**Twin Research and Human Genetics**
Volume 17 ▌ Number 3 ▌ pp. 143–150 ▌ © The Authors 2014 ▌ doi:10.1017/thg.2014.17

# Statistical Properties of Single-Marker Tests for Rare Variants

**T. Bernard Bigdeli,[1,2] Benjamin M. Neale,[3,4] and Michael C. Neale[1,2,5,6]**
[1]*Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA*
[2]*Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA*
[3]*Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA*
[4]*Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA*
[5]*Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA*
[6]*Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA*

With the dramatic technological developments of genome-wide association single-nucleotide polymorphism (SNP) chips and next generation sequencing, human geneticists now have the ability to assay genetic variation at ever-rarer allele frequencies. To fully understand the impact of these rare variants on common, complex diseases, we must be able to accurately assess their statistical significance. However, it is well established that classical association tests are not appropriate for the analysis of low-frequency variation, giving spurious findings when observed counts are too few. To further our understanding of the asymptotic properties of traditional association tests, we conducted a range of simulations of a typical rare variant (~1%) under the null hypothesis and tested the allelic $\chi^2$, Cochran–Armitage trend, Wald, and Fisher's exact tests. We demonstrate that rare variation shows marked deviation from the expected distributional behavior for each test, with fewer minor alleles corresponding to a greater degree of test statistics deflation. The effect becomes more pronounced at progressively smaller $\alpha$ levels. We also show that the Wald test is particularly deflated at $\alpha$ levels consistent with genome-wide association significance, much more so than the other association tests considered. In general, these classical association tests are inappropriate for the analysis of variants for which the minor allele is observed fewer than 80 times, largely irrespective of sample size.

■ **Keywords:** genome-wide association, next-generation sequencing, significance testing, rare variation

Genome-wide association studies (GWAS) have uncovered hundreds of loci relevant to common, complex diseases (Mailman et al., 2007). These studies assay single-nucleotide polymorphism (SNP) variation across the allele frequency spectrum, but are limited to studying SNPs with minor allele frequency (MAF) of at least 1–5%. Despite incomplete coverage of rare alleles in GWAS, a number of rare variants have been implicated in common, complex diseases. For example, recent work in type I diabetes identified a rare protective mutation in the gene *IFIH1*, with a population allele frequency of approximately 2% (Nejentsev et al., 2009). Sequencing endeavors, such as the 1,000 Genomes Project, are identifying human genetic variation down to frequencies less than 1%. This expanding collection of genetic polymorphisms is, in turn, being made accessible through extending genome-wide association SNP chips at ever decreasing frequencies and greater marker density.

With the increased focus on rare variants, the question of how best to assess their statistical significance arises. Traditional approaches, namely tests of independence for contingency tables, are not suitable when numbers of observations are too few, owing to the inexact approximation of discrete probabilities to the continuous theoretical $\chi^2$ distribution (Yates, 1934). For extremely uncommon variation, methods have been developed to test whether a set of variants are implicated in disease (Li & Leal, 2008; Madsen & Browning, 2009; Morgenthaler & Thilly, 2007; Neale et al., 2011). Such methods are better suited to loci for which classical association testing cannot be conducted because of the limited number of observations. Considered another way, a single locus that has only 10 copies of the minor allele

RECEIVED 10 November 2013; ACCEPTED 25 February 2014. First published online 17 April 2014.

ADDRESS FOR CORRESPONDENCE: Benjamin M. Neale, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA. E-mail: bneale@broadinstitute.org

in a balanced case–control study cannot achieve significance at established genome-wide levels ($5\times10^{-8}$) (Risch & Merikangas, 1996). One strategy to overcome this problem is to group multiple variants and to conduct tests of association with particular regions rather than with specific variants.

Given that the field has adopted a genome-wide association significance threshold, the accuracy of extreme *p*-values is also of great importance. For example, in the seminal work of Cohen and colleagues (2006), who identified *PCSK9* as a component of low-density lipoprotein (LDL) cholesterol, the authors used a $\chi^2$ test to assess the role of rare variation in determining risk of coronary heart disease, and reported *p*-values of .008 and .003 for African-American and Caucasian samples, respectively. If instead Fisher's exact test is applied, these *p*-values shrink to .0037 and .0024, respectively. Thus, the basic $\chi^2$ test under these circumstances is comparatively conservative in the face of rare observations. To enhance our understanding of the asymptotic properties of these traditional association tests, we undertook a range of simulations of the $\chi^2$ test, the Cochran–Armitage trend test (Neale et al., 2010; Sladek et al., 2007; Wellcome Trust Case Control Consortium, 2007), and the Wald test (Scott et al., 2007). We contrast the asymptotic properties of these tests with those of Fisher's (1922) exact test, which represents the canonical test for small sample sizes.

## Methods

To assess the asymptotic behavior of rare variant testing, we used a simple null model consisting of an SNP with 1% MAF equally likely to occur in 1,000 cases and 1,000 controls. Our naïve simulation procedure assigned genotypes for each individual randomly; we sampled $N = 2{,}000$ times (with replacement) from a binomial distribution, thereby allowing for sampling variance. That is, each individual replicate may have an observed MAF of 1%, a little more than 1%, or slightly less than 1%. To further constrain the behavior of these tests, we limited the minor allele count to 40 copies among 2,000 individuals. To determine whether the sample size matters, we increased the number of individuals to 10,000, while still fixing the number of minor alleles to 40. We also considered 20 and 80 copies of the minor allele in a sample size of 10,000. We proceeded to analyze each simulated dataset using a suite of common, association tests: the allelic $\chi^2$ test, the 1-*df* Cochran–Armitage trend test, and the Wald test for logistic regression. As our goal was to assess the asymptotic behavior of these tests, we chose to conduct a large number of simulations (one billion) for each scenario.

## Tests of Association

To best represent historical and common practices in genetic association studies, we have elected to highlight four tests in particular. The allelic $\chi^2$ test represents the most basic of these, while the Cochran–Armitage trend test has gained popularity as a 1-*df* test of genotypes. For regression-based approaches, the Wald test is a straightforward means of obtaining a $\chi^2$ approximation. Fisher's exact test, although less commonly utilized in the context of genetic association, is robust to small sample sizes, and as such provides a basis of comparison for the three traditional approaches.

The allelic $\chi^2$ test compares allele frequencies between cases and controls, and is widely used as a test of association for disease traits (Apple et al., 1994). Since the allelic test considers the allele as a relevant unit of analysis, it is assumed that the Hardy–Weinberg equilibrium (HWE) exists. This is equivalent, in the present context, to assuming that the alleles at a locus occur independently within both case and control populations. In other words, non-additive effects of the alleles at a locus are assumed to be absent. The allelic test is known to give spurious results if HWE is not met, although SNPs that show severe departures from HWE are generally unreliable and should be excluded from analysis. Interpretation of odds ratios given by this method is also with respect to alleles, as opposed to individuals, and is discussed elsewhere (Sasieni, 1997).

The Cochran–Armitage test for trend is a modification of a 2-*df* genotypic $\chi^2$ test to account for a hypothesized ordering of effects across genotype classes, consistent with additive models of disease risk (Armitage, 1955; Freidlin et al., 2002). Applied to common variants, the trend test is a more powerful test of association than standard allelic and genotypic $\chi^2$, owing to a weighting of genotypic classes that reduces the effective degrees of freedom. As the individual represents the relevant unit of analysis, the trend test has an additional advantage of not assuming HWE, although the allelic and trend tests are expected to be asymptotically equivalent when this condition is met (Sasieni, 1997). Odds ratios from the trend test may be interpreted as the increase in risk to an individual conferred by each additional copy of the reference allele.

The Wald test (Hauck & Donner, 1977; Wald, 1943) compares the maximum likelihood estimate of a statistical parameter with its expectation under the null, often as an approximation to the theoretical $\chi^2$ distribution. In the present context, we apply the Wald test to a simple logistic model ($\text{Aff} \sim \beta_0 + \beta_1 \cdot \text{SNP}$), which considers the number of minor alleles carried by an individual. Since it is often desirable to include demographic or clinical covariates in predictive disease models, we extend our regression model to incorporate a covariate predictor of fixed prevalence in the population, but for which carriers of the minor allele are at increased risk of endorsement. As for our basic logistic model, we applied the Wald test to obtain a $\chi^2$ approximation for the effect of SNP genotype.

As applied herein, Fisher's exact test compares allele frequencies between cases and controls and therefore may be considered analogous to the allelic $\chi^2$. Unlike the allelic $\chi^2$, however, Fisher's exact test is appropriate for all sample sizes. Fisher (1922) noted that if the marginal totals of

a 2×2 contingency table are held constant, then the test statistics follow a known sampling distribution (i.e., hypergeometric), from which calculation of the exact probability of observing a given set of counts by chance is straightforward. This robustness, in situations when the number of observations is otherwise limiting, is due to Fisher's exact test being based on a discrete probability distribution rather than approximation to the continuous theoretical $\chi^2$. More details regarding these tests can be found in the supplementary information.

## Generation of Asymptotic Distributions

Under each scenario, we simulated genotypic data that were identical with respect to the total number of minor alleles, *C*, the total sample size, *N*, and the proportions of cases and controls. For each replicate dataset, we sampled *N* times without replacement from a population of *N* diploid persons, in which only *C* chromosomes carry the minor allele, and assigned case status at random to exactly half of all individuals. It follows that the resultant case–control differences in allele frequency will be identically distributed, as illustrated by the observation that in the most extreme circumstance, all *C* copies of the minor allele will occur within cases or controls. By comparison, random simulation of genotypes on a per individual basis (i.e., sampling with replacement) might result in the total number of alleles being slightly greater or slightly fewer than *C*. Stated differently, each replicate dataset represents a standard 2×2 table of allele counts by outcome, but for which the marginal totals of rows and columns are fixed. Similarly, for both the trend test and the logistic model, the data may be arranged as 2×3 contingency tables of genotypic counts by outcome, in which the marginal totals are generally maintained. That is, our focus is on the asymptotic properties of standard association tests as applied to low-frequency variants, for which the occurrence of a minor allele homozygote ($MAF^2$) is an exceedingly rare event.

Since it is often desirable to include demographic or clinical covariates in predictive disease models (Bush et al., 2010; Scott et al., 2007; Thomas et al., 2009), we extended the regression models to incorporate a binary covariate predictor of fixed prevalence in the population, which carriers of the minor allele are more likely to endorse. We assume a 0.10 population endorsement rate across all scenarios, but vary this rate among carriers as 0.10, 0.20, 0.40, 0.60, and 0.80. For example, consider the scenario in which the endorsement rate among carriers is 0.80; individual covariate values were drawn at random from a binomial distribution, with the probability of 'success' specified as 0.80 for carriers of the minor allele and 0.10 for non-carriers, irrespective of case–control status. For each replicate dataset, we fitted logistic models that specified case–control status as a function of SNP genotype and a single covariate, and applied the Wald test to obtain a $\chi^2$ approximation for the effect of SNP genotype. Of particular interest is the effect of adding a predictor, unrelated to disease, on the regression of disease outcome on genotype. Note that although the number of cases and controls is fixed and equal across permutations, random simulation of a covariate will introduce variance into the observed distribution of test statistics.

Distributions for Fisher's exact test were also derived, but indirectly from the distributions for the allelic $\chi^2$. This is justified by our simulation procedure, as fixing the marginal totals constrains the number of possible configurations of the data within a 2×2 table of allele counts. That is, each unique value of the allelic $\chi^2$ corresponds to a specific set of observed counts for which the value of Fisher's exact test is known.

Due to the exceptional number of permutations required to evaluate asymptotic behavior within the critical region, we seeded 100,000 separate instances of our simulation procedure per scenario, making use of several high-performance computing clusters. Rendering of complete null distributions for each test was simplified by tabulating observed test statistics within each constituent distribution and compounding the resulting counts. We proceeded to quantify departures from expected asymptotic behavior, as defined by the theoretical $\chi^2$ distribution for $10^9$ tests.

## Results

### Common Association Tests

For each scenario, Table 1 gives the number of Cochran–Armitage trend, allelic $\chi^2$, and Wald tests (uncorrected for continuity) found to be significant at various α levels. Corresponding quantile–quantile plots are displayed in Figure 1. Expectations regarding asymptotic behavior are based on the theoretical $\chi^2$ distribution (see the Central Limit Theorem), to which approximations of binomial SNP data are, by definition, inexact. At a given threshold, the probability of observing a significant test statistics under the null is simply the proportion of the total number of permutations. Because our sampling procedure is effectively without replacement, resultant test statistics take on a finite number of discrete values. This is illustrated by the step-function-like appearance of the observed quantile plots (Figure 1).

Consider the distributions of the allelic $\chi^2$ and Fisher's exact tests, recalling that a 2×2 table of allelic counts will follow a hypergeometric distribution if marginal totals are held constant. For 40 copies of the minor allele in 1,000 cases and 1,000 controls, we observe fewer significant allelic $\chi^2$ tests than expected, with more pronounced discrepancies for progressively smaller α levels. Comparing the allelic $\chi^2$ and Fisher's exact methods, significant test counts obtained by each method are indistinguishable for all but the most extreme α levels. Given the same number of minor alleles (40) in 5,000 cases and 5,000 controls, we see an overall pattern of deflation similar to that observed for the smaller

**TABLE 1**

**Number of Significant 1-*df* Allelic $\chi^2$, Cochran–Armitage trend, Wald, and Fisher's Exact Test Statistics**

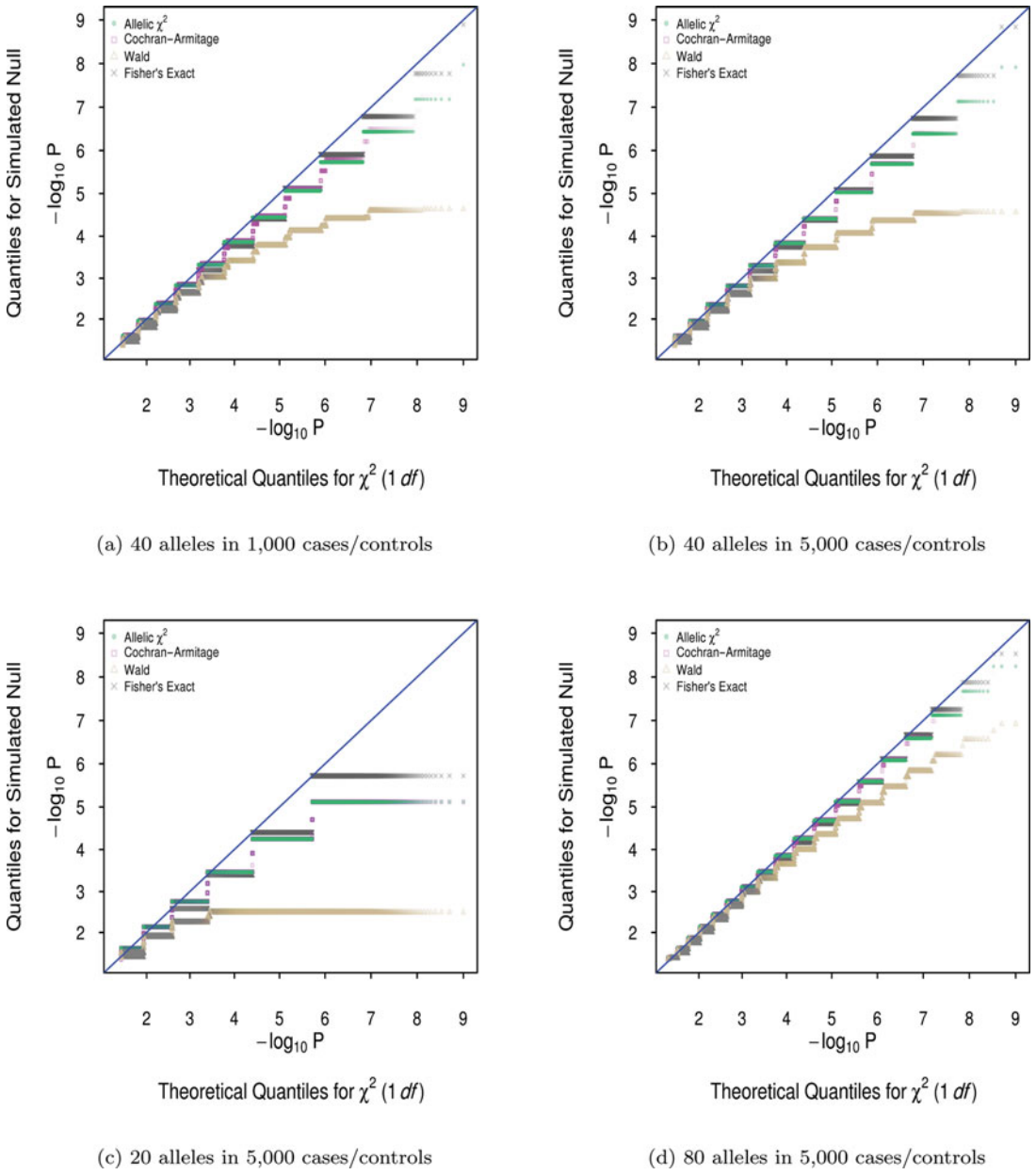
| | | | Significance threshold ($\alpha$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Null distribution | | | $\alpha < 10^{-1}$ | $\alpha < 10^{-2}$ | $\alpha < 10^{-3}$ | $\alpha < 10^{-4}$ | $\alpha < 10^{-5}$ | $\alpha < 10^{-6}$ | $\alpha < 10^{-7}$ | $\alpha < 10^{-8}$ |
| Theoretical $\chi^2$ (1 *df*) | | | 100,000,000 | 10,000,000 | 1,000,000 | 100,000 | 10,000 | 1,000 | 100 | 10 |
| | $Count_{MA}$ | *N* | | | | | | | | |
| Cochran–Armitage trend | 40 | 2,000 | 79,179,748 | 6,173,009 | 639,359 | 38,879 | 5,987 | 143 | 8 | 1 |
| Allelic $\chi^2$ | | | 79,179,748 | 6,173,016 | 640,064 | 38,933 | 7,647 | 148 | 11 | 0 |
| Wald | | | 79,179,748 | 6,167,797 | 534,239 | 5,983 | 0 | 0 | 0 | 0 |
| Fisher's exact | | | 79,179,748 | 6,173,016 | 640,064 | 38,933 | 7,647 | 148 | 11 | 1 |
| Cochran–Armitage trend | 80 | 10,000 | 92,269,671 | 9,541,111 | 1,021,754 | 68,457 | 8,128 | 699 | 55 | 3 |
| Allelic $\chi^2$ | | | 92,269,671 | 9,541,208 | 1,029,744 | 68,459 | 8,209 | 791 | 65 | 3 |
| Wald | | | 92,269,671 | 9,538,127 | 439,954 | 24,480 | 2,215 | 65 | 0 | 0 |
| Fisher's exact | | | 92,269,671 | 9,541,208 | 1,029,744 | 68,459 | 8,209 | 791 | 65 | 3 |
| Cochran–Armitage trend | 40 | 10,000 | 80,400,565 | 6,368,800 | 669,924 | 41,455 | 7,745 | 169 | 17 | 0 |
| Allelic $\chi^2$ | | | 80,400,565 | 6,368,800 | 669,932 | 41,455 | 8,147 | 170 | 18 | 0 |
| Wald | | | 80,400,565 | 6,368,746 | 178,816 | 7,745 | 0 | 0 | 0 | 0 |
| Fisher's exact | | | 80,400,565 | 6,368,800 | 669,932 | 41,455 | 8,147 | 170 | 18 | 2 |
| Cochran–Armitage trend | 20 | 10,000 | 115,139,237 | 11,657,125 | 399,840 | 39,160 | 1,884 | 0 | 0 | 0 |
| Allelic $\chi^2$ | | | 115,142,423 | 11,772,128 | 399,870 | 39,822 | 1,927 | 0 | 0 | 0 |
| Wald | | | 115,137,310 | 2,561,766 | 0 | 0 | 0 | 0 | 0 | 0 |
| Fisher's exact | | | 41,290,490 | 2,563,863 | 399,870 | 39,822 | 1,927 | 0 | 0 | 0 |

sample. Inspection of Table 1 reveals a slight increase in the number of significant tests observed, less than 2% and 5% for $\alpha$ thresholds of $10^{-2}$ and $10^{-5}$, respectively. However, the larger sample size does not see the allelic $\chi^2$ attain significance at $\alpha < 10^{-8}$. Restricting the number of minor alleles to exactly 20 copies, there is a marked decrement in the value of test statistics by either method, with neither reporting a single *p*-value less than $10^{-6}$. We observe an excess of significant findings by the allelic $\chi^2$ for the $\alpha < 10^{-2}$ critical region, with no such inflation apparent for Fisher's exact test. Increasing the number of minor alleles to 80 copies in 5,000 cases and 5,000 controls, asymptotic behavior is visibly restored. Residual deflation is only minimally apparent at genome-wide thresholds, at which both tests report significant findings.

Except those additional values indicating one or more observed minor homozygotes, quantiles for the Cochran–Armitage trend test largely parallel those of the allelic $\chi^2$. With 40 minor alleles in 1,000 cases and 1,000 controls, the trend test gives a significant result at $\alpha < 10^{-8}$, which the allelic test failed to identify. Loss of power is evident, with fewer significant permutations observed overall than with either the allelic $\chi^2$ or exact test. Differences between the allelic $\chi^2$ and trend tests are less marked, with 40 copies in 5,000 cases and 5,000 controls, due to the reduced likelihood of observing a minor allele homozygote. With only 20 copies of the minor allele, power for the trend test is diminished further. Under these conditions, the chance of observing a minor homozygote is only one in one million. Similar to the allelic $\chi^2$ and exact tests, the trend test fails to return a single *p*-value less than $10^{-6}$. An excess of significant findings in the $\alpha < 10^{-2}$ critical region is also apparent, but to a slightly lesser extent than seen for the allelic $\chi^2$. Power for the trend test is restored by increasing the number of minor alleles to 80 copies. In spite of the deflation being visibly attenuated, the trend test gives slightly fewer significant differences in the critical region than either the allelic $\chi^2$ or Fisher's exact test.

Deflation of the Wald test statistics is considerably more pronounced than those of the allelic $\chi^2$ and trend tests. With 40 minor alleles in either sample size, the Wald test fails to report a single significant finding at $\alpha < 10^{-5}$, returning *p*-values 10-, 100-, and 1,000-fold larger than expected at $\alpha$ thresholds of $10^{-5}$, $10^{-7}$, and $10^{-8}$, respectively. With the total number of minor alleles limited to 20 copies, deviation from expected distributional behavior is particularly extreme. We fail to observe any significant findings for $\alpha < 10^{-3}$, corresponding to a deflation factor of 100,000 at $\alpha < 10^{-8}$. With 80 minor alleles in 5,000 cases and 5,000 controls, the Wald test is noticeably improved but still gives *p*-values an order of magnitude larger than expected at $\alpha < 10^{-8}$.

Comparing 80, 40, and 20 copies of the minor allele in 5,000 cases and 5,000 controls, there is an overall increase in the extent of deflation for successively fewer copies of the minor allele, and an increase in the value of $\alpha$ at which this deflation is first apparent. Given the demonstrated non-effect of sample size, it follows that we may take findings for 80 minor alleles in 5,000 cases and 5,000 controls as indicative of expected null behavior for a 2% MAF SNP. Under these conditions, the allelic $\chi^2$ and trend tests exhibit similar asymptotic behavior and return empirical significance estimates, which, compared with those obtained by Fisher's exact test, are not appreciably misestimated. Equivalently, we take findings for 40 minor alleles in 5,000 cases and 5,000 controls as representative of a 1%, establishing a reasonable lower limit for the allelic $\chi^2$ and trend tests. The Wald test is particularly sensitive to the number of minor alleles, returning substantially diminished estimates of significance in the genome-wide critical region. At $\alpha < 10^{-6}$, deflation




(a) 40 alleles in 1,000 cases/controls

(b) 40 alleles in 5,000 cases/controls

(c) 20 alleles in 5,000 cases/controls

(d) 80 alleles in 5,000 cases/controls

**FIGURE 1**

**(Colour online) Quantiles for null distributions of the 1*B* allelic $\chi^2$, Cochran–Armitage trend, Wald, and Fisher's exact tests.**

of the Wald test statistics is at least 4, 40, and 400 times greater than for the allelic $\chi^2$ with 80, 40, and 20 minor alleles, respectively. Whereas both the allelic $\chi^2$ and trend tests exhibit inflation in the $\alpha < 10^{-2}$ critical region for 20 copies of the minor allele, the counts for Fisher's exact test are simply reduced compared with 40 or 80 copies, demonstrating the robustness of Fisher's exact test in situations for which traditional tests are not suitable.

### Null Covariate Effect

Table 2 gives the observed number of Wald test statistics for logistic models incorporating a null covariate effect of fixed prevalence among controls; corresponding quantile–quantile plots are displayed in Figure 2. Regression coefficients, $\alpha$ levels, and expectations regarding asymptotic behavior are as described for our previous implementation.

With random assignment of case status, inclusion of the covariate in our regression analysis should not alter the observed distribution of test statistics. While generally true, approximations at the extreme tails appear slightly less deflated for higher prevalence of the covariate among carriers of the minor allele (Figure 2). Strictly speaking, this phenomenon may be best described as countervailing inflation, occurring as a result of increased sampling variance. That

**TABLE 2**

**Number of Significant 1-*df* Wald Statistics for Logistic Regression Models Featuring a Null Covariate**

| Null distribution | | | | | Significance threshold (α) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha < 10^{-1}$ | $\alpha < 10^{-2}$ | $\alpha < 10^{-3}$ | $\alpha < 10^{-4}$ | $\alpha < 10^{-5}$ | $\alpha < 10^{-6}$ | $\alpha < 10^{-7}$ |
| Theoretical $\chi^2$ (1 *df*) | | | | | 100,000,000 | 10,000,000 | 1,000,000 | 100,000 | 10,000 | 1,000 | 100 |
| | Trait risk | | | | | | | | | | |
| Model (~Aff) | Carrier | Pop | $Count_{MA}$ | *N* | | | | | | | |
| $\beta_0 + \beta_1$ SNP | . | . | 40 | 2,000 | 79,179,748 | 6,167,797 | 534,239 | 5,983 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 10% | 10% | | | 79,782,126 | 6,180,674 | 472,267 | 6,065 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 20% | 20% | | | 84,495,240 | 6,409,079 | 409,365 | 6,276 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 40% | 40% | | | 95,357,905 | 7,262,721 | 371,535 | 7,344 | 5 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 60% | 60% | | | 96,940,157 | 7,536,009 | 397,235 | 9,753 | 28 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 80% | 80% | | | 97,342,986 | 7,757,342 | 438,797 | 13,473 | 126 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP | . | . | 80 | 10,000 | 92,269,671 | 9,538,127 | 439,954 | 24,480 | 2,215 | 65 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 10% | 10% | | | 92,271,582 | 9,529,042 | 501,357 | 34,820 | 2,224 | 65 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 20% | 20% | | | 92,781,985 | 9,407,137 | 610,113 | 40,053 | 2,156 | 65 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 40% | 40% | | | 97,423,726 | 8,784,754 | 659,591 | 41,297 | 1,935 | 68 | 2 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 60% | 60% | | | 98,325,325 | 8,741,198 | 668,054 | 42,540 | 1,993 | 80 | 1 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 80% | 80% | | | 98,425,924 | 8,793,612 | 680,545 | 44,148 | 2,197 | 93 | 1 |
| $\beta_0 + \beta_1$ SNP | . | . | 40 | 10,000 | 80,400,565 | 6,368,746 | 178,816 | 7,745 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 10% | 10% | | | 80,405,394 | 6,368,716 | 198,507 | 7,733 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 20% | 20% | | | 80,790,880 | 6,372,733 | 260,393 | 7,543 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 40% | 40% | | | 86,962,981 | 6,598,294 | 343,261 | 6,396 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 60% | 60% | | | 93,681,516 | 7,082,939 | 358,649 | 6,107 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 80% | 80% | | | 96,110,140 | 7,343,465 | 365,042 | 6,841 | 1 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP | . | . | 20 | 10,000 | 115,137,310 | 2,561,766 | 0 | 0 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 10% | 10% | | | 114,999,456 | 2,561,820 | 0 | 0 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 20% | 20% | | | 114,872,450 | 2,565,090 | 0 | 0 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 40% | 40% | | | 113,182,826 | 2,674,582 | 0 | 0 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 60% | 60% | | | 108,060,024 | 3,061,018 | 0 | 0 | 0 | 0 | 0 |
| $\beta_0 + \beta_1$ SNP + $\beta_2$ Cov | 80% | 80% | | | 101,803,263 | 3,564,315 | 0 | 0 | 0 | 0 | 0 |

Note: For each logistic model, let Aff represent outcome (i.e., case status); SNP denotes the additive genotype with respect to the minor allele; and Cov denotes the binary covariate predictor. Carrier and Pop give the risk associated with the binary covariate predictor for carriers of the minor allele and the population, respectively.

is, increasing the covariance between minor allele and covariate is accompanied by a gradual degradation of the discrete-valued function seen for our original logistic model. For very small α, at which approximations of binomial data to the continuous $\chi^2$ distribution are exceptionally poor, this additional variance imparts a slight effect on our probability estimates. Comparison of 40 minor alleles in combined samples of 2,000 and 10,000 cases and controls exemplifies our interpretation; the effect is markedly enhanced in the smaller sample, as would be expected for any sampling effect.

## Discussion

We have demonstrated the tendency of common tests of association to underestimate significance of less common variants, highlighting the inadequacy of current analytical practices for dense SNP and sequencing data. These results show convincingly that common approaches to multiple-test correction will be subject to inflated Type II error rates, particularly within the genome-wide significance levels. We note that although the sampling variance for a 1% allele yields a slightly improved asymptotic distribution (with respect to continuity), estimates of extreme *p*-values are nonetheless deflated.

Meaningful replication of novel findings demands that *p*-values be readily interpretable in the context of the entire catalogue of reported associations, and not subject to across-study differences in study design, sample size, or the number of SNPs actually assayed. Generally accepted estimates of genome-wide α levels are currently of the order of $10^{-8}$, and these will undoubtedly become even smaller as larger numbers of rare variants are tested. With respect to what constitutes an appropriate correction for genome-wide studies, a reasonable assertion is that α levels should reflect the total number of polymorphisms in the genome (Dudbridge & Gusnanto, 2008; Hoggart et al., 2008;). Table 3 gives the number of permutations required to establish significance at various significance thresholds. At the 95% confidence level, our estimates are valid for $\alpha < 10^{-6}$, at which we see a considerable discrepancy between realized and expected test statistic values for 20, 40, and 80 minor alleles. The required number of simulations to attain equivalent precision at current genome-wide α levels is prohibitively large. However, the observed trend in

**TABLE 3**

**Number of Additional Permutations Required to Establish Significance at Given α-Levels**

| Confidence level | Significance threshold (α) $\alpha < 10^{-6}$ | $\alpha < 10^{-7}$ | $\alpha < 10^{-8}$ |
|---|---|---|---|
| 90% | $270.55\times10^6$ | $2.7055\times10^9$ | $27.055\times10^9$ |
| 95% | $384.15\times10^6$ | $3.8415\times10^9$ | $38.415\times10^9$ |
| 99% | $663.49\times10^6$ | $6.6349\times10^9$ | $66.349\times10^9$ |

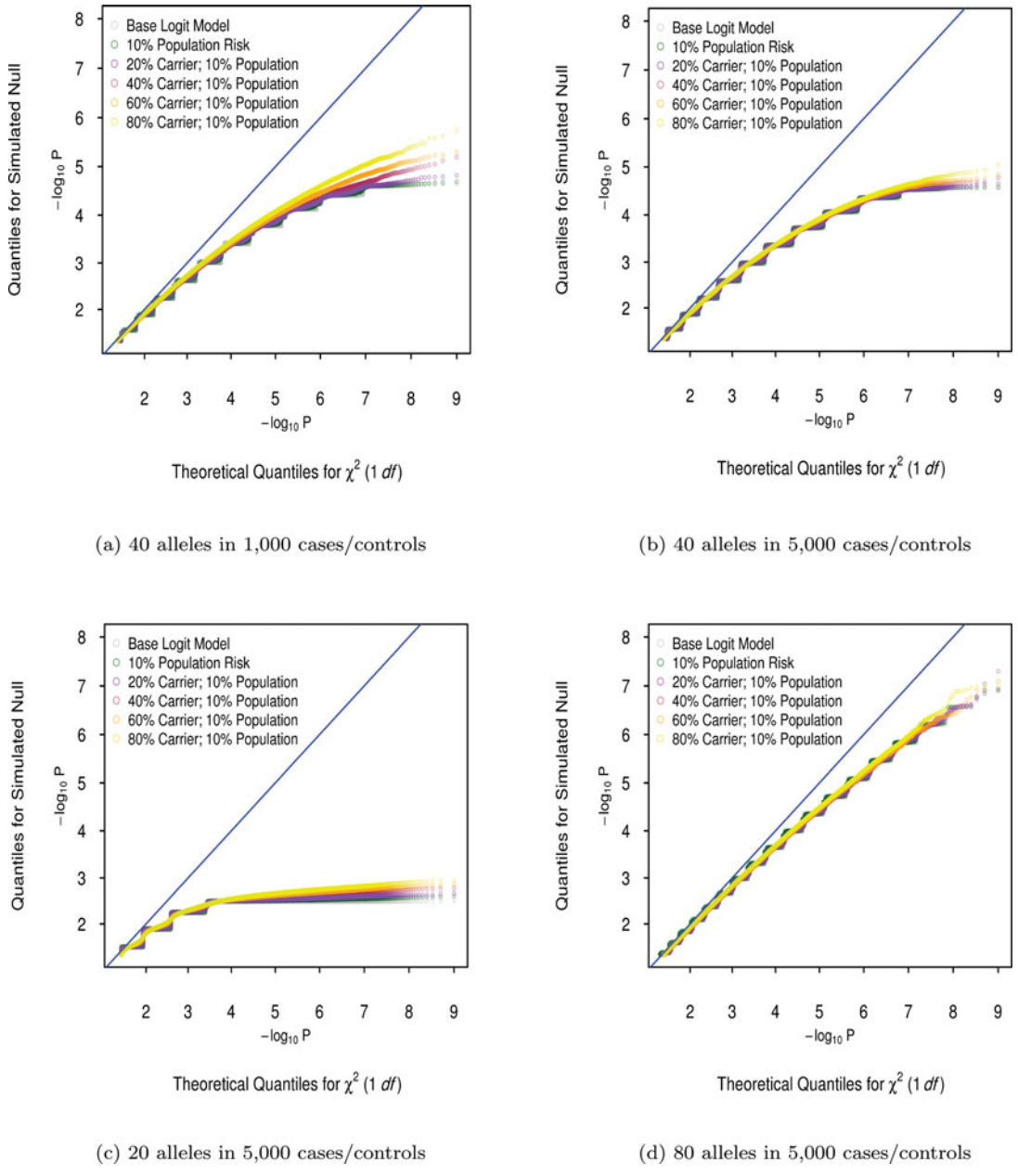



**FIGURE 2**

**(Colour online) Quantiles for null distributions of 1*B* Wald statistics for logistic regression models featuring a null covariate.**

distributional behavior is thoroughly convincing at increasingly stringent significance thresholds.

The appropriate choice of statistical test for analysis of rare variation is not entirely straightforward. Small samples are typically remedied by Yates (1934) correction to the usual $\chi^2$ formula. However, it is well established that the corrected $\chi^2$ yields a conservative estimate of significance (Little, 1989), increasing the likelihood of observing a false negative finding. It may be desirable to obtain an empirical estimate of significance, although permutation procedures are generally computationally intensive. Of major concern with respect to the veracity of reported empirical *p*-values is the choice of an appropriate null distribution. Although relatively straightforward for basic case–control designs (i.e. 'shuffling' of affection status), this might also entail 'regressing out' the effects of confounding factors, or maintaining patterns of linkage disequilibrium in a multi-marker (e.g. 'gene-based') test. Alternatively, Fisher's exact test provides an exact estimate of significance for a given set of values within a contingency table, and is an appropriate method when sample size is limited. Intrinsic differences between these approaches demand careful consideration, with non-negligible consequences for both study design and interpretation of findings. We caution readers against

casual interpretation of exact tests across studies, and recommend that empirical significance for lower-frequency common variants be assessed by permutation.

## Supplementary Material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/thg.2014.17.

## Acknowledgments

We thank Mark Daly for useful feedback and comments on the manuscript. Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org), which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003), along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam.

## References

Apple, R. J., Erlich, H. A., Klitz, W., Manos, M. M., Becker, T. M., & Wheeler, C. M. (1994). HLA DR-DQ associations with cervical carcinoma show papillomavirus-type specificity. *Nature Genetics*, *6*, 157–162.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, *11*, 375–386.

Bush, W. S., Sawcer, S. J., de Jager, P. L., Oksenberg, J. R., McCauley, J. L., Pericak-Vance, M. A., & Haines, J. L. (2010). Evidence for polygenic susceptibility to multiple sclerosis — the shape of things to come. *American Journal of Human Genetics*, *86*, 621–625.

Cohen, J. C., Boerwinkle, E., Mosley, T. H. J., & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, *354*, 1264–1272.

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*, 227–234.

Fisher, R. A. (1922). On the interpretation of $\chi^2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*, 87–94.

Freidlin, B., Zheng, G., Li, Z., & Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human Heredity*, *53*, 146–152.

Hauck Jr., W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, *72*, 851–853.

Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., & Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, *32*, 179–185.

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, *83*, 311–321.

Little, R. J. (1989). Testing the equality of two independent binomial proportions. *American Statistician*, *43*, 283–288.

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, *5*, e1000384.

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., ... Sherry, S. T. (2007). The NCBI dbGAP database of genotypes and phenotypes. *Nature Genetics*, *39*, 1181–1186.

Morgenthaler, S., & 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). *Mutation Research*, *615*, 28–56.

Neale, B. M., Fagerness, J., Reynolds, R., Sobrin, L., Parker, M., Raychaudhuri, S., ... Seddon, J. M. (2010). Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (*LIPC*). *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 7395–7400.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., ... Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, *7*, e1001322.

Nejentsev, S., Walker, N., Riches, D., Egholm, M., & Todd, J. A. (2009). Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, *324*, 387–389.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*, 1516–1517.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics*, *53*, 1253–1261.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., ... Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, *316*, 1341–1345.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., ... Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, *445*, 881–885.

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., ... Hunter, D. J. (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nature Genetics*, *41*, 579–584.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661–678.

Yates, F. (1934). Contingency tables involving small numbers and the $\chi^2$ test. *Supplement to the Journal of the Royal Statistical Society*, *1*, 217–235.