OPEN ACCESS Freely available online

PLOS ONE

# Dosage Transmission Disequilibrium Test (dTDT) for Linkage and Association Detection

**Zhehao Zhang[1]*, Jen-Chyong Wang[1], William Howells[1], Peng Lin[1], Arpana Agrawal[1], Howard J. Edenberg[2], Jay A. Tischfield[3], Marc A. Schuckit[4], Laura J. Bierut[1], Alison Goate[1], John P. Rice[1]***

**1** Washington University School of Medicine, Department of Psychiatry, St. Louis, Missouri, United States of America, **2** Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana, United States of America, **3** LSB 136, Rutgers University, Piscataway, New Jersey, United States of America, **4** Department of Psychiatry, University of California San Diego, La Jolla, California, United States of America

## Abstract

Both linkage and association studies have been successfully applied to identify disease susceptibility genes with genetic markers such as microsatellites and Single Nucleotide Polymorphisms (SNPs). As one of the traditional family-based studies, the Transmission/Disequilibrium Test (TDT) measures the over-transmission of an allele in a trio from its heterozygous parents to the affected offspring and can be potentially useful to identify genetic determinants for complex disorders. However, there is reduced information when complete trio information is unavailable. In this study, we developed a novel approach to ''infer'' the transmission of SNPs by combining both the linkage and association data, which uses microsatellite markers from families informative for linkage together with SNP markers from the offspring who are genotyped for both linkage and a Genome-Wide Association Study (GWAS). We generalized the traditional TDT to process these inferred dosage probabilities, which we name as the dosage-TDT (dTDT). For evaluation purpose, we developed a simulation procedure to assess its operating characteristics. We applied the dTDT to the simulated data and documented the power of the dTDT under a number of different realistic scenarios. Finally, we applied our methods to a family study of alcohol dependence (COGA) and performed individual genotyping on complete families for the top signals. One SNP (rs4903712 on chromosome 14) remained significant after correcting for multiple testing Methods developed in this study can be adapted to other platforms and will have widespread applicability in genomic research when case-control GWAS data are collected in families with existing linkage data.

**Citation:** Zhang Z, Wang J-C, Howells W, Lin P, Agrawal A, et al. (2013) Dosage Transmission Disequilibrium Test (dTDT) for Linkage and Association Detection. PLoS ONE 8(5): e63526. doi:10.1371/journal.pone.0063526

**Editor:** Stacey Cherny, University of Hong Kong, Hong Kong

**Received** August 10, 2012; **Accepted** April 6, 2013; **Published** May 14, 2013



**Funding:** This national collaborative study is supported by National Institutes of Health (NIH) Grant U10 AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA). Funding support for GWAS genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the National Institute on Alcohol Abuse and Alcoholism, the NIH GEI (U01HG004438), and the NIH contract ''High throughput genotyping for studying the genetic contributions to human disease'' (HHSN268200782096C). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: zhang.zhehao@wustl.edu (ZZ); john@zork.wustl.edu (JPR)

## Introduction

Linkage studies have been successfully used to identify many disease genes such as hyper-cholesterolaemia [1–3], Huntington's disease [4] and cystic fibrosis [5]. Linkage studies allow direct observation of recombination events in a family pedigree with a limited number of generations, as well as simultaneous analysis of multiple genetic markers. The LOD score (logarithm of odds), developed by Newton E. Morton [6], is a statistical test often used for linkage analysis in human. The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data by chance. However, this setup requires tailor-made likelihood statistics. When it comes to a multi-loci model, the situation can be even more cumbersome [7]. On the other hand, because of the requirement of a large number of families with several affected generations, linkage analysis can be less helpful when dealing with diseases of late-onset with a high mortality. Alternatively, association studies are used to identify disease susceptibility genes by comparing genetic variants between individuals with and without the disease of interest. High-throughput genotyping has allowed large-scale association studies over the entire human genome. In 2005, the first Genome-Wide Association Study (GWAS) was successfully applied on human age-related macular degeneration [8]. Since then, GWAS has been widely used to identify the association between genetic variants, typically single-nucleotide polymorphisms (SNPs), and heritable traits or diseases.

In general, there are two major types of designs that are commonly used in association research: population-based and family-based studies. As the most common population-based approach, the case-control setup compares an unrelated healthy control group and affected case group. The genotyped SNPs are investigated to identify the allele frequency differences between these two groups. The study then determines whether the SNPs are associated with the genetic trait or disease based on the statistical significance of the differences. The independent samples are typically easier to obtain in a case-control study than family samples. However, many case-control samples select independent cases from existing family data that were originally used in linkage analysis. Because cases can be over-sampled from groups with

higher disease prevalence, the differences of allele frequencies in an admixture of ethnic groups may produce spurious associations. Therefore, although case-control studies have shown advantages in identifying association between the disease susceptibility and markers in a candidate gene, the results may reflect type I errors (false-positive) due to unaccounted confounding factors [9–11] such as population stratification [12–16].

Unlike the population-based studies, family-based studies are resistant to type I errors arising from population stratification. The family-based Transmission/Disequilibrium Test (TDT) measures the over-transmission of an allele from heterozygous parents to their affected offspring, in which the non-transmitted parental alleles serve, in effect, as a control group. Therefore the TDT is a robust test of association in the presence of geographical or ethnical impact from the population [17]. In the original TDT [18], a parent-proband trio is considered as a basic unit, in which a proband is the first affected family member who seeks medical attention for a genetic disorder. Assuming complete genotype information for a two allele marker locus in each trio, the TDT compares the number of heterozygous parents who transmit either allele to the affected offspring. The TDT can be constructed through a 2 by 2 table (**Table 1**). Under the null hypothesis of no associaton, the proportions $b/(b+c)$ & $c/(b+c)$are tested against (0.5, 0.5) using a binomial (asymptotically chi-square) test with one degree of freedom:

$$\chi^2 = \frac{[b-(b+c)/2]^2}{(b+c)/2} + \frac{[c-(b+c)/2]^2}{(b+c)/2} = \frac{(b-c)^2}{b+c} \qquad (1)$$

Because neither genotypes nor allele frequencies are required, the TDT is considered robust to the population stratification as mentioned above.

A variety of TDT-like tests have been suggested starting with Rubinstein et al [19]. Curtis and Sham studied a multi-allelic TDT with incorporation of missing parents [20,21]. This was extended by Spielman et al and Horvath et al [22,23] with the TDT applied to different family structures in their sib-ship tests. For an allele of interest at a marker locus, the sib-ship test essentially compares the frequency of that allele among affected individuals with the frequency of the allele among unaffected individuals, which allows the TDT to be applied to diseases with late age of onset, such as non-insulin-dependent diabetes, cardiovascular diseases, Alzheimer's disease, and other diseases related to aging. Several studies also discussed the application of the TDT for mapping quantitative trait loci [24–30]. Gordon et al.'s TDTae allows for genotyping errors in the analysis and accommodates various error models [31]. As discussed above, multiple affected and unaffected siblings are often collected and used for both linkage and association analysis. The family-based association test (FBAT) generalized the TDT model on various phenotypic traits and multiple markers [32–36]. Instead of using data from only the heterozygous parents as in the TDT, the affected-family based controls (AFBAC) method [37] is developed to take advantage of all the parental information. But the trade-off of this setup is its vulnerability to population stratification as genotype frequencies are not irrelevant in this test [37,38]. Another extension of the TDT, the pedigree disequilibrium test (PDT), is specifically designed for analyzing the Linkage Disequilibrium (LD, the non-random association of alleles at two or more loci) in general pedigrees, which has been successfully applied on a number of complex traits such as diabetes [38,39]. Further, as a more powerful development to the PDT, the presence of linkage (APL) is used to handle diseases of late-onset [38,40]. However, in spite of the divergences as well as the great promise of these TDT-type analyses [9,24,41], one primary limitation that most of these extensions encounter is the dependence on completeness of the genotype information for all trio members in a single test and lack of scalability on utilizing both the linkage and association data in a study.

**Table 1.** Summary of the original TDT design in a 2×2 table.

| | Non-transmitted allele | | |
|---|---|---|---|
| **Transmitted allele** | **$M_1$** | **$M_2$** | **Total** |
| **$M_1$** | *a* | *b* | *a+b* |
| **$M_2$** | *c* | *d* | *c+d* |
| **Total** | *a+c* | *b+d* | *2n* |

The letters (*a, b, c, d*) represent the counts of over-transmissions of an allele from the parents to affected offsprings. The number *n* denotes total number of affected offsprings and 2*n* represents the total number of parents.
doi:10.1371/journal.pone.0063526.t001

Disorders can often have genotype information from only one parent of the affected individuals. As a common practice, these trios are simply discarded [42] though this can result in considerable loss of information and bias to the association study [20,38]. Several studies have been proposed to allow TDT to handle missing parental genotypic information [20,22,43–50]. Within these studies, the missing parental genotypes are mostly reconstructed based on the assumption that they are missing completely at random and do not depend on the genotypes themselves [51]. However, this assumption may not hold true and the probability that a genotype is missing may rely on the unobserved alleles [38,52,53]. Furthermore, these approaches are not designed for pedigrees with missing genotypes on the proband when both linkage and association data are available.

Against this background, we note this is a two step procedure. In the first step, the SNP data is used with both parents missing, so that the analysis depends on external allele frequency estimates and is sensitive to population stratification. In the second step, individual genotyping is performed on the parents for the top SNPs from step one, so this step is a traditional TDT and insensitive to the potential biases in step one. Accordingly, having available family DNA is needed to avoid false positive results.

## Methods

As stated above, the traditional TDT requires complete genotypic information from all members of the nuclear families. However, obtaining all genotypes cannot always be feasible for some diseases or families. Therefore the traditional TDT-type studies may not be useful to identify the presence of genetic determinants in data with relatively small amounts of complete trio information. One way to solve this type of issue is to reconstruct the missing parental genotypes under the assumption that they follow the probability distribution of the fully observed cases. However, most studies designed for this purpose do not incorporate the impact from LD and thus may introduce bias to the results. On the other hand, there are data available that have genotype information on both microsatellite and SNP markers for diseases; one example is alcoholism [54]. With both the linkage and association data, usually the microsatellite genotypic information from families for linkage analysis together with SNP data from the offspring who are genotyped for both linkage and

GWAS, we can "infer" the transmission of SNPs for the rest of the family members who have not been genotyped on SNPs. We call these family members the "missing individuals". In detail, first we generate the combined pedigrees in which each individual has both the linkage and GWAS genotype data filled in. Genotypes that those individuals do not have will be taken as missing data in the combined pedigrees. Then we use the program MERLIN [55] to read these combined pedigrees as input and infer the dosage probabilities of dense SNP genotypes for these missing individuals (see section *Genotype Inference of Familial Individuals* for more details). All the trio combinations from the inferred pedigrees are extracted on the condition that the children were affected and at least one parent in the trio was genotyped on microsatellite markers. The dosage-TDT that we have developed in this study is applied on these trio pedigrees using their inferred dosage probabilities. By incorporating the family linkage information into the GWAS data, we can potentially have higher power to detect association between our genotypic markers and the disease susceptibility alleles.

## Dosage Transmission Disequilibrium Test (dTDT)

**Common map of both linkage & GWAS data.** The common map of both linkage and GWAS data is designed in the following way. With the genetic position of the linkage markers (microsatellite as in here) and physical position of both linkage and GWAS markers (microsatellite and SNPs), the genetic positions of all the GWAS markers (SNPs) are calculated based on equation (2):

$$gm_{SNP_i} = \begin{cases} gm_{MS_1} - \dfrac{pm_{MS_1} - pm_{SNP_1}}{10^6} & (pm_{SNP_i} < pm_{MS_1}) \\ gm_{MS_{j+1}} - (gm_{MS_{j+1}} - gm_{MS_j}) \cdot \dfrac{pm_{MS_{j+1}} - pm_{SNP_i}}{pm_{MS_{j+1}} - pm_{MS_j}} & \\ \quad (pm_{MS_j} < pm_{SNP_i} < pm_{MS_{j+1}}, 1 \le j \le last-1) & \\ gm_{MS_{last}} + \dfrac{pm_{SNP_i} - pm_{MS_{last}}}{10^6} & (pm_{SNP_i} > pm_{MS_{last}}) \end{cases} \quad (2)$$

where *gm* denotes the genetic position of a marker, in *centi-Morgan* (*cM*) unit; *pm* denotes the physical position of a marker, in *base pair* units; $MS_{last}$ is the last microsatellite marker on a chromosome, $pm_{MS_{j+1}} \le pm_{MS_{last}}$. Because at both ends of a chromosome when the physical position of a SNP is either smaller than the 1st microsatellite marker or larger than the last microsatellite marker, there is only one microsatellite marker that can be referred to compute the genetic position for the SNP. We simply use the convention that $1cM \approx 10^6$ base pairs to convert a SNP's physical map to its genetic map. While a SNP is in between two microsatellite markers, we use the ratio $\dfrac{pm_{MS_{j+1}} - pm_{SNP_i}}{pm_{MS_{j+1}} - pm_{MSj}}$ and multiply this ratio with the genetic distance between these two microsatellite markers ($gm_{MS_{j+1}} - gm_{MSj}$). In this way we compute the *relative* genetic position of a SNP marker to the microsatellite marker that's next to it.

**Genotype inference of familial individuals.** Initially, many approaches implicitly imputed missing genotypes based on the potential genotype distribution in a family [56–58]. In practice, the genetic linkage implied that family members share a certain degree of similarity through their "identical-by-descent" (IBD) regions on the chromosomes. In this way, genotypes of the non-typed markers for these family members can be inferred according to their shared IBD with the other relatives. **Figure 1** illustrates the procedure of this genotype inference. As shown in the figure, a subset of microsatellite markers has been typed for all the family members except the founders (red), whereas both microsatellite and SNP markers have been typed in only a few selected common individuals (black). Genotypes of the dense SNPs for missing individuals can be inferred by comparing the haplotypes that are IBD with the other individuals in the family. Several studies have been published on the genotype imputation procedures described above [59,60]. These procedures are implemented in programs such as MERLIN [55,61] and MENDEL [62,63], using one of the pedigree analysis algorithms such as the Lander-Green [64] or Elston-Stewart [65] algorithms, or Monte Carlo sampling [66,67]. Merlin uses sparse trees to represent gene flow in pedigrees and is considered as one of the fastest packages among packages implementing the same algorithms such as Allegro [68] and Genehunter [69]. In this study, we use MERLIN to infer the dosage probabilities. The output of this program includes the most likely genotypes, the expected number of copies for the tested alleles (0, 1, or 2 with genotype observed), and the posterior probabilities (dosage probabilities) of the three alternative genotypes [70]. Because a large number of related individuals are included, this family-based genotype inference is expected to improve the power of association tests [59]. Furthermore, when a GWA scan follows a linkage study, only a proportion of individuals may need to be genotyped and the inferred genotypes can be useful for the next step in the association analysis.

**Dosage-TDT.** Because the traditional TDT is a simple representation of the $\chi^2$ statistics, it requires single counts of the transmitted/non-transmitted alleles from the heterozygous parents to the affected offspring. Thus the inferred dosage probabilities cannot be processed through this setup. In this study, we generalize the original TDT by taking all possible allele transmissions in a pedigree into account. **Table 2** shows the dosage probabilities of three alternative genotypes (1/1, 1/2 and 2/2) in a trio (named a *trio-dosage set* in this work) from the inferred results. **Table 3** lists all 11 TDT-informative allele transmissions in a trio where at least one of the parents is heterozygous. The values of $b_i$ and $c_i$ used in the $\chi^2$ calculation of the TDT in each trio $i$ are calculated by summing up the probabilities across all these 11 types of transmissions. Let $t$ denote the probability that allele 1 is transmitted by a heterozygote parent of an affected child. We can then write the dosage probabilities of a child in terms of the dosage probabilities of its parents and $t$ as follows:

$$\begin{aligned} p_{c11} &= P(c_{11}|f_{11},m_{12}) \cdot P(f_{11},m_{12}) + P(c_{11}|f_{12},m_{11}) \cdot P(f_{12},m_{11}) \\ &\quad + P(c_{11}|f_{12},m_{12}) \cdot P(f_{12},m_{12}) + P(c_{11}|f_{11},m_{11}) \cdot P(f_{11},m_{11}) \\ &= p_{f11}p_{m12} \cdot t + p_{f12}p_{m11} \cdot t + p_{f12}p_{m12} \cdot t^2 + p_{f11}p_{m11} \end{aligned} \quad (3)$$

$$\begin{aligned} p_{c12} &= P(c_{12}|f_{11},m_{12}) \cdot P(f_{11},m_{12}) + P(c_{12}|f_{12},m_{11}) \cdot P(f_{12},m_{11}) + \\ &\quad P(c_{12}|f_{12},m_{12}) \cdot P(f_{12},m_{12}) + P(c_{12}|f_{12},m_{22}) \cdot P(f_{12},m_{22}) + \\ &\quad P(c_{12}|f_{22},m_{12}) \cdot P(f_{22},m_{12}) + P(c_{12}|f_{11},m_{22}) \cdot P(f_{11},m_{22}) + \\ &\quad P(c_{12}|f_{22},m_{11}) \cdot P(f_{22},m_{11}) \\ &= p_{f11}p_{m12} \cdot (1-t) + p_{f12}p_{m11} \cdot (1-t) + 2p_{f12}p_{m12} \cdot t(1-t) + \\ &\quad p_{f12}p_{m22} \cdot t + p_{f22}p_{m12} \cdot t + p_{f11}p_{m22} + p_{f22}p_{m11} \end{aligned} \quad (4)$$

$$p_{c22} = P(c_{22}|f_{12},m_{12})\cdot P(f_{12},m_{12}) + P(c_{22}|f_{12},m_{22})\cdot P(f_{12},m_{22}) + P(c_{22}|f_{22},m_{12})\cdot P(f_{22},m_{12}) + P(c_{22}|f_{22},m_{22})\cdot P(f_{22},m_{22}) \quad (5)$$
$$= p_{f12}p_{m12}\cdot(1-t)^2 + p_{f12}p_{m22}\cdot(1-t) + p_{f22}p_{m12}\cdot(1-t) + p_{f22}p_{m22}$$

Thus, the frequencies of allele 1 and 2 of a child appearing in a trio are as follows:

$$p_{c1} = p_{c11} + \frac{1}{2}p_{c12} = p_{f11}p_{m12}\cdot\frac{1+t}{2} + p_{f12}p_{m11}\cdot\frac{1+t}{2} + p_{f12}p_{m12}\cdot t + p_{f11}p_{m11} + p_{f12}p_{m22}\cdot\frac{t}{2} + p_{f22}p_{m12}\cdot\frac{t}{2} + \frac{1}{2}p_{f11}p_{m22} + \frac{1}{2}p_{f22}p_{m11} \quad (6)$$

$$p_{c2} = \frac{1}{2}p_{c12} + p_{c22} = p_{f11}p_{m12}\cdot\frac{1-t}{2} + p_{f12}p_{m11}\cdot\frac{1-t}{2} + p_{f12}p_{m12}\cdot(1-t) + \frac{1}{2}p_{f11}p_{m22} + \frac{1}{2}p_{f22}p_{m11} + p_{f12}p_{m22}\cdot(1-\frac{t}{2}) + p_{f22}p_{m12}\cdot(1-\frac{t}{2}) + p_{f22}p_{m22} \quad (7)$$

$$\text{Additionally, } p_{c11} + p_{c12} + p_{c22} = 1 \quad (8)$$

Based on equations (3) to (8), we can derive that:

$$t = \frac{(2p_{c11} + p_{c12}) - (p_{f11} + p_{m11})}{p_{f12} + p_{m12}} \quad (9)$$

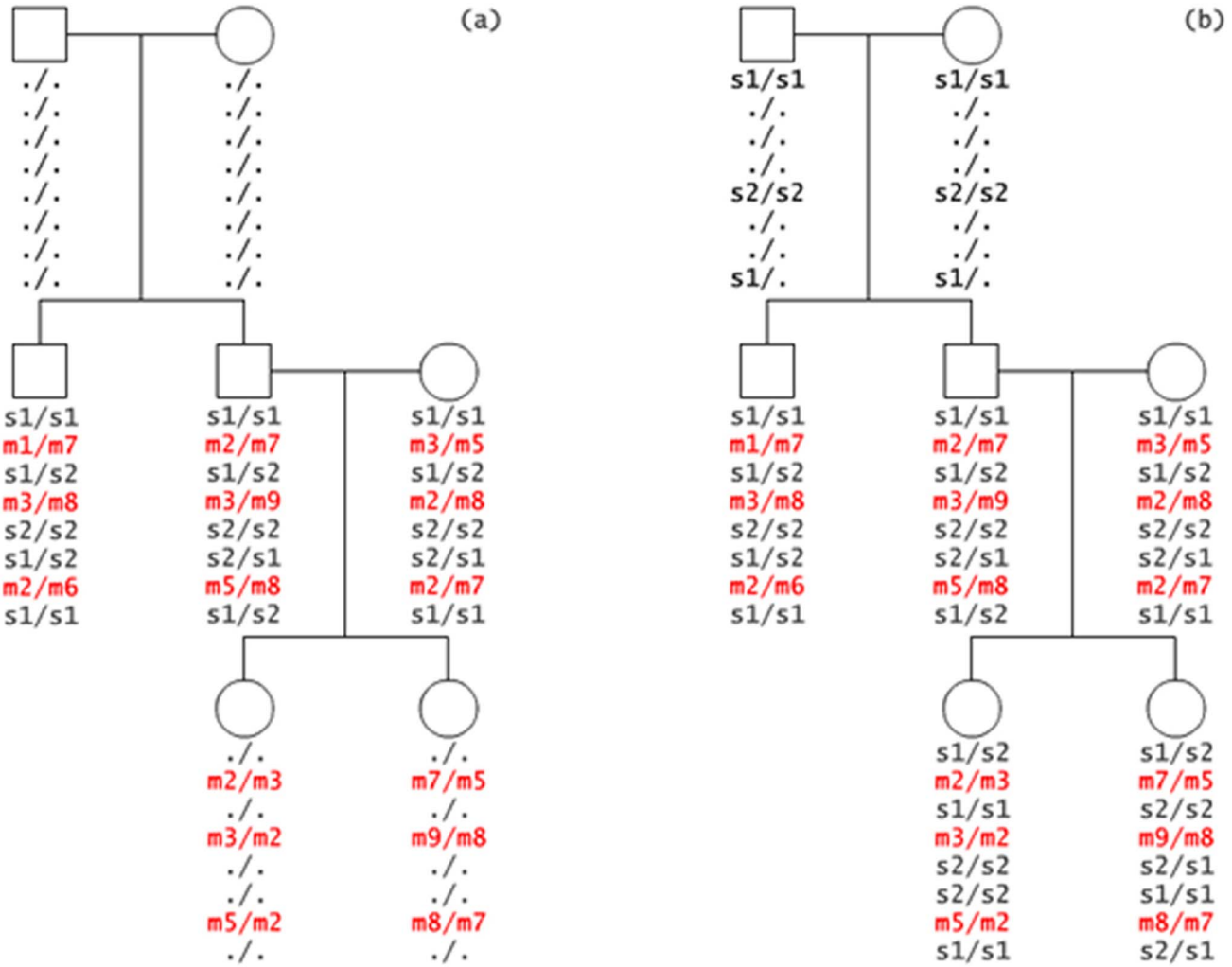


**Figure 1. Demonstration of genotype inference within a family.** (a) The observed data, which consist of genotypes at a series of microsatellite and SNP markers. A subset of microsatellite markers has been typed in all individuals except for founders (red), whereas both microsatellite and SNP markers have been typed in only a few selected common individuals (black). (b) Genotypes of dense SNPs for missing individuals are inferred by comparing the haplotypes they share with the common individuals.
doi:10.1371/journal.pone.0063526.g001

**Table 2.** Dosage probabilities in a trio (denoted as a trio-dosage set).

| | Genotype | | |
|---|---|---|---|
| | 1/1 | 1/2 | 2/2 |
| Father | $p_{f11}$ | $p_{f12}$ | $p_{f22}$ |
| Mother | $p_{m11}$ | $p_{m12}$ | $p_{m22}$ |
| Child | $p_{c11}$ | $p_{c12}$ | $p_{c22}$ |

doi:10.1371/journal.pone.0063526.t002

In summary, the sum of all 11 informative transmissions for $b_i$ and $c_i$ can be written as:

$$b_i = (p_{f12} + p_{m12}) \cdot t \tag{10}$$

$$c_i = (p_{f12} + p_{m12}) \cdot (1 - t) \tag{11}$$

where $i$ represents the $i^{\text{th}}$ trio. By substituting $t$ from equation (9) into (10) and (11) and summing up $b_i$ and $c_i$ for all trios, we can compute the $\chi^2$ values for each SNP:

$$\chi^2 = \frac{(\sum b_i - \sum c_i)^2}{\sum b_i + \sum c_i} \tag{12}$$

which follows a one degree freedom $\chi 2$ distribution under the null hypothesis of no association. We name this generalized TDT the dosage TDT (dTDT). The original TDT proposed by Spielman et al [18] can then be considered as a special case when t and the dosage probabilities in a trio-dosage set is 0 or 1.

The beauty of the above equations is the denominator from $t$ can be canceled out with the one in $b_i$ and $c_i$ thus equation (12) can be further written as:

$$\chi^2 = \frac{[\sum \delta_i - \sum (\varsigma_i - \delta_i)]^2}{\sum \delta_i + \sum (\varsigma_i - \delta_i)} \tag{13}$$

in which we denote

$$\delta_i = [(2p_{c11} + p_{c12}) - (p_{f11} + p_{m11})]_i \tag{14}$$

$$\varsigma_i = (p_{f12} + p_{m12})_i \tag{15}$$

for each trio $i$. Using form (13) can be computationally efficient.

## Simulation

The dTDT makes it possible to process the inferred dosage probabilities of the un-genotyped SNPs for those missing individuals in a nuclear family. As a follow-up study of this generalized TDT approach, we develop a simulation to investigate how the power changes for association detection with different inputs. In this simulation, we generate multiple sets of trios under various settings. Each set has 1,000 trios. Each trio has one affected child. Because we focus on the interaction between SNP and microsatellite markers, only one SNP marker and one microsatellite marker are simulated. In each trio, the microsatellite markers are assigned to both the parents and the child. SNPs are only assigned to the child. The parents who do not have such SNP markers are considered as the missing individuals and their SNP genotypes are inferred by MERLIN. The dTDT is then used to process the inferred dosage probabilities and p-values are reported from the $\chi^2$ statistics.

**Generating SNPs.** Denote the low and high risk alleles at a disease locus $D$ as $D_1$ and $D_2$, with population frequencies $p_1$ and $p_2$. Assuming Hardy-Weinberg equilibrium, the population prevalence ($K$) of the disease is

$$K = p_1^2 f_{11} + 2p_1 p_2 f_{12} + p_2^2 f_{22} \tag{16}$$

where $f_{11}$, $f_{12}$ and $f_{22}$ are the penetrances of the three genotypes $D_1D_1$, $D_1D_2$ and $D_2D_2$.

We have considered three disease models: dominant, recessive and co-dominant. The combinations of the penetrances in these three models are designed as follows: dominant ($f_{11} < f_{12} = f_{22}$), recessive ($f_{11} = f_{12} < f_{22}$) and co-dominant ($f_{11} < f_{12} = \frac{1}{2} f_{22}$).

With $K$ and $f$ predefined, we can compute $p_1$ and $p_2$ using the following equations:

$$p_1 = \frac{(f_{22} - f_{12}) - \sqrt{f_{12}^2 - f_{11} f_{22} + K \cdot (f_{11} - 2f_{12} + f_{22})}}{f_{11} - 2f_{12} + f_{22}} \tag{17}$$

$$p_2 = 1 - p_1 \tag{18}$$

Denote the haplotype frequencies of disease locus $D$ and SNP locus $S$ as $h_{11}$, $h_{12}$, $h_{21}$ and $h_{22}$. On condition of the child being

**Table 3.** Calculation of $b_i$ and $c_i$ in terms of dosage probabilities and $t$ for the $i^{\text{th}}$ trio with all 11 TDT-informative transmissions.

| | 1/1–1/2 \| 1/1 | 1/1–1/2 \| 1/2 | 1/2–1/1 \| 1/1 | 1/2–1/1 \| 1/2 | 1/2–1/2 \| 1/1 | 1/2–1/2 \| 1/2 | 1/2–1/2 \| 2/2 | 1/2–2/2 \| 1/2 | 1/2–2/2 \| 2/2 | 2/2–1/2 \| 1/2 | 2/2–1/2 \| 2/2 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $b_i$ | $p_{f11}p_{m12} \cdot t$ | | $p_{f12}p_{m11} \cdot t$ | | $2p_{f12}p_{m12} \cdot t^2$ | $2p_{f12}p_{m12} \cdot t(1-t)$ | | $p_{f12}p_{m22} \cdot t$ | | $p_{f22}p_{m12} \cdot t$ | | $(p_{f12}+p_{m12}) \cdot t$ |
| $c_i$ | | $p_{f11}p_{m12} \cdot (1-t)$ | | $p_{f12}p_{m11} \cdot (1-t)$ | | $2p_{f12}p_{m12} \cdot (1-t)$ | $2p_{f12}p_{m12} \cdot (1-t)^2$ | | $p_{f12}p_{m22} \cdot (1-t)$ | | $p_{f22}p_{m12} \cdot (1-t)$ | $(p_{f12}+p_{m12}) \cdot (1-t)$ |

$t$ denotes the possibility that allele 1 is transmitted by a heterozygote; and ($1$-$t$) is the possibility that allele 2 is transmitted.

doi:10.1371/journal.pone.0063526.t003

**Table 4.** Number of total and genotyped individuals, and corresponding number of families that these individuals are selected from.

| | Total Individuals ($ind_{total}$) | Number of Genotyped Individuals on microsatellite markers ($ind_{MS}$) | Number of Families |
|---|---|---|---|
| $Map03_{MS}$ EA | 2,037 | 1,926 (94.55%) | 219 |
| $Map03_{MS}$ AA | 335 | 283 (84.48%) | 35 |
| $Map03_{MS}$ Mixed | 87 | 74 (85.06%) | 8 |
| $Marshfield_{MS}$ EA | 1,530 | 1,090 (71.24%) | 234 |
| $Marshfield_{MS}$ AA | 570 | 347 (60.88%) | 77 |
| $Marshfield_{MS}$ Mixed | 6 | 5 (83.33%) | 1 |

doi:10.1371/journal.pone.0063526.t004

affected, the probabilities of different haplotypes in a child can be calculated through:

$$\left.\begin{aligned} p(H_{11}H_{11}|A) &= \tfrac{f_{11}h_{11}^2}{K} \\ p(H_{11}H_{12}|A) &= \tfrac{2f_{11}h_{11}h_{12}}{K} \\ p(H_{12}H_{12}|A) &= \tfrac{f_{11}h_{12}^2}{K} \end{aligned}\right\}$$

$$\left.\begin{aligned} p(H_{11}H_{21}|A) &= \tfrac{2f_{12}h_{11}h_{21}}{K} \\ p(H_{11}H_{22}|A) &= \tfrac{2f_{12}h_{11}h_{22}}{K} \\ p(H_{12}H_{21}|A) &= \tfrac{2f_{12}h_{12}h_{21}}{K} \\ p(H_{12}H_{22}|A) &= \tfrac{2f_{12}h_{12}h_{22}}{K} \end{aligned}\right\} \quad (19)$$

$$\left.\begin{aligned} p(H_{21}H_{21}|A) &= \tfrac{f_{22}h_{21}^2}{K} \\ p(H_{21}H_{22}|A) &= \tfrac{2f_{22}h_{21}h_{22}}{K} \\ p(H_{22}H_{22}|A) &= \tfrac{f_{22}h_{22}^2}{K} \end{aligned}\right\}$$

With a predefined correlation coefficient ($R$) of linkage disequilibrium (LD) between D and S, we can derive the haplotype frequencies as follows:

$$\left.\begin{aligned} h_{11} &= p_1q_1 + R\sqrt{p_1p_2q_1q_2} \\ h_{12} &= p_1q_2 - R\sqrt{p_1p_2q_1q_2} \\ h_{21} &= p_2q_1 - R\sqrt{p_1p_2q_1q_2} \\ h_{22} &= p_2q_2 + R\sqrt{p_1p_2q_1q_2} \end{aligned}\right\} \quad (20)$$

where $q_1$ and $q_2$ are the population frequencies of the SNP alleles $S_1$ and $S_2$. To simplify our model, we will assume $p_i = q_i$, where $i = 1$ or 2. The rationale behind this is that if the SNP marker and disease allele have very different frequencies, then $R^2$ is small and there is little power. Keeping both frequencies equal allow $R$ to vary the full range from $-1$ to +1.

By substituting (20) into (19), we can assign the genotypes ($S_1S_1$, $S_1S_2$, or $S_2S_2$) at locus $S$ to the affected children based on these derived frequencies:

$$\left.\begin{aligned} p(S_1S_1|A) &= \tfrac{f_{11}h_{11}^2 + 2f_{12}h_{11}h_{21} + f_{22}h_{21}^2}{K} \\ p(S_1S_2|A) &= \tfrac{2(f_{11}h_{11}h_{12} + f_{12}h_{11}h_{22} + f_{12}h_{12}h_{21} + f_{22}h_{21}h_{22})}{K} \\ p(S_2S_2|A) &= \tfrac{f_{11}h_{12}^2 + 2f_{12}h_{12}h_{22} + f_{22}h_{22}^2}{K} \end{aligned}\right\} \quad (21)$$

**Generating microsatellites.** Denote $M_i$ as the microsatellite marker from the parents. Because of a large number of polymorphisms (alleles) for a microsatellite marker, we assume that our microsatellite marker is completely informative (i.e., each parent is heterozygous at the microsatellite locus M) and assign alleles $M_1$ & $M_2$ to the father and $M_3$ & $M_4$ to the mother. Then we randomly select $M_1$ & $M_3$, $M_2$ & $M_4$, $M_1$ & $M_4$ or $M_2$ & $M_3$ equally with 0.25 probabilities as the microsatellite genotype for the child.

**Parameter settings.** Without losing biological meaning, i.e. with valid $p_i \in (0, 1.0]$ ($i = 1$ or 2 and $p_1 + p_2 = 1$), but also with a good coverage of possible natural phenomena, we predefine the following values for the parameters to generate each set of trios:

$N$: the number of trios = 1,000;

$K$: prevalence = 0.01, 0.1, or 0.2;

$R$: correlation coefficient of LD between $D$ and $S$ = 0.5, 0.7, 0.9, or 1.0 (as negative value of $R$ does not produce informative divergence from the result using positive value of $R$, we are only considering positive value of $R$ herewith);

$f$ or $g$: penetrance of disease genotype $D_iD_i$ or $D_iD_j$, where $i$ or $j = 1$ or 2 and $i \neq j$. To simplify the notation, here we use $f$ to denote $f_{11}$ and $g$ to denote $f_{12}$ or $f_{22}$. As noted, we separate the disease models as dominant ($f$, $g$, $g$), recessive ($f$, $f$, $g$), and co-dominant ($f$, $0.5g$, $g$) where $f = 0.0$, $0.1K$, $0.3K$, $0.5K$, $0.7K$, or $0.9K$ and $g = 1.1K$, 0.5, 0.7, 0.9, or 1.0.

These values are first permuted to generate all their possible combinations then any combination that produces invalid $p$, i.e.$p \notin (0,1.0]$, is excluded. Under each setting, we produce 1,000 trios based on equation (15) for SNPs in the affected offsprings. Parental SNPs are inferred by MERLIN and dTDT is used to process the inferred dosage probabilities.

## Application to Alcohol Dependence

Alcohol dependence is a serious psychiatric disorder in which an individual is characterized as having harmful consequences of repeated or compulsive alcohol use, and (sometimes) physiological dependence on alcohol (i.e., tolerance and/or symptoms of withdrawal) [71,72]. During 2001–2005, excessive alcohol use contributed to about 79,000 deaths and 2.3 million years of potential life lost in the United States [73]. Excessive alcohol


**Figure 2. Combined pedigree structure used in inference.** The pedigree has $ind_{total}$ individuals and ($mrkMS+mrk_{SNP}$) markers. Most individuals have been genotyped on microsatellite markers. $ind_{common}$ out of $ind_{total}$ individuals are selected for SNP genotyping. Microsatellite and SNP markers are mapped based on their genetic positions. Missing SNPs of ($ind_{total}$ - $ind_{common}$) will be inferred by MERLIN –infer.
doi:10.1371/journal.pone.0063526.g002

consumption, the third leading cause of preventable death in the United States, can cause damage to the central and peripheral nervous system, and to nearly every organ system in the body [74,75]. It is reported that alcohol dependence affects about 12% of American adults across their lifetime [76]. As a complex disease, alcohol dependence can be influenced by various factors such as genetic susceptibility, environmental influences and interactions among genes or between genes and environment.

The nine-site national Collaborative study On the Genetics of Alcoholism (COGA) funded by National Institute of Alcohol and Alcoholism (NIAAA) aims to identify and characterize genes that affect the susceptibility to develop alcohol dependence and related phenotypes. COGA is applying multiple strategies for genetic research. The most densely affected, multiplex alcoholic families were used in a multi-wave family-based linkage study. 2,283 out of 2,459 individuals from 262 families were genotyped using microsatellite markers in Wave I and Wave II (data denoted as Map03MS) (**Table 4**) [54]. At Wave III, another 1,442 out of 2,106 individuals from 312 families were selected for microsatellite genotyping by the Mammalian Genotyping Service (MGS) from Marshfield Clinic (data denoted as MarshfieldMS) (**Table 4**). Combined data from all three waves are denoted as LinkageMS in this study. COGA also has high throughput GWAS data with over 1 million SNP markers from 1,884 independent individuals, generated by the Center for Inherited Disease Research (CIDR) (data denoted as CIDRSNP) (Edenberg et al., 2010). The GWAS data include 566 mutual individuals chosen from the LinkageMS families.

All participants agreed to share their DNA and phenotypic information for research purposes and provided written informed consent following instructions from institutional review boards at all data collection sites. The study was approved by the institutional review board at each COGA site, with the OHRP Assurance Numbers being**:** FWA00003624 (SUNY Research Foundation), FWA00007125 (University of Connecticut), FWA00003544 (Indiana University), FWA00003007 (University of Iowa), FWA00004069 (Veterans Medical Research Foundation/UCSD), FWA00002284 (Washington University), FWA00003518 (Southwest Foundation for Biomedical Research), FWA00003913 (Rutgers, The State University of New Jersey) and FWA00005287 (Virginia Commonwealth University).

As described above, the dTDT uses the inferred dosage probabilities of dense SNPs for association detection. The COGA family data provides us such an opportunity to integrate the information from both linkage and association studies.

In this study, we first generate the combined pedigrees with both the LinkageMS and CIDRSNP genotype data from COGA. **Figure 2** demonstrates the structure of the combined pedigrees. Most individuals in the combined pedigrees were genotyped on microsatellite markers. A subset of individuals in the pedigrees was genotyped for dense SNPs. These individuals include one affected child in each of the families and other unrelated members chosen as a control group. All other individuals who have not been genotyped on SNPs are considered as the missing individuals. We use the program MERLIN [55] to read these combined pedigrees as input and infer the dosage probabilities of dense SNP genotypes for these missing individuals (described in detail below). All the trio combinations from the inferred pedigrees are extracted on the condition that the children were affected and at least one parent in the trio was genotyped on microsatellite markers. The dTDT is applied on these trio pedigrees using their inferred dosage probabilities. In addition, PLINK [77] is used to conduct a standard case-control study on the CIDRSNP data. With the idea that making use of all the available sample data would increase the power for association detection, we further combine the results from both dTDT and case-control study through the MH test [78].

**Data sets.** The Map03MS data have 219 European American (EA), 35 African American (AA) and eight mixed families. 2,283 individuals from these 262 families were genotyped on 328 microsatellite markers. The MarshfieldMS data contain 234 EA, 77 AA and one mixed families, with a total of 1,442 individuals genotyped on 394 microsatellite markers. 1,041,304 SNPs were genotyped for 1,399 EA and 485 AA individuals in the CIDRSNP GWAS. (**Tables 4, 5 and 6**).

With AA and mixed families excluded, we have 3,016 out of 3,567 EA individuals from 453 families genotyped on microsatellite markers in the LinkageMS data. 471 of these individuals in 398 linkage families were selected for SNP genotyping (known as the mutual individuals) (**Table 6**), including 41 individuals without microsatellite genotyping data. For GWAS, from each of these 398 families, one affected child (normally the proband) was selected as the case and other biologically unrelated family

**Table 5.** Number of common individuals in EA group genotyped on both microsatellite & SNP markers, and corresponding number of families.

| | Number of Common Individuals ($ind_{common}$) | Number of Families with Common Individuals | Number of Families with *m* Common Individuals | | | |
|---|---|---|---|---|---|---|
| | | | *m* = 1 | 2 | 3 | 4 |
| $Map03_{MS}$ × $CIDR_{SNP}$ EA | 260 | 208 | 167 | 32 | 7 | 2 |
| $Marshfield_{MS}$ × $CIDR_{SNP}$ EA | 211 | 190 | 169 | 21 | - | - |

In total, 471 (13.2%) out of 3,567 individuals were selected for SNP validation genotyping.
doi:10.1371/journal.pone.0063526.t005

member(s) were used as the control. In total, 1,399 EA CIDRSNP individuals consist of 847 cases and 552 controls. **Figure 3** shows the pedigree of one of the LinkageMS EA families (FAM_ID 20059). This family has four mutual individuals. Except for proband #1, all the other three (#2, 9, and 13) selected for GWAS are relatives by affinity to this family.

**Marker cleaning & mapping.** In order to match up the genetic positions of all microsatellite and SNP markers, 38 microsatellite markers in Map03 and 58 microsatellite markers in Marshfield were excluded because of their missing physical positions. ~200,000 SNPs with low minor allele frequency (≤5%) were excluded. In consideration of any possible impact from linkage disequilibrium (LD), we exclude ~1,500 SNPs that are within 1,000 base pairs flanking each microsatellite marker. We use equation (2) to create the common map for these microsatellite and SNP markers. A comparison of the numbers of microsatellite and SNP markers before and after cleaning is given in **Table 6**. **Figure 4** shows the distribution of these cleaned markers in EA families. With these cleaned and mapped markers, we create new pedigrees with the $Linkage_{MS}$ and $CIDR_{SNP}$ data combined together. One combined pedigree has $ind_{total}$ individuals (in rows) with ($mrk_{MS}$+$mrk_{SNP}$) markers (in columns). Missing SNPs of ($ind_{total}$ - $ind_{common}$) individuals were inferred by MERLIN. (**Figure 1**).

**dTDT and mantel-Haenszel test.** In this study, we apply both the dTDT and Mantel-Haenszel (MH) tests to the COGA data. The MH test was first proposed by Mantel and Haenszel in 1959 [78]. The method has been widely applied to analysis of contingency tables (normally 2 × 2) and comparison of results from different treatments. In case-control studies, a 2 × 2 table is typically used. The discrepancy between observed and expected values in each cell from the table is evaluated by $\chi^2$ test with one degree of freedom. Comparatively, because the dTDT only takes account of values of $b$ and $c$, the test can be constructed by a 1 × 2 table instead. To maximally benefit from all sample data and multiple studies, we extend the MH test to pool results on each SNP from these two contingency tables in both case-control study and dTDT. Calculation of each term in the MH test is shown in **Table 7**. Having the Observed & Expected values and Variances from case-control study and dTDT, terms in MH test can be written as the sums of corresponding values these two tests. The null hypothesis assumes no association between markers and disease.

## Results

### Simulation

We separate the simulation results into nine groups that are combinations of three disease models and three $K$ values (0.01, 0.1 and 0.2). In each group there are 100~120 settings with different $R$ and $f$ values. With each setting we generate 1,000 trios and replicate the inference and dTDT procedures. Because of the large number of these settings (1,320 in total), we attach the results as in supplement tables. A plot of the $\log_{10}(p-value)$for these models is shown in **Figure 5**. In the figure, graphs from the top row to the bottom row represent the dominant, recessive and codominant models respectively, and from left to right represent the results with three different $K$ values (0.01, 0.1 and 0.2). Each blue dot corresponds to a/under that specific setting.

Because there are many factors interacting with each other, we will start with a general comparison of different models then look at the impact from one or two factors while constraining the others constant.

In general, reading the values of $\log_{10}(p-value)$from each model, we find that a rare ($K$= 0.01) recessive disease model

**Table 6.** Summary of microsatellite and SNP markers in EA group before and after cleaning.

| | Total Number of Raw microsatellite markers | Number of microsatellite markers after cleaning ($mrk_{MS}$) | Number of CIDR SNPs after cleaning ($mrk_{SNP}$)* |
|---|---|---|---|
| $Map03_{MS}$ EA | 328 | 290 | 801,273 |
| $Marshfield_{MS}$ EA | 394 | 336 | 801,286 |

*compared to 1,041,304 SNPs before cleaning.
doi:10.1371/journal.pone.0063526.t006

produces higher power compared to common ($K = 0.1$ or 0.2), co-dominant or dominant disease models. Meanwhile, high $R$ value (0.9 or 1.0) also helps increase dTDT's ability to detect signals. This is because in a rare recessive case, markers with high LD to the disease allele both parents are heterozygous and both transmit the recessive risk allele to their offspring. Our findings from the simulation validate what we already observed in the biological phenomenon.

In the figure, each graph is broken down into four bins having R = 0.5, 0.7, 0.9 and 1.0. Interestingly, within each bin, when $\log_{10}(p-values)$ are ordered by descending $f_{11}$ (from 0.9K to 0.0) and increasing $f_{12}$ & $f_{22}$ (from $1.1K$ to 1.0), it shows a noticeable increasing trend as shown on the graphs. Meanwhile, when $f_{22}$ is small ($= 1.1K$), the blue dots are close to the bottom line on each graph. We note that in order to have enough power to detect the signals, we need to have relatively distinguishable penetrances (i.e. $f$ cannot be too close to $g$) in the model. Indeed, there is no information when all three penetrances are equal to $K$.

When we generate the trios, we use a roulette wheel algorithm to assign SNPs to the children. This randomness is reflected on the graphs as the dots spread in some irregular patterns. Reading the graphs from left to right, we can see that with low prevalence ($K = 0.01$) the dots appear in clear clusters. Each cluster corresponds to a specific $f_{11}$ value. Taking the top left graph (dominant with $K = 0.01$) as an example, $f_{11}$ changes in the order of $0.9K$, $0.7K$, $0.5K$, $0.3K$, $0.1K$ and 0.0. Within each cluster, $f_{12}$ & $f_{22}$ increase in the order of $1.1K$, 0.5, 0.9 and 1.0. This clustering holds true in the other two disease models (recessive and co-dominant) when $K$ is small ($K = 0.01$) except some $f_{11}$ valued clusters are missing because combinations with invalid $p$ were excluded. In summary, the above observation indicates that $f_{11}$ has a higher impact to the power than $f_{12}$ & $f_{22}$ do in a rare disease model. As the prevalence increases ($K = 0.1$ and 0.2), the clustering effect gradually disappears. In each $R$ valued bin when $K$ is large (0.1 or 0.2), though the penetrances are sorted in the same order as mentioned above, the dots represent certain continuity instead of clustering. This shows that, in a common disease model, $f_{11}$ is not the only or the most effective factor as it is in a rare disease model. Other factors start to interact with each other. Especially when $K = 0.2$ and $R = 1.0$ (the fourth bin in the three graphs on the right), the dots appear in clear fan-shaped sectors. This irregularity can be partially explained by the sensitivity to randomness of the model under such setting, i.e., small changes of the parameters can have high impact on the results.

We can further see how the penetrances differ by looking at the slope of the trend in each bin. Apparently as the value of $R$ increases across the bins, the slope of the trend also increases. This is because when $R$ is high (such as 0.9 and 1.0), the same degree of lift in the penetrances will add more power and move/more quickly to its next level compared to the situation when $R$ is low (such as 0.5). From another point of view, we can imagine these slopes as the (first) derivatives of a convex function in terms of $R$. On this convex curve, as $R$ moves along to its rightmost end (increases), the derivative of the function increases and the function value (power) improves faster.

## Application to Alcohol Dependence

In a recent work in a case-control study using GWAS data on the COGA sample, Edenberg et al [79] identified the most significant SNP rs10511260 on chromosome 3 with $p$-value $(P) = 3.4\times10^{-6}$. A cluster of SNPs was found in a region of chromosome 11 with $p$-values ranging from $4.8\times10^{-5}$ to $6.9\times10^{-4}$. No single SNP showed genome-wide significance $(5\times10^{-8})$. In the following sections, we will compare our results from dTDT on COGA data with these findings from Edenberg et al's work.

**dTDT on COGA data.** To apply the dTDT on each SNP from COGA, we re-build the inferred pedigrees by extracting all trio combinations in which every child in a trio must be affected and at least one parent was genotyped on microsatellite markers.

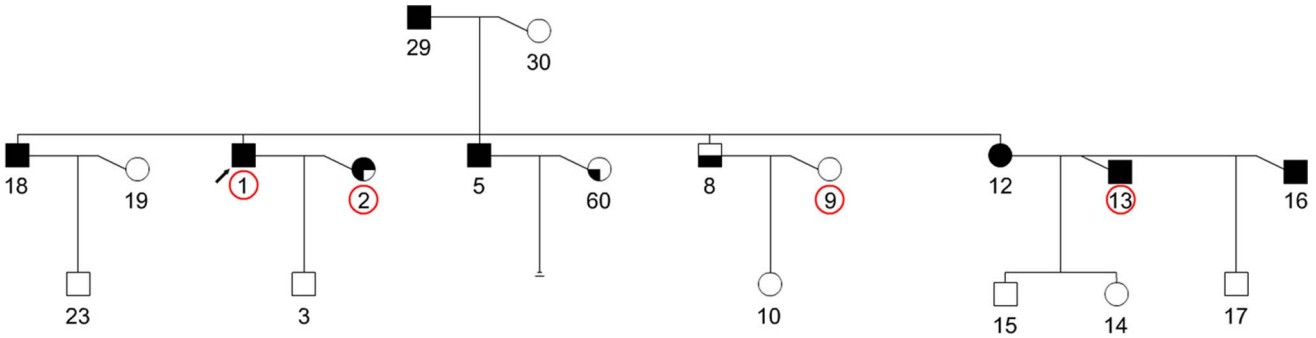


**Figure 3. Pedigree of one Map03 family (FAM_ID 20059).** Common individuals (#1, 2, 9 and 13, from left to right) are genotyped on both microsatellite and SNP markers (circled in red). Box shadowed in upper left: AB, alcohol abuse; shadowed in upper left & right: AD, alcohol dependence (DSM III-R Diagnosis). Box shadowed in lower left: PROB, probable; shadowed in lower left & right: DEF, definite (Feighner Diagnosis).
doi:10.1371/journal.pone.0063526.g003

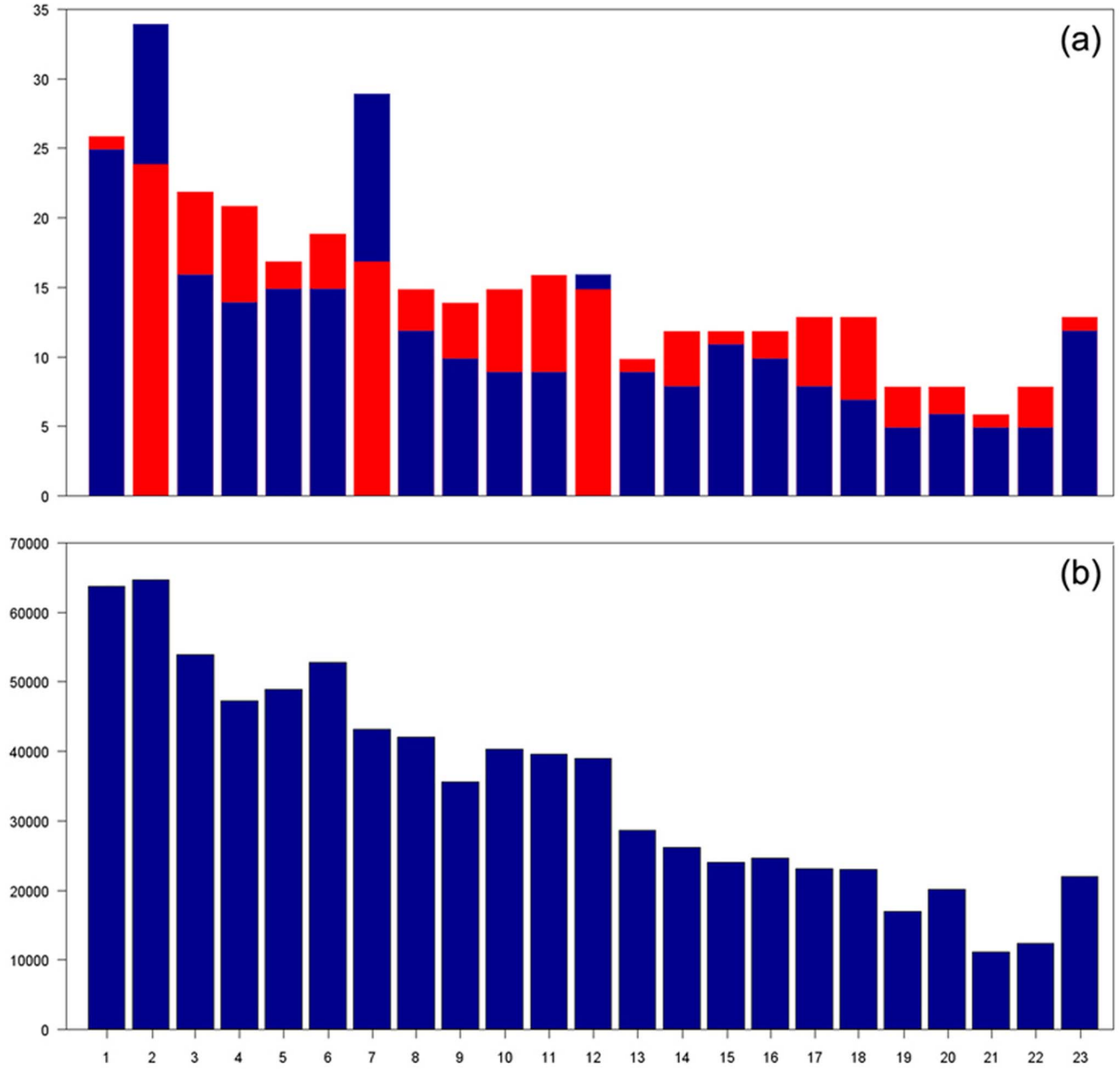


**Figure 4. Distribution of cleaned microsatellite and SNP markers on 23 chromosomes in EA families.** These markers are used in the combined pedigrees for genotype inference. (a) distribution of microsatellite markers on 23 chromosomes in Map03 (blue) and Marshfield (red) (overlapped); (b) distribution of SNPs on 23 chromosomes in Map03 and Marshfield. Because the difference of SNPs numbers in these two datasets is trivial, we only display the distribution of SNPs in Map03.
doi:10.1371/journal.pone.0063526.g004

Because one family can have more than one affected child, the parents can be found in more than one trio. For instance, the family (FAM_ID 20059) as shown in **Figure 3** has four affected children (#18, 1, 5 and 12, from left to right). Mother (#30) was genotyped on microsatellite markers. Therefore we have four trios from this family. In total, 893 trios from 323 families with 1,654 individuals in the $\text{Linkage}_{\text{MS}}$ EA group were extracted and used to build the trio-dosage pedigrees. 166 SNPs are found with $p$-values $<10^{-4}$. This is compared to 93 SNPs at the same level in Edenberg et al's paper [79]. Several clusters of SNPs on chromosome 7, 8 and 22 have $p$-values $\leq 10^{-5}$. However, further analysis with MH statistics does not yield consistent results with dTDT. This may be primarily due to the relatively small sample size used in dTDT and accuracy of the "–infer" program.

**Combining case-Control study and dTDT.** With dTDT and case-control analysis applied to the $\text{Linkage}_{\text{MS}}$ and $\text{CIDR}_{\text{SNP}}$ data respectively, we compute the $p$-values of MH test based on calculations in **Table 7**. **Figure 6** shows the Manhattan plots of $-log_{10}P$ across all 23 chromosomes from the MH test. The most significant SNP (rs11583322) on chromosome 1 gives a $p$-value $= 1.10\times10^{-8}$ that meets the GWAS significance level. This SNP lies in the gene Serine/Threonine Kinase 40 (*STK40*) that connects pluripotency factor Oct4 to the Erk/MAPK pathway controls extraembryonic endoderm differentiation [80]. **Table 8**

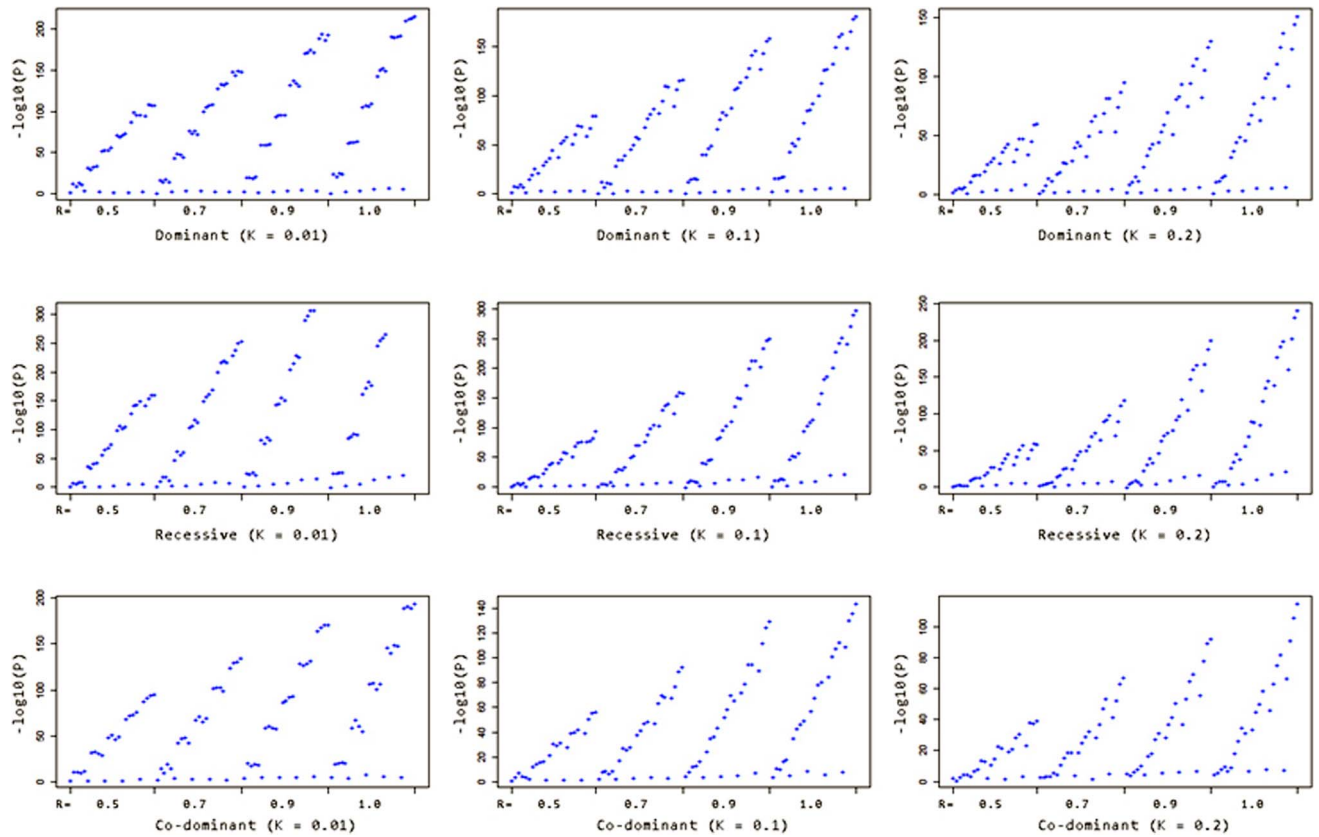


**Figure 5. –log$_{10}$(*P*) distribution of nine disease models from the simulation results.** Graphs from the top row to the bottom row represent the dominant, recessive and codominant models respectively, and from left to right represent the results with three different *K* values (0.01, 0.1 and 0.2). Each blue dot corresponds to a –log$_{10}$(*p-value*) under that specific setting. Every graph is broken down into four bins with R = 0.5, 0.7, 0.9 and 1.0. Within each bin, the –log$_{10}$(*p-values*) are ordered by descending $f_{11}$ and increasing $f_{12}$ & $f_{22}$.
doi:10.1371/journal.pone.0063526.g005

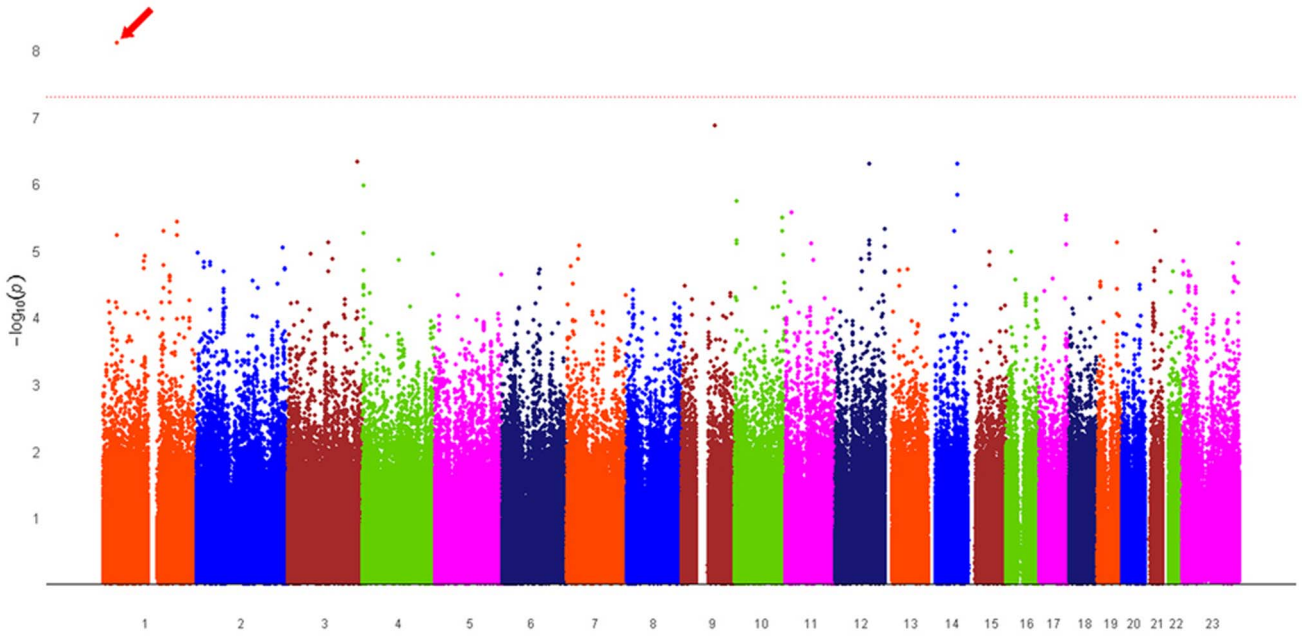


**Figure 6. Manhattan plot of –log$_{10}$P from MH test across all 23 chromosomes for Linkage$_{MS}$ EA data.** The dashed red line shows the genome-wide significant level $-\log_{10}(5 \times 10^{-8}) \approx 7.3$. SNP rs11583322 has given $-\log_{10}P \approx 8.1$ above this level that lies in gene STK40.
doi:10.1371/journal.pone.0063526.g006

**Table 7.** Calculations of Observed & Expected values, Variance, $\chi^2$ test in Case-Control study, dTDT and MH test.

| | Case-Control study | | | dTDT | | | | MH test |
|---|---|---|---|---|---|---|---|---|
| **Structure** | | **Case** | **Control** | | | **NT** | | |
| | | | | | | **allele 1** | **allele 2** | – |
| | **# allele 1** | $a_1=2\times\#cs\times f_{cs}$ | $b_1=2\times\#cn\times f_{cn}$ | **T** | **allele 1** | – | $b_2$ | |
| | **# allele 2** | $c_1=2\times\#cs-a_1$ | $d_1=2\times\#cn-b_1$ | | **allele 2** | $c_2$ | – | |
| Total (N) | $a_1+b_1+c_1+d_1$ | | | – | | | | $a_1+b_1+c_1+d_1$ |
| Observed (O) | $a_1(\text{or}c_1)$ | | | $b_2(\text{or}c_2)$ | | | | $(a_1+b_2)\text{or}(a_1+c_2)$* |
| Expected (E) | $\frac{(a_1+b_1)\times(a_1+c_1)}{N}$ | | | $\frac{(b_2+c_2)}{2}$ | | | | $\frac{(a_1+b_1)\times(a_1+c_1)}{N}+\frac{(b_2+c_2)}{2}$ |
| Variance | $\frac{(a_1+b_1)\times(a_1+c_1)\times(c_1+d_1)\times(b_1+d_1)}{N^2(N-1)}$ | | | $\frac{(b_2+c_2)}{4}$ | | | | $\frac{(a_1+b_1)\times(a_1+c_1)\times(c_1+d_1)\times(b_1+d_1)}{N^2(N-1)}+\frac{(b_2+c_2)}{4}$ |
| $\chi^2$ test | $\frac{(O-E)^2}{V}$ | | | $\frac{(b_2-c_2)^2}{(b_2+c_2)}$ | | | | $\frac{(O-E)^2}{V}$ |

Number of Cases ($\#cs$) = 847; Number of Controls ($\#cn$) = 552; $f_{cs}$: allele frequency in cases; $f_{cn}$: allele frequency in controls; T is short for Transmitted; NT is short for Non-Transmitted. $b_2=\Sigma b_i$ and $c_2=\Sigma c_i$. Using either $a_1$ or $c_1$ in Case-Control study will give the same results.
*equivalent to $(c_1+b_2)\text{or}(c_1+c_2)$.
doi:10.1371/journal.pone.0063526.t007

lists the top SNPs with MH test $p$-values $<10^{-5}$ and their corresponding case-control study and dTDT $p$-values. From the table we can see that the $p$-value of each SNP in the MH test is approximately the product of $p$-values in the other two tests. However, SNPs with high rankings by $p$-values in the MH test do not systematically correspond to high rankings in the individual tests. The $p$-values in the dTDT share the highest variance ($2.01\times10^{-3}$) among the three tests because of the randomness introduced by the inference procedure as well as the difference in sample sizes across the tests. A larger sample size will likely increase the power and generate more robust test results. In total, we have 257 SNPs in 75 genes with $p$-values $<10^{-4}$. 14 SNPs at the same level of $p$-values are found in replication of Edenberg et al's study [79]. Four of these 14 SNPs have associated genes: *CSMD2* on chromosome 1, *LZTS2* & *PDZD7* on chromosome 10, and *Gcom1* on chromosome 15. There are 34 vs. 11 SNPs that have $p$-values $<10^{-5}$. Five SNPs across chromosome 1, 3, 9, 12 and 14 show $p$-values $<5.1\times10^{-7}$ which is more statistically significant than the case-control analysis by Edenberg et al [79]. Our results also show clusters of SNPs by distance with $p$-values $<10^{-5}$ (more than five such markers in one cluster) in genes *EXOC6B*, *FTO*, *NCAM2* and *PPEF1* on chromosome 2, 5, 21 and X, respectively.

### Experimental Validation

Based on results from the MH test (5th column on **Table 8**), 19 out of the top 30 SNP markers were genotyped for 1,586 individuals from 220 Wave I & II families. We used the Sequenom MassArray technology for SNP genotyping [81]. PCR primers, extension primers, and multiplexing capabilities were determined with Sequenom MassARRAY Assay Designer software v3.1.2.2. Standard procedures were used to amplify PCR products; unincorporated nucleotides were deactivated with shrimp alkaline phosphatase. A single base pair extension step was completed with the mass extension primer and the terminator (iPLEX). The primer extension products were cleaned with resin and spotted onto a silicon SpectroChip. The chip was scanned with a mass spectrometry workstation (Bruker). The resulting genotype spectra were analyzed with Sequenom SpectroTYPER software v3.4.

Because variant rs11583322 did not work well with the Sequenom genotyping platform, we used the PrimerPicker software [82] to design the assay and followed the protocol described in KASPar SNP Genotyping System manual to run PCR reaction with an ABI GeneAmp PCR System 9700 [83]. Genotypes were accessed using an ABI 7900 HT Fast Real-Time PCR system. Because the genotypes are from linkage families, we used the program UNPHASED [84] to perform a genetic association analysis. Our colleagues in Allison Goate's lab implemented the above genotyping process. The author did the final analyses of the genotypes. Results are shown in **Table 8**.

## Discussion

### Simulation

As shown above, we can see that simulation can be a powerful tool to investigate many interactions between various factors and help discover potential rules underlying these factors.

With slight modification of the above technique, we can use our simulation to investigate how the dTDT is affected by population stratification. Simulating different populations to have different prevalences, we can choose two sets of trios using different allele frequencies. Applying the dTDT to this combined set of trios, we can test whether the power is lowered or heightened because of the prevalence difference within the populations. Since there is no information on phase of two markers in a trio, we have not introduced the recombination frequency ($\theta$) in the simulation.

### Tradeoff

When applying the dTDT to the alcohol dependence data from COGA, nearly twice the number of SNPs (166 vs. 93) were found having $p$-values $<10^{-4}$. Further, to maximally make use of the available sample data, we combine case-control study and dTDT with the MH test. This potentially increases our sample size and makes the method more robust to uncontrolled factors. As a result, we have one signal in gene STK40 with $p$-value that attains a genome-wide significance level. A large number of SNPs are found having $p$-values $<10^{-4}$. Several clusters of SNPs by distance with $p$-values $<10^{-5}$ are found in various genes across the genome.

**Table 8.** Top SNPs with *p*-values $<10^{-5}$ from MH test, and corresponding genes and *p*-values from Case-Control study, dTDT, Unphased association analysis, and IQS with threshold on dosage probabilities above 0.0 or 0.8.

| SNP | CHR | Position | Associated Gene | MH test *p*-value | Case-Control *p*-value | dTDT *p*-value | MAF | Unphased Wave I&II | IQS (>0) | IQS (>.8) |
|---|---|---|---|---|---|---|---|---|---|---|
| rs12116935 | 1 | 36,562,133 | *FAM176B* | 6.98E-06 | 6.56E-04 | 1.04E-03 | 0.38 | 6.77E-01 | 0.21 | 0.91 (395) |
| rs11583322 | 1 | 36,594,899 | *STK40* | 1.10E-08 | 7.39E-06 | 1.04E-04 | 0.38 | 8.54E-01 | 0.20 | 0.85 (345) |
| rs1932933 | 1 | 160,384,670 | *NOS1AP* | 5.82E-06 | 1.99E-04 | 4.24E-03 | 0.37 | 6.85E-01 | 0.22 | 0.95 (478) |
| rs10801629 | 1 | 196,110,990 | | 4.13E-06 | 1.18E-04 | 9.62E-03 | 0.40 | | | (482) |
| rs10922323 | 1 | 196,128,944 | | 6.46E-06 | 1.52E-04 | 1.23E-02 | 0.40 | 9.54E-01 | 0.31 | 0.91 (550) |
| rs1850344 | 3 | 108,667,763 | | 8.16E-06 | 1.33E-04 | 2.08E-02 | 0.38 | 1.13E-01 | 0.21 | 0.92 (350) |
| rs4384980 | 3 | 183,941,763 | | 5.78E-07 | 9.17E-05 | 6.93E-04 | 0.42 | 4.46E-01 | 0.21 | 0.90 (405) |
| rs2857839 | 4 | 3,006,428 | *GRK4* | 6.60E-06 | 1.11E-03 | 5.56E-04 | 0.39 | | | (471) |
| rs1801058 | 4 | 3,008,948 | *GRK4* | 6.48E-06 | 9.67E-04 | 6.57E-04 | 0.39 | 5.62E-01 | 0.22 | 0.89 (564) |
| rs2798303 | 4 | 3,010,385 | *GRK4* | 1.28E-06 | 2.17E-04 | 8.55E-04 | 0.42 | 8.24E-01 | 0.22 | 0.91 (615) |
| rs994029 | 9 | 88,565,134 | | 1.98E-07 | 3.62E-04 | 1.18E-05 | 0.37 | 3.59E-01 | 0.21 | 0.89 (713) |
| rs2398236 | 10 | 5,321,159 | | 9.07E-06 | 7.15E-04 | 1.96E-03 | 0.40 | 5.04E-01 | 0.22 | 0.90 (362) |
| rs9423593 | 10 | 5,322,349 | | 8.32E-06 | 1.04E-03 | 5.20E-04 | 0.37 | 9.04E-01 | 0.22 | 0.89 (381) |
| rs7076488 | 10 | 5,323,008 | | 2.13E-06 | 2.03E-04 | 1.49E-03 | 0.41 | | | (562) |
| rs3781458 | 10 | 126,333,921 | *FAM53B* | 4.27E-06 | 2.70E-03 | 1.31E-05 | 0.37 | | | (729) |
| rs3781452 | 10 | 126,345,119 | *FAM53B* | 6.86E-06 | 3.43E-03 | 1.81E-05 | 0.37 | | | (449) |
| rs1503452 | 11 | 16,408,708 | *SOX6* | 3.21E-06 | 2.58E-04 | 1.80E-03 | 0.37 | 5.38E-02 | 0.22 | 0.89 (477) |
| rs3924047 | 11 | 70,507,506 | *SHANK2* | 8.77E-06 | 2.19E-04 | 1.04E-02 | 0.44 | 4.95E-01 | 0.21 | 0.87 (489) |
| rs4356270 | 12 | 90,843,346 | | 8.42E-06 | 8.12E-06 | 1.91E-01 | 0.35 | | | (370) |
| rs12427267 | 12 | 90,848,103 | | 6.00E-07 | 6.04E-05 | 1.90E-03 | 0.38 | 2.15E-01 | 0.19 | 0.95 (403) |
| rs11106345 | 12 | 90,850,631 | | 7.22E-06 | 8.83E-06 | 1.81E-01 | 0.35 | | | (511) |
| rs10848190 | 12 | 129,767,988 | | 5.43E-06 | 4.83E-04 | 2.36E-03 | 0.39 | | | (327) |
| rs1035717 | 14 | 69,648,452 | *SLC8A3* | 6.36E-06 | 1.32E-03 | 2.19E-04 | 0.41 | 7.12E-01 | 0.23 | 0.95 (334) |
| rs4903712 | 14 | 77,685,346 | | 1.67E-06 | 3.40E-05 | 1.79E-02 | 0.27 | 1.99E-03 | 0.29 | 0.92 (768) |
| rs17754467 | 14 | 77,692,276 | | 5.54E-07 | 3.20E-06 | 4.50E-02 | 0.23 | 7.36E-02 | 0.20 | 0.87 (621) |
| rs1568447 | 17 | 70,348,607 | | 4.08E-06 | 5.26E-04 | 1.13E-03 | 0.38 | | | (623) |
| rs9901283 | 17 | 70,349,427 | *GRIN2C* | 9.31E-06 | 8.18E-04 | 1.67E-03 | 0.38 | | | (397) |
| rs11652088 | 17 | 70,351,427 | *GRIN2C* | 3.50E-06 | 4.58E-04 | 1.00E-03 | 0.38 | 7.25E-01 | 0.16 | 0.82 (537) |
| rs8111589 | 19 | 50,726,398 | *OPA3* | 8.27E-06 | 7.50E-05 | 4.03E-02 | 0.44 | 6.08E-01 | 0.21 | 0.90 (443) |
| rs2830045 | 21 | 26,380,280 | *APP* | 6.31E-06 | 1.77E-03 | 2.22E-04 | 0.37 | | | (396) |
| | | | **VARIANCE** | 9.09E-12 | 8.60E-07 | 2.01E-03 | | | | |
| | | | **MEAN** | 5.04E-06 | 6.94E-04 | 1.71E-02 | | | | |

Markers without Unphased and IQS data were not genotyped through the experimental validation. Numbers in the brackets next to the IQS (>0.8) column are the numbers of individuals who meet such restriction on that specific marker.

doi:10.1371/journal.pone.0063526.t008

The Quantile-Quantile (Q-Q) plots are used in GWAS to assess the inflation of FPRs by comparing the distribution of observed *p*-values against the theoretical model distribution of expected *p*-values [85]. In theory, without type I error arising from population stratification or other artifacts, the Q-Q plot shall align with the diagonal line. This is true if we use completely randomized data. By comparing the distortion of the Q-Q plot of the test results from this diagonal line, we can tell whether there are false positives or other errors due to genotyping or imputation. Before we draw any conclusion, we provide the Q-Q plots for results from all three Case-Control, dTDT and MH tests on the COGA data (**Figure 7**). From the figure we can see that most of the observed *p*-values from Case-Control and dTDT are along the diagonal line. We do not observe significant distortion, i.e., type I error, in both tests. On the other hand, the Q-Q plot of MH test lies above the diagonal line. As stated earlier, both Case-Control and the dTDT are not robust to the population stratification because of the dependence of allele frequencies in the populations. MH test is basically a combined statistic of these two tests. Though we have increased the sample size in the combined test, we have reason to believe that such sensitivity to population stratification has been inflated in the MH test. This is a tradeoff that we need to pay attention to in the future studies. However, this issue can be partially addressed by restricting accurate genotypes based on the imputation quality score (IQS) [86] (discussed below in *Dealing with errors*).

Having the observed -$\log_{10}$ (*p-value*) along the diagonal line in the Q-Q plots doesn't mean these tests agree with each other. To investigate the concordance among these three tests, we rank the -$\log_{10}$ (*p-value*) from the dTDT (or MH test) and pick the top 100

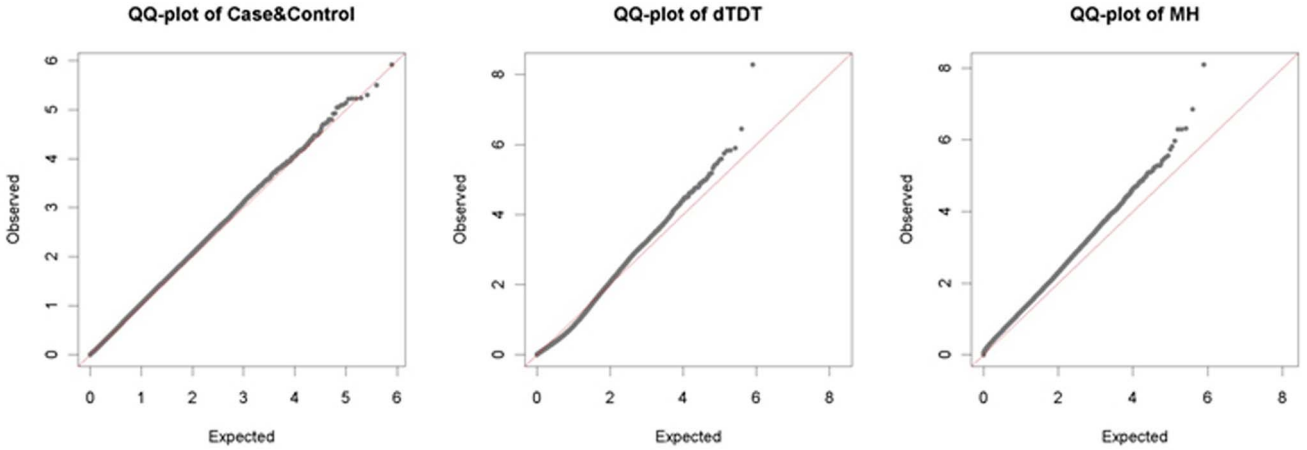


**Figure 7. QQ-plots of Case-Control, dTDT and MH tests.**
doi:10.1371/journal.pone.0063526.g007

signals. Then we plot these values with their corresponding -$\log_{10}$ (*p-value*) from the other two tests (**Figure 8**. To dTDT the other two are Case-Control and MH test; to MH test the other two are Case-Control and dTDT). From the figure, we can see that the concordance performs poorly among these tests. The top signals in either dTDT or MH test do not appear in the top list from the rest two tests as expected. There can be several reasons for this discordance: (1) genotyping errors in both linkage and association data; (2) inaccuracy of the imputation; (3) interference from population stratification. The first two issues may be addressed through experimental validation as discussed below. The last issue requires additional design for the tests that we will discuss in *Conclusions*.

## Dealing with Errors

After correcting for multiple tests ($p$-value = 0.05/19 = 0.0026), SNP rs4903712 on chromosome 14 remained significant. This was the seventh most significant SNP from the MH test. As discussed above, there are several issues affecting the dependability of our test results. As we saw from the Q-Q plots, signals from MH test have been inflated because of double counting of the population stratification factor. On the other hand, the genotyping and imputation accuracy may be taken into account as well. To address these issues, we compute the IQS for the listed top SNP markers on **Table 8** with and without setting a 0.8 threshold on the dosage probabilities. We report the number of individuals who meet such 0.8 threshold (in the last column, the numbers in the brackets next to IQS with dosage probabilities >0.8). According to the IQS, when we exclude the dosage probabilities that are below 0.8, the inference program performs very well and provides above ~0.90 IQS on average. The reason is that for dosage probabilities that are lower than 0.8, there is too much uncertainty for the program to impute, which not only heavily distorts the results (poor specificity) but also makes it difficult to filter out true

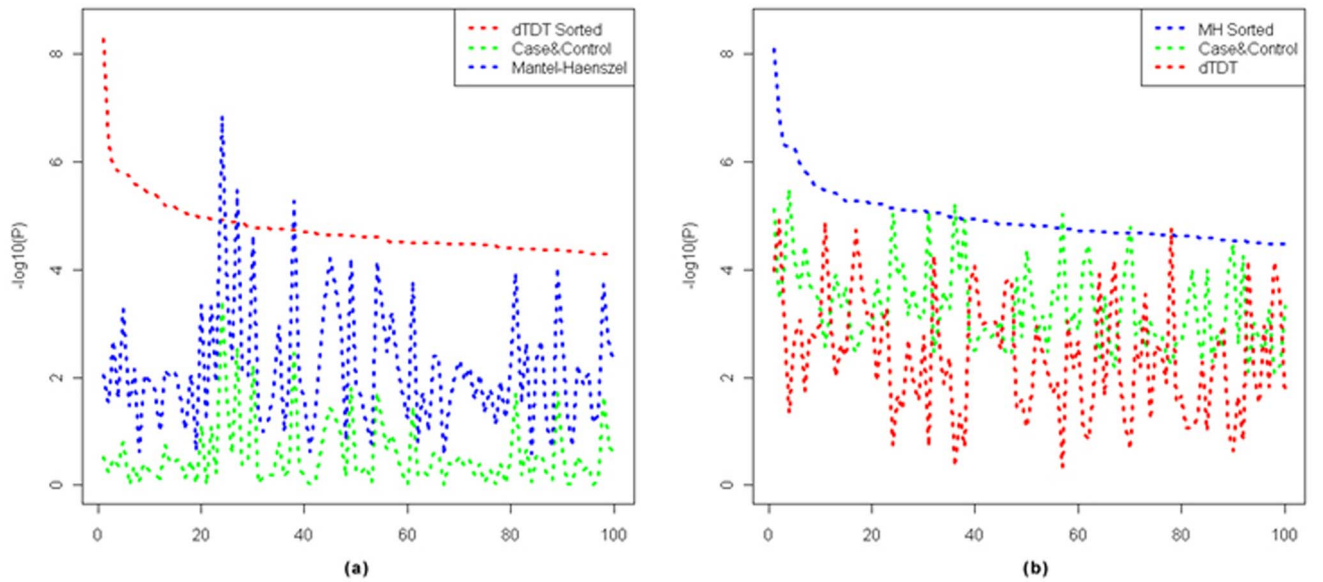


**Figure 8. Line chart of top 100 signals from Case-Control, dTDT and MH tests.** (a) line chart ranked by -$\log_{10}$ (*p-value*) from dTDT and its corresponding -$\log_{10}$ (*p-value*) from the other two tests: Case-Control & MH test; (b) line chart ranked by -$\log_{10}$ (*p-value*) from MH test and its corresponding -$\log_{10}$ (*p-value*) from the other two tests: Case-Control and dTDT.
doi:10.1371/journal.pone.0063526.g008

positives (low sensitivity). However, there is always a tradeoff when we enhance the accuracy. If we set a 0.8 threshold on dosage probabilities, the sample size dramatically reduces from 1,586 to 326 (intersection set across all markers). We further apply dTDT and MH test onto these 326 individuals with either the inferred or genotyped data but do not find significant markers at a $10^{-5}$ level due to a small sample size (data not shown).

As above, the experimental validation shows that the accuracy of the inference program can heavily impact the results of the dTDT and MH tests. The disagreement of results from these two tests on the real data could be attributed to several sources. First, the sample size of informative data is small. In total, we have 3,567 individuals from 453 EA families included for inference calculation. Within all these individuals, only 471 individuals were selected for SNP validation genotyping. This requires genotypes of more than 85% of individuals to be estimated. In addition, compared to the total number of SNPs, the number of microsatellite markers is also trivial (722 vs. 1 million).

According to **Table 8**, though the sample size may be reduced, we still recommend limiting dosage probabilities before genotyping. In our experience, a threshold at $0.7 \sim 0.8$ level can be a good cut-off. A threshold lower than this level may contribute too much noise and a threshold higher than this may reduce the sample size significantly. Meanwhile, we suggest interpreting the dTDT signals only after genotyping validation in order to lower the risk of false positives.

## Conclusions

Since the discovery of Mendel's law, genetic research has been challenged to identify genetic variants that contribute to human diseases. Along with the development of genome sequencing technologies, there have been impressive progresses within the research community over the past decade. Numerous methodologies have been developed and many disease-associated genes have been reported [87]. In this study, the work presented here embraces the recent development and addresses some of the research challenges in the field of genetic research. However, as we have seen, despite the promises of the solution we provide, it also prompts a great need to further investigate many of the issues we have presented.

As discussed, traditional case-control studies on GWA often include only unrelated individuals. By including family information in the study, we can expect an increase in power for linkage and association detection. On the other hand, because the traditional TDT requires complete genotypic information from a trio by measuring over-transmission of an allele from heterozygous parents to the affected offspring, it may be less useful in trio data like COGA where there are relatively few complete trios. To overcome these limitations, in this project, we extend the original TDT to the dTDT to accommodate dosage probabilities of a trio. The trio-dosage sets can be inferred through programs like MERLIN. Compared to a recent work from Edenberg et al, the dTDT shows increased power to detect association.

Genotype inference allows us to evaluate the evidence for association at the genetic markers that are not directly-genotyped. It helps improve the power of individual scans and is of particular usefulness for combining information from different studies such as linkage and GWAS. However, the accuracy of genotype inference may be impaired for the following reasons. First, because datasets where subjects are genotyped on different platforms may have different genotyping error rates, when we combine these datasets, inference can be problematic. Second, genotype inference for large datasets based on a small amount of shared information may encounter too much uncertainty in the procedure. For a similar reason, SNPs with low MAF may also have a higher chance of being inferred inaccurately.

On the other hand, a major advantage of the TDT is that it is not susceptible to population stratification. The dTDT is, however, sensitive in that it depends on the marker allele frequencies in the population. Because of this reason, when we combine results from both Case-Control and dTDT, the MH test potentially inflates errors due to population stratification. This can be noticed in the Q-Q plot as we present in **Figure 7**.

In summary, as we inspect the reasons for having low concordance among the tests as well as poor replication from our test results to the experiment validation, we have the following conclusions:

I. Because the linkage and GWAS data were genotyped on different platforms, they may carry different genotyping errors, which make it difficult to obtain genotype inference accurately; inference across these platforms can generate spurious associations;

II. Due to a great sparsity in the combined dataset, a large number of the markers have to be inferred without sufficient support from the common markers, which introduces too much uncertainty in the inference;

III. Because of possible population stratification, combining both the Case-Control and the dTDT to enhance the sample size may introduce false positive signals.

As one solution, when we filter out poorly-inferred SNP markers using IQS, we are able to removes thousands of false positives that can be particularly useful for SNPs with low MAF and when datasets are genotyped on different platforms. However, the tradeoff is we have to exclude a good number of individuals from our database in order to meet such restriction. But this can always be an option when enough samples are available.

Despite this area for methodological development, our work posits that the dTDT has considerable utility for linkage and association testing. By exploiting family data with inference and existing case-control information, the dTDT demonstrated here has opened a new window to possible routes for the integration of both population-based and family-based studies.

### Future Direction

To address the sensitivity issue due to population stratification, it is necessary to validate the SNP genotyping and perform a test such as the PTDT to validate results and use a program such as UNPHASED. This approach minimizes expense when the case/control sample is derived from an existing family study in which relatives have available DNA for typing. Moreover, we may implement additional application to other populations such as African Americans to compare findings with what we have from the European Americans. This will require extending the techniques described above.

We may explore using more of the family data instead of only using SNPs. Other sources of information could be captured, such as the copy number variants (CNVs) [88,89]. It is also suggested that non-genetic risk factors tend to raise the attention for complex traits and should also be incorporated into the genetic studies. Meanwhile, it is likely that COGA will obtain GWAS data in the relatives so that we can evaluate the efficiency of inference versus having GWAS genotypes available. Power calculation and sample size estimation of the new statistics are needed for general use. Due to the uncertainty inherent to the

inference procedure, we plan to develop better strategies for generating dosage probabilities.

Finally, the dTDT should not be limited to dosage probabilities from the inference programs only. As a perspective, the next-generation sequencing data will provide a challenge using the method developed in this paper. Similar methods can be used when pedigree members have a SNP chip, and a subset has sequence data. Despite the discordance and poor replication from our test results, we believe that linkage can help identify regions of interest in conjunction with association testing. Computational inference has helped us reduce the experimental cost in that we may only need to do sequencing on a limited number of family members. Keeping this in mind, we need to implement appropriate selection of the most informative families when we do genotyping. All these future directions shall be promising and have potential to inform the field.

## Supporting Information

**Table S1 Dominant (K = 0.01).**
(XLSX)

**Table S2 Dominant (K = 0.1).**
(XLSX)

**Table S3 Dominant (K = 0.2).**
(XLSX)

**Table S4 Recessive (K = 0.01).**
(XLSX)

**Table S5 Recessive (K = 0.1).**
(XLSX)

**Table S6 Recessive (K = 0.2).**
(XLSX)

**Table S7 Codominant (K = 0.01).**
(XLSX)

**Table S8 Codominant (K = 0.1).**
(XLSX)

**Table S9 Codominant (K = 0.2).**
(XLSX)

## Acknowledgments

The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, H. Edenberg, L. Bierut, includes ten different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, J. Nurnberger Jr., T. Foroud); University of Iowa (S. Kuperman, J. Kramer); SUNY Downstate (B. Porjesz); Washington University in St. Louis (L. Bierut, A. Goate, J. Rice, K. Bucholz); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield); Southwest Foundation (L. Almasy), Howard University (R. Taylor) and Virginia Commonwealth University (D. Dick). A. Parsian and M. Reilly are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting-Kai Li, currently a consultant with COGA, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their critical contributions.

## Author Contributions

Conceived and designed the experiments: JPR AMG LJB MAS JAT HJE. Performed the experiments: JCW. Analyzed the data: ZZ WBH PL. Contributed reagents/materials/analysis tools: PL JAT. Wrote the paper: JPR ZZ AA HJE.

## References

1. Ott J, Schrott HG, Goldstein JL, Hazzard WR, Allen FH Jr, et al. (1974) Linkage studies in a large kindred with familial hypercholesterolemia. Am J Hum Genet 26: 598–603.
2. Elston RC, Namboodiri KK, Go RC, Siervogel RM, Glueck CJ (1976) Probable linkage between essential familial hypercholesterolemia and third complement component (C3). Cytogenet Cell Genet 16: 294–297.
3. Berg K, Heiberg A (1978) Linkage between familial hypercholesterolemia with xanthomatosis and the C3 polymorphism confirmed. Cytogenet Cell Genet 22: 621–623.
4. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234–238.
5. Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, et al. (1985) Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science 230: 1054–1057.
6. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277–318.
7. Olson JM (1999) A general conditional-logistic model for affected-relative-pair linkage studies. Am J Hum Genet 65: 1760–1769.
8. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
9. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516–1517.
10. Strittmatter WJ, Saunders AM, Schmechel D,Pericak-Vance M, Enghild J, et al. (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late onset Alzheimer disease. Proc Natl Acad Sci USA 90: 1977–1981.
11. van der Put N, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, et al. (1995) Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. Lancet 346: 1070–1071.
12. Devlin B, Roeder K (1999) Genomic Control for Association Studies. Biometrics 55: 997–1004.
13. Knowler WC, Williams RC, Pettitt D J, Steinberg AG (1988) Gm3;5,13,14 and type-2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43: 520–526.
14. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65: 220–228.
15. Devlin B, Roeder K, Bacanu S (2001) Unbiased methods for population-based association studies. Genet Epidemiol 21: 273–284.
16. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037–2048.
17. Ronnie S, Rogus JJ (2010) The power of the Transmission Disequilibrium Test in the presence of population stratification. Eur J Hum Genet 18: 1032–1038.
18. Spielman RS, McGinnis RE, Ewens WJ. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–516.
19. Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, et al. (1981) Genetics of HLA disease associations: The use of the haplotype relative risk (HRR) and the 'haplo-delta' (Dh) estimates in juvenile diabetes from three racial groups. Hum Immunol 3: 384.
20. Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61: 319–333.
21. Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56: 811–812.
22. Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62: 450–458.
23. Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63: 1886–1897.
24. Allison DB (1997) Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet 60: 676–690.
25. Xiong MM, Krushkal J, Boerwinkle E (1998) TDT statistics for mapping quantitative trait loci. Ann Hum Genet 62: 431–452.
26. George V, Tiwari H.K, Zhu X, Elston RC (1999) A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. Am J Hum Genet 65: 236–245.
27. Zhu X, Elston RC (2001) Transmission/disequilibrium test for quantitative traits. Genetic Epidemiology 20: 57–74.
28. Yang Q, Rabinowitz D, Isasi C, Shea S (2000) Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits. Human Heredity 50: 227–233.
29. Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Hum Hered 47: 342–350.
30. Monks SA, Kaplan NL (2000) Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. Am J Hum Genet 66: 576–592.

31. Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, et al. (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. Eur J Hum Genet 12: 752–761.
32. Rabinowitz D, Laird N (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50: 211–223.
33. Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, et al. (2004) A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Stat Appl Genet Mol Biol 3: Article17.
34. Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7: 385–394.
35. Laird NM, Lange C (2008) Family-based methods for linkage and association analysis. Adv Genet 60: 219–252.
36. Won S, Wilk JB, Mathias RA, O'Donnell CJ, Silverman EK, et al. (2009) On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. PLoS Genet 5: e1000741.
37. Thomson G (1995) Mapping disease genes: family-based association studies. Am. J. Hum Genet 57: 487–498.
38. Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59: 983–989.
39. Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 67: 146–154.
40. Martin ER, Bass MP, Hauser ER, Kaplan NL (2003) Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet 73: 1016–1026.
41. McGinnis RE (1998) Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT). Ann Hum Genet 62: 159–179.
42. Weinberg CR. (1999) Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 64: 1186–1193.
43. Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65: 1170–1177.
44. Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. Am J Hum Genet 64: 861–870.
45. Sun F, Flanders WD, Yang Q, Khoury MJ (1999) Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. Am J Epidemiol 150: 97–104.
46. Schaid D, Rowland CM (1999) Quantitative trait transmission disequilibrium test: allowance for missing parents. Genet Epidemiol 17: S307–312.
47. Lee WC (2002) Transmission/disequilibrium test when neither parent is available in some families: a non-iterative approach. J Cancer Epidemiol Prev 7: 97–103.
48. Guo CY, DeStefano AL, Lunetta KL, Dupuis J, Cupples LA (2005) Expectation maximization algorithm based haplotype relative risk (EM-HRR): test of linkage disequilibrium using incomplete case-parents trios. Hum Hered 59: 125–135.
49. Guo CY, Cupples LA, Yang Q (2008) Testing informative missingness in genetic studies using case-parent triads. Eur J Hum Genet 16: 992–1001.
50. Guo C (2007) The impact of complex informative missingness on the validity of the transmission/disequilibrium test (TDT). BMC Proc 1 Suppl 1, S26.
51. Little RJA, Rubin DB (2002) Statistical analysis with missing data. Second Edition, New York: John Wiley & Sons.
52. Allen AS, Rathouz PJ, Satten GA (2003) Informative missingness in genetic association studies: case-parent designs. Am J Hum Genet 72: 671–680.
53. Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6: 127–130.
54. Rice JP, Saccone NL, Foroud T, Edenberg HJ, Nurnberger JI Jr, et al. (2003) Alcoholism: the USA Collaborative Study on the Genetics of Alcoholism (COGA). In: Encyclopedia of the Human Genome. Macmillan Publishers Ltd, Nature Publishing Group. 1–7.
55. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101.
56. George VT, Elston RC (1987) Testing of association between polymorphic markers and quantitative traits in pedigrees. Genet Epidemiol 4: 193–201.
57. Keavney B, McKenzie CA, Connell JM, Julier C, Ratcliffe PJ, et al. (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. Hum Mol Genet 7: 1745–1751.
58. Burdick JT, Chen WM, Abecasis GR, Cheung VG (2006) In silico method for inferring genotypes in pedigrees. Nat Genet 38: 1002–1004.
59. Chen WM, Abecasis GR (2007) Family based association tests for genome wide association scans. Am J Hum Genet 81: 913–926.
60. Visscher PM, Duffy DL (2006) The value of relatives with phenotypes but missing genotypes in association studies for quantitative traits. Genet Epidemiol 20: 30–36.
61. Abecasis GR, Wigginton JE (2005) Handling Marker-Marker Linkage Disequilibrium: Pedigree Analysis with Clustered Markers. Am J Hum Genet 77: 754–767.
62. Lange K, Sinsheimer JS, Sobel E (2005) Association testing with Mendel. Genet Epidemiol 29: 36–50.
63. Lange K, Weeks D, Boehnke M (1988) Programs for Pedigree Analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5: 471–472.
64. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84: 2363–2367.
65. Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21: 523–542.
66. Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61: 748–760.
67. Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58: 1323–1337.
68. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nature Genet 25: 12–13.
69. Markianos K, Daly MJ, Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet 68: 963–977.
70. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101. Available: http://www.sph.umich.edu/csg/abecasis/Merlin/. Accessed 2009 May 15.
71. Feighner JP, Robins E, Guze SB, Woodruff RA Jr, Winokur G, et al. 1972) Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry 1: 57–63.
72. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders DSM-III-R, Third Edition, Revised. (1987) Washington, DC, American Psychiatric Association.
73. Kanny D, Liu Y, Brewer RD, Centers for Disease Control and Prevention (CDC) (2011) Binge drinking - United States, 2009. MMWR Surveill Summ 60 Suppl: 101–104.
74. Testino G (2008) Alcoholic diseases in hepato-gastroenterology: a point of view. Hepatogastroenterology 55: 371–377.
75. Caan W, De Belleroche J (2002) Drink, Drugs and Dependence: From Science to Clinical Practice. New York: Routledge.
76. Hasin DS, Stinson FS, Ogburn E, Grant BF (2007) Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Arch Gen Psychiatry 64: 830–842.
77. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81: 559–575.
78. Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22: 719–748.
79. Edenberg HJ, Koller DL, Xuei X, Wetherill L, McClintock JN, et al. (2010) Genome-wide association study of alcohol dependence implicates a region on chromosome 11. Alcohol Clin Exp Res 34: 840–852.
80. Li L, Sun L, Gao F, Jiang J, Yang Y, et al. (2010) Stk40 links the pluripotency factor Oct4 to the Erk/MAPK pathway and controls extraembryonic endoderm differentiation. Proc Natl Acad Sci USA 107: 1402–1407.
81. Sequenom website. Available: http://www.sequenom.com. Accessed 2011 Aug 7.
82. Kbioscience website. Available: http://www.kbioscience.co.uk. Accessed 2011 Aug 8.
83. AppliedBiosystems website. Available: http://www.appliedbiosystems.com. Accessed 2011 Aug 7.
84. Dudbridge F (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered 66: 87–98.
85. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, et al. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nature genetics 37: 1243–1246.
86. Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, et al. (2010) A new statistic to evaluate imputation reliability. PLoS One 5: e9697.
87. Johnson AD, O'Donnell CJ (2009) An open access database of genome-wide association results. BMC Med Genet 10: 6.
88. Ionita-Laza I, Perry GH, Raby BA, Klanderman B, Lee C, et al. (2008) On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. Genet Epidemiol 32: 273–284.
89. Murphy A, Won S, Rogers A, Chu JH, Raby BA, et al. (2010) On the genome-wide analysis of copy number variants in family-based designs: methods for combining family-based and population-based information for testing dichotomous or quantitative traits, or completely ascertained samples. Genet Epidemiol 34: 582–590.