Genet. Res., Camb. (2002), **79**, *pp.* 161–174. *With 2 figures.* © 2002 Cambridge University Press
DOI: 10.1017/S0016672302005578 Printed in the United Kingdom

The effects of selected sampling on the transmission disequilibrium test of a quantitative trait locus

HONG-WEN DENG[1, 2]* AND JING LI[2]

[1] *Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, ChangSha, Hunan 410081, P. R. China*

[2] *Osteoporosis Research Center and Department of Biomedical Sciences, Creighton University, Omaha, NE 68131, USA*

(*Received 16 July 2001 and in revised form 15 October 2001*)

Summary

We investigate how sampling of parents or children based on their extreme phenotypic values selected from clinical databases would affect the power of identification of quantitative trait loci (QTL) by a transmission disequilibrium test (TDT). We consider three selective sampling schemes based on the selection of phenotypic values of parents or children in nuclear families: (1) two children, one of extreme value, the other random; (2) two children extremely discordant; (3) one parent of extreme value. Other family members not specified will be recruited randomly with regard to phenotypic values. Our study shows that the second sampling scheme can always enhance the power for QTL identification, sometimes dramatically so. The increase in the statistical power of the TDT is particularly dramatic when h^2 at the QTL under test is small or intermediate (e.g. 0·05 or 0·10). For the other two sampling schemes, under dominant effects at the QTL, the power is always increased relative to random sampling; however, under recessive or additive genetic effects, the power gain is generally minor or even decreased a little sometimes. Allele frequencies at the QTL and the selection stringency are important for determining the effect of selective sampling on the power of QTL identification. Our study is useful as a practical guideline on how to perform the TDT efficiently in practice by taking advantage of the extensive databases accumulated that are enriched with people of extreme phenotypic values.

1. Introduction

Mapping and identification of genes underlying complex traits, especially those of primary health importance, has been a challenge for geneticists. The challenge is largely due to the limited power of and the large samples required by many currently employed approaches, such as traditional sib pair linkage studies (Risch & Merikangas, 1996). A powerful approach, the transmission disequilibrium test (TDT), has been developed for identification of genes, originally for complex diseases (Spielman *et al.*, 1993), and has recently been extended to quantitative traits (e.g. Allison, 1997; Rabinowitz, 1997; Xiong *et al.*, 1998). The TDT is used widely in practice (Schaid, 1998). In testing candidate genes for association with complex traits, the TDT is not plagued by the problem of population admixture/stratification and can test for linkage in the presence of evidence for association (Ewens & Spielman, 1995; Spielman & Ewens, 1996). When markers are at or very close to the genes of complex traits, the TDT is much more powerful than traditional sib pair linkage analyses (Risch & Merikangas, 1996; Allison, 1997; Xiong *et al.*, 1998). However, the sample size required may still be too large for application of the TDT in practice. This is especially true for identifying those genes with relatively small to intermediate effects. Therefore, it is of great practical importance to develop sampling schemes that can effectively reduce the sample sizes required for the TDT.

Selective sampling based on phenotypic values or disease status of family members (such as parents and/or children) may greatly enhance the power of traditional sib pair linkage analyses (Eaves & Meyer, 1994; Risch & Zhang, 1995; Zhang & Risch, 1996). For the TDT of disease genes, ascertainment of

* Corresponding author. Osteoporosis Research Center, Creighton University, 601 N. 30th Street, Suite 6787, Omaha, NE 68131, USA. Tel: +1 (402) 280 5911. Fax: +1 (402) 280 5034. e-mail: deng@creighton.edu

nuclear families with consideration of the affected status of parents will substantially increase the power (Whittaker & Lewis, 1998; Chen & Deng, 2001). For the TDT of QTL, the investigation (Allison, 1997; Xiong *et al*., 1998) was largely for randomly ascertained nuclear families without regard to the selection of phenotypic values of family members. Although Allison (1997) considered the sampling of extreme children for his TDT_{Q2}, TDT_{Q3} and TDT_{Q4}, sampling with regard to parental phenotypic values is not considered and these TDT tests require family trios consisting of one heterozygous parent, one homozygous parent and only one child. Xiong *et al*. (1998) developed a general TDT (TDT_G) for QTL identification. The TDT_G allows for more than one child per family and does not require only one parent to be heterozygous. With multiple children from each nuclear family, the power of the TDT_G is greatly increased (Xiong *et al*., 1998; Deng *et al*., 2001). However, for the TDT_G, the investigation of its statistical power is conducted for randomly ascertained nuclear families. The power of the TDT_G is unknown under various selective sampling schemes that can be based on extreme phenotypic values of parents or children.

For quantitative traits important for human health, generally only extreme (low/high) values are of primary clinical significance. For the past few decades, extensive records have been accumulated for people with extreme phenotypic values in clinics/clinical studies for many quantitative traits (such as blood pressure, bone mass and cholesterol level) (e.g. Deng *et al*., 1998*a*, *b*, 2000*b*, *c*). These extensive records of individuals with extreme phenotypic values may form convenient and powerful resources for recruitment of parents or children for nuclear families for the TDT_G analyses. Depending on the ages of the people in the records, these people may form probands as children or as parents for the nuclear families to form various selective sampling schemes. Therefore, it is important to investigate the effects of various selective sampling schemes on the power of the TDT_G. The investigation will provide a practical guideline on efficient implementation of the TDT_G (as an example of the TDT) by taking advantage of existing data for people with extreme phenotypes.

In this study we will investigate how selective sampling of parents or children based on their extreme values would affect the power of QTL identification by the TDT_G. The purpose is to provide a theoretical basis and a practical guideline to improve the power of the TDT_G. For demonstration, we will consider three situations based on selection of phenotypic values of parents or children in nuclear families each with two children: (1) one child is of an extreme phenotypic value, the other random; (2) two children are extremely discordant; (3) one parent is of an extreme value. Other family members not specified will be selected randomly with regard to their phenotypic values.

2. Methods

First, we will introduce the TDT_G of Xiong *et al*. (1998) for QTL identification. Then we will derive the non-centrality parameters (essential for our analytical computation) for the TDT_G statistic under the three selective sampling schemes. Since the TDT_G is a valid test of linkage in the presence of population admixture (Xiong *et al*., 1998), to demonstrate the effect of selective sampling in a relatively simple way, we will assume that the study population is randomly mating so that Hardy–Weinberg equilibrium holds.

(i) *The* TDT_G

We assume that there are n nuclear families with at least one parent being heterozygous for the marker locus under study. Such families are here termed *informative families*. Assume that there are two alleles M and m at the marker under test. For the ith ($i = 1,\ldots, n$) informative nuclear family that has n_i children, we assume that the marker allele M is transmitted to n_{Mi} children from heterozygous parent(s). Let Y denote the phenotypic value of the quantitative trait under study. For the jth child in the set of n_{Mi} children, let Y_{Mij} be his/her phenotypic value. We can denote n_{mi} and Y_{mil} similarly for the allele m. n_{Mi} and n_{mi} can be simply counted based on the genotypes of parents and children. The total numbers of children receiving M and m alleles from heterozygous parents are, respectively, $n_M = \sum_{i=1}^{n} n_{Mi}$ and $n_m = \sum_{i=1}^{n} n_{mi}$. Then the mean phenotypic values among children who receive M or m alleles from heterozygous parents are, respectively,

$$\bar{Y}_M = \frac{1}{n_M} \sum_{k=1}^{n} \sum_{j=1}^{n_{Mk}} Y_{Mkj}$$

and

$$\bar{Y}_m = \frac{1}{n_m} \sum_{k=1}^{n} \sum_{l=1}^{n_{mk}} Y_{mkl}.$$

The variance in the observations in children receiving the M allele is assumed to be the same as that in children receiving the m allele. Define

$$S^2 = \frac{\sum_{k=1}^{n} \sum_{j=1}^{n_k} ((Y_{Mkj} - \bar{Y}_M)^2 + (Y_{mkj} - \bar{Y}_m)^2)}{n_M + n_m - 2}.$$

Then the TDT_G statistic can be computed as

$$TDT_G = \frac{(\bar{Y}_M - \bar{Y}_m)^2}{\left(\frac{1}{n_M} + \frac{1}{n_m}\right) S^2}, \quad (1)$$

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

where $((1/n_M)+(1/n_m))S^2$ is an unbiased estimator of the variance of $\bar{Y}_M - \bar{Y}_m$ (Xiong *et al.*, 1998). With large sample sizes, the TDT_G approximately follows a χ^2 distribution with 1 d.f.

(ii) *Theory with selective sampling for the* TDT_G

In this paper we only present the results for the first situation when the marker is a functional mutation of the QTL under study. The second situation when the marker locus is not at a QTL but is linked to and is in linkage disequilibrium (LD) with a QTL has also been studied (Deng & Li, unpublished, results available on request) with the conclusions conforming to the first situation. We will not consider the effects of background polygenes separately from random environments on the TDT_G, as it has been demonstrated (Deng *et al.*, 2001) that the effects are minor with two children sampled from each family, a situation to be investigated here.

Define a QTL under study with two alleles Q and q. Let p and $p' = 1-p$ be the frequencies of the alleles Q and q, respectively. Let a ($>0{\cdot}0$) be the mean (genotypic value) for individuals of genotype QQ, let d be that of Qq individuals, and let $-a$ be that of qq individuals. d is equal to 0, a and $-a$, respectively under additive, dominant and recessive genetic effects. Under partial dominant or partial recessive genetic effects, $-a < d < a$ but $d \neq 0$. The additive genetic variance of this locus is $\sigma_A^2 = 2pp'[a+(p'-p)d]^2$, and the dominant genetic variance is $\sigma_D^2 = (2pp'd)^2$ (Falconer, 1989). The total genetic variance due to this QTL is $\sigma_G^2 = \sigma_A^2 + \sigma_D^2$. We assume that the variance due to all other QTLs and all random environmental effects is σ_e^2. The heritability h^2 due to this QTL is $h^2 = \sigma_G^2/(\sigma_G^2+\sigma_e^2)$. Under a genetic model (such as additive, dominant and recessive), once three of the four parameters of the h^2, a and p at the QTL and σ_e^2 are given, the fourth parameter can be computed easily (Falconer, 1989; Deng *et al.*, 2000*a*). The phenotypic value of an ith individual in the population is

$$y_i = \mu + G_i + e_i,$$

where μ is the mean baseline value of the quantitative trait under study, G_i is the genotypic value at the QTL for the ith genotype, and e_i represents a random variable for all random environmental effects. G_i is equal to a, d and $-a$ respectively for genotypes of QQ, Qq and qq. As in common practice, we can assume that $\mu = 0$, and e_i follows a normal distribution with mean 0 and variance σ_e^2. Let $F(x)$ be the cumulative distribution function (c.d.f.) of a normal random variable x.

Let μ_Q and σ_Q^2 be the mean and variance, respectively, of phenotypic values of the children who receive the Q allele from heterozygous parents; μ_q and σ_q^2 are similarly defined for the q allele. Let n_Q and n_q respectively be the numbers of children who receive the Q and q alleles from heterozygous parents. The noncentrality parameter of the distribution of the statistic TDT_G is (Xiong, 1998)

$$\lambda = \frac{(\mu_Q - \mu_q)^2}{(\sigma_Q^2/n_Q + \sigma_q^2/n_q)}. \tag{2}$$

To compute analytically the statistical power of the TDT_G, λ and thus μ_Q, σ_Q^2, μ_q, σ_q^2, n_Q and n_q should be derived in terms of the parameters such as p, p', genetic effects (such as a and d) under various selective sampling schemes. Let g_o, g_f and g_m, respectively, denote the genotypes of children, fathers and mothers in informative nuclear families for the TDT_G. Then within a nuclear family, conditional on the parental genotypes of g_f and g_m, the mean value of all children is

$$\mu_1 = E(Y \mid g_f, g_m) = \sum_{g_o} E(Y \mid g_o, g_f, g_m) P(g_o \mid g_f, g_m), \tag{3a}$$

where P denotes probability throughout. Over all the informative nuclear families, the mean value of all the children is

$$\begin{aligned}\mu_2 &= \sum_{g_f}\sum_{g_m} P(g_f, g_m) * \sum_{g_o} E(Y \mid g_o, g_f, g_m) P(g_o \mid g_f, g_m) \\ &= \sum_{g_f}\sum_{g_m}\sum_{g_o} E(Y \mid g_o, g_f, g_m) P(g_o, g_f, g_m).\end{aligned} \tag{3b}$$

To focus on the main idea and its significance, we will only outline our analytical derivation and the results in the following. The tedious technical details are available from the authors on request.

One child has an extremely low value and falls below the bottom ϕ per cent of the phenotypic distribution

Let Qq_p denote the event that at least one parent is heterozygous, C_l denote the event that at least one child of the two in each nuclear family recruited has an extremely low value, and Q_o denote the event that a heterozygous parent transmits the allele Q to an offspring. The subscripts 'p' and 'o' denote respectively the parental and offspring generations. By the same derivation principle as in Equation 3, conditional on Qq_p, Q_o and C_l in informative nuclear families, we have

$$\begin{aligned}\mu_Q = \sum_{g_f}\sum_{g_m}\sum_{g_o} & E(Y \mid g_o, g_f, g_m, Qq_p, Q_o, C_l) \\ & \times P(g_o, g_f, g_m \mid Qq_p, Q_o, C_l),\end{aligned} \tag{4a}$$

$$\begin{aligned}\sigma_Q^2 = \sum_{g_f}\sum_{g_m}\sum_{g_o} & E(Y^2 \mid g_o, g_f, g_m, Qq_p, Q_o, C_l) \\ & \times P(g_o, g_f, g_m \mid Qq_p, Q_o, C_l) - \mu_Q^2.\end{aligned} \tag{4b}$$

To derive μ_Q, we consider two mutually exclusive situations. First, the child who receives the allele Q

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

has an extremely low phenotypic value ($Y \leqslant Z_L$; the event is denoted by C_{Yl}) and the other child's phenotype is randomly selected. Second, the child whose phenotype is being considered (here it is the child who receives the allele Q) does not have an extremely low phenotypic value ($Y \geqslant Z_L$; the event is denoted by C_{Yn}) and the other child has an extremely low phenotypic value (the event is denoted by C'_l). The second situation can be denoted as the joint events of $C_{Yn}C'_l$. It can be seen that

$$P(C_l) = P(C_{Yl}) + P(C_{Yn}C'_l).$$

Hence, we have

$$\begin{aligned} E(Y|g_o, g_f, g_m, Qq_p, Q_o, C_l) &= E(Y|C_{Yl}, g_o, Q_o) \\ &\times \frac{P(C_{Yl}|g_o, Q_o)}{(P(C_{Yl}|g_o, Q_o) + P(C_{Yn}|g_oQ_o)P(C'_l|g_f, g_m, Qq_p))} \\ &+ E(Y|C_{Yn}C'_l, g_o, Q_o) \\ &\times \frac{P(C_{Yn}C'_l|g_o, g_f, g_m, Qq_p, Q_o)}{(P(C_{Yl}|g_o, Q_o) + P(C_{Yn}|g_oQ_o)P(C'_l|g_f, g_m, Qq_p))}. \end{aligned} \tag{5}$$

The threshold phenotypic value Z_L for the bottom ϕ per cent of the phenotypic distribution can be computed by

$$\begin{aligned} \phi\% &= \Pr(Y \leqslant Z_L) \\ &= \Pr(Y \leqslant Z_L|QQ)P_{QQ} \\ &+ \Pr(Y \leqslant Z_L|Qq)P_{Qq} + \Pr(Y \leqslant Z_L|qq)P_{qq}. \end{aligned} \tag{6}$$

With Z_L known, the terms in Equation 5 can be expressed, respectively, as

$$\begin{aligned} E(Y|C_{Yl}, g_o, Q_o) &= E(Y|C_{Yl}, g_{o.Q}) \\ &= \frac{\int_{z_L}^{\infty} x^* f(x, \mu_{g_{o.Q}}, \sigma_e^2)dx}{F(Z_L, \mu_{g_{o.Q}}, \sigma_e^2)}, \end{aligned} \tag{7a}$$

$$\begin{aligned} E(Y|C_{Yn}, g_o, Q_o) &= E(Y|C_{Yn}, g_{o.Q}) \\ &= \frac{\int_{z_L}^{\infty} x^* f(x, \mu_{g_{o.Q}}, \sigma_e^2)dx}{1 - F(Z_L, \mu_{g_{o.Q}}, \sigma_e^2)}, \end{aligned} \tag{7b}$$

$$P(C_{Yl}|g_o, Q_o) = P(C_{Yl}|g_{o.Q}) = F(Z_L, \mu_{g_{o.Q}}, \sigma_e^2), \tag{7c}$$

$$\begin{aligned} P(C_{Yn}|g_o, Q_o) &= P(C_{Yn}|g_{o.Q}) \\ &= 1 - F(Z_L, \mu_{g_{o.Q}}, \sigma_e^2), \end{aligned} \tag{7d}$$

where $g_{o.Q}$ is the genotype of the child who receives the Q allele from a heterozygous parent. Hence, $g_{o.Q}$ can only be one of the two genotypes, QQ and Qq. $\mu_{g_{o.Q}}$ is the genotypic value of the genotype QQ or Qq and is a for the genotype QQ and d for Qq, respectively.

Given that at least one parent is heterozygous in a nuclear family, the event $(g_f, g_m, Qq_p) = (QQ_f, Qq_m) \cup (QQ_m, Qq_f) \cup (Qq_f, Qq_m)$, where 'U' denotes a union in probability and the subscripts 'f' and 'm' denote the father and the mother, respectively. Conditional on the genotypes of parents, $P(C'_l|g_f, g_m, Qq_p)$, the probability that the child's phenotypic value $Y \leqslant Z_L$ (denoted by 'l' in the subscript), can be computed. For example, if the parents are of the genotypes QQ and Qq, we can have:

$$P(C'_l|g_f, g_m, Qq_p) = 0{\cdot}5(F(Z_L, a, \sigma_e^2) + F(Z_L, d, \sigma_e^2)).$$

In Equation 4a,

$$P(g_o, g_f, g_m|Qq_p, Q_o, C_l) = \frac{P(g_o, g_f, g_m, Qq_p, Q_o, C_{Yl}) + P(g_o, g_f, g_m, Qq_p, Q_o, C_{Yn}C'_l)}{\sum_{g_f}\sum_{g_m}\sum_{g_o} P(g_o, g_f, g_m, Qq_p, Q_o, C_l)}. \tag{8}$$

Given genotypes of parents and the child who receives a Q allele from a heterozygous parent, the probabilities in the numerator of Equation 8 can be computed easily. For example, if the parents and the child have the genotypes qq, Qq and Qq, respectively, we have

$$P(g_o, g_f, g_m, Qq_p, Q_o, C_{Yl}) = pp'^3 F(Z_L, d, \sigma_e^2),$$

and

$$\begin{aligned} &P(g_o, g_f, g_m, Qq_p, Q_o, C_{Yn}C'_l) \\ &= pp'^3(1 - F(Z_L, d, \sigma_e^2))\sum_{g_o} P(C'_l|g_o)P(g_o|qq_f, Qq_m). \end{aligned}$$

With Equations 4a, 5, 7a–d and with the procedures outlined above for Equation 8, we can compute μ_Q analytically by the following equation:

$$\mu_Q = \sum_{g_f}\sum_{g_m}\sum_{g_{o.Q}} \left(\frac{\int_{-\infty}^{z_L} xf(x, \mu_{g_o}, \sigma_e^2)dx + \int_{z_L}^{\infty} xf(x, \mu_{g_o}, \sigma_e^2)dx^* P(C'_l|g_f, g_m, Qq_p)}{F(Z_L, \mu_{g_o}, \sigma_e^2) + (1 - F(Z_L, \mu_{g_o}, \sigma_e^2))P(C'_l|g_f, g_m, Qq_p)} \right) P(g_o, g_f, g_m|Qq_p, Q_o, C_l). \tag{9}$$

To derive σ_Q^2 in Equation 4b, we have

$$\begin{aligned} &E(Y^2|g_o, g_f, g_m, Qq_p, Q_o, C_l) \\ &= E(Y^2|C_{Yl}, g_o, g_f, g_m, Qq_p, Q_o)\frac{P(C_{Yl}|g_o, g_f, g_m, Qq_p, Q_o)}{(P(C_{Yl}|g_o, g_f, g_m, Qq_p, Q_o) + P(C_{Yn}C'_l|g_o, g_f, g_m, Qq_p, Q_o))} \\ &+ E(Y^2|C_{Yn}C'_l, g_o, g_f, g_m, Qq_p, Q_o)\frac{P(C_{Yn}C'_l|g_o, g_f, g_m, Qq_p, Q_o)}{(P(C_{Yl}|g_o, g_f, g_m, Qq_p, Q_o) + P(C_{Yn}C'_l|g_o, g_f, g_m, Qq_p, Q_o))}, \end{aligned} \tag{10}$$

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

where

$$E(Y^2 \mid C_{Yl}, g_o, g_f, g_m, Qq_p, Q_o) = E(Y^2 \mid C_{Yl}, g_{o.Q}) = \frac{\int_{-\infty}^{z_L} x^2 f(x, g_{o.Q}, \sigma_e^2)dx}{F(Z_L, g_{o.Q}, \sigma_e^2)}, \quad (11a)$$

$$E(Y^2 \mid C_{Yn}C_l', g_o, g_f, g_m, Qq_p, Q_o) = E(Y^2 \mid C_{Yn}, g_{o.Q}) = \frac{\int_{z_L}^{\infty} x^2 f(x, g_{o.Q}, \sigma_e^2)dx}{1-F(Z_L, g_{o.Q}, \sigma_e^2)}. \quad (11b)$$

By Equations 4b, 5, 7c, 7d and 10–11, we have

$$\sigma_Q^2 = \sum_{g_f}\sum_{g_m}\sum_{g_{o.Q}} \left(\frac{\int_{-\infty}^{z_L} x^2 f(x, \mu_{g_o}, \sigma_e^2)dx + \int_{z_L}^{\infty} x^2 f(x, \mu_{g_o}, \sigma_e^2)dx * P(C_l' \mid g_f, g_m, Qq_p)}{F(Z_L, \mu_{g_o}, \sigma_e^2) + (1-F(Z_L, \mu_{g_o}, \sigma_e^2))P(C_l' \mid g_f, g_m, Qq_p)} \right) * P(g_o, g_f, g_m \mid Qq_p, Q_o, C_l). \quad (12)$$

Similarly, we can derive the expression for μ_q and σ_q^2 as above for μ_Q and σ_Q^2.

Finally, n_Q and n_q need to be derived, in order to compute λ analytically. Let Ns be the total number of the screened families to obtain n informative families under the selective sampling scheme under consideration. Let n_H be the total number of heterozygous parents in the sample. We have

$$n = Ns * P(Qq_p, C_l) = Ns * [P(Qq_p, Q_o, C_l) + P(Qq_p, q_o, C_l)], \quad (13)$$

where $P(Qq_p, Q_o, C_l)$ is the probability that at least one parent is heterozygous (the event is denoted by Qq_p) and one child is of an extremely low value ($Y \leqslant Z_L$; the event is denoted by C_l), and the heterozygous parent transmits the allele Q to a child.

$$n_H = n + n^* \frac{P(Qq_f, Qq_m, C_l)}{P(Qq_p, C_l)}, \quad (14)$$

where

$$P(Qq_f, Qq_m, C_l) = P(Qq_f, Qq_m)P(C_{Yl} \mid Qq_f, Qq_m) + P(Qq_f, Qq_m)P(C_{Yn}C_l' \mid Qq_f, Qq_m) = 4p^2p'^2 \left(\begin{array}{l} \sum_{g_o} P(C_{Yl} \mid g_o)P(g_o \mid Qq_f, Qq_m) \\ + \sum_{g_o} P(C_{Yn} \mid g_o)P(g_o \mid Qq_f, Qq_m) \sum_{g_o} P(C_l' \mid g_o)P(g_o \mid Qq_f, Qq_m) \end{array} \right). \quad (15)$$

Assuming that each nuclear family has J children (in this study, $J = 2$), we have

$$n_Q = Jn_H \frac{P(Qq_p, Q_o, C_l)}{P(Qq_p, Q_o, C_l) + P(Qq_p, q_o, C_l)}, \quad (16a)$$

$$n_q = Jn_H \frac{P(Qq_p, q_o, C_l)}{P(Qq_p, Q_o, C_l) + P(Qq_p, q_o, C_l)}. \quad (16b)$$

Specifying a significance level (α) and a statistical power (η), we can, with the aid of a suitable statistical software package (e.g. Wolfram, 1996), obtain the value for the non-centrality parameter λ for the TDT_G statistic. With the λ value and Equations 2, 9, 13–16, we can compute the required sample sizes Ns, n and n_H for specified α and η given parameter values (p, p', a, d, σ_e^2 and ϕ) under the first selective sampling scheme that each recruited family has two children and one belongs to the bottom ϕ per cent of the phenotypic distribution.

One child's phenotype belongs to the bottom ϕ per cent and the other to the top ρ per cent of the population distribution

One parent's phenotype falls into bottom ϕ per cent of the distribution

The analytical derivations for these two selective sampling schemes are similar. The keys are to derive the μ_Q, σ_Q^2, μ_q, σ_q^2, n_Q and n_q under a specific selective sampling scheme, the results for which are given in the Appendix. With these results, the statistical power can be obtained as outlined above.

(iii) *Computer simulations*

To validate the above derivations and analytical power computation, we perform computer simulations. The validation of the power computation that is based on the complex analytical derivation by simulations is necessary; this is also true given the approximation of the test statistics to a χ^2 distribution. The comparison of simulation and analytical results can provide a mechanism to crosscheck and validate the results from the two approaches. In the absence of segregation distortion, parents of nuclear families from random mating populations are simulated, in which the p, p', a, d and h^2 at the QTL and ϕ and/or ρ and σ_e^2 are specified. Only for nuclear families with at least one parent heterozygous at the marker locus are the parents' phenotypes simulated. The simulation for phenotypes based on genotypes and other

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

parameters is standard and has been documented elsewhere (e.g. Deng *et al*., 2000*a*). For the first two selective sampling schemes, two children are simulated for their genotypes and phenotypes. The genotypes of children are simulated according to random transmission of alleles from parents to children. Once the genotypes of children are simulated, their phenotypes are simulated. Those nuclear families in which the children meet the selection criterion for a specific selective sampling scheme are retained for analyses. For the third sampling scheme, only when one parent's phenotype is in the bottom ϕ per cent are the children's genotypes and phenotypes simulated. For comparison of the statistical powers under random sampling and selective sampling, nuclear families with two children are also simulated without regard to the phenotypes of family members – a situation that has been focally investigated previously (Xiong *et al*., 1998). Once the informative families are simulated for a specific selective sampling scheme or random sampling, the TDT_G analyses (Equation 1) are performed.

For a desired statistical power η and a specified significance level α and for a specific sampling scheme, we first compute the sample size (n) of informative nuclear families needed by our analytical power computation method. The analytical power computation for random sampling can be implemented by, for example, specifying $\phi = 100$ in the first sampling scheme. Then informative nuclear families each with two children are simulated for the specific selective sampling scheme or random sampling. The TDT_G is applied to the n nuclear families. When a QTL is simulated, the simulated statistical power is the proportion of times that the TDT_G analyses are significant in a number of simulations (10000 times unless otherwise specified) performed. The statistical power (η') obtained in simulations under the significance level α can be compared with the specified level of η in the analytical power computation. Once our analytical power computation for the TDT_G is validated by computer simulations, the investigation of the power of the TDT_G under various sampling schemes for different other parameter values is conducted by our analytical method. To validate the TDT_G under the various selective sampling schemes considered, under a specified α, we also examine the size (the type I error rate) in simulations (α') with a marker locus that is not linked to and/or is not in linkage disequilibrium with any QTL.

3. Results

(i) *The accuracy of our analytical power computation*

Table 1 presents some representative data of our extensive simulation studies for a range of parameter values with different genetic models under the three selective sampling schemes and random sampling. It can be seen that, for all the three typical models of genetic effects at the QTL (recessive, additive and dominant), the sample sizes (n) computed from our analytical method under a specified statistical power (η), if employed in computer simulations, can yield the simulated statistical power (η') that is very close to the η. This is true for different sampling schemes considered. Therefore, our analytical derivation and the power computation for the TDT_G under various sampling schemes considered are validated by our computer simulations.

Table 1. *The accuracy of our analytical power computation and the validity of the* TDT_G *under selective sampling*

Sampling scheme	Genetic effect	n (η')	α' ($\alpha = 0{\cdot}05$)
One child $\in$ B10 %	Recessive	175 (0·85)	0·053
	Additive	113 (0·85)	0·047
	Dominant	109 (0·83)	0·047
One child $\in$ B10 %, the other child $\in$ T30 %	Recessive	72 (0·75)	0·046
	Additive	55 (0·80)	0·055
	Dominant	88 (0·81)	0·055
One parent $\in$ B10 %	Recessive	191 (0·79)	0·049
	Additive	137 (0·79)	0·045
	Dominant	187 (0·79)	0·053
Random sampling	Recessive	170 (0·81)	0·052
	Additive	141 (0·81)	0·045
	Dominant	331 (0·78)	0·047

n is the number of informative families needed in a specific sampling scheme in order to achieve 80 % power (η) with $\alpha = 10^{-4}$ computed by our analytical approach and η' is the power obtained by 10000 repeated simulations with the sample size n. In the studies for this table, $p = 0{\cdot}7$, $h^2 = 0{\cdot}1$, and $a = 1$. α' is the empirical size (type I error rate) for the TDT_G test obtained from 10000 repeated simulations when the marker is not a QTL or is not linked to a QTL, or is not in linkage disequilibrium with a QTL. It is the proportion of the times that the TDT_G analysis is not significant under the specified significance level of α ($= 0{\cdot}05$). In validating the TDT_G, the significance level of $\alpha = 0{\cdot}05$ is chosen to avoid unnecessary excessive simulations for the α at much lower levels such as $\alpha = 10^{-4}$. '$\in$' denotes 'belongs to'; 'B10 %' denotes bottom 10 % and 'T30 %' denotes top 30 % of the phenotype distribution.

(ii) *The validity of the* TDT_G *under selective sampling*

The last column of Table 1 presents the results of the simulated significance level α' under the null hypothesis that the marker locus is not linked to and/or is not in linkage disequilibrium with a QTL. It can be seen that, for various genetic effects at the QTL and under all the sampling schemes investigated, the simulated significance level is essentially equal to

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

Table 2. *Comparison of the* TDT_G *and sib pair linkage tests under selective sampling*

		T10%B10%	
Genetic effect	p	RZ	TDT_G
Recessive	0·1	19984 (205218)	120 (13897)
	0·5	1565 (17769)	64 (3307)
Additive	0·1	1647 (19120)	18 (2211)
	0·5	1464 (17166)	44 (2325)
Dominant	0·1	1567 (18153)	18 (2214)
	0·5	1565 (17769)	64 (3307)

'T10%B10%' denotes the sampling scheme where one sib belongs to the bottom 10% and the other to the top 10% of the phenotypic distribution. 'RZ' denotes the linkage test with extremely discordant sib pairs by the method of Risch & Zhang (1995). The numbers in the RZ column are the numbers of extremely discordant sib pairs (EDSP) and the associated numbers of sib pairs needed to be screened to obtain the EDSP required by the linkage test of Risch & Zhang (1995) under the power of 80% and significance level 0·0001. These numbers were extracted from Risch & Zhang (1996). The data for the TDT_G are the required numbers of informative nuclear families with EDSP (n) and the associated number of nuclear families that need to be screened (Ns) in order to achieve the required power (80%) under the specified significance (0·0001) with the TDT_G under the selective sampling scheme. These were obtained by our analytical power computation and confirmed by computer simulations. In the investigation for this table, $h^2 = 0{\cdot}1$, as in Risch & Zhang (1996) and Zhang & Risch (1996), σ_e^2 is set to 1, and a can be determined by the genetic effects, p, h^2 at the QTL and σ_e^2.

the specified significance level $\alpha = 0{\cdot}05$, except for the minor differences caused by sampling. Therefore, the TDT_G is valid and robust with selective sampling in that the significance level achieved in practice is the about same as that specified in the TDT_G testing.

(iii) *Comparison of the* TDT_G *and a linkage test for extremely discordant sib pairs*

The TDT_G is much more powerful than sib pair linkage analyses in the absence of selective sampling (Xiong *et al*., 1998). Selectively sampling can dramatically increase the power of sib pair linkage analyses (e.g. Risch & Zhang, 1995, 1996). Therefore, it is of interest to compare the powers of the TDT_G and sib pair linkage analyses when selective sampling is involved for both of the two approaches that may employ the same type of sampled families. Here we will use nuclear families with extremely discordant sib pairs as an example for comparison. Note that although extreme sampling for children is considered in Allison (1997), Allison's tests (TDT_{Q1-4}) originally only allow for nuclear family trios consisting of one heterozygous parent and one child, and thus have a restrictive limitation in practical application. Therefore, we will only consider the comparison of the TDT_G with sib pair linkage analysis of Risch & Zhang (1995, 1996) for extremely discordant sib pairs. It is apparent from Table 2 that with extremely discordant sib pairs, the TDT_G is much more powerful than the sib pair linkage analyses of Risch & Zhang (1995, 1996) in that the required sample sizes (for testing or for screening) are much smaller across various parameters and genetic models.

(iv) *The effects of selective sampling under various sampling schemes*

Compared with the random sampling scheme, the three selective sampling schemes all increase the power of the TDT_G under dominant effects for the allele (Q) that increase phenotypic values (plot *c* in Fig. 1; plots *c*, *f* and *i* in Fig. 2) in that the numbers of required informative nuclear families are fewer. Under recessive and additive genetic effects for allele Q (plots *a* and *b* in Fig. 1; plots *a*, *b*, *d*, *e*, *g* and *h* in Fig. 2), the selective sampling considered may or may not increase the power compared with random sampling. Extreme sampling involving discordant sib pairs always increases the statistical power. The extreme sampling involving only one extreme child or one extreme parent considered may have little effect on the power (plots *a* and *b* in Fig. 1; plots *b* and *e* in Fig. 2) or even suffer some minor loss of power (plot *a* in Fig. 1; plots *a* and *d* in Fig. 2). However, simply due to the symmetry of recessive and dominant genetic effects of alleles Q and q, when extreme sampling through one end of the phenotypic distribution does not increase the power of QTL identification, extreme sampling through the other end of the distribution for parents or children will. This is confirmed by our computer simulations. To be more specific, if under recessive genetic effects of allele Q, extreme sampling through one parent or one child with extremely low phenotypic value does not increase the power, extreme sampling through one parent or one child with extremely high phenotypic values does increase the power substantially. This is simply because when allele Q is recessive, allele q has dominant genetic effects. Therefore, except under additive genetic effect, appropriate selective sampling can generally increase the power of QTL identification.

Of particular interest is that with selective sampling, when it can increase the power, the increase in the TDT_G power is most dramatic when the h^2 at the QTL under study is relative small or intermediate (e.g. 0·05 or 0·10) (plots *a–c* in Fig. 1; plots *c*, *f–i* in Fig. 2). Compared with random sampling, when it can increase the power, the effect of selective sampling decreases with an increasing h^2 at the QTL under study (plots

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

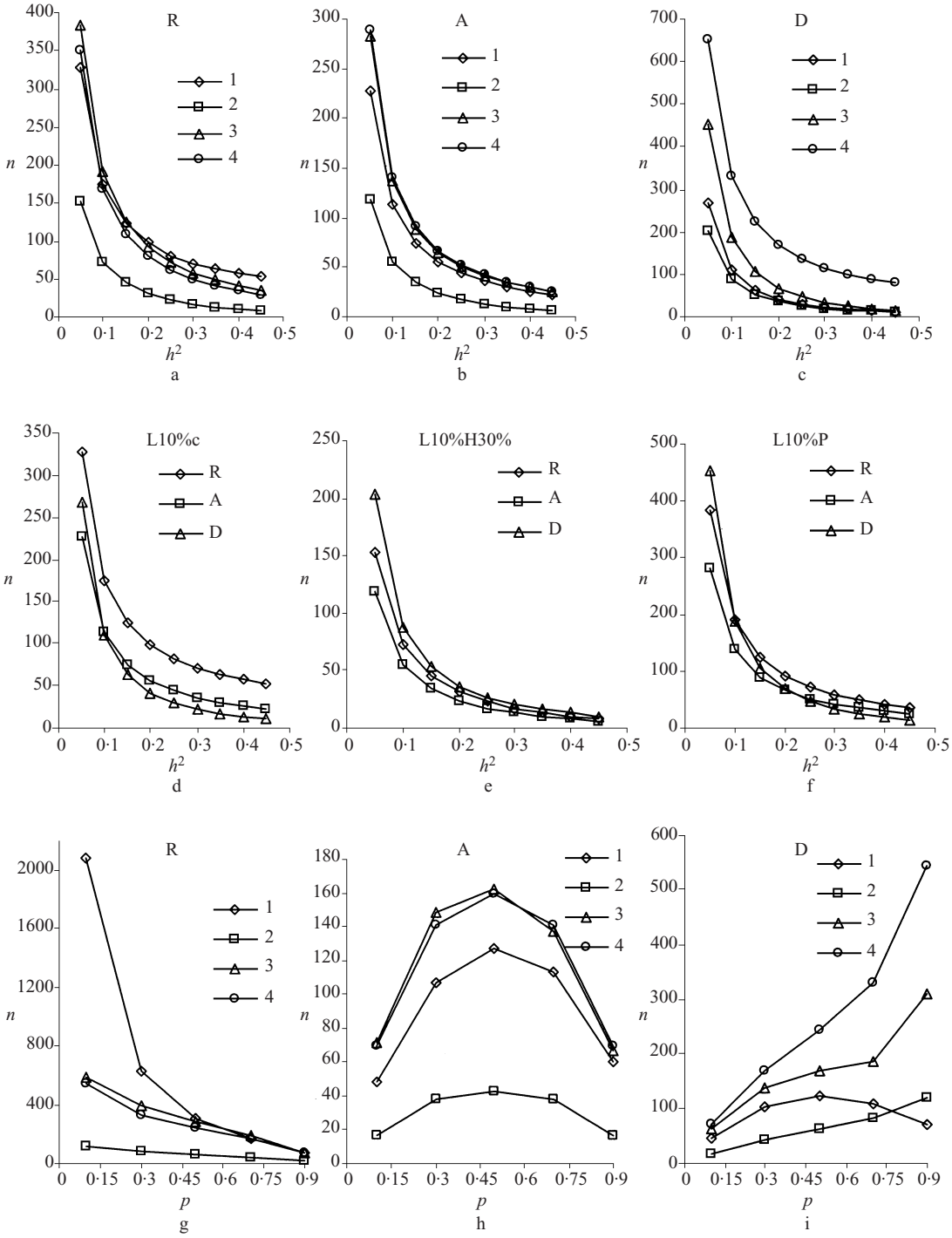

Fig. 1. Comparison of the numbers (n) of informative families required to achieve 80% power under various sampling schemes for nuclear families with two children under different heritabilities, genetic effects and the frequencies (p) of the allele Q at the QTL. The marker is at the QTL, $p = 0{\cdot}7$ and $a = 1$; various values of h^2 are achieved by varying the magnitude of σ_e^2. $\alpha = 10^{-4}$. In the first sampling scheme, one child $\in$B10%; in the second sampling scheme, one child $\in$B10% and the other child $\in$T30%; in the third sampling scheme, one parent $\in$B10%; in the fourth sampling scheme, all family members are randomly recruited. The phenotypes of other unmentioned members of the nuclear families are random. In Fig. 1, '1', '2', '3', '4' denote the first, second, third and fourth sampling schemes respectively. In Figs. 1 and 2, 'A', 'D' and 'R' denote additive, dominant and recessive genetic effects, respectively, at the QTL. Plots *a–c* present comparisons of the selective sample schemes of 1–3 with random sampling for recessive, additive and dominant genetic models, respectively. Plots *d–f* present the effects of the first, second and third sampling schemes under various genetic models at the QTL. Plots *g–i* compare various sampling schemes under various frequencies (p) of the allele Q and the QTL.

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

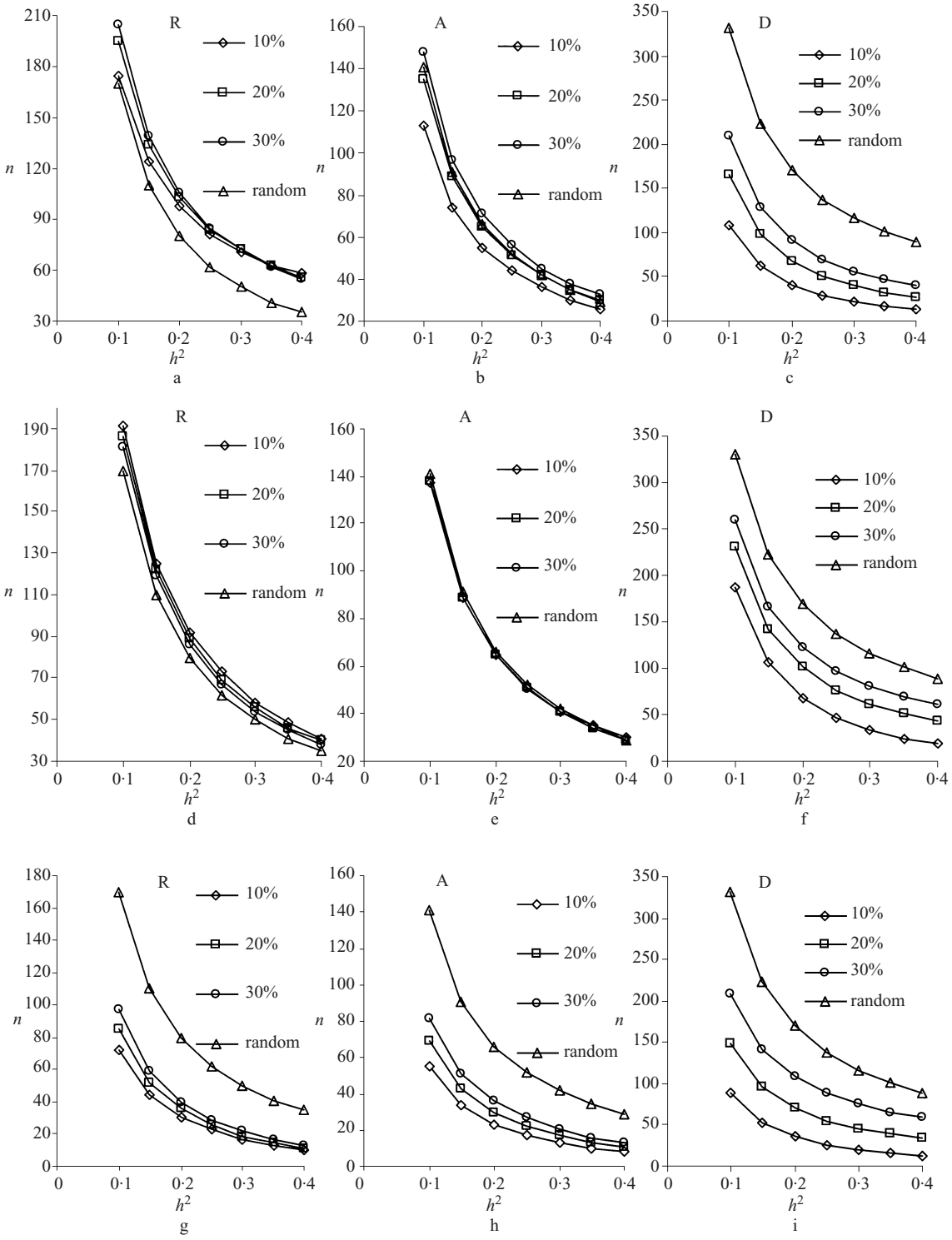

Fig. 2. Comparison of n required to achieve 80% power for various sampling schemes under different selective sampling stringency when the marker under test is a QTL. The selective stringency of sampling is measured by ϕ so that individuals with extreme phenotypes are selected from the bottom ϕ% of the phenotypic distribution. In each plot, the parameters given indicate the ϕ per cent chosen; 'random' indicates random sampling without considering extreme phenotypic values. In the selective sampling scheme in plots *a–c*, one child $\in$Bϕ per cent is selected as proband for the recruitment of nuclear families each with two children. In the selective sampling scheme in plots *d–f*, one child $\in$Bϕ per cent and the other child $\in$T30% are selected as probands for the recruitment. In the selective sampling scheme in plots *g–i*, one parent $\in$B10% is selected as proband for the recruitment. Unless otherwise specified, $p = 0{\cdot}7$, $a = 1$ and $h^2 = 0{\cdot}1$.

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

a–c in Fig. 1; plots *c*, *f–i* in Fig. 2). Among various sampling schemes, extremely discordant sib pairs are consistently the most powerful samples across different genetic models (plots *a–c* in Fig. 1).

For the same selective sampling scheme, the power depends on the mode (recessive, additive or dominant) of the genetic effect at the QTL under study (plots *d–f* in Fig. 1). Other things being equal, the power of the TDT_G depends on the allele Q frequency (p) at the QTL (plots *g–i* in Fig. 1). With recessive genetic effects, the power is relatively low when p is small and increases with an increasing p in that the required number of informative families (n) decreases with an increasing p. The trend of the power with p is opposite under dominant genetic effects at the QTL, which is expected. Under additive genetic effect, the power is the smallest (as reflected by the largest n) with intermediate allele frequency p and increases with p approaching 0 or 1. The sample sizes here are *n's* (the number of informative families required under a specific sampling scheme), not N_s (the number of families that need to be screened in order to recruit n informative families) as reported in tables 1–3 in Xiong *et al*. (1998). As is apparent, when it can increase the power (plots *c*, *f–i* in Fig. 2), the sample size n required decreases (and thus the power increases) with an increasing stringency of subject selection (plots *d–f* in Fig. 2) for the parameter values investigated when compared with random sampling.

4. Discussion

In this study we investigated how the sampling of parents or children based on their extreme phenotypic values selected from clinical databases would affect the power of QTL identification by the TDT_G. We considered three selective sampling schemes based on the selection of phenotypic values of parents or children in nuclear families: (1) one child is of extreme value, the other random; (2) two children are extremely discordant; (3) one parent is of extreme value. Our study shows that the second sampling scheme can always enhance the power for QTL identification, sometimes dramatically so. The increase in statistical power of the TDT is particularly dramatic when h^2 at the QTL under test is small or intermediate (e.g. 0·05 or 0·10). For the other two sampling schemes, except under additive genetic effects, the power may generally increase under selective sampling from the appropriate end of the phenotypic distribution. Allele frequencies at the QTL are important for determining the effect of selective sampling on power of QTL identification. Therefore, clinical records of extreme individuals may form powerful resources for QTL identification by the TDT. Our study should be useful for performing the TDT_G efficiently in practice by taking advantage of extensively accumulated data that are enriched with people of extreme phenotypic values. As a demonstration of the effects of selective sampling on the TDT, we investigated nuclear families each with two children (sib pairs). If there are more than two children in nuclear families, the phenotypic values of the other sibs may play a role in determining the TDT power, as has already been shown for disease phenotypes (Chen & Deng, 2001).

Extensive records have been accumulated in clinics/clinical studies for those individuals with extreme phenotypes that are of health and clinical significance. How to employ these records efficiently in gene identification for complex traits has been under extensive investigation for linkage analyses (e.g. Eaves & Meyer, 1994; Risch & Zhang, 1995; Zhang & Risch, 1996) and recently for TDT analyses for disease gene identification (Whittaker & Lewis, 1998; Chen & Deng, 2001). It has been shown that the TDT may have much higher power than linkage analyses (Risch & Merikangas, 1996; Allison, 1997; Xiong *et al*., 1998) for testing the significance of particular genes or genomic locations for complex traits. In addition, the TDT analyses of QTL may have much higher power with the TDT_G in nuclear families with more than one heterozygous parent and multiple children (Xiong *et al*., 1998; Deng *et al*., 2001). Therefore, our work in investigating the effect of selective sampling on the power of Xiong *et al*.'s (1998) TDT_G should be important.

As is shown, with selective sampling the TDT_G is still valid in that it ensures the type I error rate suffered in practice is the same as the specified significance level in the TDT_G testing. Importantly, it is shown that under the three selective sampling schemes investigated for parent or child(ren), the power of the TDT_G may be increased and sometimes dramatically so for the QTL with relatively small h^2. This is of particular significance given that the greatest challenge in QTL identification is the limited power of all the current approaches in identifying QTLs with small to intermediate h^2 values. It is shown that when the marker is a QTL, the required sample size n for the TDT_G may be orders of magnitude smaller than that for the sib pair linkage analyses for nuclear families with extremely discordant sib pairs (Table 2). Therefore, application of the TDT_G with appropriate selective sampling whenever possible is warranted in practice for its validity and its large power. However, the detailed sampling scheme to be adopted may depend on practical issues at hand and on the availability of extreme individuals and their ages (that determine whether they can be recruited as parents or children for studying nuclear families) and the information on the likely genetic effects of the QTL. While nuclear families with discordant sib pairs may be consistently of the greatest power for the TDT_G

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

analyses and always yield higher power than random sampling, their recruitment may also be most difficult, even with the availability of clinic records on extreme individuals. However, selective sampling for only one parent or only one child should be simple and convenient with the aid of the accumulated clinic records on extreme individuals and it may greatly enhance the power except under additive genetic effect of the QTL. Even under additive genetic effect of the QTL when power is not gained significantly or even may suffer some minor loss under selective sampling for only one parent or only one child, the power is generally not changed much. Hence, in the absence of knowledge of the genetic effects at the QTL, selective sampling for only one parent or only one child may still be attempted with the accumulated clinical records. When the QTL allele increasing the trait value is dominant, extreme sampling of only one parent or only one child through low phenotypic value may greatly increase the power. When the QTL allele increasing the trait value is recessive, extreme sampling of only one parent or only one child through high phenotypic value may greatly increase the power. Hence, which selective sampling strategy to adopt should depend on the individual investigator's resources and information about the genetic effects of the QTL being studied.

The mechanisms by which selective sampling may increase the power of the TDT_G are as follows: (1) Selective sampling may increase the number of heterozygous parents in informative nuclear families as revealed in our simulations. (2) Selective sampling may increase the difference of the mean phenotypic values of children who receive the Q and the q alleles respectively and may also decrease the phenotypic variances of the children who receive the Q and q alleles respectively. This will then increase the value of the TDT_G statistic computed and thus increase the power of the TDT_G. It should be noted that with the TDT_G, not every kind of selective sampling scheme can increase the power. For example, we (Deng & Li, unpublished) investigated the power of selective sampling when two sibs are extremely concordant for trait values – a selective sampling strategy that has been shown to increase the power of sib pair linkage analyses (Risch & Zhang, 1995; Zhang & Risch, 1996, 1997). We found that concordant sib pairs are generally not powerful samples for the TDT_G analyses. This should not come as a surprise since both sibs are selected to be extremely concordant. Their phenotypic values are selected to be so concordant that the difference in the mean phenotypic values of the children who receive the Q and q alleles respectively is selected to be small. This reduces the difference in the numerator for computing the TDT_G statistic (Equation 1), thus diminishing the power of the TDT_G. However, this does not mean that extremely concordant sib pairs are not useful for TDT analyses of QTLs. A TDT statistic that is different from the TDT_G but that can be applied to informative nuclear families with more than one child may utilize nuclear families with extremely concordant sib pairs efficiently. Such a TDT test statistic is yet to be constructed and investigated.

H.-W. D. was partially supported by grants from the NIH, Health Future Foundation, US Department of Energy, State of Nebraska, Creighton University, National Science Foundation of China, HuNan Normal University and Ministry of Education of China. We are grateful to Professor William G. Hill and two anonymous reviewers for their constructive comments that helped to improve the manuscript.

Appendix. The analytical results for two selective sampling schemes

One child's phenotype belongs to the bottom ϕ per cent and the other to the top ρ per cent of the population distribution

Let the event that one of the two children belongs to the bottom ϕ per cent of the phenotypic distribution be denoted as 'C_l' and the event that the other child belongs to the top ρ per cent of the distribution be denoted as 'C'_h'. Then:

$$\mu_Q = \sum_{g_f}\sum_{g_m}\sum_{g_o.Q}\left[\left(\frac{\int_{-\infty}^{z_L} xf(x,\mu_{g_o},\sigma_e^2)dx * P(C'_h\,|\,g_f,g_m)+\int_{z_H}^{\infty} xf(x,\mu_{g_o},\sigma_e^2)dx * P(C_l\,|\,g_f,g_m)}{F(Z_L,\mu_{g_o},\sigma_e^2)P(C'_h\,|\,g_f,g_m)+(1-F(Z_H,\mu_{g_o},\sigma_e^2))P(C_l\,|\,g_f,g_m)}\right) * P(g_o,g_f,g_m\,|\,Qq_p,Q_o,C_lC'_h)\right],$$

where the threshold phenotypic value Z_H can be computed by the following if ρ is known:

$$\rho\% = \Pr(Y \geqslant Z_H) = \Pr(Y \geqslant Z_H\,|\,QQ)P_{QQ} + \Pr(Y \geqslant Z_H\,|\,Qq)P_{Qq} + \Pr(Y \geqslant Z_H\,|\,qq)P_{qq}.$$

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

$$\mu_q = \sum_{g_f}\sum_{g_m}\sum_{g_{o.q}} \left[\left(\frac{\int_{-\infty}^{z_L} xf(x, \mu_{g_o}, \sigma_e^2)dx^* P(C_h' | g_f, g_m) + \int_{z_H}^{\infty} xf(x, \mu_{g_o}, \sigma_e^2)dx^* P(C_l | g_f, g_m)}{F(Z_L, \mu_{g_o}, \sigma_e^2)P(C_h' | g_f, g_m) + (1 - F(Z_H, \mu_{g_o}, \sigma_e^2))P(C_l | g_f, g_m)} \right) \right.$$
$$\left. {}^* P(g_o, g_f, g_m | Qq_p, q_o, C_l C_h') \right]$$

$$\sigma_Q^2 = \sum_{g_f}\sum_{g_m}\sum_{g_{o.Q}} \left[\left(\frac{\int_{-\infty}^{z_L} x^2 f(x, \mu_{g_o}, \sigma_e^2)dx^* P(C_h' | g_f, g_m) + \int_{z_H}^{\infty} x^2 f(x, \mu_{g_o}, \sigma_e^2)dx^* P(C_l | g_f, g_m)}{F(Z_L, \mu_{g_o}, \sigma_e^2)P(C_h' | g_f, g_m) + (1 - F(Z_H, \mu_{g_o}, \sigma_e^2))P(C_l | g_f, g_m)} \right) \right.$$
$$\left. {}^* P(g_o, g_f, g_m | Qq_p, Q_o, C_l C_h') \right]$$

$$\sigma_q^2 = \sum_{g_f}\sum_{g_m}\sum_{g_{o.q}} \left[\left(\frac{\int_{-\infty}^{z_L} x^2 f(x, \mu_{g_o}, \sigma_e^2)dx^* P(C_h' | g_f, g_m) + \int_{z_H}^{\infty} x^2 f(x, \mu_{g_o}, \sigma_e^2)dx^* P(C_l | g_f, g_m)}{F(Z_L, \mu_{g_o}, \sigma_e^2)P(C_h' | g_f, g_m) + (1 - F(Z_H, \mu_{g_o}, \sigma_e^2))P(C_l | g_f, g_m)} \right) \right.$$
$$\left. {}^* P(g_o, g_f, g_m | Qq_p, q_o, C_l C_h') \right]$$

$$n = Ns^*(P(Qq_p, Q_o, C_l C_h') + P(Qq_p, q_o, C_l C_h'))$$

$$n_H = n + n^* \frac{P(Qq_f, Qq_m, C_l C_h')}{P(Qq_p, C_l C_h')},$$

where

$$P(Qq_f, Qq_m, C_l C_h') = 4p^2 p'^2 \left(\begin{array}{l} \sum_{g_o} P(C_{Yl} | g_o) P(g_o | Qq_f, Qq_m) \sum_{g_o} P(C_h' | g_o) P(g_o | Qq_f, Qq_m) \\ + \sum_{g_o} P(C_{Yh} | g_o) P(g_o | Qq_f, Qq_m) \sum_{g_o} P(C_l' | g_o) P(g_o | Qq_f, Qq_m) \end{array} \right)$$

$$n_Q = Jn_H \frac{P(Qq_p, Q_o, C_l C_h')}{P(Qq_p, Q_o, C_l C_h') + P(Qq_p, q_o, C_l C_h')}$$

$$n_q = Jn_H \frac{P(Qq_p, q_o, C_l C_h')}{P(Qq_p, Q_o, C_l C_h') + P(Qq_p, q_o, C_l C_h')}.$$

One parent's phenotype falls into bottom ϕ per cent of the distribution

Denote the event that one parent belongs to the bottom ϕ per cent of the distribution as '*Pi*'. Conditional on that each nuclear family has at least one heterozygous parent and at least one parent belonging to the bottom ϕ per cent of the phenotypic distribution, we have

$$\mu_Q = \sum_{g_f}\sum_{g_m}\sum_{g_o} E(Y | g_o) P(g_o, g_f, g_m | Qq_p, Q_o, P_l)$$

$$\mu_q = \sum_{g_f}\sum_{g_m}\sum_{g_o} E(Y | g_o) P(g_o, g_f, g_m | Qq_p, q_o, P_l)$$

$$\sigma_Q^2 = \sum_{g_f}\sum_{g_m}\sum_{g_o} E(Y^2 | g_o) P(g_o, g_f, g_m | Qq_p, Q_o, P_l) - \mu_Q^2$$
$$= \sum_{g_f}\sum_{g_m}\sum_{g_o} \mu_{g_o}^2 P(g_o, g_f, g_m | Qq_p, Q_o, P_l) - \mu_Q^2 + \sigma_e^2$$

$$\sigma_q^2 = \sum_{g_f}\sum_{g_m}\sum_{g_o} \mu_{g_o}^2 P(g_o, g_f, g_m | Qq_p, q_o, P_l) - \mu_q^2 + \sigma_e^2.$$

To obtain the above μ_Q, we have

$$P(g_o, g_f, g_m | Qq_p, Q_o, P_l) = \frac{P(g_o, g_f, g_m, Qq_p, Q_o, P_l)}{\sum_{g_f}\sum_{g_m}\sum_{g_o} P(g_o, g_f, g_m, Qq_p, Q_o, P_l)}.$$

Denote the event that a father belongs to the bottom ϕ per cent and a mother is random with respect to her

https://doi.org/10.1017/S0016672302005578 Published online by Cambridge University Press

phenotype as P_{fl}, and the event that the father does not belong to the bottom ϕ per cent and the mother belongs to the bottom ϕ per cent as $P_{fn}P_{ml}$. We have in Equation 20a,

$$P(g_o, g_f, g_m, Qq_p, Q_o, P_l) = P(g_f)P(g_m)P(g_o \mid g_f, g_m, Q_o)P(P_{fl} \mid g_o, g_f, g_m, Q_o) + P(g_f)P(g_m)P(g_o \mid g_f, g_m, Q_o)P(P_{fl}P_{ml} \mid g_o, g_f, g_m, Q_o).$$

For example, if the genotype of the child is QQ, and the parents are of the genotypes QQ and Qq, we have

$$P(g_o, g_f, g_m, Qq_p, Q_o, P_l) = p^3p'[F(Z_L, a, \sigma_e^2) + (1 - F(Z_L, a, \sigma_e^2))^*F(Z_L, d, \sigma_e^2)].$$

With the above two equations, we can obtain μ_Q analytically. Similarly, we can obtain μ_q, σ_Q^2 and σ_q^2 analytically.

$$n = 4pp'N_S \left(\begin{array}{l} (p^2 + p'^2 + pp')^*F(Z_L, d, \sigma_e^2) \\ +[p^2F(Z_L, a, \sigma_e^2) + p'^2F(Z_L, -a, \sigma_e^2) + pp'F(Z_L, d, \sigma_e^2)]^*[1 - F(Z_L, d, \sigma_e^2)] \end{array} \right)$$

$$Jn_H = n + \frac{npp'F(Z_L, d, \sigma_e^2)^*(2 - F(Z_L, d, \sigma_e^2))}{\left(\begin{array}{l} (p^2 + p'^2 + pp')^*F(Z_L, d, \sigma_e^2) \\ +[p^2F(Z_L, a, \sigma_e^2) + p'^2F(Z_L, -a, \sigma_e^2) + pp'F(Z_L, d, \sigma_e^2)]^*[1 - F(Z_L, d, \sigma_e^2)] \end{array} \right)}$$

$$n_Q = Jn_H \frac{P(Qq_p, Q_o, P_l)}{P(Qq_p, Q_o, P_l) + P(Qq_p, q_o, P_l)}$$

$$n_q = Jn_H \frac{P(Qq_p, q_o, P_l)}{P(Qq_p, Q_o, P_l) + P(Qq_p, q_o, P_l)}.$$

References

Allison, D. B. (1997). Transmission-disequilibrium tests for quantitative traits. *American Journal of Human Genetics* **60**, 676–690.

Chen, W. M. & Deng, H. W. (2001). A general and accurate approach for computing the statistical power of the TDT test for complex disease genes. *Genetic Epidemiology* **21**, 53–67.

Deng, H. W., Li, J. L., Li, J., Davies, M. K. & Recker, R. R. (1998*a*). Heterogeneity of bone mass density across skeletal sites and its clinical implications. *Journal of Clinical Densitometry* **1**, 339–353.

Deng, H. W., Li, J., Li, J. L., Johnson, M., Davies, K. M. & Recker, R. R. (1998*b*). Change of bone mass in postmenopausal Caucasian women with and without hormone replacement therapy is associated with vitamin D receptor and estrogen receptor genotypes. *Human Genetics* **103**, 576–585.

Deng, H. W., Chen, W. M. & Recker, R. R. (2000*a*). QTL fine mapping by measuring and testing for Hardy–Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of populations. *American Journal of Human Genetics* **66**, 1027–1045.

Deng, H. W., Chen, W. M., Recker, S., Stegman, M. R., Li, J. L., Davies, K. M., Zhou, Y., Deng, H. Y., Heaney, R. R. & Recker, R. R. (2000*b*). Genetic determination of Colles' fractures and differential bone mass in women with and without Colles' fractures. *Journal of Bone and Mineral Research* **15**, 1243–1252.

Deng, H. W., Chen, W. M., Conway, T., Zhou, Y., Davies, K. M., Stegman, M. R., Deng, H. Y. & Recker, R. R. (2000*c*). Determination of bone mineral density of the hip and spine in human pedigrees by genetic and life-style factors. *Genetic Epidemiology* **19**, 160–177.

Deng, H. W., Li, J. & Recker, R. R. (2001). The effects of background polygenes on transmission disequilibrium test of a QTL in nuclear families with multiple children. *Genetic Epidemiology* **21**, 243–265.

Eaves, L. & Meyer, J. (1994). Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. *Behavior Genetics* **24**, 443–455.

Ewens, W. J. & Spielman, R. S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics* **57**, 455–464.

Falconer, D. S. (1989). *Introduction to Quantitative Genetics*. Harlow, UK: Longman.

Rabinowitz, D. (1997). A transmission disequilibrium test for quantitative trait loci. *Human Heredity* **47**, 342–350.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Risch, N. & Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589.

Risch, N. J. & Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *American Journal of Human Genetics* **58**, 836–843.

Schaid, D. J. (1998). Transmission disequilibrium, family controls, and great expectations. *American Journal of Human Genetics* **63**, 935–941.

Spielman, R. S. & Ewens, W. J. (1996). The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59**, 983–989.

Spielman, R. S. & Ewens, W. J. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics* **62**, 450–458.

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**, 506–516.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer.

Whittaker, J. C. & Lewis, C. M. (1998). The effects of family structure on linkage tests using allelic association. *American Journal of Human Genetics* **63**, 889–897.

Whittaker, J. C. & Lewis, C. M. (1999). Power comparison of the transmission disequilibrium test and sib-transmission disequilibrium-test statistics. *American Journal of Human Genetics* **65**, 578–580.

Wolfram, S. (1996). *The Mathematica*. Champaign, IL: Wolfram Research.

Xiong, M. M., Krushkal, J. & Boerwinkle, E. (1998). TDT statistics for mapping quantitative trait loci. *Annals of Human Genetics* **62**, 431–452.

Zhang, H. & Risch, N. (1996). Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: selected sampling by parental phenotypes. *American Journal of Human Genetics* **59**, 951–957.

Zhang, H. & Risch, N. (1997). Errata. *American Journal of Human Genetics* **60**, 748–750.