frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 01 April 2014
doi: 10.3389/fgene.2014.00062

Evaluating the impact of genotype errors on rare variant tests of association

Kaitlyn Cook[1], Alejandra Benitez[2], Casey Fu[3] and Nathan Tintle[4]*

[1] Department of Mathematics, Carleton College, Northfield, MN, USA
[2] Department of Applied Mathematics, Brown University, Providence, RI, USA
[3] Department of Mathematics, Massachusetts Institute of Technology, Boston, MA, USA
[4] Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, IA, USA

Edited by:
Joanna Biernacka, Mayo Clinic, USA

Reviewed by:
Zheyang Wu, Worcester Polytechnic Institute, USA
Nicholas Larson, Mayo Clinic, USA

***Correspondence:**
Nathan Tintle, Department of Mathematics, Statistics and Computer Science, Dordt College, 498 4th Ave. NE, Sioux Center, IA 51250, USA
e-mail: nathan.tintle@dordt.edu

The new class of rare variant tests has usually been evaluated assuming perfect genotype information. In reality, rare variant genotypes may be incorrect, and so rare variant tests should be robust to imperfect data. Errors and uncertainty in SNP genotyping are already known to dramatically impact statistical power for single marker tests on common variants and, in some cases, inflate the type I error rate. Recent results show that uncertainty in genotype calls derived from sequencing reads are dependent on several factors, including read depth, calling algorithm, number of alleles present in the sample, and the frequency at which an allele segregates in the population. We have recently proposed a general framework for the evaluation and investigation of rare variant tests of association, classifying most rare variant tests into one of two broad categories (length or joint tests). We use this framework to relate factors affecting genotype uncertainty to the power and type I error rate of rare variant tests. We find that non-differential genotype errors (an error process that occurs independent of phenotype) decrease power, with larger decreases for extremely rare variants, and for the common homozygote to heterozygote error. Differential genotype errors (an error process that is associated with phenotype status), lead to inflated type I error rates which are more likely to occur at sites with more common homozygote to heterozygote errors than vice versa. Finally, our work suggests that certain rare variant tests and study designs may be more robust to the inclusion of genotype errors. Further work is needed to directly integrate genotype calling algorithm decisions, study costs and test statistic choices to provide comprehensive design and analysis advice which appropriately accounts for the impact of genotype errors.

Keywords: SKAT, gene-based, genotype uncertainty, misclassification, dosage

INTRODUCTION

Over the past 5 years, numerous gene-based tests of rare variant association have been proposed. Several summaries and reviews of these methods are available (Asimit and Zeggini, 2010; Bansal et al., 2010; Cooper and Shendure, 2011; Dering et al., 2011; Gibson, 2012). The majority of these tests accumulate evidence of genotype-phenotype association across multiple single nucleotide variants (SNVs) within a gene either by first collapsing genotypes of some or all of the SNVs (collapsing; burden; length tests) (Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009; Han and Pan, 2010; Li et al., 2010; Morris and Zeggini, 2010; Zawistowski et al., 2010; Feng et al., 2011; Sul et al., 2011; Zhang et al., 2011; Dai et al., 2012) or by aggregating (e.g., summing) individual variant association statistics across all SNVs within a gene (variance components; joint tests) (Li and Leal, 2008; Basu and Pan, 2011; Ionita-Laza et al., 2011; Lin and Tang, 2011; Neale et al., 2011; Pan and Shen, 2011; Wu et al., 2011) (see Liu et al., 2013 for details).

A recent paper (Liu et al., 2013) introduced the terminology "length" and "joint" tests to illustrate a geometric interpretation of the gene-based, rare variant, test statistic formulation for case-control studies. Most rare variant test statistics can be written as functions of the generally stated Length (L_p) or Joint (J_p) test statistics as defined immediately below:

$$\text{General Length Test Statistic, } L_p = \left(\sum_{i=1}^{m}\left|\frac{c_i^+}{2N^+}\right|^p\right)^{1/p} - \left(\sum_{i=1}^{m}\left|\frac{c_i^-}{2N^-}\right|^p\right)^{1/p}$$

$$\text{General Joint Test Statistic, } J_p = \left(\sum_{i=1}^{m}\left|\frac{c_i^+}{2N^+} - \frac{c_i^-}{2N^-}\right|^p\right)^{1/p}$$

where, m is the number of SNVs within the gene, N^+ and N^- indicate the sample sizes of the cases and controls, respectively, c_i^+ and c_i^- indicate the observed number of minor alleles at variant i, within the case and control samples, respectively, and p reflects the choice of L^p norm. To date, most published length tests use $p = 1$, while most joint tests use $p = 2$. Thus, length tests compare

the magnitudes (lengths) of the m-dimensional minor allele frequency (MAF) vectors between cases and controls by taking the L^p norms of the vectors, with larger differences in length indicating stronger evidence of genotype-phenotype association. Joint tests compare both the lengths of the case-control vectors, as well as the angle between the vectors (evidence for association increases as the magnitude of the angle between the vectors increases). This geometric framework provides the basis for theoretical evaluation of test behavior—moving beyond comparison of rare variant test statistic behavior solely by simulation.

Genotype errors occur when calling algorithms misidentify an individual's genotype (e.g., an individual who is actually AA is identified as AT). To date, the majority of evidence showing the detrimental effects of genotype error on this new class of rare variant tests has been based on simulation results. In particular, simulation of genotype data followed by simulation of genotype errors on those genotypes finds that the power of some specific length tests decreases—sometimes dramatically—in the presence of non-differential (independent of case-control status) genotyping errors. These power declines can be particularly large for errors misclassifying the common homozygote as the heterozygote, even when the error rate is relatively low (Powers et al., 2011). Relatedly, for some specific joint and length tests, the type I error rate increases above nominal levels in the presence of differential genotyping errors, even at low error rates. The magnitude of the type I error inflation increases further as the sample size, number of rare variants or relative difference in case-control error rates at the site increases, or as the MAF of variants decreases. Similarly, these effects are enhanced for errors from the common homozygote to the heterozygote (Mayer-Jochimsen et al., 2013). At error levels observed in sequence and imputed data for rare variants, the effects of errors on power and type I error can be measurable (Awadalla et al., 2010; Ilie et al., 2011; Nielsen et al., 2011; Rogers et al., 2014). These findings are similar to findings about the effects of both non-differential (Gordon et al., 2002, 2004; Kang et al., 2004a,b; Ahn et al., 2007) and differential (Moskvina et al., 2006; Ahn et al., 2009) errors when analyzed with single marker test statistics.

While such findings based on simulation are useful, their utility in providing a deeper understanding of the reasons why errors can be so detrimental to power and type I error is limited. In this paper, we use the geometric framework as a platform for deeper understanding of the mechanisms by which genotype errors impact rare variant tests of association. In particular, we use the geometric framework to gain greater insights into the relative impact of different types of genotype errors (homozygote to heterozygote, or vice versa), MAF, the differential or non-differential nature of the genotype errors and choice of rare variant test statistic on the power and type I error rate of length and joint tests.

METHODS

DISTRIBUTIONS, POWER AND TYPE I ERROR RATES OF GENE-BASED RARE VARIANT TEST STATISTICS

We start by noting that $c_i^+ \sim \text{Binom}(2N^+, f_i^+)$ and $c_i^- \sim \text{Binom}(2N^-, f_i^-)$, where f_i^+ and f_i^- are the MAFs in the cases and controls, respectively. For a low prevalence disease, f_i^- will be approximately equal to the population MAF, f_i. We are often interested in the scaled difference of these counts, $D_i = \frac{c_i^+}{2N^+} - \frac{c_i^-}{2N^-}$. Applying basic distribution theory yields: $\mu_{D_i} = f_i^+ - f_i^-$ and $\sigma^2_{D_i} = \frac{1}{2N^+}\left(f_i^+\left(1-f_i^+\right)\right) + \frac{1}{2N^-}\left(f_i^-\left(1-f_i^-\right)\right)$. For all rare variant tests considered in this manuscript, the null hypothesis is that $f_i^+ = f_i^-$ for all i. We start by stating assumptions needed for our analytic evaluation.

Assumptions

(1) Let $\varepsilon_{01,i}$ represent the probability that the major allele is misclassified as the minor allele at site i, and let $\varepsilon_{10,i}$ represent the probability that the minor allele is misclassified as the major allele at site i. We can write the population MAF in both the cases and controls as a function of the true population minor allele frequencies and the error rate. In particular

$$f_i^{+*} = f_i^+\left(1-\varepsilon_{10,i}\right) + \left(1-f_i^+\right)\left(\varepsilon_{01,i}\right)$$
$$f_i^{-*} = f_i^-\left(1-\varepsilon_{10,i}\right) + \left(1-f_i^-\right)\left(\varepsilon_{01,i}\right)$$

where we assume that each allele has an equal chance of being misclassified and that likelihood of errors in the cases is the same as in the controls (non-differential errors). Differential errors follow a similar definition and assumption, except that the change of errors is different in cases and controls.

(2) In all proofs and simulations, we assume that the allele frequencies in the population follow Hardy-Weinberg Equilibrium.

(3) In all proofs and simulations, we assume that the variant sites within the gene are not in linkage disequilibrium (LD) as we have done in previous work (Mayer-Jochimsen et al., 2013). See the Discussion for implications.

(4) When evaluating the impact of genotype errors on J_2^* (Impact of Genotype Errors on the Type I Error and Power of J_2^*) and J_∞^* (Impact of Genotype Errors on the Type I Error and Power of J_∞), as well as when providing analytic power and sample size estimates (Asymptotic Power Formulas for L_1^* and J_2^*), we explore the impact of genotype errors on the distributions of c_i^+ and c_i^- as approximated by Normal distributions. In particular, that $c_i^+ \dot{\sim}$ Norm $\left(N^+p_i^+, N^+p_i^+\left(1-p_i^+\right)\right)$ and $c_i^- \dot{\sim}$Norm $\left(N^-p_i^-, N^-p_i^-\left(1-p_i^-\right)\right)$. It follows directly that $D_i \dot{\sim}$Norm $\left(\mu_{D_i}, \sigma^2_{D_i}\right)$, and, thus, $\frac{D_i^2}{\sigma^2_{D_i}} \dot{\sim} \chi^2_{1,\lambda}$ where $\lambda = \left(\frac{\mu_{D_i}}{\sigma_{D_i}}\right)^2$ is the non-centrality parameter. We evaluate robustness to this assumption as part of our simulation study (see Quality of Asymptotic Power and Type I Error Predictions).

Impact of genotype errors on the type I error and power of L_1^*

When $p = 1$, we can write

$$L_1 = \left(\sum_{i=1}^m \left|\frac{c_i^+}{2N^+}\right|^p\right)^{\frac{1}{p}} - \left(\sum_{i=1}^m \left|\frac{c_i^-}{2N^-}\right|^p\right)^{\frac{1}{p}}$$
$$= \sum_{i=1}^m \frac{c_i^+}{2N^+} - \sum_{i=1}^m \frac{c_i^-}{2N^-} = \sum_{i=1}^m \left(\frac{c_i^+}{2N^+} - \frac{c_i^-}{2N^-}\right) = \sum_{i=1}^m D_i$$

where we have dropped the absolute value since the observed minor allele counts will always be positive. Thus, $\mu(L_1) = \sum_{i,=1}^{m} \mu_{D_i}$ and $\sigma^2(L_1) = \sum_{i=1}^{m} \sigma^2_{D_i}$ when variant sites are independent (no LD).

When genotype errors are present (indicated by $*$), similar arguments hold. The distribution of $L_1^* = \sum_{i=1}^{m} \left(\frac{c_i^{+*}}{2N^+} - \frac{c_i^{-*}}{2N^-} \right) = \sum_{i=1}^{m} D_i^*$ has mean $\mu\left(L_1^*\right) = \sum_{i=1}^{m} \left(f_i^{+*} - f_i^{-*}\right) = \sum_{i=1}^{m} \mu_{D_i}^*$ and $\sigma^2\left(L_1^*\right) = \sum_{i=1}^{m} \sigma^2_{D_i^*} + 2\sum_{i<j} Cov\left(D_i^*, D_j^*\right)$ where, $\sum_{i=1}^{m} \sigma^2_{D_i^*} = \sum_{i=1}^{m} \left(\frac{1}{2N^+}\right)\left(f_i^{+*}\left(1 - f_i^{+*}\right)\right) + \frac{1}{2N^-}\left(f_i^{-*}\left(1 - f_i^{-*}\right)\right)$. As above, when variant sites are independent (no LD) $\sum_{i<j} Cov\left(D_i^*, D_j^*\right) = 0$.

Non-differential genotype errors and the type I error rate. When the null hypothesis is true, it is straightforward to see that $\mu(L_1) = 0$. When there are non-differential genotype errors $\mu\left(L_1^*\right) = \sum_{i=1}^{m} \mu_{D_i}^* = 0$ since $f_i^{+*} - f_i^{-*} = f_i^+\left(1 - \varepsilon_{10,i}\right) + \left(1 - f_i^+\right)\left(\varepsilon_{01,i}\right) - f_i^-\left(1 - \varepsilon_{10,i}\right) - \left(1 - f_i^-\right)\left(\varepsilon_{01,i}\right) = \left(1 - \varepsilon_{10,i}\right)\left(f_i^+ - f_i^-\right) - \left(\varepsilon_{01,i}\right)\left(f_i^+ - f_i^-\right) = 0$ for all *i*. Bross (1954) proved that estimates of the variance of D_i^* are unbiased in the presence of non-differential misclassification errors for both small and large samples. Thus, linear scaled sums of these estimates (as in L_1^*) are also unbiased, resulting in a test which controls the Type I error rate.

Non-differential genotype errors and power. Given the fact that the Type I error is maintained in the presence of non-differential errors, we now explore the impact of non-differential genotype errors on the power of L_1^*. To do this we start by noting that $\mu\left(L_1^*\right)$ can be written as:

$$\begin{aligned}\mu\left(L_1^*\right) &= \sum_{i=1}^{m}\left(\left(1 - \varepsilon_{10,i}\right)\left(f_i^+ - f_i^-\right) - \left(\varepsilon_{01,i}\right)\left(f_i^+ - f_i^-\right)\right) \\ &= \sum_{i=1}^{m}\left(\left(1 - \varepsilon_{10,i} - \varepsilon_{01,i}\right)\left(f_i^+ - f_i^-\right)\right)\end{aligned}$$

Thus, in the presence of non-differential genotype errors $\left(\varepsilon_{10,i} > 0,\ \varepsilon_{01,i} > 0\right)$, $\mu\left(D_i^*\right) = \left(1 - \varepsilon_{10,i} - \varepsilon_{01,i}\right)\left(f_i^+ - f_i^-\right) < \mu\left(D_i\right) = \left(f_i^+ - f_i^-\right)$, moving $\mu\left(D_i^*\right)$ closer to 0 (which is our expectation under the null hypothesis), with both $\varepsilon_{10,i}$ and $\varepsilon_{01,i}$ contributing equally to the shift of the mean of the alternative distribution closer to 0. When $f_i^+ \geq f_i^-$ for all i (all variants are non-causal or risk increasing), then $\left(L_1^*\right) < \mu(L_1) = \sum_{i=1}^{m}\left(f_i^+ - f_i^-\right)$, moving $\mu\left(L_1^*\right)$ closer to 0 (which is our expectation under the null hypothesis). When at least one $f_i^+ < f_i^-$ (at least one protective variant), then moving $\mu\left(D_i^*\right)$ closer to 0, will increase the overall value of $\mu\left(L_1^*\right)$ since there will be less "cancellation" occurring between risk increasing and risk reducing variants when computing the test statistic.

We will now show that in general, $\sigma^2\left(L_1^*\right) > \sigma^2(L_1)$. Recall that $\sigma^2(L_1) = \sum_{i=1}^{m} \sigma^2_{D_i}$ and that $\sigma^2_{D_i} = \frac{1}{2N^+}\left(f_i^+\left(1 - f_i^+\right)\right) + \frac{1}{2N^-}\left(f_i^-\left(1 - f_i^-\right)\right)$, with similar relationships true when errors are present (denoted by $*$). To show that $\sigma^2\left(L_1^*\right) > \sigma^2(L_1)$ it is sufficient to show that $\sigma^2_{D_i^*} > \sigma^2_{D_i}$ for all *i*, an inequality which is true when both $f_i^{+*}\left(1 - f_i^{+*}\right) > f_i^+\left(1 - f_i^+\right)$ and $f_i^{-*}\left(1 - f_i^{-*}\right) > f_i^-\left(1 - f_i^-\right)$.

To see that $f_i^{+*}\left(1 - f_i^{+*}\right) > f_i^+\left(1 - f_i^+\right)$ is true in most cases consider that $0 < f_i^+ < 0.5$ and, thus, in most situations, $0 < f_i^{+*} < 0.5$ because we have defined f as the MAF. Thus, $f_i^{+*}\left(1 - f_i^{+*}\right) > f_i^+\left(1 - f_i^+\right)$ when $f_i^{+*} > f_i^+$, an inequality that will be true in most practical cases, as shown below

$$\begin{aligned} f_i^{+*} &> f_i^+ f_i^+ - f_i^+\left(\varepsilon_{10,i} + \varepsilon_{01,i}\right) + \varepsilon_{01,i} > f_i^+ \\ \varepsilon_{01,i} &> f_i^+\left(\varepsilon_{10,i} + \varepsilon_{01,i}\right) \\ \varepsilon_{01,i} &> \varepsilon_{10,i}\left(\frac{f_i^+}{1 - f_i^+}\right) \approx \varepsilon_{10,i} f_i^+ \end{aligned}$$

Where we make use of the fact that for rare alleles f is quite small, and so, unless the value of $\varepsilon_{10,i}$ is many orders of magnitude larger than $\varepsilon_{01,i}$ the inequality will be true. Similar arguments hold when showing $f_i^{-*}\left(1 - f^{-*}\right) > f_i^-\left(1 - f_i^-\right)$.

It is also important to note that the increases to $\sigma^2\left(L_1^*\right)$ due to effect of ε_{01} are substantially more than the effects of ε_{10}. This can be seen by observing that $f_i^{+*} = f_i^+ - f_i^+\left(\varepsilon_{10,i} + \varepsilon_{01,i}\right) + \varepsilon_{01} = f_i^+\left(1 - \varepsilon_{10,i}\right) + \left(1 - f_i^+\right)\varepsilon_{01,i}$. Since f_i is small, increases in values of $\varepsilon_{01,i}$ increase variance, while increases to $\varepsilon_{10,i}$ decrease variance, but substantially less. Increases in variance, combined with shifting of the mean of the alternative distribution toward the mean of the null distribution, will result in decreases in power. The only exception is in cases where genotype errors occur on protective variants, which, as shown in the previous section, may mitigate power loss to some extent. Our evaluation shows that the relative effects of $\varepsilon_{01,i}$ on power loss are more than power loss driven by $\varepsilon_{10,i}$.

Differential genotype errors and the type I error rate. Differential genotype errors occur when the genotype error rate in the cases (ε^+) is different than it is in the controls (ε^-). In this case, it follow directly from earlier arguments that,

$$\begin{aligned}\mu\left(L_1^*\right) = \sum_{i=1}^{m} & f_i^+\left(1 - \varepsilon_{10,i}^+\right) + \left(1 - f_i^+\right)\varepsilon_{01,i}^+ - \left(f_i^-\left(1 - \varepsilon_{10,i}^-\right)\right. \\ & \left. + \left(1 - f_i^-\right)\varepsilon_{01,i}^-\right)\end{aligned}$$

Where, $+$ and $-$ indicate the different genotype error rates in the cases and controls, respectively. We note that when the null hypothesis is true, the following is true for each variant *i*.

$$\begin{aligned} & f_i\left(1 - \varepsilon_{10,i}^+\right) + \left(1 - f_i\right)\varepsilon_{01,i}^+ - \left(f_i\left(1 - \varepsilon_{10,i}^-\right) + \left(1 - f_i\right)\varepsilon_{01,i}^-\right) \\ & = f_i\left(\varepsilon_{10,i}^- - \varepsilon_{10}^+, i\right) + \left(1 - f_i\right)\left(\left(\varepsilon_{01,i}^+ - \varepsilon_{01,i}^-\right)\right) \end{aligned}$$

This quantity is not zero in the presence of differential genotype errors. This means that when differential genotype errors are

present $\mu\left(L_1^*\right) \neq 0$, which is sufficient to show that the resulting type I error rate will typically no longer be the nominal value. The exception is when the effects of differential genotype errors cancel out, which can occur if genotype error rates are larger in the cases for some variants, and larger in the controls for other variants. Examining the equation further suggests that in general the larger the difference in error rates, the larger the type I error rate will be, with differences in the $\varepsilon_{01,i}$ error rates contributing more to inflation in the type I error rate than differences in the $\varepsilon_{10,i}$ error rates, since differences in $\varepsilon_{10,i}$ only impact $\mu\left(L_1^*\right)$ through a term which is multiplied by f, typically a small quantity. Sites with higher MAF (larger f_i) will tend to increase the value of $\mu\left(L_1^*\right)$ more, however, the impact is scaled by the difference in case and control genotyping error rates, which will typically be a small quantity, meaning that the overall impact of f_i on the value of $\mu\left(L_1^*\right)$ is quite minimal.

Much of the argument about the relationship between $\sigma^2\left(L_1^*\right)$ and $\sigma^2\left(L_1\right)$ in the presence of differential genotype errors follows directly from arguments made in the previous section (Power) when examining non-differential errors. To show that, in general, $\sigma^2\left(L_1^*\right) > \sigma^2\left(L_1\right)$ it is sufficient to show that $\sigma^{2*}_{D_i} > \sigma^2_{D_i}$, an inequality which is true when both $f_i^{+*}\left(1-f_i^{+*}\right) > f_i^{+}\left(1-f_i^{+}\right)$ and $f_i^{-*}\left(1-f_i^{-*}\right) > f_i^{-}\left(1-f_i^{-}\right)$. It is typically true that $f_i^{+*}\left(1-f_i^{+*}\right) > f_i^{+}\left(1-f_i^{+}\right)$ because $\varepsilon_{01,i}^{+} > \varepsilon_{10,i}^{+}\left(\frac{f_i^+}{1-f_i^+}\right) \approx \varepsilon_{10}^{+} f_i^{+}$, with similar arguments holding in the controls—even when the error rates in the controls are different than in the cases. Thus, once again, the effect of ε_{01} on the variance is substantially more than the effect of ε_{10}. Since increases in variance will result in increases in the type I error rate, ε_{01} has a potentially large impact on the type I error rate, while ε_{10} has less impact (really only impacting $\mu\left(L_1^*\right)$.

Impact of genotype errors on the type I error and power of J_2^*

When $p = 2$, we can write

$$J_2 = \left(\sum_{i=1}^{m}\left|\frac{c_i^+}{2N^+} - \frac{c_i^-}{2N^-}\right|^2\right)^{1/2} = \sqrt{\sum_{i=1}^{m}\left(\frac{c_i^+}{2N^+} - \frac{c_i^-}{2N^-}\right)^2}$$

$$= \sqrt{\sum_{i=1}^{m}\left(D_i\right)^2}$$

Thus, $\mu\left(J_2^2\right) = \sum_{i=1}^{m}\mu_{D_i^2} = \sum_{i=1}^{m}\left(f_i^+ - f_i^-\right)^2$ and $\sigma^2\left(J_2^2\right) = \sum_{i=1}^{m}\sigma^2_{D_i^2} + 2\sum_{i<j} Cov\left(D_i^2, D_j^2\right)$, where $Cov\left(D_i^2, D_j^2\right)$ is the covariance between the differences in case and control allele counts at variant i and j, and, thus, is an indirect measure of LD. When variant sites are independent (no LD) $Cov(D_i^2, D_j^2) = 0$.

When genotype errors are present (indicated by $*$), similar arguments hold. The distribution of $J^2_{2^*} = \sum_{i=1}^{m}\left(\frac{c_i^{+*}}{2N^+} - \frac{c_i^{-*}}{2N^-}\right)^2 = \sum_{i=1}^{m}\left(D_i^*\right)^2$ has mean $\mu\left(J^2_{2^*}\right) = \sum_{i=1}^{m}\mu^*_{D_i^2}$ and $\sigma^2\left(J^2_{2^*}\right) = \sum_{i=1}^{m}\sigma^2_{D_i^{*2}} + 2\sum_{i<j} Cov\left(D_i^{*2}, D_j^{*2}\right)$. As above, when variant sites are independent (no LD) $\sum_{i<j} Cov\left(D_i^{*^2}, D_j^{*^2}\right) = 0$.

Insights into the direction and pattern of effects of genotype errors on J_2^2 are aided by utilizing χ^2 distributions. As noted in Distributions, Power and Type I Error Rates of Gene-based Rare Variant Test Statistics (Assumptions), $\frac{D_i^2}{\sigma^2_{D_i}} \dot{\sim} \chi^2_{1,\lambda}$ where $\lambda = \left(\frac{\mu_{D_i}}{\sigma_{D_i}}\right)^2$ is the non-centrality parameter. It follows directly that $J^2_{2,scaled} = \sum_{i=1}^{m}\left(\frac{D_i}{\sigma_{D_i}}\right)^2 \dot{\sim} \chi^2_{m,\lambda}$ where $\lambda = \sum_{i=1}^{m}\left(\frac{\mu_{D_i}}{\sigma_{D_i}}\right)^2$ is the non-centrality parameter. Our analyses focus on the behavior of $J^2_{2,scaled}$ which can be interpreted as a MAF-variant weighted version of J_2 in the spirit of Madsen and Browning (2009) and others.

Non-differential genotype errors and the type I error rate. When the null hypothesis is true, $\lambda = \sum_{i=1}^{m}\left(\frac{\mu_{D_i}}{\sigma_{D_i}}\right)^2 = \sum_{i=1}^{m}\left(\frac{f_i^+ - f_i^-}{\sigma_{D_i}}\right)^2 = 0$. This is also true in the presence of non-differential genotype errors since, as shown in Non-differential Genotype Errors and the Type I Error Rate, $f_i^{+*} - f_i^{-*} = 0$ for all i, and so $\lambda^* = \sum_{i=1}^{m}\left(\frac{f_i^{+*} - f_i^{-*}}{\sigma^*_{D_i}}\right)^2 = 0$. Thus, the type I error rate is maintained since the distribution of $J^2_{2,scaled}$ is the same with or with non-differential genotype errors when the null hypothesis is true.

Non-differential genotype errors and power. When the alternative hypothesis is true, $f_i^+ \neq f_i^-$ for at least one i, and the non-centrality parameter, $\lambda = \sum_{i=1}^{m}\left(\frac{f_i^+ - f_i^-}{\sigma_{D_i}}\right)^2$ will be greater than 0. Furthermore, the power of $J^*_{2,scaled}$ (non-differential genotype errors) will be lower than $J_{2,scaled}$ (no errors) if $\lambda^* < \lambda$. As shown in 2.1.1.2, $\sigma^*_{D_i} > \sigma_{D_i}$, and so we can show that, in general, $\lambda^* < \lambda$ if $\left(f_i^{+*} - f_i^{-*}\right)^2 < \left(f_i^+ - f_i^-\right)^2$ is also true, which is the case since $\left(f_i^{+*} - f_i^{-*}\right)^2 = \left(1 - \varepsilon_{10,i} - \varepsilon_{01,i}\right)^2\left(f_i^+ - f_i^-\right)^2 < \left(f_i^+ - f_i^-\right)^2$. Furthermore, we can conclude that the impact of the errors follows the same pattern as for L_1^*, namely that the relative effects of $\varepsilon_{01,i}$ on power loss are more than power loss driven by $\varepsilon_{10,i}$.

Differential genotype errors and the type I error rate. When differential genotype errors are present, then there may be inflation of the type I error rate. This inflation occurs because, due to differential genotype errors, the non-centrality parameter, $\lambda^* = \sum_{i=1}^{m}\left(\frac{f_i^{+*} - f_i^{-*}}{\sigma^*_{D_i}}\right)^2$, is no longer, necessarily, zero. This result follows directly from the fact that f_i^{+*} may not equal f_i^{-*} for all i, since $f_i^{+*} - f_i^{-*} = f_i\left(\left(\varepsilon_{10,i}^- - \varepsilon_{10}^+, i\right) + \left(\varepsilon_{01,i}^- - \varepsilon_{01,i}^+\right)\right) + \left(\varepsilon_{01,i}^+ - \varepsilon_{01,i}^-\right)$ will not necessarily equal 0. even when $f_i^+ = f_i^- = f$. Following directly from Differential Genotype Errors and the Type I Error Rate, the case-control differences in the $\varepsilon_{01,i}$ error

rates will inflate the type I error rate more than case-control differences in the $\varepsilon_{10,i}$.

Impact of genotype errors on the type I error and power of J_∞

Liu et al. (2013) showed that, while under-explored in the literature, the choice of norm for both Length and Joint statistics had practical implications. In particular, as the value of the norm increases, gene-based rare variant tests are increasingly robust to the inclusion of non-causal variants (i.e., variants for which $f_i^+ = f_i^-$). To explore how the impact of genotype errors may vary based on choice of norm, we consider using the infinity norm on a joint test. Following Liu et al. (2013), we let, $J_\infty = \underset{1 \le i \le m}{argmax} \left(\frac{c_i^+}{2N^+} - \frac{c_i^-}{2N^-} \right)$.

Non-differential genotype errors and the type I error rate. Results earlier showed that the Type I error rate is maintained because when non-differential genotype errors are present $\mu_{D_i} = \mu^*_{D_i} = 0$, and that estimates of the variance of D_i^* are also unbiased resulting in a test (J^*_∞) which maintains the type I error rate since the distribution at each variant site maintains the type I error rate and the variant sites are independent of each other.

Non-differential genotype errors and power. When there are non-differential genotyping errors, the power will be reduced because $\mu^*_{D_i} < \mu_{D_i}$. However, because J_∞ focuses only on a single variant site (namely, the site, i, showing the largest difference in minor allele frequencies), the impact of errors on power relative to L_1 and J_2 may be lessened because the power loss does not accumulate across variant sites when genotype errors are evenly distributed across variant sites. However, if non-differential genotype errors are focused only on the sites with the largest true difference in minor allele counts power loss may be substantial. The relative impact of ε_{01} and ε_{10} follow patterns described earlier (Non-differential Genotype Errors and Power).

Differential genotype errors and the type I error rate. When differential genotyping errors are present, the type I error rate will increase because $\mu^*_{D_i} \neq 0$. As with power, the impact on type I error may be lessened because the type I error effects do not accumulate across variant sites when genotype errors are evenly distributed across variant sites. However, if the differential genotype errors are contained only on a single variant—inducing the largest observed differences in minor allele frequencies—the type I error rate may inflate above levels observed for L_1 and J_2. The relative impact of ε_{01}, ε_{10} and f follow patterns described in Differential Genotype Errors and the Type I Error Rate.

ASYMPTOTIC POWER FORMULAS FOR L_1^* AND J_2^*

We can derive general power and sample size formulas for situations of both differential and non-differential errors, which yields the potential for directly computing the change in power and sample size increase necessary to mitigate the effects of genotype errors.

L_1^*

As established in the introduction to Section Distributions, Power and Type I Error Rates of Gene-based Rare Variant Test Statistics, the minor allele counts c_i^+ and c_i^- are both binomially distributed, with $c_i^+ \sim \text{Binom}(2N^+, f_i^+)$ and $c_i^- \sim \text{Binom}(2N^-, f_i^-)$. While not needed in our initial exploration of the direction and relative effects of non-differential genotype errors on the type I error rate and power, to make predictions of the actual change in power or type I error rate, we utilize the normal approximation described earlier (Distributions, Power and Type I Error Rates of Gene-based Rare Variant Test Statistics Assumptions).

Since $D_i \dot{\sim} \text{Norm}\left(\mu_{D_i}, \sigma^2_{D_i}\right)$, $L_1 = \sum_{i=1}^m D_i \dot{\sim} \text{Norm}\left(\sum_{i=1}^m \mu_{D_i}, \sum_{i=1}^m \sigma^2_{D_i}\right)$. In the presence of errors, $L_1^* = \sum_{i=1}^m D_i^* \dot{\sim} \text{Norm}\left(\sum_{i=1}^m \mu_{D_i^*}, \sum_{i=1}^m \sigma^2_{D_i^*}\right)$.

Estimated power in the presence of non-differential genotype error. To determine the test's power, first find the $z_{1-\alpha}$ quantile, C, under the null hypothesis as $C = z_{1-\alpha}\sqrt{\sum_{i=1}^m \sigma^2_{D_{i,H_0}}}$. Find the corresponding quantile, z_β, on the alternative hypothesis distribution as $z_\beta = \frac{C - \sum_{i=1}^m \mu_{D^*_{i,H_A}}}{\sqrt{\sum_{i=1}^m \sigma^2_{D^*_{i,H_A}}}}$ and compute the power, π, as $\pi = 1 - \Phi\left(z_\beta\right)$ where $\Phi(\cdot)$ is the normal cdf.

Sample size necessary in the presence of non-differential genotype error. Since power decreases in the presence of non-differential genotype error (as shown in Non-differential Genotype Errors and Power), we can find the sample size necessary for a given power in the presence of genotype errors. To assist in the following proof, let $k = N^-/N^+ = N^{-*}/N^{+*}$ and $t_i^* = \left(\frac{1}{2}\right)\left(f_i^{+*}\left(1 - f_i^{+*}\right) + \frac{1}{k} f_i^{-*}\left(1 - f_i^{-*}\right)\right)$ so that $\sigma^2_{D_i^*} = \frac{1}{2N^{+*}}\left(f_i^{+*}\left(1 - f_i^{+*}\right)\right) + \frac{1}{2kN^{+*}}\left(f_i^{-*}\left(1 - f_i^{-*}\right)\right) = \frac{t_i^*}{N^{+*}}$.
To determine N^{+*} needed for a given α and β note that

$$z_\beta = \frac{C - \sum_{i=1}^m \mu_{D^*_{i,H_A}}}{\sqrt{\sum_{i=1}^m \frac{t^*_{i,H_A}}{N^{+*}}}} = \frac{z_{1-\alpha}\sqrt{\sum_{i=1}^m \frac{t^*_{i,H_0}}{N^{+*}}} - \sum_{i=1}^m \mu_{D^*_{i,H_A}}}{\sqrt{\sum_{i=1}^m \frac{t^*_{i,H_A}}{N^{+*}}}}$$

$$z_{1-\alpha}\sqrt{\sum_{i=1}^m t^*_{i,H_0}} - z_\beta\sqrt{\sum_{i=1}^m t^*_{i,H_A}} = \frac{\sum_{i=1}^m \mu_{D^*_{i,H_A}}}{\sqrt{\frac{1}{N^{+*}}}}$$

And so,

$$N^{+*} = \left(\frac{z_{1-\alpha}\sqrt{\sum_{i=1}^m t^*_{i,H_0}} - z_\beta\sqrt{\sum_{i=1}^m t^*_{i,H_A}}}{\sum_{i=1}^m \mu_{D^*_{i,H_A}}}\right)^2$$

To find the percent sample size increase necessary to maintain power, simply compute the ratio of N^{+*} to N^+, where N^+ is determined following the same procedure as is used for N^{+*}, only using values for t_i and μ_{D_i} not in the presence of errors.

Type I error rate in the presence of differential genotype error. In the presence of differential error, we can use a similar procedure to the one described in Estimated Power in the Presence of Non-Differential Genotype Error to determine the Type I error rate. Specifically, first find the $z_{1-\alpha}$ quantile, C, under the null hypothesis as $C = z_{1-\alpha}\sqrt{\sum_{i=1}^{m}\sigma^2_{D_{i,H_0}}}$ corresponding to the nominal type I error rate α. Find the corresponding type I error rate in the presence of differential genotype errors, $z_{1-\alpha^*}$, as $z_{1-\alpha^*} = \frac{C-\sum_{i=1}^{m}\mu_{D^*_{i,H_0}}}{\sqrt{\sum_{i=1}^{m}\sigma^2_{D^*_{i,H_0}}}}$ and compute the inflated type I error rate, $1-\Phi\left(z_{1-\alpha^*}\right)$ where $\Phi(\cdot)$ is the normal cdf.

J_2^*

In Section Impact of Genotype Errors on the Type I Error and Power of j_2^* we demonstrated that $J^2_{2,scaled} = \sum_{i=1}^{m}\left(\frac{D_i}{\sigma_{D_i}}\right)^2 \dot{\sim} \chi^2_{m,\lambda}$ where $\lambda = \sum_{i=1}^{m}\left(\frac{\mu_{D_i}}{\sigma_{D_i}}\right)^2$ is the non-centrality parameter. The non-centrality parameter can be used to find the power, type I error rate and related quantities.

Estimated power in the presence of non-differential genotype error. To determine the test's power, first find $C = \chi^2_{m,\alpha}$. Then, find the value of β such that $C = \chi^2_{m,\beta,\lambda^*}$ and compute the power, π, as $\pi = 1-\beta$ and where λ^* is the non-centrality parameter in the presence of non-differential genotype errors.

Sample size necessary in the presence of non-differential genotype error. Since power decreases in the presence of non-differential genotype error (as shown in Non-differential Genotype Errors and Power), we can find the sample size necessary to attain a particular level of power in the presence of genotype errors. As was done in Sample Size Necessary in the Presence of Non-Differential Genotype Error, we will focus on obtaining the percent increase in sample size necessary (N^{+*}/N^{+}) when genotype errors are present to maintain power when genotype errors are not present, where we again let $k = N^-/N^+ = N^{-*}/N^{+*}$ and $t_i^* = \left(\frac{1}{2}\right)\left(f_i^{+*}\left(1-f_i^{+*}\right) + \frac{1}{k}f_i^{-*}\left(1-f_i^{-*}\right)\right)$ so that $\lambda = \sum_{i=1}^{m}\left(\frac{\mu_{D_i}}{\sigma_{D_i}}\right)^2 = \sum_{i=1}^{m}\left(\frac{{\mu_{D_i}}^2}{\frac{t_i}{N^+}}\right)$.

We start by noting that in order to maintain power, the value of the non-centrality parameter without errors, λ^*, must be the same as the value of the non-centrality parameter when errors are present, λ^*.

Thus, we solve the following for N^{+*}/N^{+}.

$$\lambda = \lambda^*$$

$$\sum_{i=1}^{m}\left(\frac{{\mu_{D_i}}^2}{t_i}\right) = \sum_{i=1}^{m}\left(\frac{{\mu_{D_i^*}}^2}{\frac{t_i^*}{N^{+*}}}\right)$$

$$\frac{N^{+*}}{N^{+}} = \frac{\sum_{i=1}^{m}\left(\frac{{\mu_{D_i}}^2}{t_i}\right)}{\sum_{i=1}^{m}\left(\frac{{\mu_{D_i^*}}^2}{t_i^*}\right)}$$

Type I error rate in the presence of differential genotype error. In the presence of differential error, we can use a similar procedure to the one described in Estimated Power in the Presence of Non-Differential Genotype Error to determine the Type I error rate. To determine the test's power, first find $C = \chi^2_{m,\alpha}$, the nominal type I error rate with no errors. Then, find the value of α^* (the inflated type I error) such that $C = \chi^2_{m,\alpha^*,\lambda^*}$ where λ^* is the non-centrality parameter in the presence of differential genotype errors.

SIMULATION

We conducted a simulation study In order to determine to confirm theoretical intuitions described above, evaluate the quality of asymptotic normal distributions and to demonstrate that, while not explicitly considered above, joint and length test behavior across a wider class of norms ($L_1, L_2, L_4, L_\infty, J_1, J_2, J_4, J_\infty$) follows predicted patterns.

Simulation settings

For all simulation settings we consider a situation where there were 1000 cases and 1000 controls, and the number of variants, m, was fixed at 8. Genotypes at each variant, i, were simulated independently, following the assumptions of Hardy–Weinberg Equilibrium in the controls. Genotype errors were added to the true genotypes according to three error different models: ε_{10} error only, ε_{01} error only, and both ε_{10} and ε_{01} errors. Due to the stringent priors often placed on genotype callers, calling rare minor alleles is difficult, and thus ε_{01} error rates tend to be smaller than ε_{10} error rates (Powers et al., 2011). In order to reflect these realistic differences in error rates, we considered the following seven error settings, which are given as $(\varepsilon_{01}, \varepsilon_{10})$: (0, 0), (0, 0.1), (0, 0.5), (0.01, 0), (0.05, 0), (0.01, 0.1), (0.05, 0.5). We considered five different MAF settings: all variants MAF = 1%, all variants MAF = 0.1%, all variants MAF = 0.01%, two variants at 1%/six variants at 0.1% and two variants at 1%/six variants at 0.01%. All 35 combinations of MAF and genotype error rates were then considered for additional situations using differential and non-differential errors.

For non-differential errors, we used a relative risk distribution of 1.5 for MAF = 1%, 3 for MAF = 0.1% and 5 for MAF = 0.01% for risk-increasing, and the inverse for protective variants with those MAFs. We then considered six different mixes of causal and non-causal variants (1) all variants non-causal, (2) all variants risk increasing, (3) all variants risk reducing, (4) ½ variants risk reducing and ½ risk increasing, (5) ½ variants non-causal and ½ risk increasing, and (6) ½ variants non-causal, ¼ risk increasing and ¼ risk reducing), for a total of $6 \times 35 = 210$ settings with non-differential errors, 35 of which have no risk variants. In the case of differential errors, the relative risk was set to 1, and two different magnitudes of differential error were considered: relative difference in case and control genotype error rates (error rate in cases divided by error rate in controls) of 1.2, 1.5, 1/1.2, and 1/1.5. Thus, we considered $35 \times 4 = 140$ different cases of differential genotyping error.

A follow-up simulation study was conducted for the purposes of better understanding the behavior of tests with different norms. In particular, we started with the same 35 combinations of MAF

and genotype error rate as in the main simulation study. We then considered two settings: one with 8 SNPs and the other with 16 SNPs, where in each case only one SNP in the set was causal (designated to be a SNP with a larger MAF in cases where SNPs have varying MAF). This simulation only considered non-differential error.

Calculating power and type I error

For each simulation setting listed above, we generated 1000 independent samples. We then used phenotype permutation (1000 permutations for each sample) to compute p-values for eight different test statistics: $L_1, L_2, L_4, L_\infty, J_1, J_2, J_4, J_\infty$, where the p-value is the percent of permuted values of the test statistic that exceeded the observed value. The power or type I error rate is then computed as the percentage of the 1000 samples with p-values less than 0.05. For L_1 and J_2 asymptotic power predictions were also computed for each setting.

RESULTS

OVERALL IMPACTS OF NON-DIFFERENTIAL ERRORS

Type I error is control in the presence of non-differential errors

There were 35 simulation settings where there were no causal variants and non-differential genotype errors. To assess the overall control of the type I error rate, we looked at all 280 simulation by test statistic combinations (35 settings × 8 different statistics). An empirical type I error rate between 3 and 7% was considered to be reasonable control of the type I error rate (nominal level = 5%; approximate 99% margin of error = 2%). The vast majority (86.1%; 241/280) of test-statistic combinations showed reasonable control of the nominal type I error rate (empirical type I error rate between 3 and 7%). Of the 39 remaining settings, all showed deflation of the empirical type I error rate below the nominal level (Mean = 0.01, $SD = 0.011$, Min. = 0, Max. = 0.028). Twenty-five of the thirty-nine settings occurred when all variants had MAF = 0.01%, meaning that the average number of rare variants in the gene being analyzed was only 1.6 in the cases and 1.6 in the controls across all 8 variant sites combined. Across the remaining 14 settings, the average MAF was still relatively low (mean = 0.0011). The 39 settings were fairly indiscriminate across the 8 different test statistics considered here. Overall, type I error was controlled in the presence of non-differential errors.

Non-differential genotype errors decrease power

To assess the overall relationship between non-differential genotype errors and power when causal variants were present, we regressed empirical power on (a) average MAF across all variants, (b) magnitude of errors (0,1,5; where for ε_{10}, 0 = 0%, 1 = 1%, and 5 = 5% and for ε_{01}, 0 = 0%, 1 = 10%, and 5 = 50%), (c) percent of risk increasing variants and (d) percent of risk reducing variants for each of the test statistic by type of error (ε_{01} only, ε_{10} only, or ε_{01} and ε_{10}) combinations where at least one variant increased or reduced disease risk. Overall, when focusing on the impact of genotype errors, we found that regression model coefficients for ε_{01} only and ε_{01} and ε_{10} models were quite similar, while ε_{10} only was quite different. This confirms that the impact of ε_{10} is much less than that of ε_{01}. Furthermore, as error rates increased, power decreased (e.g., 3–5% for 1% increase in ε_{01} errors). Finally, as expected, increases to the MAF and percent of risk increasing variants increased power (e.g., increase in average MAF of 0.1%, increased power 1.0–3.2%; increase of 10% in proportion of risk increasing variants increased power 1.3–7.0%), while increases to the percent of risk-reducing variants increased power for joint tests (0.5–2.4%) and decreased power (0.6–1.3%) for length tests. **Table 1** shows the coefficients for regression models across all non-differential genotype error settings.

Figure 1 further illustrates that the effect of genotype errors is compounded by the MAF. While the power is similar when no errors are present, similar magnitude errors for lower MAF decrease power at a faster pace than in cases with larger MAF variants.

OVERALL IMPACTS OF DIFFERENTIAL ERRORS

Similar to the previous section, we used regression to assess the overall impacts of the three main simulation parameters (MAF, error magnitude and ratio of case to control errors) on the type I error rate when there were differential genotype errors. **Table 2** shows the coefficients for regression models across all differential genotype error settings. In general, regression coefficients are similar for the ε_{01} only and ε_{01} and ε_{10} models, confirming that, as is the case for non-differential genotype errors, the effect of ε_{10} errors are less compared to the effects of ε_{01} errors. When ε_{01} errors are present, the type I error rate increased when increasing either the magnitude of the errors (between 6 and 13% increase in type I error rate for 1% increase in ε_{01} errors) or increasing the difference between the case and control error rates (between 9 and 12% increase in type I error rate for 10% relative increase in case error rate); changes to the MAF alone did not had little impact the type I error rate. However, as MAF decreases the effects of differential genotyping errors become even greater in magnitude, as illustrated in **Figure 2** for J_2, but a pattern that is true regardless of choice of test statistic.

THE IMPACT OF GENOTYPE ERRORS ON CHOICE OF TEST STATISTIC

While we have described the general effects of genotype errors on power and type I errors within particular test statistics, the geometric framework provides a basis for comparisons about the effects of genotype errors across two characteristics of rare variant test statistic: choice of length or joint test and choice of norm. We now consider each of these choices in turn.

Choice of length or joint test statistic

As shown both theoretically and validated by simulation, the general patterns of the effects of genotype error and allele frequency on length and joint tests are similar (see Methods, Overall Impacts of Non-Differential Errors, and Overall Impacts of Differential Errors). However, there is one important distinction worth addressing. In particular, recall the distinction between length and joint tests: length tests use the difference in case-control total allele frequency at the locus as the statistic, while joint tests compute the difference in allele frequencies at each variant site and then sum the differences across the locus.

Non-differential errors. For non-differential errors at a causal locus, if genotype errors yield a reduction in the difference in the

Table 1 | Regression model coefficients relating power loss/gain to simulation parameters.

Norm	Type	Error magnitude (per 1% for ε_{01}; 10% for ε_{10})[a]			MAF (per 0.1%)[b]			Percent risk increasing (per 10%)[c]			Percent risk reducing (per 10%)[d]		
		ε_{01} only	ε_{10} only	ε_{01} and ε_{10}	ε_{01} only	ε_{10} only	ε_{01} and ε_{10}	ε_{01} only	ε_{10} only	ε_{01} and ε_{10}	ε_{01} only	ε_{10} only	ε_{01} and ε_{10}
1	Length	−0.04**	−0.02**	−0.05***	0.015**	0.010	0.011*	0.053***	0.070***	0.046**	−0.007	−0.010	−0.006
	Joint	−0.05***	−0.04***	−0.05***	0.028***	0.031***	0.023***	0.034***	0.049***	0.030***	0.017*	0.024**	0.015*
2	Length	−0.04**	−0.03**	−0.05***	0.017**	0.014**	0.013*	0.050***	0.062***	0.043***	−0.009	−0.011	−0.007
	Joint	−0.05***	−0.04***	−0.05***	0.028***	0.032***	0.024***	0.030***	0.040***	0.026**	0.014*	0.018*	0.012
4	Length	−0.04**	−0.03***	−0.05***	0.021***	0.020***	0.017***	0.045***	0.054***	0.039***	−0.009	−0.013	−0.008
	Joint	−0.04***	−0.03***	−0.05***	0.024***	0.027***	0.019***	0.023***	0.029***	0.020**	0.009	0.012	0.008
8	Length	−0.03***	−0.03***	−0.04***	0.017**	0.018***	0.014***	0.028***	0.039***	0.026***	−0.010	0.011	−0.007
	Joint	−0.03***	−0.03***	−0.03***	0.017***	0.020***	0.014***	0.015***	0.017***	0.013**	0.005	0.005	0.005

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

[a] *Increase in power for a 1% increase in error rate for ε_{01} or a 10% increase in error rate for ε_{10}. For example, for J_2, in the presence of ε_{01} only errors, a 1% increase in genotype error rate decreases power by an average of 5% points.*

[b] *Increase in power for a 0.1% increase in average MAF across all variant sites in the gene. For example, for J_2, in the presence of ε_{01} only errors, a 0.1% increase in average MAF across all variant sites increases power by an average of 2.8% points.*

[c] *Increase in power for a 10% point increase in the number of risk increasing SNPs. For example, for J_2, in the presence of ε_{01} only errors, a 10% increase in number of number of risk increasing SNPs across all variant sites increases power by an average of 3.4% points.*

[d] *Increase in power for a 10% point increase in the number of risk reducing SNPs. For example, for J_2, in the presence of ε_{01} only errors, a 10% increase in number of number of risk reducing SNPs across all variant sites increases power by an average of 1.7% points.*

cumulative MAF between cases and controls, there will be power loss. For joint tests, if genotype errors yield a reduction in the cumulative differences in allele frequency, there will be power loss. Thus, for joint tests, total power loss is a straightforward cumulative function of the power loss at each variant site. Things are, however, more complex for length tests. In a situation where all variants are risk-increasing, total power loss is a cumulative function of the power loss at each variant site. However, length tests lose power when protective variants and risk-increasing variants are present in the same gene because the effects of the variants "cancel out." In this case, genotype errors can mitigate some of the power loss due to cancellation by bringing the difference

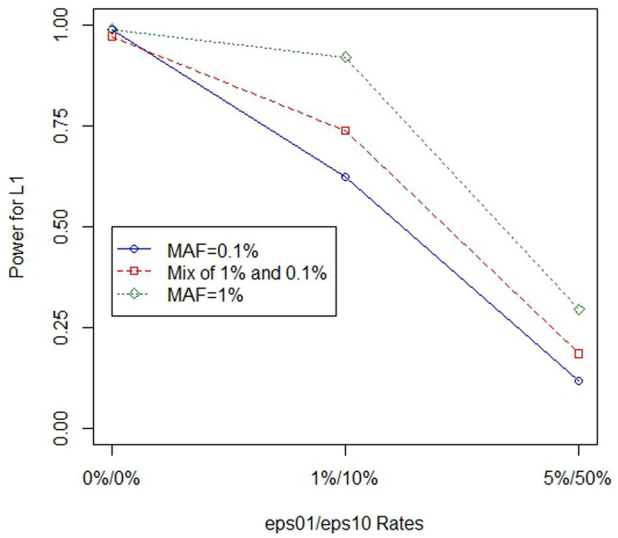

FIGURE 1 | Power for L_1 in the presence of non-differential genotype errors at different MAF. Power loss occurs when non-differential genotype errors are present at a locus. The power curves illustrated are at a site with eight causal variants. As genotype errors increase, power loss occurs. However, the power loss is most substantial when the minor allele frequency is the lowest.

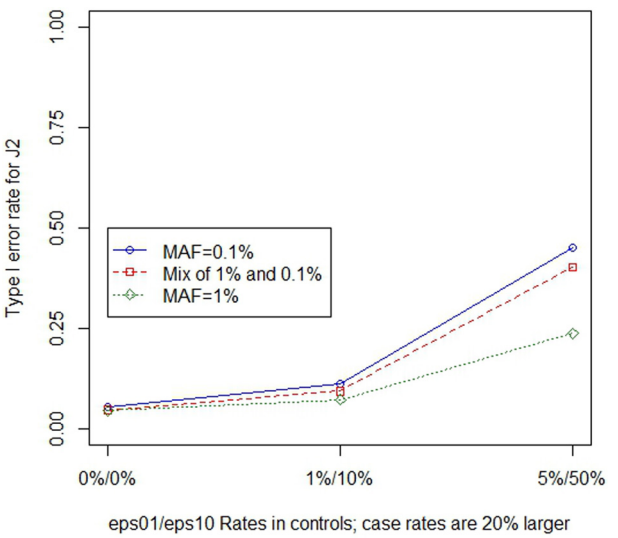

FIGURE 2 | Type 1 error rate for J_2 in the presence of differential genotype errors at different MAF. As differential error rates increase, the type I error rate increases. The type I error rate is illustrated at a site with eight non-causal variants. As differential (20% higher in cases) genotype error rates increased, the type I error rate increased. When the MAF was low, this effect was even larger.

Table 2 | Regression model coefficients relating type I error loss/gain to simulation parameters.

Norm	Type	Error magnitude (per 1% for ε_{01}; 10% for ε_{10})[a]			MAF (per 0.1%)[b]			Ratio of case and control error rates (per 10%)[c]		
		ε_{01} only	ε_{10} only	ε_{01} and ε_{10}	ε_{01} only	ε_{10} only	ε_{01} and ε_{10}	ε_{01} only	ε_{10} only	ε_{01} and ε_{10}
1	Length	0.08***	0.01*	0.07***	−0.004	0.008*	−0.005	0.12***	−0.004	0.11***
	Joint	0.13***	0.02**	0.12***	−0.005	0.009**	−0.009	0.10***	0.022*	0.10***
2	Length	0.08***	0.01*	0.07***	−0.003	0.007*	−0.005	0.11***	−0.006	0.11***
	Joint	0.13***	0.02**	0.12***	−0.005	0.009**	−0.009	0.10***	0.022*	0.10***
4	Length	0.08***	0.01*	0.08***	−0.002	0.007*	−0.004	0.10***	−0.005	0.10***
	Joint	0.12***	0.01**	0.11***	−0.004	0.008**	−0.008	0.10***	0.020*	0.09***
8	Length	0.06***	0.005	0.06***	−0.001	0.004*	−0.003	0.10***	−0.003	0.10***
	Joint	0.10***	0.01**	0.10***	−0.003	0.005**	−0.007	0.09***	0.013*	0.09***

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

[a] *Increase in type I error rate for a 1% increase in error rate for ε_{01} or a 10% increase in error rate for ε_{10}. For example, for J_2, in the presence of ε_{01} only errors, a 1% increase in genotype error rate increases the average type I error rate by 13% points when differential genotype errors are present.*

[b] *Increase in type I error rate for a 0.1% increase in average MAF across all variant sites in the gene. For example, for J_2, in the presence of ε_{10} only errors, a 0.1% increase in average MAF across all variant sites increases the average type I error rate by 0.9% points.*

[c] *Increase in type I error for a 10% increase in the relative difference between the ratio of the case to control error rate. For example, for J_2, in the presence of ε_{01} only errors, if the ratio of case error rate is 10% larger than control error rate 10% (e.g., 0.011 and 0.01), then the average type I error rate increases by 10% points.*

in case-control allele counts closer together at protective variant sites (see Section non-Differential Genotype Errors and Power for details).

Differential errors. Similar to Non-Differential Errors, the effects of differential genotype errors on joint tests is simply the accumulation of the effects at each variant site. However, the effect of differential errors on length tests becomes more complex. For example, if ε_{10} is larger in the cases than in the controls for a risk increasing variant, then differential errors can create a variant site which has more rare alleles in the controls than in the cases increasing the type I error rate for both length and joint tests. However, for length tests, the inflation of the type I error rate may be mitigated if a protective variant is present in the gene or if another variant in the gene has ε_{10} is larger in the controls than in the cases. Details follow directly from equations in Section Differential Genotype Errors and the Type I Error Rate.

Choice of norm

While the focus of the bulk of literature has been on development of L_1 or J_2 tests, recent work has shown potential advantages to the use of higher normed tests as a built in form of variant weighting which may yield higher power, while controlling the type I error rate when the proportion of non-causal variants is high. We will now explore the simulation results by evaluating the performance of test statistics using different norms.

In the main simulation, lower normed tests always outperformed higher normed tests in the main simulation where there were 8 variants, with either 50 or 100% of the variants classified as "causal" in cases where at least one variant at the locus modified disease risk. In the follow-up simulation we considered situations with 8 and 16 variants, where only one of the variants modified risk. When only one of the eight variants was causal, low norm tests outperformed high norm tests. However, when only one of sixteen variants was causal, high normed tests outperformed low norm tests in some cases. **Figures 3**, **4** illustrate the general patterns for length and joint tests, across norms. In short, while genotype errors contributed to power loss, the power loss was partially mitigated through the use of the larger norm.

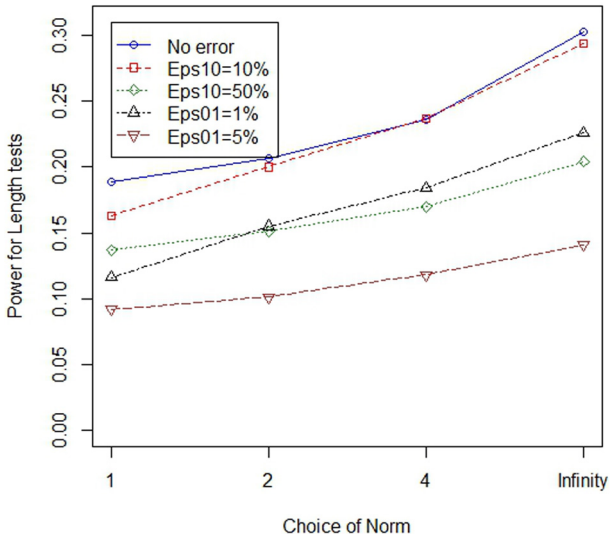

FIGURE 3 | Higher norms are more robust to genotype errors when the proportion of non-causal variants is larger: length tests. The figure illustrates power of four different (norm) length statistics, under varying error models. All test statistics experience power loss in the presence of errors. However, power loss can be mitigated through the use of high norm test statistics.

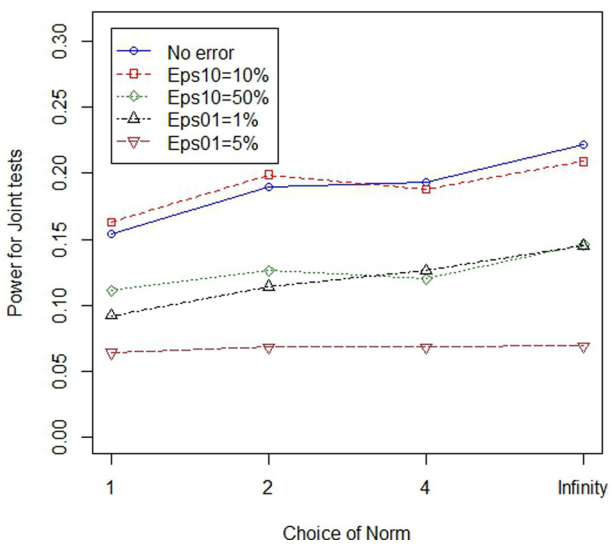

FIGURE 4 | Higher norms are more robust to genotype errors when the proportion of non-causal variants is larger: joint tests. The figure illustrates power of four different (norm) joint statistics, under varying error models. All test statistics experience power loss in the presence of errors. However, power loss can be mitigated through the use of high norm test statistics.

QUALITY OF ASYMPTOTIC POWER AND TYPE I ERROR PREDICTIONS

In order to evaluate the quality of asymptotic power and type I error predictions we compared the predicted power and type I error rates (see Simulation) to those obtained via permutation in the simulation study for $\boldsymbol{L_1}$ and $\boldsymbol{J_2}$. We use a significance level of 5% to evaluate consistency of predictions, but a follow-up analysis using lower significance thresholds of 10^{-4}, 10^{-5}, and 10^{-6} for a select group of simulation settings showed similar levels of consistency with predicted power and type I error rates as described in the following three sections (detailed results shown).

Type I error predictions in the presence of non-differential genotype error

As expected the type I error rate of the three asymptotic tests generally matched permutation tests since the asymptotic tests predicted 5% type I error rate in all cases (details not shown) and the permutation tests generally demonstrated control of the type I error rate, except in cases of extremely low (aggregate) MAF (see Type I Error is Control in the Presence of Non-Differential Errors for details), where the permutation tests showed empirical type I error rates less than the nominal level.

Power predictions in the presence of non-differential genotype error

Overall, predicted power was very close to observed power. Across 175 simulation settings with causal variants, most power predictions were within 10% of the true power (91% for $\boldsymbol{L_1}$, and 83% for $\boldsymbol{J_2}$). The quality of power predictions was strongly associated with the average MAF across the 8 sites in the control sample, as shown in **Table 3**.

Type I error predictions in the presence of differential genotype error

Similarly, predicted type I error inflation from differential genotype errors was very close to the empirical type I error rate across 140 simulation settings with no risk variants, but differential genotype errors present. The vast majority of type I error predictions were within 5% of the empirical type I error rate (91% for $\boldsymbol{L_1}$ and 84% for $\boldsymbol{J_2}$). Again, the quality of predictions was strongly associated with the average MAF in the control sample.

Software

Software (R scripts) for asymptotic power predictions and sample size computations for $\boldsymbol{L_1}$ and $\boldsymbol{J_2}$ based on the formulas and methods shown in Asymptotic Power Formulas for L_1^* and J_2^* is provided on the research group's website at: http://www.dordt.edu/statgen and following the links to the Software page.

DISCUSSION

Misclassification errors are a perennial problem in data analysis, and can be particularly magnified when using new technology which is often more error prone than mature technology. Recently, there has been substantial methodological effort devoted to the development of methods for analyzing next-generation sequencing data. However, much of this effort has ignored the problem of misclassification errors in the underlying genotype data (genotype errors). We have demonstrated that the persistent issue of genotype errors in next-generation sequencing data (Nielsen et al., 2011; Browning and Browning, 2013) has the potential to substantially reduce power and/or increase the type I error rate of the majority of related rare variant tests of association. Researchers should use the software and analytic tools described above to easily estimate the impact of genotype errors on downstream analyses. Thus, appropriately increasing sample size of next-generation studies to minimize power loss due to genotype error.

We have provided an initial theoretical justification behind recent simulation results evaluating the impact of both non-differential and differential genotype errors. In particular, we have confirmed that errors from the common homozygote to the heterozygote (ε_{01}) are particularly detrimental. The effects are further compounded depending upon whether the genotype errors are differential (increasing MAF increases type I error rate) or non-differential (decreasing MAF decreases power). In general, the effects of heterozygote to common homozygote errors (ε_{10}) are small and varied. The type I error rate is maintained in the presence of non-differential misclassification errors, with some over-conservatism when using permutation tests with extremely small allele frequencies due to the discrete nature of the permutation distribution. However, the type I error rate inflates in the presence of differential genotype errors. Our results are shown explicitly for common classes of test statistics, but are suggestive of the impact of genotype errors on all tests within the broad classes of length and joint tests regardless of the norm chosen.

To better understand why common homozygote to heterozygote errors can be so detrimental, it is useful to consider how many misclassifications are actually occurring in a dataset of interest. In the case of non-differential genotype errors, when examining rare variants (p is small), even small values of ε_{01} can yield many errors because most individuals in the dataset are common homozygotes. For example, on average, in a sample of 10,000 individuals, a rare variant with population MAF, $p = 0.001$, 9990 individuals will be the common homozygote, and so if ε_{01} is only 0.01, we expect nearly 100 (0.01*9990) misclassifications. On the other hand, even if ε_{10} is large (e.g., $\varepsilon_{10} = 0.10$), this yields, on average a small number of misclassifications (e.g., $0.10 * 10 = 1$). Notably, due to the aggregating nature of all gene-based rare variant tests as compared to single marker tests, the effects of genotype errors aggregate across variant sites within the gene, further increasing impact on power loss and type I error inflation.

Liu et al. (2013) demonstrated that the use of larger norms in rare variant tests provides increased robustness to the inclusion of non-causal variants. Our analysis demonstrates that another advantage of these tests is that they may be more robust to genotype errors than lower normed tests. Rare variant tests using a larger norm place increasing weight on sites with larger MAF in

Table 3 | Proportion of simulation settings and average MAF, within each absolute difference subcategory.

Abs. Diff.		Differential error		Non-differential error	
		$\boldsymbol{L_1}$	$\boldsymbol{J_2}$	$\boldsymbol{L_1}$	$\boldsymbol{J_2}$
<0.05	Percentage of settings (Count/Total)	98.9% (137/140)	83.6% (117/140)	88.0% (154/175)	73.1% (128/175)
	Mean control MAF	0.6%	0.6%	0.6%	0.7%
0.05–0.1	Percentage of settings (Count/Total)	2.1% (3/140)	13.6% (19/140)	5.7% (10/175)	15.4% (27/175)
	Mean control MAF	0.01%	0.1%	0.3%	0.2%
>0.1	Percentage of settings (Count/Total)	0	2.9% (4/140)	6.3% (11/175)	11.4% (20/175)
	Mean control MAF	–	0.3%	0.004%	0.1%

the cases or controls (length tests) or on the difference in MAF between cases and controls (joint tests). Because of the cumulative nature of the impact of genotype errors on rare variant tests, use of higher norms, reduces the overall impact of genotype errors. Whether high norm (e.g., infinity norm) tests are a powerful choice in practice is dependent upon underlying genetic architecture, dependent upon what percent of the variants at the locus are, in fact, causal, and how much prior understanding of the potential functional implications of those variants (e.g., synonymous vs. non-synonymous) can be used to minimize the impact of non-causal variants on the test (e.g., only including non-synonymous variants in the test). Importantly, in cases where genotype errors are larger for some variants, if the largest observed effects are at sites with a low error rate, and non-causal SNPs have a higher error rate, high normed tests may perform particularly well. Of course, high-normed tests perform less optimally compared to low-normed tests when numerous causal variants are present. Thus, use of methods in the spirit of those proposed by Derkach et al. (2013) have the potential to combine high norm tests with low-normed tests to yield a combined testing approach which is robust and powerful to numerous genetic architectures and genotype error distributions. Continued exploration of this class of high-normed rare variant tests is needed to assess its practical utility.

A related issue is that nearly all rare variant tests proposed to date do not explicitly account for genotype errors in the formulation of the test statistic. However, inclusion of genotype errors in the test statistic may also help to mitigate power loss and type I error inflation from genotype errors. While use of higher norms may, in some cases, mitigate the impact of genotype errors, development of tests which explicitly incorporate errors into the test may perform even better. There are some recently developed methods which address these weaknesses by directly incorporating sequence quality information (Daye et al., 2012) or advocating pooled study designs (Wang et al., 2012; Navon et al., 2013). However, in general, these methods remain outside of the mainstream. Expanded consideration of the impact of errors on more commonly used methods, combined with increased use of methods which explicitly model errors and/or study designs which limit the impact of errors are needed.

To explicitly incorporate errors into gene-based rare variant tests, explicit modeling of genotype error structures is needed. To do this, precise error models for genotype calling algorithms are needed. Currently, adjustments to, and practical use of, genotype calling algorithms are typically made with a generic sense of reducing errors and improving downstream analysis. Our results provide the basis for making stronger, more direct evaluation of upstream genotype calling algorithms in light of specific power and type I error implications. For example, the results here can be used to determine optimal ratios of ε_{01} to ε_{10} to minimize power loss—striking a meaningful and justified balance of sensitivity and specificity in the detection of rare alleles. Further work is needed which directly evaluates the decisions made in genotype calling algorithms with regard to their effects on genotype errors and downstream power and type I error implications and the potential development of alternate rare variant tests which explicitly incorporate genotype errors. This work may also include consider of errors involving the rare homozygote which was beyond the scope of our analysis.

Our analysis considers a situation where there is no LD between variants. The general effects of LD on the relationship between genotype errors and test performance are straightforward, while the details are quite complex. In short, the effects of genotype errors will generally be mitigated by LD structure due to (a) the potential for reduced genotype errors when using LD-aware callers and (b) the potential for increased power of multi-marker tests when LD is present between non-causal variants. While this general pattern is true, there is substantial detail related to (a) potential association between genotype error rates and LD structure and (b) potential differences in performance related to the relationship between LD and test statistic choice. Further work is needed to more specifically characterize the impact of LD on the effects of genotype errors.

Consideration of genotype errors in the design of studies is another implication of our work. In particular, we have conclusively demonstrated that power loss will be realized in the presence of non-differential genotype errors. Thus, if a researcher determines that they need N subjects to achieve an *a priori* specified level of statistical power, $1-\beta$, in their rare variant analysis, we have demonstrated that, in the presence of non-differential genotype errors, in almost all cases, the actual number of subjects needed is $N^*/N > 1$. While it is straightforward to see that the value of N^*/N increases in all the same situations that power decreases, tools are needed for researchers to quickly determine how sample size and power estimates should be modified to appropriately account for the impact of genotype errors. The asymptotic power predictions for $\boldsymbol{L_1}$ and $\boldsymbol{J_2}$ are provided as a first step toward nearly instantaneous evaluation of the impacts on power and type I error from different types and levels of genotype errors. The main utility in these formulas is in predicting the relative changes in power and type I error from genotype errors. However, even absolute power and type I error predictions were quite accurate in most cases. That said, there is room for improvement if the goal is accurate prediction of absolute power values (e.g., tweaking predictions for a particular variant weighting scheme).

Another important study design consideration relates to differential genotype errors. A growing practice is the use of publicly available databases (e.g., 1000 Genomes Project) as a source of non-diseased subjects since this can substantially reduce study costs. However, in such a case there is no guarantee that the genotype error model is the same in these publicly available databases vs. the error model in the diseased subjects—a situation potentially leading to differential genotype errors and inflated type I errors. The use of the asymptotic equations provided here can give a first level approximation of type I error inflation due to differential genotype errors. As shown, this inflation can be substantial even for modest levels of differential genotype error. Caution should be used when using publically available control samples. While overall methods for controlling the type I error (e.g., genomic control) are available, these methods can substantially reduce power compared to methods with explicitly model, account for or eliminate differential errors. A related issue is that of population stratification which also can inflate the type I error

rate. Further work is needed to more fully investigate relationships between population stratification and differential genotype error for rare variant tests of association.

To date only simulation results providing suggestive evidence of the impact of genotyping errors on rare variant tests of association has been available. Our work here, building off of the geometric framework, provides theoretical justification to these patterns. In particular, we demonstrate the potentially substantial impact of common homozygote to heterozygote errors on both power and type I error. The impact of the errors can be intensified depending on the underlying MAF and differential or non-differential nature of the genotype errors, and the test statistic used. Further work is needed to explore additional implications of these results on genotype calling algorithms, study design decisions and rare variant test statistic choice.

ACKNOWLEDGMENTS

This work was funded by the National Human Genome Research Institute (R15HG006915). We acknowledge the use of the Hope College parallel computing cluster for assistance in data analysis.

REFERENCES

Ahn, K., Gordon, D., and Finch, S. J. (2009). Increase of rejection rate in case-control studies with differential genotyping error rates. *Stat. Appl. Genet. Mol. Biol.* 8:Article25. doi: 10.2202/1544-6115.1429

Ahn, K., Haynes, C., Kim, W., Fleur, R. S., Gordon, D., and Finch, S. J. (2007). The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Ann. Hum. Genet.* 71(Pt 2), 249–261. doi: 10.1111/j.1469-1809.2006.00318.x

Asimit, J., and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* 44, 293–308. doi: 10.1146/annurev-genet-102209-163421

Awadalla, P., Gauthier, J., Myers, R. A, Casals, F., Hamdan, F. F., Griffing, A. R., et al. (2010). Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* 87, 316–324. doi: 10.1016/j.ajhg.2010.07.019

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* 11, 773–785. doi: 10.1038/nrg2867

Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* 35, 606–619. doi: 10.1002/gepi.20609

Bross, I. (1954). Misclassification in 2 X 2 Tables. *Biometrics* 10, 478–486. doi: 10.2307/3001619

Browning, B. L., and Browning, S. R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.* 93, 840–851. doi: 10.1016/j.ajhg.2013.09.014

Cooper, G. M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* 12, 628–640. doi: 10.1038/nrg3046

Dai, Y., Jiang, R., and Dong, J. (2012). Weighted selective collapsing strategy for detecting rare and common variants in genetic association study. *BMC Genet.* 13:7. doi: 10.1186/1471-2156-13-7

Daye, Z. J., Li, H., and Wei, Z. (2012). A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res.* 40, e60. doi: 10.1093/nar/gks024

Dering, C., Hemmelmann, C., Pugh, E., and Ziegler, A. (2011). Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet. Epidemiol.* (35 Suppl. 1), S12–S17. doi: 10.1002/gepi.20643

Derkach, A., Lawless, J. F., and Sun, L. (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genet. Epidemiol.* 37, 110–121. doi: 10.1002/gepi.21689

Feng, T., Elston, R. C., and Zhu, X. (2011). Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genet. Epidemiol.* 35, 398–409. doi: 10.1002/gepi.20588

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118

Gordon, D., Finch, S. J., Nothnagel, M., and 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, 22–33. doi: 10.1159/000066696

Gordon, D., Yang, Y., Haynes, C., Finch, S. J., Mendell, N., Brown, A., et al. (2004). Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double sampling. *Stat. Appl. Genet. Mol. Biol.* 3, 26. doi: 10.2202/1544-6115.1085

Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* 70, 42–54. doi: 10.1159/000288704

Ilie, L., Fazayeli, F., and Ilie, S. (2011). HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics* 27, 295–302. doi: 10.1093/bioinformatics/btq653

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* 7:e1001289. doi: 10.1371/journal.pgen.1001289

Kang, S., Finch, S. J., Haynes, C., and Gordon, D. (2004a). Quantifying the percent increase in minimum sample size necessary for SNP genotyping errors in genetic model-based association studies. *Hum. Hered.* 58, 139–144. doi: 10.1159/000083540

Kang, S., Gordon, D., and Finch, S. (2004b). What SNP genotyping errors are most costly for genetic association studies? *Genet. Epidemiol.* 26, 132–141. doi: 10.1002/gepi.10301

Li, B., and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321. doi: 10.1016/j.ajhg.2008.06.024

Li, Y., Byrnes, A. E., and Li, M. (2010). To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. *Am. J. Hum. Genet.* 87, 728–735. doi: 10.1016/j.ajhg.2010.10.014

Lin, D.-Y., and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* 89, 354–367. doi: 10.1016/j.ajhg.2011.07.015

Liu, K., Fast, S., Zawistowski, M., and Tintle, N. L. (2013). A geometric framework for evaluating rare variant tests of association. *Genet. Epidemiol.* 37, 345–357. doi: 10.1002/gepi.21722

Madsen, B. E., and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5:e1000384. doi: 10.1371/journal.pgen.1000384

Mayer-Jochimsen, M., Fast, S., and Tintle, N. L. (2013). Assessing the impact of differential genotyping errors on rare variant tests of association. *PLoS ONE* 8:e56626. doi: 10.1371/journal.pone.0056626

Morgenthaler, S., and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* 615, 28–56. doi: 10.1016/j.mrfmmm.2006.09.003

Morris, A. P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* 34, 188–193. doi: 10.1002/gepi.20450

Moskvina, V., Craddock, N., Holmans, P., Owen, M. J., and O'Donovan, M. C. (2006). Effects of differential genotyping error rate on the type I error probability of case-control studies. *Hum. Hered.* 61, 55–64. doi: 10.1159/000092553

Navon, O., Sul, J., Han, B., Conde, L., Bracci, P., Riby, J., et al. (2013). Rare variant association testing under low-coverage sequencing. *Genetics* 194, 769–779. doi: 10.1534/genetics.113.150169

Neale, B. M., Rivas, M. A, Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., et al. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.* 7:e1001322. doi: 10.1371/journal.pgen.1001322

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

Pan, W., and Shen, X. (2011). Adaptive tests for association analysis of rare variants. *Genet. Epidemiol.* 35, 381–388. doi: 10.1002/gepi.20586

Powers, S., Gopalakrishnan, S., and Tintle, N. (2011). Assessing the impact of non-differential genotyping errors on rare variant tests of association. *Hum. Hered.* 72, 153–160. doi: 10.1159/000332222

Rogers, A., Beck, A., and Tintle, N. (2014). Evaluating the concordance between sequencing, imputation and microarray genotype calls in the GAW18 data. *BMC Proc.* (in press).

Sul, J. H., Han, B., He, D., and Eskin, E. (2011). An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics* 188, 181–188. doi: 10.1534/genetics.110.125070

Wang, T., Lin, C.-Y., Zhang, Y., Wen, R., and Ye, K. (2012). Design and statistical analysis of pooled next generation sequencing for rare variants. *J. Probab. Stat.* 2012, 1–19. doi: 10.1155/2012/524724

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., and Zöllner, S. (2010). Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am. J. Hum. Genet.* 87, 604–617. doi: 10.1016/j.ajhg.2010.10.012

Zhang, Q., Irvin, M., Arnett, D., Province, M., and Borecki, I. (2011). A data-driven method for identifying rare variants with heterogeneous trait effects. *Genet. Epidemiol.* 35, 679–685. doi: 10.1002/gepi.20618

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 October 2013; accepted: 11 March 2014; published online: 01 April 2014. Citation: Cook K, Benitez A, Fu C and Tintle N (2014) Evaluating the impact of genotype errors on rare variant tests of association. Front. Genet. ***5****:62. doi: 10.3389/fgene.2014.00062*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.