*Bioinformatics*, 31(9), 2015, 1452–1459
doi: 10.1093/bioinformatics/btu860
Advance Access Publication Date: 6 January 2015
Original Paper

OXFORD

Genetic and population analysis

# A haplotype-based framework for group-wise transmission/disequilibrium tests for rare variant association analysis

**Rui Chen[1], Qiang Wei[1], Xiaowei Zhan[2], Xue Zhong[3], James S. Sutcliffe[1], Nancy J. Cox[4], Edwin H. Cook[5], Chun Li[6], Wei Chen[7,*] and Bingshan Li[1,*]**

[1]Department of Molecular Physiology and Biophysics, Vanderbilt University, TN, 37221, USA, [2]Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, Dallas, TX, USA, [3]Center for Quantitative Sciences, Vanderbilt University, TN, 37221, USA, [4]Department of Medicine, University of Chicago, Chicago, IL, USA, [5]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA, [6]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA and [7]Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on September 12, 2014; revised on December 7, 2014; accepted on December 23, 2014

## Abstract

**Motivation:** A major focus of current sequencing studies for human genetics is to identify rare variants associated with complex diseases. Aside from reduced power of detecting associated rare variants, controlling for population stratification is particularly challenging for rare variants. Transmission/disequilibrium tests (TDT) based on family designs are robust to population stratification and admixture, and therefore provide an effective approach to rare variant association studies to eliminate spurious associations. To increase power of rare variant association analysis, gene-based collapsing methods become standard approaches for analyzing rare variants. Existing methods that extend this strategy to rare variants in families usually combine TDT statistics at individual variants and therefore lack the flexibility of incorporating other genetic models.

**Results:** In this study, we describe a haplotype-based framework for group-wise TDT (gTDT) that is flexible to encompass a variety of genetic models such as additive, dominant and compound heterozygous (CH) (i.e. recessive) models as well as other complex interactions. Unlike existing methods, gTDT constructs haplotypes by transmission when possible and inherently takes into account the linkage disequilibrium among variants. Through extensive simulations we showed that type I error was correctly controlled for rare variants under all models investigated, and this remained true in the presence of population stratification. Under a variety of genetic models, gTDT showed increased power compared with the single marker TDT. Application of gTDT to an autism exome sequencing data of 118 trios identified potentially interesting candidate genes with CH rare variants.

**Availability and implementation:** We implemented gTDT in C++ and the source code and the detailed usage are available on the authors' website (https://medschool.vanderbilt.edu/cgg).

**Contact:** bingshan.li@vanderbilt.edu or wei.chen@chp.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Next generation sequencing is routinely employed to identify rare variants, e.g. variants with minor allele frequency (MAF) <0.01, associated with complex traits. Although there are examples of studies implicating rare variants in complex diseases/traits (Auer *et al.*, 2014; Cohen *et al.*, 2004; Cruchaga *et al.*, 2014; Ji *et al.*, 2008; Nejentsev *et al.*, 2009; Nelson *et al.*, 2012), it is clear that such studies remain underpowered, and elucidating the contribution of low-frequency variants to disease etiology should benefit from additional methodology advances. Major challenges include the low power of traditional analysis methods and the severe multiple testing problem. For case–control studies, group-wise methods, such as CAST (Morgenthaler and Thilly, 2007) and CMC, collapse rare variants into a single genetic unit—a gene, region or pathway—to test the association of multiple rare variants as a whole. The underlying rationale of collapsing approaches is that if multiple variants at a locus confer disease risk independently, aggregating individual variants strengthens the overall signal and avoids multiplicity in single- and multi-marker tests, resulting in increased power for association analysis of rare variants (Li and Leal, 2008). A more powerful approach is to take a weighted sum of rare variants, with weights reflecting the genetic effects of the corresponding variants (Li *et al.*, 2013; Madsen and Browning, 2009). Such a strategy has the potential to achieve much improved power by assigning weights proportional to the genetic effect of corresponding variants. A variety of methods have been developed with varying ways of selecting and weighting rare variants (Han and Pan, 2010; Li *et al.*, 2013; Neale *et al.*, 2011; Price *et al.*, 2010; Wu *et al.*, 2011) in an effort to increase power of rare variant association studies and new methods continue to emerge (see reviews Asimit and Zeggini, 2010; Li *et al.*, 2013 for more details).

This rationale in group-wise strategy to increase power, however, is able to aggregate and amplify spurious association signals as well, if such undesired signals exist. Such spurious association can be due to population stratification. For common variants, this confounding has been long recognized and can be effectively accounted for by existing methods (Kang *et al.*, 2010; Price *et al.*, 2006). Most rare variants are relatively young and often show differential allelic spectra in populations traditionally considered as homogeneous; 85% of rare variants are population specific (Tennessen *et al.*, 2012). Population stratification effects due to rare variants can be stronger than that due to common variants, and methods routinely used for controlling population stratification in association studies of common variants may fail for rare variants (Mathieson and McVean, 2012). Family-based analysis, such as the transmission/disequilibrium test (TDT) (Spielman *et al.*, 1993), uses only transmission information from parents to offspring and is therefore robust to such confounding. The collapsing approaches of rare variants originally developed for unrelated samples have been extended to gene-based TDT and show improved power over single marker TDT while controlling for population stratification/admixture (De *et al.*, 2013; He *et al.*, 2014; Jiang *et al.*, 2014), providing a robust approach to rare variant analysis. The general strategy of these methods involves initial analysis done based on single variants, and statistics across all rare variants are then aggregated, usually in a linear fashion, to form a gene- or group-wise statistic (De *et al.*, 2013; He *et al.*, 2014; Jiang *et al.*, 2014). These methods essentially implement additive genetic models, and are unable to model interactions among rare variants in a gene or pathway. Such interactions may play important roles in complex disease (Lim *et al.*, 2013; Liu and Leal, 2010). For example, the compound heterozygous (CH) model of loss of function variants has been implicated in autism (Lim *et al.*, 2013). Given that causal rare variants may act in complex mechanisms at different loci for different diseases, it is critical to develop methods that can encompass flexible models in a unified framework, and to provide tools that enable the community to carry out such analyses.

In this study, we describe a general framework for group-wise TDT (gTDT) for rare variants and implement a few classical genetic models in a user-friendly and computationally efficient tool. The gTDT is haplotype-based and models the transmission of rare variant carrying haplotypes rather than single variants. Directly modeling haplotypes is advantageous because (i) the transmission unit is haplotype and therefore it is the logical unit to model, and (ii) gene products originate from haplotypes, so it is biologically meaningful to model haplotypes (Tewhey *et al.*, 2011), (iii) it is straightforward to model interactions among variants along haplotypes or between father or mother's haplotypes and (iv) linkage disequilibrium (LD) among variants is precisely accounted for without explicit estimation of LD. Given the flexibility of the framework, a variety of models can be implemented. For example, models implemented in our tool include not only simple models such as additive and CH models but also a hybrid that fuses CH and weighted additive models to test for recessive effects while incorporating prior knowledge on different variants. It is straightforward to extend to other customized models tailored to specific diseases or genes to meet specific needs. Through extensive simulations, we showed that type I error was correctly controlled for rare/low-frequency variants (e.g. up to MAF = 0.1) under a variety of scenarios, including various levels of population stratification. We evaluated the power of gTDT for different underlying genetic models, and gTDT significantly outperformed single marker TDT. For additive models we were able to compare gTDT with other methods and observed comparable power. Overall our gTDT provides a flexible framework and an efficient computational tool to identify rare variants implicated in diseases.

## 2 Methods

### 2.1 gTDT for rare variants

The gTDT models phased genotypes and thus requires haplotype construction. If genotypes provided are phased by other means, gTDT can directly takes the haplotypes to carry out haplotype-based gTDT analysis. When unphased genotypes are provided, gTDT automatically re-constructs haplotypes by transmission within each trio when carrying out the analysis. The phasing is carried out as following. We assume all variants are bi-allelic with allele 0 and 1, representing the reference and alternative allele, respectively. We denote the offspring's ordered genotype as $(A_1, A_2)$ such that $A_1$ is transmitted from father and $A_2$ is from mother. For parental genotypes, the alleles are arranged in an order that the allele on the left is transmitted to offspring and the allele on the right is not transmitted. By this convention, the order of the alleles can be uniquely determined for all possible genotype configurations (15 in total) in a trio except the ambiguous situation in which all genotypes in a trio are heterozygous. This process is repeated for all variants in a trio so that phasing can be determined along the chromosome. In ambiguous cases, two phasing configurations are compatible, one of which can be obtained from the other by flipping the two alleles in each of the ordered genotypes in a trio. Phasing is randomly assigned for such cases in gTDT. For rare variants this will be extremely rare and therefore has little impact on the analysis (see Section 3).

**Table 1.** Genetic models implemented in gTDT

| | Equation |
|---|---|
| AD | $\delta(G)=\delta(H_1,H_2)=\sum_{j=1}^{J}(H_1^j+H_2^j)$ |
| wAD | $\delta(G)=\delta(H_1,H_2)=\sum_{j=1}^{J}w_j(H_1^j+H_2^j)$ |
| DOM | $\delta(G)=\begin{cases}1 & \text{if } \sum_{j=1}^{J}(H_1^j+H_2^j)>0\\ 0 & \text{otherwise}\end{cases}$ |
| CH | $\delta(G)=\begin{cases}1 & \text{if } \sum_{j=1}^{J}(H_1^j) \text{ and } \sum_{j=1}^{J}(H_2^j)\geq 1\\ 0 & \text{otherwise}\end{cases}$ |
| wCH | $\delta(G)=\begin{cases}\sum_{j=1}^{J}w_j(H_1^j+H_2^j) & \text{if } \sum_{j=1}^{J}(H_1^j) \text{ and } \sum_{j=1}^{J}(H_2^j)\geq 1\\ 0 & \text{otherwise}\end{cases}$ |

Suppose we are testing the group-wise association of $J$ rare variants in $N$ parents-proband trios. Let $G$ denote the phased genotypes of an individual as determined above, i.e. $G=(H_1,H_2)$, where $H_1$ and $H_2$ are the two haplotypes in the gene or genomic region. We further let $H_i^j$, $i=1,2$ and $j=1,\ldots,J$, denote the coding of the $j$th variant on haplotype $H_i$, with $H_i^j=1$ when it carries a rare allele and zero otherwise. Let $G_F=(H_{F1},H_{F2})$ and $G_M=(H_{M1},H_{M2})$ denote the father and mother's genotypes respectively, then the child's genotype is deduced as $G_c=(H_{F1},H_{M1})$ according to the phasing strategy aforementioned. To achieve the aggregation effect of multiple rare variants in a group, we define a function, $\delta(G)$, which numerically codes the phased genotype across multiple rare variants in appropriate ways to reflect the underlying genetic models. This framework provides a flexible approach to incorporating a variety of different genetic models (see examples below and Table 1). Let $P(D|G)$ denote the risk of being affected when the genotype is $G$, and define the relative risk (RR) of $G$ over a baseline genotype $G_0$ as $P(D=1|G)/P(D=1|G_0)$. Suppose the RR follows a multiplicative model, i.e.

$$\log\left(\frac{P(D|G)}{P(D|G_0)}\right)=\beta(\delta(G)-\delta(G_0)) \quad (1)$$

The null hypothesis ($H_0$: $\beta=0$) holds when no variants in the group are associated with the disease, i.e. RR = 1. The conditional-on-parental-genotype likelihood (Schaid, 1996) of a trio is:

$$P(G_c|D,G_F,G_M)=\frac{P(D|G_c,G_F,G_M)P(G_c|G_F,G_M)}{\sum_{G_c'}P(D\mid G_c',G_F,G_M)P(G_c'|G_F,G_M)}$$

The sum in the denominator is over all four possible phased offspring genotypes from the parents, and $P(G_c|G_F,G_M)=P(G_c'|G_F,G_M)=1/4$. Assuming that disease status only depends on the child's genotype in a trio, the above equation is simplified to

$$P(G_c|D,G_F,G_M)=\frac{P(D|G_c)}{\sum_{G_c'}P(D|G_c')}=\frac{\exp(\beta\delta(G_c))}{\sum_{G_c'}\exp(\beta\delta(G_c'))}$$

Differentiating the log likelihood of the above and evaluating at $\beta=0$, we get the score statistic for a trio under the null hypothesis as the following:

$$S=\delta(G_c)-\frac{\sum_{G_c'}\delta(G_c')}{4}$$

In the above equation, $\delta(G_c)$ and $\sum_{G_c'}\delta(G_c')$ are the observed and expected numerically coded genotypes of offspring. The four possible phased genotypes under the null, i.e. random transmission from parents to offspring, are $G_{c1}'=(H_{F1},H_{M1})$, $G_{c2}'=(H_{F1},H_{M2})$, $G_{c3}'=(H_{F2},H_{M1})$, $G_{c4}'=(H_{F2},H_{M2})$, of which $G_{c1}'=(H_{F1},H_{M1})$ is the same as the observed offspring genotype by construction. The variance of the score under the null, $V(S)=\frac{1}{4}\sum_{G_c'}\delta(G_c')^2-\left[\frac{1}{4}\sum_{G_c'}\delta(G_c')\right]^2$, is calculated by taking the negative of the second derivative of the log likelihood and evaluating at $\beta=0$. A score test can then be constructed by summing over all $N$ trios as

$$t=\frac{\sum_{k=1}^{N}S_k}{\sqrt{\sum_{k=1}^{N}V(S_k)}}\sim N(0,1)$$

One-sided or two-sided tests can be carried out depending on specific hypothesis. The key advantage of this framework lies in the flexible aggregation of phased genotypes to encompass a variety of models to increase power. In this study, we describe five genetic models that were implemented in our user-friendly tool. The AD model is the traditional additive model across all variants and between the two haplotypes, i.e. $\delta(G)=\delta(H_1,H_2)=\sum_{j=1}^{J}(H_1^j+H_2^j)$. In this model the aggregate coding of the genotype is the count of rare variants carried in this group. The wAD is the weighted version of the additive model, where the weight $w_j$ is assigned to the $j$th variant. The DOM is the carrier (i.e. dominant) model, where the phased genotype is coded 1 if an individual carries one or more rare variants. The CH is a compound heterozygous (i.e. recessive) model in which the two haplotypes have to carry at least one rare allele each. The wCH is a hybrid of wAD and CH, in which compound heterozygotes are coded as the weighted sum of individual variants. If equal weights are assigned to all variants, wCH becomes a hybrid of AD and CH. For the representation of all models, see Table 1. It is straightforward to extend the framework to handle other models, with possible incorporation of complex interactions among rare variants across haplotypes, by designing coding schemes of the phased genotypes. We plan to incorporate such models in our software package as the need arises.

By default, gTDT uses a frequency-dependent scheme similar to Madsen and Browning (2009), i.e. $w_j=1/\sqrt{p_j(1-p_j)}$, where $p_j$ is the MAF of the $j$th variant and is estimated based on parental genotypes. This weighting scheme up-weights rare variants under the hypothesis that the rarer variants have larger effect sizes, and is also robust against non-causal common variants by down-weighting them. More efficient weighting schemes can be constructed by incorporating prior knowledge, e.g. prediction scores from various bioinformatics tools. Our tool can readily take such weights into these models to carry out customized gTDT tests.

We implemented gTDT in C++. It takes a multi-sample VCF as input and calculates either one-sided or two-sided tests depending on the user's input. More details are available in the online manual. In this study, we used two-sided tests for the evaluation.

## 2.2 Simulation

We evaluated the type I error and power of gTDT using simulated data. In order to generate simulated data that resemble empirical data in allele frequency, LD, and population differentiation, we adopted cosi (Schaffner *et al.*, 2005) to generate haploid DNA sequence pools, each of which contained 10 000 haplotypes. To generate a trio, a pair of haplotypes was randomly drawn from a pool and assigned as the genotype of each parent, and offspring's genotype was generated by randomly selecting a haplotype from each parent. For type I error evaluation, we simulated $N$ trios and assigned the offspring as affected, regardless of offspring's genotype. For power analysis, the disease status of offspring was determined

based on the penetrance model described in Equation (1), in which the penetrance was calculated according to RR with the baseline penetrance of 0.05. Only trios with affected offspring were collected.

Under the null hypothesis, we generated 50 000 replicates of 1000 trios. Two lengths of haplotypes with 30 and 50 variants were simulated. We used the two lengths to explore the grouping of regions similar to the average gene coding sequences as well as situations where larger genes or genes with non-coding variants are included. To further test type I error in the presence of population stratification, we generated haplotype pools for both European and African populations using cosi, and then simulated trios based on these haplotypes. Next, we mixed trios from different populations at ratios of 1:4, 1:1 and 4:1 to simulate different levels of population stratification. Again 50 000 replicates with 1000 trios were generated as described above such that population stratification issues were included in simulated data.

To evaluate the power of gTDT, data were simulated under AD, wAD, CH and wCH models separately. To mimic the reality in which both causal and non-causal variants are present, we selected haplotypes with 100 variants and randomly assigned 10 or 30% of variants with MAF <0.05 as causal. For AD with equal effect sizes, we assigned $\beta = \log(4)$, i.e. RR = 4, to all causal variants and zero otherwise. For CH, we used $\beta = \log(10)$, to causal compound heterozygotes and zero otherwise. For wAD, we assigned effect sizes in an allele frequency-dependent fashion. Let $\beta_i$ denote the specific effect of the $j$th variant. We first assumed variants with MAF below 0.01 have weights $\beta_i \in [\log(1.5), \log(4)]$ and a linear relationship between $\beta_i$ and MAF. Specifically, we divided $\beta_i$ $[\log(1.5) - \log(4)]$ and MAF $[0.01 - 0.0001]$ into 10 equal intervals and then assigned $\beta_i$ to variants with corresponding MAF. For variants with MAF $\in [0.01, 0.1]$, we adopted $\beta_i \in [\log(1.2), \log(1.5)]$, and also divided MAF and $\beta_i$ into 10 equal intervals, then assigned variants with different weights as above. Finally, we assigned a constant $\beta_i = \log(1.1)$ to all the variants with MAF >0.1 if included. In this setup, the variants with smaller MAF were assigned higher effect sizes. For all simulations, power was evaluated based on 1000 replicates each of which had a sample size of 1500 trios.

## 3 Results

### 3.1 Evaluation of type I error

We first evaluated type I error at $\alpha$ level of 0.05 with different numbers of collapsed variants in homogenous population when the phasing was known through simulations. Table 2 summarizes the proportion of replicates with $P$-value $\leq \alpha$ for variants spanning minor allele frequencies <0.1 under various genetic models. Type I error rates were correctly controlled in all the scenarios, although for the CH and wCH the tests were conservative (Table 2). We also investigated the type I error rates for variants with MAF between 0.1 and 0.5 and no inflation of type I error was observed (data not shown), indicating that LD was correctly accounted for even though extensive LD exists among common variants. The conservativeness of CH and wCH is due to the rarity of compound heterozygotes of rare variants and is a common phenomenon associated with rare variants (Li and Leal, 2008; Li *et al.*, 2013). Furthermore, we obtained similar results when evaluating the type I error at $\alpha$ level of 0.005 (data not shown).

In typical data analysis pipelines, usually only unphased genotypes are generated. One solution is to phase genotypes using statistical methods, such as BEAGLE (Browning and Browning, 2009), MaCH (Li *et al.*, 2010), SHAPEIT (Delaneau *et al.*, 2013) or TrioCaller (Chen *et al.*, 2013). We investigated whether phasing by simple transmission implemented in gTDT was adequate to reconstruct haplotypes to correctly control the type I error rates. Table 2 shows the type I error rates for the same scenarios as investigated for known phasing. For variants with MAF <0.1 the type I error rate was adequately controlled for all genetic models, with similar conservativeness observed for CH and wCH. When variants with MAF >0.1 were included, inflation of type I error rates was observed (Supplementary Table S1). This is caused by incorrect phasing, which occurs when all individuals of a trio are heterozygotes and becomes common when the MAF increases. However, the collapsing approach is applied only to rare variants and such configurations are extremely unlikely to be observed, confirmed by the correct type I error rates for variants with MAF below 0.1. When it is desirable to include common variants in the collapsing, it is necessary to phase genotypes using LD-based tools in order to correctly control the type I error rates.

**Table 2.** Type-I error at $\alpha = 0.05$ level for 50 000 replicates with 1000 trios using phased/unphased data (50/30 variants) at various MAF cutoffs

| MAF | ≤0.01 | ≤0.05 | ≤0.1 |
|---|---|---|---|
| Phased | | | |
| AD | 0.0504/0.0511 | 0.0497/0.0498 | 0.0485/0.0502 |
| wAD | 0.0492/0.0498 | 0.0494/0.0502 | 0.0494/0.0496 |
| DOM | 0.0494/0.0507 | 0.0483/0.0502 | 0.0491/0.051 |
| CH | 0.0324/0.0217 | 0.0435/0.0375 | 0.0472/0.0441 |
| wCH | 0.0361/0.0245 | 0.042/0.0367 | 0.0468/0.0417 |
| Unphased | | | |
| AD | 0.0505/0.0511 | 0.0497/0.0499 | 0.0481/0.05 |
| wAD | 0.0492/0.0498 | 0.0495/0.0502 | 0.0492/0.0496 |
| DOM | 0.0494/0.0507 | 0.0483/0.0502 | 0.0493/0.0507 |
| CH | 0.0324/0.0217 | 0.0434/0.0373 | 0.0482/0.0446 |
| wCH | 0.0363/0.0245 | 0.0421/0.0362 | 0.0467/0.0418 |

Variants that are heterozygous in all individuals of a trio are non-informative in single marker tests, but have impact on the type I error of haplotype-based gTDT if not phased correctly. We use the example with two variants in a group in Supplementary Figure S1 to illustrate the impact of incorrect phasing on test statistics. While the phase of the second variant (at the bottom) is uniquely determined, the phasing of the first variant is ambiguous. There are two compatible phasing configurations (Supplementary Fig. S1). For AD, the scores for cases 1 and 2 are the same, while the corresponding variances are different, e.g. 1.67 and 0.33, respectively. This holds for wAD as well. For DOM, both the scores and the variances are different in the two cases, with $S = 0.75$, $V(S) = 0.25$ for case 1 and $S = 1$, $V(S) = 0$ for case 2. For CH, the same pattern holds as for DOM, with $S = 0.25$, $V(S) = 0.25$ for case 1 and $S = 0.5$, $V(S) = 0.33$ for case 2. The wCH is similarly affected. Based on these analyses ambiguous phasing has differential impact on different tests, and without phasing by LD, incorrect haplotype construction can lead to inflation of type I error rates as shown in Supplementary Table S1. For additive models the impact is not as dramatic as others (Supplementary Table S1), probably due to the use of random phasing implemented in gTDT, which potentially reduces the impact on the overall statistics.

Finally, we evaluated type I error in the presence of population stratification. In this investigation, we used phasing by transmission to reconstruct haplotypes for gTDT rare variant analysis. As expected, the type I error rates were well controlled at different

levels of population stratification for all models and MAF <0.1 (Table 3). The overall patterns were similar to scenarios without population stratification (Tables 2 and 3). When phasing was available, our framework had correct type I error rates regardless of allele frequencies in the presence of population stratification (data not shown).

### 3.2 Evaluation of statistical power of gTDT

We explored the power of gTDT using simulated data for different models with various effect sizes. All tests were carried out on datasets with 1500 trios. For datasets simulated under AD, the power of additive models with and without weighting showed the highest power for variants with different allele frequencies (Table 4). When the data were simulated with varying effect sizes, the weighted version of the additive model exhibited the greatest power, and was significantly more powerful than the unweighted version when high-frequency non-causal variants were included (Table 4). For scenarios in which CH was the true model, CH and wCH were more powerful than other models (Table 4). However, wCH was more powerful than the CH model (Table 4), due to the use of weights that preferentially penalize non-causal high frequency variants. Consistently, for data simulated under the CH model with varying effect sizes, wCH was the most powerful one (Table 4). It is noticeable that the power was significantly lower for CH models even though higher effect sizes were used. This is not surprising due to the rarity of compound heterozygotes and only large effects in such models can be detected. In addition, the dominant model (DOM) was generally very similar to the unweighted additive model (AD) (Table 4). For additive models we were able to compare the power of gTDT with other methods, and we observed similar power compared with rvTest (He *et al.*, 2014) in various scenarios. For example, the power of wAD was ~93% for MAF <0.05 at alpha level of 0.05, and for the same setting the power was ~94% for rvTest.

There are scenarios that are non-informative to single variant TDT but have impact on haplotype-based gTDT results: (i) parents have the same genotypes that are homozygous rare allele and (ii) parents have different homozygous genotypes. These two have no impact on additive models since such variants contribute nothing to both the score and its variance. This is not the case for other models. For example, with scenario (i), the genotype is coded as 1 for the DOM and CH model, and non-zero for wCH; with scenario (ii), it has the same effect on DOM as for (i), and has the potential to make the genotype as compound heterozygotes. Although such variants do not contribute variance to the test statistics on their own, they may carry valuable functional information, which is captured and reflected in gTDT via haplotype-based modeling.

**Table 3.** Type-I error at $\alpha = 0.05$ levels in presence of population stratification at various MAF cutoffs

| E:A | 1:1 | | 1:4 | | 4:1 | |
|---|---|---|---|---|---|---|
| MAF | 0.01 | 0.1 | 0.01 | 0.1 | 0.01 | 0.1 |
| AD | 0.0509 | 0.0499 | 0.0505 | 0.0499 | 0.0495 | 0.0494 |
| wAD | 0.0513 | 0.0503 | 0.0498 | 0.0499 | 0.0502 | 0.0499 |
| DOM | 0.0507 | 0.0498 | 0.0504 | 0.0513 | 0.0486 | 0.0495 |
| CH | 0.019 | 0.0467 | 0.0166 | 0.0473 | 0.0219 | 0.0464 |
| wCH | 0.0214 | 0.0449 | 0.0185 | 0.0462 | 0.0246 | 0.0451 |

E:A, European:African.

### 3.3 Application to autism trio exome sequencing data

We applied gTDT to exome sequencing data of 118 parent-offspring trios as part of the University of Illinois at Chicago's Autism Center of Excellence (ACE) project. The sample information has been described elsewhere (Levin-Decanini *et al.*, 2013). The exome data were generated at the Center for Inherited Disease Research, and variant calling was carried out using the standard GATK pipeline (DePristo *et al.*, 2011; McKenna *et al.*, 2010). We only kept the variants that passed the GATK's VQSR quality assessment. Variants were annotated using ANNOVAR (Wang *et al.*, 2010) (2014Mar10) and we denoted non-synonymous, stop and splicing variants as 'functional' variants.

Since TDT assumes under the null that the transmission of alleles from parents to offspring is random, i.e. the probability that either allele is transmitted is 0.5, we first checked whether the transmission rate had a systematic deviation from 0.5 in GATK calls. The transmission ratios along with other statistics are displayed in Table 5. For functional variants there was an under-transmission bias as transmission ratios are <50% across all MAF and depth cutoffs investigated (Table 5). We also investigated the transmission ratio using all variants and observed a more severe under-transmission bias compared with functional variants (Table 5). The major difference in the non-functional variants is the relatively lower depth (~50×) compared with functional variants (~73×), and for lower coverage the bias will be more severe. We reasoned that a major cause of the bias was the genotype calling error that tended to miscall rare heterozygotes as homozygous reference allele, as a strong prior was assigned favoring the reference allele in standard variant calling algorithms. For example, in a trio where a rare allele is

**Table 4.** Power at $\alpha = 0.05$ without population stratification and admixture at various MAF cutoffs with different fractions of causal variants (10%, 30%)

| %casual | Model | AD, RR = 4 | | wAD, RR ∈ [1.2, 4] | | CH, RR = 10 | | wCH, RR ∈ [1.2, 4] | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.01 | 0.1 | 0.01 | 0.1 | 0.01 | 0.1 | 0.01 | 0.1 |
| 30% | AD | 0.972 | 0.882 | 0.977 | 0.915 | 0.088 | 0.613 | 0.014 | 0.503 |
| | wAD | 0.964 | 0.901 | 0.978 | 0.986 | 0.084 | 0.522 | 0.011 | 0.356 |
| | DOM | 0.974 | 0.883 | 0.974 | 0.918 | 0.05 | 0.263 | 0.003 | 0.124 |
| | CH | 0.381 | 0.733 | 0.399 | 0.686 | 0.363 | 0.708 | 0.172 | 0.622 |
| | wCH | 0.368 | 0.726 | 0.374 | 0.757 | 0.363 | 0.724 | 0.16 | 0.609 |
| 10% | AD | 0.425 | 0.533 | 0.489 | 0.423 | 0.048 | 0.226 | 0.002 | 0.144 |
| | wAD | 0.341 | 0.490 | 0.389 | 0.455 | 0.045 | 0.131 | 0.005 | 0.059 |
| | DOM | 0.445 | 0.510 | 0.465 | 0.354 | 0.040 | 0.081 | 0.005 | 0.019 |
| | CH | 0.098 | 0.351 | 0.108 | 0.184 | 0.065 | 0.290 | 0.013 | 0.199 |
| | wCH | 0.100 | 0.283 | 0.097 | 0.191 | 0.060 | 0.251 | 0.011 | 0.160 |

**Table 5.** Under-transmission bias happens during variants calling utilizing GATK, and Polymutt eliminated this bias essentially

| Variants | Depth | MAF | GATK | | | Polymutt | | |
|---|---|---|---|---|---|---|---|---|
| | | | $T/(T+N)$ (%) | Percentage of Mendelian inconsistency | | $T/(T+N)$ (%) | Percentage of Mendelian inconsistency | |
| | | | | Marker | Trio | | Marker | Trio |
| Functional | 5 | 0.01 | 49.19 | 1.34 | 0.84 | 50.04 | 0.00 | 0.00 |
| | | 0.02 | 49.28 | 1.43 | 0.71 | 49.98 | 0.00 | 0.00 |
| | | 0.05 | 49.45 | 1.50 | 0.56 | 49.95 | 0.00 | 0.00 |
| | 10 | 0.01 | 49.56 | 0.66 | 0.44 | 49.97 | 0.00 | 0.00 |
| | | 0.02 | 49.57 | 0.71 | 0.37 | 49.92 | 0.00 | 0.00 |
| | | 0.05 | 49.63 | 0.74 | 0.29 | 49.87 | 0.00 | 0.00 |
| All | 5 | 0.01 | 48.15 | 2.50 | 1.73 | 49.99 | 0.00 | 0.00 |
| | | 0.02 | 48.21 | 2.59 | 1.43 | 49.97 | 0.00 | 0.00 |
| | | 0.05 | 48.11 | 2.75 | 1.13 | 49.93 | 0.01 | 0.00 |
| | 10 | 0.01 | 48.85 | 0.99 | 0.77 | 49.81 | 0.00 | 0.00 |
| | | 0.02 | 48.86 | 1.02 | 0.63 | 49.83 | 0.00 | 0.00 |
| | | 0.05 | 48.70 | 1.07 | 0.51 | 49.82 | 0.00 | 0.00 |

$T/(T+N)$, transmission/(transmission + non-transmission).

transmitted from mother to child (Supplementary Fig. S2), a miscall in either the mother or the child will induce under-transmission—for case I it is a Mendelian inconsistency and an under-count of the true transmitted alleles, and for case II the true transmitted allele is incorrectly regarded as non-transmission. Since GATK ignores the transmission information in trios, we applied PolyMutt (Li *et al.*, 2012), an algorithm that jointly models the trios for genotype calling, to refine the genotypes and re-evaluated the bias on the new calls. Clearly the bias was much reduced, for both functional and all variants, although the bias was not completely eliminated (the transmission ratio is still slightly <0.5, Table 5). As a result of joint modeling of trios, the Mendelian error was essentially eliminated in PolyMutt calls compared with GATK calls (Table 5).

We used PolyMutt calls for gene-based gTDT analysis given the much reduced transmission bias. We focused on functional variants that were predicted to be deleterious by various bioinformatics tools provided in ANNOVAR output; these prediction algorithms include Polyphen-2, SIFT, LRT, MutationTaster, MutationAssessor and CADD. We used MAF <0.02 for AD and DOM and MAF <0.05 for CH models (CH and wCH). Furthermore, for each variant site trios were included in the analysis if the minimum depth of the three family members was >5. Q–Q plots showed that there was no inflation of type I error in all gTDT tests (Supplementary Fig. S3). For this dataset, the *P*-values of AD-DOM were consistent and we reported results based on AD; similarly for compound heterozygotes we reported results based on CH. No genes achieve exome-wide significance, and the smallest *P*-value is $5 \times 10^{-4}$ for gene KCNV2 under AD. The top five genes for AD with over-transmission are KCNV2, SPEF2, ENAM, BCAM and DRC1. However, none of these top genes overlap with the 124 known high confidence autism genes reported in a previous study (Pinto *et al.*, 2014). We further assessed the expression levels of these five genes in different brain regions spanning 15 developmental periods based on the data from BrainSpan Atlas (Kang *et al.*, 2011). None of the five genes are expressed in brain across developmental periods (Fig. 1A), indicating that these are unlikely to be autism-associated genes.

For CH, the most significant gene is STAU2 with a *P*-value of 0.0027, and the top five genes include STAU2, ZFHX4, PDZD2, ACIN1 and ANKRD16. We expected to observe less significant *P*-values for CH models due to the rarity of compound heterozygotes for rare variants. However, when examining the gene expression in the BrainSpan Atlas data (Kang *et al.*, 2011), we observed that all these genes were expressed in different regions across developmental periods (Fig. 1B). Moreover, STAU2, ZFHX4 and PDZD2 showed varying expression patterns in the prenatal stages, which are important for the development of autism (Willsey *et al.*, 2013). Literature search revealed the involvement of these genes in brain functions or psychiatric disorders. For example, STAU2, which belongs to a family of RNA-binding proteins, is highly enriched in the brain and plays key roles in early differentiation of neurons and in the synaptic plasticity of mature neurons (Heraud-Farlow and Kiebler, 2014); ZFHX4 is found to be deleted in the 8q21.11 microdeletion syndrome associated with intellectual disability (Palomares *et al.*, 2011).

## 4 Discussion

Identifying genes harboring rare variants associated with human complex traits is of great interest. On the other hand, it is also intrinsically challenging, owing to the dramatically reduced power as well as the difficulty to account for rare variant population stratification/admixture. The framework we described in this study tackles this problem by aggregating rare variants in a group and uses a haplotype-based TDT to control for population stratification/admixture. Compared with existing methods, our framework is better able to deal with complex genetic models. Given the ubiquity of allelic heterogeneity in disease etiology, it is natural to assume that there exist complex interactions between causal variants. Our work provides a framework that enables the exploration of such complex models. In addition, customized weighting schemes can be readily provided to gTDT so that effective weights, such as functional prediction scores (Price *et al.*, 2010) or the weights based on the clever use of controls (Jiang *et al.*, 2014), can be incorporated to various models in gTDT to take advantage of specialized weights. If the focus is on rare variants (e.g. MAF < 0.05), our results show that it is adequate to use phasing by transmission implemented in gTDT. We believe that this is the most commonly employed strategy so that gTDT is able to handle the majority of situations. For variants with MAF >0.1 we recommend using LD-based methods to phase genotypes before gTDT analysis.

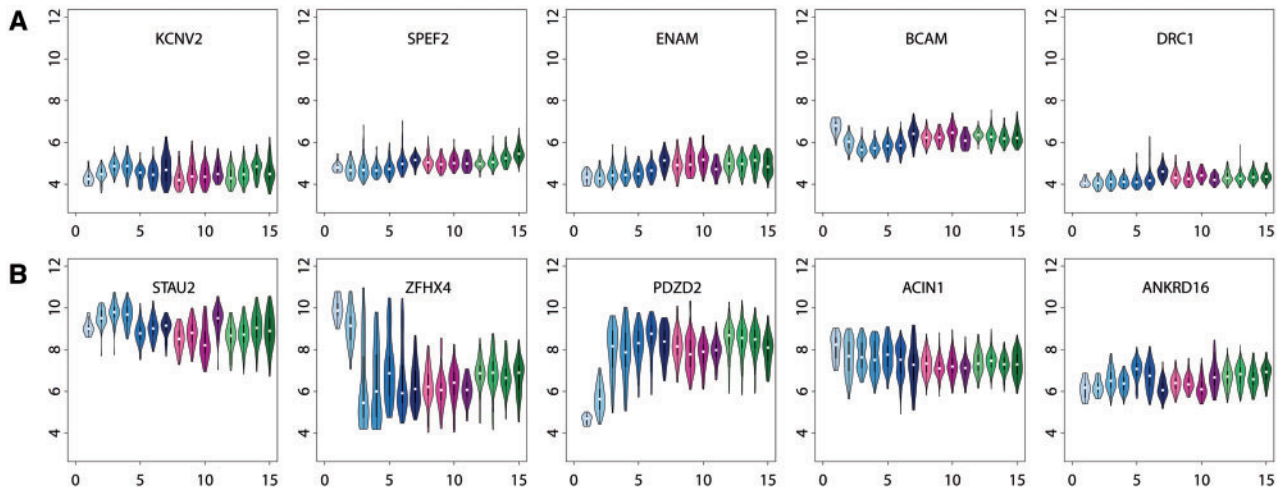


**Fig. 1.** Violin plots of the expression of most significant genes in various brain regions across developmental stages based on BrainSpan Atlas data. (**A**) Upper panel displays the gene expression for top genes under the additive model. (**B**) Displays the gene expression of top genes under the CH model. *X*-axis is the developmental stage, with 1–7 representing prenatal periods, 8–11 for childhood and 12–15 for the adulthood. *Y*-axis displays the distribution of the gene expression in different brain regions (see Kang *et al.*, 2011 for details). Expression levels under 5 are generally considered as non-expression

Haplotype-based TDT frameworks for multiple makers have been proposed previously based on genotyping data on common variants (Abad-Grau *et al.*, 2013; Clayton, 1999; Zhao *et al.*, 2000). These methods usually have multiple degrees of freedom without collapsing effects, and often need to deal with uncertainties in haplotype phasing. A notable method employed a strategy to classify haplotypes into two groups to reduce degrees of freedom (Abad-Grau *et al.*, 2013). In this method the collapsing of haplotypes is based on observed over- or under-transmission in the data instead of prior knowledge. On the other hand, our method is designed for rare variants uncovered via sequencing, with almost certain haplotype reconstruction and biological meaningful collapsing of rare variants. As a result, it offers flexibility to test different genetic models based on prior knowledge with reduced degree of freedom to increase power.

In gTDT it is natural to use gene as the analysis unit. It is equally applicable to groups of rare variants from multiple genes. A key factor to consider for gTDT analysis of pathways or gene sets is that recombination events can occur between genes. To check the validity of gTDT for such analyses, we also simulated two groups of variants that had recombination rate of 0.5 between them. We carried out the gTDT assuming no recombination and obtained correct type I error estimates (data not shown), indicating that gTDT could be applied to groups of variants beyond the gene level.

In gTDT we implemented several common genetic models. In reality the underlying genetic model is generally unknown, and is likely to vary from gene to gene. Some researches may be interested in how to combine the results from different models. If the relative probability of each model is known *a priori*, the association signal can be obtained by taking a meaningful average of individual tests, e.g. through Bayesian model averaging approaches (Stephens and Balding, 2009). Since our method is not Bayesian, and it is hardly possible to have a meaningful estimate of the priors, model averaging is not readily applicable here. An alternative is to carry out association tests of genetic models of interest and use the maximum statistic as the association signal for each gene; in this setup the significance need to be correctly calculated either through permutation or other approximation approaches (Gonzalez *et al.*, 2008). If the primary interest is to test one of the models, e.g. the CH model, model averaging or selection approaches are not necessary.

We observed an under-transmission bias in genotype calls and this bias can lead to not only inflation of type I error but also decrease in the power to detect risk alleles. This bias can be more dramatic when sequencing depth is low, as observed in other data we have analyzed, and may be influenced by other factors during complex sequencing and analysis steps. This further exacerbates the challenges for rare variant association studies. Correcting the bias may not be straightforward, as rare variants in the same gene or group will likely to have differential bias patterns due to different depth of coverage, allele frequencies, among other factors. Although PolyMutt can dramatically reduce the bias by modeling allele transmission in trios, the bias does not seem to be completely eliminated. It may require further effort to develop sophisticated methods in order to efficiently correct the bias for gTDT tests.

Application of gTDT to small-scale exome sequencing data on autism trios did not reveal promising genes under additive models. This may simply reflect the inadequate power of the this study to detect weak signals of rare variants. On the other hand, our analyses showed that CH models had potential to identify associated genes harboring interacting rare risk alleles. Studies with larger sample sizes are needed to elucidate the etiology of complex traits contributed by rare variants. Our framework is flexible to incorporate other complex interacting models to identify rare variant associations.

## Acknowledgement

The authors thank Goncalo Abecasis in the Department of Biostatistics at the University of Michigan for sharing the C++ code for processing pedigrees.

## Funding

This study was supported by the National Institute of Health [grant R01HG006857 to R.C., Q.W. and B.L., R01HG007358 to W.C., P50 HD055751 and X01 HG007235 to E.C., J.S. and N.C.] for the ACE data collection and sequencing.

*Conflict of Interest*: none declared.

## References


Abad-Grau, M.M. *et al.* (2013) Increasing power by using haplotype similarity in a multimarker transmission/disequilibrium test. *J. Bioinform. Comput. Biol.*, **11**, 1250014.

Asimit,J. and Zeggini,E. (2010) Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.*, **44**, 293–308.

Auer,P.L. *et al.* (2014) Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat. Genet.*, **46**, 629–634.

Browning,B.L. and Browning,S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.

Chen,W. *et al.* (2013) Genotype calling and haplotyping in parent-offspring trios. *Genome Res.*, **23**, 142–151.

Clayton,D. (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am. J. Hum. Genet.*, **65**, 1170–1177.

Cohen,J.C. *et al.* (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.

Cruchaga,C. *et al.* (2014) Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*, **505**, 550–554.

De,G. *et al.* (2013) Rare variant analysis for family-based design. *PLoS ONE*, **8**, e48495.

Delaneau,O. *et al.* (2013) Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.*, **93**, 687–696.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Gonzalez,J.R. *et al.* (2008) Maximizing association statistics over genetic models. *Genet. Epidemiol.*, **32**, 246–254.

Han,F. and Pan,W. (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.*, **70**, 42–54.

He,Z. *et al.* (2014) Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am. J. Hum. Genet.*, **94**, 33–46.

Heraud-Farlow,J.E. and Kiebler,M.A. (2014) The multifunctional Staufen proteins: conserved roles from neurogenesis to synaptic plasticity. *Trends Neurosci.*, **37**, 470–479.

Ji,W. *et al.* (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.

Jiang,Y. *et al.* (2014) Utilizing population controls in rare-variant case–parent association tests. *Am. J. Hum. Genet.*, **94**, 845–853.

Kang,H.J. *et al.* (2011) Spatio-temporal transcriptome of the human brain. *Nature*, **478**, 483–489.

Kang,H.M. *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

Levin-Decanini,T. *et al.* (2013) Parental broader autism subphenotypes in ASD affected families: relationship to gender, child's symptoms, SSRI treatment, and platelet serotonin. *Autism Res.*, **6**, 621–630.

Li,B. *et al.* (2012) A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet.*, **8**, e1002944.

Li,B. *et al.* (2013) Identifying rare variants associated with complex traits via sequencing. *Curr. Protoc. Hum. Genet.*, Chapter 1, Unit 1.26.

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.

Li,Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Lim,E.T. *et al.* (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron*, **77**, 235–242.

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.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Mathieson,I. and McVean,G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243–246.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Morgenthaler,S. and Thilly,W.G. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, **615**, 28–56.

Neale,B.M. *et al.* (2011) Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e1001322.

Nejentsev,S. *et al.* (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, **324**, 387–389.

Nelson,M.R. *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**, 100–104.

Palomares,M. *et al.* (2011) Characterization of a 8q21.11 microdeletion syndrome associated with intellectual disability and a recognizable phenotype. *Am. J. Hum. Genet.*, **89**, 295–301.

Pinto,D. *et al.* (2014) Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.*, **94**, 677–694.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Price,A.L. *et al.* (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, **86**, 832–838.

Schaffner,S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.

Schaid,D.J. (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genet. Epidemiol.*, **13**, 423–449.

Spielman,R.S. *et al.* (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **52**, 506–516.

Stephens,M. and Balding,D.J. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.*, **10**, 681–690.

Tennessen,J.A. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.

Tewhey,R. *et al.* (2011) The importance of phase information for human genomics. *Nat. Rev. Genet.*, **12**, 215–223.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Willsey,A.J. *et al.* (2013) Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, **155**, 997–1007.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Zhao,H. *et al.* (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am. J. Hum. Genet.*, **67**, 936–946.