Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 737-742
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Generating Simulated Nuclear Families for TDT Type Methods

Caixia Li[1], Peixing Li[1,2*]

[1]School of Mathematics, Sun Yat-sen University, Guangzhou, China
[2]Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou, China
Email: *lnslpx@mail.sysu.edu.cn

How to cite this paper: Li, C.X. and Li, P.X. (2024) Generating Simulated Nuclear Families for TDT Type Methods. *Open Journal of Statistics*, **14**, 737-742.
https://doi.org/10.4236/ojs.2024.146033

Received: November 23, 2024
Accepted: December 6, 2024
Published: December 9, 2024

Abstract

Transmission disequilibrium tests (TDT) is a well-known case-parents family-based method to detect the association between genetic polymorphisms and a disease phenotype. Various extensions of the TDT have been developed and widely applied in medical research. In this article, we introduced a simple simulation algorithm based on a transition model to generate general nuclear families rather than trios to simulate multiple tightly linked markers. The simulations show that the empirical distributions of the test statistics coincide with the expected distribution under the null hypothesis.

Keywords

Genetic Association Study, Transmission Disequilibrium Test, Simulation, Haplotype, Linkage Disequilibrium

1. Introduction

Genetic association studies aim to detect the association between one or more genetic polymorphisms and a trait. Once an allele of a gene is over-presented in a case population relative to the control, it may be established that such an allele of the gene is associated with the studied disease. In a population-based case-control design, the association can be demonstrated, if it exists, by comparing allele frequencies at the marker locus in random samples of unrelated patients and controls. However, the association between a disease and a genetic marker can arise from confounding by underlying stratification within the population [1] [2]. Population stratification can occur in case-control or other population-based designs. It's important to make comparisons between cases and controls, as far as possible, within homogeneous subpopulations. Family-based designs have been proposed to counteract confounding due to population stratification.

The best-known of the family-based design is the case-parent triad design. The case-parent triad design and transmission disequilibrium test (TDT) proposed by Spielman (1993) suggested to collect case-parent trios [3]. Alleles or haplotypes transmitted to affected offspring are compared with untransmitted alleles, providing a control sample that is inherently matched to the case sample with regard to population structure. Various extensions of the TDT and computer programs have been developed to deal with a variety of different pedigree structures with nuclear families [4]-[8]. The methods have been widely applied in medical research, even in Genome-wide association studies (GWAS) [9]-[11]. However, current simulations for the statistical power of TDT-type methods are limited to some special cases. For example, the special cases in which there is no recombination between the marker and disease susceptibility locus or there are only case-parents trios. For population-based case-control design, Su *et al.* (2011) used a resampling approach to simulate a single or multiple nearby disease SNPs on the same chromosome [12]. In this paper, we introduced a simple simulation algorithm to generate general nuclear families rather than trios to simulate multiple tightly linked markers.

2. Method

We consider a triad to contain both parents and an affected child first. Suppose a disease locus with mutant disease allele D and normal allele d, with allele frequencies, $q_D = q$, $q_d = 1-q$. For multiple tightly linked markers, we consider the haplotypes, rather than alleles for each marker. For example, for two SNP (single nucleotide polymorphism) markers with alleles 1 and 2, there will be four haplotypes 11, 12, 21, 22. The haplotype frequencies for the markers are denoted by $p_1, p_2, \cdots, p_{m-1}, p_m = 1-\sum_{i=1}^{m-1} p_i$ for m haplotypes. Linkage disequilibrium (LD) coefficients between disease locus and markers defined in Sham (1995) [4]

$$e_{si} = \frac{h_{si}}{p_i q_s}, \quad s = D,d;\ i = 1,2,\cdots,m,$$

where h_{si} is the frequency of disease-marker haplotype *si*. All $e_{si} = 1$ implies linkage equilibrium. The LD coefficients satisfy $e_{si} \geq 0$ and

$$\sum_{i=1}^{m} p_i e_{si} = \sum_{i=1}^{m} \frac{h_{si}}{q_s} = \sum_{i=1}^{m} P(i \mid s) = 1, \quad \sum_{s=D,d} q_s e_{si} = \sum_{s=D,d} \frac{h_{si}}{p_i} = \sum_{s=D,d} P(s \mid i) = 1.$$

The penetrance $P(\text{Affected} \mid \text{Genotype})$ is denoted by f_{DD}, f_{Dd}, f_{dd} $(1 \geq f_{DD} \geq f_{Dd} \geq f_{dd} \geq 0)$ for genotype *DD*, *Dd*, *dd*. And then the prevalence $P_A = q^2 f_{DD} + 2q(1-q) f_{Dd} + (1-q)^2 f_{dd}$.

If there is no recombination between the disease locus and markers, *i.e.* recombination rate $\theta = 0$, there are 4 disease-marker haplotypes in one case-parents trio, denoted by si, tj, uk, vl, $(s,t,u,v \in \{D,d\}, i,j,k,l \in \{1,2,\cdots,m\})$, where (si, tj), (uk, vl) and (si, uk) are haplotype pairs for father, mother and child, denoted by g_f, g_m, g_c respectively. According to Morris *et al.* (1997) [5], under Hardy-

Weinberg equilibrium, the probability of the corresponding haplotype pairs for the trio

$$
\begin{aligned}
&P\left(g_c = si/uk, g_f = si/tj, g_m = uk/vl \mid \text{Affected}\right)\\
&= \frac{1}{P_A} P\left(\text{Affected} \mid g_c = si/uk\right) P\left(g_c = si/uk, g_f = si/tj, g_m = uk/vl\right)\\
&= \left(f_{su}/P_A\right) h_{si} h_{tj} h_{uk} h_{vl}\\
&= \left(f_{su}/P_A\right)\left(p_i q_s e_{si}\right)\left(p_k q_u e_{uk}\right)\left(p_j q_t e_{tj}\right)\left(p_l q_v e_{vl}\right)\\
&= \left(f_{su}/P_A\right)\left(q_s q_u q_t q_v\right)\left(p_i e_{si}\right)\left(p_j e_{tj}\right)\left(p_k e_{uk}\right)\left(p_l e_{vl}\right)
\end{aligned}
$$

Note that

$$\sum_{s,u=D,d} f_{su}/P_A = 1, \quad \sum_{s=D,d} q_s = \sum_{s,u=D,d} q_s q_u = 1, \text{ and } \sum_{i=1}^{m} p_i e_{si} = 1.$$

The parents have disease allele *s* with probability q_s. And the parent has disease-marker haplotype *si* with probability $p_i e_{si}$, after having disease allele *s*. The term f_{su}/P_A is the probability that the parents transmitted disease alleles s,u (*s* from father, *u* from mother) to the affected child when they have alleles s,u. Therefore, when the parents have genotype *a*/*b* and *c*/*d* $\left(a,b,c,d \in \{D,d\}\right)$, the probabilities that they transmitted alleles (*a*, *c*), (*a*, *d*), (*b*, *c*), or (*b*, *d*) to the affected child are proportional to f_{ac}, f_{ad}, f_{bc}, f_{bd}.

If there is no recombination, consider the unaffected-parents trio. Similarly, we can get

$$
\begin{aligned}
&P\left(g_c = si/uk, g_f = si/tj, g_m = uk/vl \mid \text{Unffected}\right)\\
&= \left(\left(1 - f_{su}\right)/\left(1 - P_A\right)\right)\left(q_s q_u q_t q_v\right)\left(p_i e_{si}\right)\left(p_j e_{tj}\right)\left(p_k e_{uk}\right)\left(p_l e_{vl}\right)
\end{aligned}
$$

The term $\left(1 - f_{su}\right)/\left(1 - P_A\right)$ is the probability that the parents transmit disease alleles s,u (*s* from father, *u* from mother) to the unaffected child when they have alleles s,u. Therefore, when the parents have genotype *a*/*b* and *c*/*d* $\left(a,b,c,d \in \{D,d\}\right)$, the probabilities that they transmitted alleles (*a*, *c*), (*a*, *d*), (*b*, *c*), or (*b*, *d*) to an unaffected child are proportional to $1 - f_{ac}$, $1 - f_{ad}$, $1 - f_{bc}$, $1 - f_{bd}$.

However, if recombination is considered, there are 4 situations for the affected or unaffected child. Let (si,tj) and (uk,vl) be haplotype pairs for father and mother, where the disease alleles *s* and *u* are transmitted to the affected child. Let $\theta\left(0 \le \theta \le 0.5\right)$ be the recombination rate, it's easy to see that $P\left(g_c = si/uk \mid g_f = si/tj, g_m = uk/vl\right) = \left(1-\theta\right)^2$. Then, the disease-marker haplotype pair for the child will be (si,uk), (sj,uk), (si,ul), or (sj,uk) with probability $\left(1-\theta\right)^2$, $\left(1-\theta\right)\theta$, $\left(1-\theta\right)\theta$, or θ^2, respectively.

3. Simulating Nuclear Family Data

After specifying the parameters, including the number of nuclear families *N*, disease allele frequencies $q, 1-q$, penetrance f_{DD}, f_{Dd}, f_{dd}, marker allele frequencies p_i $(i = 1,2,\cdots,m, \sum p_i = 1)$, recombination rate θ between disease locus and markers, LD between disease locus and markers,

$$e_{D1},\cdots,e_{D(m-1)},\quad e_{Dm}=\frac{1}{p_m}\left(1-\sum_{i=1}^{m-1}p_i e_{Di}\right),$$

$$e_{d1}=\frac{1-qe_{D1}}{1-q},e_{d2}=\frac{1-qe_{D2}}{1-q},\cdots,e_{dm}=\frac{1-qe_{Dm}}{1-q}.$$

We simulate the nuclear families using the following procedure:

Step 1. Generate the paternal and maternal genotypes *a*/*b* and *c*/*d* $\left(a,b,c,d\in\{D,d\}\right)$ with probability $q_a q_b$ and $q_c q_d$.

Step 2. Transmit disease alleles (*a*, *c*), (*a*, *d*), (*b*, *c*), or (*b*, *d*) to the affected child with probabilities proportional to f_{ac}, f_{ad}, f_{bc}, f_{bd}, and to unaffected child with probabilities proportional to $1-f_{ac}$, $1-f_{ad}$, $1-f_{bc}$, $1-f_{bd}$. The alleles transmitted to affected or unaffected children are denoted by (*s*, *u*). The alleles untransmitted to affected or unaffected children are denoted by (*t*, *v*).

Step 3. Combing with the generated disease alleles *s*, *t*, *u* and *v* from step 2, generate marker haplotypes *i*, *j*, *k* and *l* ($i,j,k,l\in\{1,2,\cdots,m\}$ with probability $p_i e_{si}$, $p_j e_{tj}$, $p_k e_{uk}$ and $p_l e_{vl}$ respectively. Assign marker haplotype pairs (*i*, *j*) for the father, (*k*, *l*) for the mother.

Step 4. Generate haplotype pair (*si*, *uk*), (*sj*, *uk*), (*si*, *ul*), or (*sj*, *ul*) for each child with probability $(1-\theta)^2$, $\theta(1-\theta)$, $\theta(1-\theta)$, θ^2, respectively. Assign the corresponding marker haplotype pair (*i*, *k*), (*j*, *k*), (*i*, *l*), or (*j*, *l*) to the affected or unaffected child.

Step 5. Output: the genotypes of the markers for the family members, based on the haplotype pairs from step 3 and step 4.

In step 5, we need the haplotype information to construct the genotypes of the markers. For example, if there are three biallelic SNP markers with allele 1 and 2, then there might be 8 haplotype 111, 112, 121, 122, 211, 212, 221, 222, numbered with 1, 2, …, 8. Therefore, when the haplotype pairs 2/5 generated from step 3 which represents 112/211, that means the genotypes are 1/2, 1/1, 2/1 for the three markers respectively.

In addition, we can use the simulator to generate a population with markers satisfying Hardy-Weinberg equilibrium. If we specify all which LD coefficients $e_{si}=1$ ($s=D,d$; $i=1,2,3,4$), and recombination rate $\theta=0$, then

$$P\left(g_f=si/tj,g_m=uk/vl\right)=h_{si}h_{tj}h_{uk}h_{vl}=\left(p_i q_s\right)\left(p_j q_t\right)\left(p_k q_u\right)\left(p_l q_v\right),$$

and hence $P\left(g_f=i/j,g_m=k/l\right)=\sum_{s,t,u,v=D,d}P\left(g_f=si/tj,g_m=uk/vl\right)=p_i p_j p_k p_l$,

i.e., the genotypes of parents at the markers follow HW equilibrium. In this case, the parents from the simulators can be regarded as a sample from a population with HW equilibrium at the markers.

4. Results

To evaluate our algorithm, two extended TDT methods, based on conditional likelihood and full likelihood, respectively, are selected to demonstrate how our simulator will preserve the test statistic distribution under the null hypothesis. Based

on conditional likelihood, Clayton (1999) [6] developed a score test with the program TRANSMIT. Based on full likelihood, we presented a likelihood ratio test to compare haplotype frequencies in transmitted and non-transmitted groups and derive relative risks [8]. It can be regarded as a generalized haplotype-relative-risk method, denoted by GHRR here. Both of them can deal with multiple tightly linked markers and uncertain haplotype phases for multi-locus genotypes.

In our simulations, one disease locus with alleles D and d, and two tightly linked SNP markers genotype data for 100 case-parent trios are generated. The four marker haplotypes are assumed to be equally frequent $p_i = 1/4$. For a general heredity mode, we specify $q = 0.2$, $f_{DD} = 0.02$, $f_{Dd} = 0.005$, $f_{dd} = 0.001$. Under the simulated condition, 500 replicated samples were generated to assess the distribution of the statistics under the null hypothesis in which $e_{si} = 1$ ($s = D, d$; $i = 1, 2, 3, 4$). For each replicated sample, TRANSMIT and GHRR were applied respectively to obtain the values of the test statistic.

When the null hypothesis is true, both of the test statistics are chi-square distributed with degree of freedom 3. The quantile-quantile (QQ) plots for test statistics from 500 replicated samples under the null hypothesis are shown in **Figure 1**. The plots show that the scatters of the observed quantiles and the expected quantiles of distribution $\chi^2(3)$ are tightly close to the line $y = x$. That means that the empirical distributions of the test statistics coincide with the expected distribution under the null hypothesis.

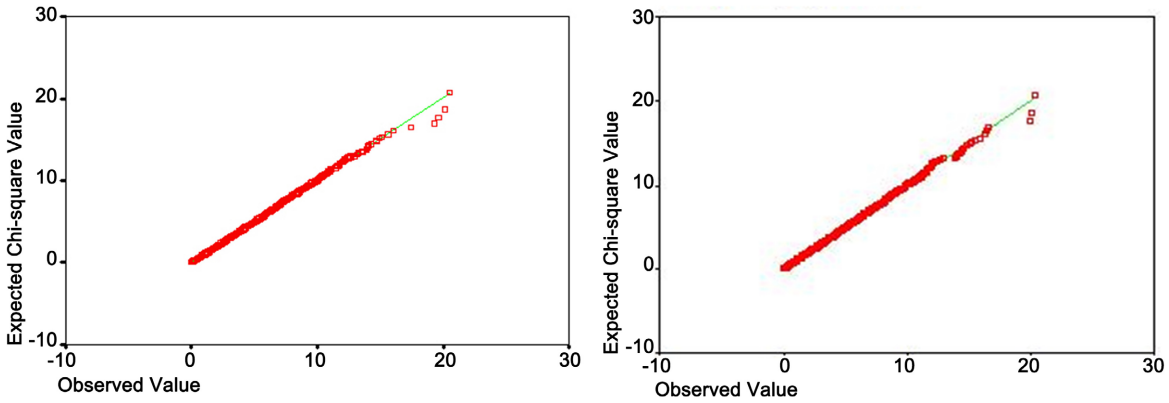

Figure 1. QQ plots for test statistics from TRANSMIT (left) and GHRR (right).

5. Conclusion

The TDT test for case-parents design measures the over-transmission of an allele in the affected from the heterozygous parents, thus avoiding problems of population stratification. TDT has been extended to various nuclear family designs. We introduced a simple simulation algorithm based on a transition model to generate general nuclear families with multiple tightly linked markers. To deal with multiple markers, we generate haplotypes first and output the genotypes for each marker at the last step. The simulations show that the empirical distributions of the test statistics coincide with the expected distribution under the null hypothesis.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 82373433), Guangdong Basic and Applied Basic Research Foundation (2020B1515310007), Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University (2020B1212060032).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ott, J., Kamatani, Y. and Lathrop, M. (2011) Family-Based Designs for Genome-Wide Association Studies. *Nature Reviews Genetics*, **12**, 465-474. https://doi.org/10.1038/nrg2989

[2] Nsengimana, J. and Bishop, D.T. (2011) Design Considerations for Genetic Linkage and Association Studies. In: Elston, R., Satagopan, J. and Sun, S., Eds., *Statistical Human Genetics*, Humana Press, 237-262. https://doi.org/10.1007/978-1-61779-555-8_13

[3] Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus. *American Journal of Human Genetics*, **52**, 506-516.

[4] Sham, P.C. and Curtis, D. (1995) An Extended Transmission/Disequilibrium Test (TDT) for Multi-Allele Marker Loci. *Annals of Human Genetics*, **59**, 323-336. https://doi.org/10.1111/j.1469-1809.1995.tb00751.x

[5] Morris, A.P., Whittaker, J.C. and Curnow, R.N. (1997) A Likelihood Ratio Test for Detecting Patterns of Disease-Marker Association. *Annals of Human Genetics*, **61**, 335-350. https://doi.org/10.1017/s0003480097006349

[6] Clayton, D. and Jones, H. (1999) Transmission/Disequilibrium Tests for Extended Marker Haplotypes. *The American Journal of Human Genetics*, **65**, 1161-1169. https://doi.org/10.1086/302566

[7] Lange, C. (2007) Family-Based Association Tests: FBAT Testing Strategies for Large-Scale Association Studies. *The Biomedical & Life Sciences Collection*, **2007**, e1001761. https://doi.org/10.69645/mumo8918

[8] Li, C. and Li, P. (2018) Haplotype Frequency Comparison for Case-Parents Data. *Open Journal of Statistics*, **8**, 721-730. https://doi.org/10.4236/ojs.2018.84047

[9] Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., *et al.* (2021) Genome-Wide Association Studies. *Nature Reviews Methods Primers*, **1**, Article No. 59. https://doi.org/10.1038/s43586-021-00056-9

[10] Bezamat, M., Carver, C.E. and Vieira, A.R. (2024) Family-Based GWAS for Dental Class I Malocclusion and Clefts. *BMC Oral Health*, **24**, Article No. 665. https://doi.org/10.1186/s12903-024-04444-x

[11] Veller, C. and Coop, G.M. (2024) Interpreting Population- and Family-Based Genome-Wide Association Studies in the Presence of Confounding. *PLOS Biology*, **22**, e3002511. https://doi.org/10.1371/journal.pbio.3002511

[12] Su, Z., Marchini, J. and Donnelly, P. (2011) HAPGEN2: Simulation of Multiple Disease SNPs. *Bioinformatics*, **27**, 2304-2305. https://doi.org/10.1093/bioinformatics/btr341