Genetic Epidemiology 30: 471–484 (2006)

Estimating the Power of Variance Component Linkage Analysis in Large Pedigrees

Wei-Min Chen* **and Gonçalo R. Abecasis**

Department of Biostatistics, University of Michigan, Ann Arbor, MI

Variance component linkage analysis is commonly used to map quantitative trait loci (QTLs) in general pedigrees. Large pedigrees are especially attractive for these studies because they provide greater power per genotyped individual than small pedigrees. We propose accurate and computationally efficient methods to calculate the analytical power of variance component linkage analysis that can accommodate large pedigrees. Our analytical power computation involves the approximation of the noncentrality parameter for the likelihood-ratio test by its Taylor expansions. We develop efficient algorithms to compute the second and third moments of the identical by descent (IBD) sharing distribution and enable rapid computation of the Taylor expansions. Our algorithms take advantage of natural symmetries in pedigrees and can accurately analyze many large pedigrees in a few seconds. We verify the accuracy of our power calculation via simulation in pedigrees with 2–5 generations and 2–8 siblings per sibship. We apply this proposed analytical power calculation to 98 quantitative traits in a cohort study of 6,148 Sardinians in which the largest pedigree includes 625 phenotyped individuals. Simulations based on eight representative traits show that the difference between our analytical estimation of the expected LOD score and the average of simulated LOD scores is less than 0.05 (1.5%). Although our analytical calculations are for a fully informative marker locus, in the settings we examined power was similar to what could be attained with a single nucleotide polymorphism (SNP) mapping panel (with >1 SNP/cM). Our algorithms for power analysis together with polygenic analysis are implemented in a freely available computer program, POLY. *Genet. Epidemiol.* 30:471–484, 2006.

Key words: analytical calculation; pedigree analysis; algorithm; polygenic analysis; QTL mapping

Contract grant sponsor: National Institutes of Health; Contract grant numbers: EY10562; HG02651; Contract grant sponsor: National Institutes of Aging; Contract grant numbers: 263-MA-410953.
*Correspondence to: Wei-Min Chen, Ph.D., Center for Statistical Genetics, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: wechen@umich.edu
Received 18 November 2005; Revised 21 February 2006; Accepted 22 March 2006
Published online 9 May 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20160

INTRODUCTION

Variance component linkage analysis is commonly used to search for quantitative trait loci (QTLs) using pedigree data, in data sets that range from sibpairs to very large extended pedigrees. The likelihood-based variance component linkage analysis [Amos, 1994; Almasy and Blangero, 1998] can offer more power and flexibility than the Haseman-Elston method [Haseman and Elston, 1972]. Although the normality assumption can lead to inflated type I error rates [Allison et al., 1999], the disadvantage of distributional assumption can be overcome by various robustness techniques [Chen et al., 2004, 2005]. Power calculation is the key for study design, both for linkage studies and for association studies where one wishes to apply ''causality'' tests [Fulker et al., 1999; Cardon and Abecasis, 2000].

Analytical power calculations have been used to evaluate different study designs in quantitative trait linkage analysis and to show, for example, that large sibships provide more power per individual than smaller ones [e.g., Dolan et al., 1999], to demonstrate the power of multivariate trait QTL linkage analysis [Evans, 2002] and to compare variance component analysis to alternative approaches [e.g., Visscher and Hopper, 2001]. An accurate and efficient method of power calculation for general pedigrees is crucial for the analysis of variance component models.

Williams and Blangero [1999] proposed an analytical approach to calculate the power of variance component linkage analysis. They also

demonstrated the analytical power of variance-component based discrete trait linkage analysis for relative pairs [2004]. Rijsdijk et al. [2001] presented a general expression for the "exact" power calculation for the variance component linkage analysis and provided approximations to the power calculation for small pedigrees. Tang and Siegmund [2001] derived the power calculation formulas for a score test that is derived from the variance component model.

Despite these advances, a number of limitations remain in existing methods for power calculation. First, almost all current approches for analytical power calculations are limited to small or moderate-sized pedigrees. When pedigrees are large, existing strategies are not feasible. In this study, we present efficient algorithms that can be applied to arbitrarily large pedigrees. Second, many of existing approaches of analytical power calculation were not accurate. Rijsdijk et al. [2001] point out Williams and Blangero [1999]'s formula does not give the correct noncentrality parameter (NCP) for many types of relatives. Unfortunately, Rijsdijk et al.'s [2001] improved approximation can only be evaluated for small pedigrees. An accurate and efficient approach to compute the analytical power for general pedigrees is presented here. Third, since the variance component test statistic does not always follow a noncentral chi-square distribution, comparing the average of test statistics alone may not be sufficient to evaluate analytical power formulas [Yu et al., 2004]. Hence, we evaluate the performance of our approximation in terms of both the analytical statistical power and the expected LOD (ELOD) scores, and compare our results to computationally intensive simulations.

In the following sections, we first propose several efficient methods to analytically compute the power of likelihood-based variance component linkage analysis in unascertained pedigrees. Then we assess the accuracy of our algorithms via computer simulations. Finally, we apply our analytical power calculation to a data set with 6,148 phenotyped individuals in which the largest pedigree includes 625 phenotyped individuals [Pilia et al., 2006].

METHODS

Variance component models [Hopper and Mathews, 1982; Lange and Boehnke, 1983; Amos, 1994] allow maximum likelihood estimation of the contribution of genes and environment to variance, under the assumption that trait values across the pedigree follow a multivariate normal distribution. Let Ω denote the covariance matrix for trait values. Assume the covariance matrix under the null hypothesis of no linkage is Ω_0. Let σ_a^2 and σ^2 denote the additive genetic variance due to the major QTL and the total variance, respectively. In an additive model, the covariance between measurements for individuals i and j in a given non-inbred pedigree is $(\Omega_0)_{ij} + (\pi_{ij} - 2\phi_{ij})\sigma_a^2$ [Chen et al., 2005], where ϕ_{ij} and π_{ij} denote the kinship coefficient and the expected proportion of alleles shared identical by descent (IBD) for individuals i and j, respectively. Let k index families and vector y_k denote trait values of all phenotyped individuals in the kth family. Without loss of generality, we assume the trait mean is 0. The test statistic for likelihood ratio test (LRT) is

$$T^{LRT} = \sum_k \ln|\Omega_0^{(k)}| + \sum_k y_k'(\Omega_0^{(k)})^{-1}y_k - \sum_k \ln|\Omega^{(k)}| - \sum_k y_k'(\Omega^{(k)})^{-1}y_k$$

where all parameters are evaluated at their MLEs. This LRT statistic is distributed as a 50:50 mixture of 0: $\chi^2(1)$ under the null hypothesis of no linkage, and approximately a noncentral chi-square under the alternative hypothesis of linkage [Williams and Blangero, 1999]. If the NCP is known, the statistical power can be calculated directly from the distribution of the noncentral chi-square [e.g., Chen and Deng, 2001]. The expected LOD, or ELOD, which is one measure of the power to detect linkage, can be calculated analytically as

$$ELOD = \left(1 + \sum_k NCP_k\right) \Big/ 2\log(10).$$

Thus, the calculation of NCP of each pedigree is crucial in the power analysis of the variance component models. Rijsdijk et al. [2001] show NCP of LRT for one family is

$$NCP = \ln|\Omega_0| - E[\ln|\Omega|] \quad (1)$$

where $E[\ln|\Omega|]$ is the expectation over all possible realization of allele-sharing coefficients at a fully informative marker closely linked to the QTL of interest. Note that, when marker data are not fully informative about IBD sharing, NCPs will necessarily be smaller than those given in (1).

Although our analytical derivations focus on the setting of full information about IBD, we contrast our analytical results with settings with incomplete information about IBD using Simulations. When the size of pedigree is small (e.g., $<\sim 14$ individuals), $E[\ln|\Omega|]$ can be computed by enumerating all possible inheritance vectors, which is known as "exact" calculation [Rijsdijk et al., 2001]. When the pedigree is large, exact evaluation of $E[\ln|\Omega|]$ is not feasible because the size of inheritance vector space increases at an exponential rate with pedigree size. One possible solution is to break large pedigrees into many smaller families. Although computationally convenient, this solution discards information from more distant relative pairs and could lead to substantial underestimates of the power of linkage analysis. Our strategy to simplify the computation of $E[\ln|\Omega|]$ for large pedigrees is to approximate $E[\ln|\Omega|]$ by its Taylor expansions.

SECOND-ORDER APPROXIMATION

A Taylor expansion of $\ln|\Omega|$ to the second order yields

$$\begin{aligned}\ln|\Omega| &\approx \ln|\Omega_0| + \frac{\partial \ln|\Omega|}{\partial \sigma_a^2}\Big|_{\sigma_a^2=0}\sigma_a^2 + \frac{1}{2}\frac{\partial^2 \ln|\Omega|}{\partial \sigma_a^4}\Big|_{\sigma_a^2=0}\sigma_a^4 \\ &= \ln|\Omega_0| + Tr\left(\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\right)\sigma_a^2 \\ &\quad - \frac{1}{2}Tr\left(\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\right)\sigma_a^4\end{aligned}$$

where off-diagonal element (i, j) of matrix $\partial\Omega/\partial\sigma_a^2$ is $\pi_{ij} - E\pi_{ij}$ and all diagonal elements are 0. To simplify notation, we use $\tilde{\pi}_{ij}$ to denote $\pi_{ij} - E\pi_{ij}$. Note $E[\tilde{\pi}_{ij}] = 0$ leads to $E[\partial\Omega/\partial\sigma_a^2] = 0$. Thus, according to equation (1), the NCP can be approximated as

$$\begin{aligned}NCP &\approx \frac{1}{2}Tr\left(E\left[\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\right]\right)\sigma_a^4 \\ &\quad - Tr\left(E\left[\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\right]\right)\sigma_a^2 \qquad (2) \\ &= \frac{1}{2}\sigma_a^4\sum_{a,b,c,d}(E\tilde{\pi}_{ab}\tilde{\pi}_{cd})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{da}\end{aligned}$$

where notation $Tr(\cdot)$ denotes the trace function which sums up all diagonal elements of a matrix, and $(\Omega_0^{-1})_{bc}$ denotes element (b, c) of the inverse of matrix Ω_0. This algorithm (denoted as Approx 2) involves the calculation of $E\tilde{\pi}_{ab}\tilde{\pi}_{cd}$, or the covariance of two allele-sharing coefficients, and thus the computation of analytical power converts to the computation of the covariance of allele-sharing coefficients. Although formula (2) gives identical results to the "exact" power calculation of Williams and Blangero [1999], it is important to note that it corresponds to a second-order Taylor approximation of the NCP. In fact, this formula and the approach of Williams and Blangero [1999] can produce inaccurate results in some settings [Rijsdijk et al., 2001] and should not be treated as exact.

The calculation of the covariance of allele-sharing coefficients can be accomplished using generalized kinship coefficients [see Lange, 2002]. Our method can be extended to accommodate any type of pedigree, but for simplicity, we restrict our attention to non-inbred pedigrees. Our implementation and the formulas presented here apply to any non-inbred pedigree, with or without marriage loops and multiple matings. We define an ordering for individuals in the pedigree where $a > b$ implies person a is not an ancestor of person b. We assume that individuals are ordered such that $a \geq \max(b, c, d)$ and $c \geq d$. Further, we let p and q be parents of person a. Then we apply the following iterative procedure to estimate $E\tilde{\pi}_{ab}\tilde{\pi}_{cd}$:

(1) If π_{ab} (or π_{cd}) is a constant (e.g., if a and b are a parent-offspring pair or are unrelated), then $E\tilde{\pi}_{ab}\tilde{\pi}_{cd} = 0$
(2) If $a > c$, then $E\tilde{\pi}_{ab}\tilde{\pi}_{cd} = (E\tilde{\pi}_{pb}\tilde{\pi}_{cd} + E\tilde{\pi}_{qb}\tilde{\pi}_{cd})/2$
(3) If $a = c$, then

$$\begin{aligned}E\tilde{\pi}_{ab}\tilde{\pi}_{ad} &= (E\tilde{\pi}_{pb}\tilde{\pi}_{qd} + E\tilde{\pi}_{qb}\tilde{\pi}_{pd})/4 \\ &\quad + \phi_{pbd} + \phi_{qbd} - \phi_{pb}\phi_{pd} - \phi_{qb}\phi_{qd}\end{aligned}$$

where ϕ_{pb} is the kinship coefficient for individuals p and b, and ϕ_{pbd} is the probability that three alleles drawn at random from each person in the set $\{p, b, d\}$ are IBD, and can be calculated using standard procedures [see Lange, 2002 or Appendix A]. Note that steps (2) and (3) are recurrence rules where individual a is replaced by its parents in further evaluations. The proof of the above algorithm is given in Appendix B.

FURTHER SIMPLIFICATION FOR SIBSHIP DATA

In the case of sibship data, where any two distinct allele-sharing coefficients are uncorre-

lated, the NCP approximation based on formula (2) can be simplified. Suppose the size of a sibship is s and the correlation between any two siblings is ρ. Let J denote a matrix consisting of 1's, and I denote an identity matrix. The inverse of the covariance matrix under the null hypothesis of no linkage is

$$\begin{aligned}\Omega_0^{-1} &= ((1-\rho)I + \rho J)^{-1}/\sigma^2 \\ &= \frac{1}{(1-\rho)\sigma^2}\left(I - \frac{\rho}{1+\rho(s-1)}J\right)\end{aligned}$$

i.e., element (u, v) of Ω_0^{-1} is

$$\frac{1+\rho(s-2)}{(1-\rho)(1+\rho(s-1))\sigma^2} \text{ when } u = v \text{ and}$$
$$\frac{-\rho}{(1+\rho(s-1))(1-\rho)\sigma^2}$$

otherwise. Since $E\tilde{\pi}_{ab}\tilde{\pi}_{cd}$ equals 1/8 when $(a, b) = (c, d)$ and 0 otherwise, for the case of sibships, formula (2) simplifies to

$$\begin{aligned}NCP &\approx \sigma_a^4 \sum_{a>b} E\tilde{\pi}_{ab}^2((\Omega_0^{-1})_{aa}(\Omega_0^{-1})_{bb} + (\Omega_0^{-1})_{ab}^2) \\ &= \frac{s(s-1)((1+(s-2)\rho)^2+\rho^2)}{16(1-\rho)^2(1+(s-1)\rho)^2}\left(\frac{\sigma_a^2}{\sigma^2}\right)^2.\end{aligned} \quad (3)$$

Formula (3) for sibship data has been reported previously [e.g., Tang and Siegmund, 2001].

THIRD-ORDER APPROXIMATION

We can further improve the accuracy of NCP approximation by considering higher orders of the Taylor expansion. The Taylor expansion of $\ln|\Omega|$ to the third order is

$$\begin{aligned}\ln|\Omega| \approx \ln|\Omega_0| &+ Tr\left(\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\right)\sigma_a^2 \\ &- \frac{1}{2}Tr\left(\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\right)\sigma_a^4 \\ &+ \frac{1}{3}Tr\left(\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\frac{\partial\Omega}{\partial\sigma_a^2}\Omega_0^{-1}\right)\sigma_a^6\end{aligned}$$

and the NCP (Approx 3) is

$$\begin{aligned}NCP \approx &\frac{1}{2}\sigma_a^4 \sum_{a,b,c,d}(E\tilde{\pi}_{ab}\tilde{\pi}_{cd})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{da} \\ &- \frac{1}{3}\sigma_a^6 \sum_{a,b,c,d,e,f}(E\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{de}(\Omega_0^{-1})_{fa}.\end{aligned} \quad (4)$$

Again, we assume that individuals are ordered such that $a \geq \max(b, c, d, e, f)$, $c \geq d$, $e \geq f$, and let p and q denote the parents of person a. $E[\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef}]$ can be calculated recursively as follows:

(1) If π_{ab},π_{cd} or π_{ef} is a constant, then $E[\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef}] = 0$
(2) If $a > c$ and $a > e$, then $E[\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef}] = (E[\tilde{\pi}_{pb}\tilde{\pi}_{cd}\tilde{\pi}_{ef}] + E[\tilde{\pi}_{qb}\tilde{\pi}_{cd}\tilde{\pi}_{ef}])/2$
(3) If $a = c > e$, then

$$\begin{aligned}E[\tilde{\pi}_{ab}\tilde{\pi}_{ad}\tilde{\pi}_{ef}] &= (E[\tilde{\pi}_{pb}\tilde{\pi}_{qd}\tilde{\pi}_{ef}] + E[\tilde{\pi}_{qb}\tilde{\pi}_{pd}\tilde{\pi}_{ef}] \\ &+ E[\tilde{\pi}_{pbd}\tilde{\pi}_{ef}] + E[\tilde{\pi}_{qbd}\tilde{\pi}_{ef}])/4 \\ &- (E[\tilde{\pi}_{pb}\tilde{\pi}_{ef}]\phi_{pd} + E[\tilde{\pi}_{qb}\tilde{\pi}_{ef}]\phi_{qd} + E[\tilde{\pi}_{pd}\tilde{\pi}_{ef}]\phi_{pb} \\ &+ E[\tilde{\pi}_{qd}\tilde{\pi}_{ef}]\phi_{qb})/2\end{aligned}$$

(4) If $a = c = e$, then

$$\begin{aligned}E[\tilde{\pi}_{ab}\tilde{\pi}_{ad}\tilde{\pi}_{af}] &= (E[\pi_{pbd}\pi_{qf}] + E[\pi_{qbd}\pi_{bf}] \\ &+ E[\pi_{pbf}\pi_{qd}] + E[\pi_{qbf}\pi_{pd}] \\ &+ E[\pi_{pdf}\pi_{qb}] + E[\pi_{qdf}\pi_{pb}])/8 + \phi_{pbdf} + \phi_{qbdf} \\ &- \phi_{ab}\phi_{ad}\phi_{af} \\ &- (E[\tilde{\pi}_{ab}\tilde{\pi}_{ad}]\phi_{af} + E[\tilde{\pi}_{ad}\tilde{\pi}_{af}]\phi_{ab} + E[\tilde{\pi}_{ab}\tilde{\pi}_{af}]\phi_{ad})/4\end{aligned}$$

where the definitions and algorithms for $E\tilde{\pi}_{abc}\tilde{\pi}_{de}$ and the kinship coefficient ϕ_{abcd} are given in Appendix A. The proof of this algorithm is given in Appendix B.

HIGHER-ORDER APPROXIMATION

The kth-order ($k > 3$) approximation of the NCP can be obtained via the following formula:

$$\begin{aligned}NCP = &\frac{1}{2}\sigma_a^4 \sum_{a,b,c,d}(E\tilde{\pi}_{ab}\tilde{\pi}_{cd})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{da} \\ &- \frac{1}{3}\sigma_a^6 \sum_{a,b,c,d,e,f}(E\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{de}(\Omega_0^{-1})_{fa} \\ &+ \cdots \\ &+ (-1)^k\frac{1}{k}\sigma_a^{2k} \sum_{i_1,i_2\cdots i_{2k-1},i_{2k}}(E\tilde{\pi}_{i_1i_2}\tilde{\pi}_{i_3i_4}\cdots\tilde{\pi}_{i_{2k-1}i_{2k}}) \\ &(\Omega_0^{-1})_{i_2i_3}\cdots(\Omega_0^{-1})_{i_{2k-2}i_{2k-1}}(\Omega_0^{-1})_{i_{2k}i_1}.\end{aligned}$$

In practice, the computation of higher-order moments for IBD coefficients becomes progressively cumbersome. Thus, although these higher-order approximations should be more accurate, in this paper, we only consider second and third-order approximations of the NCP. Because of the nonnegative definite property of the covariance matrix Ω, when genetic effect of the QTL is small enough, the exact NCP in expression (1) should be

between approximations (2) and (4), so that, power calculations based on the second-order Taylor expansion will overestimate the power, and the power calculations based on the third-order Taylor expansion will underestimate the power. To further improve the precision, it should be possible to construct an estimate of the NCP that is intermediate between the second- and third-order Taylor expansions. This approximation could be defined as (Approx 3′):

$$NCP \approx \frac{1}{2}\sigma_a^4 \sum_{a,b,c,d} (E\tilde{\pi}_{ab}\tilde{\pi}_{cd})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{da} - K\sigma_a^6 \sum_{a,b,c,d,e,f} (E\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{de}(\Omega_0^{-1})_{fa}. \quad (5)$$

Here, K should vary between 0 (which gives the second-order approximation) and 1/3 (which gives the third-order approximation). In the settings we examined, $K = 1/4$ appears to perform well in comparison to power estimates derived from simulation, but we cannot guarantee that this factor will be appropriate for every data set. In the results, we present empirical data to support our choice of K and consider other choices for K.

EFFICIENT ALGORITHMS FOR POWER CALCULATION

Both accuracy and efficiency in terms of computer running time are crucial for a practical power calculation algorithm. Here we summarize strategies to further improve the computational efficiency of our proposed power calculation. To simplify the presentation, we detail strategies for improving evaluation of the second-order approximation in equation (2). Nevertheless, the strategies are general and can be applied to higher order approximations.

First note that for the purpose of power calculation, Ω_0^{-1}, the inverse of the variance-covariance matrix under the null hypothesis of no linkage, only needs to be computed once. Thus the majority of computational effort is expended on the summation over allele-sharing coefficients among different individuals. One simple way to improve the efficiency of power approximations is to rearrange elements in the power calculation formulas (2), (4) and (5). Formula (2) for the second order $NCP \approx$ becomes

$$\begin{aligned}
&\frac{1}{2}\sigma_a^4 \sum_{a,b,c,d} (E\tilde{\pi}_{ab}\tilde{\pi}_{cd})(\Omega_0^{-1})_{bc}(\Omega_0^{-1})_{da} \\
&= \sigma_a^4 \sum_{a>b} \sum_{c>d} (E\tilde{\pi}_{ab}\tilde{\pi}_{cd})\Big((\Omega_0^{-1})_{ac}(\Omega_0^{-1})_{bd} + (\Omega_0^{-1})_{ad}(\Omega_0^{-1})_{bc}\Big) \\
&= \sigma_a^4 \sum_{a>b} (E\tilde{\pi}_{ab}^2)\Big((\Omega_0^{-1})_{aa}(\Omega_0^{-1})_{bb} + (\Omega_0^{-1})_{ab}^2\Big) \\
&+ 2\sigma_a^4 \sum_{\substack{a>b,c>d \\ (a,b)>(c,d)}} (E\tilde{\pi}_{ab}\tilde{\pi}_{cd})\Big((\Omega_0^{-1})_{ac}(\Omega_0^{-1})_{bd} + (\Omega_0^{-1})_{ad}(\Omega_0^{-1})_{bc}\Big).
\end{aligned} \quad (6)$$

The efficiency gain of expression (6) over (2) is roughly 4-fold. Similarly, this algebra applies to the third-order approximation (4) and (5) and there the efficiency gain can be as large as 48-fold. Excluding relative pairs with constant allele-sharing coefficients (such as parent-offspring pairs) in the above calculations further reduces the amount of computation. This strategy is general and suitable for any pedigree.

Additional efficiencies are possible in large pedigrees, where we might have to calculate moments of allele-sharing coefficients for many sets of 2, 3 and 4 individuals. In this case, many sets of individuals will have identical moments of allele-sharing coefficients. For example, in the pedigree in Figure 1 there are four cousin pairs (7–9, 8–9, 7–10 and 8–10) and the moments of allele-sharing coefficients are the same for all four. In fact, although there are ${}^7C_2 = 21$ distinct pairs of individuals in the pedigree, they can be organized into five equivalence groups (corresponding to sib-pairs, parent-offspring pairs, grand-parent grand-child pairs, avuncular pairs and cousin pairs, respectively). Identifying equivalent sets of individuals and calculating the moments of allele-sharing coefficients for only one representative for each group of equivalent sets can produce large computational savings, and still produce identical results for power calculations.

Identifying all equivalent subsets of 2, 3 and 4 individuals is challenging to do in general pedigrees, but we have implemented a recursive algorithm that can identify many equivalent subsets of individuals in tree-like pedigrees (pedigrees where all matings except one are between a founder and a nonfounder, such as the pedigrees in Figures 1 and 2). We skip the detailed algorithm due to its complexity. Although our implementation might not be optimal and does not identify all equivalent subsets of

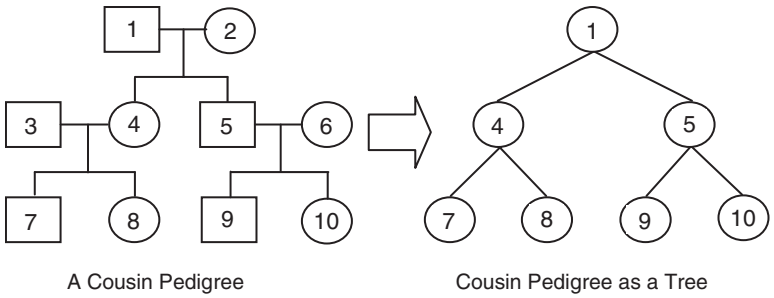

Fig. 1. Structure of the cousin pedigree.

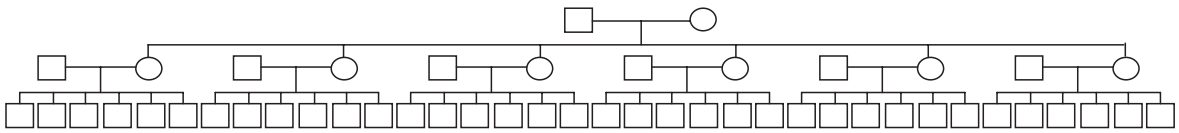

Fig. 2. An extended pedigree with three generations and six siblings in each sibship.

individuals, it already results in substantial computational savings and the savings increase with pedigree size. Results remain identical to those using a naïve implementation of Formula (6).

The benefits of the tree-based algorithm increase for pedigrees with larger sibships and/or more generations. Figure 2 shows a tree-like pedigree with three generations and six siblings in each sibship. For the example of this large pedigree with 50 individuals, our tree-based algorithm requires calculation of 252 $E\tilde{\pi}_{ab}\tilde{\pi}_{cd}$ and 4,739 $E\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef}$ taking less than 1 second, while a naïve implementation of formula (6) requires calculating 18,573 $E\tilde{\pi}_{ab}\tilde{\pi}_{cd}$ and 2,423,570 $E\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef}$ taking about 4 min. Note that, our implementation can handle all pedigrees, and automatically selects the optimizations available for each pedigree automatically.

SIMULATION STUDY

We conducted extensive computer simulations to verify the accuracy of our proposed analytical power calculations for general pedigrees.

We consider a variety of pedigrees with different sizes and structures. We first consider small pedigrees with sibships of sizes 2, 4, 6 and 8, as well as cousin pedigrees (Fig. 1), cousin pedigrees with three siblings (cousin 3) and four siblings (cousin 4) in each sibship in the third generation. We also consider more extended pedigrees with 3–5 generations, and 2–6 siblings in each sibship. A quantitative trait was simulated with a single major, diallelic QTL, with minor allele frequency 0.3 and explaining 10% of the total phenotypic variance, plus 10 additive, unlinked diallelic polygenes, each explaining 7% of the total phenotypic variance. Individual-specific environmental effects account for 20% of the phenotypic variance. A single fully informative marker was simulated to be completely linked to the QTL. The number of families simulated was chosen so that, analytically, the variance component method would have ~80% power (at level 0.01) to detect the QTL. Simulations were repeated 10,000 times for each pedigree configuration, and the average of LOD scores and the distribution of *P*-values for the likelihood-ratio test were summarized.

Power calculations for small pedigrees are shown in Table I. Simulated power as well as analytical power based on exact NCPs (Exact), and second- (Approx 2) and third-order approximations (Approx 3 and Approx 3′) are presented. All pedigrees in Table I can be analyzed using the Exact approach. Evaluating the largest pedigree (with size 14) takes >1 hour using the Exact analytical power calculation, but less than 1 sec for all approximate analytical power calculations. Our results show that both statistical power and ELOD can be approximated reasonably well by their analytical estimates, assuming the alternative distribution of the likelihood-ratio test statistic is a noncentral chi-square. As expected [Yu et al., 2004], when sibships are very large (e.g., with >8 sibs) the analytical estimate of statistical power is somewhat inflated. However, the analytical ELOD

TABLE I. Comparison of power calculations in small pedigrees

			ELOD					Statistical power				
			Analytical calculation					Analytical calculation				
Pedigree	Size	No. of family	Exact	Approx 2	Approx 3	Approx 3′	Simulation	Exact	Approx 2	Approx 3	Approx 3′	Simulation
Sib2	2	4869	2.40	2.39	2.39	2.39	2.39	0.800	0.799	0.799	0.799	0.794
Sib4	4	714	2.40	2.49	2.38	2.41	2.42	0.800	0.818	0.796	0.802	0.797
Sib6	6	272	2.39	2.58	2.35	2.40	2.40	0.799	0.835	0.790	0.802	0.789
Sib8	8	144	2.39	2.69	2.31	2.40	2.38	0.799	0.852	0.782	0.801	0.777
Cousin	10	487	2.39	2.47	2.38	2.40	2.40	0.800	0.814	0.796	0.801	0.794
Cousin 3	12	259	2.40	2.53	2.37	2.41	2.43	0.800	0.827	0.794	0.802	0.790
Cousin 4	14	160	2.40	2.60	2.34	2.41	2.43	0.800	0.838	0.789	0.802	0.787

Note: 10,000 simulations were performed, where one major gene and 10 polygenes explain 10% and 70% of the total phenotypic variance, respectively.

TABLE II. Comparison of power calculations in extended pedigrees

				ELOD				Statistical power			
				Analytical calculation				Analytical calculation			
No. of generation	No. of sib	Size	No. of family	Approx 2	Approx 3	Approx 3′	Simulation	Approx 2	Approx 3	Approx 3′	Simulation
3	3	17	140	2.676	2.410	2.477	2.464	0.850	0.803	0.816	0.800
3	4	26	60	2.968	2.423	2.559	2.528	0.891	0.805	0.831	0.807
3	5	37	33	3.444	2.449	2.698	2.763	0.937	0.810	0.854	0.839
3	6	50	20	3.885	2.275	2.677	2.894	0.963	0.774	0.851	0.855
4	2	22	180	2.955	2.758	2.807	2.791	0.890	0.863	0.870	0.863
4	3	53	30	2.744	2.211	2.345	2.351	0.861	0.759	0.789	0.772
4	4	106	9	2.946	1.779	2.071	2.230	0.889	0.639	0.724	0.738
5	2	46	60	2.591	2.332	2.397	2.378	0.837	0.787	0.800	0.784

Note: Each pedigree structure is uniquely determined by the number of generations and siblings in each sibship. 10,000 simulations were performed, where one major gene and 10 polygenes explain 10% and 70% of the total phenotypic variance, respectively.

is still very accurate. Second, consistent to the theory, the third-order approximations are more accurate than the second-order approximation, and the power predicted by the analytical approaches Exact and Approx 3′ is in between Approx 2 and Approx 3. Approach Approx 3′ is almost as accurate as the Exact analytical approach.

Power calculations for extended pedigrees are shown in Table II. An exact analytical power calculation cannot be conducted for these pedigrees. For all eight extended pedigrees with 3–5 generations in Table II, the differences between simulated ELODs and ELODs from Approx 2, Approx 3, Approx 3′ range between 0.164 and 0.991, 0.033 and 0.619, and 0.016 and 0.217, respectively, and the difference between simulated powers and the analytical powers from Approx 2, Approx 3, Approx 3′ (with $K = 1/4$) are between 0.027 and 0.108, 0.000 and 0.099, and 0.004 and 0.024, respectively. We conclude that an analytical power calculation based on the third-order of Taylor expansion (Approx 3 and Approx 3′) is always more accurate than an analytical power calculation based on the second- order of Taylor expansion (Approx 2). In the settings we examined, Approx 3′ performed best among the proposed approximation approaches, especially when sibships are large. Table II also demonstrates larger pedigrees (especially with larger sibling size) provide greater power per genotyped individual than smaller pedigrees.

Analytical power calculations provide tremendous computational advantages over power calculations via simulation for large pedigrees. For example, for all pedigrees with three generations and 3–6 siblings in each sibship in Table II, each computer simulation took 2–4 h, while the analy-

tical approach employing the algorithm for tree-like pedigrees required less than 1 sec.

When markers are not fully informative or there are missing data, the power of linkage analysis will decrease. To assess the impact of marker informativeness on estimates of power, we simulated and analyzed both microsatellite and single nucleotide polymorphism (SNP) marker data [Abecasis et al., 2002]. Four alleles with equal frequencies were simulated for the microsatellite markers and two alleles with equal frequencies for the SNPs. We simulated 1,000 data sets, each with 487 cousin pedigrees (Fig. 1) for each of three different data missing patterns: (a) no missing genotype or phenotype information, (b) no genotypes or phenotypes available for grandparents and (c) no genotypes or phenotypes available for parents and grandparents. We consider the same genetic model used for Tables I and II. Simulation results are presented in Table III, together with the corresponding analytical power estimates. Table III shows that power provided by a dense SNP map closely reflects estimates given by our analytical approach, even when data for a large proportion of individuals is missing. In contrast, microsatellite markers at ~10 cM spacing provide considerably less power than a fully informative marker.

In order to consider different choices for K in implementing Approx 3′ (equation 5), we compared simulated ELODs and analytical ELODs using Approx 3′ for different values of K. Note that $K = 0$ corresponds to Approx 2 and $K = 1/3$ corresponds to Approx 3. Figure 3 shows an analytical approach with $K > \sim 1/4$ tends to underestimate the power by a modest amount, and an analytical approach with K close to zero can overestimate the power by a large amount. Except for the case of three generation pedigree with six siblings per sibship (Fig. 2), the difference between the simulated ELOD and the ELOD of Approx 3′ with $K = 1/4$ (equation 5) is very small. A slightly larger K could be even more accurate for the small pedigrees we examined, but it would also result in a large underestimate of the ELOD for the largest pedigree we examined. Clearly, there appears to be no uniformly ''best'' choice for K.

REAL DATA EXAMPLE

Pilia et al. [2006] conducted variance component polygenic analysis to dissect heritabilities for 98 quantitative traits in a cohort study of 6,148 Sardinians. The sample includes 4,933 sib pairs, 4,256 parent-child pairs, 4,014 first cousins, 6,400 avuncular pairs in addition to other more distant relative pairs. The largest pedigree in the cohort connects 625 phenotyped individuals in five generations. No inbreeding was present in the pedigrees we analyzed. We continued the polygenic analysis using the same data by predicting ELOD scores for a hypothetical future linkage study.

We estimated the power of linkage analysis for each of the 98 quantitative traits, for a simple additive genetic model as well as for a household model. In the additive genetic model, variance was partitioned into a polygenic component σ_g^2 and an environmental component σ_e^2; so that, the variance of trait for person i was $(\Omega_0)_{ii} = \sigma_g^2 + \sigma_e^2$ and the covariance between measurements for a pair of individuals i and j with kinship ϕ_{ij} was $(\Omega_0)_{ij} = 2\ \phi_{ij}\ \sigma_g^2$. In the household model, we also allowed for shared sibling environment, σ_s^2. Let $I_{sib(i,j)}$ be an indicator variable with value 1 when individuals i and j are full sibs, and value 0 otherwise. Then $(\Omega_0)_{ii} = \sigma_g^2 + \sigma_e^2 + \sigma_s^2$ and $(\Omega_0)_{ij} = 2\ \phi_{ij}\ \sigma_g^2 + I_{sib(i,j)}\sigma_s^2$.

TABLE III. Power of cousin pedigrees with partially informative markers

	Everyone genotyped		First generation untyped		First, second generation untyped	
Marker map	ELOD	Power	ELOD	Power	ELOD	Power
Microsatellite, 10 cM, QTL in the middle	1.875	0.657	1.056	0.354	0.531	0.133
Microsatellite, 10 cM, QTL at a marker	2.154	0.739	1.202	0.417	0.566	0.146
SNP, 1 cM, QTL in the middle	2.308	0.774	1.385	0.493	0.772	0.238
SNP, 0.2 cM, QTL in the middle	2.411	0.801	1.486	0.525	0.859	0.264
Ideal map (with analytical calculation)	2.401	0.801	1.500	0.542	0.865	0.275

Note: 487 cousin pedigrees were simulated 1,000 times with the same genetic effects as before. The last row gives the analytical results for the three missing data patterns using Approx 3′ approach.

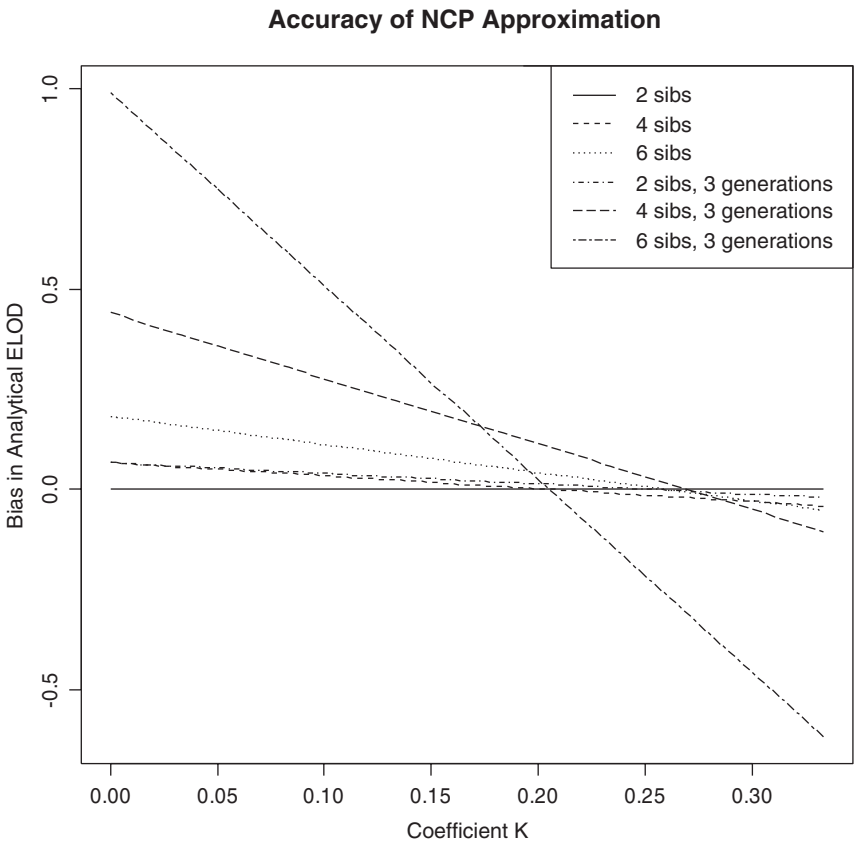

Fig. 3. Performance of ad hoc approximations of analytical ELOD (Approx 3′) for different choices of *K* in Taylor expansion of NCP.

In order to verify the accuracy of our analytical power calculation for real data, we conducted computer simulations for eight representative traits, including cholesterol and HDL for blood analysis, height and weight for anthropometric measures, IMT and PWV for cardiovascular function and NEO N and NEO O for personality measures. For each simulation, a quantitative trait was simulated with a single major, diallelic QTL, with minor allele frequency 0.5 and explaining 10% of the total phenotypic variance, plus 10 additive, unlinked diallelic polygenes with equal effect sizes so that the total genetic variance was the same as the heritability estimated from the polygenic analysis. A nonshared environmental effect was simulated independent of genetic effects using parameters estimated from polygenic analysis. A single fully informative marker was simulated to be completely linked to the QTL. 10,000 simulations were performed and LOD scores from each simulation were averaged. The average of simulated LOD scores serves as the standard for evaluating the analytical ELOD.

Table IV shows analytical and simulated ELODs for eight traits at a QTL that explains 10% of the total phenotypic variance. The number of phenotyped individual ranges from 5,657 for the two personality measures to 6,146 for trait HEIGHT. The overall heritabilities in the second column were estimated from the polygenic analysis [Pilia et al., 2006]. For all eight traits, the difference between an analytical ELOD (Approx 3′) and the simulated ELOD was always less than 0.046 (1.4%), while analytical power approximation Approx 3 underestimated the ELOD by as much as 0.140 (3.2%), and approximation Approx 2 overestimated ELOD by as much as 0.257 (8.1%). In general, traits with higher heritability have higher ELODs. One exception is that although the heritability estimate of trait NEO O is higher than that of the two cardiovascular traits in the table, the power estimate for NEO O is actually lower. One reason is that trait NEO O has a smaller sample size than other traits. The exceptions such as this show it is useful to carry out power analysis individually for all 98 traits in our data.

TABLE IV. Simulated and analytical ELODs for eight illustrative traits in Sardinia data

		Additive model					Household model		
Trait	No. of individuals	Heritability	Approx 2 ELOD	Approx 3 ELOD	Approx 3′ ELOD	Simulated ELOD	Heritability	Household effect	Approx3′ ELOD
Height	6146	0.801	4.767	4.282	4.403	4.422	0.771	0.100	5.229
Weight	6144	0.498	3.466	3.180	3.251	3.233	0.439	0.094	3.522
Cholesterol	6142	0.424	3.293	3.031	3.097	3.070	0.374	0.067	3.233
HDL	6142	0.487	3.433	3.152	3.222	3.176	0.471	0.028	3.288
IMT	6080	0.187	2.967	2.750	2.804	2.776	0.132	0.062	2.874
PWV	6048	0.226	2.954	2.738	2.792	2.766	0.224	0.002	2.794
NEO O	5657	0.329	2.769	2.571	2.621	2.623	0.285	0.063	2.713
NEO N	5657	0.258	2.700	2.511	2.558	2.561	0.211	0.059	2.627

Note: Heritability of each trait is estimated from the polygenic analysis of the Sardinia data. The household model incorporates a common environmental variance component or a dominant polygenic genetic effect only shared by siblings, and the proportion of variance that the household effect explains is shown in column ''Household Effect''.

Our proposed power calculation framework allows rather flexible modeling of the variance components. Table IV also gives the powers of the eight traits under a household model where an additional household variance component (due to both common environmental effect and dominant polygenic genetic effect) is only shared by siblings. The magnitude of the household effect was estimated from the polygenic analysis [Pilia et al., 2006] and given in Table IV. It is interesting to observe that the ELOD for some traits is much higher for the model incorporating a household variance component than under the basic additive genetic model.

An analytical power computation is much faster than a power computation via simulations. It took 2 weeks to obtain an empirical ELOD via efficient simulations for each trait, while it only took 6 h to calculate ELODs for all 98 traits (see online supplementary table). The proposed tree-based algorithm which works for tree-like pedigrees could be extended to more complicated pedigree structures like our real data example, and thus it is possible that the analytical power analysis of all traits in our data set could be accomplished in minutes.

DISCUSSION

Variance component linkage analysis has been routinely used to detect QTLs in general pedigrees. As the key for study design, estimating the power for variance component linkage analysis is very important. The trend towards larger data sets in terms of both the pedigree size and the number of traits introduces more computational challenges for statistical genetics. To our knowledge, no efficient and accurate algorithms of power analysis have been developed for large pedigrees. In this study, we propose algorithms of analytical power calculation for large pedigrees based on approximating the noncentrality of LRT statistic using Taylor expansion theory. The accuracy and efficiency of these algorithms were verified by computer simulations using simulated pedigrees of a variety of sizes as well as a recently collected large data set.

For a particular data set, in order to predict the power to detect linkage, one can choose from one of the following methods: the ''exact'' analytical power calculation (for example, as implemented in Linkage Explorer [Chen et al., 2005]), the approximate analytical power calculation (for example, as implemented in POLY program described here) or the power calculation based on simulations (for example, as implemented in SOLAR [Almasy and Blangero, 1998]). When the pedigree size is small (say, <10) and there are only one or a few traits to be analyzed, both the power calculation based on simulations and the ''exact'' analytical power calculation are practical. When pedigrees are large, when multiple traits need to be analyzed or when a linkage test is time-consuming, our proposed analytical power approximation is highly accurate and efficient, and hence is strongly recommended.

Our proposed algorithms for power calculation can be easily extended to several other scenarios. In a variance component framework, discrete traits are typically analyzed by use of a liability

threshold method [Williams and Blangero, 2004] where computationally intensive integrations of the underlying continuous traits need to be computed for the linkage test. A simulation study is not practical for the purpose of power analysis, because of computational requirements for the numerical integration procedure. In contrast, our proposed approximate power calculation only requires a one time evaluation of the test statistic and thus could make power analysis feasible. Furthermore, our proposed algorithms should also be helpful for the power analysis in other computationally challenging scenarios including multivariate trait linkage analysis, quantitative trait linkage analysis with multiple repeated measures and even linkage analysis of thousands of gene expression traits.

We present efficient algorithms to compute moments of allele-sharing coefficients, up to the third moment, relying on the pedigree structure only. These algorithms should be useful for the power analysis in other pedigree-based studies. They could also potentially enhance other stages of linkage and association studies. For example, the covariance of allele-sharing coefficients needs to be calculated in two robust linkage tests [Chen et al., 2005] and our algorithms enable these tests to handle more general pedigrees.

We show our analytical power calculation for variance component linkage analysis is highly accurate when the marker is fully informative and tightly linked to the QTL. When the marker data are not fully informative, the power of linkage tests could be overestimated by our calculation. In some simple cases, our proposed analytical power calculation could be modified to incorporate marker informativeness. For example, the NCP calculation (3) for sibship data can incorporate marker informativeness by simply multiplying the empirical moment estimate of variance of allele-sharing coefficient for any two siblings times 8. Incorporating marker informativeness in a power calculation for more general pedigrees remains an open question. Nevertheless, in the settings we examined by simulation, we found that using an SNP linkage panel with $\sim$1 SNP/cM resulted in only a small loss of power and that with $\sim$5 SNPs/cM provided nearly the same power as a fully informative marker panel. These numbers correspond to 3,000–15,000 SNPs genome wide and are comparable to the density provided by panels in current widespread use. At 10 cM density, microsatellite panels resulted in some loss of power, but the problem could be attenuated by genotyping additional markers near linkage peaks.

In this study, we only considered non-inbred pedigrees. With slight modifications, our proposed algorithms can be applied to nonstandard scenarios such as linkage analysis in the presence of inbreeding, parent-of-origin effects and sex-chromosome linkage analysis.

Our power analysis conveniently incorporates a variety of genetic models and covariance matrices, as well as missing data patterns. All algorithms for power calculation presented here have been implemented in our polygenic analysis program POLY. POLY runs on platforms where a modern C++ compiler is available, including those based on the Linux, UNIX, Windows and Mac OS X operating systems. Both executables and source code are freely available at our website.

ACKNOWLEDGMENTS

The authors were supported in part by Grants EY10562 and HG02651 from National Institutes of Health and by contract 263-MA-410953 from the National Institutes of Aging to the University of Michigan. The authors thank two anonymous reviewers for valuable suggestions for improving the manuscript.

ELECTRONIC DATABASE INFORMATION

http://www.sph.umich.edu/csg/chen/public/software/poly/ for downloading the POLY software package.

http://www.sph.umich.edu/csg/chen/public/sardinia/LOD.html for supplementary online table for analytical ELOD scores for 98 traits in the cohort study of 6,148 Sardinians.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531–544.

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigree. Am J Hum Genet 62:1198–1211.

Amos CI. 1994. Robust variance-component approach for assessing genetic linkage in pedigree. Am J Hum Genet 54:535–543.

Cardon LR, Abecasis GR. 2000. Some properties of a variance components model for fine-mapping quantitative trait loci. Behav Genet 30:235–243.

Chen WM, Deng HW. 2001. A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. Genet Epidemiol 21: 53–67.
Chen WM, Broman KW, Liang KY. 2004. Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genet Epidemiol 26:265–272.
Chen WM, Broman KW, Liang KY. 2005. Power and robustness of linkage tests for quantitative traits in general pedigrees. Genet Epidemiol 28:11–23.
Dolan CV, Boomsma DI, Neale MC. 1999. A note on the power provided by sibships of sizes 2, 3, and 4 in genetic covariance modeling of a codominant QTL. Behav Genet 29: 163–170.
Evans DM. 2002. The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. Am J Hum Genet 70:1599–1602.
Fulker DW, Cherny SS, Sham PC, Hewit JK. 1999. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267.
Haseman JX, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2: 3–19.
Hopper JL, Mathews JD. 1982. Extensions of multivariate normal models for pedigree analysis. Ann Hum Genet 46:373–383.
Lange K. 2002. Mathematical and Statistical Methods for Genetic Analysis. New York: Springer-Verlag.
Lange K, Boehnke M. 1983. Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. Am J Med Genet 14:513–524.
Pilia G, Chen WM, Scuteri A, Orrú M, Albai G, Dei M, Lai S et al. 2006. The SardiNIA project: heritabilities of cardiovascular and personality traits in a cohort of 6,148 Sardinians, submitted.
Rijsdijk FV, Hewitt JK, Sham PC. 2001 Analytic power calculations for QTL linkage analysis of small pedigrees. Eur J Hum Genet 9:335–340.
Tang HK, Siegmund D. 2001. Mapping quantitative trait loci in oligogenic models. Biostatistics 2:147–162.
Visscher PM, Hopper JL. 2001. Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. Ann Hum Genet 65:583–601.
Williams JT, Blangero J. 1999. Power of variance component linkage analysis to detect quantitative trait loci. Ann Hum Genet 63:545–563.
Williams JT, Blangero J. 2004. Power of variance component linkage analysis—II. Discrete traits. Ann Hum Genet 68:620–632.
Yu W, Knott SA, Visscher PM. 2004. Theoretical and empirical power of regression and maximum-likelihood methods to map quantitative trait loci in general pedigrees. Am J Hum Genet 75:17–26.

APPENDIX A

ALGORITHMS TO COMPUTE HIGHER MOMENTS OF ALLELE-SHARING COEFFICIENTS

We only consider the case of non-inbred pedigrees. Assume $a \geq \max(b, c, d, e, f)$, i.e., no person in the set $\{b, c, d, e, f\}$ is a descendant of person a. Let p and q denote the parents of person a.

The kinship coefficient ϕ_{abc} is defined as the probability that three alleles drawn at random from each person in the set $\{a, b, c\}$ are IBD. Then ϕ_{abc} can be calculated as follows:

(1) If $a = b = c$, then $\phi_{aaa} = 1/4$.
(2) If a is a founder and if $a > b$ and/or $a > c$, then $\phi_{abc} = 0$.
(3) If a is not a founder and $a = b > c$, then $\phi_{aac} = (\phi_{pc} + \phi_{qc})/4$.
(4) If a is not a founder and $a > b \geq c$, then $\phi_{abc} = (\phi_{pbc} + \phi_{qbc})/2$.

The kinship coefficient ϕ_{abcd} is the probability that four alleles drawn at random from each person in the set $\{a, b, c, d\}$ are IBD. ϕ_{abcd} can be calculated as follows:

(1) If $a = b = c = d$, then $\phi_{aaaa} = 1/8$.
(2) If a is a founder and if $a > b$, $a > c$ and/or $a > d$, then $\phi_{abcd} = 0$.
(3) If a is not a founder and $a = b = c > d$, then $\phi_{aaad} = (\phi_{pd} + \phi_{qd})/8$.
(4) If a is not a founder and $a = b > c \geq d$, then $\phi_{aacd} = (\phi_{pcd} + \phi_{qcd})/4$.
(5) If a is not a founder and $a > b \geq c \geq d$, then $\phi_{abcd} = (\phi_{pbcd} + \phi_{qbcd})/2$.

Allele-sharing coefficient π_{abc} is defined as 4 times the probability that three alleles drawn at random from each person of a, b, and c are IBD conditional on the genotype of each person.

Without loss of generality, we assume $a \geq b \geq c$ and $d \geq e$. Then the algorithm to calculate $E[\tilde{\pi}_{abc}\tilde{\pi}_{de}]$ is as follows:

(1) If π_{abc} is a constant (e.g., when $\phi_{abc} = 0$ or 0.25), or π_{de} is a constant (e.g., when $\phi_{de} = 0$ or $= 0.5$), then $E[\tilde{\pi}_{abc}\tilde{\pi}_{de}] = 0$.
(2) If $a = d$ and $a > b$, then

$$E[\tilde{\pi}_{abc}\tilde{\pi}_{ae}] = (E[\tilde{\pi}_{pbc}\tilde{\pi}_{qe}] + E[\tilde{\pi}_{qbc}\tilde{\pi}_{pe}])/4 + 2(\phi_{pbce} + \phi_{qbce} - \phi_{pbc}\phi_{pe} - \phi_{qbc}\phi_{qe}).$$

(3) If $a = b = d > c$, then

$$E[\tilde{\pi}_{aac}\tilde{\pi}_{ae}] = (E[\tilde{\pi}_{pc}\tilde{\pi}_{qe}] + E[\tilde{\pi}_{qc}\tilde{\pi}_{pe}])/8 + \phi_{pce} + \phi_{qce} - \phi_{pc}\phi_{pe} - \phi_{qc}\phi_{qe}.$$

(4) If $a > d$ and $a > b$, then

$$E[\tilde{\pi}_{abc}\tilde{\pi}_{de}] = (E[\tilde{\pi}_{pbc}\tilde{\pi}_{de}] + E[\tilde{\pi}_{qbc}\tilde{\pi}_{de}])/2.$$

(5) If $a>d$ and $a=b$, then

$$E[\tilde{\pi}_{aac}\tilde{\pi}_{de}]=(E[\tilde{\pi}_{pc}\tilde{\pi}_{de}]+E[\tilde{\pi}_{qc}\tilde{\pi}_{de}])/4.$$

APPENDIX B

PROOF OF ALGORITHMS FOR HIGHER MOMENTS OF ALLELE-SHARING COEFFICIENTS

Lange [2002] shows the allele-sharing coefficient π can be equivalently defined as $\pi_{ab}=2E[1_{\{G_a=G_b\}}|M_a,M_b]$, where M_a and M_b are fully informative markers for persons a and b, and G_a is an allele randomly drawn from person a. Following notations by Lange [2002], we use { } for a nonoverlapping block whose constituent genes are IBD. Since an allele may be drawn at random multiple times, we use $G_a^{(1)}$ to represent the first time of random drawing. Without loss of generality, we assume $a\geq\max(b,c,d,e,f)$, $c\geq d$, $e\geq f$ and let p and q denote the parents of person a. Then

$$\begin{aligned}E[\pi_{ab}\pi_{cd}]&=4E[E[1_{\{G_a=G_b\}}1_{\{G_c=G_d\}}|M_a,M_b,M_c,M_d]]\\&=4E[1_{\{G_a=G_b\}}1_{\{G_c=G_d\}}]\\&=4\Pr(G_a^{(1)}=G_b^{(1)},G_c^{(2)}=G_d^{(2)})\\&=4\Phi(\{G_a^{(1)},G_b^{(1)},G_c^{(2)},G_d^{(2)}\})\\&\quad+4\Phi(\{G_a^{(1)},G_b^{(1)}\},\{G_c^{(2)},G_d^{(2)}\}).\end{aligned}$$

When $a>b$ and $a>c$,

$$\begin{aligned}E[\pi_{ab}\pi_{cd}]&=4\Phi(\{G_a^{(1)},G_b^{(1)},G_c^{(2)},G_d^{(2)}\})\\&\quad+4\Phi(\{G_a^{(1)}=G_b^{(1)}\},\{G_c^{(2)}=G_d^{(2)}\})\\&=4\{1/2\Phi(\{G_p^{(1)},G_b^{(1)},G_c^{(2)},G_d^{(2)}\})\\&\quad+1/2\Phi(\{G_q^{(1)},G_b^{(1)},G_c^{(2)},G_d^{(2)}\})\}\\&\quad+4\{1/2\Phi(\{G_p^{(1)},G_b^{(1)}\},\{G_c^{(2)},G_d^{(2)}\})\\&\quad+1/2\Phi(\{G_q^{(1)},G_b^{(1)}\},\{G_c^{(2)},G_d^{(2)}\})\}\\&=1/2E[\pi_{pb}\pi_{cd}]+1/2E[\pi_{qb}\pi_{cd}].\end{aligned}$$

When $a=c$,

$$\begin{aligned}E[\pi_{ab}\pi_{cd}]&=4\Phi(\{G_a^{(1)},G_b^{(1)},G_a^{(2)},G_d^{(2)}\})\\&\quad+4\Phi(\{G_a^{(1)}=G_b^{(1)}\},\{G_a^{(2)}=G_d^{(2)}\})\\&=4\Big\{1/4\Phi(\{G_p^{(1)},G_b^{(1)},G_q^{(2)},G_d^{(2)}\})\\&\quad+1/4\Phi(\{G_q^{(1)},G_b^{(1)},G_p^{(2)},G_d^{(2)}\})\\&\quad+1/4\Phi(\{G_p,G_b^{(1)},G_d^{(2)}\})+1/4\Phi(\{G_q,G_b^{(1)},G_d^{(2)}\})\Big\}\\&\quad+4\Big\{1/4\Phi(\{G_p^{(1)},G_b^{(1)}\},\{G_q^{(2)},G_d^{(2)}\})\\&\quad+1/4\Phi(\{G_q^{(1)},G_b^{(1)}\},\{G_p^{(2)},G_d^{(2)}\})\Big\}\\&=1/4E[\pi_{pb}\pi_{qd}]+1/4E[\pi_{qb}\pi_{pd}]\\&\quad+\Phi(\{G_p,G_b^{(1)},G_d^{(2)}\})\\&\quad+\Phi(\{G_q,G_b^{(1)},G_d^{(2)}\}).\end{aligned}$$

Note G_a, $G_b^{(1)}$ and $G_b^{(2)}$ are three alleles drawn in different time. The generalized kinship coefficient $\Phi(\{G_a,G_b^{(1)},G_d^{(2)}\})$ could be replaced by a simpler notation ϕ_{abd}. The recurrence rule for the calculation of ϕ_{abd} is standard [see Lange, 2002].

Now we consider the third moment of allele-sharing coefficients. Since

$$\begin{aligned}E[\tilde{\pi}_{ab}\tilde{\pi}_{cd}\tilde{\pi}_{ef}]&=E[\pi_{af}\pi_{cd}\pi_{ef}]-E[\pi_{ab}]E[\pi_{cd}]E[\pi_{ef}]\\&\quad-E[\tilde{\pi}_{ab}\tilde{\pi}_{cd}]E[\pi_{ef}]-E[\tilde{\pi}_{ab}\tilde{\pi}_{ef}]E[\pi_{ab}]\\&\quad-E[\tilde{\pi}_{ab}\tilde{\pi}_{ef}]E[\pi_{cd}]\end{aligned}$$

we focus on the calculation of $E[\pi_{ab}\pi_{cd}\pi_{ef}]$.

$$\begin{aligned}E[\pi_{ab}\pi_{cd}\pi_{ef}]&=8E[E[1_{\{G_a=G_b\}}1_{\{G_c=G_d\}}1_{\{G_e=G_f\}}|M_a,M_b,\\&\qquad M_c,M_d,M_e,M_f]]\\&=8E[1_{\{G_a=G_b\}}1_{\{G_c=G_d\}}1_{\{G_e=G_f\}}]\\&=8\Pr(G_a^{(1)}=G_b^{(1)},G_c^{(2)}=G_d^{(2)},G_e^{(3)}=G_f^{(3)})\\&=8\Phi(\{G_a^{(1)},G_b^{(1)},G_c^{(2)},G_d^{(2)},G_e^{(3)},G_f^{(3)}\})\\&\quad+8\Phi(\{G_a^{(1)},G_b^{(1)},G_c^{(2)},G_d^{(2)}\},\{G_e^{(3)},G_f^{(3)}\})\\&\quad+8\Phi(\{G_a^{(1)},G_b^{(1)}\},\{G_c^{(2)},G_d^{(2)},G_e^{(3)},G_f^{(3)}\})\\&\quad+8\Phi(\{G_a^{(1)},G_b^{(1)},G_e^{(3)},G_f^{(3)}\},\{G_c^{(2)},G_d^{(2)}\})\\&\quad+8\Phi(\{G_a^{(1)},G_b^{(1)}\},\{G_c^{(2)},G_d^{(2)}\},\{G_e^{(3)},G_f^{(3)}\}).\end{aligned}$$

The boundary rule for the iterative procedure is, whenever at least one allele-sharing coefficient (e.g., between a and b) is a constant, $E[\pi_{ab}\pi_{cd}\pi_{ef}]=\pi_{ab}E[\pi_{cd}\pi_{ef}]$. We define $\pi_{abc}=4E[1_{\{G_a=G_b=G_c\}}|M_a,M_b,M_c]$. Then

$$\begin{aligned}E[\pi_{abc}\pi_{de}]&=8\Phi(\{G_a^{(1)},G_b^{(1)},G_c^{(1)},G_d^{(2)},G_e^{(2)}\})\\&\quad+8\Phi(\{G_a^{(1)},G_b^{(1)},G_c^{(1)}\},\{G_d^{(2)},G_e^{(2)}\}).\end{aligned}$$

Now we show the iterative recurrence rule for the calculation of $E[\pi_{ab}\pi_{cd}\pi_{ef}]$.

If $a>c$ and $a>e$, then

$$E[\pi_{ab}\pi_{cd}\pi_{ef}]=1/2E[\pi_{pb}\pi_{cd}\pi_{ef}]+1/2E[\pi_{qb}\pi_{cd}\pi_{ef}].$$

If $a=c$ and $a>e$, then

$$\begin{aligned}E[\pi_{ab}\pi_{ad}\pi_{ef}]&=4\Phi(\{G_p^{(1)},G_q^{(1)},G_b^{(1)},G_d^{(2)},G_e^{(3)},G_f^{(3)}\})\\&\quad+2\Phi(\{G_p^{(1)},G_b^{(1)},G_d^{(2)},G_e^{(3)},G_f^{(3)}\})\\&\quad+2\Phi(\{G_q^{(1)},G_b^{(1)},G_d^{(2)},G_e^{(3)},G_f^{(3)}\})\\&\quad+4\Phi(\{G_p^{(1)},G_q^{(1)},G_b^{(1)},G_d^{(2)}\},\{G_e^{(3)},G_f^{(3)}\})\end{aligned}$$

$$
\begin{aligned}
&+ 2\Phi(\{G_p^{(1)}, G_b^{(1)}, G_d^{(2)}\}, \{G_e^{(3)}, G_f^{(3)}\}) \\
&+ 2\Phi(\{G_q^{(1)}, G_b^{(1)}, G_d^{(2)}\}, \{G_e^{(3)}, G_f^{(3)}\}) \\
&+ 2\Phi(\{G_p^{(1)}, G_b^{(1)}\}, \{G_q^{(2)}, G_d^{(2)}, G_e^{(3)}, G_f^{(3)}\}) \\
&\quad + 2\Phi(\{G_q^{(1)}, G_b^{(1)}\}, \{G_p^{(2)}, G_d^{(2)}, G_e^{(3)}, G_f^{(3)}\}) \\
&+ 2\Phi(\{G_q^{(1)}, G_d^{(1)}\}, \{G_p^{(2)}, G_b^{(2)}, G_e^{(3)}, G_f^{(3)}\}) \\
&\quad + 2\Phi(\{G_p^{(1)}, G_d^{(1)}\}, \{G_q^{(2)}, G_b^{(2)}, G_e^{(3)}, G_f^{(3)}\}) \\
&+ 2\Phi(\{G_p^{(1)}, G_b^{(1)}\}, \{G_q^{(2)}, G_d^{(2)}\}, \{G_e^{(3)}, G_f^{(3)}\}) \\
&\quad + 2\Phi(\{G_q^{(1)}, G_b^{(1)}\}, \{G_p^{(2)}, G_d^{(2)}\}, \{G_e^{(3)}, G_f^{(3)}\}) \\
&= 1/4E[\pi_{pb}\pi_{qd}\pi_{ef}] + 1/4E[\pi_{qb}\pi_{pd}\pi_{ef}] \\
&\quad + 1/4E[\pi_{pbd}\pi_{ef}] + 1/4E[\pi_{qbd}\pi_{ef}].
\end{aligned}
$$

If $a = c = e$, then

$$
\begin{aligned}
E[\pi_{ab}\pi_{ad}\pi_{af}] &= 6\Phi(\{G_p^{(1)}, G_q^{(1)}, G_b^{(1)}, G_d^{(2)}, G_f^{(3)}\}) \\
&\quad + \Phi(\{G_p^{(1)}, G_b^{(1)}, G_d^{(2)}, G_f^{(3)}\}) \\
&\quad + \Phi(\{G_q^{(1)}, G_b^{(1)}, G_d^{(2)}, G_f^{(3)}\}) \\
&\quad + \Phi(\{G_p^{(1)}, G_b^{(1)}, G_d^{(2)}\}, \{G_q^{(3)}, G_f^{(3)}\}) \\
&\quad + \Phi(\{G_q^{(1)}, G_b^{(1)}, G_d^{(2)}\}, \{G_p^{(3)}, G_f^{(3)}\}) \\
&\quad + \Phi(\{G_p^{(1)}, G_b^{(1)}, G_f^{(2)}\}, \{G_q^{(3)}, G_d^{(3)}\}) \\
&\quad + \Phi(\{G_q^{(1)}, G_b^{(1)}, G_f^{(2)}\}, \{G_p^{(3)}, G_d^{(3)}\}) \\
&\quad + \Phi(\{G_p^{(1)}, G_d^{(1)}, G_f^{(2)}\}, \{G_q^{(3)}, G_b^{(3)}\}) \\
&\quad + \Phi(\{G_q^{(1)}, G_d^{(1)}, G_f^{(2)}\}, \{G_p^{(3)}, G_b^{(3)}\}) \\
&= 1/8E[\pi_{pbd}\pi_{qf}] + 1/8E[\pi_{qbd}\pi_{pf}] \\
&\quad + 1/8E[\pi_{pbf}\pi_{qd}] \\
&\quad + 1/8E[\pi_{qbf}\pi_{pd}] + 1/8E[\pi_{pdf}\pi_{qb}] \\
&\quad + 1/8E[\pi_{qdf}\pi_{pb}] \\
&\quad + \Phi(\{G_p^{(1)}, G_b^{(1)}, G_d^{(2)}, G_f^{(3)}\}) \\
&\quad + \Phi(\{G_q^{(1)}, G_b^{(1)}, G_d^{(2)}, G_f^{(3)}\}).
\end{aligned}
$$

We use ϕ_{pbdf} to denote $\Phi(\{G_p^{(1)}, G_b^{(1)}, G_d^{(2)}, G_f^{(3)}\})$, and its calculation is standard [Lange, 2002].