

Ascertainment Through Family History of Disease Often Decreases the Power of Family-based Association Studies

Manuel A. R. Ferreira · Pak Sham · Mark J. Daly · Shaun Purcell

Received: 8 September 2006 / Accepted: 15 February 2007 / Published online: 20 March 2007
© Springer Science+Business Media, LLC 2007

Abstract Selection of cases with additional affected relatives has been shown to increase the power of the case-control association design. We investigated whether this strategy can also improve the power of family-based association studies that use the transmission disequilibrium test (TDT), while accounting for the effects of residual polygenic and environmental factors on disease liability. Ascertainment of parent-offspring trios conditional on the proband having affected first-degree relatives almost always reduced the power of the TDT. For many disease models, this reduction was quite considerable. In contrast, for the same sample size, designs that analyzed more than one affected offspring per family often improved power when compared to the standard parent-offspring trio design. Together, our results suggest that (1) residual polygenic and environmental influences should be considered when estimating the power of the TDT for studies that ascertain families with multiple affected relatives; (2) if trios are selected conditional on having additional affected

offspring, then it is important to genotype and include in the analysis the additional siblings; (3) the ascertainment strategy should be considered when interpreting results from TDT analyses. Our analytic approach to estimate the asymptotic power of the TDT is implemented online at <http://pngu.mgh.harvard.edu/~purcell/gpc/>.

Keywords Association · TDT · Power · Study design · Complex disease · Family history

Introduction

Previous studies have demonstrated that the power of the case-control association design can be greatly improved by selecting “familial cases”, i.e. affected individuals with additional affected relatives (Risch 2001; Antoniou and Easton 2003; Howson et al. 2005; Li et al. 2006). The explanation for this is that by selecting familial cases, the frequency of the disease allele in the case sample will often increase (Risch 2001). However, this effect is somehow complicated by the number and magnitude of additional loci that contribute to disease risk. As the background heritability increases, the enrichment of the disease allele in familial cases is less pronounced (Risch 2001) and, consequently, the improvement in power as a result of selection is more modest (Li et al. 2006). For very high background heritability, selection may in fact reduce the power of the case-control design (Li et al. 2006). Nonetheless, across a realistic range of disease parameters, selection of familial cases seems to improve power.

We were interested in investigating whether ascertainment of familial cases can also improve the power of the most popular test used in family-based association designs, the transmission disequilibrium test (TDT) (Spielman and

Edited by David Allison

M. A. R. Ferreira (✉) · M. J. Daly · S. Purcell
Center for Human Genetic Research, Massachusetts General
Hospital, Harvard Medical School, 185 Cambridge St, Boston,
MA 02114, USA
e-mail: mferreira@chgr.mgh.harvard.edu

P. Sham
Institute of Psychiatry, King’s College, London, UK

P. Sham
Genome Research Centre, University of Hong Kong, Pokfulam,
Hong Kong

M. J. Daly · S. Purcell
Broad Institute of Harvard & MIT, Cambridge, MA, USA

Ewens 1996). Given the importance of efficient study design, particularly for large-scale association studies, surprisingly few studies have estimated the power of the TDT for different ascertainment strategies. Whittaker and Lewis (1998) showed that for a range of disease models, the power of the TDT improves when trios are selected conditional on having an affected parent. However, the authors implicitly assumed that no other genetic or shared environmental factors contributed to disease risk, which is unrealistic for common, multifactorial diseases. Indeed, Risch (2001) later showed that for diseases with high background heritability, there may be little advantage in ascertaining families with multiple affected offspring as opposed to families with one affected offspring only.

In this report, we quantify the power of the TDT for different family designs and disease models. Specifically, we address two separate but related issues: first, can the power of the TDT be improved by ascertaining parent-offspring trios conditional on the proband having affected first-degree relatives? Henceforth we refer to this strategy as ascertainment of trios for family history of disease. Second, for a fixed number of genotypes, which strategy is more powerful, to analyze families with one affected offspring or, alternatively, with multiple affected offspring included in the association analysis (“multiplex” design)? The first issue concerns ascertainment only: we compare designs that differ only in the number of affected first-degree relatives the offspring must have for the trio to be ascertained (design 1 versus designs 2–6, Table 1). In this case, we analyze the same family structure for all designs considered, i.e. two parents and one affected offspring. On the other hand, to address the second issue, we compare family designs that differ in the number of affected offspring used in the TDT (design 1 versus designs 7 and 8, Table 1). Clearly, in this case, the designs differ in terms of both the ascertainment and family structure used for the TDT.

Material and methods

Below we outline our approach to estimate the asymptotic power of the TDT as a function of the ascertainment scheme, the properties of the disease locus and the residual genetic and environmental influences on disease. Following the standard biometrical model (Falconer 1981), we assume that liability to disease is continuously distributed and determined by a major biallelic locus, with the residual variance partitioned into additive polygenic background, and both shared (i.e. familial) and unique environment components. Individuals that exceed a threshold t in liability are affected. We model the three genotypic means as a (genotype AA), d (Aa) and $-a$ (aa), where a is the additive

Table 1 Asymptotic power of the TDT for eight family designs with a fixed total sample size of 1,500 genotyped individuals ($\alpha = 0.05$)

Design ^a	Power of the TDT
1	0.918
2	0.759
3	0.602
4	0.798
5	0.665
6	0.553
7	0.903
8	0.845

^a Individuals are shown in white if their disease status is unknown and have been genotyped, in black if affected and genotyped, and in grey if affected but not genotyped. The disease prevalence was 5%, the additive test locus ($p = 0.2$) explained 1% of the variance in liability to disease, the background heritability was 50% and family environmental variance was 10%

genetic value of the major locus, d is the dominance deviation, and p and q are the frequency of alleles A and a in the population ($q = 1 - p$), respectively. For a given set of a , d and p parameters, the contribution of the major locus to the total variance in liability to disease can be calculated as $2 \cdot p \cdot q \cdot (a + d \cdot (q - p))^2 + (2 \cdot p \cdot q \cdot d)^2$.

To calculate power, we require the expected number of transmitted and untransmitted alleles from heterozygous parents to affected offspring, given ascertainment and the disease model. These quantities are obtained by calculating the probability of each possible family genotypic configuration given the phenotypic ascertainment scheme, $P(G_O, G_P, G_M | P_O, P_P, P_M = A)$, where G_O represents the set of genotypes for one or more offspring, G_P and G_M are paternal and maternal genotypes; P represents the

phenotypes (same subscript) and A denotes the phenotypic ascertainment strategy. The standard TDT strategy is to ascertain families with a single affected offspring, disregarding the affection status of other relatives. We also consider designs in which other first-degree relatives (parents or siblings) are also affected. We use Bayes theorem to obtain $P(G_O, G_P, G_M | P_O, P_P, P_M = A)$ from $P(G_O, G_P, G_M)$ and $P(P_O, P_P, P_M = A | G_O, G_P, G_M)$. $P(G_O, G_P, G_M)$ is the probability of each genotypic configuration in the unselected population; this is a simple function of genotype frequencies assuming Hardy–Weinberg equilibrium and Mendelian transmission probabilities. For example, $P(G_O = "AA", G_P = "AA", G_M = "Aa") = p^2 \cdot 2pq \cdot 0.5 = p^3q$. The probability of the families' phenotypic configuration given their genotypic configuration, $P(P_O, P_P, P_M = A | G_O, G_P, G_M)$, is given by numerical integration of the liability threshold model multivariate normal distribution with mean vector μ (a function of the major locus) and covariance matrix Σ (a function of the residual polygenic and environmental components). Thus, the measured effect of genotype is specified in a model for the means, whereas the effects of unmeasured polygenic and environmental components are modeled as random effects in the covariance structure. This approach has been well described in a number of publications (e.g. Fulker et al. 1999; Abecasis et al. 2000).

Finally, we obtain the expected number of transmitted and untransmitted alleles, b and c , as weighted sums of 0, 1 or 2 transmissions or non-transmissions per offspring per family over $P(G_O, G_P, G_M | P_O, P_P, P_M = A)$ multiplied by the total number of families (for a fixed family type). The expected TDT test statistic is then computed as $\text{asTDT} = \frac{(b-c)^2}{b+c}$; power for a given type-I error rate α corresponds to the area of a non-central χ^2_1 with non-centrality parameter = TDT that falls beyond the quantile $q(\alpha)$ from a central χ^2_1 (e.g. $q = 3.84$ for $\alpha = 0.05$). Note that our analytical approach accounts for the non-independence of parental transmissions to affected offspring and so it is valid for any modes of inheritance and number of affected offspring used in the TDT. The power estimates derived from our approach matched to the second decimal place the empirical power obtained through the analysis of 100,000 dataset replicates (not shown).

Results

To illustrate the ascertainment strategies and family structures tested, we first considered a single example disease model (Table 1). Both parents were assumed to be genotyped for all designs and the total sample size was fixed at 1,500 individuals. We considered a disease prevalence of 5%, an additive trait locus ($p = 0.2$) that ex-

plained 1% of the variance in liability to the disease (equivalent to a genotype relative risk of 1.9 for the heterozygote), a background heritability of 50% and a family environmental variance of 10%. Under this model, the power of the TDT when analyzing 500 parent-offspring trios ascertained with disregard to the affection status of first-degree relatives of the proband (design 1, which we refer to henceforth as the "standard parent-offspring trio design") was 0.918 ($\alpha = 0.05$). However, power decreased steadily as the number of affected first-degree relatives required to ascertain parent-offspring trios increased, reaching 0.553 for cases with three additional affected relatives (design 6). For this particular disease model, the standard parent-offspring trio design also outperformed the designs that genotyped >1 affected offspring per family (multiplex designs 7 and 8).

Next, we investigated whether the detrimental impact of selection on the TDT could be expected across a wider range of disease models. For this purpose, we considered three values for the disease prevalence (1%, 5% and 10%), two values for the variance in liability attributable to the test locus (1% and 2.5%), three genetic models (dominant, additive and recessive), three values for the family environment variance (V_C) (0%, 10% and 20%) and we varied the background heritability (V_A) between 0% and 75%, and the allele frequency p between 0.025 and 0.40. For simplicity, we restricted our analysis to five family structures (designs 1–3, 7 and 8).

Figure 1 compares the expected TDT test statistic between family design 1 (standard parent-offspring trio design) and two designs that ascertain trios conditional on family history (designs 2 and 3). The color of each cell indicates which design resulted in the largest test statistic, parent-offspring trios (blue) or the trios with one (green) or two (red) additional affected siblings. The ability of family history to improve the power of the TDT was greatly influenced by the family environment variance (V_C), the background heritability (V_A), the allele frequency (p) and the genetic model considered. Disease prevalence and locus effect size had only a minor impact on the results and so data are only presented for a 5% disease prevalence and 1% locus effect size. Ascertainment based on family history reduced power in most scenarios, with the exception of recessive loci, particularly for rare risk alleles and low background heritability and family-specific environmental influences.

We then performed the same analysis to compare the multiplex designs (7 and 8) against the standard parent-offspring trio design (design 1), maintaining a constant total sample size. When disease liability was not influenced by the family environment ($V_C = 0\%$), analyzing multiplex families was often the most powerful design (Fig. 2). In this case, the value of larger sibships increased as the

Fig. 1 Comparison of the TDT test statistic between the standard parent-offspring trio design and two designs that ascertain trios conditional on family history. The color of each cell indicates which design resulted in the largest test statistic, parent-offspring trios ascertained with disregard to family history (design 1, in blue) or trios ascertained conditional on having one (design 2, in green) or two (design 3, in red) additional affected siblings. The two shades of green identify the design with the second best test statistic, design 1 (dark green) or design 3 (light green). Disease prevalence was 5% and the test locus explained 1% of the variance in disease liability. Total sample size was 1,500 individuals (including parents) for the three designs

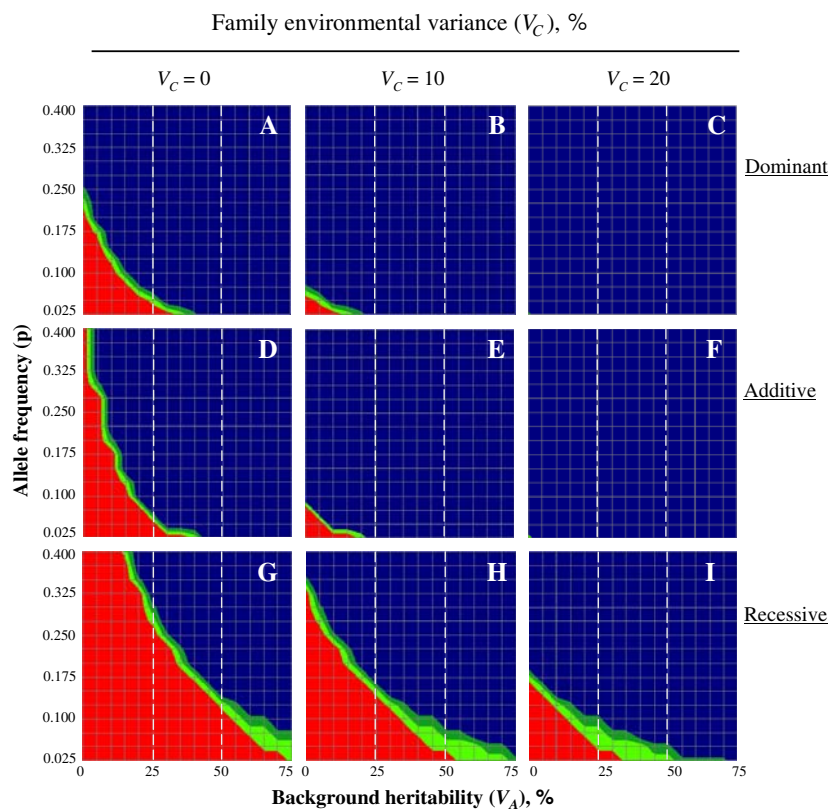
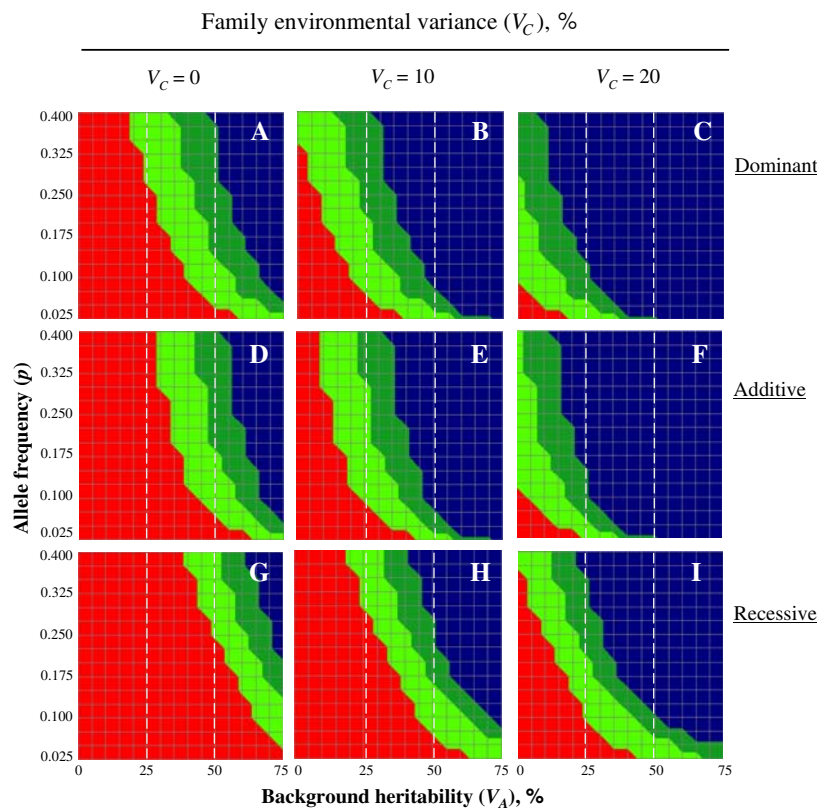


Fig. 2 Comparison of the TDT test statistic between the standard parent-offspring trio design and two multiplex designs. The color of each cell indicates which design resulted in the largest test statistic, parent-offspring trios ascertained with disregard to family history (design 1, in blue) or families with two (design 7, in green) or three (design 8, in red) genotyped affected offspring. The two shades of green identify the design with the second best test statistic, design 1 (dark green) or design 8 (light green). Disease prevalence was 5% and the test locus explained 1% of the variance in disease liability. Total sample size was 1,500 individuals (including parents) for the three designs



background heritability and allele frequency decreased. On the other hand, when $V_C > 0\%$, larger sibships were beneficial if the background heritability was low to moderate (<50%), particularly when the risk allele was relatively rare.










Discussion

It is well established that selection of familial cases often improves the power of the case-control association design (Risch 2001; Antoniou and Easton 2003; Howson et al. 2005; Li et al. 2006). Despite a commonly-held belief that such ascertainment also improves the power of family-based tests, this hypothesis has not been thoroughly tested. Our results show that selection of familial cases can improve the power of the TDT for diseases in which additional untested loci and family environmental effects do not strongly influence disease risk. These results are consistent with those reported by Whittaker and Lewis (1998) but are unlikely to be applicable to most common, complex diseases, which are expected to be influenced by many disease loci and, in some cases, by family environmental factors. In this case, our results suggest that enrichment for familial aggregation of disease will often decrease the power of the TDT. For many disease models tested, this reduction in power was quite considerable.

One intuitive explanation for these results is as follows. The power of the TDT increases with (a) increased transmission rate of the risk allele to the affected offspring and (b) increased parental heterozygosity. Ascertainment can reduce the transmission rate of the disease allele to affected offspring, especially when residual polygenic and family environmental effects are present. This occurs because ascertainment on family history acts to increase the effective prevalence of the disease (as in this case the probability of developing disease given family history, i.e. the recurrence risk, is substantially greater than the population prevalence) and the power of the TDT is influenced by prevalence. Specifically, the TDT is less powerful for higher prevalences, given the same effects on liability, which is why the undertransmission of the disease allele to unaffecteds contains little or no information for association.

On the other hand, selection of familial cases will increase the frequency of the disease variant in founders, which will in turn increase parental heterozygosity, particularly for rarer, more highly penetrant disease alleles (although not when the disease causing variant is the more common allele). However, the increase in parental heterozygosity due to ascertainment is less pronounced with increasing residual polygenic and family environmental effects, again due to the increased effective prevalence.

Table 2 Selection conditional on family history influences parental heterozygosity and the transmission rate of the risk allele

Allele frequency p	Background heritability, %	Proportion of heterozygote parents ^a			Proportion of heterozygote parents that transmit the risk allele ^b			Expected TDT test statistic		
										
0.05	0	0.131	0.177	0.234	0.653	0.652	0.651	12.215	16.386	21.385
	25	0.131	0.166	0.199	0.653	0.639	0.627	12.215	12.780	12.868
	50	0.131	0.158	0.178	0.653	0.628	0.610	12.215	10.256	8.586
	75	0.131	0.151	0.163	0.653	0.618	0.597	12.215	8.411	6.122
0.20	0	0.358	0.395	0.431	0.589	0.588	0.588	11.230	12.335	13.356
	25	0.358	0.387	0.410	0.589	0.580	0.573	11.230	9.925	8.773
	50	0.358	0.381	0.396	0.589	0.573	0.563	11.230	8.171	6.231
	75	0.358	0.376	0.386	0.589	0.568	0.555	11.230	6.845	4.659
0.40	0	0.495	0.504	0.508	0.574	0.574	0.573	10.760	10.916	10.954
	25	0.495	0.502	0.505	0.574	0.567	0.561	10.760	8.904	7.467
	50	0.495	0.500	0.502	0.574	0.561	0.552	10.760	7.413	5.447
	75	0.495	0.499	0.500	0.574	0.556	0.546	10.760	6.269	4.150

^a Defined as $\frac{b+c}{2N}$, where b and c are the number of heterozygote parents that transmit the risk or non-risk alleles to the affected child, respectively, and N the total number of families

^b Defined as $\frac{b}{b+c}$, where b and c are defined above

Individuals are shown in white if their disease status is unknown and have been genotyped, in black if affected and genotyped, and in grey if affected but not genotyped. The disease prevalence was 5%, the additive test locus explained 1% of the variance in liability to disease and family environmental variance was 0%

These effects are demonstrated in Table 2, in which we separate the two determinants of power for the TDT: parental heterozygosity and transmission distortion.

We also investigated whether testing multiplex families could be expected to provide a more powerful TDT than testing parent-offspring trios, for a fixed number of genotyped samples. In this case, our results suggest that multiplex families are often beneficial unless the polygenic or the family environmental variances are high. Similar conclusions were reported by Risch (2001). This extra efficiency derives from being able to re-use the same parental genotypes to determine transmissions to more than one affected offspring (for a fixed total sample size). For example, if parental heterozygosity is 0.177, 500 parent-offspring trios would be expected to have ~ 177 ($500 \text{ families} \times 2 \text{ parents} \times 0.177$) informative transmissions, whereas 375 families with two affected offspring would have ~ 265 ($375 \text{ families} \times 2 \text{ parents} \times 0.177 \times 2 \text{ offspring}$).

These results may have important implications for the design, analysis and interpretation of association studies of complex diseases. First, residual polygenic and environmental influences should be considered when estimating the power of the TDT for studies that ascertain families with multiple affected relatives. Failure to do so may result in severely biased power analyses. Similarly, if families are ascertained based on having multiple affected offspring but are then broken down into parent-offspring trios to estimate power, this will often lead to biased results. Second, if trios are selected for a family history of disease, then it is important to genotype and include in the analysis the additional affected relatives. Finally, our results imply that the ascertainment strategy should be considered when interpreting results from TDT analyses, since for the same sample size different strategies will lead to different power levels.

We have implemented our approach to estimate the asymptotic power of the TDT through a user-friendly module of the Genetic Power Calculator (Purcell et al. 2003) web interface (<http://pngu.mgh.harvard.edu/~purcell/gpc/>). The user can specify different ascertainment strategies, family structures and disease models to best describe individual studies. Other accurate methods to

calculate the power of the TDT have been described (Knapp 1999; Chen and Deng 2001; Iles 2002) but, to our knowledge, ours is the first to account for the effects of background loci and family environment.

Acknowledgments MARF was funded by Sidney Sax fellowship 389927 from the National Health and Medical Research Council of Australia. SP and PS acknowledge the Medical Research Council grant G9901258 and National Eye Institute grant EY-12562. We thank three anonymous reviewers for their comments and suggestions on an earlier version of this manuscript.

References

- Abecasis GR, Cardon LR, Cookson WO (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279–292
- Antoniou AC, Easton DF (2003) Polygenic inheritance of breast cancer: implications for design of association studies. *Genet Epidemiol* 25:190–202
- Chen WM, Deng HW (2001) A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genet Epidemiol* 21:53–67
- Falconer DS (1981) *Introduction to quantitative genetics*. Longman, New York
- Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* 64:259–267
- Howson JM, Barratt BJ, Todd JA, Cordell HJ (2005) Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci. *Genet Epidemiol* 29:51–67
- Iles MM (2002) On calculating the power of a TDT study—comparison of methods. *Ann Hum Genet* 66:323–328
- Knapp M (1999) A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet* 64:1177–1185
- Li M, Boehnke M, Abecasis GR (2006) Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet* 78:778–792
- Purcell S, Cherny SS, Sham PC (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19:149–150
- Risch N (2001) Implications of multilocus inheritance for gene-disease association studies. *Theor Popul Biol* 60:215–220
- Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 59:983–989
- Whittaker JC, Lewis CM (1998) The effect of family structure on linkage tests using allelic association. *Am J Hum Genet* 63:889–897