BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 18 2011, pages 2571–2577
doi:10.1093/bioinformatics/btr443

Genetics and population analysis

Advance Access publication July 28, 2011

A powerful approach for association analysis incorporating imprinting effects

Fan Xia[1], Ji-Yuan Zhou[2] and Wing Kam Fung[1,*]

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

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: For a diallelic 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 Angleman 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 the case of missing genotypes among parents, 1-PAT was proposed for incomplete nuclear families (nuclear families with a single parent) with an arbitrary number of affected children (Zhou *et al.*, 2009). Based on a mixture of complete and incomplete nuclear families, C-PAT, a combination of PAT and 1-PAT, was suggested to increase the test power (Zhou *et al.*, 2009). Other methods for the detection of imprinting effects were generally derived in the framework of log-linear model, leading to the likelihood ratio tests (Weinberg *et al.*, 1998).

When the genes are imprinted, the performance of gene mapping would be affected. As a result, studies on incorporating imprinting effects into genetic linkage or association analysis have arisen the interest of researchers (Hu *et al.*, 2007a; Shi *et al.*, 2007; Strauch *et al.*, 2000; Weinberg *et al.*, 1998; Wu *et al.*, 2005). Among the approaches, a test statistic termed TDTI (Hu *et al.*, 2007a) 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, TDT_p (the paternal version of TDT) for maternal imprinting, and TDT_m (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. 1-TDTI (Hu *et al.*, 2007b), which is essentially the T_2 of the 1-TDT (Sun *et al.*, 1999) tests, is a valid linkage/association test in the presence of imprinting effects for incomplete nuclear 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

*To whom correspondence should be addressed.

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-TDT$_{\rm p}$ (the paternal version of 1-TDT) and 1-TDT$_{\rm m}$ (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-TDT$_{\rm p}$ and 1-TDT$_{\rm m}$ to test for association in the second stage based on the findings of the 1-PAT. Thus, a test statistic 1-TDTI* is developed for use in incomplete nuclear families. Further, combining complete nuclear families with incomplete nuclear families, the C-TDTI* is suggested to efficiently utilize the information in a dataet. TDTI*, 1-TDTI* and C-TDTI* 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-TDTI* to the genome-wide association study of rheumatoid arthritis.

2 METHODS

2.1 Background and notations

Consider a marker locus (ML) with two alleles M_1 and M_2, 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_1, carried by the father, mother and affected child j of a nuclear family, respectively. Let $\phi_{d/d}$, $\phi_{d/D}$, $\phi_{D/d}$ and $\phi_{D/D}$ 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 the Mendel's law holds, $\phi_{d/D}=\phi_{D/d}$; otherwise, there may exist parent-of-origin effects at the DSL. The degree of imprinting effects is introduced as $I=(\phi_{D/d}-\phi_{d/D})/2$, where $I<0$ ($I>0$) signifies paternal (maternal) imprinting and $I=0$ implies that risks for the two heterozygous DSL genotypes are identical, i.e. no imprinting effects (Strauch *et al.*, 2000).

As in earlier work (Hu *et al.*, 2007a; Weinberg, 1999), mating symmetry in the parental generation is assumed throughout this article, that is, $\Pr(F=f,M=m)=\Pr(F=m,M=f)$ 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 TDTI*, where we first consider that there are n_C complete nuclear families and $l(i)$ 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, TDT$_{\rm p}$ and TDT$_{\rm m}$ are similarly expressed. For example, the difference statistics for PAT and TDT are respectively listed as below,

$$s_{Iij}=I_{F_i>M_i,C_{ij}=1}-I_{F_i<M_i,C_{ij}=1},$$
$$s_{ij}=T_{ij}-NT_{ij},\ i=1,\ldots n_C;j=1,\ldots,l(i),$$

where $I_{\{\text{comparison statement}\}}$ is 1 when the comparison statement holds and is 0 otherwise, and T_{ij} and NT_{ij} 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 disease gene, s_{Iij} has a zero mean. Under the null hypothesis of TDT test, which has no association between the marker allele and the disease gene or no linkage between the ML and DSL, s_{ij} has a zero mean. Further, by summing over all the case-parents trios, define $s_I=\sum_{i=1}^{n_C}\sum_{j=1}^{l(i)}s_{Iij}$ and $s=\sum_{i=1}^{n_C}\sum_{j=1}^{l(i)}s_{ij}$, and the PAT and TDT can then be expressed as follows,

$$\text{PAT}=\frac{s_I}{\sqrt{\widehat{\text{Var}_0}\,(s_I)}},\ \text{TDT}=\frac{s}{\sqrt{\widehat{\text{Var}_0}\,(s)}},$$

where $\widehat{\text{Var}_0}(s_I)$ and $\widehat{\text{Var}_0}(s)$ are the corresponding unbiased estimates of the variances of s_I and s under the null hypotheses of the PAT and the TDT tests, respectively. $\widehat{\text{Var}_0}(s_I)=\sum_i(\sum_j s_{Iij}^2+2\sum_{k<j}s_{Iij}s_{Iik})$ and $\widehat{\text{Var}_0}(s)=\sum_i(\sum_j s_{ij}^2+2\sum_{k<j}s_{ij}s_{ik})$. TDT$_{\rm p}$ and TDT$_{\rm m}$ could be similarly expressed from the corresponding difference statistics $s_{pij}=T_{pij}-NT_{pij}$ and $s_{mij}=T_{mij}-NT_{mij}$, where T_{pij} (T_{mij}) and NT_{pij} (NT_{mij}), 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 j. With the available statistics, the test statistic TDTI* is formulated as

$$\begin{aligned}\text{TDTI*}\ =\ &\text{TDT}_{\rm m}I_{\left\{\text{PAT}<-z_{\alpha_1/2}\right\}}\\&+\text{TDT}I_{\left\{|\text{PAT}|\le z_{\alpha_1/2}\right\}}\\&+\text{TDT}_{\rm p}I_{\left\{\text{PAT}>z_{\alpha_1/2}\right\}},\end{aligned}$$

where α_1 is the pre-specified significance level in testing for imprinting effects, and $z_{\alpha_1/2}$ is the upper $\alpha_1/2$ quantile of a standard normal distribution. TDTI* illustrates that when the imprinting effect is significant enough, TDT$_{\rm m}$ or TDT$_{\rm p}$ 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 TDT$_{\rm m}$, TDT$_{\rm p}$ and TDT and found TDT$_{\rm m}$ (TDT$_{\rm p}$) 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/TDT/TDTp is bivariate normal with mean vector being zero. Therefore, to find the asymptotic null distribution and P-value of the proposed test TDTI*, 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

$$\widehat{\text{Cov}_0}(\text{PAT},\ \text{TDT})=\frac{\widehat{\text{Cov}_0}(s_I,s)}{\sqrt{\widehat{\text{Var}_0}(s_I)}\sqrt{\widehat{\text{Var}_0}(s)}},$$

where $\widehat{\text{Cov}_0}(s_I,s)=\sum_i\sum_j\sum_k s_{Iij}s_{ik}$, and is the unbiased estimate of the covariance between s_I and s under the null hypothesis of no association (Supplementary Materials). The estimates of the covariances between PAT and TDT$_{\rm m}$ and between PAT and TDT$_{\rm p}$ 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 test statistic in the second stage for the TDTI* be t_1 and t_2, then

$$\begin{aligned}&\Pr\left(\text{TDTI*}\le t_2\right)\\&=\begin{cases}\Pr(\text{PAT}<-z_{\alpha_1/2},\text{TDT}_{\rm m}\le t_2),\text{if } t_1<-z_{\alpha_1/2}\\\Pr(|\text{PAT}|\le z_{\alpha_1/2},\text{TDT}\le t_2),\text{if } |t_1|\le z_{\alpha_1/2}\\\Pr(\text{PAT}>z_{\alpha_1/2},\text{TDT}_{\rm p}\le t_2),\text{if } t_1>z_{\alpha_1/2}\end{cases}\end{aligned}$$

(see Supplementary Materials for the computation of P-value of TDTI*). 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 TDTI* 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-TDTI*, where the notations involved with incomplete nuclear families are as follows. Suppose there are n_I incomplete

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

Observed genotypes of the available parent and the affected child[a]	Possible genotype of the missing parent	Transmitted allele from the missing parent
$\{M_1M_2, M_1M_1\}$	M_1M_2/M_1M_1	M_1
$\{M_2M_2, M_1M_2\}$	M_1M_2/M_1M_1	M_1
$\{M_1M_1, M_1M_2\}$	M_1M_2/M_2M_2	M_2
$\{M_1M_2, M_2M_2\}$	M_1M_2/M_2M_2	M_2

[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_M single-mother families and n_F 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 pair. 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 assoctaion 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 with the above-mentioned statistics, we define $s'_{pij}=I_{M_i<C_{ij}}-I_{M_i>C_{ij}}$ and $s'_{mij}=I_{F_i<C_{ij}}-I_{F_i>C_{ij}}$ for each case-mother pair and each case-father pair, respectively. Under the null hypothesis of no association, s'_{pij} and s'_{mij} have expectations to be zero (Supplementary Materials). Let $s'_p=\sum_{i=1}^{n_M}\sum_{j=1}^{l(i)}s'_{pij}$, $s'_m=\sum_{i=1}^{n_F}\sum_{j=1}^{l(i)}s'_{mij}$, we have

$$1\text{-TDT}=\frac{s'_m+s'_p}{\sqrt{\widehat{\text{Var}_0}\left(s'_m\right)+\widehat{\text{Var}_0}\left(s'_p\right)}},$$

where $\widehat{\text{Var}_0}\left(s'_p\right)$ and $\widehat{\text{Var}_0}\left(s'_m\right)$ are accordingly defined and similarly calculated as $\widehat{\text{Var}_0}\,(s)$.

On the other hand, motivated by the construction of TDT_p and TDT_m, we further develop two test statistics 1-TDT_p and 1-TDT_m 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. In the effort to construct these two statistics, we first illustrate the design of 1-TDT, 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-TDT_p. On the contrary, the 1-TDT merely calculated with single-father families essentially tests for association based on the transmission information of the missing mother. We take it as the maternal version of 1-TDT (1-TDT_m). 1-TDT_p and 1-TDT_m are formulated as follows,

$$1\text{-TDT}_\text{p}=\frac{s'_p}{\sqrt{\widehat{\text{Var}_0}\left(s'_p\right)}},\ 1\text{-TDT}_\text{m}=\frac{s'_m}{\sqrt{\widehat{\text{Var}_0}\left(s'_m\right)}}.$$

Indeed, 1-TDT_p and 1-TDT_m can be more efficient than 1-TDT when there are maternal imprinting effect and paternal imprinting effect, respectively (Supplementary Materials).

To investigate the imprinting effects for the selection of an optimal statistic among the three 1-TDT-type statistics, 1-PAT, formulated as $s'_I\Big/\sqrt{\widehat{\text{Var}_0}(s'_I)}$, is applied in incomplete nuclear families, where

$$\begin{aligned}s'_I\ =\ & w\sum_{i=1}^{n_M}\sum_{j=1}^{l(i)}\left(I_{M_i<C_{ij},\,C_{ij}=1}-I_{M_i>C_{ij},\,C_{ij}=1}\right)\\ &+(1-w)\sum_{i=1}^{n_F}\sum_{j=1}^{l(i)}\left(I_{F_i>C_{ij},\,C_{ij}=1}-I_{F_i<C_{ij},\,C_{ij}=1}\right),\end{aligned}$$

and $w=\sum_{i=1}^{n_F}l(i)/\sum_{i=1}^{n_I}l(i)$, which is the proportion of case-father pairs to case-parent pairs. Therefore, 1-TDTI* is constructed in an identical framework of TDTI* with expression as follows,

$$\begin{aligned}1\text{-TDTI*}\ =\ & 1\text{-TDT}_\text{m}I_{\left\{1-\text{PAT}<-z_{\alpha_1/2}\right\}}\\ &+1\text{-TDT}I_{\left\{|1-\text{PAT}|\le z_{\alpha_1/2}\right\}}\\ &+1\text{-TDT}_\text{p}I_{\left\{1-\text{PAT}>z_{\alpha_1/2}\right\}}.\end{aligned}$$

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* and 1-TDTI*, where we first specify a set of combined statistics for the association test, C-TDT, C-TDT_p and C-TDT_m, through natural combinations of TDT and 1-TDT, TDT_p and 1-TDT_p, TDT_m and 1-TDT_m, respectively. For instance,

$$\text{C-TDT}=\frac{s+s'}{\sqrt{\widehat{\text{Var}_0}\,(s)+\widehat{\text{Var}_0}\,(s')}},$$

where $s'=s'_m+s'_p$ and then $\widehat{\text{Var}_0}\left(s'\right)=\widehat{\text{Var}_0}\left(s'_m\right)+\widehat{\text{Var}_0}\left(s'_p\right)$. Based on the previous discussions, C-TDT_p, C-TDT and C-TDT_m follow standard normal distributions under the null hypothesis of no association, and are the corresponding optimal statistics to test for association in the presence of maternal imprinting, no imprinting and paternal imprinting. Further, we apply C-PAT, equal to $(s_I+s'_I)\Big/\sqrt{\widehat{\text{Var}_0}\left(s_I\right)+\widehat{\text{Var}_0}\left(s'_I\right)}$, for imprinting test in the first stage and formulate C-TDTI* as,

$$\begin{aligned}\text{C-TDTI*}\ =\ & \text{C-TDT}_\text{m}I_{\left\{\text{C}-\text{PAT}<-z_{\alpha_1/2}\right\}}\\ &+\text{C-TDT}I_{\left\{|\text{C}-\text{PAT}|\le z_{\alpha_1/2}\right\}}\\ &+\text{C-TDT}_\text{p}I_{\left\{\text{C}-\text{PAT}>z_{\alpha_1/2}\right\}}.\end{aligned}$$

Since the complete nuclear families and incomplete nuclear families are independent, the covariances between C-PAT and C-TDT type statistics could be easily estimated to find the asymptotic null distribution of C-TDTI*.

3 SIMULATION STUDY

3.1 Settings

A simulation study is carried out to check the validity and to evaluate the performance of the proposed test statistics TDTI*, 1-TDTI* and C-TDTI*. Existing test statistics TDT and TDTI, 1-TDT and 1-TDTI, C-TDT and C-TDTI, which is just a simple sum of TDTI and 1-TDTI, are selected for comparison in terms of size and power for situations with or without imprinting. Throughout the simulation, the additive genetic model is applied at the DSL with $\phi_{D/D}=3\phi_{d/d}$, and the penetrance of heterozygous genotypes d/D and D/d are, respectively, obtained as $\phi_{d/D}=2\phi_{d/d}-I$ and $\phi_{D/d}=2\phi_{d/d}+I$, where $\phi_{d/d}$ 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

Table 2. Type I error rates in 100%

I[c]	TDTI*	1-TDTI*[a]			C-TDTI*[b]			C-TDTI			C-TDT		
		0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
−0.1	4.98	5.32	4.92	5.13	5.33	4.71	4.94	5.15	4.56	4.68	5.09	4.48	4.73
−0.05	4.95	4.98	4.73	4.90	5.33	5.28	5.13	4.74	4.85	4.88	5.13	4.98	5.03
0	4.67	5.08	5.16	5.00	4.97	5.51	5.19	4.89	5.45	4.92	4.67	5.33	4.84
0.05	5.04	5.24	5.01	4.83	4.98	5.36	5.33	4.74	5.12	4.63	4.76	5.08	4.90
0.1	4.96	4.60	4.86	5.14	5.24	5.09	5.31	5.23	4.89	4.96	4.94	4.72	4.97

[a]In each replicate, the father's or mother's genotype information is randomly deleted according to β, which is assigned with three levels (0.25, 0.5 and 0.75) as below for each family.
[b]In each replicate, n_C and n_I are determined by τ, which is assigned with three levels (0.25, 0.5 and 0.75) as below. β is fixed to be 0.5 for each incomplete nuclear family. The same settings are applied to C-TDTI and C-TDT.
[c]The five levels of I assigned could represent complete paternal imprinting, incomplete paternal imprinting, no imprinting, incomplete maternal imprinting and complete maternal imprinting, respectively.

of proposed methods. In each sample, define τ to control the ratio of incomplete nuclear families and β 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_1 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 0.001 in both subpopulations. The LD coefficient between a marker allele and a disease gene is measured by Lewontin 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 β and incomplete nuclear family ratio τ. 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 minimum power, it is still comparable to TDT, which is due to the fact that the actual test statistic of TDTI* is TDT when the null hypothesis of no imprinting in the first stage test is not rejected.

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 β. In the presence of complete paternal imprinting, it is observed that for each β, 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).

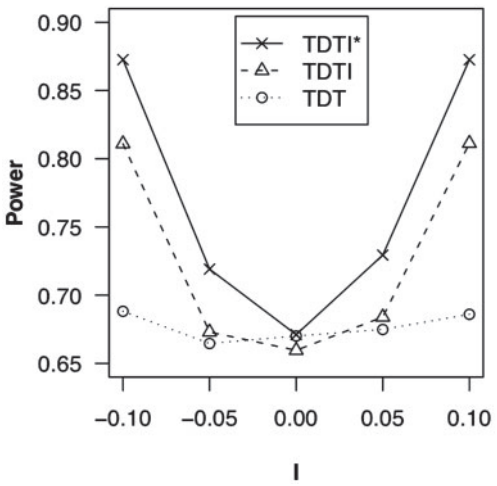

Fig. 1. Powers of TDTI*, TDTI and TDT against five levels of imprinting effects.

For both 1-TDT and 1-TDTI*, the powers decline with the increasing β value. This is because when there exists paternal imprinting effect, the case-father pairs play the major role in one-parent association analysis. 1-TDTI*, which efficiently applies 1-TDT$_m$, is then more powerful than 1-TDT. In the presence of complete maternal imprinting (results not shown), the powers of both 1-TDT and 1-TDTI* ascend with the increasing β value, and 1-TDTI* gains the highest power among the three statistics, as expected. When there is no imprinting, it appears that 1-TDT and 1-TDTI* remain the same for different β values and have similar level of testing efficiency.

For family samples combined of both complete and incomplete nuclear families, we compare the performance of C-TDTI*, C-TDTI and C-TDT in Figure 3 under different settings of τ values, while β is fixed to be 0.5 for incomplete families. The figure demonstrates that for different τ values, the power of the proposed C-TDTI* is significantly higher than the other two methods when there exist imprinting effects, which is consistent with the findings obtained from Figures 1 and 2. Meanwhile, we also find C-TDTI may even perform worse than C-TDT, and so is not recommended.

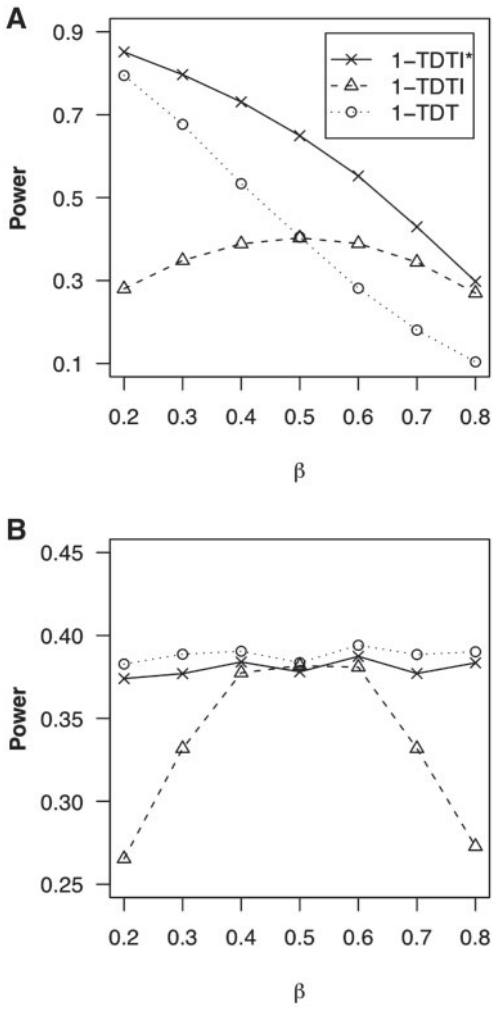

Fig. 2. Powers of 1-TDTI*, 1-TDTI and 1-TDT against father missing probability β in the presence of complete paternal imprinting (**A**) and no imprinting (**B**), respectively.

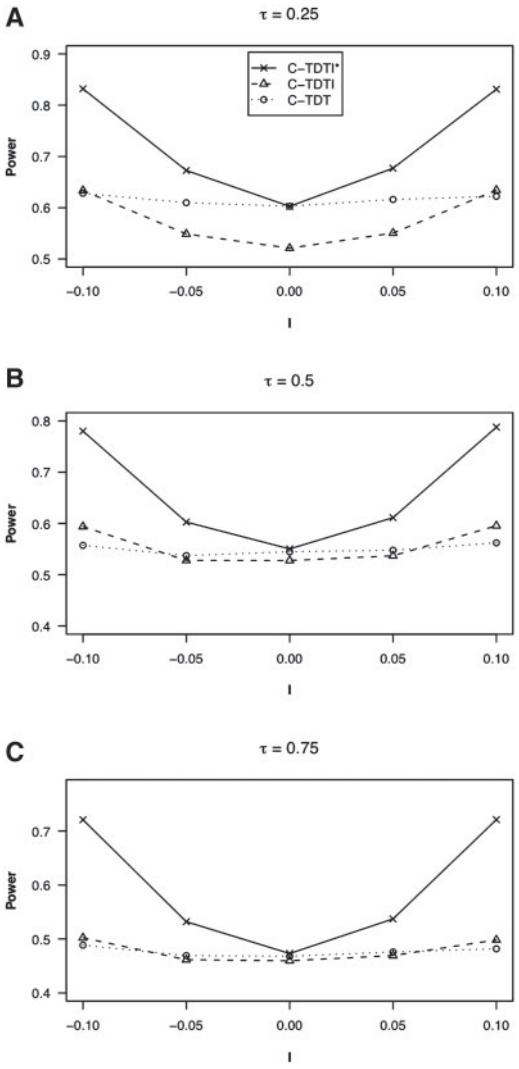

Fig. 3. Powers of C-TDTI*, C-TDTI and TDT against five levels of imprinting effects for the family samples with incomplete family ratio $\tau = 0.25$ (**A**), $\tau = 0.5$ (**B**) and $\tau = 0.75$ (**C**), respectively.

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 10 156 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-TDTI* for SNPs to make an efficient use of the available information.

The results of C-TDT and C-TDTI* based on 22 chromosomes are plotted in Figure 4. Both tests suggest that the strongest evidence for association is at SNP520297 on Chromosome 8 ($P<1\times10^{-5}$). 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-TDTI*) 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-TDTI* indeed provides notable

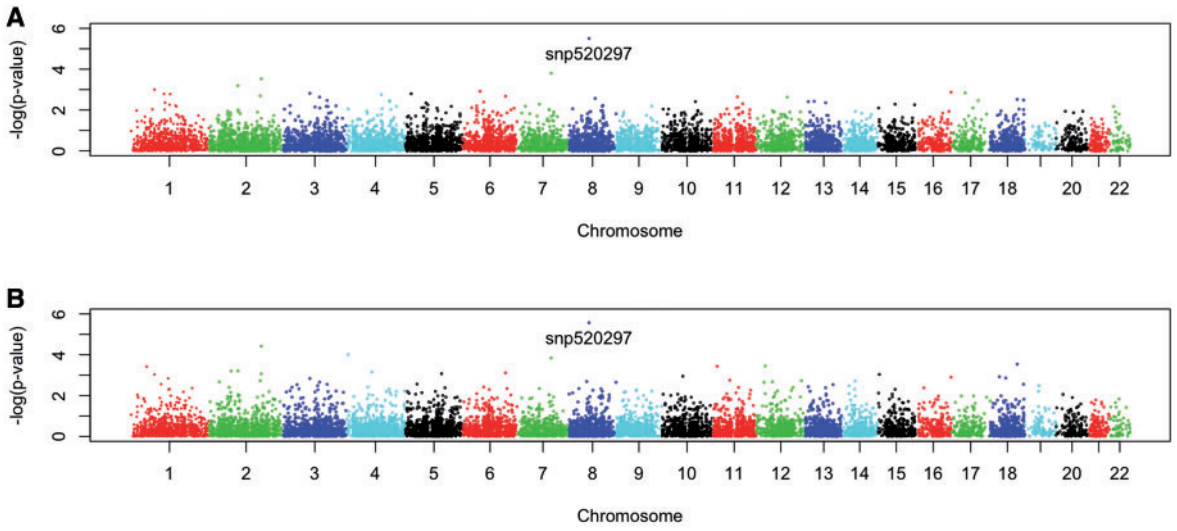

Fig. 4. Association analysis for RA over 22 chromosomes using C-TDT (**A**) and C-TDTI* (**B**) tests.

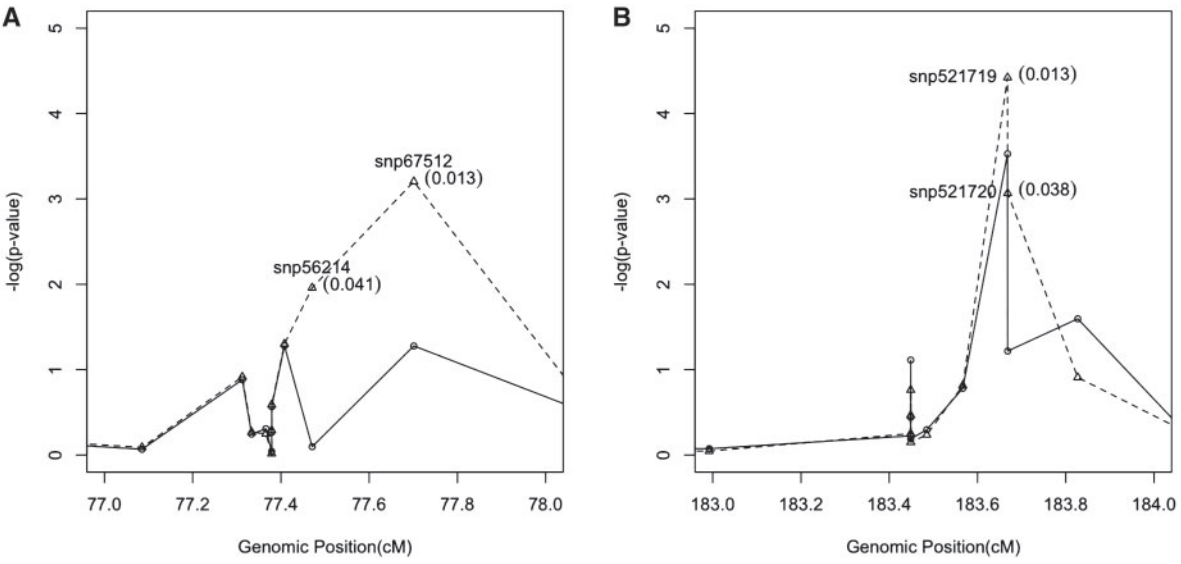

Fig. 5. Association analysis over 1 cM intervals on Chromosome 2, 77-78 cM (**A**) and 183-184 cM (**B**), respectively. The SNPs detected to be imprinted are labelled with the corresponding *P*-values of C-PAT test (*in parenthesis*).

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 one 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 of quantitative traits need to be considered.

Funding: Hong Kong RGC GRF Research Grant HKU 766511M; National Natural Science Foundation of China (grant 81072386); National Institutes of Health grant (R01 GM031575) to The Genetic Analysis Workshops; National Institutes of Health (NO1-AR-2-2263 and RO1-AR-44422 to gather RA data; National Arthritis Foundation to gather RA data).

Conflict of Interest: none declared.

REFERENCES

Amos,C.I. *et al.* (2007) Data for genetic analysis workshop (GAW) 15 problem 2, genetic causes of rheumatoid arthritis and associated traits. *BMC Proc.*, **1**, S3.

Cordell,H.J. and Clayton,D.G. (2005) Genetic association studies. *Lancet*, **366**, 1121–1131.

Falls,J.G. *et al.* (1999) Genomic imprinting: implications for human disease. *Am. J. Pathol.*, **154**, 635–647.

Gregersen,P.K. (1999) Genetics of rheumatoid arthritis: confronting complexity. *Arthritis Res.*, **1**, 37–44.

Hu,Y.Q. *et al.* (2007a) An extension of the transmission disequilibrium test incorporating imprinting. *Genetics*, **175**, 1489–1504.

Hu,Y.Q. (2007b) The transmission disequilibrium test and imprinting effects test based on case-parent pairs. *Genet. Epidemiol.*, **31**, 273–287.

John,S. *et al.* (2004) Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet.*, **75**, 54–64.

Lewontin,R.C. (1988) On measures of gametic disequilibrium. *Genetics*, **120**, 849–852.

Morison,I.M. *et al.* (2001) The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.*, **29**, 275–276.

Pfeifer,K. (2000) Mechanisms of genomic imprinting. *Am. J. Hum. Genet.*, **67**, 777–787.

Reik,W. and Walter,J. (2001) Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, **2**, 21–32.

Shi,M. *et al.* (2007) Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families. *Am. J. Hum. Genet.*, **81**, 53–66.

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

Strauch,K. *et al.* (2000) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. *Am. J. Hum. Genet.*, **66**, 1945 – 1957.

Sun,F. *et al.* (1999) Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am. J. Hum. Genet.*, **150**, 97–104.

Weinberg,C.R. *et al.* (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am. J. Hum. Genet.*, **62**, 969–978.

Weinberg,C.R. (1999) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am. J. Hum. Genet.*, **65**, 229–235.

Wu,C.-C. *et al.* (2005) Linkage analysis of affected sib pairs allowing for parent-of-origin effects. *Ann. Hum. Genet.*, **69**, 113–126.

Zhou,J.Y. *et al.* (2009) Detection of parent-of-origin effects based on complete and incomplete nuclear families with multiple affected children. *Hum. Hered.*, **67**, 1–12.

Zhou,X. *et al.* (2007) Joint linkage and imprinting analyses of GAW15 rheumatoid arthritis and gene expression data. *BMC Proc.*, **1**, S53.