Rare-Variant Extensions of the Transmission Disequilibrium Test: Application to Autism Exome Sequence Data

Zongxiao He,[1] Brian J. O'Roak,[2,3] Joshua D. Smith,[2] Gao Wang,[1] Stanley Hooker,[1] Regie Lyn P. Santos-Cortez,[1] Biao Li,[1] Mengyuan Kan,[1] Nik Krumm,[2] Deborah A. Nickerson,[2] Jay Shendure,[2] Evan E. Eichler,[2] and Suzanne M. Leal[1,*]

Many population-based rare-variant (RV) association tests, which aggregate variants across a region, have been developed to analyze sequence data. A drawback of analyzing population-based data is that it is difficult to adequately control for population substructure and admixture, and spurious associations can occur. For RVs, this problem can be substantial, because the spectrum of rare variation can differ greatly between populations. A solution is to analyze parent-child trio data, by using the transmission disequilibrium test (TDT), which is robust to population substructure and admixture. We extended the TDT to test for RV associations using four commonly used methods. We demonstrate that for all RV-TDT methods, using proper analysis strategies, type I error is well-controlled even when there are high levels of population substructure or admixture. For trio data, unlike for population-based data, RV allele-counting association methods will lead to inflated type I errors. However type I errors can be properly controlled by obtaining p values empirically through haplotype permutation. The power of the RV-TDT methods was evaluated and compared to the analysis of case-control data with a number of genetic and disease models. The RV-TDT was also used to analyze exome data from 199 Simons Simplex Collection autism trios and an association was observed with variants in *ABCA7*. Given the problem of adequately controlling for population substructure and admixture in RV association studies and the growing number of sequence-based trio studies, the RV-TDT is extremely beneficial to elucidate the involvement of RVs in the etiology of complex traits.

Introduction

Complex-trait rare-variant association studies of exome or whole-genome sequence data have been facilitated by next-generation sequencing (NGS).[1] The vast majority of NGS association studies of complex traits have been population-based studies of qualitative and quantitative traits. However, these studies are vulnerable to population substructure and admixture, which can greatly increase false-positive rates. The observation of spurious associations due to population substructure has been shown to be an even greater problem for rare variants than for common ones.[2] Even for European populations, unlike for common variants, there can be considerable differences in the rare allelic spectrum from one European ethnic group to another. These differences can be even more extreme when studying admixed populations such as African-Americans and Hispanics. Although it has been demonstrated for association studies of common variants in European populations that principal components analysis (PCA) can adequately control for population substructure,[3] it is debated whether PCA is adequate to control for population substructure when rare variants are analyzed.[2] For admixed populations, performing PCA to globally control for population admixture can be insufficient, even for the association analysis of common variants.[4]

For population-based studies in the presence of either population substructure or admixture, spurious associations can be detected as a result of sampling artifacts because of differences in allele frequencies between populations. What is desired is to detect an association due to a difference in the genotype frequencies (e.g., between cases and controls, individuals with high and low quantitative trait values) at the causal variant or variants in linkage disequilibrium (LD) with the causal variant. Family-based analysis can avoid the problem of spurious associations due to population substructure and admixture, and significant findings always imply association due to the causal variant or LD with the causal variant. The study of trios is the most basic family-based design for association testing, using genotype data from an affected proband and his parents. Since the trio design was first proposed by Falk and Rubinstein in 1987 to control for population admixture and substructure,[5] a number of adaptations have been developed including the method that is predominantly used to date, the transmission disequilibrium test (TDT).[6,7] For the TDT, only parents that are heterozygous at the marker locus are informative, and it tests whether or not the frequency of transmitted alleles is the same as the alleles not transmitted to an affected child. The only assumption for the TDT is Mendelian transmission, and an excess of an allele of one type transmitted to the affected offspring indicates a disease-susceptibility

[1]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
[3]Present address: Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR 97239, USA
*Correspondence: sleal@bcm.edu
http://dx.doi.org/10.1016/j.ajhg.2013.11.021.

transmitted	**non-transmitted**		
	M_1	M_2	**total**
M_1	a	b	$a+b$
M_2	c	d	$c+d$
total	$a+c$	$b+d$	$2n$

Figure 1. Two-by-Two Table for the McNemar's Test
Displays the manner in which transmission and nontransmission of the parental minor alleles are counted for the transmission disequilibrium test.

locus for the trait that is both linked and associated with the marker.[8] Both linkage and association between the trait and the marker are required to reject the null hypothesis.[9] This dual-alternative hypothesis protects the TDT from spurious associations where an association is observed but linkage is not present, which can occur in the presence of population admixture and/or substructure.[10] The cost of recruiting probands and their relatives and performing NGS used to be a bottleneck in performing family-based rare-variant association studies. Currently in order to study the role of de novo mutations in genetic diseases, NGS data are being generated for a large number of trios.[11]

The traditional TDT can be used to perform rare-variant association analysis by analyzing single variants. However, it has been shown that association analysis of individual rare variants (minor allele frequency [MAF] $\leq$ 1%) is underpowered,[12] as a result of the small number of observations for a rare variant and a stringent multiple-testing correction. In order to analyze rare variants, many association methods have been developed specifically to enrich the association signal and reduce the multiple-testing penalty. All of these methods group information across multiple variants within a genomic region, which is usually a gene.[13–17] In addition to aggregating rare variants within a region, these methods include (1) weighting each variant by either the frequency in controls[14] or the complete sample[18] or by predicted functionality of the variants[17] and (2) maximizing the test statistic over all variants or variant frequencies.[17–19] These methods can improve power to detect rare-variant associations compared to single variant analysis. Methods that combine the benefits of rare-variant association analysis and family-based tests provide a robust and powerful approach to identify and characterize rare disease-susceptibility variants.

We incorporated four commonly used rare-variant association methods into the TDT framework: Combined Multivariate and Collapsing (CMC),[13] Weighted Sum Statistic (WSS),[14] Burden of Rare Variants (BRV),[20] which is a revised version of Gene- or Region-based Analysis of Variants of Intermediate and Low frequency (GRANVIL),[16] and Variable Threshold (VT).[17] We also compared the power of these methods to the previously described Family Based-Sequence Kernel Association Test (FB-SKAT).[21] By using simulated genetic data, we demonstrate that type I errors are well-controlled for all extended rare-variant (RV)-TDT methods, even when applied to admixed or substructured populations. However, in the presence of LD between variants, there are some caveats in properly controlling type I errors.

The power of the four RV-TDT methods to detect associations varies only slightly and the most powerful method is dependent on the underlying disease model. However, all RV-TDT methods are more powerful than FB-SKAT. The power of the RV-TDT methods were also compared to population-based rare-variant association methods.

In order to further illustrate the application of the RV-TDT methods with NGS data, 199 autism spectrum disorder (ASD) trios from the Simons Simplex Collection were analyzed.[22] ASD, a heterogeneous disorder with substantial heritability, is defined by impaired social communication, deficits in language development, and the presence of restricted interests and/or stereotyped repetitive behaviors.[23] Genome-wide association, de novo mutation, and copy-number-variant studies have reported more than 100 different genes and genomic regions to be associated with this complex trait,[24] but for at least 70% of autism cases the underlying genetic component remains unexplained.[25] This motivates great interest in investigating the role of low-frequency and rare variants in the etiology of ASD. Application of our RV-TDT methods identified an association with rare variants within *ABCA7* (MIM 605414), which encodes ATP-binding cassette subfamily A member 7 protein and might be involved in the etiology of autism.

Material and Methods

Transmission Disequilibrium Test

The TDT was performed in the manner described by Spielman et al.[7] by using a 2 × 2 table to tally all possible transmission events (Figure 1). For the TDT, only the transmission of alleles from parents who are heterozygous is of interest; for those genotypes where the parent is homozygous the meiosis is uninformative. Transmission events from homozygous parents fall into cells a and d, which are not used in the test statistic. The informative meioses from heterozygous parents are where

(1) the minor allele is transmitted to the proband and the major allele is not transmitted, which we define as a minor-allele-transmitted event. These observations are tallied in cell c; and
(2) the major allele is transmitted to the proband and the minor allele is not transmitted, which we define as a major-allele-transmitted event. These observations are tallied in cell b.

The null hypothesis holds when the proportions $b/(b+c)$ and $c/(b+c)$ are comparable with probabilities 0.5 and 0.5 (*i.e.,* $b = c$). The hypothesis is tested by using a 1-degree of freedom asymptotical χ^2 test, McNemar's test,[26] and the χ^2 statistic is defined as

$$\chi^2 = \frac{(b-c)^2}{(b+c)}. \qquad \text{(Equation 1)}$$

For the RV-TDT methods a one-sided test is performed, because only the overtransmission of the minor allele to the affected child is of interest. For all RV-TDT methods, a de novo event is considered to be a minor-allele-transmission event.

Rare-Variant Association Methods

Four commonly used rare-variant association methods are incorporated into the TDT framework to detect the association between rare variants and the phenotype of interest.

TDT-CMC

The CMC method uses an indicator variable to denote the presence or absence of rare variant(s) and tests the association between the phenotype and rare-variant carrier status. For every parent, for each informative variant site we count whether or not a minor-allele-transmitted event occurs. For parent j with variant i, we define indicator variables c_{ij} and b_{ij} as

$$c_{ij} = \begin{cases} 1, & \text{if a minor-allele-transmitted event} \\ & \text{occurs for parent } j \text{ with variant } i \\ 0, & \text{otherwise} \end{cases}$$

$$b_{ij} = \begin{cases} 1, & \text{if a major-allele-transmitted event} \\ & \text{occurs for parent } j \text{ with variant } i \\ 0, & \text{otherwise} \end{cases} \qquad \text{(Equation 2)}$$

Then, for a genetic region L, the total minor-allele-transmitted events and major-allele-transmitted events for parent j are given by

$$c_j = \sum_{i \in L} c_{ij},\ b_j = \sum_{i \in L} b_{ij}. \qquad \text{(Equation 3)}$$

For the TDT-CMC method, for a data set with n trios ($2n$ parents), the c and b quadrants in the 2 × 2 table for gene L above are given by

$$c = \sum_{j=1}^{2n} c_j / \left(b_j + c_j\right),\ b = \sum_{j=1}^{2n} b_j / \left(b_j + c_j\right). \qquad \text{(Equation 4)}$$

Then Equation 1 will be used to attempt to reject the null hypothesis of no linkage or association between the genetic region and the disease. This approach will ensure every informative parent contributes a score of 1 to the McNemar's test. There are a few rare situations where phasing is required for the TDT-CMC, because each informative parent can only contribute a score of 1, for example, if the proband, mother, and father are all heterozygous at the same rare-variant site and the mother is heterozygous at an additional rare variant site. In this situation, if both of the mother's rare variants are on the same haplotype, then she is scored 1 for a major-allele-transmitted event (quadrant b) and the father is scored 1 for a minor-allele-transmitted event (quadrant c). On the other hand, if the mother's rare variants are on different haplotypes, she is scored $^1/_2$ for a major-allele-transmitted event (quadrant b) and $^1/_2$ for a minor-allele-transmitted event (quadrant c), whereas the father is scored 1 for a major-allele-transmitted event (quadrant b). Therefore, for the application of the TDT-CMC, haplotypes must be phased.

TDT-BRV

The TDT-BRV method counts the number of minor-allele-transmitted events and major-allele-transmitted events from every informative parent to the affected proband. Therefore unlike the TDT-CMC where each informative parent can only contribute a score of 1 to the McNemar's test, for the TDT-BRV each informative parent contributes a score that is equivalent to the number of informative sites within the region, e.g., 1, 2, 3.

The same analysis is used as for the TDT-CMC, except Equation 4 above is given in the form of

$$c = \sum_{j=1}^{2n} c_j,\ b = \sum_{j=1}^{2n} b_j. \qquad \text{(Equation 5)}$$

Because each site within an informative region can be counted independently of the other sites, it is not necessary to phase the data before performing the TDT-BRV, when analytical p values are obtained. However to control type I errors in the presence of LD, empirical p values should be estimated via haplotype permutation, which requires phasing the trio data.

TDT-WSS

For the TDT-WSS, each variant site is weighted by the estimated SD of the number of variants in the parental haplotypes that are not transmitted to the offspring

$$\widehat{\omega}_i = \sqrt{n_i \cdot q_i (1 - q_i)}, \qquad \text{(Equation 6)}$$

where q_i is the allele frequency of variant i in parental haplotypes that are not transmitted to the offspring. The remaining analysis is similar to the TDT-BRV, except the c and b in Equation 4 above are given in the form of

$$c = \sum_{j=1}^{2n} c_j = \sum_{j=1}^{2n} \sum_{i \in L} \frac{c_{ij}}{\widehat{\omega}_i},\ b = \sum_{j=1}^{2n} b_j = \sum_{j=1}^{2n} \sum_{i \in L} \frac{b_{ij}}{\widehat{\omega}_i}. \qquad \text{(Equation 7)}$$

Because internal information is applied to obtain the weights, p values must be obtained empirically by using permutation to avoid spurious associations.

TDT-VT

For the TDT-VT, the test statistic is maximized over allele frequencies and therefore a variable allele frequency threshold is applied, instead of a fixed MAF cut-off. The TDT-VT can be implemented by using either the TDT-CMC or TDT-BRV coding. The TDT-VT avoids the implicit assumption about the relationship between allele frequency and odds ratio. Because the test statistic is maximized over allele frequencies, to correct for multiple testing, p values must be obtained empirically.

We also compared the power of the RV-TDT methods to FB-SKAT. FB-SKAT is an extension of the family-based association test (FBAT) to detect rare-variant associations by using a variance component test. More details about this method can be found in Ionita-Laza et al.[21]

Simulation Based on Population Demographic Models

To evaluate type I error rates and the power of the RV-TDT methods, we generated population genetic data by using forward time simulation.[27] Two population demographic models were followed, Kryukov[28] and Boyko.[29] For the Kryukov model, a conventional four-parameter model was used to describe the demographic history (i.e., bottleneck and exponential expansion) of the European population. Purifying selection, which affects the rare-variant site frequency spectrum, is also modeled for nonsynonymous variants. Details on the population genetic model's parameters can be found in Kryukov et al.[28] For the Boyko model, a simple two-epoch and a six-parameter complex bottleneck models were used to model population demographic changes for Africans and Europeans, respectively. Details for these population genetic models can be found in Boyko et al.,[29] and particulars on the choice of parameters have been previously described by Liu and Leal.[30] For each population genetic model, 500 haplotype pools were generated, each consisting of 200,000 haplotypes of 1,500 bp in length, which is the average size of a human gene. This 1,500 bp "gene" represents only the coding regions and consists of sites deemed to be either synonymous or nonsynonymous.

Generation of Trio Data

One haplotype pool is randomly selected for each replicate. The genotypes for the proband are obtained by pairing two randomly drawn haplotypes. When genetic data are generated under the alternative hypothesis, a penetrance model is used to determine whether the inherited pair of haplotypes will cause the proband to be "affected." When an affected proband is obtained, one haplotype is selected to be the maternal haplotype and the other chosen to be the paternal haplotype. The remaining maternal and paternal haplotypes are obtained by randomly selecting two additional haplotypes from the same haplotype pool. When genetic data are generated under the null hypothesis, the same procedure is performed, except that the proband's genotypes are composed solely from two randomly sampled haplotypes.

To generate African and European admixed trios, we generated haplotype pools by using African and European population demographic models, and the haplotypes were sampled from both the African and European pools. The proportion of African and European admixture is determined by the probability that is used to select from either the African or European haplotype pool. Various degrees of population admixture are examined, i.e., 75% African and 25% European, 50% African and 50% European, and 25% African and 75% European, with the assumption of random mating. By using these probabilities, the two proband haplotypes are either selected both from the African pool, both from the European pool, or one haplotype is selected from the European pool and the other from the African pool. Each one of the proband's haplotypes is assigned to a parent and then, by using the admixture probabilities, it is determined whether the second haplotype for each parent should be selected from an African or European haplotype pool.

To generate trios in the presence of population substructure, the proband's haplotypes were constructed by sampling from haplotype pools that were either African or European, and the parents' haplotypes were drawn from a haplotype pool of the same ancestry. Population substructure was created by analyzing together "African" and "European" trios with the proportions 75% African and 25% European, 50% African and 50% European, and 25% African and 75% European.

To evaluate the effect of intermarker LD on type I error, we generated trio haplotypes with perfect LD. To create a pool, we selected 20 haplotypes that have two or more variant sites and no one variant site could be found on more than one haplotype background. Additionally, to evaluate haplotype reconstruction when there is perfect LD and population substructure, two pools were created each with 20 nonoverlapping haplotypes. Haplotypes were drawn from these two pools with either equal probability or 25% of haplotypes were sampled from one pool and 75% from the other haplotype pool.

Generation of Case-Control Data

In order to compare the power of the RV-TDT methods to the analysis of population-based data by using the original versions of the rare-variant association methods, variant data were generated by the Kryukov model. Two haplotypes were sampled from a haplotype pool and by using the penetrance model described below it was determined whether the haplotype pair should be assigned "case" or "control" status. The process was repeated until the desired numbers of cases and controls were obtained.

Generation of Phenotype Data

The disease status for a pair of haplotypes is assigned based upon their multisite genotypes consisting only of rare nonsynonymous variant sites (MAF $\leq$ 1%). Power is evaluated when 100%, 75%, and 50% of the nonsynonymous variant sites are causal. Those sites within the gene region that are nonsynonymous were randomly deemed to be causal based upon the predetermined proportions, whereas the remaining rare-variant sites are noncausal with no phenotypic effect. An odds ratio (OR) > 1 is assigned to each causal variant, and the disease probabilities of all variants within a gene are computed based upon an additive mode of inheritance.[31] Two different disease models were applied, both using a disease prevalence of 1%: the equal-effect model where the ORs of causal variants are constant and the variable-effects model where the ORs of causal variants are inversely correlated with their MAFs. For the equal-effect model, those variant sites that are deemed to be causal were evaluated by using four ORs = 1.8, 2.0, 2.2, and 2.5. For the variable-effects model, those variant sites which were deemed to be causal with the lowest observed allele frequency were assigned OR_{max} while those variant sites with the highest allele frequency were assigned OR_{min}. Interpolation was used to obtain the effect size of all causal variants with allele frequency between the highest and lowest MAF. The power was evaluated for four variable-effects models OR_{min}-OR_{max} = 1.5–2.5, 1.5–3.0, 1.5–3.5 and 1.5–4.0.

Data Analysis

For the TDT-CMC, TDT-BRV, and FB-SKAT, only rare variants with a MAF $\leq$ 1% were analyzed, whereas for the TDT-WSS, TDT-VT-CMC, and TDT-VT-BRV both rare and low frequency variants (MAF $\leq$ 5%) were analyzed. For evaluating the effect of LD on the RV-TDT methods, a MAF of $\leq$ 5% was used for all tests. All variants meeting the MAF criteria were analyzed whether or not they were deemed to be "causal." For TDT-CMC and TDT-BRV, p values were obtained both analytically and empirically, whereas for TDT-WSS, TDT-VT-CMC, and TDT-BRV, p values were only obtained empirically through genotype and haplotype permutation. For the FB-SKAT method, p values were obtained by moment matching approach with 10,000 Monte Carlo simulations. For the analysis of population-based data with the rare-variant association methods BRV, WSS, and VT, p values were obtained empirically, whereas for the CMC, p values were obtained both empirically and analytically.

Permutation

Genotype and haplotype permutation methods were evaluated. For genotype permutation, genotypes are shuffled at every variant site between each parental pair, and then a paternal and maternal haplotype were randomly chosen to form the offspring's genotypes. For haplotype permutation, the haplotypes are phased and then the parental haplotypes within each trio are shuffled, and the offspring's genotypes are obtained by pairing a randomly selected paternal and maternal haplotype.

Haplotype Phasing

For the TDT-CMC, TDT-VT-CMC and haplotype permutation, haplotypes must be phased. In order to evaluate how well haplotypes could be phased in the presence of African and European population admixture, the phase of the generated data was ignored and phasing was performed with BEAGLE.[32] Additionally, haplotype data were generated with perfect LD, and the haplotype phases were ignored and then reconstructed with BEAGLE. By using 10,000 replicates, the proportion of times parental haplotypes could be correctly phased was evaluated.

Evaluating Type I Error and Power

To evaluate type I errors, we generated 20,000 replicates each with 1,500 trios. For RV-TDT methods where p values were obtained empirically, 10,000 permutations were performed. Power was evaluated for an $\alpha = 0.05$. Two thousand replicates were generated to evaluate power for samples of 800, 1,000, 1,200, 1,500, and 1,800 trios, and to obtain p values empirically, 2,000 permutations were performed. Additionally, to compare power for the RV-TDT methods to the analysis of population-based data, 2,000 replicates were generated for three different scenarios: 1,500 trios, 1,500 cases/1,500 controls, and 1,000 cases/1,000 controls. P values were obtained empirically by performing 2,000 permutations.

Application to Autism Data

From the Simons Simplex Collection, 199 autism spectrum disorder trios were analyzed. Previously, 189 of these trios were analyzed to detect de novo events.[25] An additional 10 trios, which have not been described, were also analyzed.

All samples and phenotypic data were collected under the direction of the Simons Simplex Collection by its 12 research clinic sites: Baylor College of Medicine; Children's Hospital Boston and Harvard University; Columbia University; Emory University; McGill University; Vanderbilt University; Yale University; University California Los Angeles (UCLA); University of Illinois at Chicago; University of Michigan; University of Missouri; and University of Washington. Parents consented and children assented as required by each local institutional review board. Participants were de-identified before distribution. Research was also approved by the University of Washington Human Subject Division under nonidentifiable biological specimens/data.

Exome Capture and Genotype-Calling

Genomic DNA was extracted from whole blood.[25] Exomes were captured by using NimbleGen EZ Exome V2.0 and reads were mapped to a custom reference genome assembly (GRC build37). All exomes met the completed criteria of $\geq 8\times$ read depth in 90% of the capture target and $\geq 20\times$ read depth in 80% of capture target. Additional details on exome sequencing of the autism trio data can be found in O'Roak et al.[25]

Quality Control

Variants were selected if they passed the following GATK filters:[33] AB (allele balance for hets [ref/(ref+alt)]) $\leq$ 0.75, HRun (largest contiguous homopolymer run of variant allele in either direction) $\leq$5.0, QD (variant confidence/quality by depth) $\geq$5, SB (strand bias) ≤ -0.10, QUAL (sequencing quality) >30, and SnpCluster (at least 3 variants clustered within 10 bp).[34] Variant Association Tools (VAT) software was used to remove genotypes with a read depth <10× and also to select variants for analysis. Gene regions were assigned based upon RefSeq definition and ANNOVAR[35] was used to annotate variant sites. Only variant sites that were either nonsynonymous or putative splice site variants were analyzed.

Phasing Trios Data

Before phasing, Mendelian inconsistencies were identified and removed with the PLINK software.[36] Phased genotypes were obtained with BEAGLE software.[32] For missing genotype data, BEAGLE imputes missing data and only provides the most likely genotype. We observed that analyzing the most likely genotype can increase false-positive rates for trio data (data not shown). Therefore the imputed variant calls were removed from the analysis. Additionally to avoid spurious associations, those regions of the exome containing copy-number variants or pseudogenes were removed from the analysis. Genes on the autosomal chromosomes with $\geq$4 variant sites were analyzed, and 8,441 genes were included in the analysis.

Results

Evaluation of Type I Error

Type I error rates were estimated by the proportion of replicates with p values $\leq$ 0.05 or $\leq$ 0.005. Additionally, Quantile-Quantile (QQ) plots were generated. When the data were generated without LD, no type I error inflation was observed for any of the RV-TDT methods (Table 1; see Figure S1 available online). For both the TDT-CMC and TDT-BRV when p values were obtained analytically, the type I error is well-controlled and the p values are slightly conservative (Table 1; Figures S1A and S1D). Likewise, when p values were obtained empirically through either genotype or haplotype permutation, type 1 error is well-controlled (Table 1; Figure S1). However, when the haplotypes with perfect intermarker LD, i.e., $r^2 = 1$ were analyzed, extreme inflation in type I error was observed (Table 1; Figure 2) under several conditions. For p values that were obtained analytically for the TDT-BRV, type I error is extremely inflated, but for the TDT-CMC, the type I errors are well-controlled and even slightly conservative (Table 1; Figure 2A). For example, for the TDT-BRV when haplotypes with the variants in perfect LD were analyzed, an α level of 0.05 has a type I error rate of 0.20. This inflation of type I error is not resolved through genotype permutation (Figure 2B). Haplotype permutation resolves the problem and type I error is well-controlled (p value = 0.05) (Figure 2C). For the TDT-WSS, TDT-VT-CMC, and TDT-VT-BRV, p values must be obtained empirically to properly control type I errors, and when there is intermarker LD, although genotype permutation leads to inflated type I errors (Table 1; Figure 2B), haplotype permutation properly controls type I error rates (Table 1; Figure 2C).

To demonstrate that the RV-TDT methods can adequately control for population admixture, admixed African and European populations were generated following the Boyko population demographic models using different ratios of African and European admixture. For all methods, haplotype permutation provided

Table 1. Type I Error for RV-TDT Methods at α Levels of 0.05 and 0.005

		Models					
		Kyrukov		**Boyko**		**LD**	
Method		**α = 0.05**	**α = 0.005**	**α = 0.05**	**α = 0.005**	**α = 0.05**	**α = 0.005**
TDT-BRV	Analytical[a]	0.0493	0.0032	0.0468	0.0033	**0.1996**[b]	**0.0946**
	Genotype[c]	0.0427	0.0042	0.0396	0.0034	**0.1992**	**0.0934**
	Haplotype[d]	0.0432	0.004	0.0396	0.0034	0.051	0.0033
TDT-CMC	Analytical	0.0492	0.0032	0.0467	0.0033	0.0472	0.0034
	Genotype	0.043	0.004	0.0397	0.0035	**0.1082**	**0.0265**
	Haplotype	0.0431	0.0039	0.0396	0.0034	0.0519	0.0041
TDT-VT-BRV	Genotype	0.0448	0.0037	0.0423	0.0039	**0.2795**	**0.1269**
	Haplotype	0.0445	0.0039	0.0424	0.0038	0.0487	0.004
TDT-VT-CMC	Genotype	0.0447	0.0038	0.0423	0.004	**0.1086**	**0.0188**
	Haplotype	0.0442	0.0039	0.0425	0.0039	0.0445	0.0047
TDT-WSS	Genotype	0.0459	0.0042	0.0441	0.0038	**0.1846**	**0.081**
	Haplotype	0.0449	0.0041	0.0446	0.0037	0.0509	0.0041

		Proportion of African and European Admixture					
		0.75/0.25		**0.5/0.5**		**0.25/0.75**	
Method		**α = 0.05**	**α = 0.005**	**α = 0.05**	**α = 0.005**	**α = 0.05**	**α = 0.005**
TDT-BRV	Haplotype	0.0509	0.0044	0.0489	0.0047	0.0483	0.0050
TDT-CMC	Analytical	0.0416	0.0040	0.0463	0.0043	0.0481	0.0037
	Haplotype	0.0470	0.0041	0.0483	0.0043	0.0504	0.0047
TDT-VT-BRV	Haplotype	0.0491	0.0058	0.0483	0.0051	0.0518	0.004
TDT-VT-CMC	Haplotype	0.0495	0.0055	0.0489	0.0054	0.0520	0.0042
TDT-WSS	Haplotype	0.0519	0.0050	0.0498	0.0056	0.0527	0.0048

[a] p values obtained analytically.
[b] Inflated type I errors are highlighted in bold font.
[c] p values obtained with 10,000 genotype permutations.
[d] p values obtained with 10,000 haplotype permutations.

control of type I errors. Also for the TDT-CMC, when p values were obtained analytically, type I error was well-controlled (Table 1; Figure 3). Similar results were obtained for substructured populations (data not shown). These findings strongly support that RV-TDT methods are robust to both population admixture and substructure.

For all methods for which type I error is well-controlled, haplotypes need to be reconstructed. We examined the ability to reconstruct haplotypes for trio data, where information from the proband can aid in the reconstruction. In the presence of population admixture, even with no LD, when a ratio of 50% African and 50% European was used, 99.95% (SD 0.12%) of the haplotypes could be reconstructed correctly. Likewise, when the ratio was changed to 75% African and 25% European, 99.93% (SD 0.23%) of the haplotypes were correctly reconstructed. The results were very similar when data were generated under population substructure. For example, for population substructure where 50% of the population was African and 50% European, 99.88% (SD 0.40%) of the haplotypes were correctly reconstructed.

Power of RV-TDT Methods

The power of the RV-TDT methods were evaluated for a variety of sample sizes and effect sizes, and also compared to the power of FB-SKAT. The difference in power between the RV-TDT methods is small, although TDT-BRV and TDT-CMC are slightly more powerful than other methods under the equal-effect model (Figure 4; Figure S2); however, there is a clear difference in power between the RV-TDT methods and FB-SKAT, with the RV-TDT methods being considerably more powerful. The power of the RV-TDT methods, as a function of genetic effect size and sample size, is shown in Figure 4 and Figure S3, respectively. For example, when 75% of all nonsynonymous rare-variant sites were causal and the OR for causal variants is 2.5, the power of TDT-CMC and TDT-BRV is 71.15% and

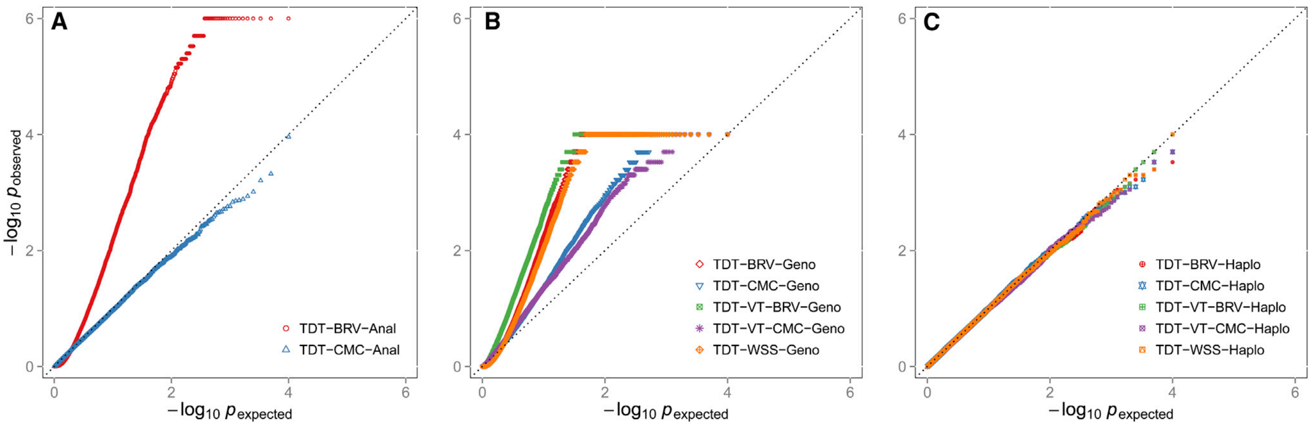

Figure 2. QQ Plot of Negative Natural Log p Values Obtained for Trio Data under the Null Hypothesis of No Association when the Variant Sites that Are Tested Are in Perfect LD

For each scenario, a total of 1,500 trios were analyzed and 20,000 replicates were generated. For the TDT-CMC and TDT-BRV, variants with MAF ≤ 1% were analyzed while for the TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS, variants with MAF ≤ 5% were analyzed.
(A) Displays the results for the TDT-BRV and TDT-CMC when p values were obtained analytically (Anal).
(B) Displays the results for the TDT-BRV, TDT-CMC, TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS. All p values were obtained empirically by performing 10,000 genotype (Geno) permutations for each replicate.
(C) Displays the results for the TDT-BRV, TDT-CMC, TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS. All p values were obtained empirically by performing 10,000 haplotype (Haplo) permutations for each replicate.

71.10%, respectively, whereas the power for TDT-WSS, TDT-VT-CMC, and TDT-VT-BRV is 67.25%, 66.00%, and 66.10%, respectively. For the same scenario, the power for FB-SKAT is 53.90% (Figure 4E). When 75% of the variant sites are causal and the OR for causal variant is between 1.5 and 4.0, the power of TDT-CMC, TDT-BRV, TDT-WSS, TDT-VT-CMC, and TDT-VT-BRV is 68.25%, 68.05%, 70.25%, 70.05%, and 70.20%, respectively, while for the same scenario the power for FB-SKAT is 35.88% (Figure 4B).

To further evaluate the power of the RV-TDT methods, we compared their power to the corresponding rare-variant association methods for population-based data. In Figure 5, the comparison of the power between the TDT-BRV and BRV is shown, used to analyze trio and case-control data, respectively. As previously observed for the analysis of common variants per genotyped individuals, the case-control design is slightly more powerful than the trio design. For example, when 75% of all nonsynonymous rare-variant sites are causal with OR = 2.0, the power of BRV when 1,500 cases and 1,500 controls were analyzed is 45.85% and the power of TDT-BRV when 1,000 trios were analyzed is 43.25%. However, if only 1,000 cases and 1,000 controls are analyzed using the BRV, the power is 33.00% (Figure 5B). Similar results were observed for the other TDT extensions (data not shown).

Applications to Autism Data Set

We applied the RV-TDT methods to analyze 199 trios from the Simons Simplex collection that had available whole-exome sequence data. The QQ plots for TDT-CMC, TDT-BRV, TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS indicate that there is no inflation of type I error (Figure S3). None of the detected associations meet exome-wide significance of 5.92×10^{-6}, i.e., an α level of 0.05 Bonferroni corrected for testing 8,441 genes. *ABCA7* showed the strongest evidence of being associated with autism (OR = 8.5 ± 0.75 stdev) with all RV-TDT methods.[37] Results from the TDT-RV methods were similar: TDT-BRV (p value = 1.4×10^{-4}), TDT-CMC (p value = 1.6×10^{-4}), TDT-CMC (analytical) (p value = 2.9×10^{-4}), TDT-VT-BRV (p value = 1.5×10^{-4}), TDT-VT-CMC (p value = 2.3×10^{-4}), and TDT-WSS (p value = 2.8×10^{-4}). None of the variants that were observed in *ABCA7* are de novo events. Ten missense variants in *ABCA7* with MAF ≤ 1% were observed in 18 trios with a total of 19 minor alleles observed in the parental generation (Table 2). There were 17 minor-allele-transmitted events and 2 major-allele-transmitted events. For seven missense variants, only a single minor-allele-transmitted event was observed. Three missense variants had multiple transmission events: c.2629G>A (p.Ala877Thr) had a minor-allele-transmitted event in five trios and a major-allele-transmitted event in one trio, c.5435G>A (p.Arg1812His) had a minor-allele-transmitted event in three trios and a major-allele-transmitted in one trio, and c.4795G>A (p.Val1599Met) had a minor-allele-transmitted event in two trios (Table 2). Only in one trio, two transmission events were observed in *ABCA7*, c.1534C>G (p.Arg512Gly), and c.4795G>A (p.Val1599Met). Of the ten missense variant sites, five occurred at conserved nucleotides (both PhyloP and GERP scores > 1) and were deemed damaging by at least three of four bioinformatics tools, and therefore could be potentially causal (Table 2). Three of the damaging missense variants were observed in the NHLBI GO Exome Sequencing Project (ESP)[38] with MAF 0.0002-0.004, while two variants are not previously reported in

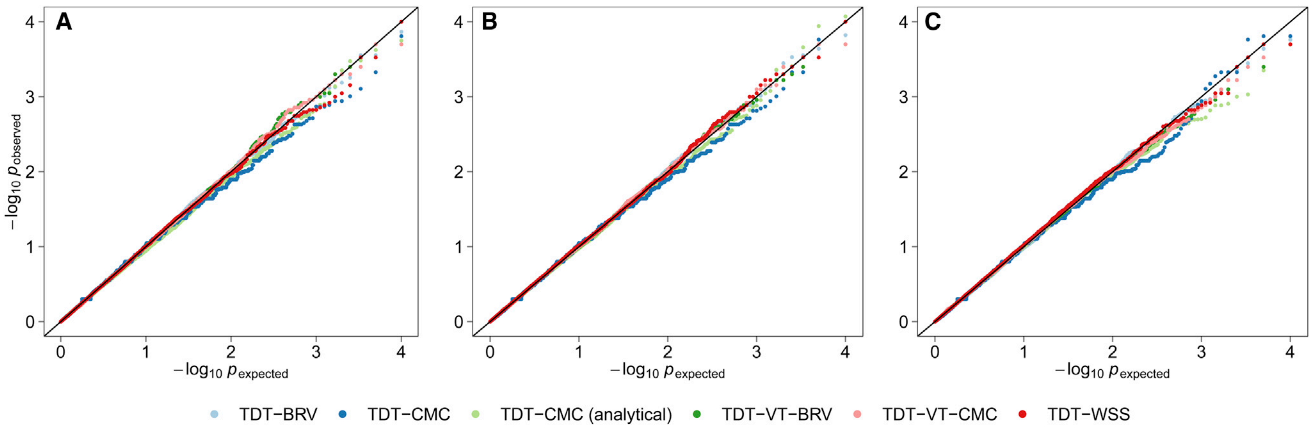

Figure 3. QQ plot of p Values Obtained from the Analysis of African and European Admixed Populations
Genetic variant data for African and European populations were generated under the Boyko model. A total of 1,500 trios were analyzed using 20,000 replicates. Type I error rates were evaluated for the TDT-BRV, TDT-CMC, TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS. For the TDT-CMC and TDT-BRV, variants with a MAF ≤ 1% were analyzed while for the TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS, variants with MAF ≤ 5% were analyzed. All p values were obtained empirically by performing 10,000 haplotype permutations for each replicate, except for the TDT-CMC analytical. The data were generated with different proportions of African and European admixture: in (A) 75% African and 25% European, (B) 50% African and 50% European, and (C) 25% African and 75% European.

publically available databases including 1000 Genomes.[39] For the additional five *ABCA7* variants, only one was not previously reported in publically available databases (Table 2).

Discussion

In this work, we incorporated rare-variant association analysis into the TDT framework to analyze sequence data, in particular rare variants. The simulation results demonstrate that our RV-TDT methods are robust to both population substructure and admixture, which highlights the potential benefits of their application to the analysis of sequence data. Current methods to control for population substructure and admixture might not be sufficient to avoid spurious associations when analyzing rare variants, in particular for admixed populations such as African-Americans and Hispanics.[2,4] The RV-TDT framework can control for both admixture and substructure and thus avoid spurious associations. Additionally proper control of population substructure and admixture can also decrease type II error and lead to an increase in power.

Although the accuracy of NGS technologies has greatly improved, there is still ~1% false-positive call rate even for high read-depth sequence data.[38] An additional advantage of analyzing trio data is that it is possible to improve the accuracy of variant calls, by using variant callers that make use of family or trio information.[40,41] The increased precision in variant calls can in turn lead to increased power to detect associations.

BEAGLE was used to phase the simulated and autism trio data. Other programs could have been used to accurately phase trio data including PHASE (v2.1)[42] and Shape-It.[43] We demonstrate that phasing is quite accurate for trio data even when “admixed” data were generated. It has also been demonstrated by others that phasing of haplotypes is considerably more accurate for trio data than population-based data.[44] For example, when PHASE (v2.1) was used, the percentages of genotypes whose phase was correctly inferred was 99.8% for simulated trio data and 99.95% for the HapMap Centre d’Etude du Polymorphisme Humain (CEPH) trio data, whereas for unrelated individuals, haplotype phasing was correctly inferred 94.8% for the simulated data and 94.1% for CEPH data.[44] Therefore, for trio data it is possible to obtain highly accurate haplotype information.

When p values are analytically obtained for TDT-BRV, it is not necessary to phase the data and additionally for genotype permutation, phasing of the haplotypes is not necessary. We demonstrate that although these methods adequately control type I error when there is no intermarker LD, there can be serious inflation of type I error in the presence of LD. For those methods, which require empirical p values, e.g., TDT-WSS and TDT-VT, haplotypes must be permuted because permuting genotypes leads to an increase in type I errors. When population-based data are analyzed with the BRV method, analytical p values have well-controlled type I errors. Conversely for trio data analyzed with the TDT-BRV, analytical p values have inflated type I errors; however, for empirical p values obtained via haplotype permutation, type I errors are well-controlled.

For rare variants it is usually assumed that there are only low levels of LD, because it is unlikely that rare variants fall on the same haplotype background. However, for the initial analysis of the autism trio data, we detected associations with several genes with TDT-BRV and genotype permutation for which no significant association was discerned by using the TDT-CMC or haplotype permutation.

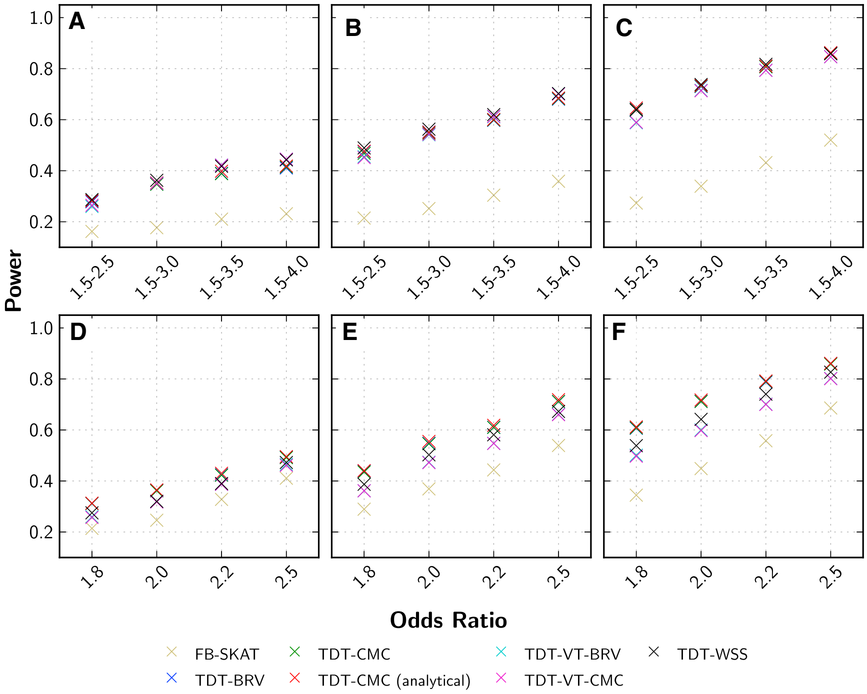

Figure 4. Comparison of Power for the RV-TDT Methods and FB-SKAT

Power was evaluated for an α level of 0.05 for 1,500 trios by generating 2,000 replicates. Analysis was performed with TDT-BRV, TDT-CMC, TDT-VT-BRV, TDT-VT-CMC, TDT-WSS, and FB-SKAT. For the TDT-CMC, TDT-BRV, and FB-SKAT, variants with a MAF $\leq$ 1% were analyzed while for the TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS, variants with MAF $\leq$ 5% were analyzed. For the TDT-BRV, TDT-CMC, TDT-VT-BRV, TDT-VT-CMC, and TDT-WSS, p values were obtained empirically by performing 2,000 haplotype permutations for each replicate. For the TDT-CMC, p values were also obtained analytically. For the FB-SKAT, p values were obtained with a moment matching approach by using 10,000 Monte Carlo simulations. Genetic variant data were generated under the Kryukov model and the proband's affection status was obtained with two different penetrance models: variable-effects model (A, B, C) and equal-effect model (D, E, F). Different proportions of the variant sites were deemed to be causal: (A and D) 50%, (B and E) 75%, and (C and F) 100%.

Upon closer inspection we observed that these associations were driven by multiple rare variants that all lay upon the same haplotype. When analytical or empirical p values via genotype permutation are obtained, each variant is treated as an independent event, but in the presence of LD, this is not the case. This led to an additional investigation on the effects of intermarker LD on type I error and the demonstration that analytical p values for the TDT-BRV, and empirical p values obtain through genotype permutation, which breaks down the LD structure, are not robust to intermarker LD and therefore should not be used. Only haplotype permutation retains the LD structure when used to obtain empirical p values and therefore properly controls type I errors.

We demonstrate that the rare-variant case-control design with an equal number of cases and controls is generally slightly more powerful than the trio design if an equal number of individuals are analyzed, e.g., 1,000 trios versus 1,500 cases and 1,500 controls. However, if an equivalent sample size of cases is analyzed, the power for the RV-TDT methods is slightly higher than the population-based design, e.g., 1,000 trios versus 1,000 cases and 1,000 controls. Additionally, for trio design, only the proband must be phenotyped, which is equivalent to one-third of the study participants, whereas for a case-control design, all study participants should be phenotyped.

A disadvantage of the trio design is that it is not usually suitable for late-onset diseases, because parents will often be deceased and no longer available for study. Additionally, nonpaternity can reduce the power to detect associations, because genotype data for the biological fathers will not be available. Its distinct advantages include: control of type I error in the presence of population substructure and admixture and the ability to investigate parent-of-origin effects. These benefits make the family-based design an excellent choice for sequence-based genetic studies, in particular for early-onset diseases. An additional advantage of using the family-based design is that both inherited and de novo events can be studied and tested simultaneously using the RV-TDT methods.

If it is of interest to detect an association with either protective or detrimental variants, although less powerful than a one-sided test, a two-sided test should be performed. If protective variants are involved in disease etiology, there is an undertransmission of minor alleles to the affected proband. In the traditional implementation of the TDT, parents are not phenotyped and some of the parents might manifest the phenotype of interest, thus reducing the power to detect an association with protective variants. If trios have been ascertained to detect de novo events where the parents have been screened to ensure they are unaffected, it could be the case that they harbor protective variants that prevent them from being diseased or they might not have the correct combination of causal variants and\or environmental exposures to induce the phenotype. Another scenario is that both protective and detrimental variants within the same gene are involved in disease etiology. It has been previously shown that in these situations, variance component tests such as SKAT can be more powerful than aggregate rare-variant methods. However, variance-component methods are less powerful when the vast majority of variants within a region are either detrimental

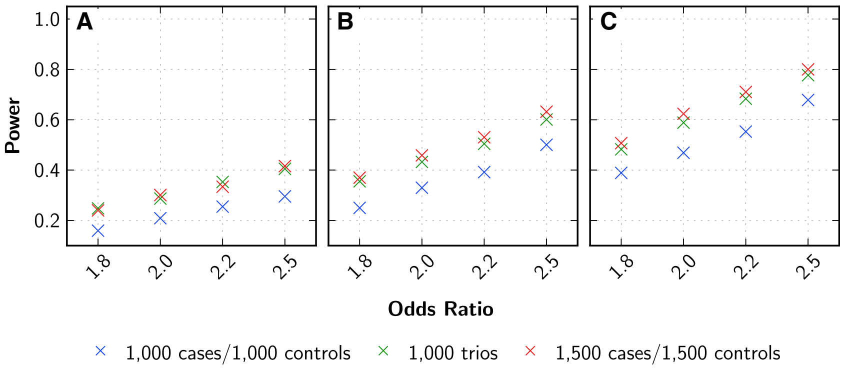

Figure 5. Comparison of Power to Detect Rare-Variant Associations with Population-Based and Trio Data
The BRV was used to analyze samples of size 1,000 cases and 1,000 controls and 1,500 cases and 1,500 controls, and the TDT-BRV was used to analyze 1,000 trios. Power was evaluated for an α level of 0.05 for both case-control and trio data by generating 2,000 replicates. P values were obtained empirically by performing 2,000 haplotype permutations for each replicate. Genetic variant data were generated with Kryukov model. Affection status was determined with an equal-effect penetrance model with ORs varying between 1.8 and 2.5. Different proportions of causal variants were used in the analysis with 50% (A), 75% (B), and 100% (C).

or protective.[45] We also demonstrate here that FB-SKAT is considerably less powerful than the other RV-TDT methods when causal variants within a gene region have an effect that is unidirectional. Additionally, the FB-SKAT did not detect an association between autism and *ABCA7* ($p = 0.11$).

Although the emphasis of this study is performing complex disease association analysis with the proposed RV-TDT methods, the RV-TDT is also suitable for analyzing Mendelian traits. For Mendelian traits, the RV-TDT is particularly beneficial to use when although a family history of disease has been recorded, only the proband and his parents are available for study. The power of the RV-TDT will be dependent on the underlying Mendelian model. Assuming that there is no locus heterogeneity for the disease under study, it is possible to analytically obtain estimates of the necessary sample sizes to detect an association for various Mendelian modes of inheritance. For 80% power, to detect an association for an autosomal recessive trait when $\alpha = 0.05$, four trios are needed, and if an exome-wide significance criterion of $\alpha = 2.5 \times 10^{-6}$ (a Bonferroni correction for testing 20,000 genes) is used, then 15 trios are necessary to detect an association with 80% power. For autosomal-dominant traits for $\alpha = 0.05$ and $\alpha = 2.5 \times 10^{-6}$, 13 trios and 59 trios, respectively, are required to detect an association with 80% power. It is also possible to use the TDT-RV methods to analyze X-linked traits, although only the mother will provide informative meioses. To detect an association for an X-linked recessive trait with a power of 80% for $\alpha = 0.05$, seven trios are necessary, but for $\alpha = 5 \times 10^{-5}$ (Bonferroni correction for testing ~1,000 genes on the X chromosome) 23 trios are necessary.

By using the RV-TDT methods, we identified variants in *ABCA7* (19p13.3) as potentially the underlying cause of autism by analyzing 199 families from the Simons Simplex collection. Previously, an ASD locus was mapped to the 19p13.12 region with a maximum nonparametric LOD (NPL) score of >2.0 in 115 multiplex U.S. families.[46,47] Additionally, by using an extended ASD pedigree consisting of 20 nuclear families from Finland, we obtained an NPL score of 3.57 within chromosome 19p13.3 at marker D19S591,[48] which is 1.9 Mb away from *ABCA7*.

The association between rare variants in *ABCA7* and autism is consistent with the finding that autistic children display abnormal rates of in vivo lipid metabolism compared with healthy controls.[49] ABCA7 is an integral transmembrane ATP–binding cassette transporter that involves the translocation of cellular lipid across membrane, such as cholesterol.[50] Current studies suggest that lipid signaling plays an important role in neuronal processes, such as synaptogenesis and neurotransmitter functions.[51,52] There is increasing evidence suggesting abnormalities of lipid metabolic pathways might affect the nervous system and contribute to autism.[49,53]

Common variants in *ABCA7* have been associated with Alzheimer disease (AD) through brain expression and genome-wide association studies with samples from patients of both African and European descent.[54–56] It has been shown that, like in AD, plasma levels of β-amyloid or α-precursor protein (APP) are significantly elevated in ASD patients.[57] While the occurrence of β-amyloid plaques in the brain is well-known as a pathologic hallmark of AD, the accumulation of β-amyloid in the brains of both pediatric and adult ASD patients was demonstrated only recently.[58] In *Abca7*$^{-/-}$ mice, phagocytotic cells have reduced ability to clear amyloid from the brain, which results in decreased memory and capacity to learn new tasks.[59] The identification of *ABCA7* as a gene that is possibly involved in autism etiology suggests the existence of a common pathway for neurodevelopmental and neurodegenerative diseases that might be targeted for prevention and treatment.

The RV-TDT methods were developed to provide a robust and powerful way to identify rare-variant complex disease associations by using trio sequence data. Given the problem of adequately controlling for population substructure and admixture in rare-variant association studies and the growing number of sequence-based trio studies, the RV-TDT is extremely beneficial in elucidating the involvement of rare variants in the etiology of complex traits. The RV-TDT methods can be used to analyze exome and genome sequence data. Additionally, these methods can be applied

Table 2. Bioinformatic Evaluation and Frequencies of Rare Missense Variants within *ABCA7*

Chr19 Position	Nucleotide Substitution[a]	PhyloP[b]	GERP[c]	Amino Acid Substitution	PolyPhen-2	SIFT	MutationTaster	Mutation Assessor	Transmitted/ Non-transmitted Events	dbSNP rsID	NHLBI-ESP EA MAF[d]	NHLBI-ESP AA MAF[d]
1,043,788[e]	**c.995G>A**	**2.90**	**4.33**	**p.Gly332Glu**	**Probably damaging**	**Damaging**	**Polymorphism**	**Functional, medium**	**1/0**	**NA**	**NA**	**NA**
1,045,109	c.1324G>A	0.28	2.54	p.Gly442Arg	Possibly damaging	Tolerated	Polymorphism	Nonfunctional, low	1/0	NA	NA	NA
1,046,317	c.1534C>G	−0.04	1.11	p.Arg512Gly	Benign	Tolerated	Polymorphism	Neutral	1/0	NA	0.0001	0
1,050,996	c.2629G>A[f]	1.18	2.59	p.Ala877Thr	Benign	Tolerated	Polymorphism	Nonfunctional, low	5/1	rs74176364	0.006	0.003
1,051,481	**c.2858C>A**	**4.96**	**4.43**	**p.Ala953Asp**	**Probably damaging**	**Damaging**	**Disease-causing**	**Functional, high**	**1/0**	**NA**	**NA**	**NA**
1,057,343	**c.4795G>A**[f]	**1.58**	**3.65**	**p.Val1599Met**	**Probably damaging**	**Damaging**	**Polymorphism**	**Functional, medium**	**2/0**	**rs117187003**	**0.004**	**0.0009**
1,058,883	**c.5344C>T**	**1.58**	**3.14**	**p.Arg1782Trp**	**Probably damaging**	**Damaging**	**Disease-causing**	**Functional, medium**	**1/0**	**NA**	**0.0003**	**0**
1,059,056	c.5435G>A[f]	1.28	0.81	p.Arg1812His	Benign	Damaging	Polymorphism	Neutral	3/1	rs114782266	0.005	0.07
1,062,248	**c.5648C>T**	**4.87**	**3.61**	**p.Thr1883Met**	**Probably damaging**	**Damaging**	**Disease-causing**	**Functional, high**	**1/0**	**NA**	**0**	**0.0002**
1,065,305	c.6322G>A[f]	2.08	3.73	p.Glu2108Lys	Benign	Tolerated	Polymorphism	Functional, medium	1/0	rs139706726	0.0002	0

[a] cDNA position is based on reference sequence NM_019112.3.

[b] PhyloP scores indicate nucleotide conservation under a null hypothesis of neutral evolution.

[c] GERP provides position-specific estimates of evolutionary constraint.

[d] Minor allele frequencies (MAF) for European-Americans (EA) and African-Americans(AA) from the NHLBI GO – Exome Sequencing Project (ESP) Exome Variant Server. NA, not available.

[e] Conserved nucleotides (both PhyloP and GERP with scores > 1) and damaging variants (deemed damaging by at least three of four bioinformatics tools) are highlighted in bold font.

[f] Included on the Illumina Human Exome BeadChip.

to analyze rare variants obtained from genotyping arrays including the "exome" chip. To analyze the autism trio data with all five RV-TDT methods, TDT-BRV, TDT-CMC, TDT-WSS, TDT-VT-CMC, and TDT-VT-BRV, obtaining empirical p values based on haplotype permutation took a total of 3.1 hr. To analyze the same data set implementing the TDT-CMC, obtaining analytical p values took 4.5 min. The analysis was performed with a single CPU, however, by using multiple processors for the analysis can greatly decrease the computational time. The RV-TDT software package and documentation are publicly available online.

Supplemental Data

Supplemental Data includes three figures and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, E. Hanson, D. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, and E. Wijsman). We appreciate obtaining access to phenotypic data on SFARI Base. This work was supported by the Simons Foundation Autism Research Initiative (SFARI 137578 and 191889 to E.E.E. and J.S.). E.E.E. is an Investigator of the Howard Hughes Medical Institute. This study was also funded by the National Institutes of Health grants HD065285, HL102926, MD005964, and HG006493.

Received: October 8, 2013
Accepted: November 26, 2013
Published: December 19, 2013

Web Resource

The URLs for data presented herein are as follows:

1000 Genomes, http://browser.1000genomes.org
BEAGLE, http://faculty.washington.edu/browning/beagle/beagle.html
dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/
Exome Variant Server (EVS), http://evs.gs.washington.edu/EVS/
Genome Analysis Toolkit (GATK), http://www.broadinstitute.org/gatk/
MutationAssessor, http://mutationassessor.org/v1/
MutationTaster, http://www.mutationtaster.org/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
PolyPhen-2, http://www.genetics.bwh.harvard.edu/pph2/
RV-TDT, http://bioinformatics.org/rv-tdt/
SIFT, http://sift.jcvi.org/
Simons Foundation Autism Research Initiative (SFARI), http://sfari.org/sfari-initiatives/simons-simplex-collection
Variant Association Tools (VAT), http://varianttools.sourceforge.net/Association/

Accession Numbers

The dbGaP accession number for the exome sequences reported in this paper is phs000482.v1.p.

References

1. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.
2. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
3. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
4. Liu, J., Lewinger, J.P., Gilliland, F.D., Gauderman, W.J., and Conti, D.V. (2013). Confounding and heterogeneity in genetic association studies with admixed populations. Am. J. Epidemiol. *177*, 351–360.
5. Falk, C.T., and Rubinstein, P. (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann. Hum. Genet. *51*, 227–233.
6. Terwilliger, J.D., and Ott, J. (1992). A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum. Hered. *42*, 337–346.
7. Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. *52*, 506–516.
8. Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nat. Rev. Genet. *12*, 465–474.
9. Ott, J. (1989). Statistical properties of the haplotype relative risk. Genet. Epidemiol. *6*, 127–130.
10. Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nat. Rev. Genet. *7*, 385–394.
11. Veltman, J.A., and Brunner, H.G. (2012). De novo mutations in human genetic disease. Nat. Rev. Genet. *13*, 565–575.
12. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. *82*, 100–112.
13. 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.
14. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
15. Liu, D.J., and Leal, S.M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. *6*, e1001156.
16. 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.
17. 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.

18. 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.
19. Bhatia, G., Bansal, V., Harismendy, O., Schork, N.J., Topol, E.J., Frazer, K., and Bafna, V. (2010). A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput. Biol. *6*, e1000954.
20. Auer, P.L., Wang, G., and Leal, S.M. (2013). Testing for rare variant associations in the presence of missing data. Genet. Epidemiol. *37*, 529–538.
21. 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.
22. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192–195.
23. State, M.W., and Levitt, P. (2011). The conundrums of understanding genetic risks for autism spectrum disorders. Nat. Neurosci. *14*, 1499–1506.
24. Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. Brain Res. *1380*, 42–77.
25. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.
26. McNEMAR, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika *12*, 153–157.
27. Hernandez, R.D. (2008). A flexible forward simulator for populations subject to selection and demography. Bioinformatics *24*, 2786–2787.
28. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA *106*, 3871–3876.
29. Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. *4*, e1000083.
30. Liu, D.J., and Leal, S.M. (2012). A unified framework for detecting rare variant quantitative trait associations in pedigree and unrelated individuals via sequence data. Hum. Hered. *73*, 105–122.
31. Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. *46*, 222–228.
32. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
33. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
34. Zook, J.M., Samarov, D., McDaniel, J., Sen, S.K., and Salit, M. (2012). Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. PLoS ONE *7*, e41356.
35. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
36. 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.
37. Kazeem, G.R., and Farrall, M. (2005). Integrating case-control and TDT studies. Ann. Hum. Genet. *69*, 329–335.
38. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
39. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
40. Chen, W., Li, B., Zeng, Z., Sanna, S., Sidore, C., Busonero, F., Kang, H.M., Li, Y., and Abecasis, G.R. (2013). Genotype calling and haplotyping in parent-offspring trios. Genome Res. *23*, 142–151.
41. Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H.M., and Abecasis, G.R. (2012). A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet. *8*, e1002944.
42. Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am. J. Hum. Genet. *76*, 449–462.
43. Delaneau, O., Marchini, J., and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.
44. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., and Donnelly, P.; International HapMap Consortium (2006). A comparison of phasing algorithms for trios and unrelated individuals. Am. J. Hum. Genet. *78*, 437–450.
45. Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M., and Richards, J.B. (2012). The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet. *8*, e1002496.
46. Liu, J., Nyholt, D.R., Magnussen, P., Parano, E., Pavone, P., Geschwind, D., Lord, C., Iversen, P., Hoh, J., Ott, J., and Gilliam, T.C.; Autism Genetic Resource Exchange Consortium (2001). A genomewide screen for autism susceptibility loci. Am. J. Hum. Genet. *69*, 327–340.
47. Buxbaum, J.D., Silverman, J., Keddache, M., Smith, C.J., Hollander, E., Ramoz, N., and Reichert, J.G. (2004). Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. Mol. Psychiatry *9*, 144–150.
48. Kilpinen, H., Ylisaukko-oja, T., Rehnström, K., Gaál, E., Turunen, J.A., Kempas, E., von Wendt, L., Varilo, T., and Peltonen, L. (2009). Linkage and linkage disequilibrium scan for autism loci in an extended pedigree from Finland. Hum. Mol. Genet. *18*, 2912–2921.
49. Tamiji, J., and Crawford, D.A. (2010). The neurobiology of lipid metabolism in autism spectrum disorders. Neurosignals *18*, 98–112.

50. Tanaka, N., Abe-Dohmae, S., Iwamoto, N., and Yokoyama, S. (2011). Roles of ATP-binding cassette transporter A7 in cholesterol homeostasis and host defense system. J. Atheroscler. Thromb. *18*, 274–281.
51. Mellon, S.H., and Griffin, L.D. (2002). Neurosteroids: biochemistry and clinical significance. Trends Endocrinol. Metab. *13*, 35–43.
52. Shackleton, C., Roitman, E., Guo, L.W., Wilson, W.K., and Porter, F.D. (2002). Identification of 7(8) and 8(9) unsaturated adrenal steroid metabolites produced by patients with 7-dehydrosterol-delta7-reductase deficiency (Smith-Lemli-Opitz syndrome). J. Steroid Biochem. Mol. Biol. *82*, 225–232.
53. Buchovecky, C.M., Turley, S.D., Brown, H.M., Kyle, S.M., McDonald, J.G., Liu, B., Pieper, A.A., Huang, W., Katz, D.M., Russell, D.W., et al. (2013). A suppressor screen in Mecp2 mutant mice implicates cholesterol metabolism in Rett syndrome. Nat. Genet. *45*, 1013–1020.
54. Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V., et al.; Alzheimer's Disease Neuroimaging Initiative; CHARGE consortium; EADI1 consortium (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat. Genet. *43*, 429–435.
55. Karch, C.M., Jeng, A.T., Nowotny, P., Cady, J., Cruchaga, C., and Goate, A.M. (2012). Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains. PLoS ONE *7*, e50976.
56. Reitz, C., Jun, G., Naj, A., Rajbhandary, R., Vardarajan, B.N., Wang, L.S., Valladares, O., Lin, C.F., Larson, E.B., Graff-Radford, N.R., et al.; Alzheimer Disease Genetics Consortium (2013). Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E ε4, and the risk of late-onset Alzheimer disease in African Americans. JAMA *309*, 1483–1492.
57. Ray, B., Long, J.M., Sokol, D.K., and Lahiri, D.K. (2011). Increased secreted amyloid precursor protein-α (sAPPα) in severe autism: proposal of a specific, anabolic pathway and putative biomarker. PLoS ONE *6*, e20405.
58. Wegiel, J., Frackowiak, J., Mazur-Kolecka, B., Schanen, N.C., Cook, E.H., Jr., Sigman, M., Brown, W.T., Kuchna, I., Wegiel, J., Nowicki, K., et al. (2012). Abnormal intracellular accumulation and extracellular Aβ deposition in idiopathic and Dup15q11.2-q13 autism spectrum disorders. PLoS ONE *7*, e35414.
59. Kim, W.S., Li, H., Ruberu, K., Chan, S., Elliott, D.A., Low, J.K., Cheng, D., Karl, T., and Garner, B. (2013). Deletion of Abca7 increases cerebral amyloid-β accumulation in the J20 mouse model of Alzheimer's disease. J. Neurosci. *33*, 4387–4394.