

# A Statistical Approach for Rare-Variant Association Testing in Affected Sibships

Michael P. Epstein,<sup>1,\*</sup> Richard Duncan,<sup>1</sup> Erin B. Ware,<sup>2</sup> Min A. Jhun,<sup>2</sup> Lawrence F. Bielak,<sup>2</sup> Wei Zhao,<sup>2</sup> Jennifer A. Smith,<sup>2</sup> Patricia A. Peyser,<sup>2</sup> Sharon L.R. Kardia,<sup>2</sup> and Glen A. Satten<sup>3</sup>

Sequencing and exome-chip technologies have motivated development of novel statistical tests to identify rare genetic variation that influences complex diseases. Although many rare-variant association tests exist for case-control or cross-sectional studies, far fewer methods exist for testing association in families. This is unfortunate, because cosegregation of rare variation and disease status in families can amplify association signals for rare variants. Many researchers have begun sequencing (or genotyping via exome chips) familial samples that were either recently collected or previously collected for linkage studies. Because many linkage studies of complex diseases sampled affected sibships, we propose a strategy for association testing of rare variants for use in this study design. The logic behind our approach is that rare susceptibility variants should be found more often on regions shared identical by descent by affected sibling pairs than on regions not shared identical by descent. We propose both burden and variance-component tests of rare variation that are applicable to affected sibships of arbitrary size and that do not require genotype information from unaffected siblings or independent controls. Our approaches are robust to population stratification and produce analytic p values, thereby enabling our approach to scale easily to genome-wide studies of rare variation. We illustrate our methods by using simulated data and exome chip data from sibships ascertained for hypertension collected as part of the Genetic Epidemiology Network of Arteriopathy (GENOA) study.

## Introduction

Many recent genetic studies of complex human traits have been designed to determine the effects of rare or less-common genetic variants on phenotype. These studies have used either next-generation sequencing technology (which enables complete interrogation of single-nucleotide variants throughout the human genome) or cheaper alternatives like the Illumina HumanExome Beadchip, which provides genotypes for nearly 250,000 markers, primarily protein-altering variants, that were identified from large-scale resequencing efforts (details on the Beadchip design available online; see [Web Resources](#)). These new technologies have spurred the development of powerful new statistical approaches for rare-variant association testing of complex traits. However, the majority of such tests are tailored for use in either case-control or population-based studies that assume unrelated subjects. Fewer rare-variant association tests exist for family-based studies, which is unfortunate because families are uniquely suited for studying rare variants.

Family-based studies of rare variants possess many valuable attributes that population studies lack. Rare susceptibility variants are likely to be enriched in affected relatives and therefore be easier to detect relative to population-based counterparts.<sup>1,2</sup> Family-based tests of rare-variant association can also be made robust to population stratification,<sup>3</sup> which is important given recent work suggesting that existing common-variant-based methods to correct for stratification might not be as effective in studies of rare variants.<sup>4,5</sup> Several projects have initiated collection

of families for rare-variant analysis, including studies of epilepsy<sup>6</sup> and simplex autism.<sup>7</sup> Additionally, other projects have begun analysis of rare variants in familial samples collected from past linkage studies. One such project is the Genetic Epidemiology Network of Arteriopathy (GENOA) study,<sup>8</sup> which has genotyped a large collection of African American sibships with hypertension for the Illumina HumanExome Beadchip. Other linkage studies were funded for rare-variant analysis via a NIH funding opportunity titled “Life After Linkage: The Future of Family Studies” (RFA-HL-12-007) that was sponsored by the National Heart, Lung, and Blood Institute.

To fully capitalize on the opportunities presented by sequencing family members originally recruited into linkage studies, it is necessary to develop new statistical tools for rare-variant analysis in such projects. Although several rare-variant methods exist for family-based analysis, some are entirely restricted to quantitative traits.<sup>9–12</sup> Other family-based tests for use in complex diseases<sup>3,13–15</sup> require information from both affected and unaffected individuals for inference. For the many linkage studies that collected only affected sibs, there might not be data from unaffected family members that can be used for inference. The use of historical controls is unappealing because of potential bias from population stratification<sup>4</sup> as well as the need to account for differences in sequencing coverage.<sup>16</sup> Thus, a statistical approach for rare-variant analysis in affected sibships that is self-contained and that does not require information from unaffected subjects is appealing.

We propose a strategy for association testing of rare variants in affected sibships that is motivated by the notion

<sup>1</sup>Department of Human Genetics, Emory University, Atlanta, GA 30322, USA; <sup>2</sup>Department of Epidemiology, University of Michigan, Ann Arbor, MI 48103, USA; <sup>3</sup>Centers for Disease Control and Prevention, Atlanta, GA 30341, USA

\*Correspondence: [mpepste@emory.edu](mailto:mpepste@emory.edu)

<http://dx.doi.org/10.1016/j.ajhg.2015.01.020>. ©2015 by The American Society of Human Genetics. All rights reserved.

that, among affected sibling pairs, we would expect to observe more copies of rare susceptibility variants on haplotypes shared identical by descent by the pair than on haplotypes not shared identical by descent. We can evaluate this hypothesis without knowledge of haplotype phase by using an efficient-score test derived from an estimating-equation model. In this model, we regress the number of copies of rare variants an affected sib pair possesses on the number of loci they share that are identical by descent. Many family-based studies already have identity by descent (IBD) information from earlier linkage studies or GWASs but, if this is not the case, such information can be obtained from sequence data or common SNPs found on the Illumina HumanExome Beadchip. Although our approach uses IBD information for association testing, it is not a test of linkage and, as we will show, can detect rare-variant association signals in regions that do not harbor a significant linkage signal.

With our framework, we derive both “linear” or “burden” tests, which collapse multiple rare variants within a region of interest into a single composite variable and then test for association between that variable and IBD status,<sup>17–21</sup> and “quadratic” or “variance-component” tests that compare pairwise similarity in IBD status to genotypic similarity within the sample.<sup>22,23</sup> Adaptive procedures that combine aspects of both the burden and variance-component classes of tests are also possible<sup>24</sup> but are not considered here. As we will show, our strategy is robust to confounding as a result of population stratification and can handle sibships of arbitrary size. Unlike existing rare-variant tests of complex diseases in pedigrees, our approach is tailored for analysis of datasets comprised entirely of affected sibships. That is, the approach does not require variant information from unaffected relatives, although such information can be used to assist in estimating IBD sharing among affected siblings. The test has a closed form and p values can be obtained analytically, which facilitates genome-wide analysis. In subsequent sections, we derive our rare-variant association test and illustrate the approach by using both simulated sequence data as well as exome-chip data from affected sibships collected by the GENOA study.

## Materials and Methods

### Assumptions and Notation

We assume we possess sequence or exome-chip data from a collection of  $N$  affected sibships of arbitrary size. We assume sibship  $j$  ( $j = 1, \dots, N$ ) is comprised of  $M_j$  affected siblings that can be partitioned into  $S_j = (M_j)(M_j - 1) / 2$  possible pairs. We consider the analysis of a gene or region and let  $R$  denote the number of polymorphic rare-variant sites observed in the gene or region (with a rare variant defined as a variant whose frequency is less than some threshold value, such as 0.01 or 0.05). We define  $T_{ij}$  as the total number of copies of rare variants at the  $R$  sites possessed by the  $i^{\text{th}}$  sib pair in the  $j^{\text{th}}$  sibship. If desired, we can up- or down-weight the contributions of specific rare variants

over others in the  $T_{ij}$  count by applying site-specific weights using information such as minor-allele frequency<sup>19,22</sup> or functional information;<sup>21</sup> in this case,  $T_{ij}$  denotes the weighted sum. We further define  $Z_{ij}$  as the estimated number of copies of the gene/region shared identical by descent by the  $i^{\text{th}}$  sib pair in the  $j^{\text{th}}$  sibship, which we can infer via the Lander-Green<sup>25</sup> algorithm (or an extension that allows for linkage disequilibrium among markers<sup>26</sup>), as implemented in software packages like MERLIN.<sup>27</sup>

### Proposed Model

Our model assumes the mean and variance in the number of copies of rare causal variants on haplotypes inherited by an affected sibling pair differ depending on whether the haplotype is inherited identical by descent or not. We define the mean and variance of the number (or weighted sum) of rare variants on a parental haplotype inherited identical by descent as  $\mu_1$  and  $\sigma_1^2$ , respectively. Similarly, we define  $\mu_0$  and  $\sigma_0^2$  as the mean and variance of the number (or weighted sum) of rare variants on a parental haplotype that is not inherited identical by descent by the affected pair. If a sib pair shares zero haplotypes identical by descent, then the pair possess two independent maternal haplotypes that are not shared identical by descent and two independent paternal haplotypes that are not shared identical by descent. If a sib pair shares one haplotype identical by descent, then the pair possesses two identical haplotypes (shared identical by descent) inherited from one parent and two distinct independent haplotypes (not shared identical by descent) inherited from the other parent. If the pair shares two haplotypes identical by descent, then the pair possesses two identical maternal haplotypes (shared identical by descent) and two identical paternal haplotypes (shared identical by descent). Based on this information, we can derive the mean and variance of  $T_{ij}$  conditional on the IBD sharing  $Z_{ij}$  as

$$E[T_{ij} | Z_{ij}] = \begin{cases} 4\mu_0 & Z_{ij} = 0 \\ 2\mu_0 + 2\mu_1 & Z_{ij} = 1 \\ 4\mu_1 & Z_{ij} = 2 \end{cases} \quad (\text{Equation 1})$$

and

$$\text{Var}[T_{ij} | Z_{ij}] = \begin{cases} 4\sigma_0^2 & Z_{ij} = 0 \\ 2\sigma_0^2 + 4\sigma_1^2 & Z_{ij} = 1 \\ 8\sigma_1^2 & Z_{ij} = 2. \end{cases} \quad (\text{Equation 2})$$

Our primary interest is to test the null hypothesis that  $H_0: \mu_0 = \mu_1$ . To simplify presentation, we reparameterize our model by defining  $\delta = \mu_1 - \mu_0$  so that we can rewrite Equation 1 as

$$E[T_{ij} | Z_{ij}] = 4\mu_0 + 2\delta Z_{ij} \quad (\text{Equation 3})$$

and reconfigure our primary null hypothesis to be  $H_0: \delta = 0$ . Based on Equations 2 and 3, we can perform inference on the parameter  $\delta$  by solving the following estimating equations for  $(\mu_0, \delta)$ :

$$\sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} (T_{ij} - 4\mu_0 - 2\delta Z_{ij}) \begin{pmatrix} 4 \\ 2Z_{ij} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (\text{Equation 4})$$

where  $W_{ij} = (\text{Var}(T_{ij} | Z_{ij}))^{-1}$ . We are primarily interested in a 1-sided test of  $H_0: \delta = 0$  versus  $H_A: \delta > 0$  because, given the non-random ascertainment of affected sibships, we anticipate rare risk variants to be observed on segments shared identical by descent. However, we can also perform the standard 2-sided test of  $H_0: \delta = 0$  versus  $H_A: \delta \neq 0$ , if desired, to allow for the possibility of rare protective variants observed in regions not shared identical by descent by an

affected sib pair. We note that analysis via Equation 4 is robust to confounding due to population stratification because, although ancestry might correlate with the number of copies of rare variants in a region, ancestry will not correlate with IBD sharing in a region under the null hypothesis that the genetic variation we are testing is neither causal nor in linkage disequilibrium with other causal regions.

Inference on  $\delta$  within Equation 4 requires estimates of the variance parameters  $\sigma_0^2$  and  $\sigma_1^2$  shown in Equation 2. To estimate these parameters, we first calculate the sample variance for the number (or weighted sum) of rare variants possessed by affected sib pairs sharing 0, 1, or 2 alleles identical by descent and denote these quantities by  $\hat{V}_0$ ,  $\hat{V}_1$ , and  $\hat{V}_2$ , respectively. Based on Equation 2, we can then estimate the variance parameters by solving the system of equations

$$\begin{bmatrix} \hat{V}_0 \\ \hat{V}_1 \\ \hat{V}_2 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 2 & 4 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} \sigma_0^2 \\ \sigma_1^2 \end{bmatrix}.$$

Defining  $\hat{V} = (\hat{V}_0, \hat{V}_1, \hat{V}_2)^T$  and  $X = \begin{bmatrix} 4 & 0 \\ 2 & 4 \\ 0 & 8 \end{bmatrix}$ , we then estimate

the variance parameters as  $\begin{bmatrix} \hat{\sigma}_0^2 \\ \hat{\sigma}_1^2 \end{bmatrix} = (X^T X)^{-1} X^T \hat{V}$ . We then use

these estimates to define  $W_{ij}$  in Equation 4. If IBD state is known only probabilistically, we assign each sib pair to its most probable IBD state when calculating  $\hat{V} = (\hat{V}_0, \hat{V}_1, \hat{V}_2)^T$ .

In the formulation of Equation 4, we assumed that we inferred the IBD sharing variable  $Z_{ij}$  with certainty. Although this assumption probably holds when inferring IBD information from sequence data, the use of sparser SNP sets from the exome chip for estimation might lead to IBD uncertainty. In this case, we specify  $Z_{ij}$  as the expected IBD dosage by using  $Z_{ij} = \hat{P}[Z_{ij} = 1] + 2\hat{P}[Z_{ij} = 2]$ , where the estimated IBD probabilities are calculated by MERLIN.<sup>27</sup> We can insert such IBD dosage variables directly into the mean model in Equation 3. To model  $W_{ij} = (\text{Var}(T_{ij}|Z_{ij}))^{-1}$  for a pair with uncertain IBD sharing, we assign the IBD state to be the most-likely IBD sharing category. Although this leads to misspecification of  $\text{Var}(T_{ij}|Z_{ij})$ , we note that such misspecification does not affect the validity of our estimation-equation testing framework and only potentially affects the power.

### Efficient Score Function and Test Statistics

Rather than solve a system of two equations (Equation 4), we can instead simplify the problem by finding the efficient score for  $\delta$  from the weighted regression of  $T_{ij}$  on  $Z_{ij}$ . After first normalizing the weights  $W_{ij}$ , we define the centered count and IBD variables as  $\tilde{T}_{ij} = T_{ij} - \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} T_{ij}$  and  $\tilde{Z}_{ij} = Z_{ij} - \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} Z_{ij}$ , respectively. As we show in Appendix A, the efficient score to test  $H_0: \delta = 0$  based on Equation 4 is proportional to

$$\mathbf{u} = \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} \tilde{T}_{ij} \tilde{Z}_{ij} = \sum_{j=1}^N \sum_{i=1}^{S_j} U_{ij}. \quad (\text{Equation 5})$$

To construct the estimated variance accounting for within-sibship correlation of sib pairs, we use a robust variance estimator that accounts for cluster membership.<sup>28</sup> This variance estimator has the form

$$\mathbf{v} = \sum_{j=1}^N \left\{ \sum_{i=1}^{S_j} W_{ij} \tilde{T}_{ij} \tilde{Z}_{ij} \right\}^2 - N \left( \frac{\mathbf{u}}{N} \right)^2. \quad (\text{Equation 6})$$

With Equations 5 and 6, we can then construct a closed-form burden test to assess  $H_0: \delta = 0$  by writing

$$Y_{\text{Burden}} = \frac{\mathbf{u}}{\sqrt{\mathbf{v}}}.$$

Under  $H_0: \delta = 0$ ,  $Y_{\text{Burden}}$  follows a standard normal distribution.

In addition to a burden test, we can also consider a variance-component test of rare-variant association analogous to SKAT.<sup>22</sup> The approach assumes that  $\delta$  follows a random effect with mean 0 and variance  $\tau$  and then considers a score test of  $H_0: \tau = 0$ . Such a variance-component test might be preferable to the burden test  $Y_{\text{Burden}}$  if we expect, in addition to rare susceptibility variants in regions shared identical by descent, the possibility that a region might harbor rare protective variants on regions not shared identical by descent by an affected sib pair. We can use the efficient score framework to construct such a test by summing the squared contribution of the efficient score function at each variant site. By definition,  $T_{ij} = \sum_{r=1}^R T_{ijr}$ , where  $T_{ijr}$  denotes the total number of rare variants possessed at variant site  $r$  ( $r = 1, \dots, R$ ) by sib pair  $i$  in sibship  $j$ . Because  $R$  is fixed and rare variants typically exhibit low linkage disequilibrium, we expect  $\text{Var}(T_{ijr}|Z_{ij}) \approx (1/R) \text{Var}(T_{ij}|Z_{ij})$ , suggesting that we choose  $W_{ijr} = W_{ij}$  as calculated previously, because the weights we use are normalized. Similar to Equation 6, we can define  $U_{ijr} = W_{ij} \tilde{T}_{ijr} \tilde{Z}_{ij}$  and then define our variance-component test statistic as

$$Y_{\text{VC}} = \sum_{r=1}^R \left\{ \sum_{j=1}^N \sum_{i=1}^{S_j} U_{ijr} \right\}^2.$$

Under the null hypothesis,  $Y_{\text{VC}}$  follows a mixture of chi-square distributions<sup>29</sup>  $\sum_r \lambda_r \chi_1^2$  where the  $\lambda$  parameters denote the eigenvalues of the  $R$  by  $R$  matrix  $\Sigma$ , where row  $r$  and column  $r'$  element of the matrix has the form

$$\Sigma_{r,r'} = \sum_{j=1}^N \left\{ \sum_{i=1}^{S_j} U_{ijr} U_{ijr'} \right\} - \frac{1}{N} \left\{ \sum_{j=1}^N \sum_{i=1}^{S_j} U_{ijr} \right\} \left\{ \sum_{j=1}^N \sum_{i=1}^{S_j} U_{ijr'} \right\}.$$

As before, this variance estimator accounts for within-sibship correlation.<sup>28</sup> We can analytically calculate the p value of  $Y_{\text{VC}}$  via Davies method.<sup>30</sup>

### Application to Simulated Datasets

We evaluated the size and power of our test for rare-variant association testing in sibships by using simulated sequence data. With COSI,<sup>31</sup> we generated 10,000 haplotypes of length 10 kb whose variation patterns emulated those observed in HapMap CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) samples. To generate data with confounding by population stratification, we similarly simulated 10,000 haplotypes whose variation patterns mimicked those observed in HapMap YRI (Yoruba in Ibadan, Nigeria) samples. For each population sample, we generated sequence data for a sibship by first randomly sampling two haplotypes for each parent: parents were either both European for simulations without confounding, or both European or both African for simulations with confounding. We then randomly transmitted one maternal and one paternal haplotype to each offspring in a sibship, keeping track of the inheritance flow in the family to derive pair-wise IBD sharing among all pairs of siblings.

We then modeled the probability of disease for an index proband within a sibship by using a logistic function that has an intercept, a term for African origin (for simulations with confounding),

**Table 1. Distribution of Sibship Size in GENOA Dataset**

Sibship Size	Count
1 (singleton)	204
2	226
3	101
4	53
5	23
6	8
7	6
8	2

and an additive model for each risk variant (for models evaluating power). For each simulation scenario, we chose the intercept to yield an overall disease prevalence of 0.03. African origin increased the odds of disease by a factor of 4 for simulations with confounding. For power simulations, we assumed 15% of rare (minor allele frequency [MAF] < 0.05) variants were causal with the magnitude of the log odds ratio given by  $c|\log_{10}(\text{MAF})|$ , where we chose  $c$  to be  $\log(4)/4$ ,  $\log(6)/4$ , or  $\log(8)/4$ , corresponding to a causal risk variant with a MAF of 0.0001 having a disease odds ratio of 4, 6, or 8, respectively. For some power simulations, we assumed that all causal variants increased risk for disease. For other power simulations, we assumed that half of causal variants increased risk and the other half were protective and decreased risk. We then randomly sampled the disease outcome for the index proband and discarded the sibship if the index proband was unaffected.

Given that the index proband is affected, we then generated disease outcomes for each of the proband's siblings assuming a total sibship size of 2, 3, or 4 with the following model. Let  $y_1$  and  $y_2$  denote the disease status ( $y = 1$  denotes affected) for the index proband and an additional sibling, respectively. Further, define  $\mathbf{g}_1$  and  $\mathbf{g}_2$  as the genotype vectors for rare susceptibility variants at the test gene possessed by the index proband and additional sibling, respectively. Also, let  $a$  be an indicator of African origin for the two siblings. We must specify  $P(y_2 = 1|y_1 = 1, \mathbf{g}_1, \mathbf{g}_2, a)$ ; the probability the additional sibling is affected conditional on the index proband being affected and conditional on the rare-susceptibility genotypes and ancestry of both siblings. To do this, we define a conditional version of Risch's sibling recurrence risk ratio<sup>32</sup> as  $\lambda_C = P(y_2 = 1|y_1 = 1, \mathbf{g}_1, \mathbf{g}_2, a)/P(y_2 = 1|\mathbf{g}_1, \mathbf{g}_2, a)$ , which models the overall risk of disease due to shared sibling effects conditional on the sibling's ancestry and genotypes at the rare risk variants in the gene of interest. We rewrite  $\lambda_C$  as a function of our probability of interest as  $P(y_2 = 1|y_1 = 1, \mathbf{g}_1, \mathbf{g}_2, a) = \lambda_C P(y_2 = 1|\mathbf{g}_1, \mathbf{g}_2, a)$ . By making the reasonable assumption that  $P(y_2 = 1|\mathbf{g}_1, \mathbf{g}_2, a) = P(y_2 = 1|\mathbf{g}_2, a)$  (risk of disease of the additional sibling is independent of the genotype information of the index proband conditional on the additional sibling's own genotype and the ancestry of the sibship), our probability of interest simplifies to  $P(y_2 = 1|y_1 = 1, \mathbf{g}_1, \mathbf{g}_2, a) = \lambda_C P(y_2 = 1|\mathbf{g}_2, a)$ . We then model this probability by varying the value of  $\lambda_C$  among values of 2, 4, and 8 (with larger value indicating increased global risk of disease due to shared sibling effects) and specifying  $P(y_2 = 1|\mathbf{g}_2, a)$  via the same logistic function used to generate the disease phenotype for the index proband. Data for each successive sib are sampled conditional on the index proband but independently of the disease status of subsequent sibs.

With this model, we generated disease outcomes for 1, 2, or 3 successive sibs independently to obtain data for sibships of size 2, 3, or 4, respectively. We retained the sibship for analysis if the resulting disease data satisfied the ascertainment criterion and discarded the sibship otherwise. Our primary ascertainment criterion was that the sibship possess at least two affected siblings but, for sibships of size 3 and 4, we also considered a criterion requiring three or more affected siblings for analysis. For a given simulation model, we prospectively generated sequence and phenotype data on sibships until we achieved a targeted number of sibships that fulfilled the ascertainment criterion.

### Application to GENOA Dataset

The GENOA study,<sup>8,33</sup> which forms a component of the NHLBI Family Blood Pressure Program,<sup>34</sup> seeks to identify genetic variants that influence risk for hypertension and arteriosclerotic/atherosclerotic complications involving the heart, brain, kidneys, and peripheral arteries. As part of the study, GENOA ascertained a cohort of African American sibs from Jackson, Mississippi, in which  $\geq 2$  siblings had essential hypertension diagnosed prior to the age of 60. 1,854 subjects were enrolled in the initial study and approximately 80% were re-recruited for a second exam many years later. Clinical data collected on such subjects include essential hypertension status as well as quantitative and qualitative outcomes related to lipid metabolism and organ function. This study was approved by the institutional review boards at the parent institutions of all participating laboratories involved in GENOA.

The GENOA investigators genotyped 1,429 samples for 233,507 SNPs on the Illumina HumanExome Beadchip. After standard data cleaning to remove problematic individuals and SNPs, the final sample for analysis consisted of 1,392 participants in 623 sibships. We summarize the distribution of sibships in Table 1. GENOA investigators used PLINK<sup>35</sup> to estimate genome-wide IBD sharing among relative pairs to confirm putative pedigree structure. To infer site-specific IBD sharing in the GENOA sibships for use in our proposed tests, we used 4,448 common autosomal SNPs specifically included on the Illumina HumanExome Beadchip to infer inheritance flow. We identified the sex-averaged linkage map position for each of these SNPs from the Rutgers Map v.3<sup>36</sup> (see Web Resources). With this information, we first used MERLIN<sup>27</sup> to identify likely genotyping errors for the 4,448 SNPs in the GENOA families and subsequently set such problematic genotypes to be missing. We then applied MERLIN on the cleaned genotype data to infer IBD sharing among GENOA family members at 1 centiMorgan (cM) intervals across the autosomes.

With our rare-variant association test, we performed two sets of analyses with affected siblings in GENOA. We first studied GENOA affected siblings with essential hypertension (MIM 145500); published sibling recurrence risk ratios for hypertension range between 1.5 and 3.5.<sup>37</sup> We analyzed rare and less-common variants (MAF  $\leq 0.05$ ) found in a set of candidate genes containing common variation previously associated with blood pressure in independent African American GWAS samples.<sup>38</sup> We next focused on GENOA affected siblings with obesity (MIM 601665); published sibling recurrence risk ratios for obesity range between 2.6 and 4.2.<sup>39</sup> We analyzed rare and less-common variants found in genes harboring common variation previously associated with body-mass index (BMI) in independent African American GWASs.<sup>40</sup> For each set of analyses, we removed unaffected siblings prior to association testing. We show the distribution of GENOA



**Table 2. Distribution of Affected Sibships in GENOA**

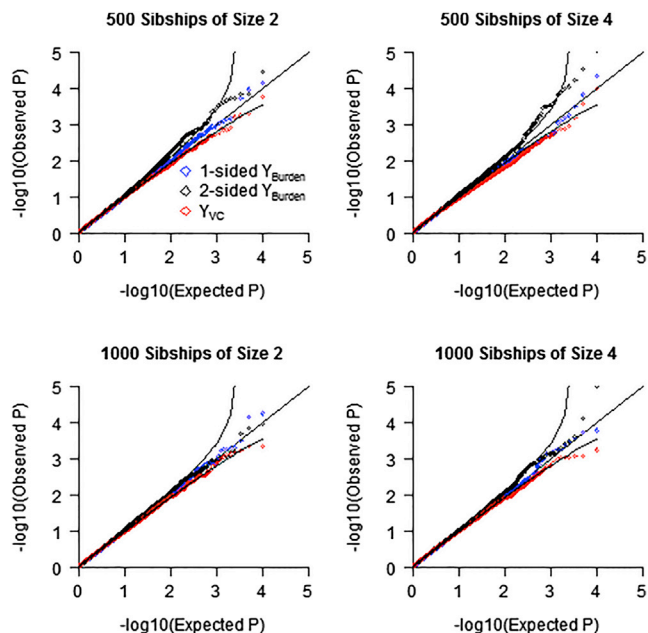
Sibship Size	Count with Hypertension	Count with Obesity
2	204	183
3	75	68
4	35	26
5	12	9
6	4	1
7	0	1
8	1	1
Total Number of Pairs	847	697
Number of Pairs with Both Hypertension and Obesity	484	

affected sibships with the hypertension and obesity phenotypes in Table 2. As shown in the table, there are a total of 847 affected sibling pairs that are hypertensive and 697 affected sibling pairs that are obese. Of these sib pairs, 484 overlap between the two datasets, so that the proportion of hypertensive sib pairs that were also obese was 0.57 while the proportion of obese sib pairs that were also hypertensive was 0.69.

We identified the physical location of each tested gene from the UCSC Genome Browser<sup>41</sup> and included all variants on the Illumina HumanExome Beadchip with  $MAF \leq 0.05$  from 20 kb upstream of the start of the gene position to 20 kb downstream beyond the end of the gene position. We found the corresponding linkage map position (sex averaged) for each gene by using Rutgers Map v.3 and extracted the IBD sharing of the GENOA affected sibships at that position from our MERLIN output. We considered both  $Y_{Burden}$  and  $Y_{VC}$  and performed both unweighted analyses as well as more standard analyses that weighted rare variants inversely proportional to their MAF (using a weight corresponding to  $(\sqrt{MAF(1-MAF)})^{-1}$  as suggested by others<sup>19</sup>).

### Simulations with Uncertainty in IBD Sharing

In the power simulations described previously, we assumed that IBD sharing among affected sib pairs was known exactly when constructing  $Y_{Burden}$  and  $Y_{VC}$ . By using next-generation sequencing data, we would expect to infer IBD with near certainty. However, for sparser data like the common SNPs found on the Illumina HumanExome Beadchip, ambiguity in IBD sharing is a possibility. To investigate the impact of IBD uncertainty on our approaches, we looked at pairwise IBD estimates among affected siblings in GENOA across the candidate genes that we explored in this article. We provide a histogram of these estimates in Figure S1. We repeated the simulations described previously but, instead of using the true IBD sharing for each pair, we induced ambiguity in IBD sharing by sampling from the distribution of IBD estimates observed in the GENOA pairs. Specifically, if a pair's true IBD sharing was 0, we sampled the observed IBD sharing from the distribution of GENOA IBD estimates between 0 and 0.5. If true IBD sharing was 1, we sampled an observed IBD from GENOA estimates between 0.5 and 1.5. Finally, if true IBD sharing was 2, we sampled an observed IBD estimate from the GENOA estimates between 1.5 and 2.



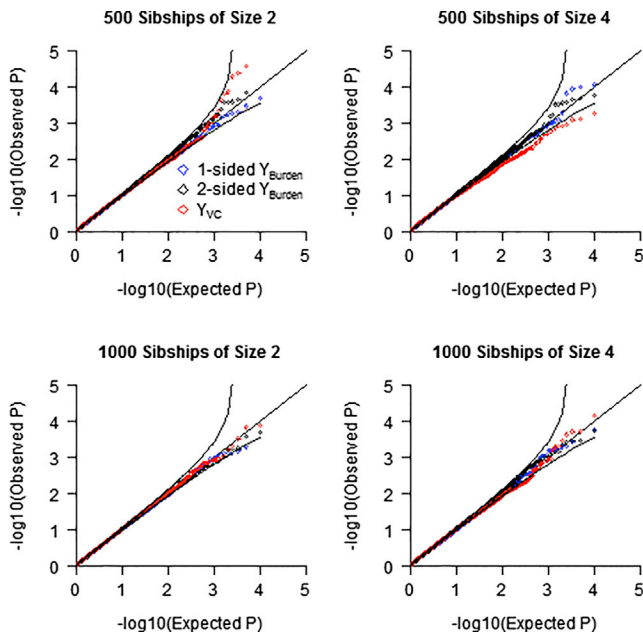
**Figure 1. Quantile-Quantile Plots of Observed p Values versus Expected p Values on  $-\log_{10}$  Scale for the  $Y_{Burden}$  and  $Y_{VC}$  Tests under the Null Hypothesis for Samples of European Ancestry** Top left shows QQ plot for 500 sibships of size 2, top right shows 500 sibships of size 4, bottom left shows 1,000 sibships of size 2, and bottom right shows 1,000 sibships of size 4. Each set of results based on 10,000 replications with  $\lambda_C = 4$ .

## Results

### Type I Error Results

We performed null simulations to ensure that  $Y_{Burden}$  and  $Y_{VC}$  had appropriate size under different conditions. Figure 1 shows quantile-quantile (QQ) plots for the two tests in a situation where the sample consisted entirely of subjects of European ancestry. The top half of Figure 1 shows the QQ plots assuming a sample size of 500 sibships of size 2 (top left) or of size 4 (top right), and the bottom half of Figure 1 shows QQ plots for sample sizes of 1,000 sibships of size 2 (bottom left) or of size 4 (bottom right). For each design, we analyzed 10,000 replicates of the data assuming  $\lambda_C = 4$  (we show QQ plots under the same models when  $\lambda_C = 8$  in Figure S2). The results show that both statistics have appropriate type I error under the null and, based on results that assumed sibships of size 4, appropriately account for the dependency among observations due to within-sibship correlation.

The type I error simulations shown in Figure 1 assumed samples entirely of European ancestry (no confounding). Figure 2 shows QQ plots for simulations with confounding, assuming a sample size of 500 sibships of size 2 (top left) or of size 4 (top right), or a sample size of 1,000 sibships of size 2 (bottom left) or of size 4 (bottom right). For each model, we analyzed 10,000 replicates of the data assuming  $\lambda_C = 4$  (we provide QQ plots for  $\lambda_C = 8$  in Figure S3). The results show that both tests are valid in



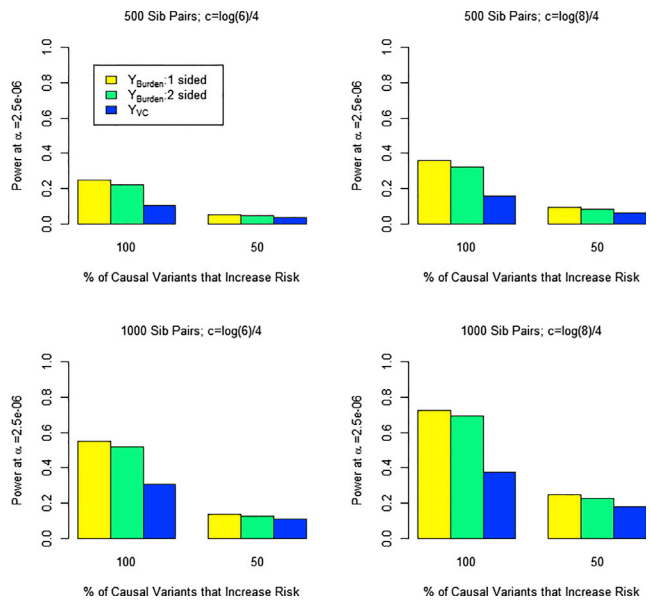
**Figure 2.** Quantile-Quantile Plots of Observed  $p$  Values versus Expected  $p$  Values on  $-\log_{10}$  Scale for the  $Y_{\text{Burden}}$  and  $Y_{\text{VC}}$  Tests under the Null Hypothesis with Samples of Both African and European Ancestry

Odds ratio of disease for African versus European ancestry is 4. Top left shows QQ plot for 500 sibships of size 2, top right shows 500 sibships of size 4, bottom left shows 1,000 sibships of size 2, and bottom right shows 1,000 sibships of size 4. Each set of results based on 10,000 replications with  $\lambda_C = 4$ .

the presence of confounding resulting from population stratification.

### Power Results

In [Figure 3](#) we compare the power of  $Y_{\text{Burden}}$  and  $Y_{\text{VC}}$  for simulated data on affected sib pairs of European ancestry. For all simulations we assumed  $\lambda_C = 4$ . Power was markedly higher for all tests when all causal alleles are risk alleles. However, even when half the causal alleles are protective, we observed that both the 1-sided and 2-sided  $Y_{\text{Burden}}$  test outperformed  $Y_{\text{VC}}$ . This result is in contradistinction to results showing that variance-component tests generally outperform burden tests in case-control and cross-sectional studies when the proportion of causal variation is small.<sup>22</sup> We believe the reason we observe this phenomenon is that the sampling of affected sibs implies the presence of cosegregating risk alleles in the family, which in turn implies rare risk variants are shared identical by descent. Protective variants preferentially found on haplotypes *not* shared identical by descent would probably have a smaller effect than risk variants for two reasons. First, each sib already has enough risk alleles to cause disease. Second, recall that haplotypes in affected sibships not shared identical by descent are like untransmitted haplotypes in a transmission-disequilibrium test; the distortion of allele frequencies in untransmitted haplotypes is typically much smaller than that in transmitted haplotypes. In a related finding, Jiang et al.<sup>42</sup> observed burden tests



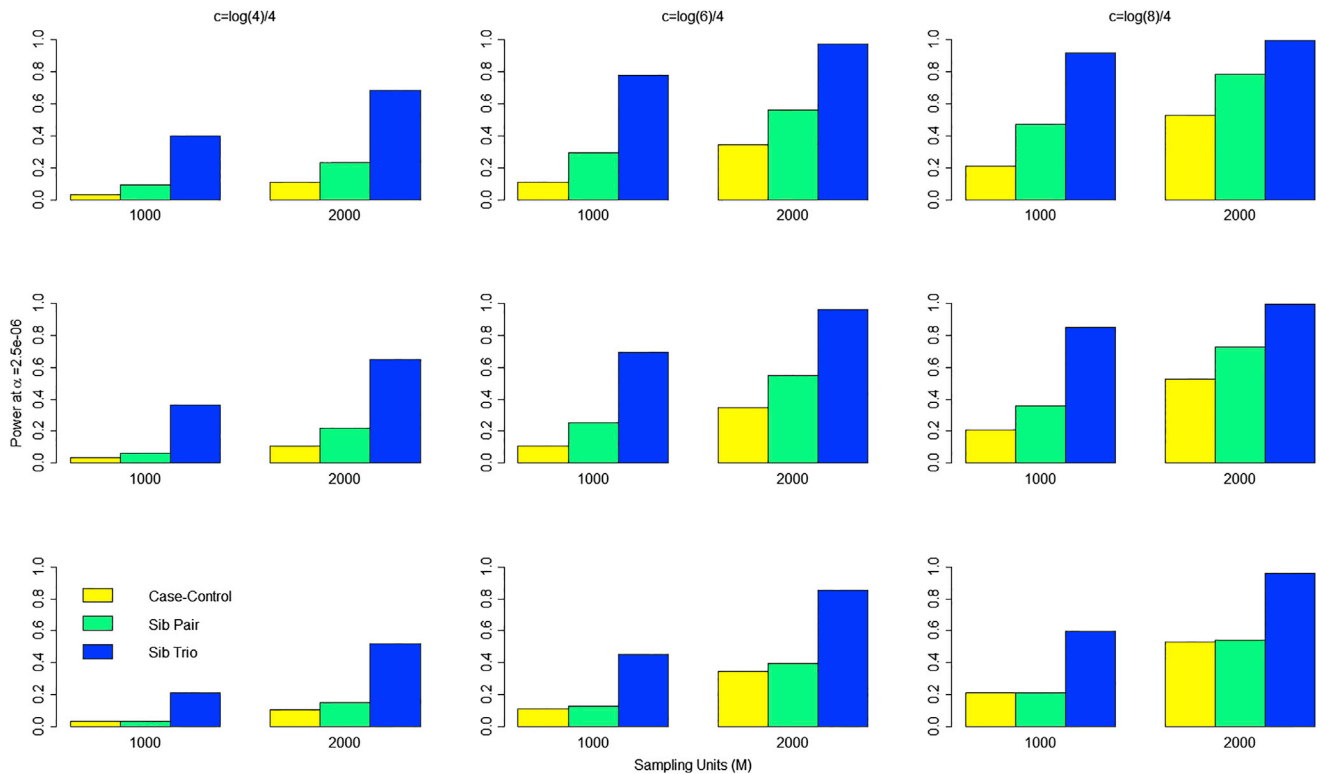
**Figure 3.** Power of  $Y_{\text{Burden}}$  and  $Y_{\text{VC}}$  Tests at  $\alpha = 2.5 \times 10^{-6}$

Data were generated under a model where 15% of variants with  $\text{MAF} \leq 0.05$  were causal with effect on  $\log(\text{OR})$  of  $c \cdot |\log_{10}(\text{MAF})|$  (for risk variants) and  $-c \cdot |\log_{10}(\text{MAF})|$  (for protective variants). Results are shown for  $c = \log(6)/4$  and 500 sib pairs (top left),  $c = \log(6)/4$  and 1,000 sib pairs (top right),  $c = \log(8)/4$  and 500 sib pairs (bottom left), and  $c = \log(8)/4$  and 1,000 sib pairs (bottom right). Each set of results based on 1,000 replications with  $\lambda_C = 4$ .

to be more powerful than variance-component tests in the study of case-parent trios, even when a gene harbored an equal mixture of rare and protective variants. Based on our findings, we recommend using the 1-sided  $Y_{\text{Burden}}$  test as the primary analytic tool for rare-variant testing in affected sibships.

We also performed simulations to compare the power of our rare-variant association tests for affected sibships to the power of a case-control study having the same sample size. We assumed samples comprised of  $M$  individuals that were collected under one of three possible designs. First, we generated sequence data prospectively under a case-control design assuming  $M/2$  case subjects and  $M/2$  control subjects, and we analyzed the data via a logistic-regression burden test that regressed disease status on the total number of rare variants possessed by a subject. Second, we generated data on  $M/2$  affected sib pairs and analyzed the data via our 1-sided  $Y_{\text{Burden}}$  test. Finally, we generated data on  $M/3$  families, each having three affected sibs (sib trios) and analyzed the data again with our 1-sided  $Y_{\text{Burden}}$  test. Although ascertainment of siblings based on three or more affected individuals is rare, we were curious to see whether increasing the genetic load by requiring three affected sibs would increase the power even as the number of sibships decreased.

[Figure 4](#) provides power comparisons at  $\alpha = 2.5 \times 10^{-6}$  (corresponding to a Bonferroni correction for testing 20,000 independent genes) of case-control, affected sib pair, and affected sib trio designs assuming 1,000 or 2,000 total sampling units. For all simulations, we used



**Figure 4. Power at  $\alpha = 2.5 \times 10^{-6}$  to Detect Rare Susceptibility Variants under Different Sampling Designs Assuming a Fixed Number of  $M$  Sampling Units**

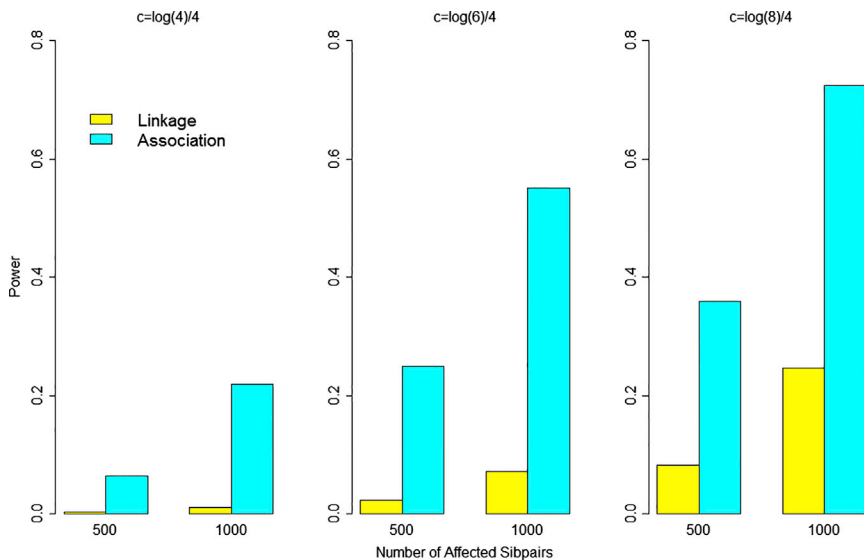
Yellow bar denotes power using  $M/2$  cases and  $M/2$  controls using logistic-regression burden test. Green bar denotes power for  $M/2$  affected sib pairs using 1-sided  $Y_{\text{Burden}}$  test. Blue bar denotes power for  $M/3$  affected sib trios using 1-sided  $Y_{\text{Burden}}$  test. Results are shown for  $c = \log(4)/4$  (left column),  $c = \log(6)/4$  (middle column), and  $c = \log(8)/4$  (right column). Top row,  $\lambda_C = 2$ ; middle row,  $\lambda_C = 4$ ; bottom row,  $\lambda_C = 8$ . Each set of results based on 1,000 replications assuming 15% of rare variants are causal with 100% of causal variants increasing risk.

only samples of European ancestry and assumed that all causal variants increased risk. We varied the value of  $\lambda_C$  among the values of 2, 4, and 8 with corresponding power results shown in the top, middle, and bottom panels of Figure 4. In all scenarios, the results show that sampling affected sibships is more powerful than the using a case-control sample assuming the same number of sampling units. Figure 4 also shows that the power of sib trios is considerably higher than sib pairs, suggesting that more stringent ascertainment criteria can improve the ability to detect rare susceptibility variation. We should note that the 1-sided  $Y_{\text{Burden}}$  test has appropriate size for affected sib trios (see Figure S4). Further, when we repeated the simulations summarized in Figure 4, using the “observed” IBD sharing estimated in the GENOA sibs in place of the true IBD sharing, we observed negligible power loss for our method, compared to using the true IBD for a pair (see Figure S5).

Figure 4 also shows that the power of the 1-sided  $Y_{\text{Burden}}$  test generally decreases with increasing value of  $\lambda_C$ . This trend occurs because an increase in  $\lambda_C$  represents an increase in the overall (polygenic) risk due to shared sibling factors, not just the risk at the test locus. As  $\lambda_C$  increases, the risk of disease increases even for sibs with no rare risk variants in the gene of interest. Consequently, the propor-

tion of risk load carried by rare variation in a gene of interest among affected sibships decreases as  $\lambda_C$  increases, which leads to reduced power with the 1-sided  $Y_{\text{Burden}}$  test. Nevertheless, even at sibling recurrence risk ratios of 8, the 1-sided  $Y_{\text{Burden}}$  test has equal or improved power compared to the case-control design for a fixed sample size.

We also compared the power of rare-variant association testing via the 1-sided  $Y_{\text{Burden}}$  test to the power of detecting linkage at the same gene using the affected sibling pairs generated to create Figure 4. We tested for linkage via a minimax test<sup>43</sup> assuming the IBD sharing of each affected sib pair was known with certainty. We assumed the genome-wide significance threshold for declaring linkage suggested by Lander and Kruglyak.<sup>44</sup> In Figure 5, we compare the power to detect linkage to the power to detect association assuming an appropriate genome-wide significance threshold for both approaches, using our simulated datasets that assumed  $\lambda_C = 4$ . The results clearly show that the power to detect a genome-wide significant association via the 1-sided  $Y_{\text{Burden}}$  test is greater than the corresponding power to detect a genome-wide significant linkage signal across simulation models. Further, although we observed correlation between the p values of the linkage and association tests, the correlation was imperfect and we identified many instances where we observed a



**Figure 5. Power to Detect Linkage and Power to Detect Association in Affected Sib Pairs under Different Sampling Designs Assuming 500 or 1,000 Affected Sib Pairs**

Yellow bar denotes power to detect linkage at genome-wide significance threshold of  $10^{-5}$ . Magenta bar denotes power to detect association using  $Y_{\text{Burden}}$  at genome-wide threshold of  $2.5 \times 10^{-6}$ . Results shown for  $c = \log(4)/4$  (left),  $c = \log(6)/4$  (middle), and  $c = \log(8)/4$  (right). Each set of results based on 1,000 replications assuming  $\lambda_C = 4$  and 15% of rare variants are causal with 100% of causal variants increasing risk.

significant association signal in the absence of a linkage signal. For example, among the simulated datasets assuming  $c = \log(6)/4$  (middle bar plot in Figure 5), we estimated that the Spearman rho coefficient between the linkage and association p values was 0.30 for 500 affected sib pairs, which indicates a significant ( $p < 0.0001$ ) yet imperfect correlation. However, of the 294 replicates that yielded an (unadjusted) linkage p value  $> 0.05$  (indicating little or no evidence of linkage), 40% yielded a genome-wide-significant association signal via the 1-sided  $Y_{\text{Burden}}$  test. These results demonstrate that our approach can detect association between rare variants and IBD sharing in the absence of a strong linkage signal.

### Application to GENOA Study

We applied our approach to test whether rare variants in previously identified candidate genes affected essential hypertension or obesity in the GENOA affected sibs. Based on our simulation findings, we used the 1-sided  $Y_{\text{Burden}}$  test as our primary analysis tool. We list each gene analyzed in Table 3 along with its corresponding p values derived from a weighted (based on MAF) or unweighted 1-sided  $Y_{\text{Burden}}$  test. We performed association testing of 5 genes with hypertension and 12 genes with obesity (with no overlap in the genes tested for each phenotype). We then performed these 17 tests twice using both the weighted and unweighted forms. Although the weighted and unweighted tests are likely to be correlated, we elected to apply a Bonferroni adjustment for 34 tests that yielded a significance threshold of  $\alpha = 0.05/34 = 0.0015$ . Although we observed no significant association between rare variants in blood-pressure genes with essential hypertension, we observed an interesting association between rare variants in *NRXN3* (MIM 600567) and obesity (unadjusted  $p = 0.0024$ , adjusted  $p = 0.0816$ ). The unweighted 1-sided  $Y_{\text{Burden}}$  test for this gene was less interesting, because it yielded an unadjusted p value of 0.043. Further, we did not observe a corresponding linkage signal at the locus

IBD sharing among affected sib pairs at the *NRXN3* confirmed no evidence of linkage as the proportion of affected sib pairs that shared 0, 1, and 2 alleles identical by descent (rounded to nearest integer) was 27%, 51%, and 22%, respectively (see Table 4). To determine why the weighted  $Y_{\text{Burden}}$  test yielded an interesting association in the absence of either a signal from the unweighted  $Y_{\text{Burden}}$  test or the linkage test, we examined the average rare-variant burden of affected sib pairs with IBD = 0, 1, and 2. As shown in Table 4, the average rare-variant burdens of the IBD = 0 and IBD = 2 affected sib pairs were similar for the unweighted analysis. However, the allele frequency of the rare variants decreased monotonically as the IBD state increases, so that for the weighted analysis, the average burden of the IBD = 2 affected sib pairs was greater than that in IBD = 1 pairs, which in turn was greater than that of the IBD = 0 affected sib pairs.

The limited number of interesting rare-variant findings in the GENOA study might indicate inherent differences between our simulated and real data examples. We don't believe these differences are due to the sibling recurrence risk ratios assumed in the simulations because they include published values reported for hypertension and obesity (1.5–3.5 and 2.6–4.2, respectively<sup>37,39</sup>). We do note there are two intrinsic differences between our simulation design and the GENOA analyses. First, the GENOA study provided Illumina HumanExome Beadchip data whereas our simulations assumed sequence data. Thus, it is possible that the issue might be that the true causal variation in the candidate genes investigated in the GENOA study are not well interrogated on the HumanExome Beadchip and must be assayed by a sequencing approach. Second, we note the sample size of the GENOA study (865 affected subjects in 331 sibships for studying hypertension; 740 affected subjects in 289 sibships for studying obesity) is smaller than the sample sizes considered in our simulations, which could explain some of the difference between the power we observed



**Table 3. Rare-Variant Analysis in GENOA Affected Sibships**

Outcome	Gene	p Value (Weighted)	p Value (Unweighted)
Hypertension	<i>GOSR2</i>	0.865	0.674
	<i>GUCY1A3</i>	0.421	0.513
	<i>PLEKHG1</i>	0.378	0.387
	<i>SH2B3</i>	0.970	0.932
	<i>ULK4</i>	0.834	0.578
Obesity	<i>ADCY3</i>	0.799	0.671
	<i>BDNF</i>	0.520	0.585
	<i>FTO</i>	0.597	0.346
	<i>GNPDA2</i>	0.636	0.451
	<i>KLHL32</i>	0.442	0.254
	<i>MAP2K5</i>	0.890	0.670
	<i>MC4R</i>	0.142	0.191
	<i>NFE2L3</i>	0.919	0.872
	<i>NRXN3</i>	0.0024	0.043
	<i>SEC16B</i>	0.739	0.632
	<i>SH2B1</i>	0.393	0.305
	<i>QPCTL</i>	0.382	0.232

p value for  $Y_{\text{Burden}}$  based on 1-sided test. Weights based on  $(\sqrt{\text{MAF}(1 - \text{MAF})})^{-1}$ .

in our simulations and the results of our analyses of the GENOA data.

## Discussion

As interest grows in sequencing or genotyping exome chips on familial samples (such as those collected previously for linkage analysis), novel statistical procedures that permit rare-variant association testing for these studies are necessary. In this work, we proposed a rare-variant association test for affected sibships of arbitrary size that examines the relationship between the distribution of rare variants in a gene and the number of genes shared identical by descent among affected sib pairs. With simulations, we showed that rare-variant testing in affected sibships is more powerful than a case-control study with the same sample size, thereby illustrating the strengths of family-based data for rare-variant association analysis. Our method also possesses additional practical features: it does not use any information from control samples; it is robust to confounding due to population stratification; and our test statistics have simple closed forms with analytic p values, making use of our approach straightforward via standard software packages (see [Web Resources](#) for R code implementing our approach). Further, we have shown that our methods perform well even when there is uncertainty in IBD estimates from sparse SNP sets. We illustrated the approach with an application to exome

**Table 4. Rare-Variant Association Analysis of *NRXN3* in GENOA Study**

IBD State	Number (Proportion) of Affected Sib Pairs	Average Frequency of Rare Variants in <i>NRXN3</i>	Average Burden Per Sib Pair	
			Unweighted	Weighted
0	187 (0.27)	.0070	0.043	0.820
1	354 (0.51)	.0052	0.080	1.568
2	156 (0.22)	.0024	0.067	2.310

IBD state of each affected sibling pair rounded to nearest integer value. Rare variants are those with sample MAF < 0.05. Weighted analysis applied weight based on  $(\sqrt{\text{MAF}(1 - \text{MAF})})^{-1}$ .

chip data from the GENOA study in efforts to identify genes harboring rare variation that increase risk for essential hypertension and obesity and identified a gene (*NRXN3*) that might warrant further study for association with obesity.

Although our rare-variant association test is based on IBD sharing in affected sibling pairs, it is not a test of linkage and does not require the presence of a linkage signal to detect association (as demonstrated in the [Results](#) section and in [Figure 5](#)). Linkage tests treat IBD sharing as the outcome variable and looks for regions where estimated IBD sharing across affected pairs deviate from their expected values based only on familial relationships. Our proposed association test compares patterns of rare variation in a gene conditional on the IBD sharing in a region. If linkage exists but there is no difference in rare-variation patterns between regions shared identical by descent and not shared identical by descent, then our proposed test will not find a significant result. Conversely, our approach can detect a significant association if the patterns of rare variation differ between haplotypes that are and are not shared identical by descent, even if there is no excess IBD sharing as would be expected under linkage.

Our approach is tailored for rare-variant analysis in affected sib pairs within sibships. In theory, we could expand our approach to incorporate discordant sib pairs in sibships where one sib is affected and the other sib is unaffected. For such discordant sib pairs, we would expect to find more rare susceptibility variants found on regions not shared identical by descent than on regions shared identical by descent. We can modify our framework in [Equation 4](#) to detect such patterns by incorporating the expected sign of  $\delta$  (positive for affected sib pairs, negative for discordant sib pairs) into the analysis. However, when we performed preliminary power simulations to investigate the potential utility of incorporating discordant pairs into analysis, we found that such discordant units provided meager power (slightly more than the empirical size) to identify regions harboring rare susceptibility variants unless the proportion of causal variants and their associated effect sizes were unreasonably large (results not shown). In fact, this is related to our finding that our variance components test was always less powerful than the

two-sided burden test even when half of the causal alleles were protective. Consequently, although we can use genetic information on unaffected sibs to improve the accuracy of IBD estimation in affected sib pairs, we do not recommend incorporating their information directly into our rare-variant testing framework.

As we demonstrated in our simulations, our burden and variance-component tests follow known asymptotic distributions to facilitate genome-wide analysis. If desired, we can also use resampling procedures to derive p values for such tests; such procedures are probably valuable when analyzing small samples. Because we assume an underlying linear model in Equation 4, we can implement a resampling procedure via a nonparametric bootstrap approach<sup>46</sup> that resamples residuals from the regression analysis in Equation 4. This can be easily accomplished, for example, in the R programming language. For multiplex sibships we should resample in a way that preserves the within-sibship correlation, i.e., by resampling residuals from a single sibship. With care, this can be done with residuals from sibships at least as large as the one we are sampling; for example, we can resample data for a sibship of size 3 by using residuals from a sibship of size 5, as long as we choose (any) three sibs from that sibship and then select the residuals that correspond to their (three) pairings.

Our regression framework in Equation 4 extends easily to allow for modeling covariates at the pair level. Although adjusting for potential covariates in non-randomly ascertained samples is controversial,<sup>47–49</sup> there might be value in such adjustment, especially to untangle the effects of rare variants from known common variants in a gene.<sup>50</sup> To perform this adjustment, we create a variable for each sib pair that counts the number of copies of the minor allele of the common variant possessed by the pair, and we include this count as a covariate in our model. We plan to explore the value of such adjustment more in future work.

Our simulations show that using affected sib trios is even more powerful than using affected sib pairs, even when the sample size is held fixed. Although ascertaining families based on three affected subjects is rare, some studies do employ this strategy when collecting extended pedigrees.<sup>6,51</sup> This motivates an extension of our approach to handle data comprised of affected relative pairs of arbitrary relatedness (e.g., first cousins, avuncular). For each type of affected relative pair (e.g., affected first cousins), we can implement the same framework discussed earlier by constructing an efficient score function that compares rare variant count to IBD sharing across the pairs. To combine results across different types of affected relative pairs (e.g., a sample comprised of a mixture of affected sibling pairs and affected first-cousin pairs), we can still use our efficient score framework but require a separate intercept parameter for each type of affected relative pair (to control for differences in genotypic correlation between different types of relative pairs). We are unsure whether an analysis based on these distant affected pairs will be more powerful than using case-control data, because second-order (and

higher-order) relatives can share at most only one haplotype identical by descent (with the majority possessing IBD = 0). We intend to explore this in more detail in an additional manuscript.

## Appendix A

### Deriving the Efficient Score Function in Equation 5 from Estimating Equations in Equation 4

Given the estimating equations in Equation 4, define

$$U_{\mu}(\mu_0, \delta) = 4 \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} (T_{ij} - 4\mu_0 - 2\delta Z_{ij}),$$

$$U_{\delta}(\mu_0, \delta) = 2 \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} (T_{ij} - 4\mu_0 - 2\delta Z_{ij}) Z_{ij},$$

$$H_{\mu\mu}(\mu_0, \delta) = \frac{\partial U_{\mu}}{\partial \mu_0}(\mu_0, \delta) = -16 \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij}$$

$$H_{\mu\delta}(\mu_0, \delta) = \frac{\partial U_{\mu}}{\partial \delta}(\mu_0, \delta) = -8 \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} Z_{ij}$$

Then the efficient score for  $\delta$  is defined as

$$U_{\delta}^{\text{eff}}(\delta) = U_{\delta}(\hat{\mu}_0(\delta), \delta) - \frac{H_{\mu\delta}(\hat{\mu}_0(\delta), \delta)}{H_{\mu\mu}(\hat{\mu}_0(\delta), \delta)} U_{\mu}(\hat{\mu}_0(\delta), \delta)$$

where  $\hat{\mu}_0(\delta)$  is the value of  $\mu_0$  that solves  $U_{\mu}(\mu_0, \delta) = 0$  (as a function of  $\delta$ ). Assuming the weights  $W_{ij}$  are normalized, it is straightforward to show that  $\hat{\mu}_0(\delta) = (\sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} (T_{ij} - 2\delta Z_{ij})) / 4$ . Hence, we find

$$U_{\delta}^{\text{eff}}(\delta) = 2 \sum_{j=1}^N \sum_{i=1}^{S_j} W_{ij} (\tilde{T}_{ij} - 2\delta \tilde{Z}_{ij}) \tilde{Z}_{ij},$$

where  $\tilde{T}_{ij}$  and  $\tilde{Z}_{ij}$  are as defined in Equation 5. Dropping the superfluous factor of 2, we thus conclude that  $\mathcal{U}$  given in Equation 5 is proportional to  $U_{\delta}^{\text{eff}}(\delta = 0)$ , the efficient score for  $\delta$  at  $\delta = 0$ . A simple interpretation of the efficient score when estimating functions derived from a likelihood is that the efficient score is the derivative of the profile likelihood.

### Supplemental Data

Supplemental Data include five figures and can be found with this article online at <http://dx.doi.org/10.1016/j.ajhg.2015.01.020>.

### Acknowledgments

We thank Dr. Karen Conneely and two anonymous reviewers for their constructive comments related to this work. This work was supported by HG007508 from the National Human Genome Research Institute. GENOA is supported by HL086694 and HL119443 from the National Heart, Lung, and Blood Institute.

This manuscript has been reviewed by GENOA investigators for scientific content and consistency of data interpretation with previous GENOA publications and significant comments have been incorporated prior to submission for publication. M.P.E. is a consultant for Amnion Laboratories. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Received: October 8, 2014

Accepted: January 30, 2015

Published: March 19, 2015

## Web Resources

The URLs for data presented herein are as follows:

Epstein software, <http://genetics.emory.edu/labs/epstein/software>  
Exome Chip Design, [http://genome.sph.umich.edu/wiki/Exome\\_Chip\\_Design](http://genome.sph.umich.edu/wiki/Exome_Chip_Design)

International HapMap Project, <http://hapmap.ncbi.nlm.nih.gov/OMIM>, <http://www.omim.org/>

Rutgers Map v.3, [http://compngen.rutgers.edu/download\\_maps.shtml](http://compngen.rutgers.edu/download_maps.shtml)

## References

1. Teng, J., and Risch, N. (1999). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* 9, 234–241.
2. Zöllner, S. (2012). Sampling strategies for rare variant tests in case-control studies. *Eur. J. Hum. Genet.* 20, 1085–1091.
3. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur. J. Hum. Genet.* 21, 1158–1162.
4. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* 44, 243–246.
5. Liu, Q., Nicolae, D.L., and Chen, L.S. (2013). Marbled inflation from population structure in gene-based association studies with rare variants. *Genet. Epidemiol.* 37, 286–292.
6. Consortium, E.K.; Epi4K Consortium (2012). Epi4K: gene discovery in 4,000 genomes. *Epilepsia* 53, 1457–1467.
7. Krumm, N., O’Roak, B.J., Karakoc, E., Mohajeri, K., Nelson, B., Vives, L., Jacquemont, S., Munson, J., Bernier, R., and Eichler, E.E. (2013). Transmission disequilibrium of small CNVs in simplex autism. *Am. J. Hum. Genet.* 93, 595–606.
8. Daniels, P.R., Kardia, S.L., Hanis, C.L., Brown, C.A., Hutchinson, R., Boerwinkle, E., and Turner, S.T.; Genetic Epidemiology Network of Arteriopathy study (2004). Familial aggregation of hypertension treatment and control in the Genetic Epidemiology Network of Arteriopathy (GENOA) study. *Am. J. Med.* 116, 676–681.
9. Chen, H., Meigs, J.B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* 37, 196–204.
10. Jiang, D., and McPeck, M.S. (2014). Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet. Epidemiol.* 38, 10–20.
11. Fang, S., Sha, Q., and Zhang, S. (2012). Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genet. Epidemiol.* 36, 499–507.
12. Jiang, Y., Conneely, K.N., and Epstein, M.P. (2014). Flexible and robust methods for rare-variant testing of quantitative traits in trios and nuclear families. *Genet. Epidemiol.* 38, 542–551.
13. Schaid, D.J., McDonnell, S.K., Sinnwell, J.P., and Thibodeau, S.N. (2013). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet. Epidemiol.* 37, 409–418.
14. Zhu, Y., and Xiong, M. (2012). Family-based association studies for next-generation sequencing. *Am. J. Hum. Genet.* 90, 1028–1045.
15. Perdry, H., Müller-Myhsok, B., and Clerget-Darpoux, F. (2012). Using affected sib-pairs to uncover rare disease variants. *Hum. Hered.* 74, 129–141.
16. Garner, C. (2011). Confounded by sequencing depth in association studies of rare alleles. *Genet. Epidemiol.* 35, 261–268.
17. 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.
18. 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.
19. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5, e1000384.
20. 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.
21. 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.
22. 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.
23. 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.
24. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., and Lin, X.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237.
25. Lander, E.S., and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* 84, 2363–2367.
26. Abecasis, G.R., and Wigginton, J.E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am. J. Hum. Genet.* 77, 754–767.
27. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30, 97–101.
28. Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics* 56, 645–646.

29. Zhang, D., and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* 4, 57–74.
30. Davies, R. (1980). Algorithm as 155: The distribution of a linear combination of chi-2 random variables. *J. R. Stat. Soc. [Ser A]* 29, 323–333.
31. 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.
32. Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* 46, 222–228.
33. Lange, L.A., Lange, E.M., Bielak, L.F., Langefeld, C.D., Kardia, S.L., Royston, P., Turner, S.T., Sheedy, P.F., 2nd, Boerwinkle, E., and Peyser, P.A. (2002). Autosomal genome-wide scan for coronary artery calcification loci in sibships at high risk for hypertension. *Arterioscler. Thromb. Vasc. Biol.* 22, 418–423.
34. Investigators, F.; FBPP Investigators (2002). Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). *Hypertension* 39, 3–9.
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. Matise, T.C., Chen, F., Chen, W., De La Vega, F.M., Hansen, M., He, C., Hyland, F.C., Kennedy, G.C., Kong, X., Murray, S.S., et al. (2007). A second-generation combined linkage physical map of the human genome. *Genome Res.* 17, 1783–1786.
37. Ward, R. (1990). Familial aggregation and genetic epidemiology of blood pressure. In *Hypertension: Pathophysiology, Diagnosis, and Management* (New York: Raven Press), pp. 81–100.
38. Franceschini, N., Fox, E., Zhang, Z., Edwards, T.L., Nalls, M.A., Sung, Y.J., Tayo, B.O., Sun, Y.V., Gottesman, O., Adeyemo, A., et al.; Asian Genetic Epidemiology Network Consortium (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. *Am. J. Hum. Genet.* 93, 545–554.
39. Allison, D.B., Faith, M.S., and Nathan, J.S. (1996). Risch's lambda values for human obesity. *Int. J. Obes. Relat. Metab. Disord.* 20, 990–999.
40. Monda, K.L., Chen, G.K., Taylor, K.C., Palmer, C., Edwards, T.L., Lange, L.A., Ng, M.C., Adeyemo, A.A., Allison, M.A., Bielak, L.F., et al.; NABEC Consortium; UKBEC Consortium; BioBank Japan Project; AGEN Consortium (2013). A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nat. Genet.* 45, 690–696.
41. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006.
42. Jiang, Y., Satten, G.A., Han, Y., Epstein, M.P., Heinzen, E.L., Goldstein, D.B., and Allen, A.S. (2014). Utilizing population controls in rare-variant case-parent association tests. *Am. J. Hum. Genet.* 94, 845–853.
43. Whittemore, A.S., and Tu, I.P. (1998). Simple, robust linkage tests for affected sibs. *Am. J. Hum. Genet.* 62, 1228–1242.
44. Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* 11, 241–247.
45. Kong, A., and Cox, N.J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* 61, 1179–1188.
46. Fox, J. (2002). Bootstrapping regression models. *An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book* (Sage).
47. Kuo, C.L., and Feingold, E. (2010). What's the best statistic for a simple test of genetic association in a case-control study? *Genet. Epidemiol.* 34, 246–253.
48. Mefford, J., and Witte, J.S. (2012). The covariate's dilemma. *PLoS Genet.* 8, e1003096.
49. Pirinen, M., Donnelly, P., and Spencer, C.C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat. Genet.* 44, 848–851.
50. Huyghe, J.R., Jackson, A.U., Fogarty, M.P., Buchkovich, M.L., Stančáková, A., Stringham, H.M., Sim, X., Yang, L., Fuchsberger, C., Cederberg, H., et al. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* 45, 197–201.
51. Ottman, R., Berenson, K., and Barker-Cummings, C. (2005). Recruitment of families for genetic studies of epilepsy. *Epilepsia* 46, 290–297.



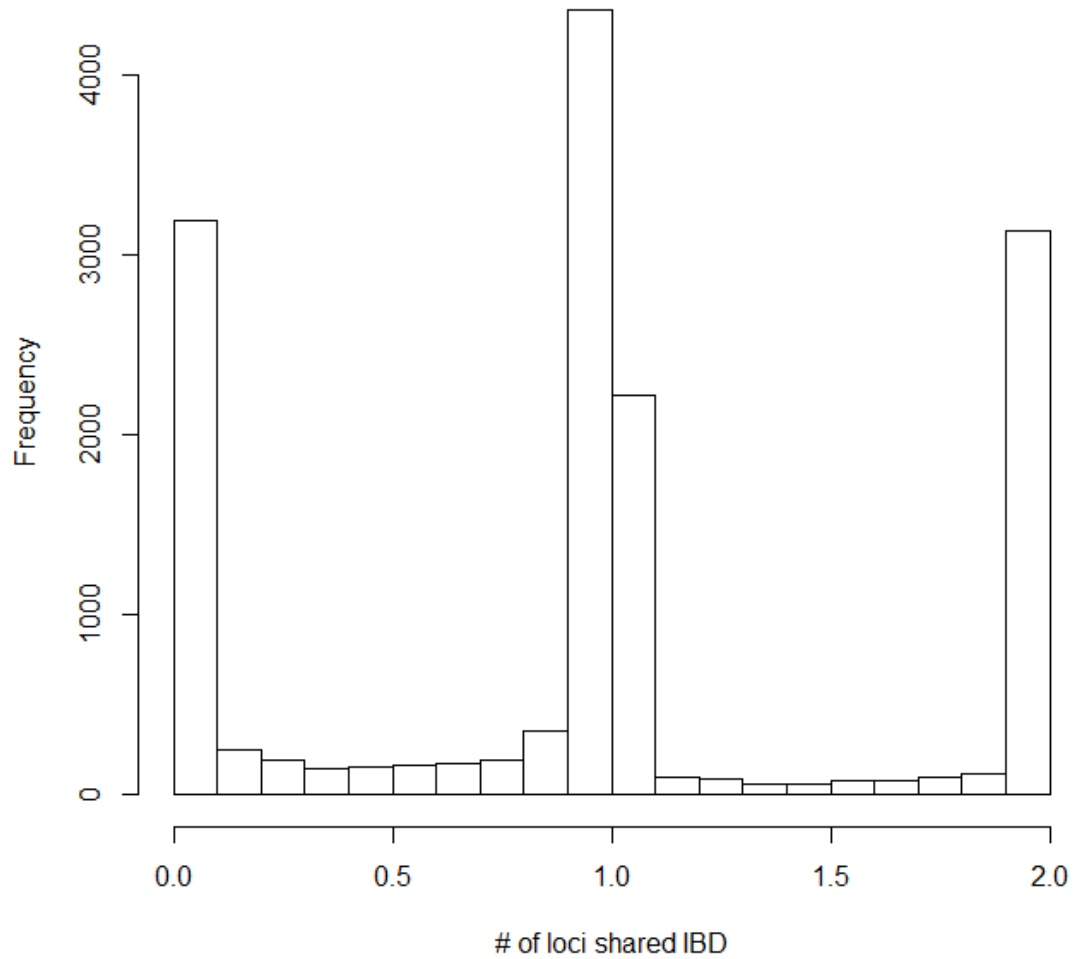
The American Journal of Human Genetics

Supplemental Data

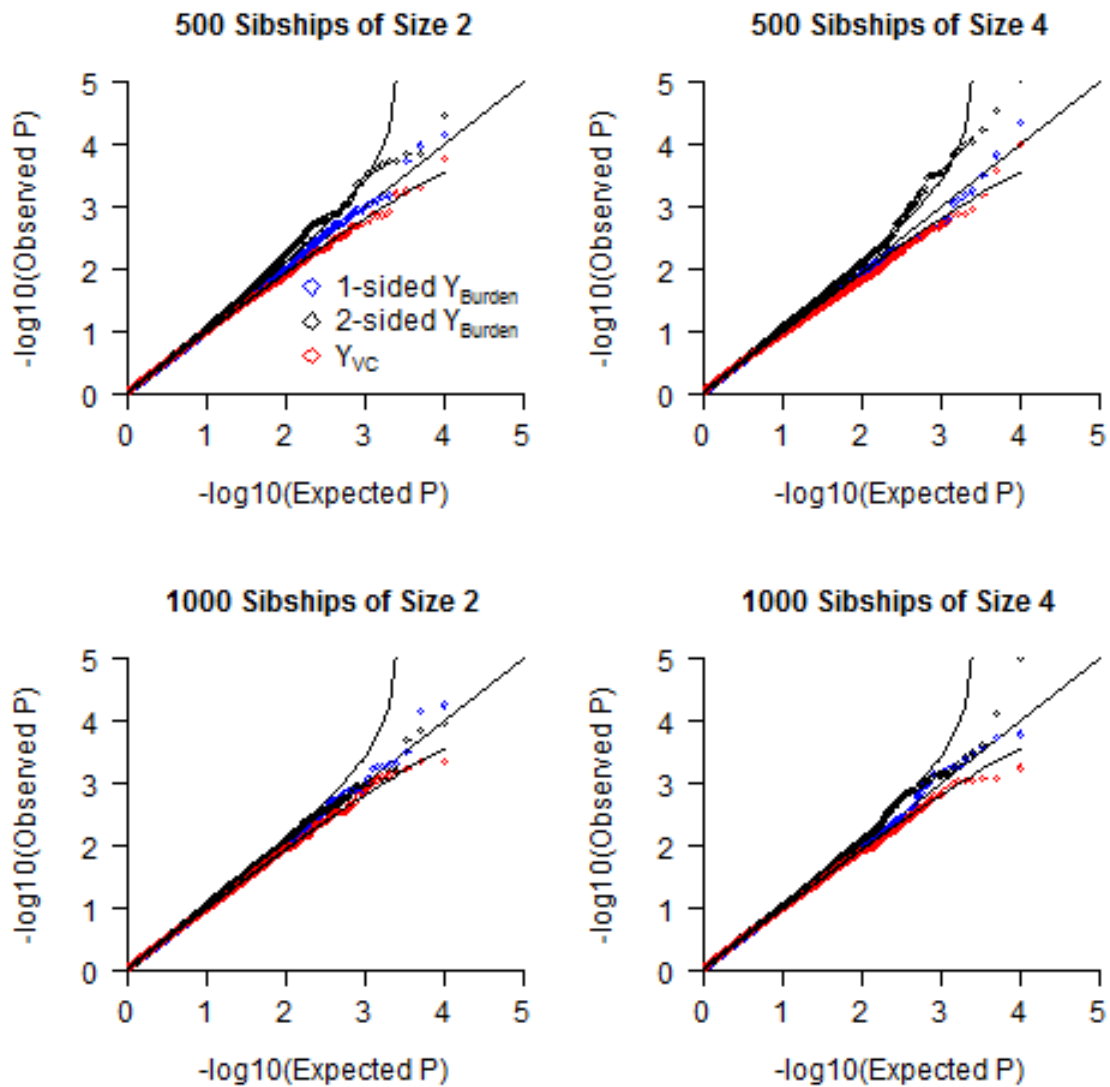
## **A Statistical Approach for Rare-Variant Association Testing in Affected Sibships**

**Michael P. Epstein, Richard Duncan, Erin B. Ware, Min A. Jhun, Lawrence F. Bielak, Wei Zhao, Jennifer A. Smith, Patricia A. Peyser, Sharon L.R. Kardia, and Glen A. Satten**

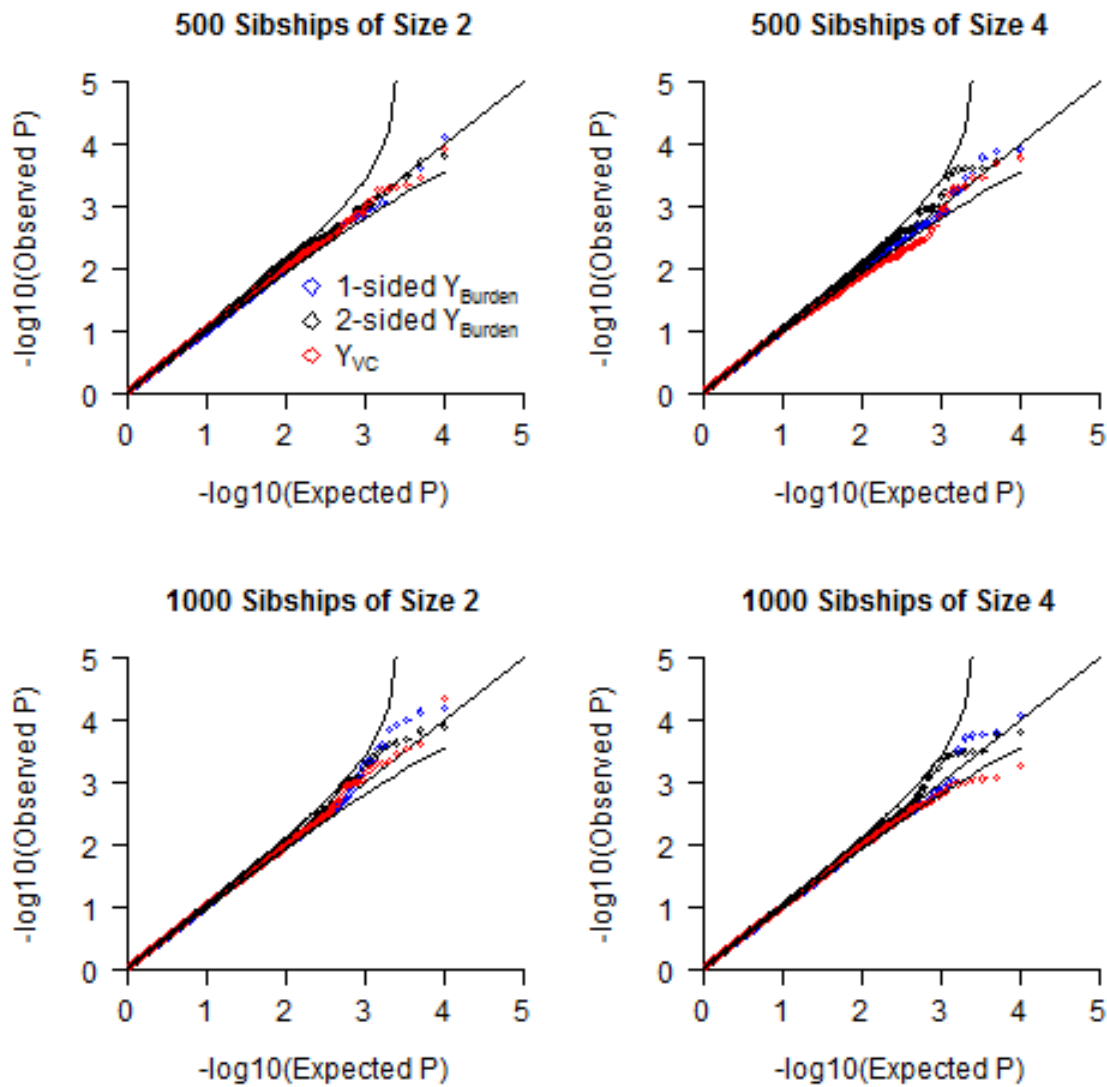
### Pairwise IBD Sharing of Affected Pairs Across Candidate Genes



**Figure S1.** Histogram of pairwise IBD estimates between GENOA affected sibling pairs for candidate genes listed in Table 3. IBD was estimated from MERLIN using 4448 common SNPs on the Illumina HumanExome Beadchip.

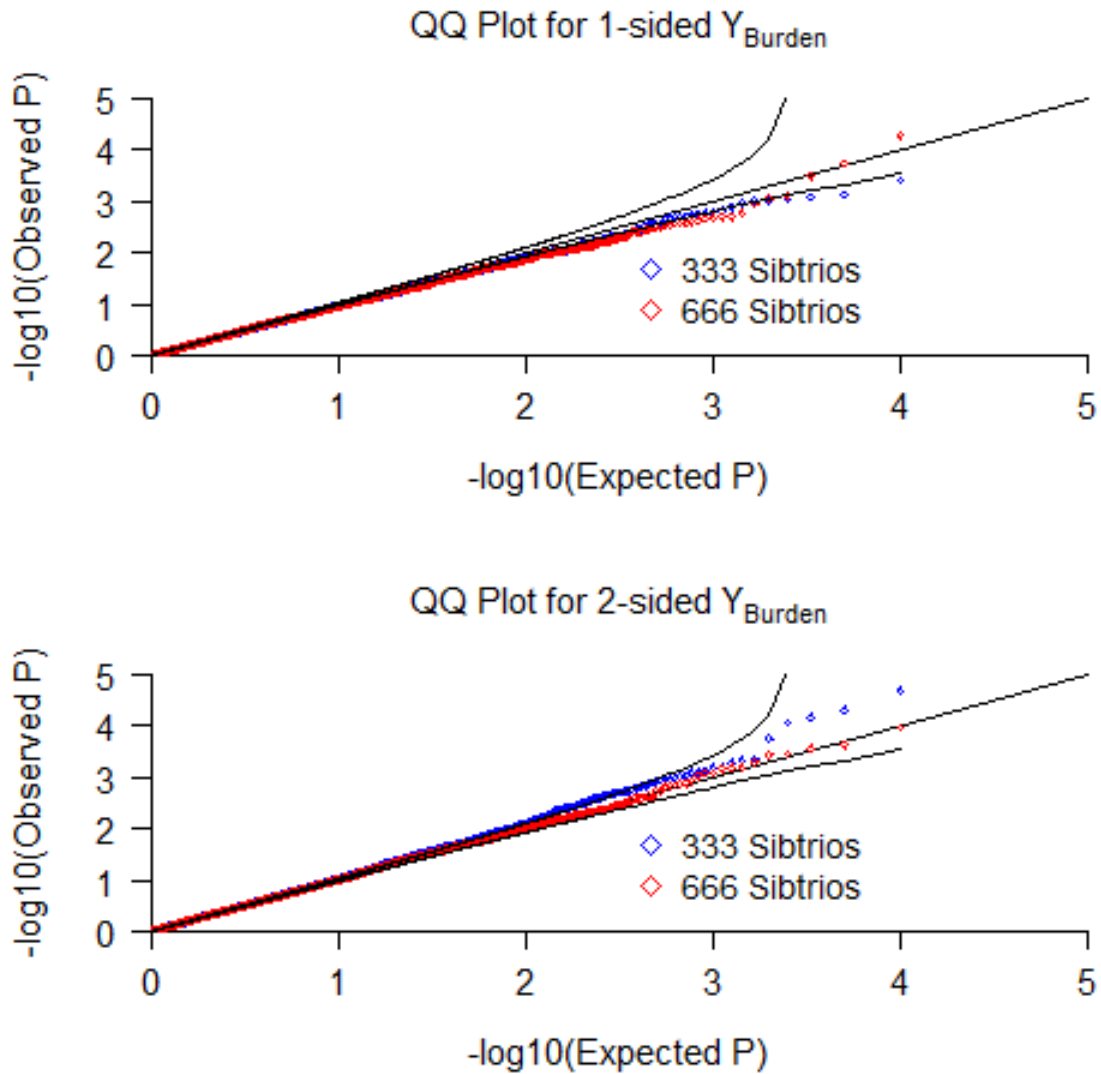


**Figure S2.** Quantile-quantile (QQ) plots of observed p-values versus expected p-values (on  $-\log_{10}$  scale) for the  $Y_{\text{Burden}}$  and  $Y_{\text{VC}}$  tests under the null hypothesis using samples of European ancestry. QQ plots are shown for 500 sibships of size 2 (upper left panel), 500 sibships of size 4 (upper right panel), 1000 sibships of size 2 (lower left panel), and 1000 sibships of size 4 (lower right panel). Each set of results based on 10,000 replications with  $\lambda_C = 8$ .

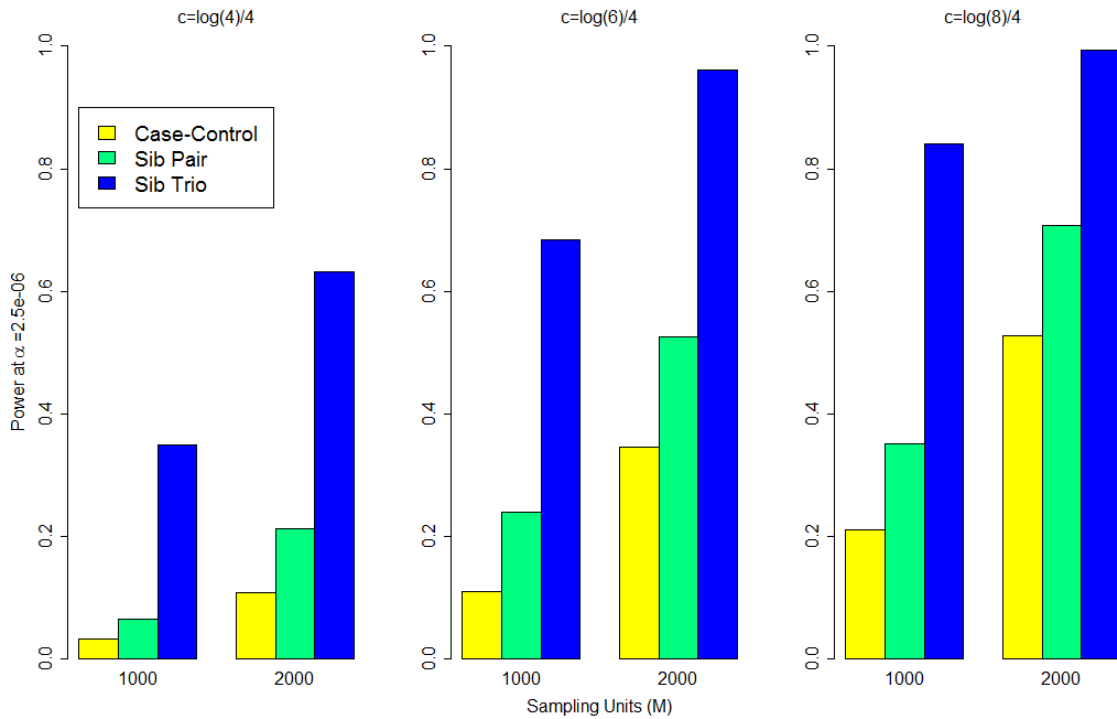


**Figure S3.** Quantile-quantile (QQ) plots of observed p-values versus expected p-values (on  $-\log_{10}$  scale) for the  $Y_{Burden}$  and  $Y_{VC}$  tests under the null hypothesis using samples of both African and European ancestry. Odds ratio of disease for African versus European ancestry is 4. QQ plots are shown for 500 sibships of size 2 (upper left panel), 500 sibships of size 4 (upper right panel), 1000 sibships of size 2 (lower left panel), and 1000 sibships of size 4 (lower right panel). Each set of results based on 10,000 replications with  $\lambda_C = 8$ .





**Figure S4.** Quantile-quantile (QQ) plots of observed p-values versus expected p-values (on  $-\log_{10}$  scale) for the  $Y_{Burden}$  test under the null hypothesis for 333 (blue) or 666 (red) sibtrios comprised of 3 affected siblings. Top panel shows the result of the 1-sided  $Y_{Burden}$  test. Bottom panel shows the result of the 2-sided  $Y_{Burden}$  test. Each set of results based on 10,000 replications with  $\lambda_C = 4$ .



**Figure S5.** Power at  $\alpha = 2.5 \times 10^{-6}$  to detect rare susceptibility variants under different sampling designs assuming a fixed number of  $M$  sampling units. Yellow bar denotes power using  $M/2$  cases and  $M/2$  controls using logistic-regression burden test. Green bar denotes power for  $M/2$  affected sib pairs using 1-sided  $Y_{Burden}$  test. Blue bar denotes power for  $M/3$  affected sibtrios using 1-sided  $Y_{Burden}$  test. Results are shown for  $c = \log(4)/4$  (left panel),  $c = \log(6)/4$  (middle panel), and  $c = \log(8)/4$  (right panel). Each set of results based on 1,000 replications assuming  $\lambda_C = 4$  and 15% of rare variants are causal with 100% of causal variants increasing risk. For 1-sided  $Y_{Burden}$  test, we induced uncertainty in IBD sharing based on the observed information in Figure S1 as described in the Discussion