Fast permutation tests and related methods, for association between rare variants and binary outcomes

Arjun Sondhi | Kenneth Martin Rice

Department of Biostatistics, University of Washington, Seattle, WA, USA

Correspondence
Arjun Sondhi, Department of Biostatistics, Box 357232, University of Washington, Seattle, 98195, WA, USA.
Email: asondhi@uw.edu

Abstract
In large-scale genetic studies, a primary aim is to test for an association between genetic variants and a disease outcome. The variants of interest are often rare and appear with low frequency among subjects. In this situation, statistical tests based on standard asymptotic results do not adequately control the type I error rate, especially if the case : control ratio is unbalanced. In this article, we propose the use of permutation and approximate unconditional tests for testing association with rare variants. We use novel analytical calculations to efficiently approximate the true type I error rate under common study designs, and in numerical studies show that the proposed classes of tests significantly improve upon standard testing methods. We also illustrate our methods in data from a recent case–control study for genetic causes of a severe side effect of a common drug treatment.

KEYWORDS
association tests, binary outcomes, permutation tests, rare variants

1 | INTRODUCTION

Association studies are often performed for binary traits, providing new knowledge of the genetic causes of human diseases, using data from case–control and cohort studies (Danjou et al., 2015; Hoffmann et al., 2015; Opherk et al., 2014; Verhaaren et al., 2015). Recent advances in sequencing technology have made it practical to type essentially every variant on the genome; to avoid spurious findings, very low type I error rates must therefore be maintained (Hoggart et al., 2008). However, for the rare variants now being studied, standard analytic approaches do not reliably achieve their nominal rates (Li & Leal, 2008; Ma et al., 2013; Xing, Lin, Wooding, & Xing, 2012) and may permit too many type I errors. The problem can be particularly severe when the ratio of cases to controls is extreme. Adjustments that maintain control at the nominal rate can be conservative, leading to loss of power relative to methods that control type I errors more accurately.

In this article, motivated by work in a case–control study of rhabdomyolysis, we develop methods with improved control of the type I error rate when testing single rare variants for association with binary traits. In Section 2, we explain a novel numerical method that approximates the actual type I error rate of a test statistic given sample size, significance level, and a variant's expected frequency; we also show how the same basic ideas can be used in permutation and approximate unconditional tests and also when adjusting for covariates. In Section 3, we give the results of the numerical studies performed, demonstrating improvements over standard asymptotic tests. Section 4 applies these methods to data from a case–control study of statin-related rhabdomyolysis, and we conclude with a short discussion, including details of an R package that implements our methods.

2 | MATERIALS AND METHODS

With rare variants, homozygotes are rare enough to be negligible for analysis, and it suffices to consider whether subjects have any copies of the variant present ($G = 1$) or not ($G = 0$). This step also means it is simple to enumerate all possible datasets; given fixed numbers of cases and controls ($m_1$ and $m_0$, respectively), we need only consider the number of cases ($r_1$) and controls ($r_0$) with the variant (adjustment for covariates is considered in Section 2.3.).
For a variant with a given minor allele frequency (MAF), following Ma et al. (2013), we define the expected number of minor allele carriers as \( EMAC = (m_0 + m_1) \times (1 - (1 - MAF)^2) \). Under the null hypothesis of no association, it follows that

\[
\begin{align*}
    r_0 &\sim \text{Binom} \left( m_0, \frac{EMAC}{m_0 + m_1} \right) \\
r_1 &\sim \text{Binom} \left( m_1, \frac{EMAC}{m_0 + m_1} \right),
\end{align*}
\]

independently. Therefore, given a value of \( EMAC \)—or equivalently \( MAF \)—and the fixed numbers of cases and controls in the study at hand, we can simply write down the probability of seeing all possible datasets under the null hypothesis.

In theory, this direct enumeration allows exact calculation of the type I error rate for any test: the type I error rate is the sum of the probabilities of the datasets for which a significant test result is returned. Formally, a dataset \( (m_0, m_1, r_0, r_1) \) returns a significant test result when its associated \( P \)-value \( P(r_0, r_1; m_0, m_1) \leq \alpha \), where \( P(r_0, r_1; m_0, m_1) = \mathbb{P}[|T| \geq T_{obs}; H_0] \) is the probability of test statistic \( T \) equaling or exceeding the observed value \( T_{obs} \) when the null hypothesis holds. The type I error rate of the test at nominal level \( \alpha \) is then defined as

\[
T1ER(\alpha) = \sum_{0 \leq r_0 \leq m_0, 0 \leq r_1 \leq m_1} f(r_0, r_1; m_0, m_1, EMAC) I(P(r_0, r_1; m_0, m_1) < \alpha),
\]

where \( f(r_0, r_1; m_0, m_1, EMAC) \) denotes the probability of observing data \( r_0, r_1 \) under the null hypothesis, given \( m_0, m_1 \), and a specific \( EMAC \). Although not discussed further, the approach is easily adapted to \( P \)-values that use lower tail areas—below \( T_{obs} \) instead of above, or two-sided tests that examine tail areas beyond \( \pm T_{obs} \).

In practice, the sum of \( (m_0 + 1) \times (m_1 + 1) \) terms in Eqn 1 may be large, making computation too slow for some purposes. However, for work on rare variants, almost all of the summands contribute negligibly to the overall type I error rate. A practical solution is therefore to truncate the summation in Eqn 1 by zeroing out terms that, in total, represent no more than a small fraction of the type I error rate.

Taking this approach, in our work we will zero out terms in Eqn 1 representing datasets for which \( r_0 + r_1 \) exceeds the upper \( 10^{-12} \) quantile of the distribution of \( r_1 + r_0 \), that is, of \( \text{Binom} \left( m_0 + m_1, \frac{EMAC}{m_0 + m_1} \right) \). Figure 1 gives a graphical description of this process. By setting these terms to zero, in this example, we therefore underestimate the type I error rate by no more than \( 10^{-12} \), which is acceptable given our focus on type I error rates near \( \alpha = 5 \times 10^{-8} \), and maintain practical computation times, even for large studies. For example, performing this calculation for \( (m_0, m_1, EMAC) = (500, 500, 15) \) at \( \alpha = 5 \times 10^{-8} \) using the standard Score test takes 0.05 s on a standard laptop. Without the zeroing-out method, the calculation takes 15.3 s (i.e., more than 300 times faster). Our choice of \( \alpha \) corresponds to testing 1 million independent variants (Pe'er, Yelensky, Altshuler, & Daly, 2008), a level that has been widely adopted as the standard for genome-wide work.

We emphasize that this approximation of the type I error rate is entirely general; any test statistic \( T \) can be used, including the familiar Wald, Score, and likelihood ratio test statistics [see, e.g., Ma et al. (2013)] or more sophisticated choices.
such as the Firth test (Firth, 1993; Heinze, Ploner, Dunkler, & Southworth, 2013). Accurate knowledge of these approximate tests enables users to better compare their performance at the nominal $\alpha$.

The formulation of type I error rate in Eqn 1 and its approximation can also directly inform construction of permutation and approximate unconditional tests, as we discuss in Sections 2.1 and 2.2. We briefly discuss adjusting for covariates in Section 2.3 for both forms of the test.

## 2.1 Permutation tests

The ability of permutation tests to provide accurate $P$-values for association testing under minimal assumptions is well known (Anderson, 2001; Huo, Heyvaert, den Noortgate, & Onghena, 2014; Nichols & Holmes, 2002; Pitman, 1937); where they are applicable, permutation tests are regarded by many analysts as the “gold standard” method. For quantitative traits, a major drawback is that permutations must, in practice, be performed using random number generation (Boyett & Shuster, 1977). For analysis of binary traits, this is not needed; we can instead enumerate all possible permutations and obtain accurate $P$-values.

Using the same notation as above, a permutation test requires an observed test statistic, $T_{\text{obs}}$, calculated on the observed data $(m_0, m_1, r_0, r_1)$. We shall consider test statistics from standard score, Wald, likelihood ratio, and Firth test approaches, thus providing a permutation version of them. The test statistic is also calculated for each possible dataset $(m_0, m_1, r'_{0}, r'_{1})$ obtained by permuting binary outcomes (e.g., case/control labels) among all study subjects or equivalently permuting the variant/nonvariant carrier status among all subjects. Under permutation, the total number of minor allele carriers is the same as in observed data, that is, $r_0 + r_1 = r'_{0} + r'_{1}$, and under the null hypothesis of no association, the probability of observing each dataset follows the hypergeometric distribution (Good, 2005):

$$f(r'_{0}, r'_{1}; m_0, m_1, r_0 + r_1) = \binom{r_0 + r_1}{r'_{1}} \binom{m_0 + m_1 - r_0 - r_1}{m_1 - r'_{1}} \binom{m_0 + m_1}{m_1}.$$

The permutation $P$-value is then defined as

$$P_{\text{perm}}(r_0, r_1; m_0, m_1) = \sum_{r'_0, r'_1} f(r'_{0}, r'_{1}; m_0, m_1, r_0 + r_1) I \left( |T_{0}'| \geq |T_{\text{obs}}| \right),$$

that is, the sum of probabilities of datasets with the same number of allele carriers that result in more extreme test statistics than $T_{\text{obs}}$. The datasets enumerated in this method are illustrated in Figure 2.

Permutation tests are exact in the sense that the observed type I error rate will always be $\leq \alpha$. This result is well known and dates back to Fisher (Janssen & Pauls, 2003). In particular, for the rare variant setting, permutation tests will be fairly conservative. A mathematical explanation is given in the Appendix.

### 2.2 Approximate unconditional tests

Approximate unconditional (AU) tests (Storer & Kim, 1990) provide type I error rates closer to the nominal level than permutation approaches. Unlike permutation tests, AU tests are not guaranteed to always strictly control the type I error rate, but this anticonservatism (where it occurs at all) is usually very mild.

Using the same notation as above, AU tests calculate a test statistic $T_{\text{obs}}$ from the observed data $(m_0, m_1, r_0, r_1)$ and from all possible datasets $(m_0, m_1, r'_{0}, r'_{1})$, but without the restriction that $r_0 + r_1 = r'_{0} + r'_{1}$. The probability of observing each dataset under the null hypothesis is calculated using fitted binomial distributions, that is,

$$f(r'_{0}, r'_{1}; m_0, m_1, r_0 + r_1) = \binom{m_0}{r'_{0}} \left( \frac{r_0 + r_1}{m_0 + m_1} \right)^{r'_{0} + r'_{1}} \left( 1 - \frac{r_0 + r_1}{m_0 + m_1} \right)^{m_0 + m_1 - r'_{0} - r'_{1}}.$$
The AU test’s P-value is then defined as

\[ P_{AU} (r_0, r_1; m_0, m_1) = \sum_{r_0', r_1'} f (r_0', r_1'; m_0, m_1, r_0, r_1) I \left( \left| T_{0,1}' \right| \leq \left| T_{obs} \right| \right) \], \quad (2) \]

that is, the sum of probabilities of datasets that result in more extreme test statistics than \( T_{obs} \). We can then apply Eqn 1 and write the type I error rate as:

\[ T1ER(\alpha) = \sum_{0 \leq r_0' \leq m_0, 0 \leq r_1' \leq m_1} f (r_0', r_1'; m_0, m_1, EMAC) I \left( P_{AU} (r_0', r_1'; m_0, m_1) < \alpha \right) , \quad (3) \]

Compared to the permutation test, the AU test’s P-value sums over many more possible datasets, allowing less crude approximation of the type I error rate. This comes at the cost of using the same data to fit the null binomial models and hence losing guaranteed control of the type I error rate. However, in our setting, a bigger practical concern is that taking a naïve approach to calculation in Eqn 2 would require \((m_0 + 1) \times (m_1 + 1)\) evaluations for each P-value, which may be a burden, as with Eqn 1. A much quicker approach that is still adequate in practice uses the same zeroing-out idea as before—we only sum elements \((r_0', r_1')\) in Eqn 2 for values of \( r_0' + r_1' \) between the upper and lower \(10^{-12}\) quantiles of the \( Binom \left( m_0 + m_1, \frac{r_0 + r_1}{m_0 + m_1} \right) \) distribution.

The datasets enumerated in this method are illustrated in Figure 3. As with the calculation of type I errors in Section 2, the zeroing out leads to a slight understatement of the P-value compared to complete enumeration. However, understating the P-value by at most \(2 \times 10^{-12}\) is a very minor concern when \(\alpha = 10^{-8}\), several orders of magnitude greater, and comes in return for a substantial speed increase. For example, computing an AU P-value under the Score test with data \((m_0, m_1, r_0, r_1) = (5000, 5000, 10, 50)\) takes 0.05 s on a standard laptop with zeroing out and 28.4 s without (i.e., over 500 times faster).

The AU approach, like the permutation approach, is completely general, and AU versions of any test can be implemented. We shall use standard Score, Wald, likelihood ratio, and Firth tests.

### 2.3 Adjusting for covariates

Both permutation and AU tests permit adjustment for covariates through stratification, that is, only using information about association from within groups of subjects for whom confounding factors (e.g., ancestry) are held constant (Clayton, Hills, & Pickles, 1993).

Extending the previous notation, for stratified permutation tests we now refer to vectors \(m_0, m_1, r_0, r_1\), each of length \(q\), where \(q\) is the number of strata defined by the levels of one or more categorical covariates. Indexing strata by \(i\), with \(1 \leq i \leq q\), for each stratum \(i\), the stratified test enumerates all possible strata-specific datasets \((m_{0i}, m_{1i}, r_{0i}, r_{1i})\) such that \(r_{0i} + r_{1i} = r_{0i} + r_{1i}\), computing a test statistic for each. The test statistics

![Figure 3](image-url)
with consequent Monte Carlo error and long compute times. While a limitation of the approach is that a finite number of strata do have to be defined, in practice we have found that stratifying data into commonly used ancestry groups (European, African-American, etc.) is generally sufficient to account for confounding issues present in analyses of complex disease traits.

3 | ANALYTICAL CALCULATION RESULTS

To illustrate analytical calculations, we set the total sample size to be $N = 10,000$—close to that seen in the example in Section 4—and considered case:control matching ratios of 1:1, 1:3, and 1:19, fixing the quantities $m_0$ and $m_1$. For EMAC ranging from 1 to 100, we enumerated all plausible observed data values of $r_0$ and $r_1$ under the null hypothesis of independence between the minor allele and case:control status. A “plausible” dataset here was defined as one where the sum $r_0 + r_1$ was in between the lower and upper $10^{-12}$ quantiles of the null Binomial distribution. Permutation and AU versions of Score, Wald, likelihood ratio, and Firth tests were examined. For comparison, we also computed the standard Score, Wald, likelihood ratio, Firth tests, and Fisher’s exact test, which is itself a permutation test. For permutation, AU, and standard tests, we also considered a regularized Wald test, which avoids undefined test statistics by adding 0.5 to each cell count when any count is zero. We set the nominal significance level at $\alpha = 5 \times 10^{-8}$ and calculated the type I error rate for each test as in Eqn 1 across the range of EMAC. Although we do not consider power calculations under an alternative hypothesis in this section, these are given in Appendix B for the AU Firth test.

As seen in Figure 4, the tests based on standard asymptotics do not adequately control the type I error rate. In the balanced design, the tests are overly conservative, with the exception of the likelihood ratio test, which is anticonservative. The Score test has very a large type I error rate under the 1:3 ratio, so it is presented separately. This is also true under the 1:19 ratio for the Score and Wald tests, which are omitted. The other tests continue to be conservative, and the likelihood ratio test’s type I error rate is too large over certain ranges. The Firth test consistently performs the best.

In Figure 5, we see that the permutation tests improve upon most of the standard tests, though remain more conservative than the regular Firth test. While these tests have the advantage of being exact, as the case:control ratio becomes more unbalanced, the type I error rate becomes more conservative. Under the 1:19 ratio, all tests perform nearly identically, with the exception of the unregularized Wald test.

In Figure 6, we see that the AU tests show a large improvement over standard and permutation tests, especially in the AU
likelihood ratio and AU Firth tests. Though they are not exact, the excess type I error rate is mild. Note that under the 1:19 ratio, the Firth and likelihood ratio tests perform identically.

4 APPLICATION: RHABDOMYOLYSIS CASE–CONTROL STUDY

The data comes from an exome-sequencing study, in which 9,763 subjects who used statins were considered (211 cases with rhabdomyolysis and 9,552 controls). The original dataset consists of 2,194,116 variants sequenced. The rationale for this design is described in detail by Marciante et al. (2011). Our interest was primarily in assessing if there existed rare genetic variants associated with developing rhabdomyolysis in statin users. We defined “rare” variants as those where ≤100 study participants carried the minor allele. Variants with less than five minor allele carriers were also removed, as these provide no ability to produce significant values at the low α threshold used in this form of study. Finally, for quality control, we filtered out variants with a genotyping rate of <0.85. We did not filter variants using the Hardy-Weinberg test. Applying these filters left 161,428 variants, and there are no covariates for which to adjust in this analysis.

We applied the AU and permutation versions of the likelihood ratio test, and the permutation and standard Firth test to all variants. The entire analysis took approximately 6.5 h on a shared server, using a single CPU. The AU version of the Firth test was not used due to its high computational burden. The resulting QQ plot and a plot of the inflation (45-degree rotated QQ plot) observed are given in Figure 7.

Based on our numerical results, it is reasonable to expect that the AU likelihood ratio test provides the best control of the type I error rate. Applied to this dataset, while some granularity in the larger P-values is present on the left-hand side of both plots, we observe that the AU likelihood ratio test results in significantly less inflation than the standard Firth test. Therefore, although no variant was found to have an AU P-value < α = 5 × 10⁻⁸, we believe that the variants that deviate from the expected null P-value distribution under the AU test are more likely to be of interest than those under the other methods.
5 | DISCUSSION

We have developed and implemented association tests for rare genetic variants that control type I error rates better than standard asymptotic tests. Of the tests proposed in this article, the AU version of the likelihood ratio and Firth tests perform the best, particularly when the ratio of cases to controls is extreme. However, the AU version of the Firth test has a notably higher computational burden than competitors. Therefore, we recommend the AU likelihood ratio test for large genome-wide studies. If an exact test is necessary, then a permutation test is recommended; though conservative, it tends to show an improvement over standard tests. We note that if the expected number of minor allele carriers is <20, then no test will perform adequately, and conservative control of the type I error rate is the best achievable property. More generally, we note that in high-throughput rare variant work, unless effect sizes are large, power will be limited even when type I error rates are not controlled conservatively. As pointed out by a reviewer, in such settings, an appealing alternative form analysis tests for association across a group of variants; examples include SKAT (Wu et al., 2011) and “burden” tests, which collapse genotypes across a region to a univariate measure. These approaches can improve power over single-variant approaches, by both combining multiple association “signals” and by reducing the multiple testing burden. However, the inference they provide is region specific and not variantspecific, leaving (for example) no strong indication of which variants may be causal. While not explored here, our permutation tests could be used with CAST (Morgenthaler & Thilly, 2007), a burden test that collapses genotype to presence/absence of a particular class of variants.

The methods described here have been implemented in an R package, AUtests. This package contains the functions basic.tests, perm.tests, and au.tests, which implement all the respective standard, permutation, and AU tests for a given vector of counts \((m_0, m_1, r_0, r_1)\), returning a vector of \(p\)-values. The AU Firth test is implemented in a separate function, au.firth, due to its increased computational time. For a typical dataset \((m_0, m_1, r_0, r_1) = (10000, 10000, 50, 50)\), on a standard laptop, the basic.tests function takes 0.03 s of CPU time, the perm.tests function takes 0.21 s, the au.tests function takes 0.39 s, and the au.firth function takes 51 s. To account for covariates, appropriately categorized, the package also contains the functions au.test.strat and perm.test.strat, which implement stratified AU and permutation likelihood ratio tests. The package is available on CRAN.
ACKNOWLEDGMENTS

Research reported in this article was supported by the National Institute on Aging of the National Institutes of Health under award numbers U01AG049505 and U01AG049507, and by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL078888. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Method development: KMR, AS; Method implementation: AS; Data analysis: AS, Writing: AS, KMR

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

ORCID

Arjun Sondhi http://orcid.org/0000-0002-7997-0477

REFERENCES

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. Canadian Journal of Fisheries and Aquatic Sciences, 58, 626–639.

Boyett, J. M., & Shuster, J. J. (1977). Nonparametric one-sided tests in multivariate analysis with medical applications. Journal of the American Statistical Association, 72, 665–668.

Clayton, D., Hills, M., & Pickles, A. (1993). Statistical models in epidemiology, volume 1. Oxford University Press, Oxford, UK.

Danjou, F., Zoledziewska, M., Sidore, C., Steri, M., Busonero, F., Mascio, A., ... Cucca, F. (2015). Genome-wide association analyses based on whole-genome sequencing in sardinia provide insights into regulation of hemoglobin levels. Nature Genetics, 47, 1264–1271.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80, 27–38.

Good, P. (2005). Permutation, parametric and bootstrap tests of hypotheses. New York: Springer.

Heinze, G., Ploner, M., Dunkler, D., & Southworth, H. (2013). logistf: Firth’s bias reduced logistic regression. R package version 1.21. Available at https://cran.r-project.org/package=logistf.

Hoffmann, T. J., Van Den Eeden, S. K., Sakoda, L. C., Jorgenson, E., Habel, L. A., Graff, R. E., ... Witte, J. S. (2015). A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer Discovery, 5, 878–891.

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.

Huo, M., Heyvaert, M., den Noortgate, W. V., & Onghena, P. (2014). Permutation tests in the educational and behavioral sciences. Methodology, 10, 43–59.

Janssen, A., & Pauls, T. (2003). How do bootstrap and permutation tests work? Annals of Statistics, 31, 768–806.

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(3), 311–321.

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J., and the GoT2D investigators (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genetic Epidemiology, 37, 539–550.

Marcianite, K. D., Durda, J. P., Heckbert, S. R., Lumley, T., Rice, K., McKnight, B., ... Psaty, B. M. (2011). Cerivastatin, genetic variants, and the risk of rhabdomyolysis. Pharmacogenetics and Genomics, 21, 280.

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/Fundamental and Molecular Mechanisms of Mutagenesis, 615, 28–56.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping, 15, 1–25.

Opherk, C., Gonik, M., Duering, M., Malik, R., Jouvent, E., Hervé, D., ... Dichgans, M. (2014). Genome-wide genotyping demonstrates a polygenic risk score associated with white matter hyperintensity volume in casadis. Stroke, 45, 968–972.

Pe’er, I., Yelenksy, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic epidemiology, 32, 381–385.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. Supplement to the Journal of the Royal Statistical Society, 4, 119–130.

Storer, B. E., & Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. Journal of the American Statistical Association, 85, 146–155.

Verhaaren, B. F., Debette, S., Bis, J. C., Smith, J. A., Ikram, M. K., Adams, H. H., ... Fornage, M. (2015). Multiethnic genome-wide association study of cerebral white matter hyperintensities on mri. Circulation: Cardiovascular Genetics, 8, 398–409.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics, 89, 82–93.

Xing, G., Lin, C.-Y., Wooding, S. P., & Xing, C. (2012). Blindly using Wald’s test can miss rare disease-causal variants in case-control association studies. Annals of Human Genetics, 76, 168–177.

How to cite this article: Sondhi A, Rice KM. Fast permutation tests and related methods, for association between rare variants and binary outcomes. Ann Hum Genet. 2018;82:93–101. https://doi.org/10.1111/ahg.12229
APPENDIX A: EXACT CONTROL OF PERMUTATION TESTS

In this appendix, we show that permutation tests give exact control of the type I error rate. Rewriting Eqn 1 as a double summation over the observed minor allele count \( t = r_0 + r_1 \) and the number of these in the controls, we obtain

\[
T1ER(\alpha) = \sum_{0 \leq t \leq m_0 + m_1} g(t; m_0, m_1) \times \sum_{r_0} f(r_0; m_0, m_1, t) I \left( P(r_0, r_1; m_0, m_1) < \alpha \right),
\]

where \( g() \) denotes the probability of the observed minor allele count, and \( f() \) gives the probability of the observed counts in cases and controls given the minor allele count \( r_0 + r_1 \) so \( f() \) supports values of \( r_0 \) between \( \max(0, t - m_1) \) and \( \min(m_0, t) \).

By construction, the inner sum always gives a value \( \leq \alpha \); the outer sum averages these, and so is similarly bounded. However, particularly for rare variants, the inner sum considers a small set of possible permutations, as illustrated in Figure 2.

While this makes the test fast enough that zeroing out is not required, it means that for small \( \alpha \), the actual type I error rate, while below \( \alpha \), will be quite conservative for many values of \( m_0 \) and \( m_1 \).

APPENDIX B: AU TEST POWER CALCULATIONS

In this section, we show power calculations for the AU Firth test, under the same matching ratios considered in the main paper and over the same range of \( EMAC \). Specifically, we perform analytical calculations similarly as in Section 3 but consider data probabilities from a binomial distribution where having a minor allele is related to case:control status through an odds ratio. We consider odds ratios ranging from 1.5 to 5. In Figure 8, we observe that power decreases as case:control ratios become more skewed. In particular, the extremely unbalanced 1:19 ratio requires a very large association in order to have meaningful power, even at higher minor allele counts.

![Figure 8](https://example.com/figure8.png)

**Figure 8** Power curves giving the probability of rejecting the null hypothesis of independence by expected minor allele count. Different curves correspond to different odds ratios. Each panel corresponds to a different case:control matching ratio with an overall sample size of 20,000. Left: 1:1, middle: 1:3, right: 1:19 [Colour figure can be viewed at wileyonlinelibrary.com]