Published in final edited form as:

Genet Epidemiol. 2014 May ; 38(4): 325–344. doi:10.1002/gepi.21805.

Valid Monte Carlo Permutation Tests for Genetic Case-Control Studies with Missing Genotypes

Daniel D. Kinnamon^{1,2} and Eden R. Martin^{2,*}

¹Division of Human Genetics, Department of Internal Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA

²Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, USA

Abstract

Monte Carlo permutation tests can be formally constructed by choosing a set of permutations of individual indices and a real-valued test statistic measuring the association between genotypes and affection status. In this paper, we develop a rigorous theoretical framework for verifying the validity of these tests when there are missing genotypes. We begin by specifying a nonparametric probability model for the observed genotype data in a genetic case-control study with unrelated subjects. Under this model and some minimal assumptions about the test statistic, we establish that the resulting Monte Carlo permutation test is exact level α if (1) the set of permutations of individual indices is a group under composition and (2) the distribution of the observed genotype score matrix under the null hypothesis does not change if the assignment of individuals to rows is shuffled according to an arbitrary permutation in the chosen set. We apply these conditions to show that frequently used Monte Carlo permutation tests based on the set of all permutations of individual indices are guaranteed to be exact level α only for missing data processes satisfying a rather restrictive additional assumption. However, if the missing data process depends on covariates that are all identified and recorded, we also show that Monte Carlo permutation tests based on the set of permutations within strata of individuals with identical covariate values are exact level α . Our theoretical results are verified and supplemented by simulations for a variety of missing data processes and test statistics.

Introduction

Monte Carlo permutation tests have been employed by many authors proposing genetic case-control association methods involving complex test statistics, rare variants, or both (see, e.g., [Chapman and Whittaker, 2008; Hoffmann et al., 2010; Hoh and Ott, 2000, 2003; Hoh et al., 2001; Ionita-Laza et al., 2011; Kinnamon et al., 2012; Madsen and Browning, 2009; Neale et al., 2011; Price et al., 2010]). Such Monte Carlo permutation tests can be formally constructed by choosing a set of permutations of individual indices and a real-valued test

*Correspondence to: Eden R. Martin, Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, P.O. Box 019132 (M-860), Miami, Florida 33101, (305) 243-2372 (office), (305) 243-2396 (fax), emartin1@med.miami.edu.

The authors have no conflicts of interest to declare.

statistic measuring the association between genotypes and affection status. Permutations are randomly sampled uniformly and with replacement from this set and applied to the indices of individuals' affection status indicators as a computational shortcut for permuting intact observed genotype score vectors among individuals. The test statistic in each resulting data set is recorded, and the null hypothesis of no association is then rejected if the p-value calculated using the discrete uniform distribution over all recorded test statistics, including the one in the original data set [Davison and Hinkley, 1997; Ernst, 2004; Phipson and Smyth, 2010], is less than or equal to the nominal α .

The primary attraction of such permutation tests is that, if the set of permutations is correctly chosen for the testing problem at hand, the resulting tests are exact level α under very few assumptions about the data generating process and test statistic [Lehmann and Romano, 2005; Romano, 1989]. In the present study, a test that is exact level α will be understood to have a type I error rate that is guaranteed to be at most the nominal α under the null hypothesis, regardless of the sample size. In a genetic context, permutation tests implemented as described above also preserve linkage disequilibrium between variants, which facilitates efficient joint inference over a large number of variants.

The standard Monte Carlo permutation tests most frequently used in genetic case-control studies with unrelated subjects are based on the set of all possible permutations of individual indices. However, it is unclear whether this set of permutations is appropriate when the data generating process results in missing genotypes. To illustrate the problem, consider a hypothetical genetic case-control study in which affection status is always observed but genotypes at some variants are missing. Suppose first that true genotypes were sampled according to a null disease model and that observed genotypes were randomly missing—that is, the probability of a missing genotype at a variant did not depend on other observed or unobserved variables in the study. In this case, individuals' observed genotype score vectors would be independent and identically distributed, which would imply that the distribution of the observed genotype score matrix containing these vectors as rows would be the same for any ordering of individuals. Thus, the distribution of the observed genotype score matrix would be invariant under permutations of its row indices, which would imply that Monte Carlo permutation tests based on the set of all permutations of individual indices would be exact level α under this missing data process [Lehmann and Romano, 2005]. However, if genotypes were missing at different rates or in different patterns across individuals because of differing data collection or assay techniques, the observed genotype score vectors would have several different distributions despite the fact that the true genotype score vectors were identically distributed. Therefore, the distribution of the observed genotype score matrix would depend on the ordering of individuals in its rows and not necessarily be invariant under permutations of its row indices. Under this missing data process, standard Monte Carlo permutation tests might have a type I error rate exceeding α .

Despite this ambiguity, it is difficult to find a method for verifying that a set of permutations yields Monte Carlo permutation tests that are exact level α with missing data. Some authors have considered permutation tests with missing data [Naddeo, 2002; Pesarin and Salmaso, 2010], but their formulations of the testing problem require assuming that, conditional on being observed, genotypes are identically distributed in cases and controls for all missing

data processes under a null disease model. Other authors provide solutions that rely on modified test statistics or data transformations to address missing data in specific applications [Entsuah, 1990; Good, 2000]. Previous studies in other fields have also presented empirical finite-sample results regarding the performance of specific types of permutation tests with missing values [Kennes et al., 2012; Mundry, 1999]. In genetics, one previous study has suggested using the set of permutations within similarly genotyped strata for quantitative trait locus (QTL) mapping in experimental crosses under selective genotyping [Manichaikul et al., 2007]. However, this development is limited to the case where a small number of missing genotype patterns arise nonrandomly by design. For case-control studies, we have previously empirically demonstrated the validity of standard Monte Carlo permutation tests based on three different test statistics when genotypes are randomly missing [Kinnamon et al., 2012]. While informative, none of these analyses provides a broadly applicable method for theoretically verifying that a particular set of permutations yields exact level- α Monte Carlo permutation tests under general missing data processes.

In the following sections, we develop a rigorous theoretical framework for verifying the validity of Monte Carlo permutation tests based on a particular set of permutations when there are missing genotypes. We begin by specifying a nonparametric probability model for the observed genotype data in a genetic case-control study with unrelated subjects. Under this model and some minimal assumptions about the test statistic, we establish that the resulting Monte Carlo permutation test is exact level α if (1) the set of permutations of individual indices is a group under composition and (2) the distribution of the observed genotype score matrix under the null hypothesis does not change if the assignment of individuals to rows is shuffled according to an arbitrary permutation in the chosen set. We apply these conditions to show that the standard Monte Carlo permutation tests most frequently used in practice are guaranteed to be exact level α only for missing data processes satisfying a rather restrictive additional assumption. However, if the missing data process depends on covariates that are all identified and recorded, we also show that Monte Carlo permutation tests based on the set of permutations within strata of individuals with identical covariate values are exact level α . Our theoretical results are verified and supplemented by simulations for a variety of missing data processes and test statistics.

Methods

Nonparametric Probability Model

A nonparametric probability model for the data generating process in a genetic case-control study underpins subsequent theoretical developments. In the following exposition, an uppercase script letter (e.g., $\mathcal{X}$) denotes a set, and $\mathcal{X}^n = \underbrace{\mathcal{X} \times \cdots \times \mathcal{X}}_n$ is the Cartesian product

taken over n copies of the same set. $I[\text{expression}]$ denotes the indicator function taking a value of 1 when the expression in brackets is true and 0 when it is false.

In a genetic case-control study, each individual is selected from the same population q based on a known affection status $a = I[\text{Individual is affected}]$. Possible confounding covariates are recorded at the time of ascertainment, and observed genotypes are subsequently obtained by some assaying process, the characteristics of which may also be recorded as covariates.

The recorded values of these discrete and/or continuous covariates in an individual are represented by a vector $\vec{x} = [x_1, \dots, x_p]$, and each individual's affection status and recorded covariates together constitute a fixed vector $[a, \vec{x}]$. The random quantity that is realized during the course of the study is the individual's vector of observed genotypes. We now derive the distribution of this quantity conditional on the fixed vector $[a, \vec{x}]$ for an individual from a particular population q .

Let $\vec{G} = [G_1, \dots, G_v]$ be a random row vector representing an individual's genotype scores at v fixed genomic positions of interest. Each genotype score G_j takes on a countable number of possible values in the space $\mathcal{G}$ and may be correlated with other genotype scores in $\vec{G}$ due to linkage disequilibrium. Following Little and Rubin, let $\vec{M} = [M_1, \dots, M_v]$ represent a random row vector of missing genotype indicators, $M_j = I[G_j \text{ missing}]$ [Little and Rubin, 2002]. The complete data vector for an individual, $[\vec{G}, \vec{M}] = [G_1, \dots, G_v, M_1, \dots, M_v]$, is then a random row vector taking values in the sample space $\mathcal{Y} = \mathcal{G}^v \times \{0, 1\}^v$. Let $f_{\vec{G}, \vec{M}}([\vec{g}, \vec{m}] | a, \vec{x}, q)$ denote the conditional probability mass function (PMF) of the complete data vector $[\vec{G}, \vec{M}]$ over $\mathcal{Y}$, where $[\vec{g}, \vec{m}] = [g_1, \dots, g_v, m_1, \dots, m_v]$ represents an arbitrary vector in $\mathcal{Y}$. This PMF describes both the distribution of true genotype scores as well as the missing data process that determines which ones are observed.

The observed genotype score vector $\vec{G}_{\text{obs}}$ for an individual is a coarsening of the genotype score vector $\vec{G}$ that results from setting each G_j to missing according to the value of its missing data indicator M_j [Little and Rubin, 2002]. One can imagine this coarsening resulting from a selection model [Little and Rubin, 2002]. First, the value of $\vec{G}$ is determined during sampling of subjects from q . The value of $\vec{G}_{\text{obs}}$ is then determined based on the value of $\vec{M}$ subsequently realized during genotype data collection. Formally, let h be an arbitrary missing genotype code not equal to any valid genotype score in $\mathcal{G}$ (e.g., -99 if genotype scores equal the number of minor alleles) and define:

$$\begin{aligned} \vec{G}_{\text{obs}} &= \eta([\vec{G}, \vec{M}]) \\ &= [G_1 I[M_1 = 0] + h I[M_1 = 1], \dots, G_v I[M_v = 0] + h I[M_v = 1]] \end{aligned}$$

Here, each observed genotype score $G_{\text{obs},j}$ takes values in the sample space $\mathcal{H} = \mathcal{G} \cup \{h\}$, and each observed genotype score vector $\vec{G}_{\text{obs}}$ takes values in the sample space $\mathcal{X} = \mathcal{H}^v$. Thus, the mapping $\eta : \mathcal{Y} \rightarrow \mathcal{X}$ takes vectors $[\vec{g}, \vec{m}]$ in the sample space $\mathcal{Y}$ to vectors $\vec{g}_{\text{obs}}$ in the sample space $\mathcal{X}$. For an arbitrary vector $\vec{g}_{\text{obs}} = [g_{\text{obs},1}, \dots, g_{\text{obs},v}]$ in the sample space $\mathcal{X}$, the value of the inverse image $\eta^{-1}(\vec{g}_{\text{obs}})$ is the set of one or more vectors $[\vec{g}, \vec{m}]$ in the

sample space $\mathcal{Y}$ that map to $\vec{\mathbf{g}}_{\text{obs}}$ through $\eta: \mathcal{Y} \rightarrow \mathcal{X}$. This set comprises all vectors $[\vec{\mathbf{g}}, \vec{\mathbf{m}}]$ in the sample space $\mathcal{Y}$ having the same missing data pattern and the same genotype scores for all observed genotypes, so:

$$\begin{aligned} \eta^{-1}(\vec{\mathbf{g}}_{\text{obs}}) &= \{[\vec{\mathbf{g}}, \vec{\mathbf{m}}] \in \mathcal{Y} : \eta([\vec{\mathbf{g}}, \vec{\mathbf{m}}]) = \vec{\mathbf{g}}_{\text{obs}}\} \\ &= \left\{ [\vec{\mathbf{g}}, \vec{\mathbf{m}}] \in \mathcal{Y} : \begin{array}{l} g_j \in \mathcal{G}, m_j = 1 \forall j \text{ with } g_{\text{obs},j} = h; \\ g_j = g_{\text{obs},j}, m_j = 0 \forall j \text{ with } g_{\text{obs},j} \neq h \end{array} \right\} \end{aligned}$$

Therefore, the conditional PMF of the observed genotype score vector given an individual's fixed vector of affection status and covariate values, $[a, \vec{\mathbf{x}}]$, and population q is:

$$\begin{aligned} P(\vec{\mathbf{G}}_{\text{obs}} = \vec{\mathbf{g}}_{\text{obs}} | a, \vec{\mathbf{x}}, q) &= P(\eta([\vec{\mathbf{G}}, \vec{\mathbf{M}}]) = \vec{\mathbf{g}}_{\text{obs}} | a, \vec{\mathbf{x}}, q) \\ &= P([\vec{\mathbf{G}}, \vec{\mathbf{M}}] \in \eta^{-1}(\vec{\mathbf{g}}_{\text{obs}}) | a, \vec{\mathbf{x}}, q) \\ &= \sum_{\eta^{-1}(\vec{\mathbf{g}}_{\text{obs}})} f_{\vec{\mathbf{G}}, \vec{\mathbf{M}}}([\vec{\mathbf{g}}, \vec{\mathbf{m}}] | a, \vec{\mathbf{x}}, q) \end{aligned} \quad (1)$$

For a genetic case-control study with n total individuals, let $\mathbf{a} = [a_1, \dots, a_n]^T$ be the nonrandom affection status indicator column vector containing a_i for individual i in row i and $\mathbf{x} = [\vec{\mathbf{x}}_1^T, \dots, \vec{\mathbf{x}}_n^T]^T$ be the nonrandom covariate matrix containing the row vector $\vec{\mathbf{x}}_i$ for individual i in row i . The observed genotype data for this study is represented by the observed genotype score matrix $\mathbf{G}_{\text{obs}} = [\vec{\mathbf{G}}_{\text{obs},1}^T, \dots, \vec{\mathbf{G}}_{\text{obs},n}^T]^T$ containing the random observed genotype score vector $\vec{\mathbf{G}}_{\text{obs},i}$ for individual i in row i . This matrix takes values $\mathbf{g}_{\text{obs}} = [\vec{\mathbf{g}}_{\text{obs},1}^T, \dots, \vec{\mathbf{g}}_{\text{obs},n}^T]^T$ in the sample space $\mathcal{X}^n$, where each row i contains an arbitrary vector $\vec{\mathbf{g}}_{\text{obs}}$ from the sample space $\mathcal{X}$. Because individuals in a genetic case-control study are unrelated, the distribution of $\vec{\mathbf{G}}_{\text{obs},i}$ conditional on $[a_i, \vec{\mathbf{x}}_i]$ and q does not depend on the observed genotype score vector in any other individual. Therefore, each $\vec{\mathbf{G}}_{\text{obs},i}$ is independently distributed according to the individual's conditional PMF in (1), and the conditional PMF of the observed genotype score matrix $\mathbf{G}_{\text{obs}}$ given the fixed matrix $[\mathbf{a}, \mathbf{x}]$ and population q is:

$$\begin{aligned}
 P(\mathbf{G}_{\text{obs}} = \mathbf{g}_{\text{obs}} | \mathbf{a}, \mathbf{x}, q) &= \prod_{i=1}^n P(\vec{\mathbf{G}}_{\text{obs}, i} = \vec{\mathbf{g}}_{\text{obs}, i} | a_i, \vec{\mathbf{x}}_i, q) \\
 &= \prod_{i=1}^n \sum_{\eta^{-1}(\vec{\mathbf{g}}_{\text{obs}, i})} f_{\vec{\mathbf{G}}, \vec{\mathbf{M}}}(\vec{\mathbf{g}}, \vec{\mathbf{m}} | a_i, \vec{\mathbf{x}}_i, q)
 \end{aligned}$$

This conditional PMF describes the distribution of the observed genotype score matrix $\mathbf{G}_{\text{obs}}$ for a specific population q and a group of individuals with known characteristics $[\mathbf{a}, \mathbf{x}]$. However, this population is one of many possible populations in a space $\mathcal{Q}$ that arose under an unknown evolutionary process [Weir, 1996]. Because $f_{\vec{\mathbf{G}}, \vec{\mathbf{M}}}$ necessarily varies across populations q having different frequencies for the haplotypes defined by the v fixed genomic positions of interest, we would like to describe the marginal distribution of the observed genotype score matrix $\mathbf{G}_{\text{obs}}$ under the evolutionary process rather than its conditional distribution in a specific population q . Let f_Q denote the PMF for the random population Q over the space of possible populations $\mathcal{Q}$ under the unknown evolutionary process. The marginal PMF of the observed genotype score matrix $\mathbf{G}_{\text{obs}}$ for a group of individuals with known characteristics $[\mathbf{a}, \mathbf{x}]$ under this evolutionary process is:

$$\begin{aligned}
 \mathbb{F}(\mathbf{g}_{\text{obs}} | \mathbf{a}, \mathbf{x}) &= \sum_{\mathcal{Q}} P(\mathbf{G}_{\text{obs}} = \mathbf{g}_{\text{obs}} | \mathbf{a}, \mathbf{x}, q) f_Q(q) \\
 &= \sum_{\mathcal{Q}} \left[\prod_{i=1}^n \sum_{\eta^{-1}(\vec{\mathbf{g}}_{\text{obs}, i})} f_{\vec{\mathbf{G}}, \vec{\mathbf{M}}}(\vec{\mathbf{g}}, \vec{\mathbf{m}} | a_i, \vec{\mathbf{x}}_i, q) \right] f_Q(q) \quad (2)
 \end{aligned}$$

The PMF $\mathbb{F}$ describes the entire stochastic process that generated the observed genotype data in a genetic case-control study with unrelated individuals and missing genotypes.

The PMF $f_{\vec{\mathbf{G}}, \vec{\mathbf{M}}}$ can be factored according to a selection model [Little and Rubin, 2002]:

$$f_{\vec{\mathbf{G}}, \vec{\mathbf{M}}}(\vec{\mathbf{g}}, \vec{\mathbf{m}} | a, \vec{\mathbf{x}}, q) = f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | a, \vec{\mathbf{x}}, q) f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, a, \vec{\mathbf{x}}, q)$$

The PMF $f_{\vec{\mathbf{G}}}$ is the genotype sampling process, and $f_{\vec{\mathbf{M}}}$ is the missing data process. The present study focuses on the impact of the missing data process on inference in the absence of other confounders such as population stratification. In such situations, the genotype sampling process does not depend on an individual's covariates. Also assume that the missing data process does not depend intrinsically on an individual's disease status; that is, any dependency of missing data indicators on affection status arises only through the association, if any, of affection status with an individual's true genotypes and covariates. It follows that:

$$f_{\vec{\mathbf{G}}, \vec{\mathbf{M}}}(\vec{\mathbf{g}}, \vec{\mathbf{m}} | a, \vec{\mathbf{x}}, q) = f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | a, q) f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, \vec{\mathbf{x}}, q)$$

In a genetic case-control study, the genotype sampling process is given by:

$$f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | a, q) = \frac{f_A(a | \vec{\mathbf{g}}, q) f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | q)}{\sum_{\mathcal{G}^V} f_A(a | \vec{\mathbf{g}}, q) f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | q)}$$

where f_A is the population disease model and $f_{\vec{\mathbf{G}}}$ is the unconditional genotype sampling process for population q . The null hypothesis specifies that the evolutionary process results in a population disease model that does not depend on the genotypes at the genomic positions of interest, which implies that $f_A(a | \vec{\mathbf{g}}, q) = f_A(a | q)$ and $f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | a, q) = f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | q)$ for all q . In other words, the genotype sampling process does not depend on affection status for all populations q under the null hypothesis. Based on (2) and the above assumptions, the PMF of the observed genotype score matrix $\mathbf{G}_{\text{obs}}$ under the null hypothesis for a group of individuals with known characteristics $[\mathbf{a}, \mathbf{x}]$ is:

$$\mathbb{F}_0(\mathbf{g}_{\text{obs}} | \mathbf{a}, \mathbf{x}) = \sum_{\mathcal{Q}} \left[\prod_{i=1}^n \sum_{\eta^{-1}(\vec{\mathbf{g}}_{\text{obs}, i})} f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | q) f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, \vec{\mathbf{x}}, q) \right] f_Q(q) \quad (3)$$

The PMF $\mathbb{F}_0$ describes the data generating process in a genetic case-control study with unrelated individuals and missing genotypes under the null hypothesis.

Constructing Monte Carlo Permutation Tests

Let $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ represent an arbitrary permutation of individual indices, which are simply the integers from 1 to n . The composite of two arbitrary permutations $\pi_2 \circ \pi_1$ is the result of first applying permutation π_1 and then applying permutation π_2 to the result; that is, $(\pi_2 \circ \pi_1)(i) = \pi_2(\pi_1(i))$. The composite $\pi_2 \circ \pi_1$ is also a permutation, and every π has a unique inverse permutation π^{-1} such that $(\pi \circ \pi^{-1})(i) = (\pi^{-1} \circ \pi)(i) = i$ for all i [Robinson, 2003]. Permutations are applied to rows of data vectors and matrices through the function u_π , which puts row $\pi(i)$ of its argument in row i of its result of the same dimensions. Thus, for an arbitrary affection status indicator vector $\mathbf{a}$, $u_\pi(\mathbf{a}) = [a_{\pi(1)}, \dots, a_{\pi(n)}]^T$.

Monte Carlo permutation tests for genetic case-control studies can be formally constructed by choosing a set of permutations of individual indices as well as a real-valued test statistic that measures the association between genotypes and affection status. Let $\mathcal{S}$ represent an arbitrary set of such permutations and $t(\mathbf{G}_{\text{obs}}, \mathbf{a}) = t(\left[\vec{\mathbf{G}}_{\text{obs}, 1}, a_1\right], \dots, \left[\vec{\mathbf{G}}_{\text{obs}, n}, a_n\right])$ represent an

arbitrary real-valued test statistic that measures the association between $\vec{\mathbf{G}}_i$ and a_i using the observed genotype data. Without loss of generality, assume that t takes on larger values for matrices $\mathbf{g}_{\text{obs}}$ in the sample space $\mathcal{X}^n$ that are more inconsistent with the null hypothesis of no association. We also make the reasonable assumption that t is symmetric in its vector arguments $[\vec{\mathbf{G}}_{\text{obs}, i}, a_i]$, which means that reordering the rows of $\mathbf{G}_{\text{obs}}$ and $\mathbf{a}$ in the same way should not change the value of the test statistic. This assumption is satisfied by many frequently used test statistics, including the ones considered in the simulations.

Suppose that a with-replacement random sample of $W = (1 - \alpha)/\alpha$ permutations, $\Pi = (\Pi_1, \dots, \Pi_W)$, is drawn from the set $\mathcal{S}$. In this sample, each random variable Π_w independently takes the value of a permutation π in $\mathcal{S}$ according to the discrete uniform distribution over $\mathcal{S}$. The Monte Carlo permutation p-value based on the set of permutations $\mathcal{S}$ and the test statistic t is defined as [Davison and Hinkley, 1997; Ernst, 2004; Phipson and Smyth, 2010]:

$$\hat{p}_{t, W}^{\mathcal{S}}(\Pi, \mathbf{G}_{\text{obs}}, \mathbf{a}) = (W + 1)^{-1} \left(\sum_{w=1}^W I[t(\mathbf{G}_{\text{obs}}, u_{\Pi_w}(\mathbf{a})) \geq t(\mathbf{G}_{\text{obs}}, \mathbf{a})] + 1 \right)$$

The Monte Carlo permutation test is based on comparing the Monte Carlo permutation p-value to an arbitrary nominal $\alpha \in (0, 1)$ and rejecting the null hypothesis with probability given by the test function:

$$\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\Pi, \mathbf{G}_{\text{obs}}, \mathbf{a}) = \begin{cases} 1 & \hat{p}_{t, W}^{\mathcal{S}}(\Pi, \mathbf{G}_{\text{obs}}, \mathbf{a}) \leq \alpha \\ 0 & \hat{p}_{t, W}^{\mathcal{S}}(\Pi, \mathbf{G}_{\text{obs}}, \mathbf{a}) > \alpha \end{cases} \quad (4)$$

As is clear from the notation, each choice of a set of permutations $\mathcal{S}$ and a test statistic t leads to a different Monte Carlo permutation p-value and test. The optimal choice of test statistic will often be dictated by domain knowledge and the alternative hypothesis of interest. Therefore, the present study focuses on finding conditions on the set of permutations sufficient to guarantee an exact level- α Monte Carlo permutation test under our nonparametric probability model and our minimal assumptions about the test statistic made above.

Simulation Procedure

The main theoretical developments in this study are verified and supplemented by simulation results generated according to the following methodology. For each simulation, 1,000 replicate case-control samples with complete genotypes were obtained through a two stage procedure described previously [Kinnamon et al., 2012]. In summary, 1,000 replicate populations each comprising 10,000 haplotypes at a 100 kb locus were first generated based on a neutral coalescent model with recombination to be used for all simulations. Within a given simulation, a balanced case-control sample of total size 2,000 was drawn from each

population under random mating and a population-specific disease model determined according to the general user-specified parameters for that simulation. The disease risk for an individual with no minor alleles at any risk variant was 5% in all simulations. Each sample included genotypes only at segregating positions with nonzero minor allele frequencies (MAFs) in the haplotype population.

The missing data process was modeled at each segregating position in the haplotype population as follows. The genotype score G_j and missing data indicator M_j for position j in the complete data vector for individual i , $[\vec{G}, \vec{M}]_i$, are denoted by G_{ij} and M_{ij} respectively.

The same notational convention applies to observed genotype scores in $\vec{G}_{\text{obs}, i}$. Assuming that individuals were assayed in C batches over time, the conditional distribution of the missing genotype indicator for a segregating position j in individual i assayed in batch c was modeled as:

$$(M_{ij} | g_{ij}, x_i = c, q_k) \sim \text{Bernoulli}(\gamma_{0,c} + \gamma_{1,c}(g_{ij} - 2p_{jk})) \quad (5)$$

In this model, $p_{jk} > 0$ is the MAF for position j in the k^{th} population q_k ; g_{ij} is the number of minor alleles in individual i 's genotype for position j ; and the $\gamma_{\cdot,c}$ are user-specified parameters for the missing data model in batch c applying to all segregating positions, individuals, and populations. In a case-control sample from a randomly mating population under the null hypothesis, $\gamma_{0,c}$ can be interpreted as the expected rate of missing genotypes in batch c for all data sets. Regardless of the disease model, $\gamma_{1,c}$ is the linear change in the missing genotype probability with each additional minor allele in the true genotype for individuals in batch c . Each missing genotype indicator M_{ij} was obtained as an independent draw from the conditional distribution above, and the observed genotype score $g_{\text{obs},ij}$ was then set to g_{ij} if $M_{ij} = 0$ or a missing code if $M_{ij} = 1$.

We considered $C=4$ batches with 450 cases/50 controls in Batch 0, 350 cases/150 controls in Batch 1, 150 cases/350 controls in Batch 2, and 50 cases/450 controls in Batch 3. This structure simulates a study in which the decision to genotype balanced numbers of cases and controls was made after large-scale genotyping of cases had already begun. We defined seven missing data processes (Models 0-6), which are summarized in Table I. Model 0 refers to the original case-control sample with complete genotype data. For Models 1-6 with missing genotypes, the $\gamma_{0,c}$ were chosen so that, under a null disease model, the expected rate of missing genotypes was $\sum_{c=0}^3 \gamma_{0,c} / 4 = 0.10$ in all data sets. Model 1 represents randomly missing genotypes. Model 2 results in nonrandomly missing genotypes and could arise if the same assay techniques were applied to all batches and the quality control measures preferentially set genotype calls containing minor alleles to missing. Models 3-6 also result in nonrandomly missing genotypes and could arise in studies in which each batch was genotyped with different assay techniques or at different laboratories.

For each of the 1,000 replicate case-control samples with complete data, one data set was obtained under each missing data process (Models 0-6) according to the above procedure. The genotype score $g_{\text{obs},ij}$ for nonmissing genotypes in each data set was the number of minor alleles in the genotype (0, 1, or 2), where the minor allele was determined based on the frequency in the original haplotype population to ensure comparability across data sets from the same case-control sample. All segregating positions in the original haplotype population that were completely missing or monomorphic in each data set were excluded from the analysis of that data set. For all data sets in a given replicate, all Monte Carlo permutation p-values based on a particular $\mathcal{S}$ were calculated using the same with-replacement sample of $W = 10,000$ permutations drawn according to the discrete uniform distribution over that $\mathcal{S}$. These samples of permutations were independent across replicates.

For a given disease model, this simulation design yielded 1,000 independent realizations of each Monte Carlo permutation p-value from its marginal distribution under each missing data process considered. At the same time, the simulation design allowed blocked comparisons of the effect of the missing data process on a particular Monte Carlo permutation p-value because the same underlying case-control sample and random sample of permutations was used for all data sets in each replicate.

Simulation Test Statistics

All simulations considered three different test statistics defined below. Inference based on efficiently combining single-variant Cochran-Armitage trend chi-square statistics over variants in a locus has been previously shown to be powerful in the presence of rare variants [Kinnamon et al., 2012], so we considered two test statistics within this general framework. Let $\mathbf{G}_{\text{obs},j} = [G_{\text{obs},1j}, \dots, G_{\text{obs},nj}]^T$ denote the column of the observed genotype score matrix $\mathbf{G}_{\text{obs}}$ corresponding to position j and:

$$\begin{aligned} O_{ij} &= I[G_{\text{obs},ij} \neq h] \\ N_{\text{obs},j} &= \sum_{i=1}^n O_{ij} \\ \hat{\sigma}_{\mathbf{G}_{\text{obs},j}, \mathbf{a}} &= (N_{\text{obs},j} - 1)^{-1} \left[\sum_{i=1}^n O_{ij} G_{\text{obs},ij} a_i - N_{\text{obs},j}^{-1} \sum_{i=1}^n O_{ij} G_{\text{obs},ij} \sum_{i=1}^n O_{ij} a_i \right] \\ \hat{\sigma}_{\mathbf{G}_{\text{obs},j}}^2 &= (N_{\text{obs},j} - 1)^{-1} \left[\sum_{i=1}^n O_{ij} G_{\text{obs},ij}^2 - N_{\text{obs},j}^{-1} \left(\sum_{i=1}^n O_{ij} G_{\text{obs},ij} \right)^2 \right] \\ \hat{\sigma}_{\mathbf{a},j}^2 &= (N_{\text{obs},j} - 1)^{-1} \left[\sum_{i=1}^n O_{ij} a_i^2 - N_{\text{obs},j}^{-1} \left(\sum_{i=1}^n O_{ij} a_i \right)^2 \right] \end{aligned}$$

Let $\mathcal{J} \subseteq \{1, \dots, v\}$ denote the subset of indices corresponding to genomic positions included in the analysis for a given data set. The CA max statistic is defined as:

$$t_{\text{CA max}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) = \max_{j \in \mathcal{J}} \left\{ N_{\text{obs}, j} \frac{\left[\hat{\sigma}_{\mathbf{G}_{\text{obs}, j}, \mathbf{a}} \right]^2}{\hat{\sigma}_{\mathbf{G}_{\text{obs}, j}}^2 \hat{\sigma}_{\mathbf{a}, j}^2} \right\}$$

Permutation tests based on the CA max statistic are closely related to standard procedures for performing multiplicity-adjusted inference on the most significant single-variant test in the locus. The principal difference is that Monte Carlo permutation tests use permutation-based multiplicity adjustment, rather than the usual Bonferroni adjustment, for the single-variant test. The CA sum statistic is defined as:

$$t_{\text{CA sum}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) = \sum_{j \in \mathcal{J}} N_{\text{obs}, j} \frac{\left[\hat{\sigma}_{\mathbf{G}_{\text{obs}, j}, \mathbf{a}} \right]^2}{\hat{\sigma}_{\mathbf{G}_{\text{obs}, j}}^2 \hat{\sigma}_{\mathbf{a}, j}^2}$$

The CA sum statistic is closely related to the Sequence Kernel Association Test [Wu et al., 2011] and C-alpha [Neale et al., 2011] statistics [Kinnamon et al., 2012]. The third statistic was a frequency-weighted burden (FW burden) statistic in the spirit of the Weighted Sum Statistic [Madsen and Browning, 2009]. For each data set, a Bayesian estimate of the MAF, $\hat{p}_j$, among all individuals with observed genotypes and a frequency weight, $\hat{w}_j$, were calculated for each $j \in \mathcal{J}$ as follows:

$$\hat{p}_j = \left(\sum_{i=1}^n O_{ij} G_{\text{obs}, ij} + 1 \right) \left(2N_{\text{obs}, j} + 2 \right)^{-1}$$

$$\hat{w}_j = \left(2\hat{p}_j(1 - \hat{p}_j) \right)^{-1/2}$$

A genetic score $\Gamma_i = \sum_{j \in \mathcal{J}} \hat{w}_j O_{ij} G_{\text{obs}, ij}$, which sums the frequency-weighted minor allele counts over all observed genotypes in each individual, was then calculated. Now let $n_a = \sum_{i=1}^n a_i = \sum_{i=1}^n a_i^2$. The FW burden statistic is the Wald chi-square statistic for the difference in mean genetic score between cases and controls, which is equivalently expressed as the product of n and the squared Pearson correlation between Γ_j and a_j :

$$t_{\text{FW burden}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) = \frac{\left[n_a^{-1} \sum_{i=1}^n \Gamma_i a_i - (n - n_a)^{-1} \sum_{i=1}^n \Gamma_i (1 - a_i) \right]^2}{\left[n_a^{-1} + (n - n_a)^{-1} \right] n^{-1} \left[\sum_{i=1}^n \Gamma_i^2 - n^{-1} \left(\sum_{i=1}^n \Gamma_i \right)^2 \right]}$$

$$= n \frac{\left[\sum_{i=1}^n \Gamma_i a_i - n^{-1} \sum_{i=1}^n a_i \sum_{i=1}^n \Gamma_i \right]^2}{\left[\sum_{i=1}^n a_i^2 - n^{-1} \left(\sum_{i=1}^n a_i \right)^2 \right] \left[\sum_{i=1}^n \Gamma_i^2 - n^{-1} \left(\sum_{i=1}^n \Gamma_i \right)^2 \right]}$$

Analysis of Simulation Results

Our simulations generated 1,000 independent realizations of each Monte Carlo permutation p-value $\hat{p}_{t,W}^{\delta}(\Pi, \mathbf{G}_{\text{obs}}, \mathbf{a})$ from its marginal distribution under each missing data process for a particular evolutionary process and disease model. For simulations under a null disease model, we produced quantile-quantile (QQ) plots graphing the 1,000 observed order statistics for each Monte Carlo permutation p-value under each missing data process against their expected values under a theoretical uniform distribution, both on a $-\log_{10}$ scale. To facilitate comparisons between the observed order statistics and their theoretical expectations, we also plotted simultaneous 95% concentration intervals designed to contain all of the observed order statistics in 95% of random samples of the same size from a theoretical uniform distribution. The concentration interval for the k^{th} order statistic was defined as the interval between the $\delta_{adj}/2$ and $1 - \delta_{adj}/2$ quantiles of the Beta $(k, 1,001 - k)$ distribution [Burton et al., 2007; Quesenberry and Hales, 1980; Stirling, 1982], where $\delta_{adj} = 0.05/1,000$ was obtained using the Bonferroni correction to ensure a simultaneous coverage probability of 0.95 over all 1,000 intervals. In addition, we estimated the type I error rate and its Clopper-Pearson exact 95% confidence interval [Clopper and Pearson, 1934] for each Monte Carlo permutation test under each missing data process at nominal α levels of 0.05, 0.01, and 0.001.

Results

Sufficient Conditions for an Exact Level- α Monte Carlo Permutation Test

As shown earlier, constructing a Monte Carlo permutation test of the null hypothesis in a genetic case-control study involves choosing a set of permutations of individual indices $\mathcal{S}$ and a real-valued test statistic t on which to base the test function in (4). In the Appendix, we establish that two conditions on the set of permutations $\mathcal{S}$ are together sufficient to guarantee an exact level- α Monte Carlo permutation test under our nonparametric probability model and our minimal assumptions about the test statistic. These conditions are:

Group Condition—The set of permutations $\mathcal{S}$ is a group under composition.

Invariance Condition—For all permutations π in the set $\mathcal{S}$, the induced PMF of the random matrix $u_{\pi}(\mathbf{G}_{\text{obs}})$ under $\mathbb{F}_0$:

$$P_{\mathbb{F}_0}(u_{\pi}(\mathbf{G}_{\text{obs}}) = \mathbf{g}_{\text{obs}} | \mathbf{a}, \mathbf{x}) = \sum_{\mathcal{Q}} \left[\prod_{i=1}^n \sum_{\eta^{-1}(\vec{\mathbf{g}}_{\text{obs},i})} f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | q) f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, \vec{\mathbf{x}}_{\pi(i)}, q) \right] f_{\mathcal{Q}}(q) \quad (6)$$

must be identical to $\mathbb{F}_0$.

The Group Condition serves to guarantee that the set of permutations $\mathcal{S}$ has certain properties useful in establishing results about the associated Monte Carlo permutation test. These properties are [Robinson, 2003]:

1. The composite of every two permutations in $\mathcal{S}$ is also in $\mathcal{S}$.
2. Composition of permutations in $\mathcal{S}$ satisfies the associative law.
3. The identity permutation is in $\mathcal{S}$.
4. The inverse of each permutation in $\mathcal{S}$ is also in $\mathcal{S}$.

The Group Condition is easily verified in practice. The set of all permutations of the integers 1 to n is a group known as the symmetric group [Robinson, 2003], so the set of all permutations of individual indices trivially satisfies the Group Condition. To show that any subset of permutations of individual indices satisfies the Group Condition, it is sufficient to show that the subset is a subgroup of the symmetric group [Robinson, 2003].

The intuitive statement of the Invariance Condition is that, under the null hypothesis, the distribution of the observed genotype score matrix should not change if the assignment of individuals to rows is shuffled according to an arbitrary permutation in the chosen set. A version of this condition applying to more general transformation groups has been given previously by other authors [Hoeffding, 1952; Lehmann and Romano, 2005; Romano, 1989, 1990].

Our nonparametric probability model makes no assumptions about the particular form of the PMFs $f_{\vec{G}}$, $f_{\vec{M}}$, and f_Q in the PMF $\mathbb{F}_0$ other than whether they depend on other fixed or random quantities. Moreover, our minimal assumptions about the test statistic t impose only the reasonable constraint that it be symmetric in its vector arguments. Therefore, the Group and Invariance Conditions can be used to verify the validity of Monte Carlo permutation tests based on a set of permutations $\mathcal{S}$ for very general classes of data generating processes and test statistics.

Because the number of fixed genomic positions v is arbitrary in our development, these conditions also apply for single-variant inference ($v=1$) provided that affection status indicators are permuted over all individuals in the study rather than within the subset of individuals with observed genotypes at the position of interest. The former approach is more likely to be used in practice for the following reasons. First, it reduces computational complexity by allowing the same sample of permutations to be applied to all positions with genotypes. Second, it preserves LD-induced correlations between single-variant test statistics and p-values at different positions, which allows for efficient multiple testing corrections. We now demonstrate how the Group and Invariance Conditions can be applied to determine whether specific classes of Monte Carlo permutation tests are guaranteed to be exact level α in the presence of missing genotypes.

Standard Permutation Tests

The standard permutation set, denoted by $\mathcal{S}_n$ is the set of all $n!$ possible permutations of individual indices. Standard permutation tests, which are the most frequently used Monte

Carlo permutation tests for genetic case-control studies, are based on the test function in (4) with $\mathcal{S} = \mathcal{S}_n$. As noted earlier, the standard permutation set $\mathcal{S}_n$ trivially satisfies the Group Condition. The remainder of this section explores whether the Invariance Condition is also satisfied for the standard permutation set.

If $\vec{\mathbf{x}}_{\pi(i)} = \vec{\mathbf{x}}_i$ for all i and permutations π in $\mathcal{S}_n$, the induced PMF of $u_\pi(\mathbf{G}_{\text{obs}})$ under $\mathbb{F}_0$ given in (6) is identical to $\mathbb{F}_0$ for all permutations π in $\mathcal{S}_n$, and the standard permutation set satisfies the Invariance Condition. Conversely, if $\vec{\mathbf{x}}_{\pi(i)} \neq \vec{\mathbf{x}}_i$ for some i and permutation π in $\mathcal{S}_n$, the induced PMF in (6) differs from $\mathbb{F}_0$ for that permutation, and the standard permutation set cannot satisfy the Invariance Condition. Thus, the standard permutation set satisfies the Invariance Condition if and only if $\vec{\mathbf{x}}_{\pi(i)} = \vec{\mathbf{x}}_i$ for all i and permutations π in $\mathcal{S}_n$. Because there is no guarantee that this condition holds, standard permutation tests are not guaranteed to be exact level α under our nonparametric probability model and our minimal assumptions about the test statistic.

Suppose that we make the additional assumption that the data generating process has a missing data process independent of an individual's covariates. Under this assumption:

$$f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, \vec{\mathbf{x}}_{\pi(i)}, q) = f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, \vec{\mathbf{x}}_i, q) = f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, q)$$

The induced PMF of the random matrix $u_\pi(\mathbf{G}_{\text{obs}})$ under $\mathbb{F}_0$ given by (6) is then clearly identical to $\mathbb{F}_0$ for all permutations π in $\mathcal{S}_n$. Thus, the Invariance Condition is satisfied and standard permutation tests are exact level α under this additional assumption. Missing data processes satisfying this assumption include complete data with $f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, q) = I[\vec{\mathbf{m}} = \mathbf{0}_v]$ and randomly missing genotypes with $f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, q) = f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | q)$. This assumption is also satisfied for missing data processes that depend on the values of both observed and unobserved genotypes but are conditionally independent of covariates. Interestingly, such missing data processes are nonignorable for likelihood inference [Little and Rubin, 2002].

Assuming that the missing data process is independent of covariates is quite restrictive. In practice, the rate and pattern of missing genotypes are likely to depend on variables such as DNA source, age of sample, sequence capture method, sequencing platform, reagent batch, genotype calling algorithm, and quality control parameters, among others. While standard permutation tests are not necessarily invalid when the missing data process depends on covariates, they come without a theoretically guaranteed level. This lack of guaranteed level can be problematic; we subsequently demonstrate with a simulation that some standard permutation tests may have highly inflated type I error rates for some data generating processes. Therefore, we illustrate in the following section how the Group and Invariance Conditions can also be applied to demonstrate that appropriately stratified Monte Carlo permutation tests are exact level α under our nonparametric probability model and our minimal assumptions about the test statistic.

Restricted Permutation Tests

Suppose that the covariate vectors $\vec{x}_i$ assume $C \ll n$ distinct values indexed by c in the data set. Let $\mathcal{J}_c$ denote the set of indices i for the $n_c - 1$ individuals having value c of $\vec{x}_i$ and $\tilde{\mathcal{S}}_n \subseteq \mathcal{S}_n$ be the set of $n_1! \dots n_C!$ permutations that exchange individual indices within but not between strata defined by $\mathcal{J}_c$:

$$\tilde{\mathcal{S}}_n = \{\pi \in \mathcal{S}_n : \forall i \forall c, \pi(i) \in \mathcal{J}_c \text{ if and only if } i \in \mathcal{J}_c\}$$

We refer to $\tilde{\mathcal{S}}_n$ as the restricted permutation set. Restricted permutation tests are Monte Carlo permutation tests based on the test function in (4) with $\mathcal{S} = \tilde{\mathcal{S}}_n$. Such stratified permutation tests have been suggested previously to control for the effect of stratified randomization [Lehmann and Romano, 2005] or a secondary experimental factor [Edgington, 1983] in randomized experiments. In genetics, stratified permutation tests have been proposed to adjust for the presence of a major QTL [Doerge and Churchill, 1996] or nonrandomly missing genotypes arising from extreme-phenotype selective genotyping [Manichaikul et al., 2007] in QTL association studies. The PRESTO software package has also implemented stratified permutation tests to adjust for population stratification in genetic case-control studies [Browning, 2008].

We now proceed to verify the Group Condition for the restricted permutation set. The identity permutation, $\pi_{\text{id}}(i) = i$, is an element of $\mathcal{S}_n$ that trivially satisfies the condition $\pi_{\text{id}}(i) \in \mathcal{J}_c$ if and only if $i \in \mathcal{J}_c$ for all i and c , so $\pi_{\text{id}} \in \tilde{\mathcal{S}}_n$. For arbitrary $\pi_1, \pi_2 \in \tilde{\mathcal{S}}_n$, $\pi_1, \pi_2 \in \mathcal{S}_n$ because $\tilde{\mathcal{S}}_n \subseteq \mathcal{S}_n$, so $\pi_2 \circ \pi_1 \in \mathcal{S}_n$. Also, $(\pi_2 \circ \pi_1)(i) = \pi_2(\pi_1(i)) \in \mathcal{J}_c$ if and only if $i \in \mathcal{J}_c$ for all i and c because $\pi_2(\pi_1(i)) \in \mathcal{J}_c$ if and only if $\pi_1(i) \in \mathcal{J}_c$ and $\pi_1(i) \in \mathcal{J}_c$ if and only if $i \in \mathcal{J}_c$ for all i and c . Thus, for all $\pi_1, \pi_2 \in \tilde{\mathcal{S}}_n$, $\pi_2 \circ \pi_1 \in \tilde{\mathcal{S}}_n$. Finally, for an arbitrary $\pi \in \tilde{\mathcal{S}}_n$, $\pi^{-1} \in \mathcal{S}_n$ because $\tilde{\mathcal{S}}_n \subseteq \mathcal{S}_n$ and $\mathcal{S}_n$ is a group under composition [Robinson, 2003]. Because $\pi(i) \in \mathcal{J}_c$ if and only if $i \in \mathcal{J}_c$ for all i and c , it also follows that $\pi(i) \in \mathcal{J}_c$ if and only if $\pi^{-1}(\pi(i)) \in \mathcal{J}_c$ for all i and c . Consequently, $\pi^{-1}(i) \in \mathcal{J}_c$ if and only if $i \in \mathcal{J}_c$ for all i and c , which implies that $\pi^{-1} \in \tilde{\mathcal{S}}_n$ as well. Thus, for all $\pi \in \tilde{\mathcal{S}}_n$, $\pi^{-1} \in \tilde{\mathcal{S}}_n$. Based on the preceding, $\tilde{\mathcal{S}}_n$ is a subgroup of $\mathcal{S}_n$, which also means that it is a group under composition [Robinson, 2003].

We now verify the Invariance Condition for the restricted permutation set. For an arbitrary i , $i \in \mathcal{J}_c$ for some c . For an arbitrary $\pi \in \tilde{\mathcal{S}}_n$, $\pi(i) \in \mathcal{J}_c$ for this i , which implies that $\vec{x}_{\pi(i)} = \vec{x}_i$. Therefore, the definition of $\tilde{\mathcal{S}}_n$ ensures that $\vec{x}_{\pi(i)} = \vec{x}_i$ for all i and $\pi \in \tilde{\mathcal{S}}_n$, which implies that the induced PMF of $u_\pi(\mathbf{G}_{\text{obs}})$ under $\mathbb{F}_0$ given by (6) is identical to $\mathbb{F}_0$ for all permutations π in $\tilde{\mathcal{S}}_n$. Therefore, the Invariance Condition is satisfied by the restricted

permutation set, and restricted permutation tests are exact level α under our nonparametric probability model and our minimal assumptions about the test statistic.

Note that continuous covariates are naturally accommodated by the restricted permutation test because permutations are performed within strata of individuals with the same $\vec{x}_i$ value regardless of whether the elements of this vector are discrete or continuous. However, as $\vec{x}_i$ assumes a larger number of distinct values in the data set, the number of strata C increases and the number of permissible permutations in the restricted permutation set $\tilde{\mathcal{S}}_n$ shrinks. At the extreme, C equals n because every individual has a different $\vec{x}_i$. In such cases, each $\mathcal{J}_c$ would contain a single individual i , and $\tilde{\mathcal{S}}_n$ would be the trivial set comprising only the identity permutation. Although the Group and Invariance Conditions would be satisfied by the restricted permutation set, the resulting Monte Carlo permutation p-value would be identically 1 for every test statistic t and matrix $\mathbf{g}_{\text{obs}}$ in the sample space $\mathcal{X}^n$. Thus, restricted permutation tests would be exact level α with zero power. In fact, it is easy to see that, when C equals n , the only set of permutations $\mathcal{S}$ for which $\vec{x}_{\pi(i)} = \vec{x}_i$ for all i and π in $\mathcal{S}$ is the trivial set comprising the identity permutation. Thus, it may not be possible to construct any permutation test that is exact level α with nonzero power when the missing data process depends on highly heterogeneous continuous covariates.

In many situations, groups of individuals will be sequenced in batches and the primary variables suspected to influence the missing data process (e.g., DNA source, sequencing platform, capture platform) will be naturally discrete. The result will be covariate vectors assuming a small number of distinct values, and restricted permutation tests will be reasonable procedures. An important question in such circumstances is whether using a restricted permutation test would sacrifice power relative to a standard permutation test based on the same test statistic if the missing data process were actually independent of covariates. Because general finite-sample theoretical results are difficult to obtain, we present results from a simulation in the next section to gain some additional insight.

Simulations

In the first simulation, we examined the effects of the missing data process on the type I error rate of different Monte Carlo permutation tests. Figures 1, 2, and 3 contain QQ plots that compare the 1,000 observed order statistics for each Monte Carlo permutation p-value under each missing data process to their expected values under a theoretical uniform distribution. Each figure considers Monte Carlo permutation p-values based on one of the three test statistics described earlier and is faceted by the permutation set (standard or restricted) and missing data process (Models 0-6). A gray line representing a concentration interval is associated with each observed order statistic. These concentration intervals are designed to collectively contain all of the observed order statistics in 95% of random samples of size 1,000 from a uniform distribution. If the observed order statistic lies above (below) its concentration interval on a $-\log_{10}$ scale, it is substantially smaller (larger) than would be expected if the Monte Carlo permutation p-value had a uniform marginal distribution. If all observed order statistics in a facet are within or below their concentration

intervals, the data are consistent with the marginal distribution of the corresponding Monte Carlo permutation p-value being stochastically larger than uniform for that missing data process. In that case, the data are consistent with the corresponding Monte Carlo permutation test being exact level α for that missing data process. Conversely, if a large number of observed order statistics in a facet lie above their concentration intervals, the data suggest that the marginal distribution of the corresponding Monte Carlo permutation p-value is not stochastically larger than uniform for that missing data process. In that case, the data suggest that the corresponding Monte Carlo permutation test is not exact level α for that missing data process. To complement these plots, Tables II, III, and IV provide the estimated type I error rate and its Clopper-Pearson exact 95% confidence interval for each Monte Carlo permutation test under each missing data process at nominal α levels of 0.05, 0.01, and 0.001.

We showed that standard permutation tests are guaranteed to be exact level α when the missing data process is independent of covariates. Only the missing data processes in Models 0-2 satisfy this assumption, and both Figures 1-3 and Tables II, III, and IV are consistent with the standard permutation tests considered being exact level α for these missing data processes. However, the standard permutation set does not satisfy the Invariance Condition for missing data processes violating this assumption. Figures 1-3 and Tables II-IV provide compelling evidence that the standard permutation tests considered had highly inflated type I error rates for some nominal α when the missing data process varied across batches (Models 3-6). The type I error inflation ranged from nonexistent to substantial depending on the particular combination of test statistic, missing data process, and nominal α . For example, at a nominal $\alpha=0.01$, the range of estimated type I errors for missing data processes in Models 3-6 was [0.009, 0.022] for the CA max statistic, [0.009, 0.020] for the CA sum statistic, and a remarkably high [0.417, 0.955] for the FW burden statistic. If the actual batch was identified as a potential factor influencing the missing data process and recorded on each subject, it could be used in defining the restricted permutation set $\tilde{\mathcal{S}}_n$. This set would satisfy the Invariance Condition for all missing data processes in our simulations, and all resulting restricted permutation tests would be exact level α . The restricted permutation set was defined in this manner in the simulations, and the results in Figures 1-3 as well as Tables II-IV are consistent with this theoretical prediction.

In the second simulation, we examined the difference in power between standard and restricted permutation tests based on the same test statistic under various missing data processes for which both tests are exact level α (Models 0-2). We considered a logistic population disease model with 50 rare risk variants (representing ~5% of all segregating positions in the locus in the average haplotype population) with population MAF<0.01 and odds ratio (OR)=2. In each haplotype population, the risk variants were randomly chosen from among the variants meeting the MAF criterion.

Table V shows the estimated power at nominal α levels of 0.05, 0.01, and 0.001 for various Monte Carlo permutation tests under this disease model and missing data processes in Models 0-2. Restricted permutation tests did have lower power for all test statistics and nominal α , but the reduction was generally modest. This finding stands to reason because

the standard and restricted permutation p-values for a particular data set were correlated with one another (Figures 4, 5, and 6). Importantly, power varied much more substantially with the choice of test statistic for the same permutation set than it did with the choice of permutation set for the same test statistic under a given missing data process. This observation suggests that validity under the suspected data generating process, rather than power, should be the primary determinant of the choice of permutation set.

Discussion

While Monte Carlo permutation tests are mainstays in the analysis of genetic case-control studies with unrelated subjects, the genetics literature is remarkably silent on how they should be applied in the presence of missing genotypes. At the same time, technological advances such as whole-exome and whole-genome sequencing are likely to increase the rate of missing genotypes in genetic case-control studies rather than reduce it. While imputation has been a successful solution for missing array-based genotypes at common variants, it may not enjoy the same level of success for rare variants and de novo mutations not present in the available training data (see, e.g., [Li et al., 2011]). Analysis based on estimated genotype dosages arising from these procedures is therefore subject to its own set of potential biases, particularly when assay-related covariates lead to varying imputation qualities across groups of individuals in the data set. We therefore sought to provide geneticists with the tools necessary to identify exact level- α permutation tests that use all available genotype data from a genetic case-control study with unrelated subjects and missing genotypes.

We began by showing how to construct a Monte Carlo permutation test by choosing a set of permutations of individual indices and a test statistic measuring the association between genotypes and affection status. Provided that two conditions are satisfied by the chosen set of permutations, we demonstrated theoretically that the resulting Monte Carlo permutation test is guaranteed to be exact level α under our nonparametric probability model and our minimal assumptions about the test statistic. One of these conditions, the Invariance Condition, is not satisfied by the set of all permutations of individual indices under our nonparametric probability model, and our simulation demonstrated that standard permutation tests based on this set may have extremely inflated type I error for some missing data processes. However, if the missing data process is influenced by known covariates that are recorded, subjects can be divided into strata having the same covariate vector values, and the set of permutations that shuffles affection status indicators only within strata can be used. We showed that the resulting tests, which are termed restricted permutation tests, are exact level α under our nonparametric probability model. We confirmed this theoretical prediction in a simulation, which demonstrated that restricted permutation tests can recover control of type I error lost by standard permutation tests based on the same test statistic.

While our results demonstrate that modification of the permutation procedure can prevent inflated type I error, they also indicate that design of data collection and assaying procedures at the beginning of the study can still play an important role in determining power. Standard permutation tests are not guaranteed to be exact level α for designs in which the values of covariates influencing the missing data process are perfectly separated by affection status. One such design is the shared controls design, in which all control genotypes are obtained

from an existing data set or a publicly available database. Only sets comprising permutations in which indices are not exchanged between cases and controls will satisfy the Invariance Condition for this design. Each observed genotype score vector will therefore have the same affection status indicator value for every permutation in these sets, and the test statistic for each permutation will be the same as the test statistic in the original data set. It follows that the p-value will be 1 over the entire sample space, and the resulting Monte Carlo permutation tests will be exact level α with zero power. Therefore, permutation inference may not be feasible in shared controls designs, and this fact should be carefully considered before undertaking such a design.

In contrast, our results demonstrate that designs in which both controls and cases are assayed concurrently in large batches allow for valid permutation inference even if the missing data process changes substantially between batches. We showed that restricted permutation tests are exact level α with missing genotypes for these types of designs, and our simulation results suggested that their powers are not necessarily substantially lower than those of standard permutation tests based on the same test statistics even when the latter are valid. However, it is worth emphasizing that the power of restricted permutation tests might be substantially lower if individuals were assayed in small batches or the missing data process depended upon a continuous covariate with substantial heterogeneity across individuals.

In fact, it may not always be possible to find a set of permutations yielding exact level α tests with reasonable power, especially if the missing data process depends on continuous covariates. Restricted permutation tests, which use the set of permutations within strata of individuals having the same covariate vector values, are exact level α with continuous covariates under our nonparametric probability model. However, covariate vectors containing continuous variables, or even more than a few discrete variables, may induce a large number of sparse strata. If discrete categories of a continuous covariate are likely to capture its effect on the missing data process, our results suggest that a reasonable method for reducing the number of strata is to discretize the covariate and use the resulting category in place of the continuous covariate. Nonetheless, it is not possible to find any set of permutations yielding an exact level α test with nonzero power when the number of strata equals the number of individuals in the data set. This observation severely limits the number of covariates that can reasonably be accommodated within the general framework of exact permutation inference.

If one is willing to accept asymptotic, rather than exact, control of the type I error rate, additional options are available. Other studies have shown that the choice of test statistic can determine the asymptotic validity of permutation tests when the underlying distribution is not invariant under transformations [Chung and Romano, 2011; Romano, 1990]. Such results offer clues as to why the type I error inflation of standard permutation tests varied substantially among the test statistics considered in our simulation for the same missing data process. Because the choice to use permutation tests is often made when asymptotic approximations are questionable, we would advocate first investing effort to find an appropriate set of permutations satisfying the Group and Invariance Conditions. If the only sets satisfying these conditions are trivial or very small, however, a potential alternative approach could involve identifying a test statistic appropriate for the alternative hypothesis

of interest and verifying that it yields an asymptotically level- α standard permutation test in the presence of missing genotypes. Exploring such asymptotic approaches could be a productive line of future inquiry.

Another potential alternative could involve modifying parametric bootstrap procedures for obtaining covariate-adjusted p-values in the presence of continuous or discrete covariates [Epstein et al., 2012; Kimmel et al., 2007]. The developments in these studies assume a prospective data generating process in which genotypes and covariates are treated as fixed and affection status indicators are treated as random. Unfortunately, it is not immediately apparent whether such procedures would yield valid inferences for the retrospective data generating process with missing genotypes considered in this paper. Therefore, further work is required to determine the validity and applicability of these procedures under such conditions.

A disadvantage common to all of these techniques is having to correctly identify the covariates likely to influence the missing data process. Omitting important covariates can inflate the type I error rate, whereas including unimportant ones can reduce power. As with any statistical modeling problem, knowledge of the domain and study design should be used to identify a promising set of candidate covariates a priori. This approach may be supplemented by empirical comparisons of the distributions of missing data indicator vectors across groups defined by candidate covariates within controls and cases. However, such comparisons will naturally be crude in a missing data context, particularly given that it is not possible to determine empirically how the missing data process depends on an individual's true genotypes. A second disadvantage of resampling-based tests is computational intensity. However, the additional computational cost may be justified when the resulting test controls the type I error rate exactly, rather than asymptotically. It may also be justified when analytic parametric or semiparametric approaches rely on restrictive or unverifiable assumptions. Regardless, the cost-benefit tradeoff for resampling-based tests will likely continue to improve as advances in parallel computing and the use of graphics processing units (see, e.g., [Shterev et al., 2010]) further reduce the computational burden.

We close by noting that the approach in this paper can be applied more broadly to derive permutation tests that address problems such as population stratification in genetic case-control studies with missing genotypes. One would begin by defining a testing problem based on a set of assumptions regarding the genotype sampling process and the missing data process in our nonparametric probability model. One could then derive the null PMF of the observed genotype score matrix and its induced PMF under permutation, propose a set of permutations, and verify that the set satisfies both the Group and Invariance Conditions. Given that the set of permutations satisfies these conditions, one could finally select a test statistic symmetric in its vector arguments. Although the particular choice of test statistic is beyond the scope of this paper, theoretical results as well as practical heuristics for selecting test statistics with optimal finite-sample or asymptotic power for the alternative of interest can be found in several sources [Hoeffding, 1952; Lehmann and Romano, 2005; Lehmann and Stein, 1949; Pesarin and Salmaso, 2010]. Any Monte Carlo permutation test derived according to this procedure is guaranteed to be exact level α for all data generating processes satisfying the assumptions made in defining the testing problem.

Appendix

Suppose that the nonparametric probability model in the main text holds and that a Monte Carlo permutation test has been constructed by choosing a set of permutations of individual indices $\mathcal{S}$ and a test statistic t that is symmetric in its vector arguments. In this appendix, we establish that, if the set $\mathcal{S}$ satisfies the Group and Invariance Conditions defined in the main text, this test is exact level α .

We proceed by proving three lemmas that build to a theorem stating the desired result. In the lemmas and theorem that follow, let R be the number of permutations in the set $\mathcal{S}$. For an arbitrary $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$, define:

$$p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) = R^{-1} \sum_{\mathcal{S}} I[t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq t(\mathbf{g}_{\text{obs}}, \mathbf{a})]$$

For an arbitrary nominal $\alpha \in (0, 1)$, also define:

$$\phi_{t, \alpha}^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) = \begin{cases} 1 & p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \leq \alpha \\ 0 & p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) > \alpha \end{cases}$$

We begin with a useful consequent of the Group Condition.

Lemma 1. If the set of permutations $\mathcal{S}$ is a group under composition (Group Condition), then $\sum_{\mathcal{S}} E_{\mathbb{F}_0} [\phi_{t, \alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{G}_{\text{obs}}), \mathbf{a}) | \mathbf{a}, \mathbf{x}] \leq R\alpha$.

Proof. Consider an arbitrary $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$ and $\pi' \in \mathcal{S}$. From the definition,

$u_{\pi}(\mathbf{g}_{\text{obs}}) = [\vec{\mathbf{g}}_{\text{obs}, \pi(1)}^T, \dots, \vec{\mathbf{g}}_{\text{obs}, \pi(n)}^T]^T$. Moreover, row i of $u_{\pi}(u_{\pi'}(\mathbf{g}_{\text{obs}}))$ contains the value of row $\pi(i)$ of $u_{\pi'}(\mathbf{g}_{\text{obs}})$, and row $\pi(i)$ of $u_{\pi'}(\mathbf{g}_{\text{obs}})$, the value of row $\pi'(\pi(i)) = (\pi' \circ \pi)(i)$ of $\mathbf{g}_{\text{obs}}$. It follows that:

$$\begin{aligned} u_{\pi}(u_{\pi'}(\mathbf{g}_{\text{obs}})) &= [\vec{\mathbf{g}}_{\text{obs}, (\pi' \circ \pi)(1)}^T, \dots, \vec{\mathbf{g}}_{\text{obs}, (\pi' \circ \pi)(n)}^T]^T \\ &= u_{\pi' \circ \pi}(\mathbf{g}_{\text{obs}}) \end{aligned}$$

From the definition of $p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a})$ above, we have that:

$$\begin{aligned}
 p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) &= R^{-1} \sum_{\mathcal{S}} I \left[t(u_{\pi'}(u_{\pi'}(\mathbf{g}_{\text{obs}})), \mathbf{a}) \geq t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \right] \\
 &= R^{-1} \sum_{\mathcal{S}} I \left[t(u_{\pi' \circ \pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \right]
 \end{aligned}$$

Because $\mathcal{S}$ is a group under composition, $\circ$ must be a binary operation on $\mathcal{S}$, which means that $\pi' \circ \pi \in \mathcal{S}$ for all $\pi', \pi \in \mathcal{S}$ [Robinson, 2003]. For an arbitrary $\pi' \in \mathcal{S}$, this binary operation defines a mapping $\lambda_{\pi'}: \mathcal{S} \rightarrow \mathcal{S}$, where $\lambda_{\pi'}(\pi) = \pi' \circ \pi$. Because $\mathcal{S}$ is a group under composition, $(\pi')^{-1} \in \mathcal{S}$ as well [Robinson, 2003], and $\lambda_{(\pi')^{-1}}: \mathcal{S} \rightarrow \mathcal{S}$. Composition of permutations in $\mathcal{S}$ must also satisfy the associative law [Robinson, 2003], so:

$$\begin{aligned}
 \lambda_{\pi'} \left(\lambda_{(\pi')^{-1}}(\pi) \right) &= \pi' \circ \left((\pi')^{-1} \circ \pi \right) = \left(\pi' \circ (\pi')^{-1} \right) \circ \pi = \pi \\
 \lambda_{(\pi')^{-1}} \left(\lambda_{\pi'}(\pi) \right) &= (\pi')^{-1} \circ (\pi' \circ \pi) = \left((\pi')^{-1} \circ \pi' \right) \circ \pi = \pi
 \end{aligned}$$

Therefore, $\lambda_{(\pi')^{-1}}$ is the unique inverse for $\lambda_{\pi'}$, which implies that $\lambda_{\pi'}$ is a one-to-one mapping of $\mathcal{S}$ onto itself [Robinson, 2003]. Because $\lambda_{\pi'}$ is one-to-one, different $\pi \in \mathcal{S}$ map to different $\pi' \circ \pi \in \mathcal{S}$, and the above sum contains R terms each corresponding to a distinct $\pi' \circ \pi \in \mathcal{S}$. Moreover, because any collection of R distinct elements of $\mathcal{S}$ must constitute the entire set, the above sum can equivalently be taken over all $\pi \in \mathcal{S}$. Therefore, $p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a})$ can be equivalently written as:

$$p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = R^{-1} \sum_{\mathcal{S}} I \left[t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \right]$$

Following previous developments [Hoeffding, 1952; Lehmann and Romano, 2005; Romano, 1989, 1990], consider the R ordered values of $t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a})$ over all $\pi \in \mathcal{S}$:

$$t^{(1)}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \leq t^{(2)}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \leq \dots \leq t^{(R)}(\mathbf{g}_{\text{obs}}, \mathbf{a})$$

If $\alpha \in (0, R^{-1})$, it is clear that $p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq R^{-1} > \alpha$ and $\phi_{t, \alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = 0$ for all $\pi' \in \mathcal{S}$. This result implies that:

$$\sum_{\pi' \in \mathcal{S}} \phi_{t, \alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = 0 < R\alpha$$

If $\alpha \in [R^{-1}, 1)$, $R\alpha$ can be separated into integral and decimal components according to $R\alpha = \lfloor R\alpha \rfloor + \varepsilon$, where $\lfloor \cdot \rfloor$ is the floor function and $\varepsilon \in [0, 1)$ is a real number representing the decimal component. Define $r^* = R - \lfloor R\alpha \rfloor + 1$ and partition $\mathcal{S}$ into three subsets:

$$\begin{aligned}\mathcal{S}_t^+(\mathbf{g}_{\text{obs}}, \mathbf{a}) &= \left\{ \pi \in \mathcal{S} : t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) > t^{(r^*)}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right\} \\ \mathcal{S}_t^0(\mathbf{g}_{\text{obs}}, \mathbf{a}) &= \left\{ \pi \in \mathcal{S} : t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = t^{(r^*)}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right\} \\ \mathcal{S}_t^-(\mathbf{g}_{\text{obs}}, \mathbf{a}) &= \left\{ \pi \in \mathcal{S} : t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) < t^{(r^*)}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right\}\end{aligned}$$

For all $\pi' \in \mathcal{S}_t^+(\mathbf{g}_{\text{obs}}, \mathbf{a})$, there are at most $\lfloor R\alpha \rfloor - 1$ $\pi \in \mathcal{S}$ for which $t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a})$, so $p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \leq R^{-1}(\lfloor R\alpha \rfloor - 1) < \alpha$ and $\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = 1$. For all $\pi' \in \mathcal{S}_t^-(\mathbf{g}_{\text{obs}}, \mathbf{a})$, there are at least $\lfloor R\alpha \rfloor + 1$ $\pi \in \mathcal{S}$ for which $t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a})$, so:

$p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq R^{-1}(\lfloor R\alpha \rfloor + 1) = R^{-1}(R\alpha - \varepsilon + 1) = \alpha + R^{-1}(1 - \varepsilon) > \alpha$ and $\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = 0$. Finally, for all $\pi' \in \mathcal{S}_t^0(\mathbf{g}_{\text{obs}}, \mathbf{a})$, there are two possibilities. If $t^{(r^*)}(\mathbf{g}_{\text{obs}}, \mathbf{a}) = t^{(r^*-1)}(\mathbf{g}_{\text{obs}}, \mathbf{a})$, then there are at least $\lfloor R\alpha \rfloor + 1$ $\pi \in \mathcal{S}$ for which $t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq t(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a})$. Therefore, $p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) > \alpha$ and $\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = 0$ for all $\pi' \in \mathcal{S}_t^0(\mathbf{g}_{\text{obs}}, \mathbf{a})$, which implies that:

$$\sum_{\pi' \in \mathcal{S}} \phi_{t,\alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = \sum_{\pi' \in \mathcal{S}_t^+(\mathbf{g}_{\text{obs}}, \mathbf{a})} 1 \leq \lfloor R\alpha \rfloor - 1 < R\alpha$$

because there are at most $\lfloor R\alpha \rfloor - 1$ $\pi' \in \mathcal{S}_t^+(\mathbf{g}_{\text{obs}}, \mathbf{a})$ by definition. Alternatively, if $t^{(r^*)}(\mathbf{g}_{\text{obs}}, \mathbf{a}) > t^{(r^*-1)}(\mathbf{g}_{\text{obs}}, \mathbf{a})$, there must be exactly $\lfloor R\alpha \rfloor$ $\pi \in \mathcal{S}$ with $t(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \geq t^{(r^*)}(\mathbf{g}_{\text{obs}}, \mathbf{a})$. Therefore, $p_t^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = R^{-1}\lfloor R\alpha \rfloor \leq \alpha$ and $\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = 1$ for all $\pi' \in \mathcal{S}_t^0(\mathbf{g}_{\text{obs}}, \mathbf{a})$. There are also exactly $\lfloor R\alpha \rfloor$ $\pi' \in \mathcal{S}_t^0(\mathbf{g}_{\text{obs}}, \mathbf{a}) \cup \mathcal{S}_t^+(\mathbf{g}_{\text{obs}}, \mathbf{a})$, which implies that:

$$\sum_{\pi' \in \mathcal{S}} \phi_{t,\alpha}^{\mathcal{S}}(u_{\pi'}(\mathbf{g}_{\text{obs}}), \mathbf{a}) = \sum_{\pi' \in \mathcal{S}_t^0(\mathbf{g}_{\text{obs}}, \mathbf{a})} 1 + \sum_{\pi' \in \mathcal{S}_t^+(\mathbf{g}_{\text{obs}}, \mathbf{a})} 1 = \lfloor R\alpha \rfloor \leq R\alpha$$

Therefore, $\sum_{\mathcal{S}} \phi_{t,\alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{g}_{\text{obs}}), \mathbf{a}) \leq R\alpha$ for an arbitrary $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$. Based on the preceding property, it follows that:

$$\mathbb{E}_{\mathbb{F}_0} \left[\sum_{\mathcal{S}} \phi_{t,\alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{G}_{\text{obs}}), \mathbf{a}) \middle| \mathbf{a}, \mathbf{x} \right] = \sum_{\mathcal{S}} \mathbb{E}_{\mathbb{F}_0} \left[\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{G}_{\text{obs}}), \mathbf{a}) \middle| \mathbf{a}, \mathbf{x} \right] \leq R\alpha$$

□

The consequent of Lemma 1 shifts interest to the quantity $\mathbb{E}_{\mathbb{F}_0} [\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{G}_{\text{obs}}), \mathbf{a}) \middle| \mathbf{a}, \mathbf{x}]$. For an arbitrary $\pi \in \mathcal{S}$, the random matrix $u_{\pi}(\mathbf{G}_{\text{obs}})$ takes on values in the space $\{u_{\pi}(\mathbf{g}_{\text{obs}}) : \mathbf{g}_{\text{obs}} \in \mathcal{X}^n\}$. For an arbitrary $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$, it is clear that $u_{\pi}(\mathbf{g}_{\text{obs}}) \in \mathcal{X}^n$ because $\vec{\mathbf{g}}_{\text{obs}, \pi(i)} \in \mathcal{X}$ for all i , so $u_{\pi} : \mathcal{X}^n \rightarrow \mathcal{X}^n$. Because every π has a unique inverse permutation π^{-1} , it follows from the development above that u_{π} and $u_{\pi^{-1}}$ satisfy:

$$u_{\pi^{-1}}(u_{\pi}(\mathbf{g}_{\text{obs}})) = u_{\pi}^{-1}(u_{\pi^{-1}}(\mathbf{g}_{\text{obs}})) = \mathbf{g}_{\text{obs}}$$

Thus, $u_{\pi^{-1}}$ is the unique inverse for u_{π} , which implies that u_{π} is a one-to-one mapping of $\mathcal{X}^n$ onto itself [Robinson, 2003]. Therefore, $\{u_{\pi}(\mathbf{g}_{\text{obs}}) : \mathbf{g}_{\text{obs}} \in \mathcal{X}^n\} = \mathcal{X}^n$, and it follows that:

$$\mathbb{E}_{\mathbb{F}_0} [\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{G}_{\text{obs}}), \mathbf{a}) \middle| \mathbf{a}, \mathbf{x}] = \sum_{\mathcal{X}^n} \phi_{t,\alpha}^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \mathbb{P}_{\mathbb{F}_0} \left(u_{\pi}(\mathbf{G}_{\text{obs}}) = \mathbf{g}_{\text{obs}} \middle| \mathbf{a}, \mathbf{x} \right)$$

This observation leads to a second lemma.

Lemma 2. If the set of permutations $\mathcal{S}$ is a group under composition (Group Condition) and, for all permutations π in the set $\mathcal{S}$, the induced PMF of the random matrix $u_{\pi}(\mathbf{G}_{\text{obs}})$ under $\mathbb{F}_0$ is identical to $\mathbb{F}_0$ (Invariance Condition), then $\mathbb{P}_{\mathbb{F}_0} (p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \leq \alpha \middle| \mathbf{a}, \mathbf{x}) \leq \alpha$.

Proof. If the induced PMF of $u_{\pi}(\mathbf{G}_{\text{obs}})$ under $\mathbb{F}_0$ is identical to $\mathbb{F}_0$ for all $\pi \in \mathcal{S}$, it follows that:

$$\begin{aligned} \mathbb{E}_{\mathbb{F}_0} [\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{G}_{\text{obs}}), \mathbf{a}) \middle| \mathbf{a}, \mathbf{x}] &= \sum_{\mathcal{X}^n} \phi_{t,\alpha}^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \mathbb{P}_{\mathbb{F}_0} (u_{\pi}(\mathbf{G}_{\text{obs}}) = \mathbf{g}_{\text{obs}} \middle| \mathbf{a}, \mathbf{x}) \\ &= \sum_{\mathcal{X}^n} \phi_{t,\alpha}^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \mathbb{F}_0(\mathbf{g}_{\text{obs}} \middle| \mathbf{a}, \mathbf{x}) \\ &= \mathbb{E}_{\mathbb{F}_0} [\phi_{t,\alpha}^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \middle| \mathbf{a}, \mathbf{x}] \end{aligned}$$

for all $\pi \in \mathcal{S}$. If the set of permutations $\mathcal{S}$ is also a group under composition, Lemma 1 implies that $\sum_{\mathcal{S}} E_{\mathbb{F}_0}[\phi_{t,\alpha}^{\mathcal{S}}(u_{\pi}(\mathbf{G}_{\text{obs}}), \mathbf{a}) | \mathbf{a}, \mathbf{x}] \leq R\alpha$. It immediately follows that

$$E_{\mathbb{F}_0}[\phi_{t,\alpha}^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) | \mathbf{a}, \mathbf{x}] \leq \alpha \text{ and } P_{\mathbb{F}_0}(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \leq \alpha | \mathbf{a}, \mathbf{x}) \leq \alpha. \quad \square$$

Similar group invariance conditions for obtaining exact level- α randomized and nonrandomized permutation tests have been presented previously in similar proofs by a number of authors [Hoeffding, 1952; Lehmann and Romano, 2005; Romano, 1989, 1990].

We now prove a final lemma relating the consequent of Lemma 2 to the level of the Monte Carlo permutation test.

Lemma 3. If the set of permutations $\mathcal{S}$ is a group under composition (Group Condition) and $P_{\mathbb{F}_0}(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \leq \alpha | \mathbf{a}, \mathbf{x}) \leq \alpha$, then $E_{\mathbb{F}_0}[\hat{\phi}_{t,\alpha,W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{G}_{\text{obs}}, \mathbf{a}) | \mathbf{a}, \mathbf{x}] \leq \alpha$.

Proof. This proof builds on previous similar previous developments [Dwass, 1957; Jöckel, 1986; Phipson and Smyth, 2010] to obtain an exact result for nonrandomized tests. The quantity $(W+1)\alpha$ can be separated into integral and decimal components according to $(W+1)\alpha = \lfloor (W+1)\alpha \rfloor + \varepsilon$, where $\lfloor \cdot \rfloor$ is the floor function and $\varepsilon \in [0, 1)$ is a real number representing the decimal component. For an arbitrary matrix $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$, let $Y_w = \mathbb{I}[\ell(\mathbf{g}_{\text{obs}}, u_{\Pi_w}(\mathbf{a})) - \ell(\mathbf{g}_{\text{obs}}, \mathbf{a})]$ so that the Monte Carlo permutation p-value can be written as:

$$\hat{p}_{t,W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a}) = (W+1)^{-1} \left(\sum_{w=1}^W Y_w + 1 \right)$$

For $\sum_{w=1}^W Y_w \leq \lfloor (W+1)\alpha \rfloor - 1$:

$$\hat{p}_{t,W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a}) \leq (W+1)^{-1} \left(\lfloor (W+1)\alpha \rfloor \right) \leq \alpha$$

For $\sum_{w=1}^W Y_w \geq \lfloor (W+1)\alpha \rfloor$:

$$\begin{aligned} \hat{p}_{t,W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a}) &\geq (W+1)^{-1} \left(\lfloor (W+1)\alpha \rfloor + 1 \right) \\ &= (W+1)^{-1} ((W+1)\alpha - \varepsilon + 1) \\ &= \alpha + (W+1)^{-1} (1 - \varepsilon) \\ &> \alpha \end{aligned}$$

Therefore, $\hat{\phi}_{t,\alpha,W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a}) = 1$ if and only if $\sum_{w=1}^W Y_w \leq \lfloor (W+1)\alpha \rfloor - 1$. The rejection rate of the Monte Carlo permutation test for an arbitrary $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$ can therefore be written as:

$$\begin{aligned} \mathbb{E}\left[\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\mathbf{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a})\right] &= \mathbb{P}\left(\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\mathbf{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a}) = 1\right) \\ &= \mathbb{P}\left(\sum_{w=1}^W Y_w \leq \lfloor (W+1)\alpha \rfloor - 1\right) \end{aligned}$$

The distribution of $\sum_{w=1}^W Y_w$ must now be determined. Let $Y_w \in \{0, 1\}$ be the realized value of Y_w , $\mathcal{S}_t^1(\mathbf{g}_{\text{obs}}, \mathbf{a}) = \{\pi \in \mathcal{S} : t(\mathbf{g}_{\text{obs}}, u_{\pi}(\mathbf{a})) \geq t(\mathbf{g}_{\text{obs}}, \mathbf{a})\}$, and $\mathcal{S}_t^0(\mathbf{g}_{\text{obs}}, \mathbf{a}) = \mathcal{S} \setminus \mathcal{S}_t^1(\mathbf{g}_{\text{obs}}, \mathbf{a})$. Because each random Π_w is independently drawn with replacement from $\mathcal{S}$:

$$\begin{aligned} \mathbb{P}(Y_1 = y_1, \dots, Y_W = y_W) &= \mathbb{P}(\Pi_1 \in \mathcal{S}_t^{y_1}(\mathbf{g}_{\text{obs}}, \mathbf{a}), \dots, \Pi_W \in \mathcal{S}_t^{y_W}(\mathbf{g}_{\text{obs}}, \mathbf{a})) \\ &= \mathbb{P}(\Pi_1 \in \mathcal{S}_t^{y_1}(\mathbf{g}_{\text{obs}}, \mathbf{a})) \dots \mathbb{P}(\Pi_W \in \mathcal{S}_t^{y_W}(\mathbf{g}_{\text{obs}}, \mathbf{a})) \\ &= \mathbb{P}(Y_1 = y_1) \dots \mathbb{P}(Y_W = y_W) \end{aligned}$$

Thus, the Y_w are independent. Moreover, because each Π_w is independently drawn according to the discrete uniform distribution over the R permutations in the set $\mathcal{S}$:

$$\begin{aligned} \mathbb{P}(Y_w = 1) &= \mathbb{P}(\Pi_w \in \mathcal{S}_t^1(\mathbf{g}_{\text{obs}}, \mathbf{a})) \\ &= R^{-1} \sum_{\mathcal{S}} I\left[t(\mathbf{g}_{\text{obs}}, u_{\pi}(\mathbf{a})) \geq t(\mathbf{g}_{\text{obs}}, \mathbf{a})\right] \end{aligned}$$

for all w . Because $t(\mathbf{g}_{\text{obs}}, \mathbf{a}) = t(\left[\vec{\mathbf{g}}_{\text{obs}, 1}, a_1\right], \dots, \left[\vec{\mathbf{g}}_{\text{obs}, n}, a_n\right])$ is symmetric in its vector arguments $\left[\vec{\mathbf{g}}_{\text{obs}, i}, a_i\right]$, it follows for an arbitrary $\pi \in \mathcal{S}$ that:

$$\begin{aligned} t(\mathbf{g}_{\text{obs}}, u_{\pi}(\mathbf{a})) &= t\left(\left[\vec{\mathbf{g}}_{\text{obs}, 1}, a_{\pi(1)}\right], \dots, \left[\vec{\mathbf{g}}_{\text{obs}, n}, a_{\pi(n)}\right]\right) \\ &= t\left(\left[\vec{\mathbf{g}}_{\text{obs}, \pi^{-1}(1)}, a_1\right], \dots, \left[\vec{\mathbf{g}}_{\text{obs}, \pi^{-1}(n)}, a_n\right]\right) \\ &= t\left(u_{\pi^{-1}}(\mathbf{g}_{\text{obs}}), \mathbf{a}\right) \end{aligned}$$

where the second equality results from sorting the vector arguments so that the affection status indicators appear in the same order as in the original data set. The above probability can therefore be equivalently expressed as:

$$\mathbb{P}(Y_w = 1) = R^{-1} \sum_{\mathcal{S}} I\left[t\left(u_{\pi^{-1}}(\mathbf{g}_{\text{obs}}), \mathbf{a}\right) \geq t(\mathbf{g}_{\text{obs}}, \mathbf{a})\right]$$

Consider the mapping $\lambda(\pi) = \pi^{-1}$ that takes an arbitrary permutation π to its unique inverse permutation π^{-1} . Because $\mathcal{S}$ is a group under composition, $\pi^{-1} \in \mathcal{S}$ for all $\pi \in \mathcal{S}$ [Robinson, 2003], so $\lambda : \mathcal{S} \rightarrow \mathcal{S}$. Because π is also by definition the unique inverse for π^{-1} , it follows that $\lambda(\lambda(\pi)) = \lambda(\pi^{-1}) = \pi$ for all $\pi \in \mathcal{S}$. Therefore, λ is its own unique inverse, which implies that λ is a one-to-one mapping of $\mathcal{S}$ onto itself [Robinson, 2003]. Because λ is one-to-one, different $\pi \in \mathcal{S}$ map to different $\pi^{-1} \in \mathcal{S}$, and the above sum contains R terms each corresponding to a distinct $\pi^{-1} \in \mathcal{S}$. Moreover, because any collection of R distinct elements of $\mathcal{S}$ must constitute the entire set, the above sum can equivalently be taken over all $\pi \in \mathcal{S}$. This result implies that:

$$\begin{aligned} P(Y_w = 1) &= R^{-1} \sum_{\mathcal{S}} I \left[t \left(u_{\pi}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right) \geq t(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right] \\ &= p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \end{aligned}$$

for all w . Thus, for an arbitrary $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$, each Y_w is an independent Bernoulli($p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a})$) random variable, and it immediately follows that the distribution of $\sum_{w=1}^W Y_w$ is Binomial($W, p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a})$).

Based on the preceding results:

$$\begin{aligned} E \left[\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a}) \right] &= \sum_{y=0}^{\lfloor (W+1)\alpha \rfloor - 1} \binom{W}{y} \left(p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right)^y \left(1 - p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right)^{W-y} \\ &= \psi_{\alpha, W} \left(p_t^{\mathcal{S}}(\mathbf{g}_{\text{obs}}, \mathbf{a}) \right) \end{aligned}$$

The law of iterated expectations then yields an expression for the unconditional rejection rate of the Monte Carlo permutation test under $\mathbb{F}_0$:

$$\begin{aligned} E_{\mathbb{F}_0} \left[\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{G}_{\text{obs}}, \mathbf{a}) \middle| \mathbf{a}, \mathbf{x} \right] &= E_{\mathbb{F}_0} \left[E \left[\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{g}_{\text{obs}}, \mathbf{a}) \middle| \mathbf{G}_{\text{obs}} = \mathbf{g}_{\text{obs}} \right] \middle| \mathbf{a}, \mathbf{x} \right] \\ &= E_{\mathbb{F}_0} \left[\psi_{\alpha, W} \left(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \right) \middle| \mathbf{a}, \mathbf{x} \right] \end{aligned}$$

The well-known relationship between the binomial survival function and the incomplete beta function [Zelen and Severo, 1972] can be used to obtain an alternative expression for $\psi_{\alpha, W}(p)$ similar to one given by Jöckel [Jöckel, 1986]:

$$\begin{aligned}\psi_{\alpha, W}(p) &= 1 - \sum_{y=\lfloor (W+1)\alpha \rfloor}^W \binom{W}{y} p^y (1-p)^{W-y} \\ &= 1 - B(\lfloor (W+1)\alpha \rfloor, W - \lfloor (W+1)\alpha \rfloor + 1) \int_0^p z^{\lfloor (W+1)\alpha \rfloor - 1} (1-z)^{W - \lfloor (W+1)\alpha \rfloor} dz\end{aligned}$$

where $B(a, b) = \int_0^1 z^{a-1} (1-z)^{b-1} dz$ is the beta function [Davis, 1972]. Note that $\psi_{\alpha, W}(1) = 0$ by the definition of the beta function. Based on the relationships between the beta function, the gamma function, and factorials [Davis, 1972], the above expression can be further simplified:

$$\begin{aligned}\psi_{\alpha, W}(p) &= 1 - \frac{\Gamma(W+1)}{\Gamma(W - \lfloor (W+1)\alpha \rfloor + 1) \Gamma(\lfloor (W+1)\alpha \rfloor)} \int_0^p z^{\lfloor (W+1)\alpha \rfloor - 1} (1-z)^{W - \lfloor (W+1)\alpha \rfloor} dz \\ &= 1 - \frac{W!}{(W - \lfloor (W+1)\alpha \rfloor)! (\lfloor (W+1)\alpha \rfloor - 1)!} \int_0^p z^{\lfloor (W+1)\alpha \rfloor - 1} (1-z)^{W - \lfloor (W+1)\alpha \rfloor} dz \\ &= 1 - W \binom{W-1}{\lfloor (W+1)\alpha \rfloor - 1} \int_0^p z^{\lfloor (W+1)\alpha \rfloor - 1} (1-z)^{W - \lfloor (W+1)\alpha \rfloor} dz\end{aligned}$$

This expression for $\psi_{\alpha, W}(p)$ has been obtained previously by Dwass [Dwass, 1957].

While other studies of the finite-sample performance of Monte Carlo permutation tests have approximated the unconditional rejection rate under the null hypothesis by assuming that $p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a})$ follows a uniform distribution [Dwass, 1957; Phipson and Smyth, 2010], we will depart from this practice in order to obtain an exact bound. Let $P \sim \text{Uniform}(0, 1)$. Because $P_{\mathbb{F}_0}(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \leq \alpha | \mathbf{a}, \mathbf{x}) \leq \alpha$ for an arbitrary nominal α , it holds for all nominal $\alpha \in (0, 1)$, and it immediately follows that $P_{\mathbb{F}_0}(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \leq p | \mathbf{a}, \mathbf{x}) \leq p$ and $P_{\mathbb{F}_0}(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) > p | \mathbf{a}, \mathbf{x}) \geq 1 - p = P(P > p)$ for all $p \in (0, 1)$. Therefore, the random variable $p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a})$ is stochastically larger than P under the null hypothesis [Ross, 1996]. Moreover, the function $1 - \psi_{\alpha, W}(p)$ is increasing in p for all $p \in [0, 1]$ because:

$$\frac{d(1 - \psi_{\alpha, W}(p))}{dp} = W \binom{W-1}{\lfloor (W+1)\alpha \rfloor - 1} p^{\lfloor (W+1)\alpha \rfloor - 1} (1-p)^{W - \lfloor (W+1)\alpha \rfloor} \geq 0$$

It follows that [Ross, 1996]:

$$\begin{aligned}
\mathbb{E}_{\mathbb{F}_0} \left[1 - \psi_{\alpha, W} \left(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \right) \middle| \mathbf{a}, \mathbf{x} \right] &\geq \mathbb{E} \left[1 - \psi_{\alpha, W}^{(P)} \right] \\
1 - \mathbb{E}_{\mathbb{F}_0} \left[\psi_{\alpha, W} \left(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \right) \middle| \mathbf{a}, \mathbf{x} \right] &\geq 1 - \mathbb{E} \left[\psi_{\alpha, W}^{(P)} \right] \\
\mathbb{E}_{\mathbb{F}_0} \left[\psi_{\alpha, W} \left(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \right) \middle| \mathbf{a}, \mathbf{x} \right] &\leq \mathbb{E} \left[\psi_{\alpha, W}^{(P)} \right]
\end{aligned}$$

The expectation $\mathbb{E}[\psi_{\alpha, W}^{(P)}]$ can be calculated using integration by parts [Dwass, 1957; Jöckel, 1986]:

$$\begin{aligned}
\mathbb{E}[\psi_{\alpha, W}^{(P)}] &= \int_0^1 \psi_{\alpha, W}^{(p)} dp \\
&= [\psi_{\alpha, W}^{(p)} p]_0^1 - \int_0^1 p \psi'_{\alpha, W}^{(p)} dp \\
&= \psi_{\alpha, W}^{(1)} 1 - \psi_{\alpha, W}^{(0)} 0 + W \left(\frac{W-1}{\lfloor (W+1)\alpha \rfloor - 1} \right) \int_0^1 p^{\lfloor (W+1)\alpha \rfloor (1-p)^{W - \lfloor (W+1)\alpha \rfloor}} dp \\
&= W \left(\frac{W-1}{\lfloor (W+1)\alpha \rfloor - 1} \right) \text{B} \left(\lfloor (W+1)\alpha \rfloor + 1, W - \lfloor (W+1)\alpha \rfloor + 1 \right) \\
&= \frac{W! \Gamma(\lfloor (W+1)\alpha \rfloor + 1) \Gamma(W - \lfloor (W+1)\alpha \rfloor + 1)}{(\lfloor (W+1)\alpha \rfloor - 1)! (W - \lfloor (W+1)\alpha \rfloor)! \Gamma(W+2)} \\
&= \frac{W! (\lfloor (W+1)\alpha \rfloor)! (W - \lfloor (W+1)\alpha \rfloor)!}{(\lfloor (W+1)\alpha \rfloor - 1)! (W - \lfloor (W+1)\alpha \rfloor)! (W+1)!} \\
&= \frac{\lfloor (W+1)\alpha \rfloor}{(W+1)} \\
&\leq \alpha
\end{aligned}$$

Based on the preceding results, it follows that:

$$\begin{aligned}
\mathbb{E}_{\mathbb{F}_0} \left[\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{G}_{\text{obs}}, \mathbf{a}) \middle| \mathbf{a}, \mathbf{x} \right] &= \mathbb{E}_{\mathbb{F}_0} \left[\psi_{\alpha, W} \left(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \right) \middle| \mathbf{a}, \mathbf{x} \right] \\
&\leq \mathbb{E} \left[\psi_{\alpha, W}^{(P)} \right] \\
&\leq \alpha
\end{aligned}$$

□

A related finite-sample result for Monte Carlo tests has been obtained previously [Jöckel, 1986], but this result applies to randomized tests rather than the nonrandomized test in (4). Other related results on the exact level of Monte Carlo permutation tests have also been given previously with a different proof [Romano, 1989] or without proof [Lehmann and Romano, 2005].

We now present the main result.

Theorem. If the set of permutations $\mathcal{S}$ is a group under composition (Group Condition) and, for all permutations π in the set $\mathcal{S}$, the induced PMF of the random matrix $u_{\pi}(\mathbf{G}_{\text{obs}})$ under

$\mathbb{F}_0$ is identical to $\mathbb{F}_0$ (Invariance Condition), then the Monte Carlo permutation test based on $\mathcal{S}$ and t is exact level α .

Proof. Under the assumptions of the theorem, it follows from Lemma 2 that

$P_{\mathbb{F}_0}(p_t^{\mathcal{S}}(\mathbf{G}_{\text{obs}}, \mathbf{a}) \leq \alpha | \mathbf{a}, \mathbf{x}) \leq \alpha$. It then follows from Lemma 3 that

$E_{\mathbb{F}_0}[\hat{\phi}_{t, \alpha, W}^{\mathcal{S}}(\boldsymbol{\Pi}, \mathbf{G}_{\text{obs}}, \mathbf{a}) | \mathbf{a}, \mathbf{x}] \leq \alpha$, which implies that the Monte Carlo permutation test based on $\mathcal{S}$ and t has a type I error rate of at most α under $\mathbb{F}_0$. Because this result does not depend on the sample size, the Monte Carlo permutation test based on $\mathcal{S}$ and t is exact level α . $\square$

We conclude by deriving an expression for the induced PMF of the random matrix $u_{\pi}(\mathbf{G}_{\text{obs}})$ under $\mathbb{F}_0$ to use in our statement of the Invariance Condition in the main text. For arbitrary $\pi \in \mathcal{S}$ and $\mathbf{g}_{\text{obs}} \in \mathcal{X}^n$, $u_{\pi}(\mathbf{G}_{\text{obs}}) = \mathbf{g}_{\text{obs}}$ if and only if $\mathbf{G}_{\text{obs}} = u_{\pi}^{-1}(\mathbf{g}_{\text{obs}})$ because u_{π} is one-to-one and onto. The induced PMF of $u_{\pi}(\mathbf{G}_{\text{obs}})$ for this $\pi \in \mathcal{S}$ can therefore be obtained from (3):

$$\begin{aligned} P_{\mathbb{F}_0}(u_{\pi}(\mathbf{G}_{\text{obs}}) = \mathbf{g}_{\text{obs}} | \mathbf{a}, \mathbf{x}) &= P_{\mathbb{F}_0}(\mathbf{G}_{\text{obs}} = u_{\pi}^{-1}(\mathbf{g}_{\text{obs}}) | \mathbf{a}, \mathbf{x}) \\ &= \mathbb{F}_0(u_{\pi}^{-1}(\mathbf{g}_{\text{obs}}) | \mathbf{a}, \mathbf{x}) \\ &= \sum_Q \left[\prod_{i=1}^n \sum_{\eta^{-1}(\vec{\mathbf{g}}_{\text{obs}, \pi^{-1}(i)})} f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | q) f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, \vec{\mathbf{x}}_{\pi(i)}, q) \right] f_Q(q) \end{aligned}$$

The observation for individual $k = \pi^{-1}(i)$ in this product has the PMF for individual $\pi(k) = \pi(\pi^{-1}(i)) = i$, so the observation for individual i must have the PMF for individual $\pi(i)$. Thus, the induced PMF of $u_{\pi}(\mathbf{G}_{\text{obs}})$ for an arbitrary $\pi \in \mathcal{S}$ under $\mathbb{F}_0$ can be equivalently written as:

$$\begin{aligned} P_{\mathbb{F}_0}(u_{\pi}(\mathbf{G}_{\text{obs}}) = \mathbf{g}_{\text{obs}} | \mathbf{a}, \mathbf{x}) &= \\ \sum_Q \left[\prod_{i=1}^n \sum_{\eta^{-1}(\vec{\mathbf{g}}_{\text{obs}, i})} f_{\vec{\mathbf{G}}}(\vec{\mathbf{g}} | q) f_{\vec{\mathbf{M}}}(\vec{\mathbf{m}} | \vec{\mathbf{g}}, \vec{\mathbf{x}}_{\pi(i)}, q) \right] f_Q(q) \end{aligned}$$

Acknowledgments

This research was supported by Grant Number RC2HG005605 from the National Human Genome Research Institute (ERM, DDK), the Dr. John T. Macdonald Foundation Department of Human Genetics at the University of Miami Miller School of Medicine (DDK), as well as the Dr. Louis J. Elsas Research Award in Biochemical Genetics from the Dr. John T. Macdonald Foundation Department of Human Genetics at the University of Miami

Miller School of Medicine (DDK). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health. We also wish to thank the Center for Computational Science at the University of Miami for the computing infrastructure and support that made our simulations possible.

References

- Browning BL. 2008; PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics*. 9(1):309. [PubMed: 18620604]
- Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, et al. 2007; Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 447(7145):661–678. [PubMed: 17554300]
- Chapman J, Whittaker J. 2008; Analysis of multiple SNPs in a candidate gene or region. *Genet Epidemiol*. 32(6):560–566. [PubMed: 18428428]
- Chung, EY, Romano, JP. Technical Report No 2011-05, General Department Series. Stanford, CA: Stanford University; 2011. Exact and asymptotically robust permutation tests. at http://statistics.stanford.edu/~ckirby/reports/2010_2019/reports2011.html. [Accessed on April 30, 2012]
- Clopper CJ, Pearson ES. 1934; The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 26(4):404–413.
- Davis, PJ. Gamma function and related functions. In: Abramowitz, M, Stegun, IA, editors. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, 10th printing, with correction. Washington, DC: U.S. Department of Commerce, National Bureau of Standards; 1972. 253–294.
- Davison, AC, Hinkley, DV. *Bootstrap methods and their application*. New York: Cambridge University Press; 1997.
- Doerge RW, Churchill GA. 1996; Permutation tests for multiple loci affecting a quantitative character. *Genetics*. 142(1):285–294. [PubMed: 8770605]
- Dwass M. 1957; Modified randomization tests for nonparametric hypotheses. *Ann Math Stat*. 28(1): 181–187.
- Edgington ES. 1983; The role of permutation groups in randomization tests. *J Educ Stat*. 8(2):121–135.
- Entsuah RA. 1990; Randomization procedures for analyzing clinical trial data with treatment related withdrawals. *Commun Stat - Theory Methods*. 19(10):3859–3880.
- Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. 2012; A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *Am J Hum Genet*. 91(2):215–223. [PubMed: 22818855]
- Ernst MD. 2004; Permutation methods: a basis for exact inference. *Stat Sci*. 19(4):676–685.
- Good, PI. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. 2nd. New York: Springer; 2000.
- Hoeffding W. 1952; The large-sample power of tests based on permutations of observations. *Ann Math Stat*. 23(2):169–192.
- Hoffmann TJ, Marini NJ, Witte JS. 2010; Comprehensive approach to analyzing rare genetic variants. *PLoS ONE*. 5(11):e13584. [PubMed: 21072163]
- Hoh J, Ott J. 2000; Scan statistics to scan markers for susceptibility genes. *Proc Natl Acad Sci U S A*. 97(17):9615–9617. [PubMed: 10931953]
- Hoh J, Ott J. 2003; Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet*. 4(9):701–709. [PubMed: 12951571]
- Hoh J, Wille A, Ott J. 2001; Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res*. 11(12):2115–2119. [PubMed: 11731502]
- Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. 2011; A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet*. 7(2):e1001289. [PubMed: 21304886]

- Jöckel KH. 1986; Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Ann Stat.* 14(1):336–347.
- Kennes LN, Hilgers RD, Heussen N. 2012; Choice of the reference set in a randomization test based on linear ranks in the presence of missing values. *Commun Stat - Simul Comput.* 41(7):1051–1061.
- Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. 2007; A randomization test for controlling population stratification in whole-genome association studies. *Am J Hum Genet.* 81(5):895–905. [PubMed: 17924333]
- Kinnamon DD, Hershberger RE, Martin ER. 2012; Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. *PLoS ONE.* 7(2):e30238. [PubMed: 22363423]
- Lehmann, EL, Romano, JP. Testing statistical hypotheses. 3rd. New York: Springer; 2005.
- Lehmann EL, Stein C. 1949; On the theory of some non-parametric hypotheses. *Ann Math Stat.* 20(1): 28–45.
- Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong X, Aponte JL, Mooser VE, Chisoe SL, Whittaker JC, et al. 2011; Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS ONE.* 6(9):e24945. [PubMed: 21949800]
- Little, RJA, Rubin, DB. Statistical analysis with missing data. 2nd. Hoboken: John Wiley & Sons; 2002.
- Madsen BE, Browning SR. 2009; A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5(2):e1000384. [PubMed: 19214210]
- Manichaikul A, Palmer AA, Sen , Broman KW. 2007; Significance thresholds for quantitative trait locus mapping under selective genotyping. *Genetics.* 177(3):1963–1966. [PubMed: 17720902]
- Mundry R. 1999; Testing related samples with missing values: a permutation approach. *Anim Behav.* 58(5):1143–1153. [PubMed: 10564619]
- Naddeo S. 2002; Permutation-based pairwise comparisons for assessing the homogeneity of probability distributions and missing data rates. *Metron - Int J Stat.* 60(3-4):125–138.
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melandier M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011; Testing for an unusual distribution of rare variants. *PLoS Genet.* 7(3):e1001322. [PubMed: 21408211]
- Pesarin, F, Salmaso, L. Permutation tests for complex data: theory, applications and software. Chichester, UK: Wiley; 2010.
- Phipson B, Smyth GK. 2010; Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol.* 9(1)
- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010; Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 86(6):832–838. [PubMed: 20471002]
- Quesenberry CP, Hales C. 1980; Concentration bands for uniformity plots. *J Stat Comput Simul.* 11(1): 41–53.
- Robinson, DJS. An introduction to abstract algebra. New York: Walter de Gruyter; 2003.
- Romano JP. 1989; Bootstrap and randomization tests of some nonparametric hypotheses. *Ann Stat.* 17(1):141–159.
- Romano JP. 1990; On the behavior of randomization tests without a group invariance assumption. *J Am Stat Assoc.* 85(411):686–692.
- Ross, SM. Stochastic processes. 2nd ed.. New York: John Wiley & Sons; 1996.
- Shterev I, Jung SH, George S, Owzar K. 2010; permGPU: Using graphics processing units in RNA microarray association studies. *BMC Bioinformatics.* 11(1):329. [PubMed: 20553619]
- Stirling WD. 1982; Enhancements to aid interpretation of probability plots. *J R Stat Soc Ser Stat.* 31(3):211–220.
- Weir, BS. Genetic data analysis II: methods for discrete population genetic data. Sunderland, MA: Sinauer Associates; 1996.

- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011; Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 89(1):82–93. [PubMed: 21737059]
- Zelen, M, Severo, NC. Probability functions. In: Abramowitz, M, Stegun, IA, editors. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, 10th printing, with corrections. Washington, DC: U.S. Department of Commerce National Bureau of Standards; 1972. 925–996.

Figure 1. Quantile-quantile plots of Monte Carlo permutation p-values based on the CA max statistic

Observed order statistics (vertical axis) plotted against their expectations under a theoretical uniform distribution (horizontal axis), both in $-\log_{10}$ units. Order statistics are based on 1,000 realizations from the marginal distribution of each Monte Carlo permutation p-value under each missing data process (Models 0-6) and a null disease model; plots are faceted by permutation set (standard or restricted) and missing data process. Gray lines represent simultaneous 95% concentration intervals for the order statistics under the uniform distribution. Observed order statistics falling within (outside of) their corresponding concentration intervals are plotted in black (red).

Figure 2. Quantile-quantile plots of Monte Carlo permutation p-values based on the CA sum statistic

Observed order statistics (vertical axis) plotted against their expectations under a theoretical uniform distribution (horizontal axis), both in $-\log_{10}$ units. Order statistics are based on 1,000 realizations from the marginal distribution of each Monte Carlo permutation p-value under each missing data process (Models 0-6) and a null disease model; plots are faceted by permutation set (standard or restricted) and missing data process. Gray lines represent simultaneous 95% concentration intervals for the order statistics under the uniform distribution. Observed order statistics falling within (outside of) their corresponding concentration intervals are plotted in black (red).

Figure 3. Quantile-quantile plots of Monte Carlo permutation p-values based on the FW burden statistic

Observed order statistics (vertical axis) plotted against their expectations under a theoretical uniform distribution (horizontal axis), both in $-\log_{10}$ units. Order statistics are based on 1,000 realizations from the marginal distribution of each Monte Carlo permutation p-value under each missing data process (Models 0-6) and a null disease model; plots are faceted by permutation set (standard or restricted) and missing data process. Gray lines represent simultaneous 95% concentration intervals for the order statistics under the uniform distribution. Observed order statistics falling within (outside of) their corresponding concentration intervals are plotted in black (red).

Figure 4. Standard versus restricted Monte Carlo permutation P-values based on the CA max statistic

The restricted P-value (vertical axis) is plotted against the corresponding standard P-value (horizontal axis) for the same data set, both in $-\log_{10}$ units. Points represent 1,000 replicates from each covariate-independent missing data process (Models 0–2) under a disease model with 50 rare risk variants ($\text{MAF} < 0.01$; $\text{OR} = 2$); plots are faceted by missing data process.

Figure 5. Standard versus restricted Monte Carlo permutation P-values based on the CA sum statistic

The restricted P-value (vertical axis) is plotted against the corresponding standard P-value (horizontal axis) for the same data set, both in $-\log_{10}$ units. Points represent 1,000 replicates from each covariate-independent missing data process (Models 0–2) under a disease model with 50 rare risk variants ($\text{MAF} < 0.01$; $\text{OR} = 2$); plots are faceted by missing data process.

Figure 6. Standard versus restricted Monte Carlo permutation P-values based on the FW burden statistic

The restricted P-value (vertical axis) is plotted against the corresponding standard P-value (horizontal axis) for the same data set, both in $-\log_{10}$ units. Points represent 1,000 replicates from each covariate-independent missing data process (Models 0–2) under a disease model with 50 rare risk variants ($\text{MAF} < 0.01$; $\text{OR} = 2$); plots are faceted by missing data process.

Table I
Summary of missing data processes used in the simulations

The symbols $(\gamma_{0,c}, \gamma_{1,c})$ refer to the user-specified parameters for the missing data model given in equation (5). In a case-control sample from a randomly mating population under the null hypothesis, $\gamma_{0,c}$ can be interpreted as the expected rate of missing genotypes in batch c for all data sets. Regardless of the disease model, $\gamma_{1,c}$ is the linear change in the missing genotype probability with each additional minor allele in the true genotype for individuals in batch c .

Missing Data Process	Batch 0 ($\gamma_{0,0}, \gamma_{1,0}$)	Batch 1 ($\gamma_{0,1}, \gamma_{1,1}$)	Batch 2 ($\gamma_{0,2}, \gamma_{1,2}$)	Batch 3 ($\gamma_{0,3}, \gamma_{1,3}$)
Model 0	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
Model 1	(0.10, 0.00)	(0.10, 0.00)	(0.10, 0.00)	(0.10, 0.00)
Model 2	(0.10, 0.03)	(0.10, 0.03)	(0.10, 0.03)	(0.10, 0.03)
Model 3	(0.10, 0.00)	(0.10, 0.02)	(0.10, 0.04)	(0.10, 0.06)
Model 4	(0.07, 0.00)	(0.09, 0.00)	(0.11, 0.00)	(0.13, 0.00)
Model 5	(0.07, 0.03)	(0.09, 0.03)	(0.11, 0.03)	(0.13, 0.03)
Model 6	(0.07, 0.00)	(0.09, 0.02)	(0.11, 0.04)	(0.13, 0.06)

Table II
Type I error rates of Monte Carlo permutation tests based on the CA max statistic

Estimates and Clopper-Pearson exact 95% confidence intervals are based on 1,000 realizations from the marginal distribution of each Monte Carlo permutation p-value under each missing data process (Models 0-6) and a null disease model. Estimated type I error rates in bold had exact 95% confidence intervals entirely above the nominal α level.

Missing Data Process	Permutation Set	Estimate (Exact 95% Confidence Interval)		
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Model 0	Standard	0.038 (0.027, 0.052)	0.011 (0.006, 0.020)	0.002 (0.000, 0.007)
	Restricted	0.045 (0.033, 0.060)	0.009 (0.004, 0.017)	0.002 (0.000, 0.007)
Model 1	Standard	0.045 (0.033, 0.060)	0.006 (0.002, 0.013)	0.001 (0.000, 0.006)
	Restricted	0.049 (0.036, 0.064)	0.007 (0.003, 0.014)	0.000 (0.000, 0.004)
Model 2	Standard	0.060 (0.046, 0.077)	0.012 (0.006, 0.021)	0.002 (0.000, 0.007)
	Restricted	0.055 (0.042, 0.071)	0.013 (0.007, 0.022)	0.001 (0.000, 0.006)
Model 3	Standard	0.083 (0.067, 0.102)	0.018 (0.011, 0.028)	0.002 (0.000, 0.007)
	Restricted	0.052 (0.039, 0.068)	0.009 (0.004, 0.017)	0.002 (0.000, 0.007)
Model 4	Standard	0.054 (0.041, 0.070)	0.010 (0.005, 0.018)	0.002 (0.000, 0.007)
	Restricted	0.049 (0.036, 0.064)	0.007 (0.003, 0.014)	0.002 (0.000, 0.007)
Model 5	Standard	0.046 (0.034, 0.061)	0.009 (0.004, 0.017)	0.001 (0.000, 0.006)
	Restricted	0.046 (0.034, 0.061)	0.011 (0.006, 0.020)	0.002 (0.000, 0.007)
Model 6	Standard	0.077 (0.061, 0.095)	0.022 (0.014, 0.033)	0.002 (0.000, 0.007)
	Restricted	0.050 (0.037, 0.065)	0.008 (0.003, 0.016)	0.001 (0.000, 0.006)

Table III
Type I error rates of Monte Carlo permutation tests based on the CA sum statistic

Estimates and Clopper-Pearson exact 95% confidence intervals are based on 1,000 realizations from the marginal distribution of each Monte Carlo permutation p-value under each missing data process (Models 0-6) and a null disease model. Estimated type I error rates in bold had exact 95% confidence intervals entirely above the nominal α level.

Missing Data Process	Permutation Set	Estimate (Exact 95% Confidence Interval)		
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Model 0	Standard	0.052 (0.039, 0.068)	0.010 (0.005, 0.018)	0.002 (0.000, 0.007)
	Restricted	0.044 (0.032, 0.059)	0.006 (0.002, 0.013)	0.001 (0.000, 0.006)
Model 1	Standard	0.052 (0.039, 0.068)	0.008 (0.003, 0.016)	0.002 (0.000, 0.007)
	Restricted	0.045 (0.033, 0.060)	0.006 (0.002, 0.013)	0.001 (0.000, 0.006)
Model 2	Standard	0.053 (0.040, 0.069)	0.008 (0.003, 0.016)	0.002 (0.000, 0.007)
	Restricted	0.044 (0.032, 0.059)	0.006 (0.002, 0.013)	0.001 (0.000, 0.006)
Model 3	Standard	0.077 (0.061, 0.095)	0.020 (0.012, 0.031)	0.004 (0.001, 0.010)
	Restricted	0.042 (0.030, 0.056)	0.005 (0.002, 0.012)	0.001 (0.000, 0.006)
Model 4	Standard	0.052 (0.039, 0.068)	0.009 (0.004, 0.017)	0.003 (0.001, 0.009)
	Restricted	0.042 (0.030, 0.056)	0.007 (0.003, 0.014)	0.001 (0.000, 0.006)
Model 5	Standard	0.054 (0.041, 0.070)	0.010 (0.005, 0.018)	0.002 (0.000, 0.007)
	Restricted	0.043 (0.031, 0.057)	0.008 (0.003, 0.016)	0.001 (0.000, 0.006)
Model 6	Standard	0.081 (0.065, 0.100)	0.018 (0.011, 0.028)	0.003 (0.001, 0.009)
	Restricted	0.050 (0.037, 0.065)	0.006 (0.002, 0.013)	0.001 (0.000, 0.006)

Table IV
Type I error rates of Monte Carlo permutation tests based on the FW burden statistic

Estimates and Clopper-Pearson exact 95% confidence intervals are based on 1,000 realizations from the marginal distribution of each Monte Carlo permutation p-value under each missing data process (Models 0-6) and a null disease model. Estimated type I error rates in bold had exact 95% confidence intervals entirely above the nominal α level.

Missing Data Process	Permutation Set	Estimate (Exact 95% Confidence Interval)		
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Model 0	Standard	0.047 (0.035, 0.062)	0.011 (0.006, 0.020)	0.001 (0.000, 0.006)
	Restricted	0.046 (0.034, 0.061)	0.011 (0.006, 0.020)	0.000 (0.000, 0.004)
Model 1	Standard	0.047 (0.035, 0.062)	0.011 (0.006, 0.020)	0.000 (0.000, 0.004)
	Restricted	0.049 (0.036, 0.064)	0.010 (0.005, 0.018)	0.001 (0.000, 0.006)
Model 2	Standard	0.054 (0.041, 0.070)	0.011 (0.006, 0.020)	0.003 (0.001, 0.009)
	Restricted	0.048 (0.036, 0.063)	0.010 (0.005, 0.018)	0.002 (0.000, 0.007)
Model 3	Standard	0.627 (0.596, 0.657)	0.417 (0.386, 0.448)	0.195 (0.171, 0.221)
	Restricted	0.059 (0.045, 0.075)	0.017 (0.010, 0.027)	0.000 (0.000, 0.004)
Model 4	Standard	0.776 (0.749, 0.801)	0.600 (0.569, 0.631)	0.367 (0.337, 0.398)
	Restricted	0.058 (0.044, 0.074)	0.017 (0.010, 0.027)	0.002 (0.000, 0.007)
Model 5	Standard	0.785 (0.758, 0.810)	0.623 (0.592, 0.653)	0.397 (0.367, 0.428)
	Restricted	0.055 (0.042, 0.071)	0.012 (0.006, 0.021)	0.000 (0.000, 0.004)
Model 6	Standard	0.980 (0.969, 0.988)	0.955 (0.940, 0.967)	0.886 (0.865, 0.905)
	Restricted	0.060 (0.046, 0.077)	0.015 (0.008, 0.025)	0.000 (0.000, 0.004)

Table V
Estimated power under a disease model with rare risk variants

Estimates are based on 1,000 realizations from the marginal distribution of each Monte Carlo permutation p-value under each covariate-independent missing data process (Models 0-2) and a disease model with 50 rare risk variants (MAF<0.01; OR=2).

α	Missing Data Process	Permutation Set	CA max	CA sum	FW burden
0.05	Model 0	Standard	0.732	0.674	0.327
		Restricted	0.696	0.644	0.306
	Model 1	Standard	0.727	0.677	0.323
		Restricted	0.675	0.633	0.297
	Model 2	Standard	0.710	0.669	0.325
		Restricted	0.664	0.639	0.292
0.01	Model 0	Standard	0.574	0.448	0.204
		Restricted	0.533	0.436	0.169
	Model 1	Standard	0.555	0.451	0.196
		Restricted	0.505	0.427	0.165
	Model 2	Standard	0.559	0.447	0.200
		Restricted	0.508	0.431	0.161
0.001	Model 0	Standard	0.412	0.248	0.121
		Restricted	0.360	0.243	0.087
	Model 1	Standard	0.391	0.247	0.120
		Restricted	0.347	0.244	0.082
	Model 2	Standard	0.384	0.245	0.117
		Restricted	0.329	0.232	0.083