A Permutation Procedure to Correct for Confounders in Case-Control Studies, Including Tests of Rare Variation

Michael P. Epstein,[1,*] Richard Duncan,[1] Yunxuan Jiang,[2] Karen N. Conneely,[1] Andrew S. Allen,[3] and Glen A. Satten[4]

Many case-control tests of rare variation are implemented in statistical frameworks that make correction for confounders like population stratification difficult. Simple permutation of disease status is unacceptable for resolving this issue because the replicate data sets do not have the same confounding as the original data set. These limitations make it difficult to apply rare-variant tests to samples in which confounding most likely exists, e.g., samples collected from admixed populations. To enable the use of such rare-variant methods in structured samples, as well as to facilitate permutation tests for any situation in which case-control tests require adjustment for confounding covariates, we propose to establish the significance of a rare-variant test via a modified permutation procedure. Our procedure uses Fisher's noncentral hypergeometric distribution to generate permuted data sets with the same structure present in the actual data set such that inference is valid in the presence of confounding factors. We use simulated sequence data based on coalescent models to show that our permutation strategy corrects for confounding due to population stratification that, if ignored, would otherwise inflate the size of a rare-variant test. We further illustrate the approach by using sequence data from the Dallas Heart Study of energy metabolism traits. Researchers can implement our permutation approach by using the R package BiasedUrn.

Introduction

Association mapping of rare variants (those with a minor allele frequency [MAF] $< 1\%$) and less-common variants (those with a MAF between 1% and 5%) requires different analytic methods from those typically used for the detection of common genetic variants in a genome-wide association study (GWAS). For a case-control study, GWAS-based statistical methods usually consider genetic variants individually and develop an association test based on allele-frequency differences of the variant between cases and controls. For rare-variant analysis, this strategy will most likely have inadequate power given that power decreases with decreasing allele frequency for fixed sample and effect sizes. Thus, many recent publications recommend association tests that aggregate rare and less-common variants within a gene or region for analysis. Many "burden" tests pool such variants into a composite variable and then test for association between that composite variable and disease status. The composite variable could be a binary indicator of whether a subject possesses a rare variant (defined as a variant below some allele-frequency threshold value) within the region of interest[1–3] or could be a sum over the number of rare variants that a subject possesses across that region.[4–6] Other tests that remain powerful when testing regions containing risk and protective rare variants include the replication-based test (RBT),[7] the C-alpha test,[8] and the weighted haplotype and imputation-based test (WHaIT),[9] among others.[10–13]

An open issue with rare-variant association tests is their validity in the presence of confounders such as population stratification. Confounding from population stratification occurs when genetic variation is correlated with variation in disease risk across latent subpopulations or geographic gradients. This confounding is likely to arise in association studies of rare variants because such variants might be unique to a particular ancestral group.[14,15] Like common-variant tests employed in GWASs, certain rare-variant tests like SKAT[12] and others implemented in a logistic-regression framework[2,5,13] can incorporate summary measures of such variables as covariates. Unfortunately, many other rare-variant association tests that exist today are implemented in statistical frameworks that do not allow such straightforward corrections for confounders. These include the RBT method,[7] which creates a statistic that detects enrichment of rare variants in cases versus controls and vice versa and that can also incorporate adaptive weights, and the C-alpha test,[8] which uses a general homogeneity score statistic[16] to test whether the variance in the proportion of cases that possess rare variants within a region differs from the expected binomial distribution if all variants are neutral. Many other rare-variant tests that do not correct for covariates also exist.[4,9,17–20]

In addition to those rare-variant tests that cannot directly adjust for confounders, there are other rare-variant tests that offer only limited mechanisms to correct for such variables. One such example is the variable-allele-frequency threshold test,[3] which proposes

[1]Department of Human Genetics, Emory University, Atlanta, GA 30322, USA; [2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA; [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA; [4]Centers for Disease Control and Prevention, Atlanta, GA 30333, USA
*Correspondence: mpepste@emory.edu

http://dx.doi.org/10.1016/j.ajhg.2012.06.004.

adjusting for covariates by the replacement of disease-outcome variables in the test statistic with residuals from a regression analysis of disease outcome on covariates under a linear model. Such a strategy is not ideal because studies have shown that applying a linear model to binary disease-outcome data can lead to an inappropriate correction for confounding.[21–23] Another test with restricted ability to correct for confounders is the cumulative minor-allele test (CMAT),[6] which allows for the adjustment of a single categorical covariate. Such an adjustment will be insufficient if there is a need to model multiple continuous covariates, such as summary ancestry measures based on significant eigenvectors from a principal-component analysis of genome-wide SNP data.

For case-control studies, we propose a method that enables the adjustment of any association test, including the rare-variant association tests discussed previously, for an arbitrary number of categorical and continuous confounding covariates. These covariates can include summary measures of ancestry to correct for confounding due to population stratification. Our strategy involves a permutation procedure that repeatedly shuffles the disease outcomes of study participants in a way that generates permuted (replicate) data sets with the same extent of confounding found in the original data set. The distribution of any test statistic, calculated with these replicate data sets, is a valid null distribution for the test statistic. Whereas other rare-variant tests (including the RBT, CMAT, and WHaIT) already use permutations to establish significance thresholds, such permutations are performed by random shuffling of case or control status among individuals; this shuffling does not preserve the confounding present in the data set and thereby invalidates the use of the distribution of permuted statistics for inference when confounding exists.

To implement our approach, we first model the odds of disease given confounding covariables (typically by using logistic regression). We then use Fisher's noncentral hypergeometric distribution[24] to resample disease status such that the odds of a subject being selected as a case are equal to his or her odds of disease conditional on confounder variables (which we previously defined as the stratification score in Epstein et al.[25]). The sampling is carried out with the open source R package BiasedUrn.[26,27] Using simulated sequence data, we apply our permutation strategy to three existing rare-variant tests (RBT, CMAT, and C-alpha) and show that it corrects for confounding due to population stratification that would otherwise inflate the size of these rare-variant statistics. In addition, we show that even the standard single-locus tests commonly used for analyzing GWASs might benefit from our approach when the MAF is small enough that only a few risk alleles are observed in the population. We also illustrate the approach with an application to sequence data from the Dallas Heart Study of energy metabolism traits.[28,29]

Table 1. Sampling a Permuted Data Set

	$j = 1$	$j = 2$	…	$j = N$	Total
Case	r_{k1}	r_{k2}	…	r_{kN}	N_1
Control	$1 - r_{k1}$	$1 - r_{k2}$	…	$1 - r_{kN}$	N_0
Total	1	1	…	1	

Material and Methods

We assume a case-control study with N_1 case participants and N_0 control participants and let $N = N_0 + N_1$. For $j = 1, ..., N$, we let D_j indicate the j^{th} study participant's disease status for which $D_j = 1$ represents a case participant and $D_j = 0$ represents a control participant. For the j^{th} study participant, we let $\boldsymbol{C}_j$ be the covariate vector that we will adjust for when evaluating the relationship between rare variants and disease outcome by using an appropriate test statistic. The vector $\boldsymbol{C}_j$ can include summary measures of ancestry, such as eigenvectors from principal-component analysis[30,31] or spectral-graph analysis[32] of GWAS SNP data. Additionally, $\boldsymbol{C}_j$ can include other potential confounders, such as age, smoking status, and body mass index (BMI).

We use permutation to establish the significance of an observed association between genetic variants and disease status. If confounding exists, then certain subjects will have greater odds of being a case than will other subjects even after differences in causal risk factors are accounted for. Therefore, we propose sampling a permuted data set in such a way that the odds of a subject being selected as a case are equal to his or her odds of disease conditional on confounder variables. For permutation k, we let $\boldsymbol{r}_k = (r_{k1}, r_{k2}, ..., r_{kN})^{\mathrm{T}}$ be the N-dimensional vector whose j^{th} component is 1 if the j^{th} study participant is selected as a case and 0 if the participant is selected as a control. On the basis of the study design, each $\boldsymbol{r}_k$ will have N_1 components valued at 1 and N_0 components valued at 0. As shown in Table 1, $\boldsymbol{r}_k$ corresponds to the vector of cell occupation counts for the first row of the table given that all row and column marginal totals are fixed. Thus, the distribution of $\boldsymbol{r}_k$ is governed by a multivariate hypergeometric distribution.[24] Therefore, we sample $\boldsymbol{r}_k$ by using Fisher's noncentral hypergeometric distribution with noncentrality parameter $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_N)$, where $\widehat{\theta}_j$ is subject j's estimated odds of disease conditional on confounder variables $\boldsymbol{C}_j$ (previously defined as the stratification score[25]).

A key aspect of our approach is that the confounding role of covariates is maintained in each permuted data set even though the association between risk genotypes (or, more generally, exposure) and disease is broken. Recall that a covariate is a confounder if it influences both genotype and disease. Because we reassign only disease status, any relationship between confounding covariates and genotype found in the original data is maintained in each permutation data set. Furthermore, sampling from Table 1 with Fisher's noncentral hypergeometric distribution with each participant's odds of disease θ_j given covariates $\boldsymbol{C}_j$ maintains the relationship between confounding covariates and disease status. In particular, the odds that individual j is chosen to be a case is θ_j. However, any association between genotype and disease is broken because genotype is not considered in the calculation of the stratification score θ_j or in the resampling of disease status.

We estimate $\widehat{\theta}_j$ with the logistic regression model

$$\log\left(\frac{P[D_j = 1 \mid \boldsymbol{C}_j]}{P[D_j = 0 \mid \boldsymbol{C}_j]}\right) \equiv \log(\theta_j) = \alpha + \gamma^{\mathrm{T}} \cdot \boldsymbol{C}_j, \qquad \text{(Equation 1)}$$

where α is an intercept and γ is a vector of disease-risk parameters corresponding to the elements in $\boldsymbol{C}_j$. We use the maximum-likelihood estimates of these parameters to construct the estimated odds of disease $\widehat{\theta}_j$ for subject j as

$$\widehat{\theta}_j = \exp\left(\widehat{\alpha} + \widehat{\gamma}^{\mathrm{T}} \cdot \boldsymbol{C}_j\right). \qquad \text{(Equation 2)}$$

Using $\widehat{\theta}_j$ in Equation 2, we construct the probability mass function of Fisher's noncentral hypergeometric distribution as

$$f\left(\boldsymbol{r}_k; \widehat{\boldsymbol{\theta}}, N_1\right) = \frac{g\left(\boldsymbol{r}_k; \widehat{\boldsymbol{\theta}}\right)}{\sum\limits_{\boldsymbol{s} \in \Xi} g\left(\boldsymbol{s}; \widehat{\boldsymbol{\theta}}\right)}, \qquad \text{(Equation 3)}$$

where $g(\boldsymbol{r}_k; \widehat{\boldsymbol{\theta}}) = \prod_{j=1}^{N} \widehat{\theta}_j^{r_{kj}}$ and Ξ denotes the set of all possible $\boldsymbol{r}_k$ configurations consistent with Table 1. We observe that $f(\boldsymbol{r}_k; \widehat{\boldsymbol{\theta}}, N_1)$ in Equation 3 does not depend on $\widehat{\alpha}$ in Equation 2 because the intercept cancels from numerator and denominator.

We point out that, within $\boldsymbol{C}_j$, specific covariates that do not predict disease status should have little impact on $\widehat{\theta}_j$ because their disease-risk parameters γ should be small. Note that use of the estimated odds (Equation 2) in place of the true odds is justified given that our method can be considered as a type of parametric bootstrap. Specifically, our sampling procedure can be thought of as a bootstrap in which we prospectively assign each participant a disease outcome on the basis of the logistic model (Equation 1) but then reject all data sets that do not contain N_1 case and N_0 control participants.

Up until now, we have avoided specification of an alternate hypothesis so that the replicate data sets we generate can be used for any hypothesis test provided that the composite null hypothesis (Equation 1) is correctly specified. If we are willing to assume a parametric alternative hypothesis, we can exploit the connection between our resampling approach and the parametric bootstrap to generate replicate data sets under a specified alternative hypothesis. For example, we can assume that

$$\log\left(\frac{P[D_j = 1 \mid G_j, \boldsymbol{C}_j]}{P[D_j = 0 \mid G_j, \boldsymbol{C}_j]}\right) \equiv \log(\theta_j) = \alpha + \beta \cdot G_j + \gamma^{\mathrm{T}} \cdot \boldsymbol{C}_j, \qquad \text{(Equation 4)}$$

where G_j counts the number of minor alleles found at a risk locus. Although standard asymptotic methods can be used for making inference about β, we can question these methods when the MAF of the variant is less common (<5%).

To generate a resampling-based confidence interval for $(\widehat{\beta} - \beta)$ under the alternative hypothesis (Equation 4), we first estimate coefficients α, γ, and β. We then use these estimates in Equation 4 to estimate $\widehat{\theta}_j$ and then use these estimates of θ_j to generate replicates as described previously. We then fit Equation 4 to each replicate data set and collect $\widehat{\beta}_r$, the β estimate obtained from the r^{th} replicate. Because these replicates correspond to a parametric bootstrap sample in which all data sets that do not have N_1 case and N_0 control participants are rejected, we can base inference on the observed distribution of the $\widehat{\beta}_r$ values. For example, we can construct a confidence interval for $\widehat{\beta}$ by using the quantiles of the resampling distribution of $\widehat{\beta}_r$. Other bootstrap-based confidence regions described in Efron and Tibshirani[33] can also be calculated.

Software

We generated random variates from Fisher's noncentral hypergeometric distribution by using the R package BiasedUrn created by Fog.[26] The name "Biased Urn" refers to a related use of Fisher's multivariate hypergeometric distribution in an urn-model problem in which each of N balls has a specified odds of being selected and in which we sample N_1 of these balls without replacement. As distributed in the Comprehensive R Archive Network (CRAN), the BiasedUrn package has set the maximum number of columns in Table 1 to 32. So that the package is amenable for the sample sizes expected from resequencing studies, the package is recompiled so that it can generate permutation data sets for any case-control study in which $N \leq 10{,}000$. On our website, we provide instructions on how to recompile and install this package with the increased value for N (see Web Resources). We also provide sample R code implementing the approach in Appendix A. Because we use the R package BiasedUrn for calculations involving Fisher's noncentral hypergeometric distribution, we refer to our sampling procedure as biased urn sampling throughout the remainder of the paper.

Results

Biased Urn Sampling Preserves Structure in the Original Data Set

We first performed a proof-of-principle simulation on the basis of an existing GWAS to ensure that our proposed biased urn sampling strategy with Fisher's noncentral hypergeometric distribution would preserve the structure present within a data set (such structure is not preserved with random [naïve] permutations.) We used data from a case-control GWAS of African American subjects with schizophrenia (these data are available for download from the database of Genotypes and Phenotypes [dbGaP][34] [see Web Resources and Acknowledgments]). The GWAS data set initially consisted of data from 921 case and 954 control participants genotyped for 845,814 SNPs on the Affymetrix 6.0 platform. After we used the PLINK software package[35] to implement quality-control procedures similar to those described in Fellay et al.,[36] our final sample consisted of data from 907 case and 937 control participants genotyped for 808,169 SNPs. We then used a reduced set of 41,182 SNPs in approximate linkage equilibrium (pairwise $r^2 \leq 0.04$ as determined by PLINK) to infer eigenvectors from principal-component analysis;[31] these eigenvectors serve as summary measures of ancestry. On the basis of the principal-component analysis, we used Tracy-Widom statistics to identify eight eigenvectors that were significant at level $\alpha = 0.01$.

We fit the logistic regression model (Equation 1) to our sample of 907 cases and 937 controls, and we let $\boldsymbol{C}_j$ represent the vector of the eight significant eigenvectors for study participant j. The maximum-likelihood estimates of $\widehat{\gamma} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_8)$ are shown in the second column of Table 2. We next examined whether permuted data sets generated with our biased urn sampling procedure maintained the structure found within the original case-control data set. We generated a permuted data set by using our

Table 2. Regression Coefficient Estimates Under Biased Urn and Random Permutation Schemes

		Permutation Scheme	
		Biased Urn	Random
	Original Data	Mean (SD)	Mean (SD)
γ_1	−8.39	−8.47 (2.02)	−0.06 (2.03)
γ_2	1.41	1.44 (2.14)	−0.08 (2.03)
γ_3	−2.13	−2.28 (2.00)	−0.11 (2.04)
γ_4	−4.86	−4.96 (2.05)	−0.09 (2.03)
γ_5	−0.88	−0.93 (2.02)	−0.08 (2.02)
γ_6	0.69	0.80 (2.09)	0.01 (2.03)
γ_7	−1.22	−1.24 (2.01)	0.00 (2.04)
γ_8	−0.76	−0.80 (1.99)	0.03 (1.96)

The results for each permutation scheme are based on 1,000 permutations of the data set. The following abbreviation is used: SD, standard deviation.

Table 3. Type-I Error Results Under Confounding for 10 kb Regions

Test	Odds Ratio of Disease (YRI versus CEU)	$\alpha = 0.05$		$\alpha = 0.005$	
		Biased Urn	Random	Biased Urn	Random
CMAT	1	0.0521	0.0511	0.0046	0.0045
	2	0.0450	0.0850	0.0047	0.0123
	4	0.0485	0.1607	0.0053	0.0503
	8	0.0551	0.2366	0.0058	0.1004
RBT	1	0.0469	0.0468	0.0043	0.0042
	2	0.0487	0.0591	0.0045	0.0066
	4	0.0501	0.0962	0.0055	0.0169
	8	0.0546	0.1994	0.0055	0.0463
C-alpha	1	0.0491	0.0542	0.0043	0.0051
	2	0.0460	0.1712	0.0049	0.0364
	4	0.0453	0.4890	0.0042	0.2251
	8	0.0527	0.7603	0.0055	0.5011

These results are based on 10,000 replicates each assuming 300 cases and 300 controls. The significance of each replicate was established with 5,000 permutations. The following abbreviations are used: YRI, Yoruba in Ibadan, Nigeria; CEU, Utah residents with ancestry from northern and western Europe from the CEPH collection; CMAT, cumulative minor-allele test; and RBT, replication-based test.

biased urn sampling procedure and then refit the logistic regression model (Equation 1) to the new data set to obtain a new estimate of γ. We repeated this process 1,000 times and recorded the mean value of γ across the permuted data sets in the third column of Table 2. The results clearly show that permuted data sets generated with our biased urn procedure maintain the same population structure found within the original data set; however, when we repeated the same analysis by using the standard approach of randomly permuting the disease status without regard for confounding, we saw that the structure present in the case-control sample was not preserved in the permuted data sets (see fourth column of Table 2). This confirms that the use of random permutations for assessing the significance of rare-variant association tests like RBT and WHaIT, which ignore confounders, could lead to erroneous inference if the observed association is due to confounding in the original case-control sample.

Simulations to Assess Validity of Rare-Variant Tests in the Presence of Confounding

We compared the performance of the biased urn sampling procedure with that of random permutations on different rare-variant tests by using simulated resequencing data sets that were subjected to confounding arising from population stratification. We used the coalescent simulator cosi[37] to produce large sets of haplotypes (ranging from 10 kb to 100 kb) whose variation patterns mimicked those observed in HapMap YRI (Yoruba in Ibadan, Nigeria) and CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) samples. Randomly pairing haplotypes to form diplotypes, we prospectively generated case-control data sets in which a subject's odds of disease were a function of the average percentage of African ancestry across the region, and we thereby induced confounding. We assumed an overall disease prevalence of 0.01. We considered both discrete population models that treated the YRI and CEU haplotypes as separate groups and admixture models that created haplotypes that were a mixture of YRI and CEU ancestry. We generated admixed haplotypes by using the model of Price et al.;[38] for a given haplotype, this model assumes crossover events occurring after initial admixture by using an exponential distribution with parameter λ (corresponding to the number of generations since initial admixture; we assumed this value to be 6). These crossover events divide the haplotype into distinct segments. With probability τ, we filled a segment with the corresponding segment from a European haplotype; otherwise, we filled it with the corresponding segment from an African haplotype. τ was sampled from a Beta(3,12) distribution.[39]

For each simulated data set, we applied the CMAT,[6] RBT,[7] and C-alpha[8] test to test for association between disease and the rare variants (defined as those variants with a sample MAF threshold $<$ 5%) within the region. For each test, we first established significance by using 5,000 random permutations that did not adjust for confounding due to population stratification. Next, we established the significance of the test by using 5,000 data sets generated with biased urn sampling that adjusted for this confounding. To do this, we simulated genotype data for each subject on at least 10,000 SNPs that were selected from HapMap and showed marked allele-frequency differences between HapMap YRI and CEU samples. We then constructed principal components[30,31] for each subject on the basis of the SNP data and used them to construct the

Table 4. Type-I Error Results Under Confounding for 100 kb Regions

Test	Odds Ratio of Disease (YRI versus CEU)	$\alpha = 0.05$		$\alpha = 0.005$	
		Biased Urn	Random	Biased Urn	Random
CMAT	1	0.0466	0.0482	0.0045	0.0048
	2	0.0474	0.1040	0.0048	0.0192
	4	0.0480	0.2308	0.0050	0.0868
	8	0.0544	0.3035	0.0052	0.1439
RBT	1	0.0477	0.0497	0.0048	0.0053
	2	0.0445	0.0691	0.0046	0.0080
	4	0.0461	0.1463	0.0042	0.0308
	8	0.0515	0.3986	0.0057	0.1406
C-alpha	1	0.0440	0.0501	0.0044	0.0049
	2	0.0402	0.2834	0.0040	0.0727
	4	0.0410	0.7962	0.0050	0.5049
	8	0.0422	0.9771	0.0038	0.8729

These results are based on 10,000 replicates each assuming 300 cases and 300 controls. The significance of each replicate was established with 5,000 permutations. The following abbreviations are used: YRI, Yoruba in Ibadan, Nigeria; CEU, Utah residents with ancestry from northern and western Europe from the CEPH collection; CMAT, cumulative minor-allele test; and RBT, replication-based test.

Table 5. Power Results for 10 kb Regions

Test	Permutation Scheme	Relative Risk of Rare Variant		
		1.5	2.0	2.5
CMAT	Biased urn	0.135	0.244	0.290
	Random	0.141	0.241	0.289
RBT	Biased urn	0.144	0.263	0.373
	Random	0.144	0.273	0.383
C-alpha	Biased urn	0.267	0.552	0.735
	Random	0.279	0.572	0.754

Results assume 300 cases and 300 controls. Results were evaluated at $\alpha = 0.05$ and are based on 1,000 replicates. The significance of each replicate was established with 5,000 permutations. Simulations assumed no confounding due to population stratification. Ten percent of rare variants (MAF $\leq$ 1%) were assumed to be causal in the region. Each causal rare variant was assumed to have an identical relative-risk value. For the RBT, we tested a one-sided hypothesis of excess rare variants in cases compared to controls. The following abbreviations are used: CMAT, cumulative minor-allele test; and RBT, replication-based test.

stratification score in Equation 2; we then used this stratification score within the biased urn procedure in Equation 3. We also investigated similar sampling based on known ancestry.

Table 3 provides empirical type-I error rates for the CMAT, RBT, and C-alpha test for 10 kb regions on the basis of biased urn and random permutation procedures, whereas Table 4 provides such rates for 100 kb regions. For each table, the results show that biased urn sampling and random permutations both maintain the appropriate significance level when no confounding exists within the simulated data sets (the odds ratio of African to European ancestry is 1). However, when we induce confounding in the simulated data sets (when the odds ratio of African to European ancestry > 1), we see that biased urn sampling maintains appropriate type-I error, whereas random permutations yield inflated size. This inflation increases with the degree of confounding. These results are based on simulations under discrete population models; we see similar findings for admixture models (results not shown).

To ensure that biased urn sampling's preservation of type-I error under confounding did not reduce power, we performed a simulation under an alternative model, in which we assumed stratification but no confounding due to stratification in samples (by assuming the odds ratio of African to European ancestry was 1). Within a simulated region, we assumed that 10% of variants with a MAF < 0.01 were causal and that each variant independently increased disease risk under a log-additive model; we assumed that the relative risk of each causal variant was identical. Table 5 provides power results for biased urn sampling and random permutations for different values of relative risk. These results show that the power of biased urn and random permutations are quite similar in these situations, suggesting that biased urn will yield results analogous to random permutations when confounding is absent, whereas the procedure has appropriate control of size when confounding is present.

Resampling-Based Confidence Intervals of Variant Risk Estimates

For less-common variants, we examined the confidence intervals of risk estimates based on our biased urn sampling under the alternative hypothesis specified in Equation 4 and compared such intervals to those derived on the basis of asymptotic theory. Assuming a disease prevalence of 0.01, we prospectively generated 5,000 data sets comprising 300 cases and 300 controls as described previously. We assumed a risk variant with a MAF of 0.02 (and an effect size of $\beta = 1$ on the log-disease-odds scale) and further induced confounding by letting the disease odds ratio of each African chromosome be 4. For each data set, we calculated a 95% resampling-based confidence interval for the variant risk estimate by using 10,000 biased urn replicates generated from Equation 3 on the basis of the model in Equation 4. We calculated resampling-based confidence intervals various ways, including by using the quantiles of the resampled estimates and a bias-corrected calculation.[40] We also calculated an asymptotic 95% confidence interval for the estimate of the variant effect on the basis of a standard logistic-regression model adjusting for the effect of the confounding.

Our simulations revealed that the resampling-based confidence intervals had appropriate coverage and were smaller in magnitude than the corresponding asymptotic interval. We observed that our 95% resampling-based

Table 6. CMAT Analysis of Sequence Data from the Dallas Heart Study

		p Value of CMAT	
Trait	Gene	Random Permutations	Biased Urn Permutations
Triglycerides	*ANGPTL3*	<0.0001	0.0141
	ANGPTL4	<0.0001	0.0015
	ANGPTL5	0.0201	0.0974
BMI	*ANGPTL3*	0.5930	0.7418
	ANGPTL4	0.6984	0.7058
	ANGPTL5	0.0077	0.0301

Biased urn permutations are corrected for effects of age, gender, and race. Analysis is based only on nonsynonymous variants in each gene. Each p value is based on 10,000 permutations. The following abbreviations are used: CMAT, cumulative minor-allele test; and BMI, body mass index.

Table 7. RBT Analysis of Sequence Data from the Dallas Heart Study

		p Value of RBT	
Trait	Gene	Random Permutations	Biased Urn Permutations
Triglycerides	*ANGPTL3*	0.0006	0.0126
	ANGPTL4	<0.0001	0.0034
	ANGPTL5	0.0231	0.1102
BMI	*ANGPTL3*	0.5174	0.6890
	ANGPTL4	0.9180	0.9348
	ANGPTL5	0.0046	0.0170

Biased urn permutations are corrected for effects of age, gender, and race. Analysis is based only on nonsynonymous variants in each gene. Each p value is based on 10,000 permutations. The following abbreviations are used: RBT, replication-based test; and BMI, body mass index.

confidence interval calculated with quantiles had appropriate coverage of 0.949. The coverage of the asymptotic 95% confidence interval was also appropriate (0.953); however, the resampling-based confidence intervals were somewhat shorter than the asymptotic intervals and shifted away from the null. We found that the mean 95% resampling-based confidence interval calculated with quantiles was (0.122, 2.120), whereas the corresponding asymptotic confidence interval was wider at (−0.029, 2.087) and further contained the null value of 0. We observed similar trends for 99% confidence intervals (results not shown).

Application to the Dallas Heart Study

The Dallas Heart Study is a multiethnic population-based study that previously examined the relationship between sequence variation within *ANGPTL3* (MIM 604774), *ANGPTL4* (MIM 605910), and *ANGPTL5* (MIM 607666) and various quantitative metabolism-related traits.[28,29] Coding regions of these three genes were sequenced in a group of 3,476 subjects (1,830 African Americans, 1,045 European Americans, and 601 Hispanics). In this application, we studied two metabolic outcomes: triglyceride levels and BMI. Prior to analysis, we first removed data from 216 subjects who were being treated with statins. Then, for each outcome, we selected subjects in the top and bottom 20% of the outcome distribution (after removing subjects with missing outcomes) to mimic a case-control study design. To study triglycerides, we obtained 570 case and 570 control participants. To study BMI, we obtained 563 case and 563 control participants.

We applied the CMAT[6] to test for association between rare nonsynonymous (NS) variants in *ANGPTL3*, *ANGPTL4*, and *ANGPTL5* and our case-control representations of triglycerides and BMI. Within the triglyceride sample, we found 36 NS variants in *ANGPTL3*, 39 in *ANGPTL4*, and 27 in *ANGPTL5*. Within the BMI sample, we saw 36 NS variants in *ANGPTL3*, 38 in *ANGPTL4*, and 27 in *ANGPTL5*. For each CMAT statistic, we established significance by using both random permutations and biased urn sampling that adjusted for the effects of age, gender, and race. The results of these analyses are shown in Table 6. The results clearly show that the CMAT p values based on random permutations are smaller than their corresponding p values based on biased urn sampling adjusting for confounders. Notably, failing to adjust for confounders appears to lead to a spurious association between rare NS variants in *ANGPTL5* and triglyceride levels. Subsequent investigation revealed that, as expected, race was a confounder because it was associated with both case-control status ($p < 0.0001$) and the presence of rare NS variants in *ANGPTL5* ($p = 0.0002$). We also repeated the analyses by using a two-sided version of the RBT[7] and observed similar findings and trends (Table 7). We also applied the C-alpha[8] test and observed similar trends with the exception that there was no evidence of association between rare NS variants in *ANGPTL5* and either BMI or triglyceride levels with the use of either random permutations or biased urn sampling (Table 8).

Discussion

In this article, we propose a simple biased urn sampling procedure (based on the use of Fisher's noncentral hypergeometric distribution) that resamples subjects under the null hypothesis of no association in a way that preserves the confounding present in the actual data set. This procedure is particularly valuable for rare-variant association tests, many of which are not easily adjusted for probable confounders like population stratification and whose applicability is thus limited in case-control resequencing studies. With a combination of simulated and real data sets, we have illustrated how our approach corrects for confounding in three common rare-variant association tests. In addition, we have shown how resampling-based

Table 8. C-Alpha Analysis of Sequence Data from the Dallas Heart Study

		p Value of C-Alpha Test	
Trait	**Gene**	**Random Permutations**	**Biased Urn Permutations**
Triglycerides	*ANGPTL3*	<0.0001	0.0010
	ANGPTL4	0.0001	0.0363
	ANGPTL5	0.1572	0.2043
BMI	*ANGPTL3*	0.9168	0.9814
	ANGPTL4	0.8314	0.8472
	ANGPTL5	0.2310	0.2872

Biased urn permutations are corrected for effects of age, gender, and race. Analysis is based only on nonsynonymous variants in each gene. Each p value is based on 10,000 permutations. The following abbreviation is used: BMI, body mass index.

confidence intervals of risk estimates for individual susceptibility variants can be calculated when a parametric alternative hypothesis is specified.

Our procedure adjusts rare-variant association testing for confounders by using permutation, which is a common resampling procedure used for statistical inference. Another resampling procedure called the parametric bootstrap[41] has been proposed by Lin and Tang[13] for adjusting logistic-regression-based rare-variant association tests for the effects of covariates. The parametric bootstrap of Lin and Tang creates replicate data sets from a prospective model in which the disease outcome of each subject in a data set is generated on the basis of the subject's probability of disease conditional on covariates (which we write as $\widehat{\theta}_j/(1+\widehat{\theta}_j)$ by using the notation in Equation 2). For logistic regression, it is known that a retrospective analysis of such prospectively generated data can give the same results (except for the intercept, which is typically not of interest). Furthermore, it is also known that logistic regression is indifferent to whether row and/or column totals in Table 1 are held fixed. However, because the parametric bootstrap does not preserve the number of cases and controls within each generated data set, it is unclear whether it can be applied to tests that are not based on logistic regression. Our biased urn procedure, on the other hand, possesses the useful feature that it preserves the number of cases and controls within each sample by design and so corresponds to retrospective sampling. For this reason, it can be applied to any test that is appropriate for case-control data as long as a valid model for the stratification score is used. Such preservation of case-control numbers in replicate data sets is particularly valuable for exome-sequencing studies of Mendelian traits, studies which often possess only a handful of cases for analysis.[19] Finally, we note that we could apply the parametric bootstrap in such a manner that we only accept data sets that preserve the original number of cases and controls, but such a procedure will be much less computationally efficient than biased urn sampling. In our simulations, we observed that this approach required ~25× more computation time than biased urn sampling across different sample sizes.

Our biased urn sampling is implemented in a recompiled version of the R BiasedUrn package. Appendix A provides sample code for applying the approach. The computation time required for generating permuted data sets depends on sample size but is reasonable even for studies composed of thousands of participants. On a single 3.20 GHz Intel Xeon central processing unit running Windows XP on a Dell PowerEdge 2950 server with 2 GB of random-access memory, the generation of 10,000 permuted data sets for sample sizes of 1,000, 5,000, and 10,000 required ~30 s, ~8 min, and ~30 min of computation time, respectively. Furthermore, the process can be implemented in parallel for the reduction of computation time.

Appendix A: Sample R Code for Implementing Biased Urn Sampling Procedure

```
library('BiasedUrn') #load (modified) package
# Assume one has already scanned in required data set.
# dis: array of disease outcomes (1 affected, 0 unaffected) for N subjects
# z: covariate matrix of dimension N × C
n.case < - sum(dis) # number of cases
n.perm < - 1000 # number of permutations
# step 1: fit logistic-regression model in Equation 1
model < - glm (dis ~z, family= binomial())
# step 2: construct estimated disease odds in Equation 2
d.odds < - exp (model$linear.predictors)
# step 3: generate N x n.perm matrix of permuted data sets
m1 < - c(rep(1, length(dis)))
perm.hg < - rMFNCHypergeo(n.perm, m1, n.case, d.odds)
```

Acknowledgments

Any opinions expressed in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. This work was supported by National Institutes of Health grants HG003618 (to M.P.E and R.D.) and MH084680 (to A.S.A). We thank Jonathan Cohen for providing us with sequence data from the Dallas Heart Study. We thank Michael Boehnke and three anonymous reviewers for their valuable comments that improved this article. We thank Cheryl Strauss for assistance with technical editing of the article. Funding support for "Genome-Wide Association Study of Schizophrenia" was provided by the National Institutes of Health, and the genotyping of samples was a courtesy of the Genetic Association Information Network. The data set used for the analyses described in this manuscript was obtained from the database of Genotype and Phenotype (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000021.v2.p1. Samples and associated phenotype data for "Genome-Wide Association Study of Schizophrenia" were provided by Pablo V. Gejman, MD.

Received: February 25, 2012
Revised: May 3, 2012
Accepted: June 5, 2012
Published online: July 19, 2012

Web Resources

The URLs for data presented herein are as follows:

BiasedUrn package in CRAN, http://cran.r-project.org/web/packages/BiasedUrn/index.html

cosi, http://www.broadinstitute.org/~sfs/cosi/

dbGaP, http://view.ncbi.nlm.nih.gov/dbgap

Epstein Software, http://www.genetics.emory.edu/labs/epstein/software

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org

PLINK Version 1.07, http://pngu.mgh.harvard.edu/purcell/plink/

References

1. Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutat. Res. *615*, 28–56.
2. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
3. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
4. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
5. Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. *34*, 188–193.
6. Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., and Zöllner, S. (2010). Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes. Am. J. Hum. Genet. *87*, 604–617.
7. Ionita-Laza, I., Buxbaum, J.D., Laird, N.M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. *7*, e1001289.
8. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.
9. Li, Y., Byrnes, A.E., and Li, M. (2010). To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. Am. J. Hum. Genet. *87*, 728–735.
10. Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. Hum. Hered. *70*, 42–54.
11. Hoffmann, T.J., Marini, N.J., and Witte, J.S. (2010). Comprehensive approach to analyzing rare genetic variants. PLoS ONE *5*, e13584.
12. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
13. Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. Am. J. Hum. Genet. *89*, 354–367.
14. Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. *11*, 773–785.
15. 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
16. Zelterman, D., and Chen, C.-F. (1988). Homogeneity tests against central-mixture alternatives. J. Am. Stat. Assoc. *83*, 179–182.
17. Sul, J.H., Han, B., He, D., and Eskin, E. (2011). An optimal weighted aggregated association test for identification of rare variants involved in common diseases. Genetics *188*, 181–188.
18. Kinnamon, D.D., Hershberger, R.E., and Martin, E.R. (2012). Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. PLoS ONE *7*, e30238.
19. Ionita-Laza, I., Makarov, V., Yoon, S., Raby, B., Buxbaum, J., Nicolae, D.L., and Lin, X. (2011). Finding disease variants in Mendelian disorders by using sequence data: methods and applications. Am. J. Hum. Genet. *89*, 701–712.
20. Ionita-Laza, I., Makarov, V., and Buxbaum, J.D.; the ARRA Autism Sequencing Consortium. (2012). Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three data sets. Am. J. Hum. Genet. *90*, 1002–1013.
21. Kimmel, G., Jordan, M.I., Halperin, E., Shamir, R., and Karp, R.M. (2007). A randomization test for controlling population stratification in whole-genome association studies. Am. J. Hum. Genet. *81*, 895–905.
22. Allen, A., Epstein, M.P., and Satten, G.A. (2010). Score-based adjustment for confounding by population stratification in genetic association studies. Genet. Epidemiol. *34*, 383–385.
23. Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A comparison of association methods correcting for population stratification in case-control studies. Ann. Hum. Genet. *75*, 418–427.
24. Harkness, W.L. (1965). Properties of the extended hypergeometric distribution. The Annals of Mathematical Statistics *36*, 938–945.
25. Epstein, M.P., Allen, A.S., and Satten, G.A. (2007). A simple and improved correction for population stratification in case-control studies. Am. J. Hum. Genet. *80*, 921–930.
26. Fog, A. (2011). BiasedUrn: Biased Urn model distributions. R package version 1.04 (http://cran.r-project.org/web/packages/BiasedUrn/index.html).
27. Fog, A. (2008). Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. Communications in Statistics–Simulation and Computation *37*, 241–257.
28. Victor, R.G., Haley, R.W., Willett, D.L., Peshock, R.M., Vaeth, P.C., Leonard, D., Basit, M., Cooper, R.S., Iannacchione, V.G., Visscher, W.A., et al; Dallas Heart Study Investigators. (2004). The Dallas Heart Study: A population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am. J. Cardiol. *93*, 1473–1480.
29. Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L.A., Boerwinkle, E., Hobbs, H.H., and Cohen, J.C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J. Clin. Invest. *119*, 70–79.

30. Chen, H.-S., Zhu, X., Zhao, H., and Zhang, S. (2003). Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann. Hum. Genet. *67*, 250–264.
31. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
32. Lee, A.B., Luca, D., Klei, L., Devlin, B., and Roeder, K. (2010). Discovering genetic ancestry using spectral graph theory. Genet. Epidemiol. *34*, 51–59.
33. Efron, B., and Tibshirani, R. (1993). An Introduction to the Bootstrap (New York, London: Chapman and Hall).
34. Manolio, T.A., Rodriguez, L.L., Brooks, L., Abecasis, G., Ballinger, D., Daly, M., Donnelly, P., Faraone, S.V., Frazer, K., Gabriel, S., et al; GAIN Collaborative Research Group; Collaborative Association Study of Psoriasis; International Multi-Center ADHD Genetics Project; Molecular Genetics of Schizophrenia Collaboration; Bipolar Genome Study; Major Depression Stage 1 Genomewide Association in Population-Based Samples Study; Genetics of Kidneys in Diabetes (GoKinD) Study. (2007). New models of collaboration in genome-wide association studies: The Genetic Association Information Network. Nat. Genet. *39*, 1045–1051.
35. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
36. Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. Science *317*, 944–947.
37. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. *15*, 1576–1583.
38. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.
39. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska,, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., et al. (2004). A high-density admixture map for disease gene discovery in african americans. Am. J. Hum. Genet. *74*, 1001–1013.
40. Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. Biometrika *72*, 45–58.
41. Davison, A.C., and Hinkley, D.V. (1997). Bootstrap Methods and Their Application (Cambridge: Cambridge University Press).