

# QTL Mapping Under Ascertainment

J. Peng<sup>1,\*</sup> and D. Siegmund<sup>2</sup>

<sup>1</sup>Department of Statistics, University of California, Davis, CA 95616

<sup>2</sup>Department of Statistics, Stanford University, Stanford, CA 94305

---

## Summary

Mapping quantitative trait loci (QTL) using ascertained sibships is discussed. It is shown that under the standard normality assumption of variance components analysis the efficient scores are unchanged by ascertainment, and two different schemes of ascertainment correction suggested in the literature are asymptotically equivalent. The use of conditional maximum likelihood estimators derived under the normality assumption to estimate nuisance parameters is shown to result in only a small loss of power compared to the case of known parameters, even when the distribution of phenotypes is non-normal and/or the ascertainment criterion is ill defined.

---

Keywords: Gene mapping, Quantitative trait, Ascertainment, Efficient score.

## Introduction

The basic theory associated with standard linkage methods for QTL mapping usually presupposes the pedigrees are randomly sampled, although many linkage studies involve more complicated sampling schemes. The sampling procedure consists of two main steps. First is recruitment of pedigrees; second is phenotyping and genotyping of individuals from the recruited pedigrees. Both the first and second step can involve selective sampling. Roughly speaking there are three possible sampling schemes. Under one scheme we randomly sample pedigrees from the population under study, then phenotype and genotype all individuals recruited. This *random sampling* produces a random sample of trait values. Sample-based estimates of unknown segregation parameters (e.g., trait mean, variance, covariance of sibs) are typically consistent regardless of distributional assumptions. However, random sampling does not favour pedigrees carrying large amounts of linkage information, so a large number of pedigrees must be screened with many individuals phenotyped and genotyped. We can modify random sampling by genotyping only those pedigrees that satisfy certain phenotypic rules. This is *selective*

*genotyping*. For example, after randomly sampling pedigrees we might only genotype pedigrees with at least one member having a phenotype greater than a threshold  $b$ , or we can genotype only pedigrees with at least one pair of relatives having discordant phenotypes. See, for example, Risch & Zhang (1995), Tang & Siegmund (2001), or T. Cuenco *et al.* (2003). This second sampling scheme typically requires that we screen very large samples to find enough pedigrees meeting the rule. Compared to random sampling it costs more to recruit pedigrees, but saves costs of genotyping. Since we still have a population sample of the phenotype, segregation parameters can be consistently estimated as before.

The last procedure is *ascertained sampling*. For many studies pedigrees are sampled on the basis of phenotypes of one or more members, called *probands*, and phenotypes are determined only for ascertained pedigrees. The criteria to identify a set of probands are *ascertainment criteria*. The above examples for selective genotyping can be modified to serve as examples for ascertained sampling, if we replace the word 'genotype' by 'ascertain' in their definitions. In practice the actual ascertainment procedure could be very complicated and not completely clear. For example, we may sample from patients listed in hospital records and recruit the sampled individuals' relatives. Those sampled individuals are probands and the ascertainment criterion involves the procedure

\*E-mail: jie@wald.ucdavis.edu

for a person being listed, which may be unclear to us. Another example arises if we have affected sibpairs from a previous study for a certain disease. We may use these affected sib pairs as probands, and recruit their family members to study related quantitative traits.

With ascertained sampling it is often easier to recruit pedigrees, and the required sample sizes are smaller than with random sampling since we over-sample pedigrees carrying more linkage information, e.g. pedigrees with members having extreme phenotypes. Hence recruiting, phenotyping and genotyping are less expensive. However, since we do not have a population sample of the phenotype, it may be difficult for us to get consistent estimates of the segregation parameters. One possibility is to use estimates from previous studies (e.g. Sham *et al.* 2002). However this could be troublesome since it is difficult to be confident that populations in the different studies have the same segregation parameters. For example, two populations could have different covariate effects that causes different population mean values. In this paper, following suggestions of (for example) Hopper & Mathews (1982) and Beaty & Liang (1987), we propose to use the conditional maximum likelihood estimates under the normality assumption given the phenotypes of probands. A concern with this approach is that the procedure depends heavily on the normality assumption, which cannot be expected to be satisfied in general. Another concern is that it may be difficult to define a proper ascertainment correction,

since in many situations the ascertainment procedure is unclear.

An outline of the paper is as follows. We first assume the ideal situation of normality and well-defined ascertainment criteria and study the score statistic and its relative efficiency. We show that two ascertainment corrections proposed in the literature (Hopper & Mathews, 1982; Elston & Sobel, 1979) are asymptotically equally efficient. A special case of this result when there is only one proband was given by Peng & Siegmund (2004), but the argument here is quite different as a reflection of the greater generality. We then discuss various more complicated ascertainment procedures and describe the effect of misspecification of nuisance parameters. Finally, we discuss the robustness of the power of the score statistic when the distribution of the phenotype is mildly non-normal and/or the ascertainment criteria are ill-defined.

The reader more interested in the conclusions than our justification for reaching those conclusions may want to turn first to the results section and Tables 1 and 2. There s/he will find evidence, under a variety of less than ideal conditions concerning the phenotype distribution and precision in specification of the ascertainment rule, that an ascertainment correction based on normal theory and conditioning on the phenotypes of probands is effective in preserving the power to detect linkage, whereas failure to make a correction leads to a considerable loss of power.

**Table 1** power of the robust score statistic at the 0.05 genome-wide significance level. (The population parameters are  $\mu = 0, \sigma_Y = 1, \rho = 0.25$ , and the linkage parameters are  $\alpha_0 = 0.1, \delta_0 = 0$ . Four hundred sibtrios are ascertained. The ascertainment rule A-sample has parameters  $\eta = 0.2, p = 0.2$ . The multivariate-*t* trait has degrees of freedom  $k = 20$  and the multivariate gamma trait has a 'shape' parameter  $c = 100$ . There are 31 equally spaced, fully informative markers with spacing  $\Delta = 5\text{cM}$ , and the trait locus is on the 16th marker. Evaluation of the power of the statistic (10) using the true nuisance parameters, the CMLE, and the uncorrected MLE is based on 5000 iterations.)

Model	Estimates	A-first			A-max			A-sample		
		75th	90th	95th	75th	90th	95th	75th	90th	95th
Multi-N	True	0.395	0.678	0.848	0.383	0.632	0.811	0.388	0.639	0.813
	<b>CMLE</b>	<b>0.395</b>	<b>0.674</b>	<b>0.847</b>	<b>0.354</b>	<b>0.616</b>	<b>0.799</b>	<b>0.389</b>	<b>0.636</b>	<b>0.813</b>
	MLE	0.230	0.370	0.520	0.260	0.396	0.528	0.212	0.346	0.464
Multi-G	True	0.386	0.684	0.840	0.359	0.623	0.783	0.366	0.633	0.799
	<b>CMLE</b>	<b>0.382</b>	<b>0.683</b>	<b>0.835</b>	<b>0.330</b>	<b>0.603</b>	<b>0.775</b>	<b>0.363</b>	<b>0.628</b>	<b>0.796</b>
	MLE	0.218	0.401	0.534	0.235	0.405	0.500	0.215	0.367	0.487
Multi-T	True	0.322	0.602	0.752	0.318	0.557	0.701	0.312	0.557	0.710
	<b>CMLE</b>	<b>0.322</b>	<b>0.597</b>	<b>0.746</b>	<b>0.299</b>	<b>0.540</b>	<b>0.694</b>	<b>0.308</b>	<b>0.552</b>	<b>0.708</b>
	MLE	0.167	0.314	0.421	0.197	0.340	0.429	0.179	0.315	0.394

**Table 2** Mean and standard deviation of the CMLE. (The mean and standard deviation are estimated via 5000 iterations. True nuisance parameters under estimation are  $\mu = 0, \sigma_Y = 1, \rho = 0.25$ . For each parameter the number in the first line is the mean of the CMLE and the number in the second line is the standard deviation.)

Model	$\theta$	A-first			A-max			A-sample		
		75th	90th	95th	75th	90th	95th	75th	90th	95th
Multi-N	$\mu$	-0.0038	-0.0044	-0.0001	-0.2915	-0.1398	-0.0767	-0.0569	-0.0293	-0.0180
		0.1061	0.1485	0.1768	0.1202	0.1414	0.1697	0.1061	0.1414	0.1626
	$\sigma_Y$	1.0023	1.0066	1.0087	0.9382	0.9229	0.9332	0.9938	0.9907	0.9950
		0.0283	0.0354	0.0354	0.0283	0.0283	0.0283	0.0283	0.0283	0.0354
	$\rho$	0.2487	0.2474	0.2449	0.2383	0.2158	0.2169	0.2456	0.2404	0.2404
Multi-G	$\mu$	-0.0092	-0.0120	-0.0060	-0.3042	-0.1623	-0.0867	-0.0610	-0.0376	-0.0259
		0.1064	0.1528	0.1862	0.1178	0.1513	0.1755	0.1022	0.1431	0.1736
	$\sigma_Y$	1.0302	1.0527	1.0689	0.9548	0.9555	0.9770	1.0189	1.0339	1.0494
		0.0325	0.0338	0.0347	0.0320	0.0318	0.0318	0.0324	0.0332	0.0339
	$\rho$	0.2516	0.2498	0.2471	0.2394	0.2181	0.2144	0.2475	0.2431	0.2420
Multi-T	$\mu$	-0.0055	-0.0103	-0.0111	-0.2802	-0.1451	-0.0725	-0.0547	-0.0317	-0.0142
		0.1078	0.1579	0.1905	0.1174	0.1449	0.1715	0.1035	0.1448	0.1734
	$\sigma_Y$	1.0257	1.0666	1.1006	0.9548	0.9702	1.0048	1.0131	1.0452	1.0756
		0.0353	0.0378	0.0392	0.0346	0.0341	0.0347	0.0340	0.0363	0.0377
	$\rho$	0.2492	0.2493	0.2483	0.2312	0.2102	0.2049	0.2438	0.2394	0.2369
		0.0561	0.0618	0.0636	0.0556	0.0568	0.0592	0.0544	0.0586	0.0607

**Methods**

**Variance-Component Model**

We first introduce the variance component model for mapping QTL and derive the efficient score based on this model when there is ascertainment. We assume Hardy-Weinberg equilibrium and linkage equilibrium throughout. Therefore, for both markers and QTL, haplotypes within the same locus and genotypes among different loci are stochastically independent. Our model goes back to the classic paper of Fisher (1918) for the case of diallelic genes; the general case was discussed by Kempthorne (1957). The basic model for the phenotype  $Y$  having a mean value  $\mu$  is

$$Y = \mu + \alpha_m + \alpha_f + \delta_{m,f} + e, \tag{1}$$

where  $\alpha_x = \alpha_x(\tau)$  denotes the additive genetic effect of allele  $x$  at locus  $\tau$ , and  $\delta_{x,y}$  denotes the dominance deviation of alleles  $x, y$ . The subscript  $m(f)$  denotes the allele contributed by the mother (father). By standard analysis of variance arguments we may assume that  $E(\alpha_m) = E(\alpha_f) = E(e) = E[\delta_{m,f}|m] = E[\delta_{m,f}|f] = 0$ . By the assumption of Hardy-Weinberg equilibrium  $m$  and  $f$  are independent (unless the parents are inbred), so the different genetic effects in the model (1) are un-

correlated. We also assume the residual term  $e$ , which may include the genetic effects from other QTL, is uncorrelated with the explicitly modelled genetic effects from  $\tau$ . It follows that  $\sigma_Y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2$ , where  $\sigma_A^2 = 2 E(\alpha_m^2), \sigma_D^2 = E(\delta_{m,f}^2), \sigma_e^2 = E(e^2)$  are the additive variance, dominance variance and residual variance, respectively. We later discuss briefly the possibility that the model also allows covariates in the mean value.

Although the analysis below can be applied to more general pedigrees in this paper we focus on sibships of the same size,  $s$ , because of their simplicity and the possibility to derive explicit formulae. Under model (1) it is easy to calculate variance components, as well as covariance matrices.

Consider a pair of siblings satisfying model (1). Denote by  $v = v(\tau)$  the number of alleles identical by descent at  $\tau$ . Letting  $Y_i$  denote the phenotypic value of the  $i$ th sibling ( $i = 1, 2$ ), we have

$$\begin{aligned} \text{Cov}(Y_1, Y_2 | v) &= \sigma_A^2 v/2 + \sigma_D^2 1_{\{v=2\}} + \sigma_e^2 r \\ &= \text{Cov}(Y_1, Y_2) + \alpha_0(v - 1) \\ &\quad + \delta_0(1/2 - 1\{v = 1\}) \end{aligned} \tag{2}$$

where  $r = \text{corr}(e_1, e_2)$  accounts for the correlation between sibs that arises from other QTL and from a

shared environment, while

$$\alpha_0 = \frac{\sigma_A^2 + \sigma_D^2}{2}, \quad \delta_0 = \frac{\sigma_D^2}{2}.$$

If there are gene-gene interactions (2) still holds, but  $\alpha_0$ ,  $\delta_0$  will also include interaction variance components (Tang & Siegmund, 2002). Note that  $0 \leq \delta_0 \leq \alpha_0$ .

### Efficient Scores

The null hypothesis we want to test is  $\alpha_0 = 0$  (which implies that  $\delta_0 = 0$  as well). The *working assumption* for the variance-component model is that the conditional distribution of the phenotypes in a pedigree, given the pairwise IBD sharing at a QTL, is multivariate normal. The usefulness of this assumption is due to the mathematical tractability of the multivariate normal distribution. It cannot be expected to be exactly true, so it is important to check that the procedures we obtain are reasonably robust. For a given sibship let  $\mathbf{Y}^T = (Y_1, \dots, Y_s)$  denote the vector of phenotypes and  $M$  denote the matrix of all marker genotypes. The observed data for a sampled sibship is  $(\mathbf{Y}, M)$ . Also let  $A_v$  denote the  $s \times s$  matrix with entries  $v_{ij} = 1$  for  $i \neq j$  and zeroes along the diagonal. Let  $L = P(\mathbf{Y}, M)$  denote the likelihood of a randomly sampled sibship. Let  $\ell'(\alpha_0)$  denote  $\partial \log(L)/\partial \alpha_0$ . Under the normality assumption, when the markers are fully informative, i.e. the IBD sharing matrix  $A_v$  is observable, the efficient score at a putative trait locus  $t$ , evaluated with  $\alpha_0 = 0$ , is (Tang & Siegmund, 2001)

$$\ell'(0) = -\frac{1}{2} \text{tr}(\Sigma^{-1} A_{v(t)}) + \frac{1}{2} (\mathbf{Y} - \mu \mathbf{1})^T \Sigma^{-1} A_{v(t)} \Sigma^{-1} (\mathbf{Y} - \mu \mathbf{1}), \quad (3)$$

where  $\Sigma = E[(\mathbf{Y} - \mu \mathbf{1})(\mathbf{Y} - \mu \mathbf{1})^T]$  is the phenotypic covariance matrix. When markers are not fully informative the likelihood can be written as

$$L = P(\mathbf{Y}, M) = \sum P(\mathbf{Y} | A_v) P(A_v | M) P(M),$$

where the summation is taken over all possible configurations of the IBD sharing matrix  $A_v$  and we use the fact that  $P(\mathbf{Y} | M, A_v) = P(\mathbf{Y} | A_v)$ . Since (3) is linear in  $A_v$ , it is easy to see that the efficient score for  $\alpha_0$  is obtained by replacing  $A_v$  with its conditional expectation

$A_{\hat{v}} = E(A_v | M)$  in (2.3), i.e.,

$$\ell'(0) = -\frac{1}{2} \text{tr}(\Sigma^{-1} A_{\hat{v}}) + \frac{1}{2} (\mathbf{Y} - \mu \mathbf{1})^T \Sigma^{-1} A_{\hat{v}} \Sigma^{-1} (\mathbf{Y} - \mu \mathbf{1}). \quad (4)$$

Now let  $\mathcal{A}$  denote the event that a particular sibship is ascertained. We assume the following *measurable ascertainment assumption*: each pedigree is ascertained through its phenotypes, and possibly some additional randomization, but not its genotypes. The conditional likelihood of the data given  $\mathcal{A}$  is

$$L_{\mathcal{A}} = P(\mathbf{Y}, M | \mathcal{A}) = [P(\mathbf{Y}, M)/P(\mathcal{A})] I(\mathcal{A}) = [L/P(\mathcal{A})] I(\mathcal{A}), \quad (5)$$

where  $I(\cdot)$  is the indicator function. The following proposition shows that the efficient score of the conditional likelihood  $L_{\mathcal{A}}$  is the same as the efficient score of the unconditional likelihood  $L$ , given by (4). In what follows we shall also be interested in a version of Proposition 1 and its companion, Proposition 2 below, when we condition on the value of a subset of the coordinates of  $\mathbf{Y}$  rather than on the event of ascertainment. In this case  $P(\mathcal{A})$  would be the joint probability density of that subset of coordinates. Slight variants of the proofs of Propositions 1 and 2 hold in this case as well. Let  $\ell'_{\mathcal{A}}(\alpha_0)$  denote  $\partial \log(L_{\mathcal{A}})/\partial \alpha_0$ .

*Proposition 1.* Under the normality assumption and the measurable ascertainment assumption,

$$\ell'_{\mathcal{A}}(0) = \ell'(0) I(\mathcal{A}).$$

*Proof.* By (5) it suffices to show that  $\partial \log[P(\mathcal{A})]/\partial \alpha_0$  equals zero when it is evaluated at  $\alpha_0 = 0$ . Let  $E_0$  denote expectation when  $\alpha_0 = 0$ . Then

$$\begin{aligned} \partial \log[P(\mathcal{A})]/\partial \alpha_0 &= \partial \log[E(I(\mathcal{A}))]/\partial \alpha_0 \\ &= \partial \log[E_0(L(\alpha_0)/L(0); \mathcal{A})]/\partial \alpha_0 \\ &= E(\ell'; \mathcal{A})/P(\mathcal{A}), \end{aligned}$$

where  $\ell'(\alpha_0) = \partial \log[L(\alpha_0)]/\partial \alpha_0$ . By the measurable ascertainment assumption,

$$E_0(\ell'(0); \mathcal{A} | \mathbf{Y}) = I(\mathcal{A}) E_0(\ell'(0) | \mathbf{Y}).$$

Since by (4) and  $E_0(A_{\hat{v}}) = E_0(A_v) = 0$ , we see that  $E_0(\ell'(0) | \mathbf{Y}) = 0$ , and hence  $E_0(\ell'(0); \mathcal{A} | \mathbf{Y}) = 0$ .

This implies that  $\partial \log[P(\mathcal{A})]/\partial \alpha_0$  equals zero when  $\alpha_0 = 0$ .

The preceding discussion also applies to  $\delta_0$  with the efficient score with respect to  $\delta_0$  under the fully informative case given by

$$-\frac{1}{2} \text{tr}(\Sigma^{-1} D_v) + \frac{1}{2} (\mathbf{Y} - \mu \mathbf{1})^T \Sigma^{-1} D_v \Sigma^{-1} (\mathbf{Y} - \mu \mathbf{1}),$$

where  $D_v$  is a matrix similar to  $A_v$  with off-diagonal elements  $1/2 - I(v_{ij} = 1)$ . The corresponding efficient scores  $\ell'_\delta(0)$  of the likelihood  $L$  and  $L_{\mathcal{A}}$  are the same, and are given by replacing  $I(v_{ij} = 1)$  with  $P(v_{ij} = 1 | M)$  in the preceding display. Proposition 1 says that the efficient scores for the linkage parameters are the same for all three sampling schemes if the segregation parameters  $\mu$  and  $\Sigma$  are given. Estimation of these parameters is discussed below.

**Remark.** Wang (2005) obtained a decomposition of the efficient score under ascertainment into the sum of two terms, but did not point out that this sum reduces to the efficient score under random sampling.

It has been shown by Tang & Siegmund (2001) that, under the assumptions of normality and fully informative markers, the segregation parameters are orthogonal to the linkage parameters  $\alpha_0, \delta_0$  under the null hypothesis of no linkage:  $\alpha_0 = 0$ . The reason is that the efficient scores for the nuisance parameters under the null hypothesis do not involve the identity by descent counts  $v_{ij} - 1$ , and the efficient scores for the linkage parameter  $\alpha_0, \ell'(0)$ , is linear in  $v_{ij} - 1$ , while  $E(v_{ij} - 1) = 0$ . When markers are partially informative the same properties hold if we replace  $v_{ij} - 1$  by  $\hat{v}_{ij} - 1$ . Therefore, for the unconditional likelihood  $L$ , we have the nuisance parameters orthogonal to the linkage parameter  $\alpha_0$  under the null hypothesis of no linkage. By a similar argument to Proposition 1 we now show this is true for the conditional likelihood  $L_{\mathcal{A}}$ .

**Proposition 2.** Under the normality assumption and the measurable ascertainment assumption, for the conditional likelihood  $L_{\mathcal{A}}$ , the nuisance parameters are orthogonal to the linkage parameters under the null hypothesis of no linkage, i.e. with  $\theta$  denoting one of the segregation parameters  $\mu, \sigma_Y^2$  or  $\rho = \text{Corr}(Y_i, Y_j)$ , and  $\lambda$  denoting either of the linkage parameters  $\alpha_0, \delta_0$ , we have  $E_0[(\partial^2 \log(L_{\mathcal{A}})/\partial \lambda \partial \theta)|_{\alpha_0=0} | \mathcal{A}] = 0$ .

**Proof.** We only prove the result for  $\alpha_0$ ; the result for  $\delta_0$  can be proved similarly. Observe that

$$\begin{aligned} \partial^2 \log(L_{\mathcal{A}})/\partial \alpha_0 \partial \theta &= \partial^2 \log(L)/\partial \alpha_0 \partial \theta \\ &\quad - \partial^2 \log[P(\mathcal{A})]/\partial \alpha_0 \partial \theta. \end{aligned}$$

By the discussion above for the unconditional likelihood  $L$ ,

$$E_0[(\partial^2 \log(L)/\partial \alpha_0 \partial \theta)|_{\alpha_0=0} | \mathcal{A}] = 0.$$

By Proposition 1, since  $\partial \log[P(\mathcal{A})]/\partial \alpha_0$  equals zero under  $\alpha_0 = 0$ , so does  $\partial^2 \log[P(\mathcal{A})]/\partial \alpha_0 \partial \theta$ , which completes the proof.

An important consequence of Proposition 2 follows from standard likelihood theory: if we estimate segregation parameters by (conditional) maximum likelihood, in the evaluation of noncentrality parameters for the asymptotic theory to follow these segregation parameters can be regarded as known.

### Test Statistics

By Proposition 1, at a putative trait locus  $t$  the efficient score for one ascertained sibship is given by (4). To use this score as a test statistic we will standardize it to have unit variance under the null hypothesis, and substitute estimates for the unknown segregation parameters  $\mu, \Sigma$ . The variance of the efficient score  $\ell'(0)$  under the null hypothesis is given by

$$I_{\alpha\alpha} = E_0(\ell'^2(0) | \mathcal{A}). \tag{6}$$

Therefore at a putative locus  $t$ , for a sample of ascertained sibships, the *standardized score* is

$$Z_t = \sum_n \ell'_n(0) / \left[ \sum_n I_{\alpha\alpha,n} \right]^{1/2}, \tag{7}$$

where the summation is taken over all ascertained sibships. A second statistic involving  $\ell'_\delta(0)$  can be defined similarly. Since  $\alpha_0 = (\sigma_A^2 + \sigma_D^2)/2, \delta_0 = \sigma_D^2/2$ , a two dimensional statistic rarely has more power than (7) unless the dominance effect  $\sigma_D^2$  is quite large. Therefore for the following discussion we only consider (7). For a more detailed discussion of the two dimensional statistics see Tang & Siegmund (2001) and Wang & Huang (2002). Since we do not know the location of the QTL

$\tau$  we can scan the genome by the *test statistic*

$$Z_{\max} = \max_t Z_t,$$

where the maximum is taken over all marker loci  $t$  throughout the genome. The expectation of the score statistic at the trait locus  $\tau$  is the *noncentrality parameter*

$$\xi = E(Z_\tau | \mathcal{A}).$$

This quantity is calculated by Tang & Siegmund (2001) under the assumptions of random sampling and fully informative markers.

Explicit formulæ can be obtained for sibships of size  $s$  which are ascertained through  $r \leq s$  siblings. However, the formula is quite complicated and will not be presented here. See Peng (2004) Appendix A.3 for a complete discussion.

When markers are fully informative it is easy to see that at a marker  $t$ ,  $E(Z_t | \mathcal{A}) = \xi(1 - 2\phi)$ , where under the Haldane model for recombination,  $1 - 2\phi = \exp(-0.04|t - \tau|)$  and  $|t - \tau|$  is the distance in centiMorgans (cM) between the marker  $t$  and the trait locus  $\tau$ . In the case of complete marker polymorphism the testing process  $\{Z_t\}$  is asymptotically an Ornstein-Uhlenbeck process. An approximation to the genome-wide false positive rate,  $P_0(Z_{\max} \geq z)$ , is given by Feingold *et al.* (1993). When markers are only partially informative the process is still approximately a Gaussian process, but its p-value is slightly smaller. One can use the fully informative approximation, which is slightly conservative, or a Monte Carlo method to approximate the p-value.

*Remark.* In (1) we consider a model without covariate effects. The simplest generalization of (1) is the model

$$Y = \mu + b^T w + \alpha_m + \alpha_f + \delta_{m,f} + e, \tag{8}$$

where  $w$  is the covariate under consideration. In this model we still keep assumptions of the model (1) and in addition assume that  $w$  is independent of the residual term  $e$  and the IBD sharing matrix  $A_v$ . For the phenotype  $\mathbf{Y}$  of a pedigree we assume a multivariate normal distribution after conditioning on the matrix  $\mathbf{w} = (w_1, \dots, w_s)$  of covariates and the IBD sharing matrix  $A_v$ .

The observed data consists of the phenotypes  $\mathbf{Y}$ , the covariates  $\mathbf{w}$  and the genotypes  $M$ . The likelihood is  $L_{\mathcal{A}} = P(\mathbf{Y}, \mathbf{w}, M | \mathcal{A}) = [P(\mathbf{Y}, \mathbf{w}, M) / P(\mathcal{A})] I(\mathcal{A})$ . The

measurable ascertainment assumption is that  $\mathcal{A}$  is determined by conditions on the phenotype  $Y$  and covariates  $w$ ; then the efficient score with respect to  $\alpha_0$  when  $\alpha_0 = 0$  is still given by (4), with  $\mathbf{Y}$  replaced by  $\mathbf{Y} - b \mathbf{w}^T$ . It follows that  $E_0(\ell'(0) | Y, w) = 0$ . Proposition 1 continues to apply, and the nuisance parameters (now including the regression coefficient  $b$ ) are orthogonal to the linkage parameters under the null hypothesis of no linkage. The analysis becomes much more complicated if there is gene-covariate interaction (cf. Peng *et al.* 2005).

### Two Ascertainment Corrections

Two different ascertainment corrections have been suggested in the literature in order to estimate the segregation parameters. Elston & Sobel (1979) suggest conditioning on the event of ascertainment, while Hopper & Mathews (1982) suggest conditioning on the exact phenotypic values of the probands. Andrade & Amos (2000) compared these methods by simulation and found them to be comparable. Here we show that under the normality and measurable ascertainment assumptions they are asymptotically equivalent when the sample size is large, so the results obtained by simulation are exactly as expected.

Suppose a sibship of size  $s$  is ascertained on the basis of phenotypic values of a specific subset of  $r$  siblings defined independently of the phenotypes with  $r < s$ . In other words, we sample those sibships satisfying  $(Y_1, \dots, Y_r)^T \in S$ , where  $S$  is a subset of  $\mathcal{R}^r$ . For an ascertained sibship let  $\mathbf{Y}_{(1)} = (Y_1, \dots, Y_r)^T$  be the probands' phenotypes, and  $\mathbf{Y}_{(2)} = (Y_{r+1}, \dots, Y_s)^T$  be the other sibs' phenotypes. Let  $L_V = P(\mathbf{Y}, M | \mathbf{Y}_{(1)})$  be the conditional likelihood suggested by Hopper & Mathews. As defined before  $L_A = P(\mathbf{Y}, M | \mathbf{Y}_{(1)} \in S)$  is the conditional likelihood suggested by Elston & Sobel. By a similar argument to Proposition 1 one can see that the efficient score of  $L_V$  with respect to  $\alpha_0$  when  $\alpha_0 = 0$  is again  $\ell'(0)$  ((4)). Note that, when  $r = 1$ , this ascertainment rule is called single ascertainment (Ewens & Green 1988) and as A-first in the results section. The Elston-Sobel likelihood  $L_A$  corresponds to equation (4) of Ewens & Green's paper; and the Hopper-Mathews likelihood  $L_V$  corresponds to equation (8) of that paper. Ewens & Green (1988) reach the conclusion

that under single ascertainment both (4) and (8) give consistent estimates of the parameters. Below we show that these two approaches give asymptotically the same noncentrality for general  $r < s$ . Let

$$I_{\alpha\alpha}^{HM} = E_0(\ell'^2(0)|\mathbf{Y}_{(1)}), \quad I_{\alpha\alpha}^{ES} = E_0(\ell'^2(0)|\mathbf{Y}_{(1)} \in S)$$

be the normalizing factors for the two conditional likelihoods. Then the standardized score statistics for  $N$  ascertained sibships are

$$Z_t^{HM} = \sum_{n=1}^N \ell'_n(0) / \left[ \sum_n I_{\alpha\alpha,n}^{HM} \right]^{1/2},$$

$$Z_t^{ES} = \sum_{n=1}^N \ell'_n(0) / \left[ \sum_n I_{\alpha\alpha,n}^{ES} \right]^{1/2},$$

respectively. Let  $\theta = (\mu, \Sigma)$  denote the vector of segregation parameters, and let  $\hat{\theta}^{HM}$  and  $\hat{\theta}^{ES}$  denote the maximum likelihood estimators under the condition  $\alpha_0 = 0$  derived from  $L_V$  and  $L_A$ , respectively. By Proposition 2, the law of large numbers, and standard likelihood theory we find that these two statistics evaluated at their corresponding maximum likelihood estimators have the same asymptotic noncentrality parameter.

**Proposition 3.** Under the normality and measurable ascertainment assumptions, for local alternatives, so  $\alpha_0$  is inversely proportional to  $N^{1/2}$ , the two statistics  $Z_t^{HM}$ ,  $Z_t^{ES}$  evaluated at  $\hat{\theta}^{HM}$  and  $\hat{\theta}^{ES}$ , respectively, have the same asymptotic noncentrality parameter,

$$\xi = \alpha_0 [NE_0(\ell'^2(0)|\mathbf{Y}_{(1)} \in S)]^{1/2},$$

as  $N \rightarrow \infty$ . (A similar result holds for  $\delta_0$ .)

A proof of Proposition 3 for the special case of single ascertainment is sketched by Peng & Siegmund (2004). Given Propositions 1 and 2 the proof in the general case is similar. For more details see Peng (2004).

Although these two ascertainment corrections are asymptotically equivalent, they involve different estimates of the nuisance parameters,  $\hat{\theta}^{HM}$  and  $\hat{\theta}^{ES}$ . Moreover, when conditioning on the exact phenotypic values of the probands the number  $r$  of the probands must be less than the sibship size  $s$ , while in principal  $r$  can equal  $s$  if one conditions on the event of ascertainment. Although Elston & Sobel's correction can theoretically be applied to any ascertainment procedure it does require a clear definition of the procedure, since we need to calculate the probability of ascertainment,  $P_0(\mathcal{A})$ , in

order to obtain efficient scores for the nuisance parameters. However, in practice we might not know how a pedigree is ascertained. Since Hopper & Mathews' suggested correction requires only that we know through whom the pedigree is ascertained, i.e., the probands, we propose to use this correction for parameter estimation while remaining aware of its limitations.

In practice an ascertainment rule can be ill-defined. Alternatively each member of a pedigree can be eligible to become one of the probands, and the proband may be determined by the phenotypes of the entire pedigree. For example, a sibship may be ascertained if at least one sibling has a phenotypic value exceeding some level, and the proband is selected at random from among those eligible (This ascertainment rule is called A-max in the results section and complete ascertainment in Ewens & Green, 1988). However, if the ascertainment is stringent in the sense that it is unlikely that two or more different subsets of the pedigree satisfy the ascertainment criterion, then  $L_V \sim L_A$  under some mild distributional assumptions, and conditioning on the phenotypic values of the probands often gives reasonable results. For instance, in the case of A-max, when the threshold  $T$  goes to infinity,  $L_V \sim L_A$ . A theoretical discussion of this point for general  $r < s$  is given by Peng (2004). Our conclusion agrees with that of Ewens & Green (1988, p. 442) who write "when the threshold  $T$  is high, it is unlikely that more than one phenotypic value will exceed  $T$ , so that conditioning on the maximum phenotype is approximately the same as conditioning on a unique proband phenotype." In this paper we use numerical examples in the results section to demonstrate this point. We also observe that, as  $T$  becomes larger, the bias caused by incorrectly using  $L_V$  becomes smaller, and even when the bias is not so small (for example, when  $T$  is the 75th quantile) the effect on power is still very small (cf. Tables 1 and 2).

*Remark.* An appealing design for some traits that avoids some of these fundamental difficulties is to ascertain *nuclear families* through the parent(s). It can be shown that the noncentrality parameter is of the same form as that of a randomly sampled sibship, but with the sibling correlation  $\rho$  replaced by the conditional correlation and the phenotypic variance  $\sigma_Y^2$  replaced by the conditional variance (Peng, 2004). However, since these noncentralities do not depend on phenotypes, compared

with randomly sampled nuclear pedigrees, there is no gain in genotyping efficiency by ascertainment through parent(s). Hence selective genotyping has no advantage, and the advantage of ascertained sampling is only in the possible ease of recruiting pedigrees.

### Conditional Maximum Likelihood Estimates

When the size of the proband set  $r$  is less than the size of the pedigree  $s$  the conditional likelihood can be used to derive the maximum likelihood estimates. Suppose the  $n$ th pedigree has size  $s_n$  and a proband-set of size  $r_n < s_n$ , and let  $\mathbf{Y}_{n,(1)}$  denote the phenotypes of the probands. Then the conditional likelihood under the null hypothesis of no linkage is

$$L_V(0) = C \prod_n P_0(\mathbf{Y}_n | \mathbf{Y}_{n,(1)}), \tag{9}$$

where  $C$  is independent of the nuisance parameters. Under the normality assumption  $\log(L_V(0))$  and its derivatives are all easily calculable. Therefore, we can find the maximum likelihood solution to (9) by iterative algorithms such as Fisher's scoring or Newton-Raphson. See Andrade *et al.* (1999) for a detailed discussion. For several special cases there is an explicit solution to (9), so no numerical search for maxima is needed. For example, if  $s_n = s, r_n = r$  for all  $n$  then there exist explicit expressions for the maximum likelihood estimators for  $L_V(0)$  (cf. Peng, 2004).

In what follows we refer to the (C)MLE to mean the (conditional) maximum likelihood estimate derived under the normality assumption, although it may not be the true (C)MLE when the normality assumption does not hold.

### Robustness of the Test Statistics

When the normality assumption is violated the standardization of the statistics discussed above is, in general, incorrect, and leads to the possibility of an inflated false positive error probability. This problem can be mitigated by using the conditional variance of  $\ell'(0)$ , given the phenotypic values  $\mathbf{Y}_1, \dots, \mathbf{Y}_N$  for all the ascertained sibships. Then the normalized score at marker  $t$  becomes

$$Z_t = \sum_n \ell'_n(0) / \left[ \sum_n I_{\alpha\alpha,n}^R \right]^{1/2}, \tag{10}$$

where  $I_{\alpha\alpha,n}^R = E_0[\ell_n'^2(0) | \mathbf{Y}_n]$ . (This quantity has been evaluated explicitly by Tang & Siegmund (2001) for sibships and Peng (2004) for nuclear families.) If the estimator  $\hat{\theta}$  is a function (only) of the phenotypes, the statistic (2.10) is asymptotically normally distributed with mean zero, variance one, and the testing process  $\{Z(t)\}$  is approximately a Gaussian process. In the case of fully informative markers it is a discretely observed Ornstein-Uhlenbeck process. In the following we call this statistic the *robust score statistic*. If the normality assumption is in fact true, and  $\hat{\theta}$  is the CMLE, (10) has the same asymptotic noncentrality as (7).

Suppose the data consist of  $N$  sibships of the same size  $s$  which are ascertained through  $r (< s)$  probands. Then the CMLE are functions of various moments. Hence by the law of large numbers, as the sample size  $N$  becomes infinitely large, the CMLE  $\hat{\theta}_N$  converges to a limit  $\theta^*$  in probability. It is not hard to show that, under local alternatives, the asymptotic distribution of the robust score statistic (10) with  $\hat{\theta}_N$  is the same as that of (10) with the limiting value  $\theta^*$ . Actually for any estimator  $\hat{\theta}$  that converges to a limiting value  $\theta^*$  and is  $N^{1/2}$ -consistent for this limit, the asymptotic distribution of the robust score statistic under  $\hat{\theta}$  is the same as that under the limiting value  $\theta^*$ . This property is based on standard, albeit complicated, arguments that depend crucially on the fact that the derivative of the "efficient score"  $\ell'(\alpha_0)$  with respect to the nuisance parameters, evaluated at  $\alpha_0 = 0$ , is linear in  $A_{\hat{v}}$  and hence has zero mean for any specifications of the nuisance parameters. It does not depend on the normality assumption. Therefore, if a  $N^{1/2}$ -consistent estimate of the nuisance parameters is used, asymptotically we will not lose efficiency due to parameter estimation (although we may lose efficiency due to our choice of statistic, which was also based on the normality assumption).

### Relative Efficiency and Sensitivity to Misspecification of Nuisance Parameters Under the Normality Assumption

We start with the simplest possible situation for ascertainment, where each sibship contains one proband which is the oldest sibling, with phenotypic value  $Y_1$ , and ascertainment is based on a threshold  $T$ , so a sibship is ascertained if the proband's phenotype satisfies

$Y_1 \geq T$ : single ascertainment. We discuss the behaviour of the asymptotic noncentrality parameter  $\xi$  as a function of sibship size and threshold level, and its sensitivity to possibly misspecified values of the nuisance parameters. Peng (2004) contains additional details. A Monte Carlo study of some related issues, including mild non-normality, is given in the following section.

Suppose we have sibships of size  $s$  ascertained by  $S = \{Y_1 \geq T\}$ . Under the normality assumption, and the ideal case of fully informative markers, the large sample noncentrality per ascertained sibship is calculable and given in Peng & Siegmund (2004). For the special case of sibpairs it is given by the simple expression

$$\alpha_0 [\rho^2 / (1 - \rho^2)^2 + E[(Y_1 - \mu)^2 / \sigma_Y^2 | Y_1 \geq T] / 2(1 - \rho^2)]^{1/2}.$$

Suppose there are two sampling schemes where the  $i$ th scheme ascertains  $N_i$  sibships of size  $s_i$  based on the ascertainment criterion  $S_i = \{Y_1 \geq T_i\}$  ( $i = 1, 2$ ). In order to achieve approximately the same power for linkage detection, the sample size  $N_i$  of these two schemes should satisfy  $N_1 \xi_1^2 \approx N_2 \xi_2^2$ , where  $\xi_i$  is the large sample noncentrality per ascertained sibship for the  $i$ th scheme. Therefore  $\xi_i^2$  could be viewed as an indicator of the efficiency of scheme  $i$ . A smaller  $\xi$  means that a larger sample size is needed. Since the two schemes might have different sibship sizes  $s_i$  it may be more appropriate to compare the number of individuals that are phenotyped and genotyped, which is  $G_i = N_i s_i$ . Therefore to achieve the same power we should require

$$G_2 / G_1 \approx (\xi_1^2 / s_1) / (\xi_2^2 / s_2).$$

Thus we define  $\mathcal{E} = \xi^2 / s$  as the *efficiency* for a sampling scheme.

Under single ascertainment  $\mathcal{E}$  is a function of the sibship size  $s$  and the ascertainment threshold  $T$ :  $\mathcal{E} = \mathcal{E}(s, T)$ . It can be shown that  $\mathcal{E}$  is an increasing function of  $s$  and  $T$  ( $>0$ ), which means that larger sibships and/or more stringent ascertainment rules lead to smaller phenotyping and genotyping costs. Of course, it is possible that these sampling schemes lead to greater overall recruitment costs, since eligible pedigrees constitute a smaller proportion of the population. If we

increase the threshold from  $T_1$  to  $T_2$  but the sibship size  $s$  remains unchanged the relative efficiency of the two sampling schemes  $\mathcal{E}(s, T_2) / \mathcal{E}(s, T_1)$  can be shown to be a decreasing function of  $s$ . This implies that for larger sibships there is less gain in efficiency by increasing the ascertainment threshold. Similarly if we leave the threshold  $T$  unchanged, but increase sibship size from  $s_1$  to  $s_2$ , we can show  $\mathcal{E}(s_2, T) / \mathcal{E}(s_1, T)$  to be a decreasing function of  $T$ . Thus a more stringent ascertainment criterion gains less efficiency by using larger sibships. For example, when  $\mu = 0$ ,  $\sigma_Y = 1$ , and  $\rho = 0.25$ , for sib pairs there is a 6.3-fold gain in efficiency by increasing  $T$  from  $-\infty$  (i.e. random sampling) to 2.33; but there is only a 3-fold change when  $s = 5$ . For random sampling there is a 4.6-fold change of efficiency by changing from sib pairs to sibships of size  $s = 5$ , while there is only a 2.5-fold change when  $T = 1.64$ . Hence, whenever a large sibship is found, it should be retained even if it does not meet the ascertainment rule, because larger sibships are much more informative than smaller sibships (Tang & Siegmund, 2001), and ascertainment will not increase their efficiency as much as for smaller sibships.

*Remark.* Similar results hold for ascertainment of sibships with at least one extreme sib, i.e.,  $S = \{\max Y_i > T\}$ ; ascertainment through discordant sibpairs, i.e.,  $S = \{\max_{1 \leq i, j \leq s} |Y_i - Y_j| > T\}$ ; or ascertainment through affected sibpairs, i.e.,  $S = \{Y_{(2)} > T\}$ , where  $Y_{(2)}$  is the second largest phenotype in the sibship. For example, for discordant pair ascertainment, when  $\mu = 0$ ,  $\sigma_Y = 1$ , and  $\rho = 0.25$ , for sib pairs there is an 8.4-fold gain in efficiency by increasing  $T$  from  $-\infty$  to 2.33; but there is only a 2.3-fold change when  $s = 5$ . For random sampling there is a 4.7-fold change in efficiency by changing from sib pairs to sibships of size  $s = 5$ , while there is only a 1.3-fold change when  $T = 2.33$ .

Misspecification of the nuisance parameters is likely to occur when pedigrees are not randomly sampled, especially if one fails to correct for ascertainment or uses estimates obtained outside the current study. Let  $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{\rho})$  denote possibly incorrectly specified (constant) values of the nuisance parameters  $\theta$ . Let  $\hat{\ell}'(0)$  denote  $\ell'(0)$  given in (4) evaluated at  $\hat{\theta}$ . Under local alternatives, as  $\alpha_0 \sim 0$  the large sample noncentrality

parameter depends critically on

$$E_{\alpha_0, \theta}[\hat{\ell}'(0)|\mathcal{A}] = E_{0, \theta}[\hat{\ell}'(0)L(\alpha_0)/L(0)|\mathcal{A}] \\ \sim \alpha_0 E_{0, \theta}[\hat{\ell}'(0)\ell'(0)|\mathcal{A}],$$

and hence equals

$$\xi(\hat{\theta}) = N^{1/2} \alpha_0 \frac{E_{0, \theta}[\hat{\ell}'(0)\ell'(0)|\mathcal{A}]}{\{E_{0, \theta}[\hat{\ell}'^2(0)|\mathcal{A}]\}^{1/2}}.$$

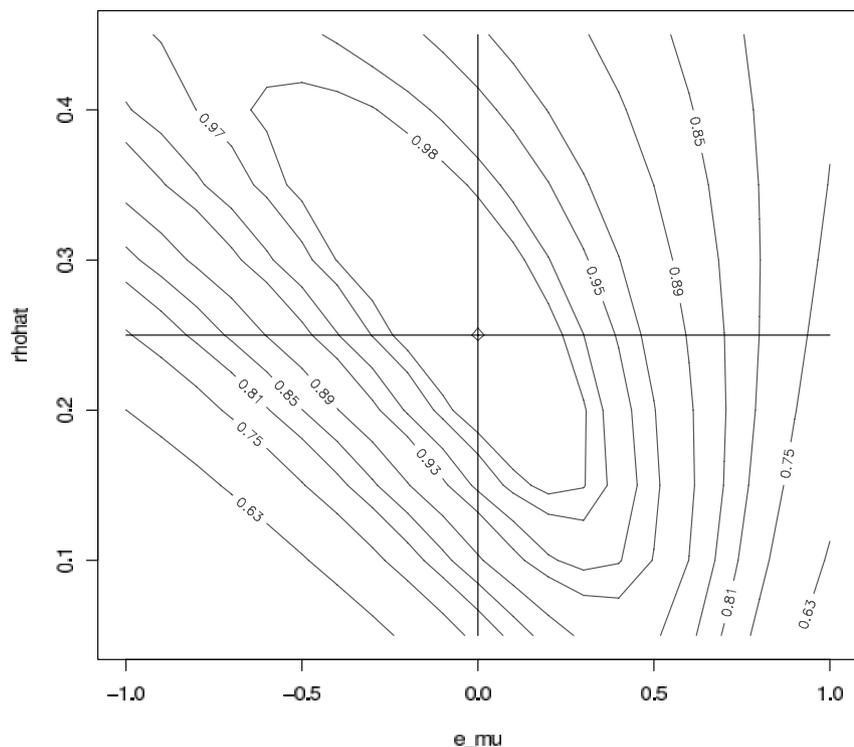
It is easy to see  $[\partial\xi/\partial\hat{\theta}]|_{\hat{\theta}=\theta} = 0$ , and by the Cauchy-Schwarz inequality  $[\partial^2\xi/\partial\hat{\theta}^2]|_{\hat{\theta}=\theta}$  is negative definite. Therefore, asymptotically the true nuisance parameters (or their  $N^{1/2}$ -consistent estimates) yield the largest noncentrality.

For an ascertainment scheme which samples sibships of a fixed size  $s$  with  $Y_1 > T$ , suppose the relative efficiency at  $\hat{\theta}$  is defined as

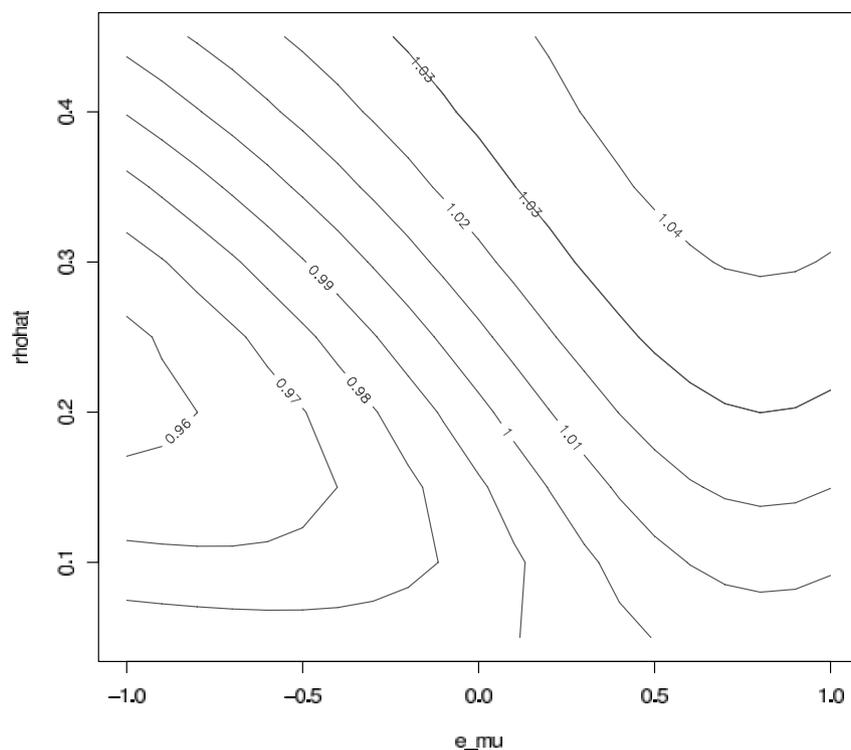
$$R(\hat{\theta}, s, T) = \xi^2(\hat{\theta})/\xi^2(\theta).$$

Note that  $R$  is a function of the relative errors  $e_\mu = (\hat{\mu} - \mu)/\sigma_Y$ ,  $e_\sigma = (\hat{\sigma} - \sigma_Y)/\sigma_Y$ , and of  $\rho$ ,  $\hat{\rho}$

(cf. Peng, 2004 for an explicit formula for  $R$  and other details). Numerical study as well as mathematical analysis shows that  $R$  is an increasing function of  $s$  and  $T$ . This means that larger sibships and/or more stringent ascertainment rules are more robust to misspecification. These findings again suggest the value of large sibships. We also find that the relative efficiency  $R$  is more robust to the specification of  $\sigma_Y^2$  than to that of  $\mu, \rho$ . For example, when  $\mu = 0, \sigma_Y = 1, \rho = 0.25$  and  $T = 1.28$ , for sibships of size  $s = 4$ , if we fix  $\hat{\sigma}$  at the true value 1, and let  $\hat{\mu}, \hat{\rho}$  vary in intervals  $[-1, 1], [0.05, 0.45]$ , respectively, then the relative efficiency varies from 0.63 to 1 (Figure 1). On the other hand, if we set  $\hat{\sigma} = 0.5$ , there is at most a 5% additional change of relative efficiency at each  $\hat{\mu}, \hat{\rho}$ , which suggests that as a function of  $e_\sigma$  the relative efficiency  $R$  is quite flat (Figure 2). Therefore we need to be more careful about the specification of  $\mu$  and  $\rho$ , which are in fact more difficult to estimate as we see below. Again, similar results hold for ascertainment of sibships with at least one phenotype exceeding the



**Figure 1** Contour plot of  $R(\hat{\theta}, 4, 1.28)$ . True parameters are  $\mu = 0, \sigma_Y = 1, \rho = 0.25$ . The x-axis is  $e_\mu$ , the y-axis is  $\hat{\rho}$ .  $\hat{\sigma}$  is fixed at the true value  $\sigma_Y = 1$ . The contour curve is defined by  $R = c$ , and the contour levels  $c$  correspond to the 0, 10, ..., 90, 100 percentiles of  $R$  in the plotted region of  $e_\mu$  and  $\hat{\rho}$ .



**Figure 2** Contour plot of  $\frac{R(\hat{\theta}_1, 4, 1.2, 8)}{R(\hat{\theta}_2, 4, 1.2, 8)}$ . True parameters are  $\mu = 0, \sigma_Y = 1, \rho = 0.25$ . The x-axis is  $e_\mu$ , the y-axis is  $\hat{\rho}$ .  $\hat{\sigma}_1 = \sigma_Y = 1, \hat{\sigma}_2 = 0.5$ . The contour curve is defined by  $\frac{R(\hat{\theta}_1, 4, 1.2, 8)}{R(\hat{\theta}_2, 4, 1.2, 8)} = c$  and the contour levels  $c$  correspond to the 0, 10,  $\dots$ , 90, 100 percentiles of the above ratio.

threshold or for ascertainment based on two-extreme sibs.

## Results

We have already mentioned that for a stringent ascertainment rule the Hopper–Mathews ascertainment correction of conditioning on the probands' phenotypes (i.e. use the conditional likelihood  $L_V$  with  $Y_{(1)}$  being the phenotype(s) of the proband(s)) should be adequate under some mild distributional assumptions. When applying this correction to estimate the nuisance parameters we also assume the multivariate normal distribution as the working model. In this section we use simulations to study the effect of violation of these two assumptions on the estimates and the power to detect linkage. We generate phenotypes of sibships from multivariate normal, multivariate- $t$  and multivariate gamma distributions, and then use three different ascertainment procedures, all of which involve a single proband with

an extreme phenotype, to ascertain the pedigrees falling into the sample. We compare the power of the robust score statistic (10) for three different estimates of the nuisance parameters: (a) the true values; (b) the CMLE based on an assumption of normality and the use of Hopper–Mathews conditional likelihood  $L_V$ ; and (c) the ordinary MLE, as if there had been random ascertainment.

## Ascertainment Procedures

We consider the following three ascertainment procedures:

1. *A-first*: Sample from sibships having the oldest sibling's phenotype above a threshold  $T$ , i.e., the ascertainment rule is  $S = \{Y_1 > T\}$ . For this scheme conditioning on the proband's phenotype is the correct ascertainment correction.
2. *A-max*: Sample from sibships with at least one sibling's phenotype above a threshold  $T$ , i.e., the

ascertainment rule is  $S = \{\max \mathbf{Y} > T\}$ . If more than one sibling has phenotype above the threshold one of them is selected at random to be the proband.

3. *A-sample*: Sample from sibships with at least one “listed” member. Here “listed” means that first s/he is “affected”, which is defined as the following: for a given nominal threshold  $T$ , for each individual, an actual threshold  $T'$  is sampled from  $N(T, \eta^2)$  and this individual is affected if  $Y > T'$ . Then a random fraction  $p_0$  of affected individuals are listed. If more than one sib qualifies as a proband (a “listed” individual), one of them is selected at random to be the proband.

In A-first and A-max “affected” is modelled as having a phenotype  $Y$  greater than the threshold  $T$ . In A-sample it is possible that different thresholds are used for different individuals. This same model for disease status is also proposed by Williams *et al.* (1999), who give a different interpretation/motivation. When  $p_0 = 1$ , A-sample is essentially the same as A-max (except for the definition of affecteds) and when  $p_0$  is very small, A-sample is very close to A-first (single ascertainment according to Ewens & Green, 1988).

### Phenotypic Distributions

To find out how the assumption of normality affects the resulting estimates and power we study phenotypic data generated from three different models. For all three models, given the sibship correlation  $\rho$ , the covariance of the phenotypes of a randomly sampled sibship is given by  $\Sigma = \sigma_Y^2((1 - \rho)I + \rho \mathbf{1} \mathbf{1}^T)$ . The first model is the *Multi-N* model, which is our working model. The second model is the *Multi-T* model, which assumes a multivariate T distribution of the phenotypes. The Multi-N model can be viewed as a special case of this second model with an infinite number of degrees of freedom. The last model is the *Multi-G* model, and under this model the marginal distribution of the phenotype is gamma (up to a location transformation).

For the Multi-G model we assume the major genetic effect  $\alpha_x$  and the residual effect  $e$  in the model (1) to be independent gamma random variables with a common scale parameter (before being standardized to have mean zero). The residual vector  $e = (e_1, \dots, e_s)^T$  in a given

sibship has a multivariate gamma distribution, defined as follows:  $e_i = \gamma_{m_i} + \beta_{f_i} + \epsilon_i$ , where  $m_i, f_i$  are i.i.d with  $P(m_i = 1) = P(m_i = 2) = 1/2$ ;  $\gamma_1, \gamma_2, \beta_1, \beta_2$  are i.i.d. from  $\text{gamma}(a_1, b)$ ; and  $\epsilon_i$  are i.i.d from  $\text{gamma}(a_2, b)$ . It is easy to verify that the marginal distribution of  $e$  is  $\text{gamma}(a, b)$  and  $\text{Corr}(e_i, e_j) = r$ , where  $a = 2a_1 + a_2$ ,  $r = a_1/a$ . For a given constant  $c > 0$ , if we choose

$$a = \frac{c}{\sigma_Y^2/\alpha_0 - 2}, \quad b = \left[ \frac{\alpha_0}{a} \right]^{1/2}, \quad r = \frac{\sigma_Y^2 \rho / \alpha_0 - 1}{\sigma_Y^2 / \alpha_0 - 2},$$

then  $\text{Var}(Y) = \sigma_Y^2$ ,  $\text{Corr}(Y_i, Y_j) = \rho$  and  $\text{Var}(\alpha_x) = \alpha_0$ . The ‘shape’ parameter  $c$  controls the skewness of the phenotypic distribution.

The Multi-T model works as an example of heavy-tailed traits, where the kurtosis of the trait is controlled by an additional parameter: the degrees of freedom  $k$ . The Multi-G model serves as an example of skewed traits. For these experiments the parameters of the Multi-T and Multi-G models were chosen to provide some sense of the effects of non-normality on the ascertainment corrections, without departing so far from normality that it would obviously be necessary to use a different statistic. For a more detailed description of these models, see Peng (2004).

### Numerical Results

To generate phenotypic data we have simulated sib-trios under the above three phenotypic models with the population parameters  $\mu = 0, \sigma_Y = 1, \rho = 0.25$ , and the linkage parameters  $\alpha_0 = 0.1, \delta_0 = 0$ . Four hundred sib-trios are then ascertained according to the ascertainment rules A-first, A-max and A-sample ( $\eta = 0.2, p_0 = 0.2$ ) with thresholds  $T$  being the 75th, 90th, 95th quantiles of the phenotypic distribution. For the Multi-T model the degrees of freedom are set to be  $k = 20$ , which results in a phenotypic kurtosis of 0.375. For the Multi-G model the “shape” parameter is set to be  $c = 100$ , which corresponds to a marginal distribution that is gamma (up to a location transformation) with shape parameter 125 and scale parameter 0.089, thus skewness 0.179 and kurtosis 0.048.

For genotypic data we use 31 equally spaced, fully informative markers on an idealized human chromosome of length 150cM, and locate the trait locus  $\tau$  on the 16th marker. The maxima of the statistics (10)

over all markers is taken as the test statistic for linkage. Power is defined as detection of linkage anywhere on the correct chromosome. It is calculated at a genome-wide 0.05 significance level, so the significance level for a single chromosome is about  $0.05/23 \approx 0.0022$ . The rejection threshold  $b = 3.809$  comes from the approximation with adjustment for skewness given in Tang & Siegmund (2001). This threshold has been checked for accuracy by simulations (data not shown).

After generating the data we apply three procedures: (a) true values, (b) CMLE, and (c) ordinary MLE for random ascertainment to estimate the nuisance parameters, which are plugged into (10) to get the test statistic. One thing we want to emphasize is that  $T$ ,  $T'$ ,  $\eta$  and  $p_0$  are model parameters which are only used to ascertain pedigrees, and are not used after data are generated. The information available for the CMLE procedure are the ascertained sibships, their phenotypes, and the identity of the proband. This is the information that we would have in practice.

As can be seen from Table 1, for the multivariate normal trait, using the CMLE incurs essentially no loss of power for the ascertainment criteria A-first and A-sample, while there is a small loss of power for A-max, especially when the ascertainment threshold  $T$  is small. This is consistent with our expectation, since under normality the CMLE is consistent when using A-first for ascertainment. A-sample is stringent when the value of  $p_0$  is small, and hence it is well approximated by the ascertainment rule A-first. In contrast, A-max is not stringent unless the threshold  $T$  is large (e.g., when  $T$  is the 95th quantile of the phenotype distribution, the ascertainment rate for sib-trios under A-max is 13%); but we observe only a 1% – 3% loss of power. We also find that the CMLE does not cost additional loss of power for the two non-normal traits under these three ascertainment rules. We do observe a large drop of power when no ascertainment correction is made and the unconditional MLE is used.

In Table 2 we report the mean and standard deviation of the CMLEs, which are estimated via 5000 iterations.

As can be seen from Table 2, for the multivariate normal trait, reasonably unbiased estimates are obtained under the ascertainment criteria A-first and A-sample. Under A-max an increase in the ascertainment threshold  $T$  results in a less biased estimate for the phenotypic

mean  $\mu$ , which is consistent with our expectation. Similar results are observed for the two non-normal traits.

We have also considered extremely non-normal versions of the Multi-T and Multi-G models. For the Multi-T model we use  $k = 4$  to generate data, which gives a trait with kurtosis  $\infty$ . We still find that using the CMLE results in only a small loss of power compared with using the true nuisance parameters. Similarly, for the Multi-G model if we use  $c = 0.1$  we get a phenotypic distribution with skewness 5.66 and kurtosis 30, and again the statistic based on the CMLE has about the same power as if the segregation parameters had been known. However, in these cases the power of the score statistic (10) based on the working normality assumption is very low, so one would presumably want to find a more satisfactory statistic, perhaps by data transformation, before worrying about the issue of ascertainment. See Peng (2004) for references and a comparative study of a number of methods for transformations.

## Discussion

In this paper we have discussed a variance component model of linkage analysis to map quantitative traits by using sibships ascertained through an arbitrary number of probands. The method can be applied to more general pedigrees.

We have proposed using an ascertainment correction by conditioning on probands' phenotypes (CMLE), and then using the score statistic (10) derived from the normality assumption. Exact conditions under which these corrections are successful are difficult to specify. We have introduced the informal criterion that an ascertainment rule is "stringent" if it is unlikely for there to be two different subsets of a pedigree eligible to serve as probands. Although ill-defined if the ascertainment rule involves all the members in the pedigree, when ascertainment is stringent the CMLE procedure leads to only a small bias in parameter estimation. Theoretical and simulation results have also been developed in Peng (2004) for more general situations involving multiple probands, such as ascertainment of pedigrees with at least  $r$  extreme members for  $1 \leq r < s$ , and ascertainment of pedigrees with at least one pair of discordant siblings, and similar conclusions have been obtained. When the ascertainment

rule is less stringent CMLE estimators can be biased, but our test statistic still has approximately the same power as knowing the true values of the nuisance parameters, and substantially more power than estimating those parameters without an ascertainment correction.

Motivated by the analysis of Tang & Siegmund (2001) showing the rapidly increasing power of large sibships as a function of sibship size, we have shown through analysis of noncentrality parameters and numerical calculations that ascertainment through siblings having extreme phenotypes can increase the power per ascertained sibship, but that this beneficial effect decreases as the sibship size increases, while the power of randomly selected sibships increases with size. A reasonable overall strategy to minimize genotyping costs would involve obtaining large sibships having, say, four or more siblings, without regard to an ascertainment criterion, while applying an ascertainment criterion to sib pairs and trios.

We have assumed that, given the IBD sharing matrix and the covariates (if any), the phenotypes of siblings follow a multivariate normal distribution. This normality assumption yields the efficient score (4) and plays an important role in the ascertainment corrections, which involve the conditional distribution of the phenotypes of the non-probands given those of the probands. We have shown by simulation that for mildly non-normal phenotype distributions the ascertainment correction based on the assumption of normality has only a small effect on the power to detect linkage (Table 1). We have also found that when the traits are very non-normal, so the power of the score statistic derived under the normal model is very low even when the nuisance parameters are known, the ascertainment correction based on the CMLE still yields comparable power. However, in this case one would ideally first of all want to find a suitable statistic for the non-normal data and then to consider adjustment for ascertainment.

We have also shown that the statistic (10) is robust in the sense that when an estimator  $\hat{\theta}$  converges to a limit  $\theta^*$  as the sample size goes to  $\infty$ , the statistic has the same limiting distribution as if the limiting value  $\theta^*$  is known and used. This result does not depend on the normality assumption. Therefore if a  $N^{1/2}$ -consistent estimator of the nuisance parameters is used, asymptotically we will not lose efficiency due to parameter

estimation. This consistent estimator is not required to be maximum likelihood.

## Acknowledgements

This research was partially supported by NIH Grant RO1 HG00848 and by a Stanford Graduate Fellowship.

## References

- Andrade, M., Amos, C. I. & Thiel, T. J. (1999) Methods to estimate genetic components of variance for quantitative traits in family studies. *Genet Epidemiol* **17**, 64–76.
- Andrade, M. & Amos, C. I. (2000) Ascertainment issues in variance component models. *Genet Epidemiol* **19**, 333–344.
- Beaty, T. H. & Liang, K. Y. (1987) Robust inference for variance component models in families ascertained through probands: I. conditioning on the proband's phenotype. *Genet Epidemiol* **4**, 203–210.
- Cox, D. R. & Hinkley, D. V. (1974) *Theoretical Statistics*. Chapman and Hall, London.
- Elston, R. C. & Sobel, E. (1979) Sampling considerations in the gathering and analysis of pedigree data. *Am J Hum Genet* **31**, 62–69.
- Ewens, W. J. & Green, R. M. (1988) A resolution of the ascertainment sampling problem: IV. continuous phenotypes. *Genet Epidemiol* **5**, 433–444.
- Feingold, E., Brown, P. O., & Siegmund, D. (1993) Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent. *Am J Hum Genetics* **53**, 234–251.
- Fisher, R. A. (1918) The correlation of relatives on the assumption of Mendelian inheritance. *Proc Roy Soc Edinburgh*.
- Hopper, J. L. & Mathews, J. D. (1982) Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* **46**, 373–383.
- Kempthorne, O. (1957) *Genetic Statistics*, John Wiley and Sons, New York.
- Peng, J. (2004) Score statistics to map genes in humans. PHD thesis. Stanford University, USA.
- Peng, J. & Siegmund, D. (2004) Mapping quantitative traits with random and with ascertained sibships. *Proc Natl Acad Sci USA* **101**, 7845–7850.
- Peng, J., Tang, H. K., & Siegmund, D. (2005) Genome scans with gene-covariate interaction. *Genet Epidemiol* **29**, 173–184.
- Risch, N. & Zhang, H. (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589.
- Sham, P. C., Purcell, S., Cherny, S. S., & Abecasis, G. R. (2002) Powerful regression-based quantitative-trait analysis of general pedigrees. *Am J Hum Genet* **71**, 238–253.

- Tang, H.-K. & Siegmund, D. (2001) Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2**, 147–162.
- Tang, H.-K. & Siegmund, D. (2002) Mapping multiple genes for complex or quantitative traits. *Genet Epidemiol* **19**, 313–327.
- T. Cuenco, T. K., Szatkiewicz, J. P. & Feingold, E. (2003) Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. *Am J Hum Genet* **73**, 863–873.
- Wang, K. (2005) A likelihood approach for quantitative-trait-locus mapping with selected pedigrees. *Biometrics* **61**, 465–473.
- Wang, K. and Huang J. (2002) A score-statistic approach for the mapping of quantitative-trait loci with sibships of arbitrary size. *Am J Hum Genet* **70**, 412–424.
- Williams, J., Eerdewegh, P. V., Almasy, L. & Blangero, J. (1999) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. likelihood formulation and simulation results. *Am J Hum Genet* **65**, 1134–1147.

Received: 2 October 2005

Accepted: 15 February 2006