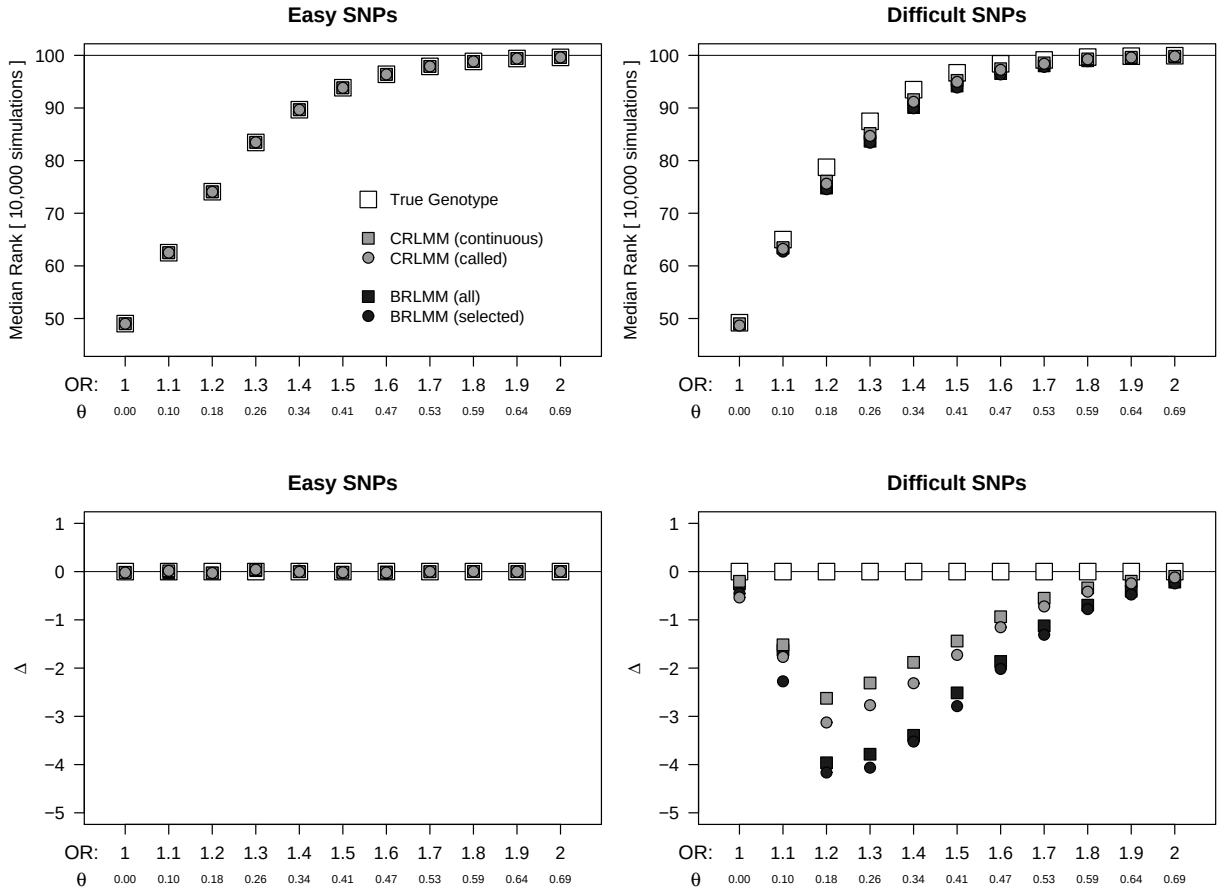

Figure 5: Simulation results using easy to call SNPs (left column) and hard to call SNPs (right column). Shown are the median percentile ranks derived from 10,000 iterations, for increasing effect sizes. A SNP was randomly selected from the set of easy or difficult SNPs respectively, and the signal was generated using the indicated effect sizes for the selected SNP. The association test statistics were calculated for all 32,443 SNPs, and for the signal carrier the percentile rank among all SNPs was calculated (100% indicates the top selection). The procedure was repeated 10,000 times, and the median percentile rank is shown in the upper rows using 5 different measures for the genotypes: the true genotypes (open square), the CRLMM "fuzzy" continuous genotypes (light grey square), the CRLMM genotype calls (light grey circle), the BRLMM genotype calls forcing complete data (dark grey square), and the BRLMM genotype calls allowing for missing data, using the default cut-off of 0.5 for genotype call confidence (dark grey circle). The lower panels show the differences in the results between the true genotypes and the respective genotype measures. As expected, for easy to call SNPs no substantial differences are observed. For difficult to call SNPs better results are achieved with an improved genotype calling algorithm (CRLMM versus BRLMM), but further improvement can be achieved using the entire information available in the continuous genotypes.

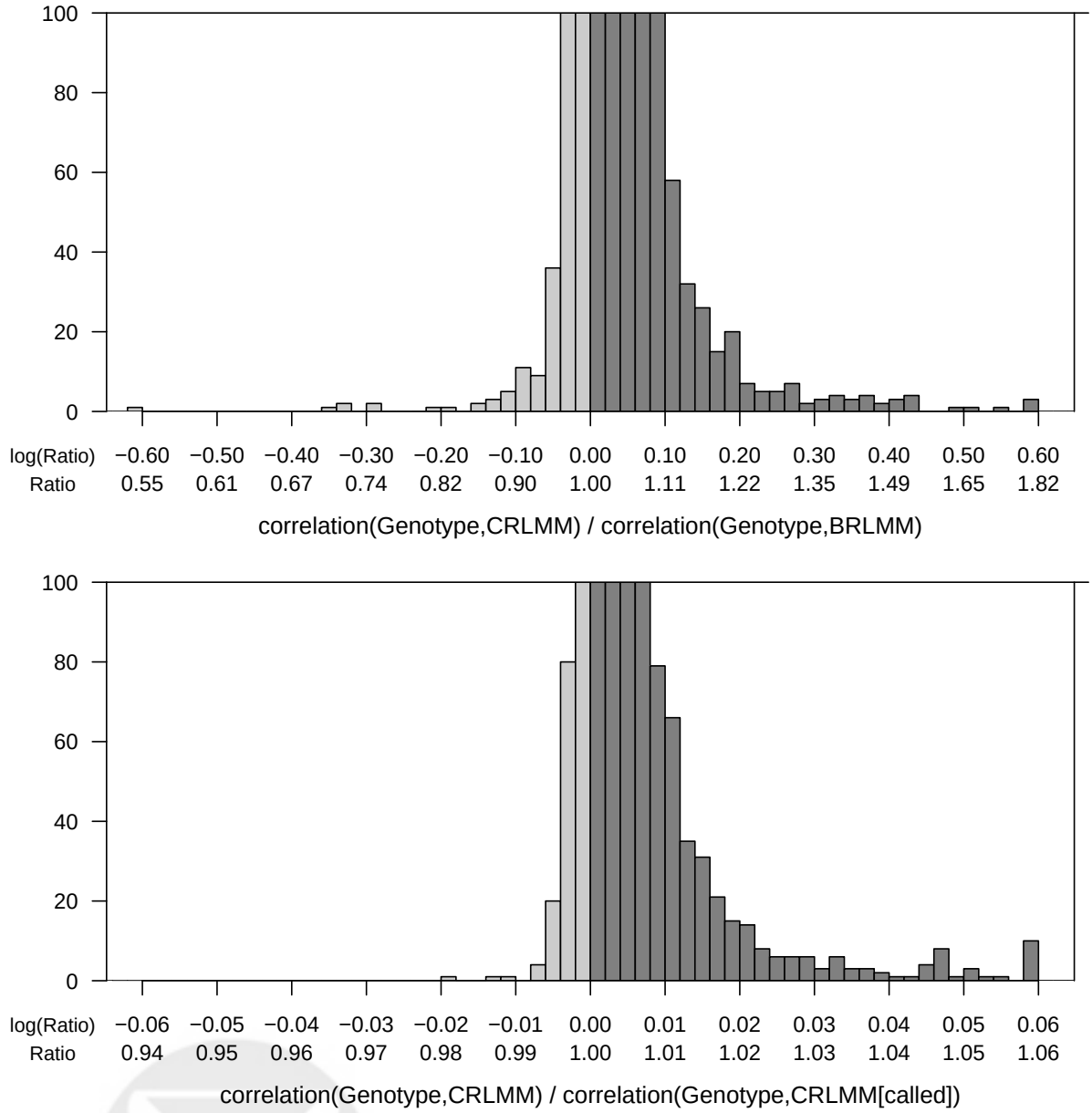

Figure 6: [*Top*] For all 32,443 SNPs we calculated the correlations of the true genotypes with the CRLMM fuzzy genotype probabilities, and the correlations of the true genotypes with the BRLMM genotype calls, forcing complete data. The (logarithm of the) ratios of these correlations is shown as a histogram. A ratio larger than one for any particular SNP indicates more power in the score test if there is signal in the SNP. A ratio less than one can for example arise if there is uncertainty in the genotype estimates, but BRLMM gets all genotype calls correct. However, the excess of ratios larger than one clearly shows the merit of using the CRLMM genotype probabilities. For clarity, the histogram was truncated at 100 counts. [*Bottom*] Histogram of the (log-) ratios of the correlations of the true genotypes with the CRLMM genotype probabilities, and the correlations of the true genotypes with the discrete CRLMM genotype calls (posterior modes). The excess of ratios larger than one shows the merit of using the fuzzy CRLMM genotypes. This histogram was also truncated at 100 counts.