A powerful approach for association analysis incorporating imprinting effects

Fan Xia1, Ji-Yuan Zhou2 and Wing Kam Fung1,*

1Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong and 2Department of Biostatistics, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou 510515, China

ABSTRACT

Motivation: For a diabetetic marker locus, the transmission disequilibrium test (TDT) is a simple and powerful design for genetic studies. The TDT was originally proposed for use in families with both parents available (complete nuclear families) and has further been extended to 1-TDT for use in families with only one of the parents available (incomplete nuclear families). Currently, the increasing interest of the influence of parental imprinting on heritability indicates the importance of incorporating imprinting effects into the mapping of association variants.

Results: In this article, we extend the TDT-type statistics to incorporate imprinting effects and develop a series of new test statistics in a general two-stage framework for association studies. Our test statistics enjoy the nature of family-based designs that need no assumption of Hardy-Weinberg equilibrium. Also, the proposed methods accommodate complete and incomplete nuclear families with one or more affected children. In the simulation study, we verify the validity of the proposed test statistics under various scenarios, and compare the powers of the proposed statistics with some existing test statistics. It is shown that our methods greatly improve the power for detecting association in the presence of imprinting effects. We further demonstrate the advantage of our methods by the application of the proposed test statistics to a rheumatoid arthritis dataset.

Contact: wingfung@hku.hk

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on April 18, 2011; revised on July 5, 2011; accepted on July 22, 2011

1 INTRODUCTION

Genetic imprinting, as an example of epigenetic factors, occurs when certain genes increase levels of expression through inheritance from one of the parental chromosomes (Pfeifer, 2000; Reik and Walter, 2001). Morison et al. (2001) have constructed an imprinted gene database and >1% of all mammalian genes are believed to show imprinting effects (http://igc.otago.ac.nz). Abnormal expressions of imprinted genes result in numerous human genetic disorders, including Prader–Willi, Beckwith–Wiedemann and Angelman syndromes (Falls et al., 1999). When the candidate gene of a disease under study is a marker allele per se or is in linkage disequilibrium (LD) with a marker allele, the parental-asymmetric test (PAT) was proposed and extended to examine imprinting effects in the absence of maternal effects using complete nuclear families (nuclear families with both parents) with one or multiple affected children (Weinberg, 1999; Zhou et al., 2009). In this article, among the approaches, a test statistic termed TDTI (Hu et al., 2007) was constructed based on the extensions of the transmission/disequilibrium test (TDT) (Spielman et al., 1993) and was shown to be more powerful than the conventional TDT in the presence of imprinting effects. The TDTI first employs a parent-of-origin effects test statistic (POET) to test for imprinting effects and then chooses the most appropriate statistic from a set of TDT-type statistics, which are TDT suitable for no imprinting effects, TDTp (the paternal version of TDT) for maternal imprinting, and TDTm (the maternal version of TDT) for paternal imprinting, to test for association. Note that the TDTI only accommodates complete nuclear families and thus is not suitable for incomplete families.

On the other hand, genetic association analysis works over short distance in genome and provides fine mapping of genetic loci detected by linkage (Cordell and Clayton, 2005). In this article, we aim to develop a class of test statistics to test for association incorporating imprinting effects in the presence of linkage. First, we propose a test statistic, termed TDTI+, for use in complete nuclear families. Like TDTI, TDTI+ is also derived under a two-stage process, where we employ a more powerful test PAT instead of POET in the first stage to test if there is paternal, maternal or
no imprinting effects at a marker of interest, and we then test for association in the second stage by selecting an appropriate statistic according to the findings of the imprinting test. Next, we propose the 1-TDTp (the paternal version of 1-TDT) and 1-TDTm (the maternal version of 1-TDT) for incomplete nuclear families. By applying the 1-PAT for detecting imprinting effects in the first stage, we select an appropriate test statistic among 1-TDT, 1-TDTp and 1-TDTm to test for association in the second stage based on the findings of the 1-PAT. Thus, a test statistic 1-TDT* is developed for use in incomplete nuclear families. Further, combining complete nuclear families and incomplete nuclear families, the C-TDT* is suggested to efficiently utilize the information in a data set. TDT*, 1-TDT* and C-TDT* are expressed in general forms, which enable a flexible use of the statistics in practice to accommodate families with arbitrary numbers of children. We investigate the validity and the performance of the proposed set of statistics by simulation, and apply C-TDT* to the genome-wide association study of rheumatoid arthritis.

2 METHODS

2.1 Background and notations

Consider a marker locus (ML) with two alleles M1 and M2, and a disease susceptibility locus (DSL) under study with disease allele D and normal allele d. Let \( F = M \) and \( C_j \) denote the numbers of\( M_i \), carried by the father, mother and affected child \( j \) of a nuclear family, respectively. Let \( \phi_{ij}, \phi_{ij}^0, \phi_{ij}^1 \) and \( \phi_{ij}^D \) denote the penetrance of the ordered genotypes \( D/d, D/D, D/d \) and \( D/D \) at the DSL, respectively, of which the left allele of the slash in each genotype is assumed to be inherited from the father and the right one is from the mother. When Mendel’s law holds, \( \phi_{ij} = \phi_{ij}^0 \); otherwise, there may exist parent-of-origin effects at the DSL. The degree of imprinting effects is introduced as \( I = (\phi_{ij} - \phi_{ij}^0)/2 \), where \( I = 0 \) if \( \phi_{ij} = \phi_{ij}^0 \) signifies paternal (maternal) imprinting and \( I = 0 \) implies that risks for the two heterozygous DSL genotypes are identical, i.e., no imprinting effects (Strach et al., 2000).

As an earlier work (Hu et al., 2007a; Weinberg, 1999), mating symmetry in the parental generation is assumed throughout this article, that is, \( \Pr(F=m, M=m) = \Pr(F=m, M=f) \) for all \( f, m = 0, 1, 2 \). We further assume that there are no maternal effects and the target allele \( M_1 \) is in positive LD with the disease if there exists association.

2.2 Test statistic when both parents are available

We begin by constructing TDT*, where we first consider that there are \( n_C \) complete nuclear families and \( n_I \) affected children in family \( i \), \( i = 1, \ldots, n_C \). Let every child be matched with the parents, and the resulting trio is termed as case-parents trio. A set of difference statistics is then defined based on each case-parents trio, through which PAT, TDT, TDTm and TDTp are similarly expressed. For example, the difference statistics for PAT and TDT are respectively listed as below:

\[
\begin{align*}
\hat{z}_P &= \frac{1}{2} \left( t_{IP} - t_{IP}^0 \right), \quad t_{IP} = t_{IP}^0, \quad t_{IP}^0, \\
\hat{z}_T &= \frac{1}{2} \left( t_{IT} - t_{IT}^0 \right), \quad t_{IT} = t_{IT}^0, \quad t_{IT}^0, \\
\hat{z}_M &= \frac{1}{2} \left( t_{IM} - t_{IM}^0 \right), \quad t_{IM} = t_{IM}^0, \quad t_{IM}^0, \\
\hat{z}_D &= \frac{1}{2} \left( t_{ID} - t_{ID}^0 \right), \quad t_{ID} = t_{ID}^0, \quad t_{ID}^0,
\end{align*}
\]

where \( t_{IP} \) is the comparison statement, \( i \) is 1 when the comparison statement holds in each case-parents trio, otherwise it is 0, and \( t_{IP}^0 \) and \( t_{IP}^0 \) denote the numbers of \( M_1 \) being transmitted and not being transmitted from the heterozygous parents in each case-parents trio, respectively. Under the null hypothesis of the PAT test, which is no imprinting effects at the DSL or no association between the marker allele and the disease gene, \( s_{IP} \) has a zero mean. Under the null hypothesis of TDT test, which has no association between the marker allele and the disease allele or no linkage between the ML and DSL, \( s_{IT} \) has a zero mean. Further, by summing over all the case-parents trios, define \( s_T = \sum_{i=1}^{n_C} s_{IT}^i \) and \( s_I = \sum_{i=1}^{n_C} s_{IP}^i \), and the PAT and TDT can then be expressed as follows,

\[
\begin{align*}
\text{PAT} &= \frac{s_T}{\sqrt{\text{Var}_0(s_T)}}, \\
\text{TDT} &= \frac{s_I}{\sqrt{\text{Var}_0(s_I)}}, \\
\text{1-PAT} &= \frac{s_T-s_I}{\sqrt{\text{Var}_0(s_T)-\text{Var}_0(s_I)}},
\end{align*}
\]

where \( \text{Var}_0(s_T) \) and \( \text{Var}_0(s_I) \) are the corresponding unbiased estimates of the variances of \( s_T \) and \( s_I \) under the null hypotheses of the PAT and the TDT tests, respectively. \( \text{Var}_0(s_T) = \sum (\sum s_{IT}^2 + \sum s_{IP}^2) \) and \( \text{Var}_0(s_I) = \sum (\sum s_{IP}^2 + \sum s_{IT}^2) \), TDTp and TDTm could be similarly expressed from the corresponding difference statistics \( s_T = T_{IP} - T_{IP}^0 \) and \( s_I = T_{IT} - T_{IT}^0 \), where \( T_{IP} \) and \( T_{IP}^0 \) respectively denote the numbers of allele \( M_1 \) being transmitted and not being transmitted from the heterozygous father (mother) in family \( i \) to his (her) child. With the available statistics, the test statistic \text{TDT}^* is formulated as

\[
\text{TDT}^* = \text{TDT} \left[ \frac{\alpha_{TDT} - \alpha_{TDT}^*}{\alpha_{TDT}} \right] + \text{TDT} \left[ \frac{\alpha_{TDT}^* - \alpha_{TDT}^*}{\alpha_{TDT}} \right]
\]

where \( \alpha_{TDT} \) is the pre-specified significance level in testing for imprinting effects, and \( \alpha_{TDT}^* \) is the upper \( \alpha_{TDT}/2 \) quantile of a standard normal distribution. \text{TDT}^* illustrates that when the imprinting effect is significant enough, TDTm or TDTp is accordingly chosen to be the test statistic, otherwise TDT will be used for the following stage of association test. Note that the paper of TDTI (Hu et al., 2007a) has theoretically compared the powers of TDTm, TDTp and TDT and found TDTm (TDTp) is more powerful than TDT when there exists paternal (maternal) imprinting effect. When there is no association, PAT and the three TDT-type statistics asymptotically follow N(0, 1), but are not independent. The joint distribution of PAT and TDTm/TDTp/TDTm/TDTp is bivariate normal with mean vector being zero. Therefore, to find the asymptotic null distribution and P-value of the proposed test \text{TDT}^*, the covariance matrices between the two test statistics in the first and second stages need deriving. As an example, we have the estimate of the covariance between PAT and TDT as

\[
\text{Cov}(\text{PAT}, \text{TDT}) = \text{Cov}(\text{PAT}, \text{TDT})^* = \frac{\text{Var}_0(s_T)}{\sqrt{\text{Var}_0(s_T) \text{Var}_0(s_I)}}
\]

where \( \text{Cov}(\text{PAT}, \text{TDT})^* = \sum_1^n \sum_1^n s_{IP} s_{IT} \) and is the unbiased estimate of the covariance between \( s_T \) and \( s_I \) under the null hypothesis of no association (Supplementary Materials). The estimates of the covariances between PAT and TDTm and between PAT and TDTp can be similarly derived and are omitted for brevity in this article. Finally, let the estimated values of PAT in the first stage and the selected TDT-type statistic in the second stage for the \text{TDT}^* be \( t_1 \) and \( t_2 \), then

\[
\begin{align*}
\Pr(\text{TDT}^* \leq t_1) &= \Pr(\text{PAT} < -\alpha_{TDT}^*/2, \text{TDT} \leq t_2, \text{TDT} \leq t_2, \text{TDT} \leq t_2), \text{if } t_1 < -\alpha_{TDT}^*/2 \\
&= \Pr(\text{PAT} \leq -\alpha_{TDT}^*/2, \text{TDT} \leq t_2, \text{TDT} \leq t_2, \text{TDT} \leq t_2), \text{if } t_1 \geq -\alpha_{TDT}^*/2 \\
&= \Pr(\text{PAT} > \alpha_{TDT}^*/2, \text{TDT} \leq t_2, \text{TDT} \leq t_2, \text{TDT} \leq t_2), \text{if } t_1 > \alpha_{TDT}^*/2,
\end{align*}
\]

(see Supplementary Materials for the computation of P-value of \text{TDT}^*). It is worth noting that under the assumption of \( M_1 \) in positive LD with \( D \), a negative (positive) value of PAT indicates that the disease is paternally (maternally) imprinted. On the contrary, when the target allele \( M_1 \) is in negative LD with the disease allele, a negative (positive) value of PAT represents maternal (paternal) imprinting effect. However, the \text{TDT}^* remains unchanged under the two situations (Supplementary Materials), so our method is practical in real data analysis irrespective of the positive or negative LD between the target allele and the disease allele.

2.3 Test statistic when only one parent is available

We continue to construct 1-TDT*, where the notations involved with incomplete nuclear families are as follows. Suppose there are \( n_C \) incomplete
Table 1. Informative genotypes for case-parent pairs in 1-TDT

<table>
<thead>
<tr>
<th>Observed genotypes of the available parent and the affected child$^a$</th>
<th>Possible genotype of the missing parent</th>
<th>Transmitted allele from the missing parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>${M_1, M_2, M_4, M_5}$</td>
<td>$M_6, M_2, M_3, M_5$</td>
<td>$M_2$</td>
</tr>
<tr>
<td>${M_1, M_2, M_3, M_4}$</td>
<td>$M_6, M_3, M_2, M_5$</td>
<td>$M_3$</td>
</tr>
<tr>
<td>${M_1, M_2, M_3, M_4}$</td>
<td>$M_6, M_2, M_3, M_5$</td>
<td>$M_2$</td>
</tr>
<tr>
<td>${M_1, M_2, M_3, M_4}$</td>
<td>$M_6, M_3, M_2, M_5$</td>
<td>$M_3$</td>
</tr>
</tbody>
</table>

$^a$The left and right genotypes in each pair of curly brackets are of the available parent and the affected child, respectively.

nuclear families, among which we have $n_0$ single-mother families and $n_m$ single-father families. For each incomplete nuclear family, let every child be paired with the available parent and the resulting pair is termed as case-parent pairs. The test statistic 1-TDT, consisting of $T_1$ and $T_2$, was proposed by Sun et al. (1999) to test for association/linkage based on case-parent pairs. For association study, we find that $T_1$ is more powerful, so we employ $T_1$ to be the 1-TDT test in this article. To keep consistent and expect expectations to be zero (Supplementary Materials). Let $s = \sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij}$, we have

$$s = \sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij},$$

where $s_{ij}$ and $s_{ij}'$ are accordingly defined and similarly calculated as $s_{ij}=\sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij}$. The left and right genotypes in each pair of curly brackets are of the available parent and the resulting pair is termed as case-parent pairs.

On the other hand, motivated by the construction of TDT and TDTm, we further develop two test statistics 1-TDTp and 1-TDTm based on the stratification of 1-TDT into paternal and maternal versions and expect them to provide higher power than 1-TDT in the presence of imprinting effect. For the effort to construct these two statistics, we first illustrate the design of 1-TDTp and 1-TDTm, which is based on the four types of informative genotypes for case-parent pairs in Table 1.

From Table 1, the 1-TDT could be divided into two parts. The 1-TDT calculated merely with single-mother families essentially tests for association based on the transmission information of the missing father. As a result, it can be seen as a paternal version of 1-TDT and is defined as 1-TDTp. The contrary, the 1-TDTp merely calculated with single-father families essentially tests for association based on the transmission information of the missing father. We take it as the maternal version of 1-TDT (1-TDTm). 1-TDTp tests for association based on the transmission information of the missing case-parent pair in Table 1. In the effort to construct these two statistics, we first illustrate the design stratification of 1-TDT into paternal and maternal versions and expect them to provide higher power than 1-TDT in the presence of imprinting effect. For the effort to construct these two statistics, we first illustrate the design of 1-TDTm, which is based on the four types of informative genotypes for case-parent pairs in Table 1.

Table 1. Informative genotypes for case-parent pairs in 1-TDT

<table>
<thead>
<tr>
<th>Observed genotypes of the available parent and the affected child$^a$</th>
<th>Possible genotype of the missing parent</th>
<th>Transmitted allele from the missing parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>${M_1, M_2, M_4, M_5}$</td>
<td>$M_6, M_2, M_3, M_5$</td>
<td>$M_2$</td>
</tr>
<tr>
<td>${M_1, M_2, M_3, M_4}$</td>
<td>$M_6, M_3, M_2, M_5$</td>
<td>$M_3$</td>
</tr>
<tr>
<td>${M_1, M_2, M_3, M_4}$</td>
<td>$M_6, M_2, M_3, M_5$</td>
<td>$M_2$</td>
</tr>
<tr>
<td>${M_1, M_2, M_3, M_4}$</td>
<td>$M_6, M_3, M_2, M_5$</td>
<td>$M_3$</td>
</tr>
</tbody>
</table>

$^a$The left and right genotypes in each pair of curly brackets are of the available parent and the affected child, respectively.

nuclear families, among which we have $n_0$ single-mother families and $n_m$ single-father families. For each incomplete nuclear family, let every child be paired with the available parent and the resulting pair is termed as case-parent pairs. The test statistic 1-TDT, consisting of $T_1$ and $T_2$, was proposed by Sun et al. (1999) to test for association/linkage based on case-parent pairs. For association study, we find that $T_1$ is more powerful, so we employ $T_1$ to be the 1-TDT test in this article. To keep consistent and expect expectations to be zero (Supplementary Materials). Let $s = \sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij}$, we have

$$s = \sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij},$$

where $s_{ij}$ and $s_{ij}'$ are accordingly defined and similarly calculated as $s_{ij}=\sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij}$. The left and right genotypes in each pair of curly brackets are of the available parent and the resulting pair is termed as case-parent pairs.

On the other hand, motivated by the construction of TDT and TDTm, we further develop two test statistics 1-TDTp and 1-TDTm based on the stratification of 1-TDT into paternal and maternal versions and expect them to provide higher power than 1-TDT in the presence of imprinting effect. For the effort to construct these two statistics, we first illustrate the design of 1-TDTm, which is based on the four types of informative genotypes for case-parent pairs in Table 1.

From Table 1, the 1-TDT could be divided into two parts. The 1-TDT calculated merely with single-mother families essentially tests for association based on the transmission information of the missing father. As a result, it can be seen as a paternal version of 1-TDT and is defined as 1-TDTp. The contrary, the 1-TDTp merely calculated with single-father families essentially tests for association based on the transmission information of the missing father. We take it as the maternal version of 1-TDT (1-TDTm). 1-TDTp and 1-TDTm are formulated as follows,

$$1-TDTp = \frac{s'}{\sqrt{Vap(s')}} + \frac{s''}{\sqrt{Vap(s'')}}$$

where $s'=\sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij}'$ and $s''=\sum_{i=1}^{nF} \sum_{j=1}^{nM} s_{ij}''$. For instance,

$$C-TDT = \frac{\hat{C}_{TDT} + \hat{C}_{TDT}' + \hat{C}_{TDT}''}{\sqrt{\Var(C_{TDT}) + \Var(C_{TDT}') + \Var(C_{TDT}'')},}$$

sharing the same property of TDTI*, 1-TDTI* will not be affected under the situation where the target allele is in positive or negative LD with the disease gene. 1-PAT and the three 1-TDT-type statistics are not independent, and the covariances are derived in order to find the asymptotic null distribution of 1-TDTI*.

2.4 Test statistic combining complete and incomplete nuclear families

It is common in practice to collect complete nuclear families as well as incomplete nuclear families. As such, a combined statistic, which has not been discussed before, is of importance for enhancing testing power. We develop a test statistic C-TDTI* in an identical framework of TDTI* with expression as follows,

$$1-TDTI* = \frac{1-TDTI}{\sqrt{\Var(1-TDTI)} + \Var(1-TDTI')},$$

for situations with or without imprinting. Throughout the simulation, the additive genetic model is applied at the DSL with $\phi_{TDTI} = \phi_{TDTI}' = \phi_{TDTI}''$, and the penetrance of heterozygous genotypes $D$ and $D'$ are, respectively, obtained as $\phi_{TDTI} = 2\phi_{TDTI}' - 1$ and $\phi_{TDTI}' = 2\phi_{TDTI}' + 1$, where $\phi_{TDTI}'$ is further fixed at 0.1.

Family samples each consisting of 200 families each with one affected child and 100 families each with two affected children are utilized in the simulation study to investigate the performance of TDTI* with expression as follows,

$$1-TDTI* = \frac{1-TDTI}{\sqrt{\Var(1-TDTI)} + \Var(1-TDTI')}$$

where $\hat{C}_{TDTI} = \frac{\sum_{i=1}^{nF} \sum_{j=1}^{nM} C_{TDTI} + \hat{C}_{TDTI}''}{\sqrt{\Var(C_{TDTI}) + \Var(C_{TDTI}') + \Var(C_{TDTI}'')},}$. For instance,
of proposed methods. In each sample, define \( \tau \) to control the ratio of incomplete nuclear families and \( \beta \) as the probability of the father being missing for each incomplete nuclear family. The population stratification demographic model, consisting of two different homogeneous subpopulations, is considered throughout the simulation. The proportion of the family samples from the first (second) population is taken to be 0.4 (0.6) and the frequencies of disease allele \( D \) and maker allele \( M \) in the first (second) population are taken to be 0.1 (0.5) and 0.3 (0.8), respectively.

The recombination fraction between the DSL and ML is fixed to be 0.5 for each incomplete nuclear family. The LD coefficient between a marker allele and a disease gene is measured by Lewontin’s \( D' \) (Lewontin, 1988), which is taken to be 0.9 (0.63) in the first (second) population in the power study. All the simulation results are estimated from 10,000 replicates and are assessed at the significance level \( \alpha = 5\% \).

### 3.2 Sizes and powers of test statistics

Table 2 shows that the type I error rates of TDTI*, 1-TDTI*, C-TDTI* and C-TDTI are consistent with the nominal 5% level, regardless of imprinting degree \( I \), missing father probability \( \beta \) and incomplete nuclear family ratio \( \tau \). In addition, C-TDT is still a valid test for association study when there exist imprinting effects (TDT and 1-TDT also control the size well in the simulation, results are omitted for brevity).

For family samples with complete nuclear families, Figure 1 plots the empirical power of three statistics, TDTI*, TDTI and TDT, against five levels of imprinting effects. It can be seen that TDTI* achieves the highest power in the presence of imprinting effects, and this superiority becomes significant with the increase of \( |I| \).

When there is no imprinting effects, though TDTI* reaches its own maximum power occurs when \( \beta = 0.5 \), which is in accordance with the simulation results in the 1-TDTI paper (Hu et al., 2007b).

### 3.3 Comparison of the performance of the three statistics

Table 2 shows that the type I error rates of TDTI*, 1-TDTI*, C-TDTI* and C-TDTI are consistent with the nominal 5% level, regardless of imprinting degree \( I \), missing father probability \( \beta \) and incomplete nuclear family ratio \( \tau \). In addition, C-TDT is still a valid test for association study when there exist imprinting effects (TDT and 1-TDT also control the size well in the simulation, results are omitted for brevity).

For family samples only with incomplete nuclear families, Figure 2 compares the power of 1-TDTI*, 1-TDT and 1-TDTI against father missing probability \( \beta \). In the presence of complete paternal imprinting, it is observed that for each \( \beta \), the 1-TDTI* is more powerful than the other two methods. As for 1-TDTI, its own maximum power occurs when \( \beta = 0.5 \), which is in accordance with the simulation results in the 1-TDTI paper (Hu et al., 2007b).
Transmission-based tests incorporating imprinting effects

4 APPLICATION TO UK RHEUMATOID ARTHRITIS DATA

We apply the proposed methods to SNP-based association analysis of rheumatoid arthritis (RA). RA is a complex disease, and some epigenetic processes, e.g. genetic imprinting, are suspected to complicate the study of its genetic components (Gregersen, 1999). The dataset used in this article is from the United Kingdom (John et al., 2004), which is provided by Genetic Analysis Workshop 15 (Amos et al., 2007). In the dataset, there are 10156 SNP markers over the 22 autosomes, genotyped on 157 families. Because of a large proportion of missing genotype information of individuals for each SNP marker, the reduced dataset for analysis is a mixture of complete nuclear families and incomplete nuclear families, as well as a mixture of families with one affected child and families with multiple affected children. We henceforth employ both C-TDT and C-TDT*I for SNPs to make an efficient use of the available information.

The results of C-TDT and C-TDT*I based on 22 chromosomes are plotted in Figure 4. Both tests suggest that the strongest evidence for association is at SNP520297 on Chromosome 8 ($P<1 \times 10^{-5}$). Meanwhile, the comparison of the two subfigures roughly reveals the existence of difference between the results of the two test statistics. It inspires us to have a closer look at the results on each chromosome. In this article, we merely present the association results on two 1 cM intervals on Chromosome 2 (Fig. 5). In fact, an existing paper on joint analysis of linkage and imprinting for RA has revealed higher logarithmic odds (LOD) scores on Chromosome 2 under the model allowing for imprinting than under the no imprinting model (Zhou et al., 2007). We find that the association test allowing for imprinting effects (C-TDT*I) also presents stronger evidence of association than the conventional association test (C-TDT) for some SNPs on Chromosome 2. Specifically, for the SNPs detected by C-PAT to be imprinted on Chromosome 2, such as SNP56214, SNP67512, SNP521719 and SNP521720, C-TDT*I indeed provides notable
Fig. 4. Association analysis for RA over 22 chromosomes using C-TDT (A) and C-TDTI* (B) tests.

Fig. 5. Association analysis over 1 cM intervals on Chromosome 2, 77-78 cM (A) and 183-184 cM (B), respectively. The SNPs detected to be imprinted are labelled with the corresponding P-values of C-PAT test (in parenthesis).

smaller P-values than C-TDT. Because C-TDT and C-TDTI* are shown to have similar and correct type I error rates in the simulation study, the results exhibit considerable increase in information by using the test statistic C-TDTI*, which incorporates the imprinting effects into association analysis.

5 CONCLUSION

We have proposed a series of two-stage test statistics for association analysis incorporating imprinting effects and introduced a general approach to investigate the asymptotic distributions of these statistics. Our method enjoys nice properties. The transmission-based test statistics are calculated within families to protect against possible population stratification. The test statistics allow for a flexible use of complete and incomplete families, and thus provide an efficient way for dealing with the situation in which one parental genotype is missing. Moreover, the method accommodates nuclear families with multiple affected children, and the dependency between siblings in each family is considered in the calculation of estimates for variances and covariances involved in the test statistics. The simulation study shows that the proposed statistics control the size well, as well as achieve higher power in association testing than the existing methods when there exist imprinting effects from modest to large. We have also demonstrated the practical advantages of our method in the empirical association study on UK RA data.
 Inspired by the test TDTI, TDTI* is constructed for complete nuclear families, and is shown to be more powerful than TDTI when testing for association. Since the two tests distinguish from each other by the use of imprinting test statistics in the first stage, additional simulations have been carried out to compare the power of PAT and POET for imprinting effects under various scenarios. It was found that PAT is more powerful than POET for various choices of marker and disease allele frequencies, genetic model and LD coefficient (see Supplementary Materials for results). So the above-mentioned findings regarding TDTI* and TDTI are not obtained by chance.

So far, our simulation results for the sizes and the powers of the tests are based on the asymptotic normality assumption, which might be questionable when the study is small. To assess the validity of the proposed tests for smaller datasets, we have also conducted additional simulations with sample size being half or a quarter of the ones presented previously in the article, and we did not find any evidence of notable deviation from the nominal level of 5% (Supplementary Materials). Since most practical studies would have sample sizes at least matching the smallest of those in our simulation study, we conclude that the use of asymptotic normality assumption is in general reasonable and efficient.

As the future work, we plan to extend our method to the association analysis on quantitative trait loci, where the distributions is in general reasonable and efficient.

REFERENCES


