Am. J. Hum. Genet. 68:1238–1249, 2001

Transformation of Sib-Pair Values for the Haseman-Elston Method

Daolong Wang,[1] Shili Lin,[2] Rong Cheng,[1] Xin Gao,[1] and Fred A. Wright[1]

[1]Division of Human Cancer Genetics and [2]Department of Statistics, The Ohio State University, Columbus

The squared sib-pair phenotype difference (SQD) has been used as a dependent variable in the Haseman-Elston (H-E) regression quantitative-trait locus (QTL) linkage method, but it has been shown that the SQD does not make full use of linkage information. In this study, we examine the efficiency of SQD in H-E regression compared to other proposed functions of the sib-pair phenotypes. A new function of sib-pair phenotypes, the product of pair values corrected with family mean (PCF), is shown to have desirable properties in many realistic situations. Consistent results were obtained using a combination of large-sample analytic approximations, simulation, and analyses of quantitative-trait data from Genetic Analysis Workshop 10. The advantages of PCF are further improved in the presence of family-specific effects arising from environmental factors or when additional QTLs influence the trait. All of the phenotype functions are incorporated in our new, freely available linkage-mapping program MULTIGENE 1.0 for the PC environment.

Introduction

Haseman and Elston (H-E) (1972) proposed a simple regression method for detecting linkage between a quantitative trait locus (QTL) and genetic markers. This method uses the squared difference (SQD) of sib-pair phenotypes as the dependent variable and the proportion of alleles shared identical by descent (IBD) by the sibs at the marker as an independent variable. A significant linear-regression result is evidence for linkage. Many other statistical approaches to mapping QTLs have been developed, but the H-E method is attractive in its simplicity and has been used as a standard for comparison (e.g., Amos et al. 1989; Cardon and Fulker 1994; Fulker and Cardon 1994; Kruglyak and Lander 1995; Olson 1995; Risch and Zhang 1995; Blangero and Almasy 1997; Williams and Blangero 1999).

One of the drawbacks of the H-E method is that the SQD discards information that may be important for detection of linkage (Amos 1994; Fulker and Cherny 1996; Wright 1997; Drigalenko 1998). Methods have been proposed to make fuller use of the information from the data by means of the bivariate sibling phenotype distributions, including variance-component methods (e.g., Amos 1994; Fulker and Cherny 1996) and a likelihood method applicable to collections of truly independent sib pairs (Wright 1997). Additional work has explored the use, as dependent variables, of other functions of the pair values, which we term "transformation schemes" in this study. Drigalenko (1998) suggested the use of the product of pair values, emphasizing the usefulness of sib-pair trait sums (Wright 1997). Elston et al. (2000) examined, by simulation, the properties of the product of mean-corrected pair values (using the population grand mean) and showed that this transformation was more powerful than SQD in many situations. Recently, Xu et al. (2000) suggested the use of an appropriately chosen weighted average of SQD and the mean-corrected squared sum of pair values to increase the power for detecting QTLs.

This study evaluates several existing transformation schemes reported in the literature, as well as a new proposed scheme. A common theme throughout this study is that the ability to detect linkage arises from the variation in the expected value of the dependent variable as a function of IBD status. This is often referred to as "explained variance," in regression analysis, or "between-group variance," in analysis of variance, with IBD status as the independent variable. The power to detect linkage depends on the *residual ratio*—that is, the ratio of residual (i.e., unexplained) variation to total variation in the dependent variable. Here, "residual variation" includes all sources of variation in the dependent variable that remain after the regression model is fitted, including genetic variability at the locus that is not fully captured by consideration of IBD status alone. On the basis of this criterion, our proposed transformation scheme, the product of pair values corrected with *family mean,* was found to have the smallest residual ratio among the four schemes studied in many situations that we deem to be realistic. We use the Genetic Analysis Workshop 10 (GAW10) nuclear-family data (problem 2A) to illustrate the analytic results and to study further

Received December 11, 2000; accepted for publication March 15, 2001; electronically published April 17, 2001.

Address for correspondence and reprints: Dr. Fred A. Wright, 500A Medical Research Facility, 420 W. 12th Avenue, Columbus, OH 43210. E-mail: wright-4@medctr.osu.edu

0002-9297/2001/6805-0018$02.00

the relative performances of these transformation schemes. In addition, we describe via simulation how family-specific effects and the presence of additional QTLs also will tend to favor the proposed transformation over other schemes.

Four Transformation Schemes of Sib-Pair Values

Suppose we have M nuclear families in the data, and each family has $n_i (i = 1, \ldots, M)$ siblings. Let Z_{ij} be the phenotypic value of an individual $j (j = 1, \ldots, n_i)$ in nuclear family i. All sibling pairs will be considered in the family, and we use z_{1k} and z_{2k} to denote the phenotype values of the two siblings in sib pair k.

SQD

The traditional SQD of pair values is computed as

$$\mathrm{SQD}_k = (z_{1k} - z_{2k})^2 \ .$$

Product of Pair Values without Correction (PRO)

Noting drawbacks in the use of SQD, Drigalenko (1998) suggested a dependent variable consisting of the product of sib-pair phenotypes. The product appearing in Drigalenko (1998) is not mean-corrected—that is,

$$\mathrm{PRO}_k = z_{1k} z_{2k} \ ,$$

and it is shown that this product is (up to a constant factor) equivalent to the difference of SQD and the squared sum, or $\frac{1}{2}[(z_{1k} - z_{2k})^2 - (z_{1k} + z_{2k})^2] = -2z_{1k}z_{2k}$. We find it instructive to contrast this approach with other approaches described here, two of which rely on forms of mean correction. Using the product can capture some of the extra information that is lost when using the sib-pair difference. However, the sensitivity of PRO to the grand mean makes it generally less efficient than other approaches. Appendix A examines transformations of the form $(z_{1k} - c)(z_{2k} - c)$, for which PRO is a special case with $c = 0$. It is shown that among choices of constant c, the residual ratio is minimized for $c = \mu$, where μ is the overall phenotype grand mean. The resulting alternative transformation $(z_{1k} - \mu)(z_{2k} - \mu)$ leads directly to the approach adopted by Elston et al. (2000), described below.

Product of Pair Values Corrected with Grand Mean (PCG)

Elston et al. (2000) suggested the use of the product of pair values corrected by the estimated phenotype grand mean, or

$$\mathrm{PCG}_k = (z_{1k} - \hat{\mu})(z_{2k} - \hat{\mu}) \ ,$$

where $\hat{\mu} = \sum_i \sum_j Z_{ij}/N$ and $N = \sum_i n_i$. This approach explicitly estimates the covariances of z_{1k} and z_{2k} for each IBD value. This appears to be sensible, as the covariance is the parameter in the joint distribution of $\{z_{1k}, z_{2k}\}$ that varies with IBD status.

Product of Pair Values Corrected with Family Mean (PCF)

We consider a new transformation scheme that corrects each phenotypic value by the corresponding family mean instead of by the grand mean. For a given family i, the family mean, $\hat{\mu}_i$, is the average of all siblings in the family:

$$\hat{\mu}_i = \frac{1}{n_i} \sum_j Z_{ij} \ .$$

The PCF value for each sib pair in the family is, then,

$$\mathrm{PCF}_k = (z_{1k} - \hat{\mu}_i)(z_{2k} - \hat{\mu}_i)$$

Note that, when a family contains only two siblings, PCF is equivalent to SQD, since $PCF_k = -SQD_k/4$.

Theoretical Aspects of the Transformation Schemes

To explore the theoretical features of the transformation schemes, we follow the assumptions of Haseman and Elston (1972). Specifically, we assume that the quantitative trait of interest is influenced by a single, diallelic QTL with alleles B and b. The allelic frequencies are denoted as p for B and q for b ($p + q = 1$). The QTL has only an additive genetic effect, denoted as a, on the trait. The genetic contribution to phenotype for individual j in family i can be written as

$$g_{ij} = \begin{cases} m + a & \text{if the genotype is BB} \\ m & \text{if the genotype is Bb} \\ m - a & \text{if the genotype is bb.} \end{cases} \ .$$

The phenotype value of an individual can be written as $Z_{ij} = g_{ij} + \varepsilon_{ij}$, where random residuals ϵ_{ij} are independent and identically distributed $N(0, \sigma_\varepsilon^2)$. We assume Hardy-Weinberg equilibrium and linkage equilibrium. For simplicity in our derivations, we further assume that all nuclear families in the data are independent and that all families have the same sibship size, n. This last assumption provides for an appropriate analytic comparison of the transformation schemes but is not necessary for actual mapping applications.

Regression Models

All of the conditional expectations of the dependent variable on the independent variable can be written in a general regression form:

$$\mathrm{E}[y_{k(\pi)}] = \beta_0 + \beta_1 x_\pi \ ,$$

where $y_{k(\pi)}$ is the dependent variable (SQD_k, PRO_k, PCG_k, or PCF_k, as applicable) for the kth sib pair sharing π alleles, and x_π is a function of π. The regression models for the transformation schemes were derived and are presented in table 1 (see Appendix B for sketch of derivations; full derivations are available at the Statistical Genetics Laboratory Web site). The regression coefficient β_1 can be shown to be $\sigma_a^2 = 2pqa^2$, the additive genetic variance of the QTL, for each transformation scheme (Haseman and Elston 1972; Drigalenko 1998; Elston et al. 2000). The intercept β_0 and x_π, however, differ among the schemes. PCG has terms of order $(1/N)$, reflecting slight negative correlations induced during correction by the estimated grand mean (noted by Elston et al. 2000). From these expressions, it can be derived that SQD and PCF do indeed coincide for $n = 2$. Also, aside from the small-order terms, it can be shown that PCG and PRO coincide when the phenotypic mean is 0 (e.g., when $m = 0$ and $p = q$).

In the regression analyses, all sib pairs are used, and the transformed phenotype is regressed on IBD sharing status. In actual data analysis, x_π usually will not be determined unambiguously, and x_π will be replaced by its expected value or its probability distribution given the markers.

Residual Ratio a Measure of Detection Power

The regression model may also be considered as a one-way analysis of variance, with the dependent variable observed for three separate groups of sib pairs, corresponding to IBD 0, 1, or 2. Use of the analysis of variance (ANOVA) entails a slight loss in power compared with the regression approach, because the latter uses the fact that the dependent variable for the IBD 1 group should be intermediate to that of the other two groups. The model may be rewritten (including residual term) as

$$y_{k(\pi)} = \nu + \tau_\pi + e_{k(\pi)} \ ,$$

where $\nu = \mathrm{E}(y_k) = \sum_\pi f_\pi \mathrm{E}[y_{k(\pi)}]$ is the overall mean of y_k, and f_π is the frequency of sib pairs sharing π alleles, $\tau_\pi = \mathrm{E}[y_{k(\pi)}] - \nu$ is the effect of IBD value π on the dependent variable, and $e_{k(\pi)} = y_{k(\pi)} - \mathrm{E}[y_{k(\pi)}]$ is the residual effect for the kth sib pair sharing π alleles IBD. A useful variance decomposition (McCulloch and Searle 2001, p. 11) can then be applied to the total variance of y_k,

$$\mathrm{var}(y_k) = \mathrm{var}\{\mathrm{E}[y_{k(\pi)}]\} + \mathrm{E}\{\mathrm{var}[y_{k(\pi)}]\} \ ,$$

where the variance and expectation are taken over the values of π. These correspond essentially to the *between-group variance* and *within-group variance* (Ott et al. 1993, p. 774), where "group" refers to IBD status. The between-group variance, denoted by σ_b^2, can be expressed as

$$\begin{aligned}\sigma_b^2 &= \mathrm{var}\{\mathrm{E}[y_{k(\pi)}]\} \\ &= \sum_\pi f_\pi \tau_\pi^2 \\ &= \sum_\pi f_\pi \{\mathrm{E}[y_{k(\pi)}] - \nu\}^2 \ .\end{aligned}$$

The within-group variance (or *residual variance*), denoted by σ_w^2, is

$$\begin{aligned}\sigma_w^2 &= \mathrm{E}\{\mathrm{var}[y_{k(\pi)}]\} \\ &= \sum_\pi f_\pi \mathrm{E}[e_{k(\pi)}^2] \\ &= \sum_\pi f_\pi \mathrm{E}\left(\{y_{k(\pi)} - \mathrm{E}[y_{k(\pi)}]\}^2\right) \ .\end{aligned}$$

For each transformation scheme except PRO, variances σ_b^2 and σ_w^2 can be expressed as

$$\sigma_u^2 = c_{u1}\sigma_a^4 + c_{u2}\sigma_a^2\sigma_\varepsilon^2 + c_{u3}\sigma_\varepsilon^4 \ ,$$

where u stands for b or w, as appropriate. For PRO, the variances include additional terms:

Table 1

Regression Models for Various Transformation Schemes

Transformation Scheme	β_0	β_1	x_0	x_1	x_2
SQD	$2\sigma_\varepsilon^2$	σ_a^2	2	1	0
PCF	$-\frac{1}{n}\sigma_\varepsilon^2$	σ_a^2	$-\frac{1}{2} + \frac{1}{2n} - \frac{1}{n^2}$	$-\frac{1}{2n}$	$\frac{1}{2} - \frac{3}{2n} + \frac{1}{n^2}$
PCG	$-\frac{1}{N}\sigma_\varepsilon^2$	σ_a^2	$-\frac{1}{N}$	$\frac{1}{2} - \frac{2}{N} + \frac{1}{N^2}$	$1 - \frac{3}{N} + \frac{2}{N^2}$
PRO	$m^2 + 2(p-q)ma + a^2$	σ_a^2	-2	$-\frac{3}{2}$	-1

$$\sigma_u^2 = c_{u1}\sigma_a^4 + c_{u2}\sigma_a^2\sigma_\varepsilon^2 + c_{u3}\sigma_\varepsilon^4 + c_{u4}m^2\sigma_a^2 + c_{u5}m^2\sigma_\varepsilon^2 + c_{u6}m\sigma_a^3 + c_{u7}m\sigma_a\sigma_\varepsilon^2 .$$

The coefficients $c_{u.}$ are given in Appendix C. The expressions above are given for generality. However, for the between-group variances, the expression reduces to simply $\sigma_b^2 = c_{b1}\sigma_a^4$. In other words, for each transformation scheme, only c_{b1} is nonzero among the coefficients for σ_b^2, because σ_b^2 is due to genetic variation only. However, none of the coefficients for σ_w^2 is zero, including several terms involving σ_a^2. This reflects that, in addition to the true random residual phenotype variation, σ_w^2 includes contributions from genetic variance that are not recovered or explicitly estimated when IBD status is used as an independent variable.

The *residual ratio* is the ratio of residual variance (σ_w^2) to the total variance ($\sigma_t^2 = \sigma_b^2 + \sigma_w^2$) and is a measure of the relative efficiency of a specific transformation scheme in detection of the QTL. The residual ratio bears a one-to-one correspondence with the noncentrality parameter in the ANOVA *F*-test (Neter et al. 1985, p. 547). A smaller residual ratio indicates greater information for QTL mapping, leading to greater power in detecting the QTL. For SQD, PCF, and PCG, the residual ratio, denoted by ρ, can be written as

$$\rho = \sigma_w^2/\sigma_t^2 = \frac{c_{w1}\sigma_a^4 + c_{w2}\sigma_a^2\sigma_\varepsilon^2 + c_{w3}\sigma_\varepsilon^4}{c_{t1}\sigma_a^4 + c_{t2}\sigma_a^2\sigma_\varepsilon^2 + c_{t3}\sigma_\varepsilon^4} = \frac{c_{w1} + c_{w2}\phi^2 + c_{w3}\phi^4}{c_{t1} + c_{t2}\phi^2 + c_{t3}\phi^4} ,$$

where $c_{t.} = c_{b.} + c_{w.}$, and $\phi = \sigma_\varepsilon/\sigma_a$. Alternatively, the residual ratio can also be written as

$$\rho = 1 - \sigma_b^2/\sigma_t^2 = 1 - \frac{c_{b1}}{c_{t1} + c_{t2}\phi^2 + c_{t3}\phi^4} . \quad (1)$$

For PRO, the residual ratio can be written as

$$\rho = 1 - \frac{c_{b1}}{c_{t1} + c_{t2}\phi^2 + c_{t3}\phi^4 + c_{t4}\psi^2 + c_{t5}\psi^2\phi^2 + c_{t6}\psi + c_{t7}\psi\phi^2} , \quad (2)$$

where $\psi = m/\sigma_a$.

Trends of Residual Ratios of the Transformation Schemes

To investigate the trends of residual ratios for the four transformation schemes, we calculated ρ for the four transformation schemes using equations (1) and (2) and varying several parameters, including the heritability

$$h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_\varepsilon^2) = 1/(1 + \sigma_\varepsilon^2/\sigma_a^2) = 1/(1 + \phi^2) ,$$

frequency of allele B, sibship size n, and ψ. All of the results are limiting efficiencies for large total number of sibs N. Of the transformation schemes, only PCF depends on sibship size n, essentially because the marginal (across families) within-group variances do not depend on n, and our H-E approaches do not attempt to explicitly use the within-family phenotype correlations.

The four plots in figure 1 show the residual ratios for the four transformation schemes under four QTL heritabilities, $h^2 = 1.0, .8, .5$, and .2. For $h^2 = 1.0$, all variation in the phenotypic values is attributable to the QTL. However, >70% of the total genetic variation is still distributed as residual variance for any of the transformation schemes, because IBD status can account only incompletely for phenotype variation. We emphasize that h^2 values approaching 1.0 are very extreme and unlikely to be observed in practice. Also, in such extreme cases, a likelihood approach would be considerably more powerful, recognizing the near-perfect correlation of sibs with IBD 2. For large heritabilities the discreteness of the two-allele genetic model will also produce somewhat different results than those based on normal approximations. Thus, in figures 1*A* and 1*B*, SQD has a smaller residual ratio than PCG, despite analytic work based on normal assumptions that indicates PCG should be somewhat more powerful than SQD (Wright 1997; Drigalenko 1998).

Overall, as the heritability decreases, the residual ratio increases. In fact, when $h^2 = .2$, the residual ratio is near 1.0 across the range of allele frequencies, reflecting the inherently low power of the H-E method under these situations. The allele frequency p of allele B is another important factor affecting residual ratios. The lowest residual ratio is always achieved when the two alleles of the QTL are equally frequent—that is, when $p = .5$ for all the four schemes and four heritabilities.

The transformation schemes themselves also play a role in the magnitudes of residual ratios (fig. 1). No matter how large the QTL contribution, the orderings of the residual ratios between PRO and PCG is PRO $\geqslant$ PCG; that is, PCG recovers a larger proportion of genetic information than PRO. An additional observation is that PCF always offers an improvement over SQD, except when $n = 2$ and the two schemes are identical. Also, the relative rankings between the two sets, {PRO, PCG} and {SQD, PCF}, depend greatly on the QTL contribution h^2. With a very high QTL contribution (fig. 1*A* and *B*), PRO and PCG tend to produce much higher residual ratios than SQD and PCF for the whole range of allele frequencies, although, for such high heritabilities, this behavior is highly model-dependent. For mod-

Figure 1 Residual ratios calculated from the formulas. Plots A, B, C, and D were obtained under $h^2 = 1.0$, .8, .5, and .2, respectively. For PRO, ψ was 0 for all of the plots. For PCG, the grand mean is assumed to be known (i.e., limit as the total number of siblings N approaches infinity). The number of siblings per family is denoted by n.

erate or low QTL contribution, smaller residual ratios for PRO and PCG are observed around the central region of the range of allele frequencies (fig. 1*C* and 1*D*). As an interesting note, we observe that for $h^2 = .2$, the residual ratio for PCG is approximately twice as far from 1.0 as that for SQD. This corresponds to the limiting result (Wright 1997) that the squared difference uses only half of the available linkage information as h^2 approaches 0.

As one might expect, the size of sibships (n) has a great influence on the residual ratio of PCF (fig. 1). For large values of n, PCF has a lower residual ratio than does PCG, even when the QTL contribution is moderate (fig. 1*C*).

Applications to GAW10 Data and Additional Simulations

The above results were obtained under the assumption that all families are of the same size. To examine the analytic results under more realistic situations with varying family sizes, we applied the four transformation schemes to analyze the data on trait "Q4" of the Genetic Analysis Workshop 10 (GAW10) simulated nuclear-family data (problem 2A) (MacCluer et al. 1997). All 200 replicates of the simulated data were used, each of which contains 239 families with an average of 2.87 ± 1.04 siblings per family. Trait "Q4" is controlled by three unlinked diallelic QTLs located on three of the 10 chromosomes. The three QTLs—MG4, MG5, and MG6—have phenotypic contributions of 28%, 16%, and 11%, respectively. There are a total of 367 markers with an average spacing of 2.03 cM across the genome. There are no epistatic effects among the QTLs, no differences in living conditions between families, and no sex or age effects.

The Hidden Markov model algorithm (Kruglyak and Lander 1995) was used to compute the distribution of alleles shared IBD, and the expected values of x_π were used in the regressions. All possible sib pairs of each family were included and analyzed as independent pairs, which produces valid linkage tests (Elston et al. 2000) under the null hypothesis of no linkage. All the analyses were conducted with our newly developed computer program MULTIGENE 1.0 for the PC environment (Windows 95/98/NT). It is freely available from our Web site.

Example Data Set and Average LODs

Figure 2 presents a typical example in which the schemes are applied to data from replicate 19. Regression results were converted to LOD-equivalents as

$$\text{LOD} = \frac{1}{2} s \log_{10}(\text{SS}_\text{R}/\text{SS}_\text{F}) \ ,$$

Figure 2 Replicate 19 of GAW 10 problem 2A shows the results from the four transformation schemes. The triangles show the true locations of the QTLs.

where SS_R is the residual sum of squares from the reduced regression model, SS_F, the residual sum of squares from the full regression model, and s is the number of sib pairs (Johnson and Wichern 1982). When a LOD of 1.8 was considered suggestive linkage (based on the proposals of Lander and Kruglyak [1995]), SQD, PCF, and PCG exhibited peaks near the locations of MG4 and MG5. Peaks were also found on chromosome 10 but at farther distances from the true location of MG6. For PRO, however, no peak >1.8 was found within 10 cM of the three QTLs. For MG4 and MG5, the contributions to the trait are relatively large, and the LOD scores at the identified peaks obtained with PCF were nearly twice as high as those obtained with SQD or PCG. MG6 has the smallest contribution to the trait, and the results from SQD, PCF and PCG were all similar.

The results from replicate 19 reflect a general trend for the four transformation schemes, which can be seen in the average LOD values at the three QTL locations across all 200 replicates. PCF produced the highest LOD scores (3.23, 1.60, and 1.09 for MG4, MG5, and MG6, respectively), which were nearly twice as large as those obtained with SQD (LODs 1.83, 0.72, and 0.49, respectively) or PCG (LODs 1.57, 0.69, and 0.47, respectively). SQD and PCG yielded similar results. PRO always gave the lowest average LOD values (0.26, 0.23, and 0.27, respectively). As expected, the phenotypic contributions of the QTLs played an important role in determining the LOD differences among the transformation schemes; larger QTL contributions corresponded to larger differences.

Detection Power

Direct comparisons of LOD scores are reasonable in evaluation of transformation schemes at a single putative QTL location, because the tests based on the schemes involve the same number of df. However, the correlation of successive LOD values across the genome may differ among the methods (see Lander and Kruglyak 1995), in ways that are imperfectly understood. Thus, to refine the comparison of the transformation schemes, we compared the four schemes across all 200 replicates of the GAW 10 data set. We estimated a separate LOD threshold for each of the transformation schemes by using the original data and counting as false positives those with

peaks occurring on a chromosome without a QTL or those >10 cM from a true QTL location. This is an approximate approach that has appeared to work well in our experience (D.W. and F.A.W., unpublished data).

To empirically control the type I error, we set the suggestive linkage level as that producing an average of one genomewide false positive per genome scan, and the significant linkage level as having a rate of .05 genomewide false positives per scan. Empirical LOD thresholds were determined from the replicates, with suggestive LODs 1.81, 1.68, 1.64, and 1.65 for PCF, SQD, PCG, and PRO, respectively. For these same schemes the corresponding significant LOD thresholds were 3.41, 3.49, 3.11, and 2.81. Considering the number of replicates, these empirical thresholds are (with the exception of the significant threshold for PRO) reasonably close to those based on analytic approximations.

The detection powers under these empirical LOD thresholds are given in table 2. Under both significance levels, PCF yielded the highest power for MG4 and MG5, but the power for MG6, the locus that contributes the least to the trait, was comparable with the other transformation schemes. PRO was the least powerful, whereas SQD and PCG had similar power. These results are quite consistent with our analytic results.

Size of Sibships

Using simulation, we can explore the effect of extreme sibship sizes on the relative performances of the transformation schemes in the more realistic situation that sample size N is finite. Figure 3 shows the estimates of residual ratios (based on application of a one-way ANOVA) for 1,000 simulations for a QTL with heritability $h^2 = .5$, assuming that the QTL IBD status can be observed directly (i.e., a fully informative marker at the QTL). The sibship sizes increase from 2 to 100, while the total number of siblings N remains fixed at 1,000. Note that, as in the analytic results, the residual ratios for SQD, PCG, and PRO do not change much over the range of sibship sizes and even less over the range of realistic sibship sizes (say, $\leqslant 10$). The residual ratio for PCF, in contrast, drops fairly dramatically at first and then drops more gradually. There is a limit to the improvement offered by PCF, as the family-specific mean will, for very large sibship sizes n, approach a fixed value based on the genetic constitution of the parents to the sibship.

Table 2

Detection Power with the Four Transformation Schemes

	Detection Power							
	Suggestive Linkage				Significant Linkage			
QTL	PCF	SQD	PCG	PRO	PCF	SQD	PCG	PRO
MG4	.67	.65	.54	.04	.25	.18	.19	.02
MG5	.22	.18	.16	.02	.04	.03	.01	.00
MG6	.07	.10	.13	.06	.00	.00	.00	.01

Figure 3 Results from 1,000 simulations, illustrating the influence of sibship size on the residual ratio. The sibship sizes varied, but the total sample size N was fixed at 1,000. IBD values were assumed to be known for each sib pair. The QTL phenotypic contribution was $h^2 = .5$ with two equally likely alleles, and $m = 0$.

Family-Specific Effects

We have explored how the performances of transformation schemes depend on several factors, including QTL contributions, allele frequencies, and size of sibship (for PCF). What may not be apparent is that our analytic results represent a worst-case scenario for the relative performance of PCF, in that up to now we have assumed no family-specific phenotype effects (e.g., environmental influences) and no additional QTLs that would contribute background genetic variation in the phenotype. The presence of such effects is realistic for complex quantitative traits and will tend to induce phenotype correlation among siblings that is not captured by consideration of genotypes at the QTL under study. The correction of phenotypes by family-specific means can correct for much of this correlation, to the point that, in extreme cases of familial correlation, the SQD can perform better than PCG (Palmer et al. 2000).

To study the influences of familial effects, we conducted two different sets of simulations. For both simulations, the QTL under study has $p = .5$, $\sigma_a^2 = 50$, and $\sigma_e^2 = 50$, and we examine a fully informative marker at the QTL. For the first simulation, we added a normal random effect $N(0, \sigma_F^2)$ to the phenotype of each individual in the family as a family-specific environmental effect, with σ_F^2 ranging from 0 to 100. For the second simulation, we included no such environmental effects but modified the phenotypes to reflect the effect of an

additional QTL. This second QTL consisted of two equally likely alleles, was unlinked to the QTL under study, and had additive genetic variance ranging from 0 to 100. In each simulation, 1,000 replicates were conducted. In each replicate, 500 nuclear families with four siblings each were sampled. The residual ratios, $\hat{\rho} = \hat{\sigma}_e^2/\hat{\sigma}_t^2$, were estimated from the H-E regression model and the average estimated ratios over the 1,000 simulations are plotted in figure 4. It is apparent that family-specific environmental effects have a great influence on the residual ratios of PCG and PRO. However, they have no influence on SQD and PCF. For these transformation schemes, family-specific environmental effects cancel in the differences of pair values and in the corrections with family means but not in PCG or PRO. In contrast, genetic background influences the residual ratios of all the four transformation schemes, as only a portion of this variation is recoverable via phenotype transformation. However (under a similar principle as with family-specific environmental effects), in the presence of an additional QTL, the residual ratio of PCG rises faster than those of SQD or PCF.

Summary and Future Extensions

The H-E sib-pair QTL regression method and others reliant on IBD status are fundamentally limited in their power to detect linkage, as much of the underlying genetic variation is not reflected in IBD status. Nonetheless, these approaches have considerable appeal, in that they require few assumptions about the underlying genetic model. We have explored how the transformation of sibling phenotypes can greatly affect the power to detect linkage. Furthermore, the product of sib-pair values corrected by family-specific means (PCF) appears to offer advantages over a range of realistic conditions. These advantages are achieved without the need to explicitly consider the underlying genetic model and involve nothing more complicated than a simple transformation of phenotype.

The analytic results are supported by analyses using data from GAW10 and additional simulations. We have also recently used these four transformation schemes to analyze simulated data from GAW12, and PCF once again was found to yield the highest power (Wang et al. 2000). We propose that PCF be considered as an alternative transformation scheme of sib-pair values to improve detection power with the H-E method.

Modest further improvements in power are to be expected from explicit modeling of correlations of phenotype for collections of sib pairs within families (Elston et al. 2000). In addition, our results suggest that improvements over PCF might be made by use of a hybrid approach that corrects by a mean which is likely to be "most representative" of the mean for that family. For small sibships, when family-specific effects are not strong, the PCG may be preferable to PCF (fig. 1*D*). Essentially, this results from the fact that a small sibship may give a poor estimate of the "true" phenotypic mean for that family, which derives from the parental genotypes. In such instances, the estimated grand mean may be preferable. However, as the sibship grows, the estimated family mean gradually forms a better estimate of the true family mean. These ideas are similar to shrinkage estimation in mixed-model analysis (McCulloch and Searle 2001, p. 51), and we are investigating hybrid approaches that correct by weighted averages of the grand mean and family mean.

Figure 4 Results of 1,000 simulations, illustrating the influence of family-specific effects on residual ratios. The sibship size was $n = 4$ in all cases, with total sample size $N = 2{,}000$. IBD values were obtained unambiguously at the QTL.

Acknowledgments

We thank Drs. Mark E. Irwin and William J. Lemon for helpful comments on the manuscript. The use of the GAW10 simulated data was permitted by the Southwest Foundation for Biomedical Research. We are grateful to Drs. Jeff Williams and Jean MacCluer and Ms. Vanessa D. Olmo for handling our request for the GAW10 data. This research was supported in part by National Institutes of Health grants GM58934 and P30CA16058 and by an award from the Ohio State James Cancer Hospital and Solove Research Institute (to F.A.W.) and National Science Foundation grant DMS-9971770 (to S.L.).

Appendix A

Product of Sib-Pair Values Corrected by a Constant

We consider the transformation

$$\mathrm{PCC} = (z_1 - c)(z_2 - c) ,$$

for a single sib pair (the subscript k is suppressed). PRO is the special case of PCC where $c = 0$. We note that the overall mean and variance of the phenotype of a randomly selected individual does not depend on IBD status (denoted by $p = 0$, 1, or 2), or

$$\mathrm{E}(z_1|\pi) = \mathrm{E}(z_2|\pi) = \mathrm{E}(Z) = \mu$$

and

$$\mathrm{var}(z_1|\pi) = \mathrm{var}(z_2|\pi) = \mathrm{var}(Z) = \sigma^2 .$$

Then

$$\mathrm{E}[(z_1 - c)(z_2 - c)|\pi] = \mathrm{E}(z_1 z_2|\pi) - c\mathrm{E}(z_1|\pi) - c\mathrm{E}(z_2|\pi) + c^2 = \mathrm{E}(z_1 z_2|\pi) - 2c\mu + c^2 .$$

Thus, compared with PRO, PCC produces a constant shift in the expectations of the dependent variable for each value of π. The overall *between-group* variance thus will remain constant, regardless of c. However, the *total* variance of PCC does depend on c. We have

$$\begin{aligned}
\mathrm{var}[(z_1 - c)(z_2 - c)] &= \mathrm{var}\{[(\mathrm{z}_1 - \mu) + (\mu - \mathrm{c})][(z_2 - \mu) + (\mu - c)]\} \\
&= \mathrm{var}[(z_1 - \mu)(z_2 - \mu) + (\mu - c)(z_1 - \mu) + (\mu - c)(z_2 - \mu) + (\mu - c)^2] \\
&= \mathrm{var}[(z_1 - \mu)(z_2 - \mu)] + 2(\mu - c)^2\sigma^2 + 2(\mu - c)\mathrm{cov}[(z_1 - \mu)(z_2 - \mu),(z_1 - \mu)] \\
&+2(\mu - c)\mathrm{cov}[(z_1 - \mu)(z_2 - \mu),(z_2 - \mu)] + 2(\mu - c)^2\mathrm{cov}[(z_1 - \mu),(z_2 - \mu)] ,
\end{aligned}$$

so the total variance can be expressed as a quadratic in $\mu - c$. The power will maximized for the choice of c that minimizes the total variance. The coefficient to the linear term in the equation is proportional to $d = \mathrm{cov}[(z_1 - \mu)(z_2 - \mu),(z_1 - \mu)]$, noting that, by symmetry of z_1 and z_2, d also equals $\mathrm{cov}[(z_1 - \mu)(z_2 - \mu),(z_2 - \mu)]$. It is convenient to reexpress the data as $x_1 = z_1 - \mu, x_2 = z_2 - \mu$. For most models of interest, the joint density f of x_1 and x_2 follows a special symmetry, such that $f(x_1,x_2) = f(-x_1, -x_2)$. Essentially, this results from the symmetry of z_1 and z_2 and from the fact that marginal densities of z_1 and z_2 are symmetric about their mean—that is, the median equals the mean. These properties hold approximately for the model examined in this paper and hold for the normal models that are often applied in QTL mapping. In real-data analysis, the phenotypes may be skewed, and it is standard practice to perform normalizing transformations to the data to reduce skew. From these assumptions, we now show that $d = 0$, which, from the quadratic equation, implies that the total variance is minimized for $\mu - c = 0$ or $c = \mu$. We have

$$\begin{aligned}\mathrm{cov}[(z_1-\mu)(z_2-\mu),(z_1-\mu)] &= \mathrm{E}(x_1x_2\cdot x_2)-\mathrm{E}(x_1x_2)\mathrm{E}(x_1)=\mathrm{E}(x_1^2x_2)\\ &=\int_{X_1}\int_{X_2}x_1^2x_2f(x_1,x_2)dx_1dx_2=\int_{X_1}\int_{X_2<0}x_1^2x_2f(x_1,x_2)dx_1dx_2+\int_{X_1}\int_{X_2>0}x_1^2x_2f(x_1,x_2)dx_1dx_2\\ &=\int_{X_1}\int_{X_2>0}(-x_1)^2(-x_2)f(-x_1,-x_2)dx_1dx_2+\int_{X_1}\int_{X_2>0}x_1^2x_2f(x_1,x_2)dx_1dx_2\\ &=\int_{X_1}\int_{X_2>0}x_1^2(-x_2)f(x_1,x_2)dx_1dx_2+\int_{X1}\int_{X2>0}x_1^2x_2f(x_1,x_2)dx_1dx_2=0\ .\end{aligned}$$

Appendix B

Conditional Expectations

The expectations conditioning on IBD sharing value π for all the four transformation schemes can be expressed generally as

$$\mathrm{E}[y_{k(\pi)}]=\sum_F f_F\sum_C f_{C(F,\pi)}\mathrm{E}[y_{k(F,C)}]\ ,$$

where F stands for family type ($F = 1,\ldots, 6$ corresponding to the six possible joint parental genotypes); f_F is the expected frequency of the family type F in the population; C stands for sib pair type ($C = 1,\ldots, 6$ for the joint sib genotypes); $f_{C(F,\pi)}$ is the frequency of C type of sib pair in the Fth type of family, conditional on IBD value (π); and $\mathrm{E}[y_{k(F,C)}]$ is the expectation, given family type F and pair type C, which is independent of π. Frequencies f_F and $f_{C(F,\pi)}$ were found under the assumptions of random mating and Hardy-Weinberg equilibrium, and are given in table B1.

Table B1

Frequencies of Family Types (F) and Sib-Pair Types (C)

		$f_{C(F,\pi)}$					
π AND FAMILY TYPE	f_F	BB-Bb	Bb-Bb	bb-Bb	BB-BB	BB-bb	bb-bb
0:							
BB × Bb	$4p^3q$	1	…	…	…	…	…
Bb × Bb	$4p^2q^2$	…	.5	…	…	.5	…
bb × Bb	$4pq^3$	…	…	1	…	…	…
BB × BB	p^4	…	…	…	1	…	…
BB × bb	$2p^2q^2$	…	1	…	…	…	…
bb × bb	q^4	…	…	…	…	…	1
1:							
BB × Bb	$4p^3q$	.5	.25	…	.25	…	…
Bb × Bb	$4p^2q^2$	.5	…	.5	…	…	…
bb × Bb	$4pq^3$	…	.25	.5	…	…	.25
BB × BB	p^4	…	…	…	1	…	…
BB × bb	$2p^2q^2$	…	1	…	…	…	…
bb × bb	q^4	…	…	…	…	…	1
2:							
BB × Bb	$4p^3q$	…	.5	…	.5	…	…
Bb × Bb	$4p^2q^2$	…	.5	…	.25	…	.25
bb × Bb	$4pq^3$	…	.5	…	…	…	.5
BB × BB	p^4	…	…	…	1	…	…
BB × bb	$2p^2q^2$	…	1	…	…	…	…
bb × bb	q^4	…	…	…	…	…	1

NOTE.—Expressions for f_c after Haseman and Elston (1972).

The frequencies f_F and $f_{C(F,\pi)}$ are the same for all the schemes, but $E[y_{k(F,C)}]$ varies with transformation schemes. For all the schemes except PCF, $E[y_{k(F,C)}] = E[y_{k(C)}]$, implying that the expectations depend only on the genetic constitution of the sib pair. However, we need to obtain $E[y_{k(F,C)}]$ for each type of family for PCF because family means differ with family types. Full derivations are available at our Web site.

Appendix C

Table C1

Coefficients c_{ui} for Variance Components

Transformation Scheme	c_{ui}
PCF	$c_{b1} = \frac{1}{8}(1 - \frac{4}{n} + \frac{8}{n^2} - \frac{8}{n^3} + \frac{4}{n^4})$, $c_{b2} = c_{b3} = 0$
	$c_{w1} = \frac{1}{4pq}\left(\frac{1}{4} - \frac{1}{2n}(1 - 2pq) - \frac{1}{4n^2}(1 + 14pq) + \frac{1}{2n^3}(3 + 8pq) - \frac{2}{n^4}pq\right)$, $c_{w2} = c_{w3} = 1 - \frac{2}{n} + \frac{2}{n^2}$
SQD	$c_{b1} = \frac{1}{2}$, $c_{b2} = c_{b3} = 0$, $c_{w1} = \frac{1}{2pq}$, $c_{w2} = c_{w3} = 8$
PCG	$c_{b1} = \frac{1}{8}(1 - \frac{4}{N} + \frac{8}{N^2} - \frac{8}{N^3} + \frac{4}{N^4})$, $c_{b2} = c_{b3} = 0$
	$c_{w1} = \frac{1}{4pq}\left(\frac{1}{2}(2 - pq) + \frac{1}{N}(-8 + 22pq) + \frac{1}{N^2}(26 - 106pq) + \frac{1}{N^3}(-34 + 164pq) + \frac{1}{N^4}(14 - 72pq)\right)$
	$c_{w2} = 2 - \frac{7}{N} + \frac{10}{N^2} - \frac{4}{N^3}$, $c_{w3} = 1 - \frac{2}{N} + \frac{2}{N^2}$
PRO	$c_{b1} = \frac{1}{8}$, $c_{b2} = c_{b3} = c_{b4} = c_{b5} = c_{b6} = c_{b7} = 0$
	$c_{w1} = \frac{3}{4pq} - \frac{17}{8}$, $c_{w2} = \frac{1}{pq} - 2$, $c_{w3} = 1$, $c_{w4} = 3$, $c_{w5} = 2$, $c_{w6} = c_{w7} = \frac{4(p-q)}{\sqrt{2pq}}$

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Statistical Genetics Laboratory Web site, http://pegasus.med.ohio-state.edu/

References

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535–543

Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE (1989) A more powerful robust sib-pair test of linkage for quantitative traits. Genet Epidemiol 6:435–449

Blangero J, Almasy L (1997) Multipoint oligogenic linkage analysis of quantitative traits. Genet Epidemiol 14:959–964

Cardon LR, Fulker DW (1994) The power of interval mapping of quantitative trait loci, using selected sib pairs. Am J Hum Genet 55:825–833

Drigalenko E (1998) How sib pairs reveal linkage. Am J Hum Genet 63:1242–1245

Elston RC, Buxbaum S, Kevin BJ, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19:1–17

Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. Am J Hum Genet 54: 1092–1103

Fulker DW, Cherny SS (1996) An improved multipoint sib-pair analysis of quantitative traits. Behav Genet 26:527–532

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Johson RA, Wichern DW (1982) Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, NJ

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

MacCluer J, Blangero J, Dyer T, Speer M (1997) GAW10: simulated family data for a common oligogenic disease with quantitative risk factors. Genet Epidemiol 14:737–742

McCulloch CE, Searle SR (2001) Generalized, linear, and mixed models. John Wiley & Sons, New York

Neter J, Wasserman W, Kutner MH (1985) Applied linear statistical models, 2d ed. Richard D. Irwin, Homewood, IL

Olson JM (1995) Multipoint linkage analysis using sib pair: an interval mapping approach for dichotomous outcomes. Am J Hum Genet 56:788–798

Ott RL (1993) An introduction to statistical methods and data analysis. Duxbury Press, Belmont, CA

Palmer LJ, Jacobs KB, Elston RC (2000) Haseman and Elston revisited: the effects of ascertainment and residual familial correlations on power to detect linkage. Genet Epidemiol 19:456–460

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

Wang D, Gao X, Lin S, Skrivanek Z, Irwin ME, Wright FA (2000) Comparison of several methods for linkage analysis. In: Genetic Analysis Workshop 12, volume 2: simulated data

participant contributions. Genetic Analysis Workshop 12, pp 509–513

Williams JT, Blangero J (1999) Comparison of variance components and sib pair-based approaches to quantitative trait linkage analysis in unselected samples. Genet Epidemiol 16: 113–134

Wright FA (1997) The phenotypic difference discards sib-pair QTL linkage information. Am J Hum Genet 60: 740–742

Xu X, Weiss S, Xu X-P, Wei LJ (2000) A unified Haseman-Elston method for testing linkage with quantitative traits. Am J Hum Genet 67:1025–1028