Optimal Allele-Sharing Statistics for Genetic Mapping using Affected Relatives

Mary Sara McPeek

Department of Statistics, University of Chicago, Chicago, IL

Running Title: Optimal Allele-Sharing Statistics

Address for correspondence: Mary Sara McPeek

Department of Statistics, University of Chicago,

5734 S. University Ave., Chicago, IL 60637

phone: (773) 702-7554

fax: (773) 702-9810

email: mcpeek@galton.uchicago.edu

Abstract

The choice of allele-sharing statistic can have a great impact on the power of robust affected relative methods. Similarly, when allele-sharing statistics from several pedigrees are combined, the weight applied to each pedigree's statistic can affect power. Here we describe the direct connection between the affected relative methods and traditional parametric linkage analysis, and we use this connection to give explicit formulae for the optimal sharing statistics and weights, applicable to all pedigree types. One surprising consequence is that under any single gene model, the value of the optimal allele-sharing statistic does not depend on whether observed sharing is between more closely or more distantly related affected relatives. This result also holds for any multigene model with loci unlinked, additivity between loci, and all loci having small effect. For specific classes of two-allele models, we give the most powerful statistics and optimal weights for arbitrary pedigrees. When the effect size is small, these also extend to multigene models with additivity between loci. We propose a useful new statistic, $S_{rob\,dom}$, which performs well for dominant and additive models with varying phenocopy rates and varying predisposing allele frequency. We find that the statistic $S_{-\#alleles}$, performs well for recessive models with varying phenocopy rates and varying predisposing allele frequency. We also find that for models with large deviation from null sharing, the correspondence between allele-sharing statistics and the models for which they are optimal may also depend on which method is used to test for linkage.

Key words: complex traits, robust linkage analysis, relative pairs, affected pedigree member, discordant sib pairs

Introduction

Genetic mapping of a trait by linkage analysis involves finding regions of the genome with a tendency to be shared identical by descent (IBD) by close relatives affected with the trait and not shared between affected and unaffected relatives. Full parametric linkage analysis entails specification of a model for inheritance of the trait, with the location of the gene then estimated by the method of maximum likelihood. For a single-gene two-allele model, the full parametric model would include the frequency, a, of the trait-causing allele, as well as penetrances f_0, f_1, f_2, for noncarriers, heterozygote carriers, and homozygote carriers, respectively. More complicated models might include 1) a single gene model with h alleles with frequencies a_1, ..., a_h, respectively, and penetrance f_{ij} for an individual whose genotype consists of the ith and jth alleles or 2) multigene versions of the above models, in which loci are unlinked and effects are additive across loci. In cases when the parameters of the model are not known, they are sometimes impossible to estimate, and the maximum likelihood analysis has been found to be very sensitive to model misspecification [Clerget-Darpoux et al. 1986]. Allele-sharing methods have been proposed as a way to avoid these difficulties. This class of methods includes the sib pair method, originated by Penrose [1935] and developed by many others [e.g., Day and Simons (1976); Green and Woodrow (1977); Fishman et al. (1978); Suarez (1978); Hodge (1984); Lange (1986); Fimmers et al. (1989)], the affected pedigree member method (APM) of Weeks and Lange [1988] which uses identity by state (IBS) information, work on affected relative pairs by Risch [1990b], and general affected relative methods that use IBD information [Whittemore and Halpern (1994); Kruglyak et al.

(1996); Whittemore (1996); Kong and Cox (1997)]. These methods are clearly not model-free, but they are believed to be more robust than full parametric likelihood analysis in those cases when the model is not known.

We now describe the framework for these allele-sharing methods. For any genome location x and any pedigree with n members and l founders, $l < n$, following Thompson [1974], we number the founders' alleles 1 through $2 \times l$, and we define the gene-identity state g at each location x in the genome by $g(x) = (p_1, m_1, p_2, m_2, ..., p_n, m_n)$ where p_i represents the founder allele inherited by individual i from his or her father and m_i that from his or her mother. We consider two gene-identity states to be equivalent if one can be obtained from the other by simply permuting the allele labels. The resulting equivalence classes of gene-identity states are called IBD configurations and denoted by c. We consider an allele-sharing statistic $S(c, \Phi)$ to be a function of the allele configuration c and the phenotype information Φ in the pedigree (more generally, S might depend on g rather than c).

Allele-sharing methods generally consider sharing among affecteds only, and with the exception of a brief discussion of discordant sib pairs, we limit this study to statistics S depending on affecteds only. Note that in principle, S may use information on sharing with unaffecteds as well. For instance, full parametric linkage analysis may be seen as the case where S is chosen to be the likelihood ratio [Kruglyak et al. 1996], which of course depends on the genotype information on both affecteds and unaffecteds. The three rationales for considering affecteds only are, first, that this effectively eliminates one penetrance parameter from the model, leading to greater robustness when the model is unknown. (When only affecteds are considered, the two-allele model can be parametrized by the frequency a of the predisposing allele and the relative risks f_2/f_0 and f_1/f_0.) Second, affecteds contribute most of the information to the study, and elimination of unaffecteds from consideration does not usually cause a severe loss of power. Third, for many complex diseases or traits, some individuals classified as unaffected may simply not yet have developed the disease or trait. For instance Alzheimer's disease, many cancers, and many phenotypes related to heart disease tend to have a late age of onset. Thus, in some cases, the designation of an individual as "unaffected" may be much more uncertain than the designation of an individual as "affected." Note that although S depends only on affecteds, genotype information on unaffecteds may be used for inferring IBD information on affecteds.

We concentrate on allele-sharing methods based on IBD rather than IBS sharing, as the former are more powerful [Kruglyak et al. (1996); Sobel and Lange (1996)]. In practice, of course, full IBD information is not available, but instead, the conditional distribution of the allele configuration $c(x)$ at any given location x, conditional on the marker data, may be computed. For instance, the software package GENEHUNTER of Kruglyak et al. [1996] can compute, for small to moderate-size pedigrees, the conditional distribution of what they call the inheritance vector, which is equivalent to the allele configuration, at a location x, given the multipoint marker information for all pedigree members. In that case, instead of considering $S(c(x), \Phi)$, one could consider, e.g., its null expectation conditional on the multipoint marker information, $\bar{S}(x, \Phi) = \sum_{w \in C} S(w, \Phi) P_O[c(x) = w|\text{data}]$, where $P_O[c(x) = w|\text{data}]$ is calculated under the null hypothesis of no gene for the trait linked to that location.

Under the null hypothesis of no gene for the trait linked to location x, the distribution of an allele-sharing statistic S is in principle known. The hope is that S will show significant deviation from its null distribution when there is a gene at that location affecting the trait. Proposed tests for detecting this deviation are described below. Not surprisingly, the power to detect linkage using any particular statistic S can vary greatly depending on the underlying genetic model for the trait. For instance, Figures 1a and b depict the power to detect dominant alternatives with various phenocopy rates and penetrances using 30 trios consisting of affected sib pairs each with affected parent, while Figures 1e and f show the power to detect recessive alternatives under the same conditions. Four different allele-sharing statistics are compared (definitions given in **Definitions of Allele-Sharing Statistics** below). Note that those that perform best in the dominant case perform worst in the recessive case and vice versa, although the statistics perform similarly in the dominant and additive cases, as shown by a comparison of Figures 1a - d. In this paper, we investigate the relationship between allele-sharing statistics and two-allele models, with extension to special cases of multiple unlinked genes.

Prior work on choice of sharing statistic

For the special case of sib pairs, there are several relevant studies. Schaid and Nick [1990], Knapp [1991], and Schaid and Nick [1991] derived an expression for the optimal sib pairs allele-sharing statistic in terms of the probabilities of sharing 0, 1, or 2 alleles under an alternative model. Knapp et al. [1994] pointed out that for affected sib pairs, using the first method for testing linkage described below in **Methods for Testing Linkage**, S

equal to the number of shared alleles (equivalent to S_{pairs} defined below) is optimal for the recessive model with full penetrance and no phenocopies, $f_2 = 1$, $f_1 = f_0 = 0$. However, they did not discuss the fact that it is optimal for many other models as well, nor that it is no longer optimal for the recessive case if there are phenocopies, as shown in Feingold and Siegmunch [1997]. These results are verified as special cases of our results below, and we give a more exhaustive list of situations in which S_{pairs} is optimal. Feingold and Siegmund [1997] includes an investigation of the power of sharing statistics for sib pairs with an emphasis on recessive and partially recessive models. For this, they use a Gaussian approximation, which is equivalent to assuming small effect size, and assume multiple unlinked genes acting additively between, but not necessarily within, loci.

Kruglyak et al. [1996] performed simulations comparing the power of two statistics, S_{pairs} and S_{all}, defined below, for a scheme where the particular pedigree was randomly determined and allowed to vary across realizations, and the method used to test linkage was the first method described below. Their results indicated that S_{all} performed much better than S_{pairs} in the dominant case and for the two complex models they consider, and that the two statistics performed equally well in the recessive case.

As to the choice of the weighting factors γ_i, Kruglyak at al. [1996] suggest equal weights (but note that they are first dividing each pedigree's statistic by its null standard deviation). Sobel and Lange [1996] suggest summing the statistics they consider, without normalizing by the standard deviation. In the case of S_{pairs}, they suggest using weight $\sqrt{2/[n(n-1)]}$, where n is the number of affecteds in the pedigree, to downweight large pedigrees. In both studies, the authors imply that these choices are *ad hoc*.

Teng and Seigmund [1997] consider both choice of sharing statistic and of weights. For relative pairs, they restrict consideration to the case of additivity within and between loci, with large-sample asymptotics, i.e. small effect size, assumed to hold. They consider a few special cases of multiple affected relatives and make the additional assumption of a two-allele model at each locus in those cases. A statistic that they find to work well can be generalized to the statistic $S_{everyone}$ considered here. We note that while this statistic may work well for the large-sample asymptotics with the particular small pedigrees considered in Teng and Siegmund [1997], if one instead uses smaller samples with larger pedigrees, $S_{everyone}$ is very sensitive to genotyping errors and loses most of its power in the presence of phenocopies or with segregation of multiple copies of the predisposing allele within a pedigree. While Teng and Siegmund [1997] consider each special case of pedigree type separately, we are instead able to describe general optimal statistics with explicit algorithms for computing them in any pedigree.

Methods for Testing Linkage

1. $\mathbf{Z^{tot}}$ [Kruglyak et al. 1996] Given a sharing statistic S, a pedigree, and a genome location x, consider the normalized version of S, $Z(c(x), \Phi) = \frac{S(c(x),\Phi)-\mu_o}{\sigma_o}$, where μ_o is the expected value of S and σ_o the standard deviation of S under the null hypothesis of no gene for the trait linked to that location. In the case of incomplete IBD data on location x,

let $\bar{Z}(x,\Phi) = \frac{\bar{S}(x,\Phi)-\mu_0}{\sigma_0}$. (With complete IBD data on x, $\bar{Z}(x,\Phi) = Z(c(x),\Phi)$.) Note that σ_o is the null standard deviation of Z, which will tend to be larger than the null standard deviation of $\bar{Z}$. Thus, inference based on $\bar{Z}$ can be overly conservative [Kong and Cox 1997]. Consider p pedigrees, with $\bar{Z}$ for the ith pedigree denoted by $\bar{Z}_i$. To combine the $\bar{Z}_i$'s for different pedigrees into an overall $\bar{Z}^{tot}$, choose appropriate weight γ_i for the ith pedigree, with $\sum_{i=1}^{p}\gamma_i^2 = 1$, and let $\bar{Z}^{tot} = \sum_{i=1}^{p}\gamma_i\bar{Z}_i$. Kruglyak et al. [1996] propose using equal weights for all pedigrees, $\gamma_i = 1/\sqrt{p}$ for all i, and they suggest comparing $\bar{Z}^{tot}$ to a standard normal distribution or computing an exact p-value for $\bar{Z}^{tot}$ in order to test linkage. Both methods are implemented in their GENEHUNTER package.

2. $\mathbf{LR^{lin}}$ Whittemore [1996] showed that in the complete data case, the test statistic $Z^{tot} = \sum_{i=1}^{p}\gamma_i Z_i$ is the *efficient score statistic* corresponding to the likelihood

$$(*) \quad L_A^{lin}(c(x),\Phi) = \Pi_{i=1}^{p}[L_O(c_i(x),\Phi_i)(1+\delta\gamma_i Z_i)] =$$

$$L_O(c(x),\Phi) \times \Pi_{i=1}^{p}(1+\delta\gamma_i Z_i).$$

Here L_A denotes the likelihood under an alternative model involving a gene at the given location contributing to the trait. The superscript lin, for "linear", denotes the particular class of alternative models given by $(*)$, L_O denotes the likelihood under the null hypothesis that no predisposing gene is linked to the given location, $c_i(x)$ is the allele-sharing configuration for the affecteds in the ith pedigree, $c(x)$ is the configuration for all the pedigrees together, and Φ_i and Φ are the affection status information for the ith pedigree and all pedigrees combined, respectively. This model is not biologically-based, but is a convenient mathematical representation of the deviation from null sharing. The parameter δ measures the magnitude of deviation of the alternative likelihood from the null likelihood in the direction specified by the $\gamma_i Z_i$'s, and δ must be estimated. Among other things, Z^{tot} being the efficient score statistic corresponding to the given likelihood implies that the test based on Z^{tot} is asymptotically equivalent to the test based on the maximized log-likelihood ratio for the given likelihood (let $\log(\hat{LR}^{lin})$ denote this maximized log-likelihood ratio). Furthermore, the framework of maximum likelihood estimation provides for computation of lod scores and creation of confidence intervals for the true gene location.

3. **LR$^{\mathbf{exp}}$** Kong and Cox [1997] have suggested a different likelihood,

$$(**) \quad L_A^{exp}(c(x), \Phi) = \Pi_{i=1}^{p}[L_O(c_i(x), \Phi_i)\; e^{\delta\gamma_i Z_i}/E_O(e^{\delta\gamma_i Z_i})] =$$

$$L_O(c(x), \Phi) \times exp(\sum_{i=1}^{p} \delta\gamma_i Z_i)/E_O[exp(\sum_{i=1}^{p} \delta\gamma_i Z_i)]$$

(where E_O denotes expected value under the null hypothesis), which also has Z^{tot} as efficient score statistic, and is another mathematically convenient representation of the deviation from null sharing. The likelihood L_A^{exp}, where *exp* is for "exponential", has an advantage over the likelihood L_A^{lin} in that for the latter, δ is restricted to the range $(-1/(\gamma Z)_{max}, -1/(\gamma Z)_{min})$, where $(\gamma Z)_{max}$ and $(\gamma Z)_{min}$ are the largest and smallest possible values, respectively, of $\gamma_i Z_i$, $i = 1, ..., p$. Thus, for models with large deviation from null sharing, power may be lost using $\log(\hat{LR}^{lin})$ when the sharing statistic S used is not close to the optimal and hence the parameter δ maximizes on the boundary. There are no such restrictions on δ in L_A^{exp}. We note that in the case of complete data, the statistic $\log(\hat{LR}^{exp})$ (the maximized log-likelihood ratio under likelihood L_A^{exp}) is just a monotone transformation of Z^{tot}, so the two methods give identical tests for linkage if exact p-values are used. This is not the case, however, for the linear model, nor for either model with incomplete data, nor if approximate p-values are used.

Guiding principles for the optimal choice of S and the γ_i's

By drawing a connection between the test statistics described above and the likelihood ratio for the affecteds under a parametric model, we can derive completely general, exact formulae for the optimal S and γ_i's. For the test based on $\log(\hat{LR}^{lin})$, the S and γ_i's given below are asymptotically most powerful against the alternative, while for the tests based on $\log(\hat{LR}^{exp})$ and Z^{tot}, the S and γ_i's given below are most powerful for any sample size.

A consequence of the work of Whittemore [1996] is that for a test based on $\log(\hat{LR}^{lin})$, the asymptotically optimal S is $S = L_A(c(x), \Phi)/L_O(c(x), \Phi) - 1$, where $L_A(c(x), \Phi)$ is the likelihood under the true alternative sharing distribution, as opposed to the mathematically convenient alternative likelihoods L_A^{lin} and L_A^{exp}. S is optimal in the sense that for a given pedigree, the choice of parameter $\delta = \sqrt{\sum_{j=1}^{p} \sigma_{0j}^2}$ in likelihood L_A^{lin} corresponds to the true alternative likelihood L_A. In that case, the likelihood ratio in the allele-sharing framework would equal the true likelihood ratio for the affecteds in the full parametric framework, giving greatest power to detect the alternative. Since the parameter δ is estimated, the equivalence of the allele-sharing and parametric likelihoods, with S chosen as above, is asymptotic. (Note that $S = b(L_A/L_O - 1) + d$ would serve just as well, where b and d are any constants). To combine pedigrees in this situation, we find that the asymptotically optimal weights are $\gamma_i = \sigma_{o\,i}/\sqrt{\sum_{j=1}^{p} \sigma_{o\,j}^2}$, where $\sigma_{o\,i}$ is the standard deviation of $S = L_A/L_O - 1$ in pedigree i under the null hypothesis. These are asymptotically optimal weights in the sense that when S and the γ_i's are so chosen, then the case $\delta = \sqrt{\sum_{j=1}^{p} \sigma_{o\,j}^2}$ in likelihood L_A^{lin} corresponds to the true alternative likelihood L_A. Thus, the likelihood ratio in the allele-sharing framework, with multiple pedigrees combined in this way, would equal the true likelihood ratio for the affecteds in the full parametric framework.

For $\log(\hat{LR}^{exp})$ and for the efficient score statistic Z^{tot}, the optimal choice of S is instead $S = \log(L_A/L_O)$ (here again, $b\log(L_A/L_O) + d$ would serve just as well), while the corresponding choices of γ_i's are the same as above, except that the null standard deviations are now for the new choice of $S = \log(L_A/L_O)$. In the case of complete data, these are non-asymptotic results. Although the non-asymptotic optimality of these S and γ_i's is not surprising for Z^{tot}, it is somewhat surprising that such a non-asymptotic result would hold in the case of $\log(\hat{LR}^{exp})$, since the parameter δ is estimated. In fact, when S and the γ_i's are so chosen, $\log(\hat{LR}^{exp})$ is a monotone transformation of the true likelihood ratio for any sample size, even though the parameter δ is estimated by maximizing the likelihood. In the case of incomplete data, the result for $\log(\hat{LR}^{exp})$ would be asymptotic, as above for the case of $\log(\hat{LR}^{lin})$. Note that when the optimal S is used and pedigrees are combined, the optimal weight $\gamma_i = \sigma_{o\,i}/\sqrt{\sum_{j=1}^{p} \sigma_{o\,j}^2}$ is equivalent to combining pedigrees on the unnormalized S scale, rather than on the normalized Z scale as was done in Kruglyak et al. 1996, i.e. Z^{tot} should be a normalized version of $\sum_i S_i$ rather than a normalized version of $\sum_i Z_i$ as in

Kruglyak et al. 1996.

Although the $\log(\hat{LR}^{lin})$ has a different optimal choice of S from the other two test statistics, these two optimal choices of S, $L_A/L_O - 1$ and $\log(L_A/L_O)$, are approximately equal for alternative models with small deviation from null sharing. However, for alternative models with large deviation from null sharing, these statistics may be quite different.

To choose the γ_i's when the S used is not the optimal S, we note that $E_A(Z^{tot})$ is maximized when γ_i is taken equal to $E_A(Z_i)/\sqrt{\sum_{j=1}^{p} E_A(Z_j)}$. This coincides with the choice of γ_i given above when the optimal S is used. For alternative models with small deviation from null sharing, when the S used is not the optimal S, the same choice of $\gamma_i \propto E_A(Z_i)$ also approximately maximizes $E_A(\log(\hat{LR}^{lin}))$ and $E_A(\log(\hat{LR}^{exp}))$.

The principles given above, which connect allele-sharing statistics with parametric likelihoods, can be applied to any specific disease model to determine the optimal S. The resulting S will be applicable to every pedigree type, not just special cases. Similarly, the principles can be applied in reverse to determine for which disease models a particular S is optimal. As described in **Methods for Testing Linkage**, Whittemore [1996] and Kong and Cox [1997] have shown the equivalence of the allele-sharing methods to likelihood-based methods using a sharing statistic and a model misfit parameter. Combining this with our results on optimal statistics, we can view allele sharing methods as equivalent to picking a particular parametric disease gene model and then introducing a parameter δ to absorb model misfit. A method will perform well when the model chosen is close to the true model, but may perform very poorly if it is far from the true model, as illustrated in Figure 1.

Exchangeability of Relatives in optimal S

All of the sharing statistics discussed below treat relatives exchangeably. By this we mean that if the genotypes of some affected individuals were permuted among them, with the two alleles of each individual's genotype treated as a unit, never separated, then, assuming that a biologically possible IBD configuration resulted, the values of the allele-sharing statistics would not be changed. For instance, consider the case of an affected sib pair with an affected first cousin, with the possible IBD configurations shown in Table I, where the allele labels are arbitrary. In configuration c_2, one sib shares an allele with the cousin, and in c_3, the sibs share one allele. Intuitively, one might think that since sharing of one allele between a sib and cousin is more unusual than sharing of one allele between the sibs, the former should receive more weight in an evaluation of linkage under many genetic models of interest, e.g. when the risk to relatives of a proband is high, and perhaps less weight when this risk is low. However, all of the allele-sharing statistics S considered here depend only on the collection of genotypes and not on which relative has which genotype, i.e. not on whether the observed sharing is between close or distant relatives (although calculation of the null mean and variance, μ_o and σ_o^2, does use the relationship information). One might think that it would be very important to take into account the closeness of the relatives who share among the affecteds, and one might see the failure to do this as a flaw in the proposed allele-sharing statistics S. Surprisingly, we can show that under rather general conditions, relatives actually should be treated in this exchangeable way. This result follows from the fact that the optimal S is some function of the likelihood ratio, here either $L_A/L_O - 1$ or $\log(L_A/L_O)$ (see Appendix A for proof).

Note that the exchangeability result does not contradict the results of Risch [1990b] and Feingold et al. [1993] that affected first-cousin pairs are more powerful for detecting linkage than affected sib pairs (assuming a single-gene model with full IBD information and at least a moderately large relative risk to offspring of affecteds). Here, the configuration c_2 contains more information than just sharing between first cousins. It also contains an affected sib pair that shares no alleles at the given location. Thus, it is weaker evidence for linkage than the observation of sharing between affected first cousins.

Conditions under which the optimal S should treat relatives exchangeably include models in which a single gene affecting the trait has h alleles with frequencies a_1, ..., a_h and penetrance f_{ij} for an individual with genotype (i,j), and also multigene models where the genes are unlinked and multiallelic, with additivity between but not necessarily within loci, and with small effect size. The result does not depend on the values of h, the a_i's, the f_{ij}'s, nor on the type of pedigree. For all of the models in this class, $P_A(\Phi|c(x)) = P_A(\Phi|c'(x))$ where $c'(x)$ is obtained from $c(x)$ by any permutation of the genotypes of the individuals, with the two alleles of each genotype treated as a unit, never separated. The models under which the result would not hold would be those under which $P_A(\Phi|c(x)) \neq P_A(\Phi|c'(x))$. Multigene models outside the class described above, and environmental effects on penetrance that could be expected to be more similar for close relatives than for more distant relatives could cause such dependence. In the case of an affected sib pair with affected first cousin, these complications could cause the optimal sharing statistic to give more weight either to c_3 (sib-sib sharing) or to c_2 (sib-cousin sharing), depending on the specific model.

In a typical allele-sharing analysis in which genome locations are tested individually, without attempting to look for interactions between loci, we argue that, among other considerations, a gene is detectable to the extent that it shows some similarity to a single gene model in at least some proportion of the families. Thus, the implication of the above result is that, in those cases when the allele-sharing method is likely to have power to detect a gene, treating relatives exchangeably is the right approach, to a first approximation, even for complex traits. Note that this result depends on an IBD analysis (incomplete data allowed) with linkage tested at every point, as in, e.g. the GENEHUNTER package [Kruglyak et al. 1996].

Definitions of Allele-Sharing Statistics

Before presenting results on optimal allele-sharing statistics, we introduce the following statistics, which can be applied to individual pedigrees with arbitrary numbers of affecteds. Table I gives sample calculations for the first seven statistics on the list for the the case of an affected sib pair with affected first cousin. (For that pedigree type, two of the statistics, S_{g-prs} and $S_{-\#geno}$ are equivalent.)

1. $\mathbf{S_{pairs}}$ [Weeks and Lange (1988); Fimmers et al. (1989); Whittemore and Halpern (1994); Kruglyak et al. (1996); Sobel and Lange (1996); Teng and Siegmund (1997)], counts,

for each pair of affected relatives, the number of alleles they share, and then sums that over all pairs of affected relatives. For a pair of relatives with respective IBD genotypes (i, j) and (k, l), the number of alleles they share is calculated as $\delta(i, k) + \delta(i, l) + \delta(j, k) + \delta(j, l)$, where $\delta(x, y) = 1$ if $x = y$, 0 otherwise.

2. $\mathbf{S_{all}}$ [Whittemore and Halpern (1994); Kruglyak et al. (1996); Teng and Siegmund (1997)] Consider a vector of length m, where m is the number of affecteds, whose ith component is one of the two alleles of the ith person at the given location. There are 2^m such possible vectors ω. For each ω, let $h(\omega) = \prod_{j=1}^{\#\text{alleles}} g_j!$, where g_j is the number of times allele j occurs in ω, i.e. $h(\omega)$ is the number of permutations that preserve ω. Define $S_{all} = 1/2^m \times \sum_{\omega \in \Omega} h(\omega)$. The value assigned to a configuration by S_{all} increases with the number of people sharing the same allele. Whittemore and Halpern [1994] proposed this statistic to weight more heavily group sharing of a single allele over pairwise sharing of different alleles by different affected pairs.

3. $\mathbf{S_{-\#alleles}}$ (negative of Statistic A in Sobel and Lange [1996]) equals -1 times the number of distinct-by-descent alleles appearing among the affecteds. Sobel and Lange [1996] suggest that this statistic would be useful for recessive traits.

4. $\mathbf{S_{everyone}}$ If all affecteds in the pedigree have a common ancestor in the pedigree, let $S_{everyone}(c)$ = the number of alleles shared by all affecteds. If not all affecteds have a common ancestor, but it is possible to choose two pedigree members such that all affecteds are descendants of at least one of them, then let $S_{everyone}(c)$ = the number of ways to choose two alleles from among those in c so that all affecteds have at least one of them, and so on. In general, if it is not possible to choose i pedigree members such that all affecteds are descendants of at least one of them, but it is possible to choose $i + 1$ such, then let

$S_{everyone}(c)$ = the number of ways to choose $i+1$ alleles from among those in c so that all affecteds have at least one of them. In certain special cases, Teng and Siegmund [1997] have proposed statistics that are equivalent to $S_{everyone}$, but they have not proposed a general definition, such as the one given here, that would be applicable to all types of pedigrees of affected relatives.

5. $\mathbf{S_{-\#geno}}$ counts -1 for each distinct genotype appearing in the observed IBD configuration of affecteds in a pedigree.

6. $\mathbf{S_{g\text{-}prs}}$ counts the number of pairs of affecteds in a pedigree who have the same genotype.

7. $\mathbf{S_{fewest}}$ equals one if the observed IBD configuration of affecteds in a pedigree contains the fewest possible distinct-by-descent alleles for that pedigree type, and it equals zero otherwise.

8. $\mathbf{S_{\#al\ triples}}$ equals the number of ways to choose three alleles i, j, k from the set of those appearing among the affecteds in a pedigree so that $(i, j), (i, k)$, and (j, k) each appear as genotypes of at least one affected.

9. $\mathbf{S}_{\#\mathit{aff}\ \mathbf{HBD}}$ (for inbred pedigrees) is the number of affected individuals who are homozygous by descent (HBD) at the given locus. Let $S_{-\#aff\ HBD} = -S_{\#aff\ HBD}$.

10. $\mathbf{S_{-\#al\ HBD}}$ (for inbred pedigrees) equals -1 times the number of distinct-by-descent alleles that occur at least once in HBD form among the affecteds in a pedigree.

11. $\mathbf{S_{rob\ dom}} = \sum_{i \in A}(7^{c1(i)} - 1)$, where A is the set of all alleles observed for the particular locus among the affecteds in the pedigree, and $c1(i)$ is equal to the number of affecteds in the pedigree with at least one copy of allele i.

Optimal S's for All Pedigree Types

The principles given above for choice of S and γ_i, giving the direct connection with the parametric likelihood, are completely general and could be applied to any specific case. What is somewhat remarkable is that for many cases of interest, the resulting S can be given in a very simple form that is applicable to all pedigree types. Following are some examples. Proofs are given in Appendix B. We assume for convenience that the penetrances satisfy $f_0 \leq f_1 \leq f_2$.

Rare dominant with phenocopies:

If a dominant model is assumed with predisposing allele frequency $a \to 0$ (i.e. each allele is introduced no more than once into a pedigree), and with phenocopy rate f_0 satisfying $f_1 > f_0 > 0$, then let $r = f_1/f_0$ be the relative risk of having the trait with and without the predisposing allele. Then the optimal allele-sharing statistic for any pedigree in this case is

$$S = \sum_{i \in A} (r^{c1(i)} - 1),$$

where A is the set of all alleles observed for the particular locus among the affecteds in the pedigree, and $c1(i)$ is equal to the number of affecteds in the pedigree with at least one copy of allele i. In practice, we find that the power to detect linkage is not very sensitive to the choice of r. We somewhat arbitrarily choose $r = 7$, and call the resulting statistic $S_{rob\,dom}$ for "robust dominant." Figures 1a-d show that $S_{rob\,dom}$ performs well for a variety of additive and dominant models with varying predisposing allele frequency and phenocopy rate.

Allele with small effect, single or multigene:

If the phenocopy rate is close to the penetrance of the homozygote carrier in the two-allele model, i.e. if $f_0 \to f_2$, then for all outbred pedigrees, the optimal sharing statistic is given by

$$\alpha\, S_{pairs} + (1-\alpha)\, S_{g\text{-}prs},$$

where $\alpha = (a\bar{m}+m\bar{a})^2/(m^2\bar{a}+\bar{m}^2 a)$, $m = (f_1-f_0)/(f_2-f_0)$, a is allele frequency, $\bar{m} = 1-m$, $\bar{a} = 1-a$, and $0 < a < 1$. This still holds if there are assumed to be multiple unlinked genes, all with small effect (i.e. $f_0 \to f_2$ at each locus), with additivity between loci, while the individual locus follows a two-allele model. This sharing statistic is optimal for tests based on any of $\log(\hat{LR}^{lin})$, $\log(\hat{LR}^{exp})$, and Z_{tot}. In the dominant case, this becomes $\bar{a}S_{pairs} + aS_{g\text{-}prs}$, in the recessive case $aS_{pairs} + \bar{a}S_{g\text{-}prs}$, and in the additive case, simply S_{pairs}. If the number of pairs sharing a genotype cannot vary among the different possible configurations of the outbred pedigree, e.g for an affected first cousin pair, where the number of shared genotypes is always 0, then when $f_0 \to f_2$, the optimal statistic is S_{pairs}. For inbred pedigrees, when $f_0 \to f_2$, the optimal statistic is $S_{\#aff\ HBD}$ when $m < 1/2$, $S_{-\#aff\ HBD}$ when $m > 1/2$, and S_{pairs} when $m = 1/2$ (additive).

Rare gene with no phenocopies:

In the dominant case with $f_0 \to 0$ and $a \to 0$, i.e. a rare dominant with no phenocopies, the optimal sharing statistic for tests based on $\log(\hat{LR}^{lin})$ is $S_{everyone}$. In outbred pedigrees in which it is possible for all affecteds to share an allele IBD, this result holds also for any

$m > 0$ (i.e. any non-recessive model). The corresponding optimal statistic in the recessive case is S_{fewest}.

Table II gives optimal allele-sharing statistics S in a number of other special cases. These results hold for arbitrary pedigrees. For outbred pedigrees in which it is not possible for a pair of affecteds to share an IBD genotype, the statistics statistics $S_{-\#geno}$ and $S_{g\text{-}prs}$ will each be the same for all possible c, and thus, are not useful as sharing statistics. As noted in the second column of sharing statistics in Table II, $S_{pairs} - S_{\#al\ triples}$ and S_{pairs} should be substituted for $S_{-\#geno}$ and $S_{g\text{-}prs}$, respectively, in such cases. With the exceptions of $S_{everyone}$ and S_{fewest}, the statistics given are optimal for tests based on Z^{tot}, $\log(\hat{LR}^{lin})$ or $\log(\hat{LR}^{exp})$. For models under which the deviation from null sharing is great, the optimal sharing statistics for tests based on Z^{tot} and $\log(\hat{LR}^{exp})$ will be different from those for $\log(\hat{LR}^{lin})$. Under such models, the same allele-sharing statistic can give substantially different power when used with $\log(\hat{LR}^{lin})$ as opposed to Z^{tot} or $\log(\hat{LR}^{exp})$, as shown in Figure 2. The cases for which $S_{everyone}$ and S_{fewest} are listed as optimal are models for which the deviation from null sharing is great. In those cases $S_{everyone}$ and S_{fewest} are optimal for the test based on $\log(\hat{LR}^{lin})$, whereas $\log(S_{fewest}/\mu_o(S_{fewest}))$ and $\log(S_{everyone}/\mu_o(S_{everyone}))$ are optimal for tests based on Z^{tot} or $\log(\hat{LR}^{exp})$. Note that in these two cases, $\log(S/\mu_o(S))$ takes on the value $-\infty$ with positive probability under the null hypothesis, corresponding to the fact that some allele-sharing configurations possible under the null hypothesis are impossible when $a \to 0$ and $f_0 \to 0$. The statistics S_{fewest} and $S_{everyone}$ would obviously not be very robust to genotyping errors or other slight deviations from the model.

When choosing the weight γ_i to assign to the normalized optimal sharing statistic Z_i from the ith pedigree using a model given above or in Table II, it is important to make the distinction between pedigrees that have the same optimal S for that model and those that do not. For instance, under model 2 in Table II, an inbred and an outbred pedigree do not have the same optimal S, while under model 6 they do. When the optimal statistic S is used and the same statistic is optimal for all pedigrees to be combined, then except for $S_{everyone}$ and S_{fewest}, the optimal weight γ_i for the normalized statistic Z_i from the ith pedigree is $\gamma_i = \sigma_{oi}(S)$. This is so because in these cases, for the given S, both $L_A(c_i)/L_O(c_i) - 1$ and $\log(L_A/L_O)$ are proportional to $S - \mu_{oi}(S)$, with the constant of proportionality not depending on i. When $S_{everyone}$ or S_{fewest} is optimal and is used, taking γ_i equal to $\sigma_{oi}(S)/\mu_{oi}(S)$ is optimal for tests based on $\log(\hat{LR}^{lin})$, because $L_A(c_i)/L_O(c_i) = S/\mu_{oi}(S)$ in these cases. (For tests based on $\log(\hat{LR}^{exp})$ and on Z^{tot} in this case, sharing statistic $\log(S/\mu_o(S))$, where $S = S_{everyone}$ or S_{fewest} respectively, is optimal. Here, $\mu_{oi}(\log(S/\mu_o(S))) = \infty$, so the statistic would not be normalized, nor would it be weighted when combined with other pedigrees. As noted before, this last statistic is entirely non-robust to deviations from the model.)

When the optimal statistic S is used for each pedigree and different pedigrees have different optimal S's, then care must be taken when combining the statistics from these pedigrees. For instance, inbred and outbred pedigrees have different optimal sharing statistics under models 2, 3, 4, 7, and 8 of Table II, and outbred pedigrees that can have variation in $S_{g\text{-}prs}$ and $S_{-\#geno}$ and those that cannot have different optimal sharing statistics under models 2, 4, and 7. In model 2, for example, $\log(L_A/L_O)$ and $L_A/L_O - 1$ are both equal to $(S_{-\#geno} - \mu_o(S_{-\#geno}))(1-a)^2 + o(1-a)^2$ for outbred pedigrees for which $S_{-\#geno}$ can take on different values. They are both equal to $(S_{pairs} - S_{\#al\,triples} - \mu_o(S_{pairs} - S_{\#al\,triples}))(1-a)^3 + o(1-a)^3$ for outbred pedigrees for which $S_{-\#geno}$ cannot vary, and they are both equal to $(S_{-\#al\,HBD} - \mu_o(S_{-\#al\,HBD}))(1-a) + o(1-a)$ for inbred pedigrees, where $\mu_o(S)$ is the expected value of S under the null hypothesis for a particular pedigree (this information on the likelihood ratio is given in Appendix B for all models discussed). Thus, in this case, outbred pedigrees have negligible value relative to inbred pedigrees, and outbred pedigrees for which $S_{-\#geno}$ cannot vary have negligible value relative to those for which it can. This is true for all models described here for which the optimal statistics are different for these two pedigree types. This is also the result if, e.g., a collection of inbred and outbred pedigrees is regarded as a single (inbred) pedigree and the optimal statistics are applied: the outbred part of the pedigree does not contribute to the sharing statistic under models 2, 3, 4, 7, and 8.

Note that the optimal weight assigned to a pedigree type can vary greatly with the model. Figure 3 gives the optimal weight for an affected sib quartet relative to an affected sib pair, assuming that the optimal S is used, under dominant and recessive models with varying allele frequency and relative risk $f_2/f_0 = 10$. In addition, for some models in which the predisposing allele frequency is high, an affected sib pair may actually receive greater weight than an affected sib trio or quartet, because the latter cases are more likely to involve multiple copies of the predisposing allele segregating in the family. (Of course, this depends on the assumption that the affection status of any additional siblings is unknown, so that the overall size of the sibship from which each affected sib pair, trio, or quartet is drawn is not a consideration.)

Example 1: Concordant and Discordant Sib Pairs

For sib pairs, $S_{everyone} = S_{pairs} = S_{-\#alleles} = S_{all}$ and $S_{g\text{-}prs} = S_{-\#geno} = S_{fewest}$. The former assign value i to pairs who share i alleles and the latter assign value 1 to pairs who share two alleles and 0 to those who do not. Consider any single-gene two-allele model parametrized by m, r, and a, where $m = (f_1 - f_0)/(f_2 - f_0)$, $m = 0$ corresponding to recessiveness, $m = 1$ corresponding to dominance, and $m = 1/2$ corresponding to additivity, $r = f_2/f_0$ is the relative risk of a homozygote carrier to a homozygote non-carrier, and a is allele frequency. Then for the test statistic $\log(\hat{LR}^{lin})$, the optimal sharing statistic S in the case of affected sib pairs is given by

$$S_{optimal} = \alpha S_{pairs} + (1 - \alpha) S_{fewest},$$

where α is as given above under **Allele with small effect**. Note that it does not depend on r. This statistic is also approximately optimal for tests based on $\log(\hat{LR}^{exp})$ and Z^{tot} when the deviation from null sharing is small. This same statistic is optimal in any outbred pedigree for models in which the relative risk r approaches 1. For discordant sib pairs, the optimal statistic is just the negative of the optimal statistic for concordant sib pairs.

There are several other common parametrizations of the allele-sharing distribution for sib pairs. To see the connections between those and the two-allele model used here, see Appendix C.

Example 2: Outbred Sib Trio

In this case, $S_{g\text{-}prs} = S_{-\#geno} + S_{fewest}$, $S_{pairs} = 2S_{everyone}$, and $S_{-\#alleles} = S_{everyone}$. For any two-allele model with $0 < a < 1$, the optimal sharing statistic for a test based on $\log(\hat{LR}^{lin})$, approximately optimal for tests based on $\log(\hat{LR}^{exp})$ and Z^{tot} when deviation from null sharing is small, is given by

$$S_{opt} \propto a\bar{a}(\bar{m} - m)(1 - 2m\bar{\rho} - \rho m^2)S_{fewest} +$$

$$(a\bar{m} + \bar{a}m)[a(1 + \bar{\rho}(1 - 4m) - 2\rho m^2) + m(m\rho + 2\bar{\rho})]S_{everyone} +$$

$$a\bar{a}(\bar{m} - m)(\bar{\rho}(\bar{m} - m) - \rho m^2)S_{-\#geno},$$

where $\rho = 1 - r^{-1}$, i.e. it is a linear combination of three statistics, giving a two-parameter model.

Example 3: Single Inbred Individual

In this case, the two possible configurations are 1 1 and 1 2, and the possible sharing statistics are $S_{\#\text{aff } HBD}$ and $S_{-\#\text{aff } HBD}$. When $0 < a < 1$ and $r > 1$, for any two-allele model, $S_{\#\text{aff } HBD}$ is optimal whenever $m < .5$. Then it is clear that $S_{-\#\text{aff } HBD}$ must be optimal whenever $m > .5$, because if the trait follows a two-allele model with $m < .5$, then the dual trait, defined as the lack of the original trait, also follows a two-allele model with $m > .5$. Single inbred individuals convey no information for linkage when the model is additive ($m = .5$).

Example 4: Lethal Embryonic

For a rare ($a \to 0$) recessive that is always lethal, the dual trait is a common ($a \to 1$) dominant with no phenocopies. Thus, model 2 in Table II applies to the surviving family members.

To What Models do S_{pairs} and S_{all} Correspond?

For some small pedigrees such as sib pairs or sib trios, these two statistics, S_{pairs} and S_{all}, coincide. For both affected sib pairs and affected sib trios, the two-allele models under which this statistic is optimal for $\log(\hat{LR}^{lin})$ are 1) any additive model or 2) any nonrecessive model ($m > 0$) with allele frequency $a \to 0$, or 3) any nondominant model ($m < 1$) with $a \to 1$. The two-allele models under which this statistic is optimal for $\log(\hat{LR}^{exp})$ and for Z^{tot} using sib pairs are 1) $m < .5$ and $r = \bar{m}^2/m^2$ (e.g. recessive model with no phenocopies) or 2) $m > .5$ and $r = [2(a\bar{m} + \bar{a}m)^2 - \bar{m}^2]/[2(a\bar{m} + \bar{a}m)^2 - m^2]$ (e.g. dominant model with $r = 1 + 1/(2\bar{a}^2)$, i.e. either low relative risk or allele frequency close to one or 3) $m = .5$ and $r \to 1$, i.e. additive with low relative risk. Note that this class of models, for which $S_{pairs} = S_{all}$ is optimal for sib pairs, using $\hat{LR}^{exp}$ and Z^{tot}, is just the class of models where the number of alleles shared by the sib pair is binomial. These results agree with the previous work of Knapp et al. [1994] who found that when testing with Z^{tot}, the optimal statistic for the single gene recessive model with no phenocopies is S_{pairs}. The results also agree with Feingold and Siegmund [1997] who found, using a Gaussian approximation which is equivalent to assuming small effect size, that S_{pairs} is no longer optimal for detecting recessive alternatives in that case, although it works well for models that are far from recessive.

We now turn to general pedigrees. For outbred pedigrees and two-allele models, we have seen that S_{pairs} is the optimal statistic for use with $\hat{LR}^{lin}$, $\hat{LR}^{exp}$, or Z^{tot} when the relative risk approaches 1 and at least one of the following holds: 1) the model is additive, 2) the predisposing allele frequency approaches 0 and the model is nonrecessive ($m \neq 0$), 3) the predisposing allele frequency approaches 1 and the model is nondominant ($m \neq 1$), or 4) it is not possible for anyone in the pedigree to share an IBD genotype. Thus, in practice, one might expect it to work well for non-recessive conditions in which the predisposing allele has small effect.

To discover if S_{all} is optimal for any two-allele models in general pedigrees, we have performed adaptive searches of the two-allele-model parameter space for various pedigrees. S_{all} does not appear to be exactly optimal for any two-allele model in general pedigrees. When $\hat{LR}^{lin}$ is used as a test statistic, the two-allele model for which the optimal sharing statistic (i.e. $LR - 1$) most closely matches S_{all} is always an additive model with different predisposing allele frequencies and relative risks for different pedigrees, but empirically with allele frequency a in the range of .03 to .15 and relative risk f_2/f_0 in the range of 5.7 to 8.1. However, when $\hat{LR}^{exp}$ or Z^{tot} is used as a test statistic, the optimal sharing statistic ($\log(LR)$) is in general not as close to S_{all} as $LR - 1$ can be, and the closest fits vary widely with the pedigree types.

Discussion

We have investigated the correspondence between allele-sharing statistics and the two-allele models for which they are optimal, with extension to multigene models with unlinked loci, additivity between loci, and small gene effects. From an understanding of this connection, the robust affected relative methods of Kruglyak et al. [1996], Whittemore [1996], and Kong and Cox [1997] can be seen as equivalent to picking a particular parametric disease gene model (or class of models) and introducing a parameter δ to absorb model misfit. They are not fundamentally different from parametric linkage methods except in the particular parametric form in which the model misfit is specified. (In single-point parametric linkage methods, the recombination fraction parameter θ was often, in effect, the model misfit parameter.) In practice, much of the robustness of the affected relative method is due to choices such as using affecteds only or using sib pairs, which reduce the dimension of the parameter space.

We find that for any single gene model, the optimal S treats relatives exchangeably. This result also extends to multigene models with unlinked loci, additivity between loci, and small gene effects. Thus, even when robust affected relative methods are applied to extended families, there is no need for the sharing statistic S to take into account whether it is the close or the more distant relatives in a family who exhibit sharing. We argue that in cases in which the robust affected relative methods discussed here are likely to have power to detect a gene, the exchangeability result should still provide a useful approximate rule of thumb, even if the true model does not fall into the above classes.

We are able to find simple expressions for the optimal S, applicable to any pedigree type, for a variety of two-allele and some multigene models. While previous theoretical work in this area depends on asymptotic scenarios and small effect sizes [e.g. Kong and Cox (1997), Teng and Siegmund (1997)], the theory given here applies to realistic sample sizes with large effects as well. We propose a new statistic, $S_{rob\,dom}$, which is easy to compute and robust across a variety of models. Our power calculations for the case of affected sib pair with affected parent (Figure 1a-d) give the following order of performance, in decreasing order of power, against a variety of dominant and additive models: $S_{rob\,dom}$, S_{all}, S_{pairs}, $S_{-\#alleles}$, while for recessive models (Figures 1e and f), this order is reversed. We have done similar calculations for a variety of outbred pedigree types, and have found first, that the orderings in terms of power given above for the four statistics hold more or less across the board, and second, that the differences in power among the statistics may be quite small or quite large, depending on the particular pedigree (results not shown). In the case of a large inbred pedigree, $S_{rob\,dom}$ was found to be powerful against a wide variety of dominant, additive, and recessive alternative models, although S_{pairs} performed slightly better in the recessive case [Mark Abney, unpublished results]. In the case of the large inbred pedigree used, exact computation of S_{all} was impossible, so this statistic was not considered.

These results suggest use of $S_{rob\,dom}$ in practice, especially for non-recessive models. In many cases, the power of S_{all} will be nearly equivalent to $S_{rob\,dom}$, but S_{all} is more difficult to calculate. In the recessive case, $S_{rob\,dom}$ and S_{all} may or may not perform well, depending on the pedigree type. There is no one statistic that performs well over all disease models in general, but S_{pairs} is perhaps the compromise choice. As seen in Figure 1, it maintains a similar level of performance over many disease models, although that level may be very low for some pedigree types. Another advantage of S_{pairs} is that its distribution is much less skewed than those of $S_{rob\ dom}$ and S_{all}, so the normal approximation is much more accurate for calculating p-values. In outbred pedigrees, $S_{-\#alleles}$ is also a good choice in the recessive case only.

Kruglyak et al. [1996] performed a simulation study comparing S_{pairs} and S_{all}. For their particular simulation scheme, with the pedigree randomly determined and allowed to vary across realizations, they found that S_{all} performed much better than S_{pairs} in the dominant case and for the two complex models they consider, and that the two statistics performed equally well in the recessive case. Our findings are not necessarily inconsistent with theirs, but we would caution that the statistic S_{all}, and likewise $S_{rob\ dom}$, has a very skewed distribution and thus great care must be taken not to overrate its power if approximations are used. The calculations shown here result from consideration of every possible outcome, with no simulation or approximation involved.

The power calculations shown here are single-locus calculations, rather than taking into account testing on a whole region of the genome as in Feingold et al. [1993], Feingold and Siegmund [1997] and Teng and Siegmund [1997]. This simplification allows us to consider non-asymptotic as well as asymptotic models, with a unified approach that is applicable to all pedigree types, without requiring separate analysis of many special cases of relationship.

Simulation studies have indicated that the relative performance of the statistics changes little when one takes into account testing across a region of the genome [Mark Abney, unpublished results].

For models with nonnegligible deviation from null sharing, the choice of Z^{tot} or $\log(\hat{LR}^{exp})$ on the one hand or $\log(\hat{LR}^{lin})$ on the other, as the basis for a test of linkage, can also affect which statistics are optimal against which alternative models. Similarly, we find that the optimal choice of weights γ_i can be heavily influenced by the model. We find that when the optimal sharing statistic S is used, pedigrees are appropriately combined by adding the values of S, rather than the normalized values Z as in Kruglyak et al. [1996]. In other words, γ_i should be taken to be proportional to the null standard deviation of S in the ith pedigree. When a non-optimal S is used, the optimal weight γ_i depends on the model and is approximately proportional to $E_A(Z_i)$.

In the special case of weights for relative pairs, Teng and Siegmund's [1997] approach is equivalent to using weights proportional to $E_A(Z)$. This coincides with our suggestion above for the case of non-optimal S. Since the alternative model is unknown, they suggest using a crude estimate of λ_O, the relative risk to an offspring of an affected, to calculate the weights. They find that a choice of $\hat{\lambda}_O = 4$ works well in a variety of scenarios. For pedigrees with more than two affecteds, Teng and Siegmund [1997] consider selected examples and find ways to convert them, on a case-by-case basis, to effective numbers of different kinds of relative pairs. Their empirical results on optimal weightings in specific examples are consistent with our recommendation that for the optimal S, pedigrees should be combined on the S scale, before dividing by the null standard deviations.

Finally, for the statistic S_{pairs}, we have described the two-allele models for which it is optimal. S_{all} does not appear to be optimal for any two-allele model in general, but the closest fits occurred among additive models.

Acknowledgments

This work was supported in part by National Institutes of Health grant R29 HG01645-01 This manuscript was prepared using computer facilities supported in part by National Science Foundation grant DMS 89-05292 awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

Appendix A: *Derivation of Exchangeability Result*

Let c denote the sharing configuration among the affecteds in the pedigree at a particular location x in the genome; Φ the affection status, where the unaffecteds are coded as "unknown", e.g. for an affected sib pair with an additional unaffected sib, Φ would denote the event that the two sibs are affected; $L_O(\cdot)$ the likelihood under the null hypothesis that a gene at location x has no effect on the trait; $L_A(\cdot)$ the likelihood under the true alternative model, assumed to involve a single gene at location x. Note that $L_A(c,\Phi)/L_O(c,\Phi) = P_A(c)P_A(\Phi|c)/[P_O(c)P_O(\Phi|c)]$, and using the fact that $P_A(c) = P_O(c)$ and $P_A(\Phi) = P_O(\Phi) = P_O(\Phi|c)$, we get $L_A(c,\Phi)/L_O(c,\Phi) = P_A(\Phi|c)/P_A(\Phi)$, where $P_A(\Phi) = \sum_c P_O(c)P_A(\Phi|c) = E_O(P_A(\Phi|c)|\Phi)$.

Consider two pedigrees that are identical in terms of structure and affected members, but differ only in their IBD sharing. Further, assume that the configuration of IBD sharing among affecteds in one pedigree can be obtained from the configuration of IBD sharing among affecteds in the other pedigree by permutation of individuals' IBD genotypes, where the two alleles of an individual's genotype are treated as a unit, never separated. Let c_1 be the IBD configuration in pedigree 1 and c_2 be the IBD configuration in pedigree 2. Then in each case, for the optimal S, we have $S(c_i) = f(L_A(c_i,\Phi)/L_O(c_i,\Phi)) = f(P_A(\Phi|c)/P_A(\Phi))$, where $f(x) = x - 1$ or $\log(x)$. $P_A(\Phi)$ is clearly the same for both pedigrees. The assumption of conditional independence of phenotypes given genotypes and the fact that the IBD genotypes in one pedigree can be obtained from the other by permuting individuals imply that $P_A(\Phi|c)$ is the same for both pedigrees as well. Thus, the sharing score assigned to the two different c's by the most powerful sharing function S must be the same, i.e. affected relatives within a pedigree should be treated as exchangeable.

The extension to multiple unlinked genes with additivity between loci and small effect at each locus, i.e. $f_0 \to f_2$ at each locus, follows from a Taylor expansion of the likelihood ratio around $\rho = 1 - f_0/f_2$ for each locus. Details are available from the author.

Appendix B: *Derivation of Optimal Sharing Statistics Against Particular Alternatives*

Applying the connection between optimal allele-sharing statistics and parametric likelihoods laid out in the **Guiding principles...** section, it remains to calculate a general expression for the likelihood ratio under each model. Let k denote the number of affected individuals in the pedigree; a the frequency of the predisposing allele; f_i the probability of being affected given i copies of the predisposing allele, $i = 0, 1, 2$; $\#al(c)$ the number of distinct-by-descent alleles occurring in configuration c. Using the fact that $L_A(c,\Phi)/L_O(c,\Phi) = P_A(\Phi|c)/P_A(\Phi)$, where

$P_A(\Phi) = \sum_c P_O(c)P_A(\Phi|c) = E_O(P_A(\Phi|c)|\Phi)$, we get the following cases by simple Taylor expansion of the likelihood:

1. **Rare dominant with phenocopies** (i.e. $a \downarrow 0$, $0 < f_0 < f_1 = f_2 \leq 1$): Then, letting $S = \sum_{i\in A}(r^{c1(i)} - 1)$, we get $L_A(c,\Phi)/L_O(c,\Phi) = 1 + a[S - E_O(S)] + o(a)$.

2. **Allele with small effect** (i.e. $f_0 \uparrow f_2 \leq 1$, $0 < a < 1$), **single or multigene with loci unlinked and additivity between loci**: If the pedigree is outbred, $L_A(c,\Phi)/L_O(c,\Phi) =$

$1+a\bar{a}\rho^2[(a\bar{m}+\bar{a}m)^2(S_{pairs}(c)-E_O(S_{pairs}))+a\bar{a}(\bar{m}-m)^2(S_{g\text{-}prs}(c)-E_O(S_{g\text{-}prs}))]+o(\rho^2)$, where $\bar{a}=1-a$, $\bar{m}=1-m$, with $S_{g\text{-}prs}(c)-E_O(S_{g\text{-}prs})=0$ if it is not possible to have variation in the number of pairs of affecteds who share a genotype. If the pedigree is inbred and the model is not additive ($m \neq .5$), then $L_A(c,\Phi)/L_O(c,\Phi)=1+\rho a\bar{a}(\bar{m}-m)S_{\#\text{aff }HBD}+o(\rho^2)$ If the pedigree is inbred and the model is additive, then $L_A(c,\Phi)/L_O(c,\Phi)=1+\frac{1}{4}\rho^2 a\bar{a}[S_{pairs}-E_O(S_{pairs})]+o(\rho^2)$.

3. **Rare dominant, no phenocopies** (i.e. $f_0=0<f_1=f_2\leq 1, a\downarrow 0$): For $1\leq i\leq \#al(c)$, let $b_i(c)$ denote the number of ways to choose i distinct-by-descent alleles from among those in c so that all of the affecteds have at least one of the i. (Note that when $i=\#al(c)$, $b_i(c)$=1.) Let $d(c)$ be the smallest $i, 1\leq i\leq \#al(c)$, such that $b_i(c)>0$. Let $d=$ the smallest possible value of $d(c)$ for the particular type of affected relatives, e.g. if all affecteds have a common ancestor in the pedigree, $d=1$. If there is not a single common ancestor, but everyone is descended from at least one of two ancestors in the pedigree, then $d=2$, and so on. Let $S_{everyone}(c)=b_d(c)$. Then $\lim_{a\to 0} L_A(c,\Phi)/L_O(c,\Phi)=S_{everyone}(c)/E_O(S_{everyone})$, where $E_O(S_{everyone})=\sum_{c'}P_O(c')S_{everyone}(c')$. The same result is obtained for rare non-recessive with no phenocopies in outbred pedigrees in which it is possible for everyone to share an allele IBD.

4. **Dominant, no phenocopies, predisposing allele frequency approaches one** (i.e. $f_0=0<f_1=f_2\leq 1, a\uparrow 1$): If the pedigree is outbred and it is possible to have variation in the number of genotypes present, then

$L_A(c,\Phi)/L_O(c,\Phi)=1+[S_{-\#geno}(c)-E_O(S_{-\#geno})](1-a)^2+o(1-a)^2$. If the pedigree is outbred and it is not possible to have variation in the number of genotypes present, e.g. if it is not possible for any pair of affecteds to have the same IBD genotype, then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + [S_{pairs}(c) - S_{\#al\ triples} - E_O(S_{pairs}) + E_O(S_{\#al\ triples})](1-a)^3 + o(1-a)^3$. If the pedigree is inbred, then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + [S_{-\#al\ HBD}(c) - E_O(S_{-\#al\ HBD})](1-a) + o(1-a)$.

5. **Rare dominant with small effect** ($f_0 \uparrow f_1 = f_2 \leq 1$, $a \downarrow 0$), **single or multigene with loci unlinked and additivity between loci**: Let $\rho = 1 - f_0/f_2$. If the pedigree is outbred, $L_A(c,\Phi)/L_O(c,\Phi) = 1 + a\rho^2[S_{pairs}(c) - E_O(S_{pairs})] + o(a^3) + o(a^2\rho) + o(\rho^3)$. If the pedigree is inbred, $L_A(c,\Phi)/L_O(c,\Phi) = 1 - a\rho[S_{\#\text{aff}\ HBD}(c) - E_O(S_{\#\text{aff}\ HBD})] + o(a^2) + o(\rho^2)$.

6. **Dominant with small effect; predisposing allele frequency approaches one** ($f_0 \uparrow f_1 = f_2 \leq 1$, $a \uparrow 1$), **single or multigene with loci unlinked and additivity between loci**: Let $\rho = 1 - f_0/f_2$. If the pedigree is outbred and it is possible to have variation in the number of pairs of affecteds who share a genotype, then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + (1-a)^2\rho^2[S_{g\text{-}prs}(c) - E_O(S_{g\text{-}prs})] + o(1-a)^4 + o((1-a)^2\rho^2) + o(\rho^4)$. If the pedigree is outbred and it is not possible to have variation in the number of genotypes present, then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + (1-a)^3\rho^2[S_{pairs}(c) - E_O(S_{pairs})] + o(1-a)^5 + o((1-a)^3\rho^2) + o((1-a)^2\rho^3) + o(\rho^5)$. If the pedigree is inbred, then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + [S_{-\#\text{aff}\ HBD}(c) - E_O(S_{-\#\text{aff}\ HBD})](1-a)\rho + o(1-a)^2 + o(\rho^2)$.

7. **Rare recessive, no phenocopies** ($0 = f_0 = f_1 < f_2 \leq 1$, $a \downarrow 0$): Then $\lim_{a\to 0} L_A(c,\Phi)/L_O(c,\Phi) = S_{fewest}/E_O(S_{fewest})$.

8. **Recessive, no phenocopies, predisposing allele frequency approaches one** ($0 = f_0 = f_1 < f_2 \leq 1$, $a \uparrow 1$): Then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + (1-a)[S_{-\#al}(c) - E_O(S_{-\#al})] + o(1-a)^2$.

9. **Rare recessive with small effect** ($f_0 = f_1 \uparrow f_2 \leq 1$, $a \downarrow 0$), **single or multigene with loci unlinked and additivity between loci**: If the pedigree is outbred and it is possible to have variation in the number of pairs of affecteds who share a genotype, then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + a^2\rho^2[S_{g\text{-}prs}(c) - E_O(S_{g\text{-}prs})] + o(a^4) + o(a^2\rho^2) + o(\rho^4)$. If the pedigree is outbred and it is not possible to have variation in the number of pairs of affecteds who share a genotype, then $L_A(c,\Phi)/L_O(c,\Phi) = 1 + a^3\rho^2[S_{pairs}(c) - E_O(S_{pairs})] + o(a^5) + o(a^3\rho^2) + o(a^2\rho^3) + o(\rho^5)$. If the pedigree is inbred,$L_A(c,\Phi)/L_O(c,\Phi) = 1 + a\rho[S_{\#\text{aff}\ HBD}(c) - E_O(S_{\#\text{aff}\ HBD})] + o(a^2) + o(\rho^2)$.

10. **Recessive with small effect; predisposing allele frequency approaches one** ($f_0 = f_1 \uparrow f_2 \leq 1$, $a \uparrow 1$), **single or multigene with loci unlinked and additivity between loci**:

If the pedigree is outbred, $L_A(c,\Phi)/L_O(c,\Phi) = 1 + (1-a)\rho^2[S_{pairs}(c) - E_O(S_{pairs})] + o(1-a)^3 + o((1-a)^2\rho) + o(\rho^3)$. If the pedigree is inbred, $L_A(c,\Phi)/L_O(c,\Phi) = 1 + (1-a)\rho[S_{\#\text{aff}\ HBD}(c) - E_O(S_{\#\text{aff}\ HBD})] + o(1-a)^2 + o(\rho^2)$.

Appendix C: *Alternative Parametrizations of Sharing Distribution for Sib Pairs*

Note that the sharing distribution for affected sib pairs involves only two independently varying quantities, $P(\text{share } 2|\text{both sibs affected})$ and $P(\text{share } 1|\text{both sibs affected})$, with $P(\text{share } 0|\text{both sibs affected}) = 1 - P(\text{share } 2|\text{both sibs affected}) - P(\text{share } 1|\text{both sibs affected})$. This may be parametrized by $0 \leq \alpha \leq 1$ and $0 \leq \delta \leq \alpha$, where $P(\text{share } 2|\text{both sibs affected}) = (1-\alpha)/4$, $P(\text{share } 1|\text{both sibs affected}) = (1-\delta)/2$, and $P(\text{share } 0|\text{both sibs affected}) = (1+\alpha+2\delta)/4$, e.g. as in Feingold and Siegmund [1997]. Alternatively, it may be parametrized by λ_S, the relative risk to siblings of affecteds, and λ_O, the relative risk to offspring of affecteds, where $\alpha = 1 - 1/\lambda_S$ and $\delta = 1 - \lambda_O/\lambda_S$ [Risch 1990b]. For a single-gene two-allele model, if we let $n_0 = [a^2 + 2a\bar{a}(1-\rho\bar{m}) + \bar{a}^2\bar{\rho}]^2$, $n_1 = a[a + \bar{a}(1-\rho\bar{m})]^2 + \bar{a}[a(1-\rho\bar{m}) + \bar{a}\bar{\rho}]^2$, and $n_2 = a + \bar{a}\bar{\rho}^2$, then $\alpha = 1 - n_0/(.25n_0 + .5n_1 + .25n_2)$ and $\delta = 1 - n_1/(.25n_0 + .5n_1 + .25n_2)$. Alternatively, the two-allele model may be parametrized by K, the population prevalence of the trait, V_A, the additive variance of the trait, and V_D, the dominant variance of the trait, with $\alpha = (V_A/2 + V_D/4)/(K^2 + V_A/2 + V_D/4)$ and $\delta = (V_D/4)/(K^2 + V_A/2 + V_D/4)$ [Suarez et al. 1978]. Feingold and Siegmund [1997] point out that the two-allele assumption is not necessary for these last formulae to hold.

References

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393-399.

Day NE, Simons MJ (1976) Disease susceptibility genes — their identification by multiple case family studies. Tissue Antigens 8:109-119

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234-251

Feingold E, Siegmund D (1997) Strategies for mapping heterogeneous recessive traits by allele-sharing methods. Am J Hum Genet 60:965-978

Fimmers R, Seuchter SA, Neugebauer M, Knapp M, Baur MP (1989) Identity-by-descent analysis using all genotype solutions. In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based on affected pedigree members: Genetic Analysis Workshop 6, Liss, New York pp 123-128

Fishman PM, Suarez B, Hodge SE, Reich T (1978) A robust method for the detection of linkage in familial diseases. Am J Hum Genet 30:308-321

Green JR, Woodrow JC (1977) Sibling method for detecting HLA-linked genes in disease. Tissue Antigens 9:31-35

Hodge SE (1984) The information contained in multiple sibling pairs Genet Epidemiol 1:109-22

Knapp M (1991) A powerful test of sib-pair linkage for disease susceptibility. Genet Epidemiol 8:141-143

Knapp M, Seuchter SA, Baur MP (1994) Linkage analysis in nuclear families. Hum Hered 44:44-51

Kong A, Cox NJ (1997) Allele Sharing Models: Lodscores and Accurate Linkage Tests. Am J Hum Genet 61:1179-1188

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347-1363

Lange K (1986) The affected sib-pair method using identity by state relations. Am J Hum Genet 39:148-150

Penrose L S (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6:133-38

Risch N (1990b) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229-241

Schaid DJ and Nick TG (1990) Sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 7:359-370

Sobel E and Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323-1337

Suarez BK (1978) The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. Tissue Antigens 12:87-93

Teng J and Siegmund D (1997) Combining information within and between pedigrees for mapping complex traits. Am J Hum Genet 60:979-992

Thompson E (1974) Gene indentities and multiple relationships. Biometrics 30:667-680

Weeks D and Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315-326

Whittemore AS (1996) Genome scanning for linkage: an overview. Am J Hum Genet 59:704-716

Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118-127

Table I Example 1: Outbred Sib Pair and First Cousin

Configuration (sib,sib,cousin)	Null Prob.	$S_{pairs} - \mu_0$	$S_{all} - \mu_0$	$S_{-\#alleles} - \mu_0$	$S_{everyone} - \mu_0$	$S_{-\#geno} - \mu_0$	$S_{fewest} - \mu_0$
c_1 1 2 3 4 5 6	.125	-1.5	-.41	-1.375	-.125	-.25	-.0625
c_2 1 2 3 4 1 5	.125	-.5	-.16	-.375	-.125	-.25	-.0625
c_3 1 2 1 3 4 5	.3125	-.5	-.16	-.375	-.125	-.25	-.0625
c_4 1 2 1 3 2 4	.125	.5	.09	.625	-.125	-.25	-.0625
c_5 1 2 1 2 3 4	.1875	.5	.09	.625	-.125	.75	-.0625
c_6 1 2 1 3 1 4	.0625	1.5	.59	.625	.875	-.25	-.0625
c_7 1 2 1 2 2 3	.0625	2.5	.84	1.625	.875	.75	.9375

Table II Optimal Sharing Statistics in Special Cases

Model	Outbred Pedigree ($S_{-\#geno}$, $S_{g\text{-}prs}$ can vary)	Outbred Pedigree ($S_{-\#geno}$, $S_{g\text{-}prs}$ cannot vary)	Inbred Pedigree
1. dominant, $f_0 \to 0$, $a \to 0$	$S_{everyone}$	$S_{everyone}$	$S_{everyone}$
2. dominant, $f_0 \to 0$, $a \to 1$	$S_{-\#geno}$	$S_{pairs} - S_{\#al\ triples}$	$S_{-\#al\ HBD}$
3. dominant $a \to 0$, $f_0 \to f_2$, single or multigene	S_{pairs}	S_{pairs}	$S_{-\#aff\ HBD}$
4. dominant, $f_0 \to f_2$, $a \to 1$, single or multigene	$S_{g\text{-}prs}$	S_{pairs}	$S_{-\#aff\ HBD}$
5. recessive, $f_0 \to 0$, $a \to 0$	S_{fewest}	S_{fewest}	S_{fewest}
6. recessive, $f_0 \to 0$, $a \to 1$	$S_{-\#alleles}$	$S_{-\#alleles}$	$S_{-\#alleles}$
7. recessive, $f_0 \to f_2$, $a \to 0$, single or multigene	$S_{g\text{-}prs}$	S_{pairs}	$S_{\#aff\ HBD}$
8. recessive, $f_0 \to f_2$, $a \to 1$, single or multigene	S_{pairs}	S_{pairs}	$S_{\#aff\ HBD}$
9. additive, $f_0 \to f_2$, single or multigene	S_{pairs}	S_{pairs}	S_{pairs}

Figure Legends

Figure 1a: Sib pair plus parent: power of various sharing statistics S against a dominant alternative with varying predisposing allele frequency and no phenocopies

Figure 1b: Sib pair plus parent: power of various sharing statistics S against a dominant alternative with varying phenocopy rate and predisposing allele frequency .02

Figure 1c: Sib pair plus parent: power of various sharing statistics S against an additive alternative with varying predisposing allele frequency and no phenocopies

Figure 1d: Sib pair plus parent: power of various sharing statistics S against an additive alternative with varying phenocopy rate and predisposing allele frequency .02

Figure 1e: Sib pair plus parent: power of various sharing statistics S against a recessive alternative with varying predisposing allele frequency and no phenocopies

Figure 1f: Sib pair plus parent: power of various sharing statistics S against a recessive alternative with varying phenocopy rate and predisposing allele frequency .02

(Legend for all of Figure 1):

In Figures 1 a)-f), sample size = 30, power is computed at a single point assumed to have no recombination with the gene, significance level $= 2 \times 10^{-5}$, and exact p-values are computed using Z^{tot} or equivalently $\log(LR^{exp})$.

Figure 2a: Sib quartet: power of S_{all} against a recessive alternative using different testing methods

Figure 2b: Sib quartet: power of $.85 \times S_{fewest} + .15 \times S_{-\#alleles}$ against a recessive alternative using different testing methods

(Legend for all of Figure 2):

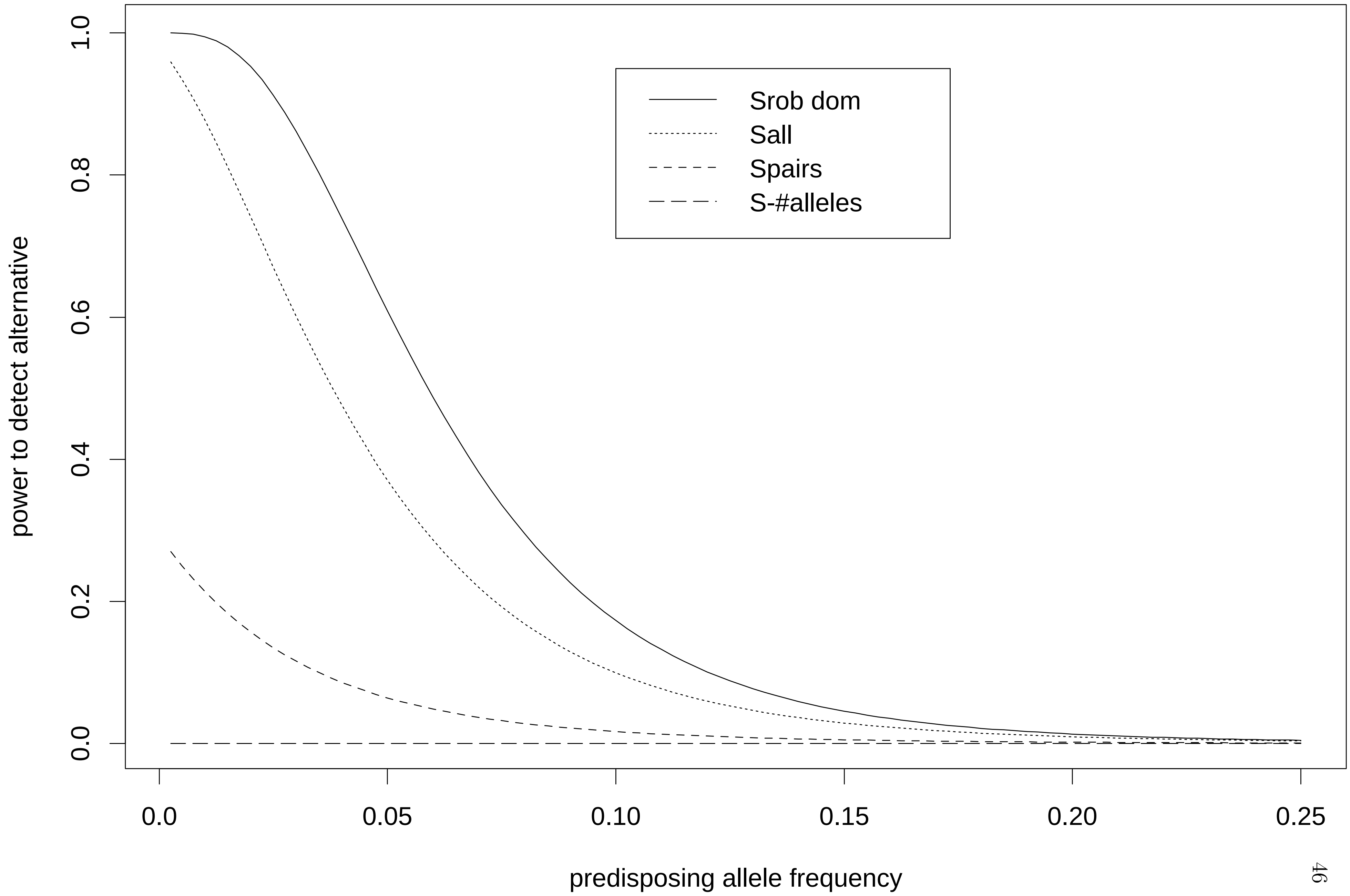
In both Figures 2a and 2b, sample size = 20, phenocopy rate = 0, power is computed at a single point assumed to have no recombination with the gene, significance level = 2×10^{-5}, exact p-values computed.

Figure 3

Optimal weight of sib quartet relative to sib pair for different models, using the optimal sharing statistic S

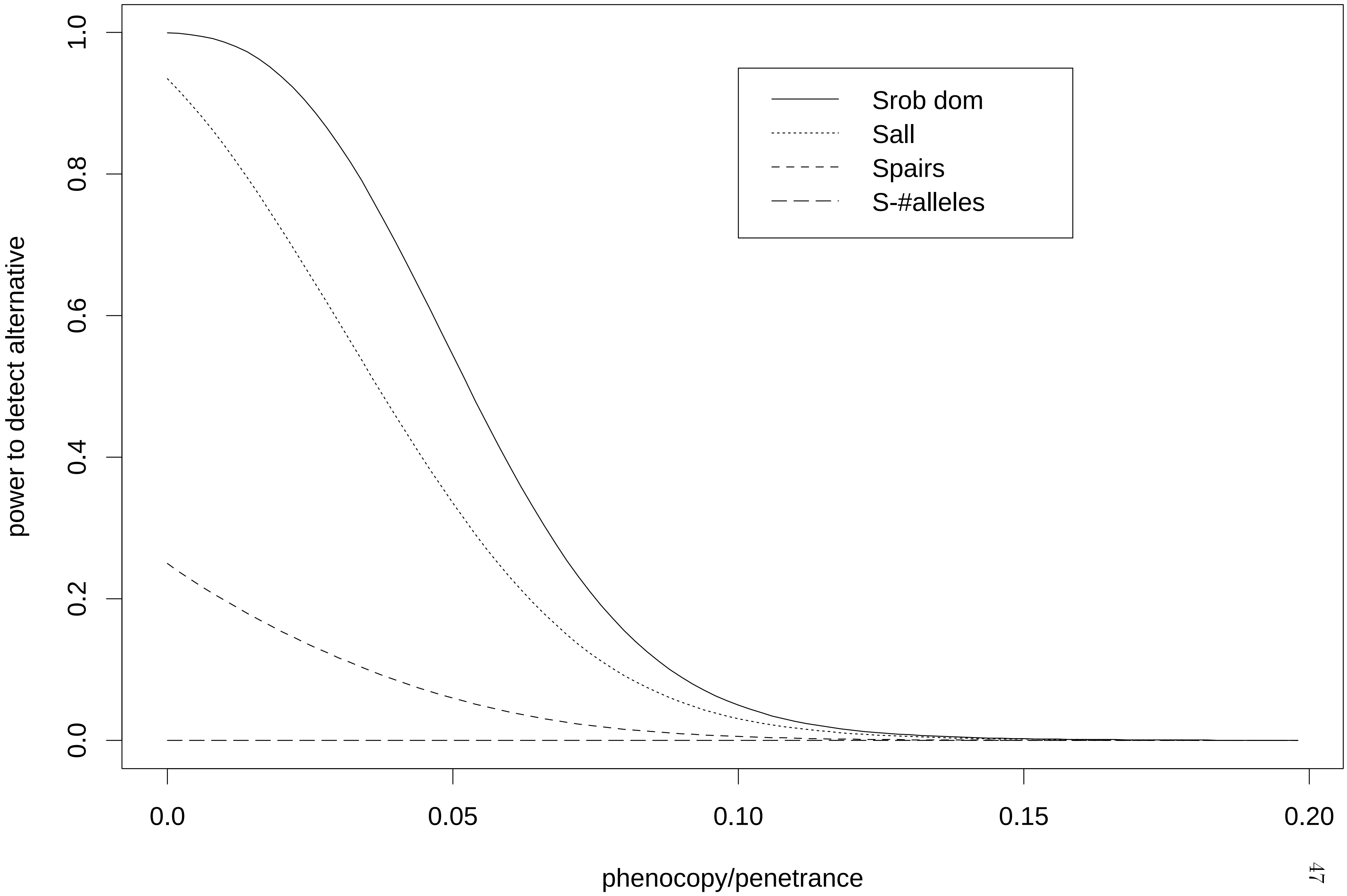
For both the dominant and recessive cases, the risk to a homozygote carrier relative to a homozygote non-carrier is assumed to be 10. The results are for Z^{tot} or, equivalently, $\log(LR^{exp})$.

Figure 1 a



Srob dom
Sall
Spairs
S-#alleles
power to detect alternative
predisposing allele frequency
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.05
0.10
0.15
0.20
0.25

Figure 1 b



power to detect alternative
1.0
0.8
0.6
0.4
0.2
0.0
Srob dom
Sall
Spairs
S-#alleles
0.0
0.05
0.10
0.15
0.20
phenocopy/penetrance

Figure 1 c

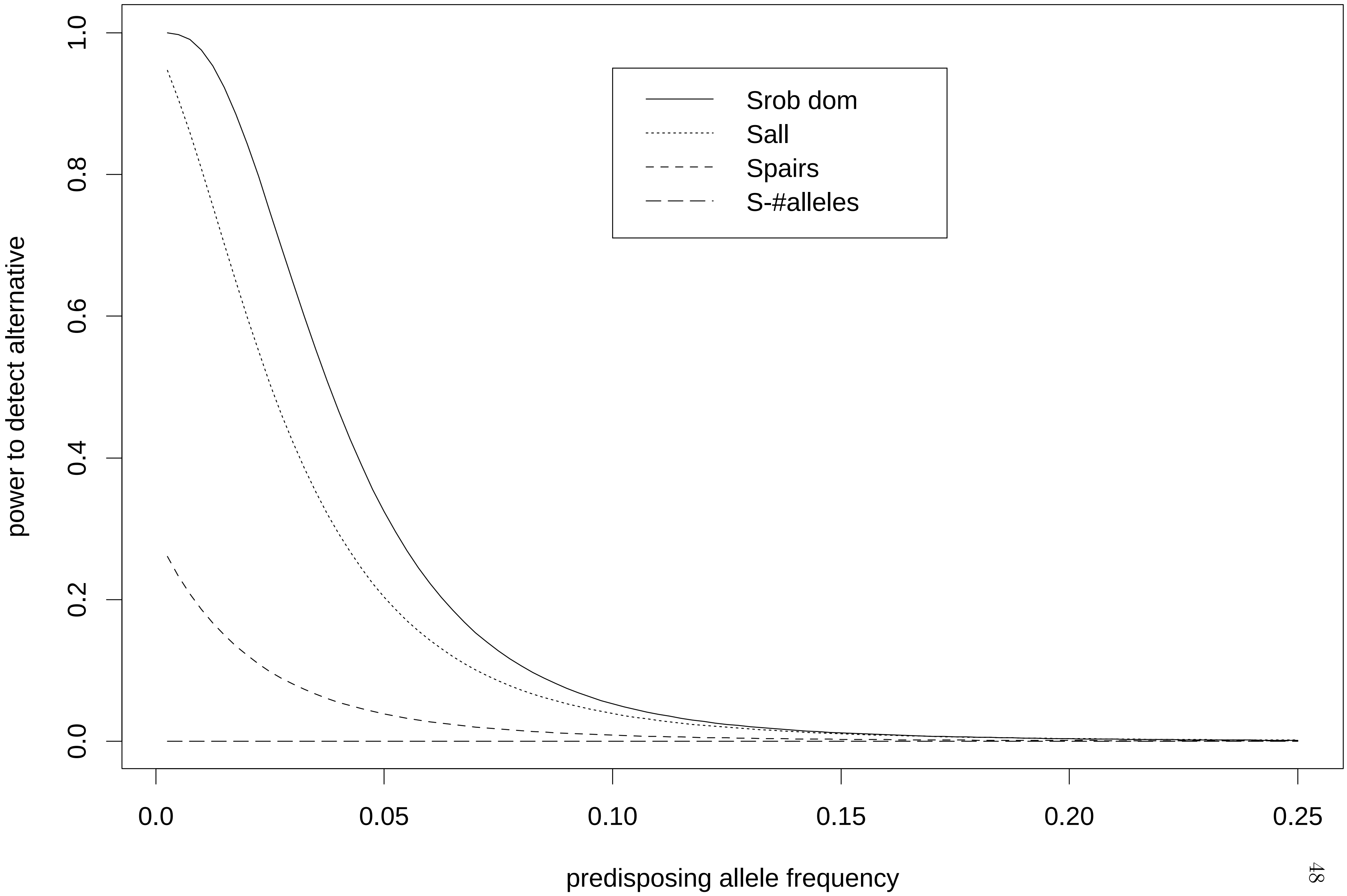

Srob dom
Sall
Spairs
S-#alleles
power to detect alternative
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.05
0.10
0.15
0.20
0.25
predisposing allele frequency

Figure 1 d

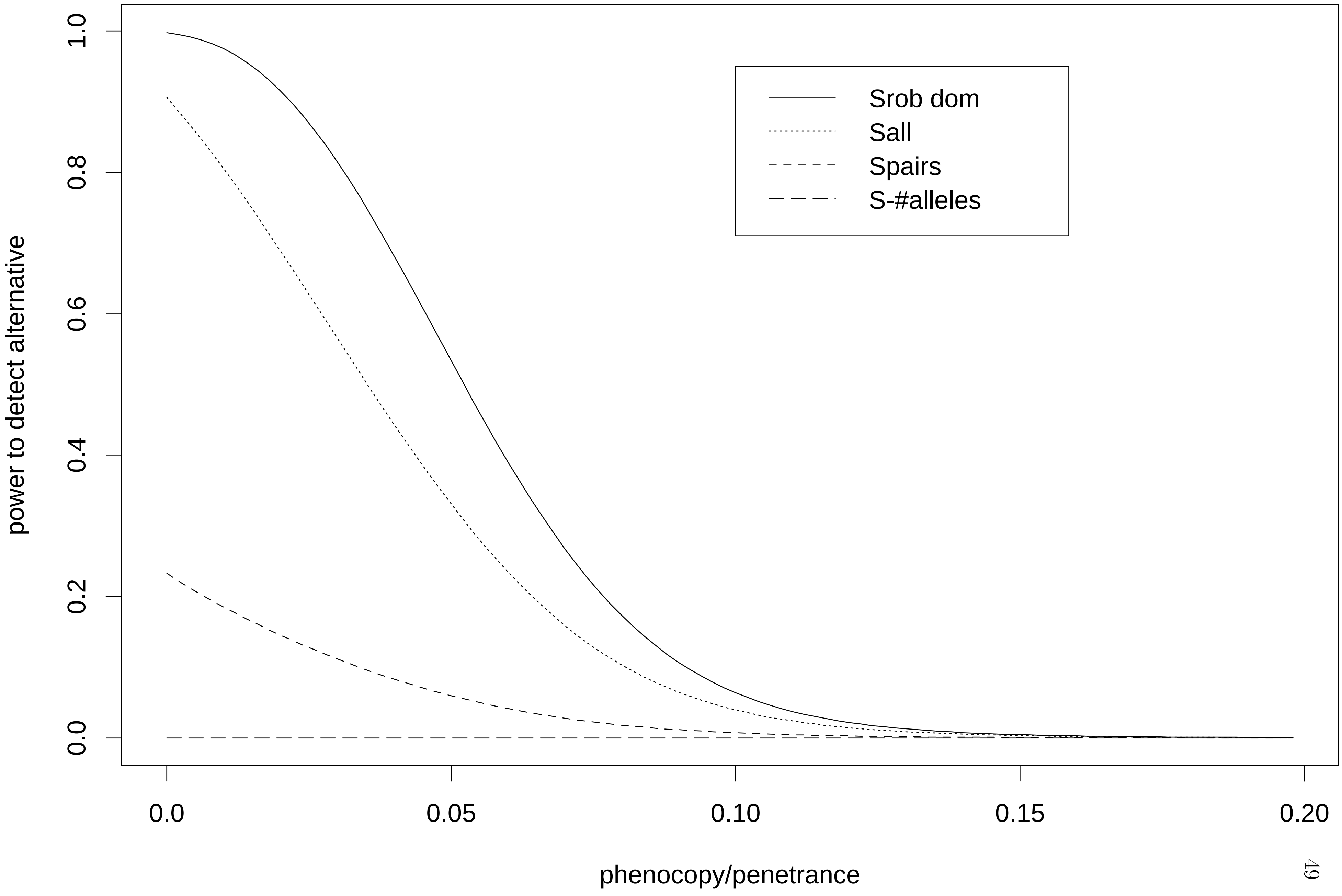
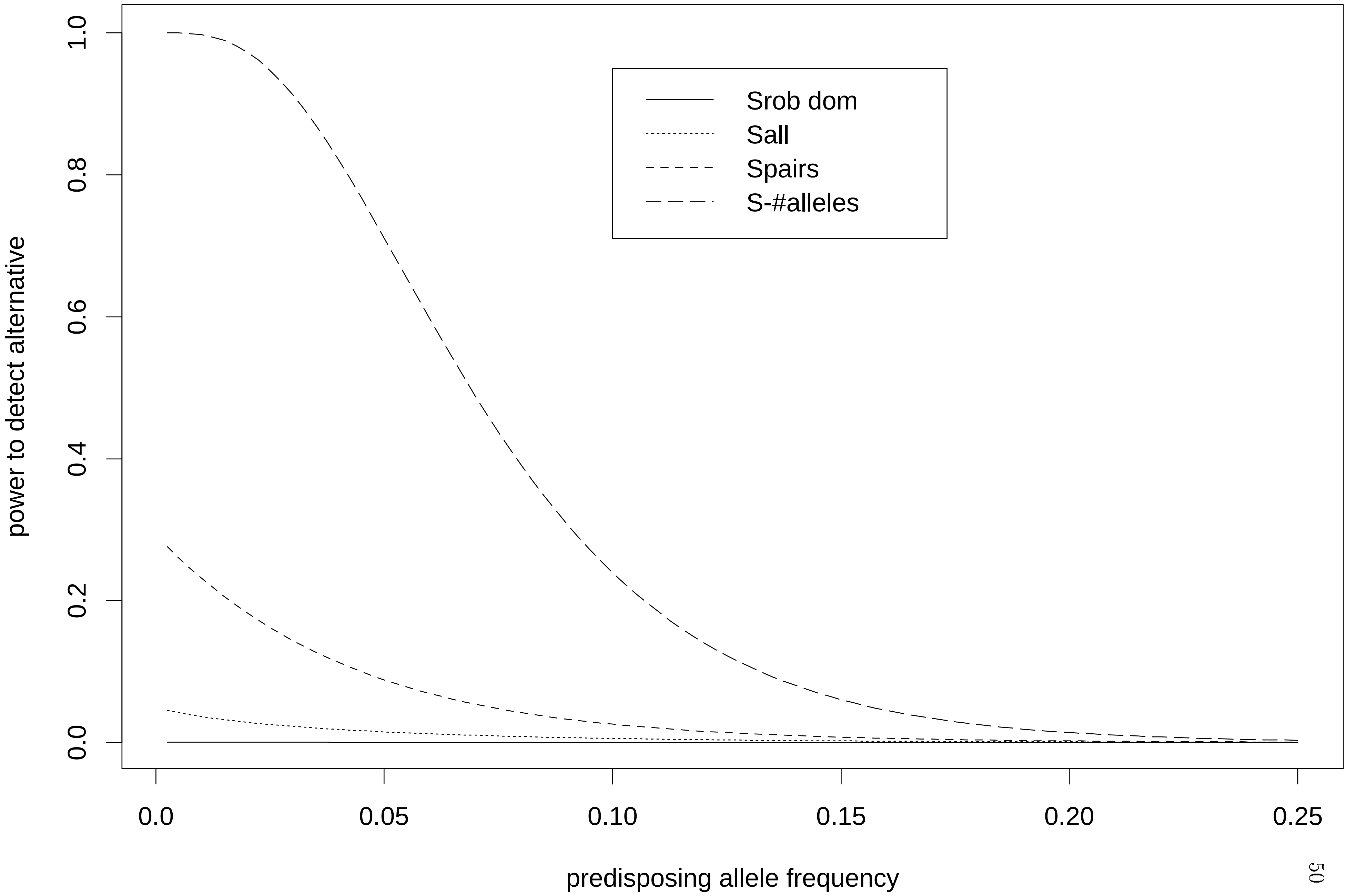

Srob dom
Sall
Spairs
S-#alleles
power to detect alternative
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.05
0.10
0.15
0.20
phenocopy/penetrance

Figure 1 e


Srob dom
Sall
Spairs
S-#alleles
power to detect alternative
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.05
0.10
0.15
0.20
0.25
predisposing allele frequency

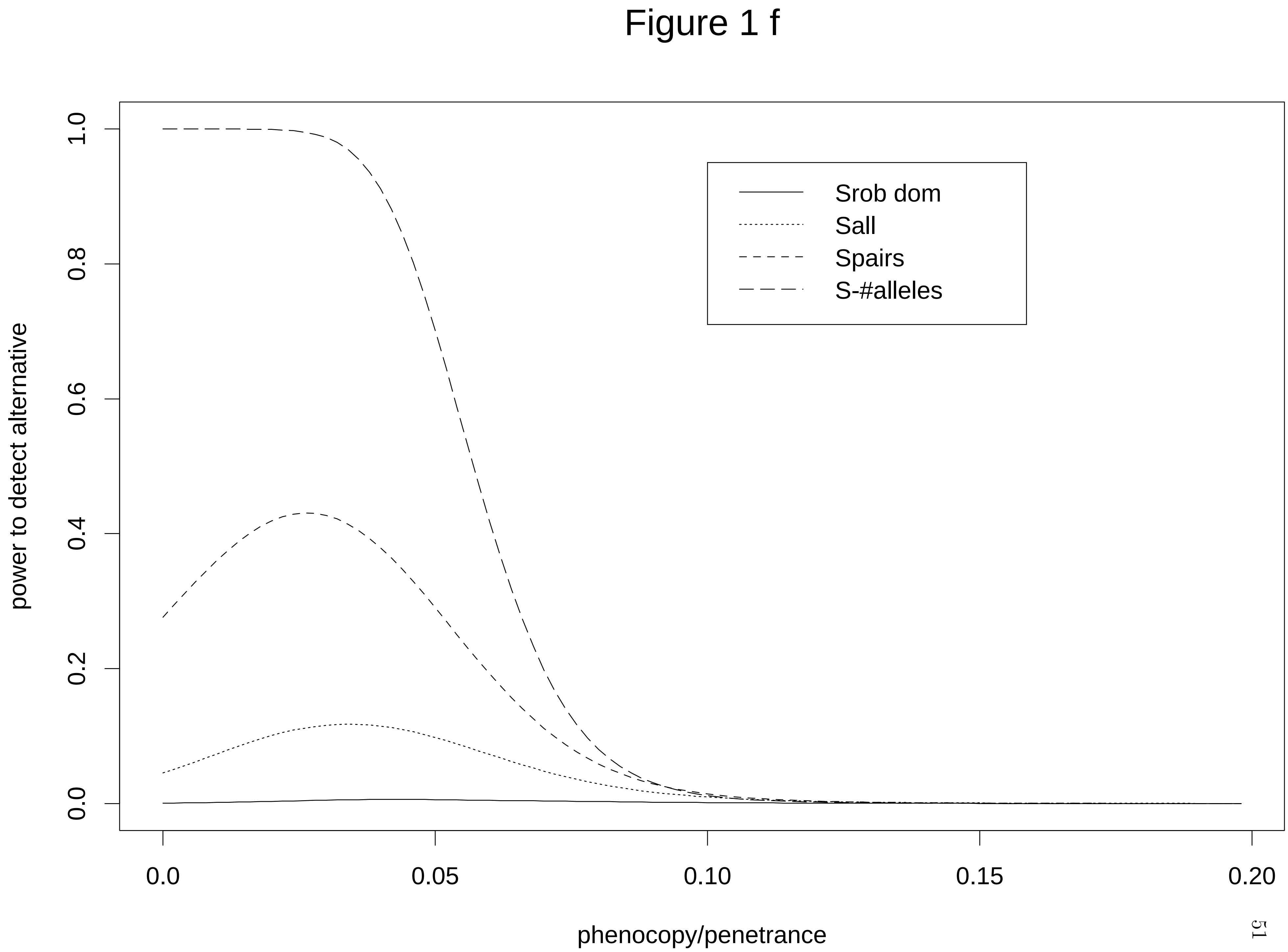


Figure 1 f
power to detect alternative
phenocopy/penetrance
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.05
0.10
0.15
0.20
Srob dom
Sall
Spairs
S-#alleles

Figure 2 a

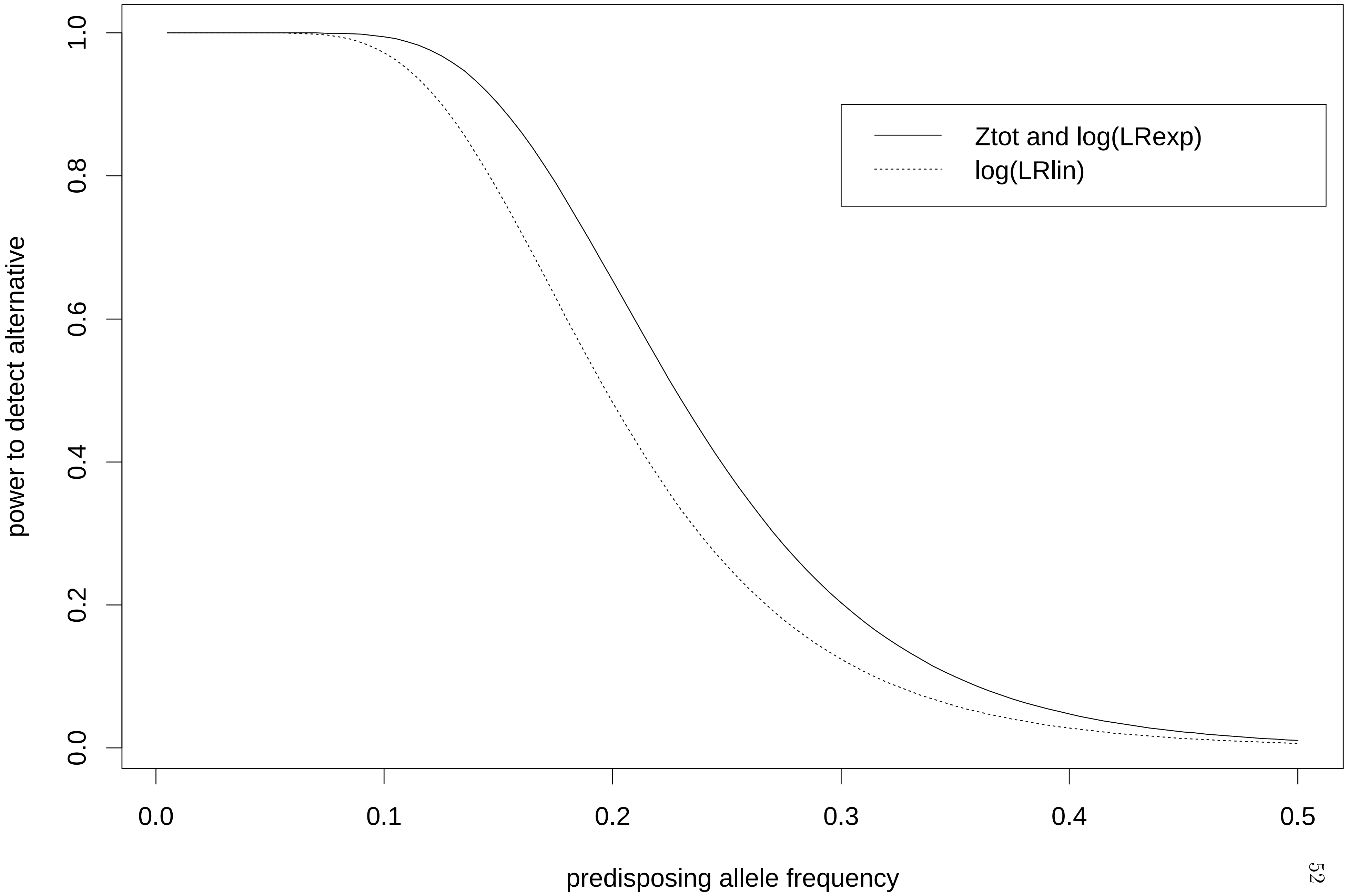

Ztot and log(LRexp)
log(LRlin)
power to detect alternative
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
predisposing allele frequency

Figure 2 b

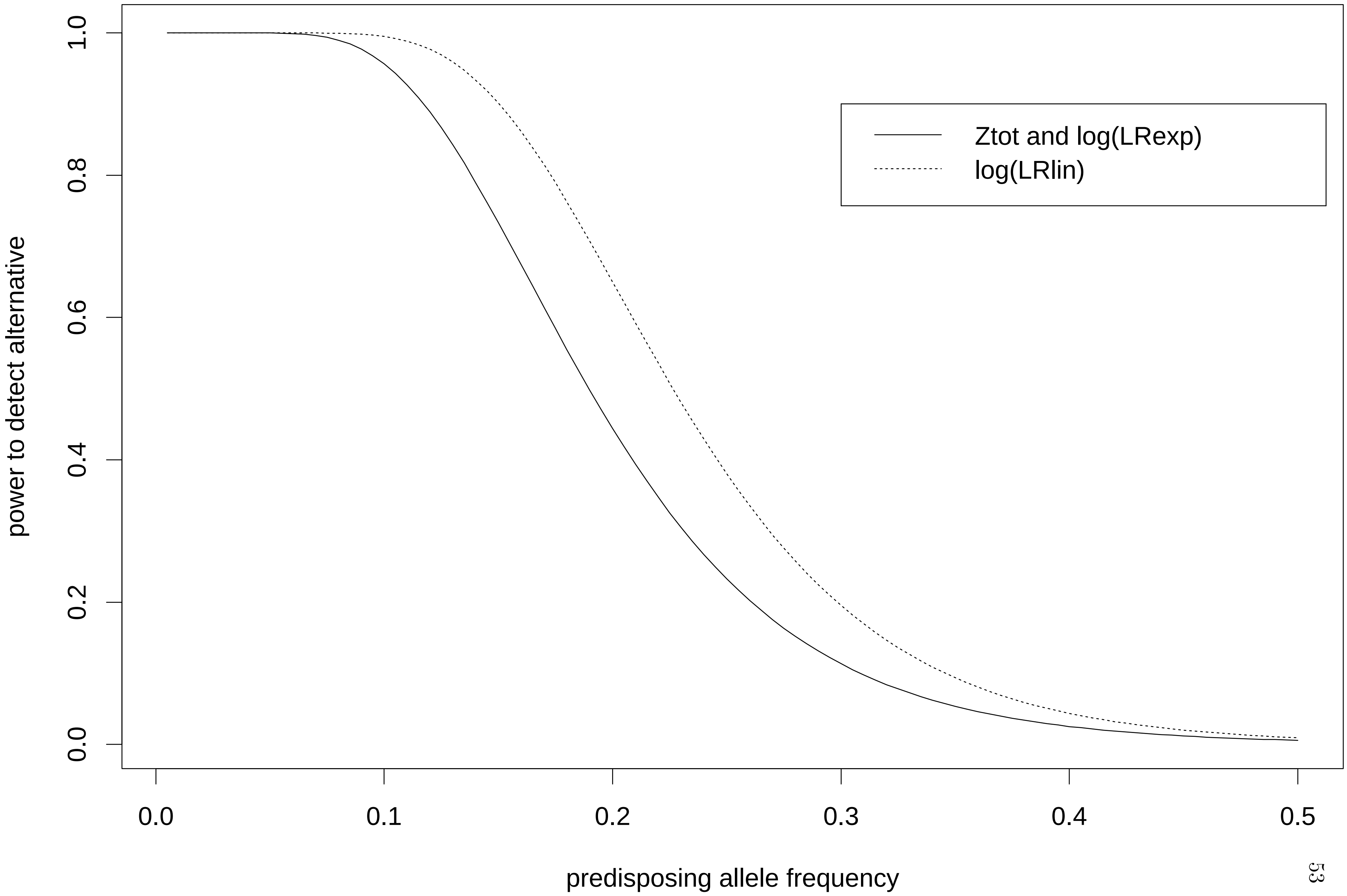

Ztot and log(LRexp)
log(LRlin)
power to detect alternative
predisposing allele frequency
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5

Figure 3: Optimal weight of sib quartet relative to sib pair
optimal S, rho=.9, Ztot