How allele frequency and study design affect association test statistics with misrepresentation errors

VALENTINA ESCOTT-PRICE

MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK
escottpricev@cf.ac.uk

MANSOUREH GHODSI

School of Mathematics, Cardiff University, Cardiff CF24 4AG, UK
Business School, Bournemouth University, Bournemouth BH8 8EB, UK

KARL MICHAEL SCHMIDT

School of Mathematics, Cardiff University, Cardiff CF24 4AG, UK

SUMMARY

We evaluate the effect of genotyping errors on the type-I error of a general association test based on genotypes, showing that, in the presence of errors in the case and control samples, the test statistic asymptotically follows a scaled non-central $\chi^2$ distribution. We give explicit formulae for the scaling factor and non-centrality parameter for the symmetric allele-based genotyping error model and for additive and recessive disease models. They show how genotyping errors can lead to a significantly higher false-positive rate, growing with sample size, compared with the nominal significance levels. The strength of this effect depends very strongly on the population distribution of the genotype, with a pronounced effect in the case of rare alleles, and a great robustness against error in the case of large minor allele frequency. We also show how these results can be used to correct $p$-values.

Keywords: Case–control study; General association test; Genomic control; Genotyping errors; Non-central $\chi^2$ distribution; Type-I error.

1. Introduction

The effects of genotyping errors have been investigated in a number of different contexts, for instance, the genotyping errors leading to false-positive results, false-negative results, or both (Brush and Almasy, 2003; Gordon and others, 2002, 2007; Kang and others, 2004; Moskvina and others, 2005, 2006). Recent genetic research has moved into the area of rare mutations, and thus the reliability of data quality has gained even more importance. For genetic association studies using case and control subjects, Moskvina and others (2006) have shown that differences in the genotyping error rate between the case and control samples can result in substantially increased type-I error, even at low genotyping error rates

*To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
(<0.01). This effect is most pronounced for SNP markers with small minor allele frequency (MAF) and markers in strong linkage disequilibrium. More generally, it has been observed in genome-wide association studies with large sample sizes that inflation of the overall test statistics is still present after correction for possible confounders (e.g. population stratification); see Freedman and others (2004).

The purpose of this paper is to give a theoretical analysis of the effects of allele misrepresentations on the type-I error of case-control studies. In contrast to the previous studies, where the occurrence of such, partly quite startling, effects has been demonstrated in real or simulated data examples, the present work will take a theoretical approach, showing how the distortion of the test statistic arises, what form it takes, and how the various study parameters (sample sizes, population genotype frequency distributions, size and type of genotyping errors, and disease model) affect the outcome. This provides an overview of all possible situations and allows pinpointing cases of strong distortion as well as identifying a range of situations where genotyping errors have little influence on the statistical test. Moreover, this analytical rather than simulation approach allows for the easy investigation of the effects for SNPs with very rare variants, where simulations become impracticable. It also shows that the quantitative effect of genotyping errors can vary strongly between different SNPs, even if the genotyping error rate is constant.

For an association test based on genotypes, we give a formula for the distribution of the test statistic in the presence of genotyping error for the additive, dominant, and recessive disease models. We consider the allele-based error model studied by Moskvina and Schmidt (2006). In addition to assessing the strength of the presence of genotyping error for the additive, dominant, and recessive disease models. We consider the general association test statistic

\[
T = \frac{N(N\sum_{i=0}^{2} w_i n_i - n \sum_{i=0}^{2} w_i N_i)^2}{n(N-n)(N \sum w_i^2 N_i - (\sum w_i N_i)^2)} \sim \chi^2_{df=1},
\]

with weights \(w_i \geq 0, w_1 + w_2 + w_3 > 0\), corresponding to the disease model. For example, the special choice \((w_0, w_1, w_2) = (0, 1, 2)\) represents the additive disease (allelic association) model

\[
T = \frac{N(N(n_1 + 2n_2) - n(N_1 + 2N_2))^2}{n(N-n)(N(N_1 + 4N_2) - (N_1 + 2N_2)^2)}.
\]

This is the standard Armitage trend test; see Armitage (1955). Choosing weights \((w_0, w_1, w_2) = (1, 0, 0)\) and \((w_0, w_1, w_2) = (0, 1, 1)\) results in the recessive disease model

\[
T = \frac{N(N N_0 - n N_0)^2}{n(N-n)(NN_0 - N_0^2)}.
\]
and the dominant disease model

\[ T = \frac{N(N(n_1 + n_2) - n(N_1 + N_2))^2}{n(N - n)(N(N_1 + N_2) - (N_1 + N_2)^2)}, \tag{2.3} \]

respectively. Note that (2.3) is in fact equal to (2.2). In addition, the general test (2.1) can accommodate other models, for example, the overdominant \((w_0, w_1, w_2) = (0, 1, 0)\).

### 2.1 Distribution of the test statistic under the alternate hypothesis

As a foundation for our later considerations, we prove, in Section 5, part A, the following key theorem on the distribution of the general test statistic (2.1) when the genotype frequency distributions in the case and control population are different.

**Theorem 1** Let \(p = (p_0, p_1, p_2)\) and \(p' = (p'_0, p'_1, p'_2)\) be the genotype frequency distributions in the case population and in the control population, respectively. Then, asymptotically for large number of cases and controls, the general association test statistic \(T\) satisfies the distribution

\[ T \sim \chi^2 \gamma(\xi + \sqrt{\lambda}), \tag{2.4} \]

where \(\xi \sim N(0, 1)\) is a standard normal random variable, \(\lambda = (mn/\beta)(\sum_{i=0}^{2} w_i(p_i - p'_i))^2\) is the non-centrality parameter and

\[ \gamma = \frac{\beta}{N(\sum_{i=0}^{2} w_i^2 \hat{p}_i - (\sum_{i=0}^{2} w_i \hat{p}_i)^2)} \tag{2.5} \]

is a scaling factor, with \(\hat{p}_i = (n/N)p_i + (m/N)p'_i\) and

\[ \beta = m(w_0 - w_1)^2 \frac{p_1 p_0}{1 - p_2} + m \frac{p_2}{1 - p_2}((w_1 - w_2)p_1 - (w_2 - w_0)p_0)^2 \]

\[ + n(w_0 - w_1)^2 \frac{p'_1 p'_0}{1 - p'_2} + n \frac{p'_2}{1 - p'_2}((w_1 - w_2)p'_1 - (w_2 - w_0)p'_0)^2. \tag{2.6} \]

Note that \(\hat{p}_i\) are estimated genotype frequencies for the combined case and control sample. Under the null hypothesis \(p_i = p'_i\), it is obvious that \(\lambda = 0\); it also follows after some calculation, using (5.1), that \(\gamma = 1\). This confirms the fact that \(T \sim \chi^2_{df=1}\) under the null hypothesis.

### 2.2 Genotyping errors

In the following, we study the effect of genotyping errors on the test statistic \(T\) under the null hypothesis that the genotype frequency distributions in the case and control populations are the same. This identity of distributions in the populations is disturbed by genotyping errors in the samples. In particular, if the genotyping error rates are different between the case and the control samples, different effective genotype frequency distributions result. The samples appear further removed from the identical distribution than expected due to the sampling error, and consequently the type-I error is inflated.

Typically genotyping errors are modeled by an error matrix \(D\) which gives the frequencies of misattribution of genotypes (see Moskvina and Schmidt, 2006 for an overview). The observed genotype frequency
distribution \( \tilde{p} = (\tilde{p}_0, \tilde{p}_1, \tilde{p}_2)^T \) arises from the actual distribution \( p = (p_0, p_1, p_2)^T \) via

\[
\tilde{p} = (I_3 + D)p, \tag{2.7}
\]

where \( I_3 \) is the \( 3 \times 3 \) unit matrix. As the genotype frequencies always add up to 1, each column of the matrix \( D \) adds up to 0.

Specifically for the allele-based error model of Moskvina and Schmidt (2006), where the frequency of a misrepresentation of allele 0 as allele 1 is \( \delta(1 + \epsilon) \) and the frequency of a misrepresentation of allele 1 as 0 is \( \delta(1 - \epsilon) \) (here \( \delta \) is the averaged allele misrepresentation frequency and \( \epsilon \) is a measure for the asymmetry), the error matrix takes the form

\[
D = 2\delta \begin{pmatrix}
-(1 + \epsilon) & \frac{1}{2}(1 - \epsilon) & 0 \\
1 + \epsilon & -1 & 1 - \epsilon \\
0 & \frac{1}{2}(1 + \epsilon) & -(1 - \epsilon)
\end{pmatrix}. \tag{2.8}
\]

In the symmetric case \( \epsilon = 0 \), the misrepresentation frequency is the same for both alleles.

In this modeling of genotyping errors, the error-affected samples are considered as sampled from a population with genotype frequency distribution \( \tilde{p} \) instead of \( p \). Although this does not directly mimic the process in which genotyping errors arise after sampling, it does give the same result, as simulation shows.

In this paper, we focus on the symmetric allele-based error model, taking \( \epsilon = 0 \). As shown by Moskvina and Schmidt (2006), the action of this model is conveniently described in terms of the additive/dominance variables (cf. Cockerham and Zeng, 1996; Cordell, 2002) in the genotype frequency distribution space, instead of \( p_0, p_1, p_2 \) themselves. In fact, due to the constraint \( p_0 + p_1 + p_2 = 1 \), the genotype frequency distributions have only two independent dimensions, and form a triangle, since \( p_i \geq 0 \).

Thus, given a genotype frequency distribution \( p \), we define the pair of coordinates \( \alpha_1 = p_0 - p_2 \) and \( \alpha_2 = p_0 - p_1 + p_2 \); then \( \alpha = (\alpha_1, \alpha_2) \in \nabla \), where

\[
\nabla = \left\{ (\alpha_1, \alpha_2) \in \mathbb{R}^2 \mid \alpha_2 \in [-1, 1], \alpha_1 \in \left[ -\frac{\alpha_2 + 1}{2}, \frac{\alpha_2 + 1}{2} \right] \right\},
\]

and conversely

\[
p_0 = \frac{1}{4} + \frac{\alpha_1}{4}, \quad p_1 = \frac{1}{2} - \frac{\alpha_2}{2}, \quad p_2 = \frac{1}{4} - \frac{\alpha_1}{2} + \frac{\alpha_2}{4} \tag{2.9}
\]

is the genotype frequency distribution corresponding to \( \alpha = (\alpha_1, \alpha_2) \in \nabla \).

### 2.3 Scaling factor

In Section 5, part B, we derive the following formula for the scaling factor \( \gamma \) of (2.5) expressed in terms of the sample sizes and the \( \alpha \) coordinates of the case and control genotype frequency distributions. The approximation is valid when the genotype frequency distributions in cases and controls are close together (\( \alpha \approx \alpha' \)); this will apply to the null hypothesis under the assumption of small genotyping error rates.

**Lemma 1** Let \( \alpha, \alpha' \in \nabla \) be the coordinate vectors for the genotype frequency distributions in the case and control populations, respectively. Then, for \( \alpha \approx \alpha' \), the scaling factor \( \gamma \) is approximately

\[
\gamma \approx 1 + \frac{m - n}{N} \frac{1}{S(\hat{\alpha})} (\partial_1 S(\hat{\alpha})(\alpha - \alpha') + \partial_2 S(\hat{\alpha})(\alpha_2 - \alpha_2')) \tag{2.10}
\]
Specifically for the additive disease model, the scaling factor \( \gamma \) becomes

\[
\gamma \approx 1 + \frac{2}{\mathsf{S}_{\text{add}}(\hat{\alpha})} \frac{m-n}{N} \left( 2\hat{\alpha}_1 \hat{\alpha}_2 - 2\hat{\alpha} \hat{\alpha}_1 \hat{\alpha}_2 \right). 
\]
but in practice, the true genotype frequency distribution coordinates $\hat{\alpha}$ are unknown, as only estimates for the coordinates of the error-affected genotype frequency distributions, $\alpha$, $\alpha'$ (and hence for $\dot{\alpha}$) are available. Noting that

$$\dot{\alpha}_1 = \left(1 - 2 \frac{n \delta_{ca} + m \delta_{co}}{N}\right) \hat{\alpha}_1, \quad \dot{\alpha}_2 = \left(1 - 4 \frac{n \delta_{ca} + m \delta_{co}}{N}\right) \hat{\alpha}_2,$$

and solving for $\dot{\alpha}_1, \dot{\alpha}_2$, we obtain an approximate scaling factor

$$\gamma \approx 1 + \frac{m - n}{N} \frac{2}{S_{add}(\dot{\alpha})} \left(\hat{\alpha}_2 - 2 \hat{\alpha}_1^2 \right) (\delta_{co} - \delta_{ca}), \quad (2.19)$$

where the last approximation is good for small genotyping error rates $\delta_{ca}, \delta_{co}$.

Similar considerations for the dominant/recessive model give, as an analog to (2.19),

$$\gamma \approx 1 + \frac{m - n}{N} \frac{4 \hat{\alpha}_1^2}{S_{rec}(\dot{\alpha})} (\hat{\alpha}_1 + \hat{\alpha}_2) (\delta_{co} - \delta_{ca}). \quad (2.20)$$

The $\dot{\alpha}$-dependent terms in (2.19) and (2.20) describe the dependence of the scaling factor on the averaged genotype frequency distribution.

When we restrict our attention to genotype frequency distributions in Hardy–Weinberg equilibrium (HWE), i.e. those where $\dot{\alpha}_2 = \dot{\alpha}_1^2$, the formulæ simplify to

$$\gamma \approx 1 + \frac{m - n}{N} \frac{4 \hat{\alpha}_1^2}{\hat{\alpha}_1^2 - 1} (\delta_{co} - \delta_{ca}), \quad (2.21)$$

for the additive model, and

$$\gamma \approx 1 + \frac{m - n}{N} \frac{\hat{\alpha}_1 (1 - \hat{\alpha}_1 (2 + \hat{\alpha}_1))}{(1 - \hat{\alpha}_1^2)(3 + \hat{\alpha}_1)} (\delta_{co} - \delta_{ca}), \quad (2.22)$$

for the recessive/dominant model; note that, in this case, $(1 + \hat{\alpha}_1)/2$ is the frequency of allele 0 (the recessive allele in the recessive/dominant model) and $(1 - \hat{\alpha}_1)/2$ is the frequency of allele 1.

### 2.4 Non-centrality parameter

In Section 5, part C, we derive the following approximate expression for the non-centrality parameter $\lambda$ (note that $\ddot{\alpha}$ below subtly differs from $\dot{\alpha}$ as defined in (2.12)).

**Lemma 2** Let $\alpha, \alpha' \in \nabla$ be the coordinate vectors for the genotype frequency distribution in the case and control populations, respectively. Then the non-centrality parameter $\lambda$ in (2.4) is

$$\lambda \approx N \frac{m n}{N} \frac{\frac{1}{2}(w_0 - w_2)(\alpha_1 - \alpha'_1) + \frac{1}{4}(w_0 - 2w_1 + w_2)(\alpha_2 - \alpha'_2)^2}{S(\ddot{\alpha})}, \quad (2.23)$$

where $S$ is the same function as in (2.11) and $\ddot{\alpha} = (m/N)\alpha + (n/N)\alpha'$. 


In the special case of the additive disease model \((w_i = i, i = 0, 1, 2)\), the non-centrality parameter is approximately

\[
\lambda \approx N \frac{m}{N} \frac{n}{N} \frac{1}{S_{\text{add}}(\tilde{\alpha})} (\alpha_1 - \alpha'_1)^2,
\]

(2.24)

where \(S_{\text{add}}\) is as in (2.13). Similarly, for the recessive disease model \((w_0 = 1, w_1 = 0, w_2 = 0)\) we find

\[
\lambda \approx N \frac{m}{N} \frac{n}{N} \frac{1}{S_{\text{rec}}(\tilde{\alpha})} \left( \frac{2\alpha_1 - 2\alpha'_1 + \alpha_2 - \alpha'_2}{4} \right)^2,
\]

(2.25)

with \(S_{\text{rec}}\) as in (2.15). These approximations are good for \(\alpha \approx \alpha'\).

For the symmetric allele-based genotyping error model, we again replace \(\alpha, \alpha'\) with \(\hat{\alpha}\) using (2.16) and (2.17). For the additive disease model, the non-centrality parameter becomes (see (2.24))

\[
\lambda \approx N \frac{m}{N} \frac{n}{N} \frac{1}{2} \frac{4\hat{\alpha}^2_1}{S_{\text{add}}(\hat{\alpha})} (\delta_{\text{co}} - \delta_{\text{ca}})^2,
\]

(2.26)

Again, the true genotype frequency distribution coordinates \(\hat{\alpha}\) are unknown in practice, but can be estimated from \(\tilde{\alpha}\), e.g. \(\tilde{\alpha}_1 = (1 - (2/N)(m\delta_{\text{ca}} + n\delta_{\text{co}}))\hat{\alpha}_1\). This gives the approximate expression for the non-centrality parameter (with \(S_{\text{add}}\) as in (2.13))

\[
\lambda \approx N \frac{m}{N} \frac{n}{N} \frac{4\hat{\alpha}^2_1}{S_{\text{add}}(\tilde{\alpha})} \left( \frac{1}{1 - (2/N)(m\delta_{\text{ca}} + n\delta_{\text{co}})} \right)^2 (\delta_{\text{co}} - \delta_{\text{ca}})^2
\]

\[
\approx N \frac{m}{N} \frac{n}{N} \frac{4\hat{\alpha}^2_1}{S_{\text{add}}(\tilde{\alpha})} \left( 1 + 4m\delta_{\text{ca}} + n\delta_{\text{co}} \right) (\delta_{\text{co}} - \delta_{\text{ca}})^2,
\]

(2.27)

with the last approximation valid if the genotyping error rates \(\delta_{\text{ca}}, \delta_{\text{co}}\) are small.

Analogously, for the dominant/recessive disease model,

\[
\lambda \approx N \frac{mn}{N^2} \frac{\hat{\alpha}_1 + \hat{\alpha}_2}{S_{\text{rec}}(\tilde{\alpha})} \left( \hat{\alpha}_1 + \hat{\alpha}_2 \right) \left( \frac{4}{N} (m\delta_{\text{ca}} + n\delta_{\text{co}}) (\hat{\alpha}_1 + 2\hat{\alpha}_2) \right) (\delta_{\text{co}} - \delta_{\text{ca}})^2,
\]

(2.28)

with \(S_{\text{rec}}\) as in (2.15).

When we restrict our attention to genotype frequency distributions in HWE \((\tilde{\alpha}_2 = \hat{\alpha}_1^2)\), these formulae simplify to

\[
\lambda \approx N \frac{mn}{N^2} \frac{8\hat{\alpha}^2_1}{1 - \hat{\alpha}^2_1} \left( 1 + \frac{4}{N} (m\delta_{\text{ca}} + n\delta_{\text{co}}) \right) (\delta_{\text{co}} - \delta_{\text{ca}})^2,
\]

(2.29)

for the additive model and

\[
\lambda \approx N \frac{mn}{N^2} \frac{16\hat{\alpha}^2_1}{(1 - \alpha_1)(3 + \alpha_1)} \left( 1 + \frac{4}{N} (m\delta_{\text{ca}} + n\delta_{\text{co}}) \frac{1 + 2\hat{\alpha}_1}{1 + \hat{\alpha}_1} \right) (\delta_{\text{co}} - \delta_{\text{ca}})^2,
\]

(2.29)

for the dominant/recessive model; again \((1 + \hat{\alpha}_1)/2\) is the frequency of allele 0 (the recessive in the second model), \((1 - \hat{\alpha}_1)/2\) the frequency of allele 1 in this case.
3. Discussion

From our analysis in Section 2, we can draw general conclusions about the dependence of the distorted distribution (2.4) of the test statistic on the sample sizes, genotype frequency distribution, and genotyping error rates.

It is clear from (2.10) that the scaling factor $\gamma$ only depends on the difference between the relative sizes of the samples, $n/N$ and $m/N$, but not on the overall sample size $N$. Thus, increasing the size of both samples in equal proportion will not affect $\gamma$. In the special case where the case and control samples have the same size, $\gamma \approx 1$ in the approximation of close genotype distributions (corresponding to small genotyping error frequency under the null hypothesis).

If the sample sizes are different, then the scaling factor depends on the genotyping error rates and the population genotype frequency distribution, which, in $\alpha$ coordinates, will be close to $\dot{\alpha}$ for a small genotyping error frequency. Formula (2.18) shows that only the difference between the genotyping error rates for the two samples is relevant; $\gamma \approx 1$ if the genotyping error rates are equal. The dependence of $\gamma$ on the population genotype frequency can be very pronounced; for example, the formula (2.13) for the additive model shows that the denominator $S_{\text{add}}(\dot{\alpha})$ becomes very small as $\dot{\alpha}_1 \approx \pm 1$, $\dot{\alpha}_2 \approx 1$, and as $\dot{\alpha}_1 \approx 0$, $\dot{\alpha}_2 \approx -1$, corresponding to genotype frequency distributions where one genotype is strongly prevalent. (However, genotype frequency distributions with $\dot{\alpha}_1 \approx 0$, $\dot{\alpha}_2 \approx -1$ are very far out of HWE and hence unlikely to occur in practice.) The dependence of the scaling factor on the genotype frequency distribution is illustrated in Figures 1 and 2, showing that there is a large effect near the corners of $\nabla$, corresponding to genotype frequency distributions with rare genotypes. For the additive model (see (2.19), (2.21), and Figure 1), the
Fig. 2. The behavior of $(2\hat{\alpha}_1 + \hat{\alpha}_2 - 1)(\hat{\alpha}_1 + \hat{\alpha}_2)/2S_{\text{rec}}(\hat{\alpha})$ in the scaling factor for the recessive disease model. The black line corresponds to HWE and the gray lines to different choices of fixed $\alpha_2$, as in Figure 1.

curve corresponding to genotype frequency distributions in HWE shows unlimited growth of the factor at small MAFs. For the recessive/dominant model (see (2.20), (2.22), and Figure 2), the factor exhibits a curious behavior in the limit where the recessive allele is rare, i.e. $\alpha_1 \approx -1$. The function tends to a finite limit $-2$ as $\alpha_1$ tends to $-1$ along the line $\alpha_2 = 1$. However, it is discontinuous at the corresponding top-left corner of $\nabla$, and, for other angles of approach to the limit, becomes unbounded much as at the right-hand corner, where the dominant allele is rare. In particular, the curve corresponding to genotype frequency distributions in HWE again shows unbounded growth of the factor in both the limits of rare recessive and of rare dominant allele, with an even stronger effect for rare recessive alleles. Thus, the situation of a rare recessive allele is not only susceptible to strong inflation of type-I error due to genotyping errors, but in contrast to the situation of a rare dominant allele, the extent of this susceptibility also depends sensitively on the genotype frequency distribution.

Regarding the non-centrality parameter $\lambda$, (2.23) shows that $\lambda$ is generally directly proportional to the total sample size $N$ (in contrast, the scaling factor $\gamma$ only depends on the relative sample sizes). It is also proportional to the product of the relative sample sizes $(m/N)(n/N)$, which takes its maximum value when the number of cases and controls are equal; in this situation $\gamma \approx 1$. Equations (2.26) and (2.27) exhibit a simple pattern of dependence of the non-centrality parameter $\lambda$ on the differential genotyping error rate $\delta_{ca} - \delta_{co}$, but differ in the dependence on the SNP genotype frequency distribution, as represented by the coordinates $\tilde{\alpha}$. Figure 3 shows the behavior of $\tilde{\alpha}_1^2/S_{\text{add}}(\tilde{\alpha})$ of (2.26) for the additive disease model. The non-centrality parameter grows without bound in the limit of rare variants, but remains small in most cases of larger MAF. In the formula for the dominant/recessive disease model (2.27), the dependence on $\tilde{\alpha}$ is more complicated and contains a term involving the genotyping error rates. However, in the approximation
V. ESCOTT-PRICE AND OTHERS

Fig. 3. The behavior of $\hat{\alpha}_1^2 / S_{\text{add}}(\tilde{\alpha})$ in the non-centrality parameter for the additive model. The black line corresponds to HWE and the gray lines to different fixed $\tilde{\alpha}_2$, analogous to Figure 1.

of small genotyping error rates, we can neglect this term; Figure 4 shows the behavior of the leading contributing factor $(\tilde{\alpha}_1 + \tilde{\alpha}_2)^2 / S_{\text{rec}}(\tilde{\alpha})$. Similarly to the scaling factor (Figure 2), this is unbounded both for rare recessive and dominant alleles, but tends to the finite limit 0 as $\tilde{\alpha}_1$ approaches $-1$ along the line $\tilde{\alpha}_2 = 1$. When restricted to genotype frequency distributions in HWE, the factor $(\tilde{\alpha}_1 + \tilde{\alpha}_2)^2 / S_{\text{rec}}(\tilde{\alpha}) = 16(\tilde{\alpha}_1^2/(1 - \tilde{\alpha}_1)(3 + \tilde{\alpha}_1))$ tends to infinity as $\tilde{\alpha}_1 \to 1$; however, it takes the finite value 4 at $\tilde{\alpha}_1 = -1$, so a rare recessive allele may lead to pronounced non-centrality, but with strong sensitivity on the genotype frequency distribution.

From these observations, we can draw the following conclusions regarding the study design. Keeping the case and control samples of comparable size will eliminate the scaling factor, which reacts more strongly to (small) differential genotyping error rates than the non-centrality parameter; however, this maximizes the non-centrality, which grows linearly with overall sample size. The difference in genotyping error rates between the samples is crucial. Hence, efforts to lower the error rate in one sample only, e.g. by very careful genotyping of the case sample but reliance on standard stock, which may be degraded due to storage, for the controls, may be counterproductive.

4. Correcting the test statistic

Our results give a new perspective on the current practice of summary adjustment for errors in the distribution of SNP $p$-values by GC. In the GC method (Devlin and Roeder, 1999), each SNP marker test statistic
Misrepresentation errors in association tests

Fig. 4. The behavior of $(\tilde{\alpha}_1 + \tilde{\alpha}_2)^2 / S_{\text{rec}}(\tilde{\alpha})$ in the non-centrality parameter for the dominant/recessive model. The black line corresponds to HWE and the gray lines to different fixed $\tilde{\alpha}_2$ as in Figure 3.

is corrected by the same scaling factor

$$GC = \frac{\text{median}(T_1, \ldots, T_M)}{0.456}$$

(where $M$ is the number of SNP markers and $T_i$ is the association test statistic for the $i$th marker). It has been observed that the sample size should be taken into account, and an adjustment of the correction factor has been suggested by Freedman and others (2004). The GC value plays an important role in the estimation of the overall inflation of the test statistics in a genome scan, but in the light of our results, its use for the correction of single SNP associations may be misleading. As the distortion parameters $\lambda$ and $\gamma$ depend on the genotype frequency distribution, the correction of the statistic should be based on the individual parameters for each separate SNP; in particular, rare variant SNPs will generally require stronger correction, and should possibly carry lower weight in a multi-SNP analysis.

Hence, we suggest an alternative approach to correction: on the basis of the sample sizes, the genotype frequency distribution and an estimate for the allele misrepresentation rate, the test statistic for the individual SNP marker may be adjusted by both scaling and non-centrality. Note that, due to its linear growth in total sample size, the influence of the non-centrality will tend to dominate over the scaling factor in large-scale studies, and hence the test statistic cannot be effectively corrected by a GC factor only. Thus, the standard GC seems inappropriate, especially in situations where the distortion can be expected to be at least partly attributed to differential genotyping errors, for example, when carefully genotyped case data from an up-to-date chip/platform are used together with singly genotyped stock datasets, such as the 1958...
birth cohort data, as controls. While batch effects within the case and the control samples (when combined from separate chips/platforms) can be dealt with by adding the batch as a regression covariate, our correction will apply to the situation when there is no separate batch variable because cases and controls form different batches.

To illustrate the procedure, we created samples from the human genome data of the HapMap3 and 1000 Genomes projects. We used the datasets for the same 165 individuals (CEU + TSI), for whom 1278013 SNPs have been genotyped in both projects by different methods (Affymetrix array genotyping vs. next generation sequencing). In 326915 SNPs (25.5%), the genotypes differ between projects in at least one individual. Assuming a symmetric allele-based genotyping error model, we estimated the genotyping error frequency $\delta$ as the ratio of allele mismatches to the total number of alleles (330) for each SNP. For 528 SNPs, $\delta > 0.1$ and the errors were of a nature incompatible with the allele-based model; these SNPs were discarded.

From the genotype frequency distribution (in HWE) of the remaining 326387 SNPs, we created random multinomial case and control samples, and then introduced random symmetric allele-based genotyping errors, with the estimated $\delta$ for each SNP, in the control sample, leaving the case sample error-free. Association test statistics and $p$-values for each marker, using the additive disease model, were calculated for the initial clean samples, after the introduction of genotyping errors, and after correction using formulae (2.4), (2.13), (5.3), and (2.28). The corrected $p$-values were obtained by applying a non-central $\chi^2$ distribution with non-centrality parameter $\lambda$ to scaled statistics $T/\gamma$. Nominal GC factors were calculated for all three stages and are shown, along with corresponding qq-plots, in Figure 5. For comparison, we also include the qq-plot resulting from correcting the error-affected $p$-values by the standard GC method.
The qq-plots clearly show that our adjustment for genotyping errors is effective, while standard GC leaves the upper portion of the plot severely distorted. Of course, the error-affected qq-plot would appear flatter if further SNPs without genotyping errors were included. Moreover, the distortion in the example with sample size 5000 may be exaggerated in that our estimate of genotyping errors was obtained from much smaller samples. Also, in practice one will need to rely on more tentative estimates for the misrepresentation error frequencies. Nevertheless, the above method provides a rational approach to the correction of test statistics for random misrepresentation errors and indicates that linear scaling of the test statistic should be complemented with proper use of the non-central distribution.

Finally, we suggest that our application of Theorem 1 can be extended from genotyping errors to more general confounders which may contribute to the inflation of the test statistics in genome-wide association studies, e.g. population stratification.

5. Derivations

5.1 Proof of Theorem 1

Let \((m_0, m_1, m_2)\) be the vector of genotype counts in a sample of size \(m\) from the case population, and \((n_0, n_1, n_2)\) be the vector of genotype counts in a sample of size \(n\) from the control population. Consider the standardized random vectors \(X = (x_0, x_1, x_2)\) and \(Y = (y_0, y_1, y_2)\), where

\[
x_i = \frac{n_i - np_i}{\sqrt{np_i}}, \quad y_i = \frac{m_i - mp_i}{\sqrt{mp_i}}, \quad i \in \{0, 1, 2\};
\]

then clearly \(\sum_{i=0}^2 x_i \sqrt{p_i} = 0\), \(\sum_{i=0}^2 y_i \sqrt{p'_i} = 0\). Hence, \(X\) lies in the 2D plane orthogonal to the vector \(\sqrt{\bar{p}} := (\sqrt{p_0}, \sqrt{p_1}, \sqrt{p_2})\) and \(Y\) lies in the 2D plane orthogonal to the vector \(\sqrt{\bar{p}'} := (\sqrt{p'_0}, \sqrt{p'_1}, \sqrt{p'_2})\). Asymptotically for large \(n\) and \(m\), the distributions of \(X\) and \(Y\) are multivariate standard normal distributions in their respective planes (Cramér, 1946, Section 30.1). Hence, the squared expression in the numerator in (2.1) becomes

\[
\tilde{T} = \left( m \sum_{i=0}^2 w_i n_i - n \sum_{i=0}^2 w_i m_i \right)^2 = \left( mn \sum_{i=0}^2 w_i (p_i - p'_i) + m \sqrt{nW \cdot X - n \sqrt{mW' \cdot Y}} \right)^2,
\]

where \(W = (w_0 \sqrt{p_0}, w_1 \sqrt{p_1}, w_2 \sqrt{p_2})\) and \(W' = (w_0 \sqrt{p'_0}, w_1 \sqrt{p'_1}, w_2 \sqrt{p'_2})\).

We can write \(X = \theta \gamma_1 + \varphi \gamma_2\), where \(\theta\) and \(\varphi\) are vectors such that \(\theta, \varphi \perp \sqrt{\bar{p}}, \theta \perp \varphi\), and \(||\theta||, ||\varphi|| = 1\), and similarly \(Y = \theta' \eta_1 + \varphi' \eta_2\), where \(\theta'\) and \(\varphi'\) are vectors such that \(\theta', \varphi' \perp \sqrt{\bar{p}'}\), \(\theta' \perp \varphi'\), and \(||\theta'||, ||\varphi'|| = 1\), and \(\gamma_1, \gamma_2, \eta_1, \eta_2\) are independent standard normal random variables. Specifically, we can choose \(\theta, \theta', \varphi, \) and \(\varphi'\) as follows:

\[
\theta = \frac{1}{\sqrt{1 - p_2}} \begin{pmatrix} \sqrt{p_1} \\ -\sqrt{p_0} \\ 0 \end{pmatrix}, \quad \theta' = \frac{1}{\sqrt{1 - p_2'}} \begin{pmatrix} \sqrt{p'_1} \\ -\sqrt{p'_0} \\ 0 \end{pmatrix}, \quad \varphi = \sqrt{\bar{p}} \times \theta, \quad \varphi' = \sqrt{\bar{p}'} \times \theta'
\]

(where \(\times\) is the vector product, which is orthogonal on both factors). Then

\[
W \cdot \theta = (w_0 - w_1) \sqrt{\frac{p_0 p_1}{1 - p_2}}, \quad W' \cdot \theta' = (w_0 - w_1) \sqrt{\frac{p'_0 p'_1}{1 - p'_2}}.
\]
and further

\[ W \cdot \varphi = W \cdot (\sqrt{p} \times \theta) = (W \times \sqrt{p}) \cdot \theta \]

\[ = (((w_1 - w_2)\sqrt{p_1}p_2, (w_2 - w_0)\sqrt{p_0}p_2, (w_0 - w_1)\sqrt{p_0}p_1) \cdot \frac{1}{\sqrt{1 - p_2}} \left( \frac{\sqrt{p_1}}{\sqrt{p_0}} \right) \]

\[ = \sqrt{\frac{p_2}{1 - p_2}} ((w_1 - w_2)p_1 - (w_2 - w_0)p_0), \]

and analogously

\[ W' \cdot \varphi' = \sqrt{\frac{p_2'}{1 - p_2'}} ((w_1 - w_2)p_1' - (w_2 - w_0)p_0'). \]

Since \(\gamma_1, \gamma_2, \eta_1, \eta_2 \sim N(0, 1)\), their linear combination

\[ \tilde{\xi} = m\sqrt{n}(W \cdot \theta \gamma_1 + W \cdot \varphi \gamma_2) - n\sqrt{m}(W' \cdot \theta \eta_1 + W' \cdot \varphi' \eta_2) \sim N(0, mn\beta), \]

where

\[ \beta = m((W \cdot \theta)^2 + (W \cdot \varphi)^2) + n((W' \cdot \theta')^2 + (W' \cdot \varphi')^2) \]

\[ = m(w_0 - w_1)^2 \frac{p_1p_0}{1 - p_2} + m \frac{p_2}{1 - p_2} ((w_1 - w_2)p_1 - (w_2 - w_0)p_0)^2 \]

\[ + n(w_0 - w_1)^2 \frac{p_1'p_0'}{1 - p_2'} + n \frac{p_2'}{1 - p_2'} ((w_1 - w_2)p_1' - (w_2 - w_0)p_0')^2 \]

(see (2.6)). Now \(\tilde{T}\) can be written as

\[ \tilde{T} = \left( mn \sum_{i=0}^{2} w_i (p_i - p_i') + \tilde{\xi} \right)^2 = mn\beta (\xi + \sqrt{\lambda})^2, \]

where \(\xi \sim N(0, 1)\) and

\[ \lambda = \frac{mn}{\beta} \left( mn \sum_{i=0}^{2} w_i (p_i - p_i') \right)^2. \]

Thus, \(\tilde{T}/mn\beta\) follows a non-central \(\chi^2\) distribution with \(df = 1\) and non-centrality parameter \(\lambda\). Hence, the normalized test statistic satisfies

\[ T = \frac{N\tilde{T}}{mn(N\sum_{i=0}^{2} w_i^2 N_i - (\sum_{i=0}^{2} w_i N_i)^2)} = \gamma (\xi + \sqrt{\lambda})^2, \]

with

\[ \gamma = \frac{N\beta}{N \sum_{i=0}^{2} w_i^2 N_i - (\sum_{i=0}^{2} w_i N_i)^2}. \]

Equation (2.5) now follows by estimating \(N_i \sim N \hat{p}_i\) and applying Slutski’s lemma.
5.2 Proof of Lemma 1

We start from (2.6) and use the key observation that
\[
(w_0 - w_1)^2 p_1 p_0 + p_2 ((w_1 - w_2) p_1 - (w_2 - w_0) p_0)^2
= (1 - p_2)(w_0^2 p_0 + w_1^2 p_1 + w_2^2 p_2 - (w_0 p_0 + w_1 p_1 + w_2 p_2)^2)
\]
(5.1)
(for any \(w_0, w_1, w_2\) and for \(p_0, p_1, p_2\) such that \(p_0 + p_1 + p_2 = 1\)) to rewrite the expression for \(\beta\). Expressed in the \(\alpha\) coordinates, then,
\[
\beta = mS(\alpha) + nS(\alpha').
\]
(5.2)

Since, in the case of equal genotype frequency distribution in the case and control populations, the scaling factor is equal to 1,
\[
1 = \frac{mS(\hat{\alpha}) + nS(\hat{\alpha})}{N(\sum_{i=0}^{2} w_i^2 \hat{p}_i - (\sum_{i=0}^{2} w_i \hat{p}_i)^2)},
\]
and it follows that \(\sum_{i=0}^{2} w_i^2 \hat{p}_i - (\sum_{i=0}^{2} w_i \hat{p}_i)^2 = S(\hat{\alpha})\). Thus, we can deduce that generally
\[
\gamma = \frac{mS(\alpha) + nS(\alpha')}{NS(\hat{\alpha})}.
\]
(5.3)

Now write \(\alpha = \hat{\alpha} + h, \alpha' = \hat{\alpha} + h'\). Then, by the mean value theorem, there is \(t \in (0, 1)\) such that
\[
S(\hat{\alpha} + h) = S(\hat{\alpha}) + \nabla S(\hat{\alpha} + th) \cdot h \approx S(\hat{\alpha}) + \nabla S(\hat{\alpha}) \cdot h,
\]
for small \(|h|\); here \(\nabla S(\hat{\alpha}) = \left(\frac{\partial S(\hat{\alpha})}{\partial \alpha}, \frac{\partial S(\hat{\alpha})}{\partial \alpha'}\right)\) is the gradient of the function \(S\). Equation (2.10) now follows, as
\[
h = (m/N)(\alpha - \alpha') \quad \text{and} \quad h' = -(n/N)(\alpha - \alpha'),
\]
so
\[
\gamma \approx 1 + \frac{\nabla S(\hat{\alpha})}{S(\hat{\alpha})} \cdot \frac{mh + nh'}{N} = 1 + \frac{m - n \nabla S(\hat{\alpha})}{N} \cdot (\alpha - \alpha').
\]

5.3 Proof of Lemma 2

In view of (2.9) and (5.2), the non-centrality parameter, expressed in terms of the coordinates \(\alpha\), is
\[
\lambda = \frac{mn(\frac{1}{4}(w_0 - w_2)(\alpha_1 - \alpha_1') + \frac{1}{4}(w_0 - 2w_1 + w_2)(\alpha_2 - \alpha_2'))^2}{mS(\alpha) + nS(\alpha')}.
\]

The denominator is not exactly equal to \(NS(\hat{\alpha})\), where \(\hat{\alpha} = (m/N)\alpha + (n/N)\alpha'\), due to the presence of non-linear (quadratic) terms in (2.11), as noted. For \(j, k \in \{1, 2\}\),
\[
\tilde{\alpha}_j \tilde{\alpha}_k = \frac{m^2}{N^2} \alpha_j \alpha_k + \frac{n^2}{N^2} \alpha_j' \alpha_k' + \frac{mn}{N^2} (\alpha_j \alpha_k' + \alpha_j' \alpha_k) = \frac{m}{N} \alpha_j \alpha_k + \frac{n}{N} \alpha_j' \alpha_k' - \frac{mn}{N^2} (\alpha_j - \alpha_j')(\alpha_k - \alpha_k'),
\]
and hence, by (2.11), \(S(\hat{\alpha}) = (m/N)S(\alpha) + (n/N)S(\alpha') + (mn/N^2)q\), where
\[
q := \left(\frac{1}{4}(w_0 - w_2)(\alpha_1 - \alpha_1') + \frac{1}{4}(w_0 - 2w_1 + w_2)(\alpha_2 - \alpha_2')^2\right).
\]
Thus, 

\[ \lambda = \frac{mnq}{N(S(\ddot{\alpha}) - (mn/N^2)q)} = N\left(\frac{1}{1 - (mn/N^2)(q/S(\ddot{\alpha}))} - 1\right) \approx N\frac{mnq}{N^2 S(\ddot{\alpha})}, \]

where the approximation is good if \( \alpha - \alpha' \) is small; this is (2.23).

**Acknowledgments**

It is a pleasure to thank Prof. N. Leonenko for helpful comments. *Conflict of Interest*: None declared.

**Funding**

M.G.’s research was supported by a Cardiff University Richard Whipp studentship.

**References**


[Received May 16, 2013; revised September 26, 2013; accepted for publication October 2, 2013]