Am. J. Hum. Genet. 75:906–909, 2004

Report

A Powerful and Robust New Linkage Statistic for Discordant Sibling Pairs

Jin P. Szatkiewicz[1] and Eleanor Feingold[1,2]

Departments of [1]Biostatistics and [2]Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh

Previously, Szatkiewicz and colleagues evaluated the performance of a wide variety of statistics for quantitative-trait–locus linkage, using discordant sibling pairs. They found that the most powerful statistics, in general, were a score statistic and a "composite statistic." However, whereas these two statistics have equal power under ideal conditions, each has limitations that reduce its power in certain circumstances. The score statistic depends on estimates of trait parameters and can lose a lot of power if those estimates are incorrect. The composite statistic is not sensitive to trait-parameter estimates but does depend on arbitrary weights that must be chosen on the basis of the ascertainment scheme. In this report, we elucidate the algebraic relationship between the score and composite statistics and then use that relationship to suggest a new statistic that combines the best properties of both. We call our new statistic the "robust discordant pair" (RDP) statistic. We report simulation studies to show that the RDP statistic does, indeed, have all of the strengths and none of the weaknesses of the score and composite statistics.

Szatkiewicz et al. (2003) used simulation studies to evaluate the type I error and power of various statistics for QTL linkage, using discordant sibling pairs. They considered a number of statistics from the literature, as well as several new variants. The bottom line was that three statistics were best, by virtue of having correct type I error rates across the board and having power higher than that of other statistics for all or most trait models tested. All three of those statistics are what we call "combination" type—that is, they combine linkage information from two sources: (1) the marginal identity-by-descent (IBD)–sharing distribution and (2) the correlation between IBD sharing and trait values. This is in contrast to IBD-sharing statistics, which only use information from the first source, and to "correlation-based" statistics such as Haseman-Elston regression (Haseman and Elston 1972), which only use information from the second source.

One of the three best statistics found by Szatkiewicz et al. (2003) is a score statistic. Tang and Siegmund (2001) derived the basic statistic, and Szatkiewicz et al. (2003) proposed a variant ("SCORE3" in their terminology) that has an entirely empirical variance in the denominator. The formula for the score statistic with the empirical variance estimate is

$$\frac{\sum A_i\left(\pi_i - \frac{1}{2}\right)}{\sqrt{\frac{1}{n}\left(\sum A_i^2\right)\left[\sum\left(\pi_i - \frac{1}{2}\right)^2\right]}} ,$$

where n is the number of (independent) sibling pairs, π_i is the estimated IBD-sharing proportion for pair i, and

$$A_i = \frac{Y_{i\mathrm{S}}}{(1+r)^2} - \frac{Y_{i\mathrm{D}}}{(1-r)^2} + \frac{4r}{1-r^2} .$$

The parameter r is the sibling correlation for the trait. $Y_{i\mathrm{S}}$ and $Y_{i\mathrm{D}}$ are the squared trait sum and the squared trait difference, respectively, and are calculated on the basis of trait values that are standardized to have a mean of zero and a variance of one. That is, if X_{i1} and X_{i2} are the trait values for pair i, then

$$Y_{i\mathrm{S}} = \left[\frac{(X_{i1}-\mu)}{\sigma} + \frac{(X_{i2}-\mu)}{\sigma}\right]^2$$

Received May 25, 2004; accepted for publication August 25, 2004; electronically published September 13, 2004.

Address for correspondence and reprints: Dr. Eleanor Feingold, 130 DeSoto Street A310, Pittsburgh, PA 15261. E-mail: feingold@pitt.edu

0002-9297/2004/7505-0018$15.00

and

$$Y_{iD} = \left[\frac{(X_{i1} - \mu)}{\sigma} - \frac{(X_{i2} - \mu)}{\sigma}\right]^2 ,$$

where μ is the population trait mean and σ is the population trait SD. Thus, the formula for A_i effectively involves three population trait parameters: the mean, the variance, and the sibling correlation. Szatkiewicz et al. (2003) showed that, with good estimates of these three parameters, this statistic has approximately the maximum possible power for reasonably Gaussian trait models. However, if the parameters (particularly the mean) are not well estimated, the score statistic can lose a substantial amount of power.

The second good statistic for discordant pairs is almost identical to the score statistic. It is one version of the HE-COM statistic proposed by Sham and Purcell (2001). Sham and Purcell proposed two different statistics under the name "HE-COM." Szatkiewicz et al. (2003) referred to these two statistics as "S&P1" and "S&P2." Here, we use the more informative names "HE-COM-correlation" and "HE-COM-combination." HE-COM-combination has the same numerator as the score statistic—$\sum A_i(\pi_i - 1/2)$—but uses a slightly different variance estimate in the denominator. HE-COM-correlation has the numerator $\sum B_i(\pi_i - \bar{\pi})$, where B_i is equal to the first two terms of A_i. The important difference between HE-COM-combination and HE-COM-correlation is the use of 1/2 versus $\bar{\pi}$ to mean-center the IBD sharing. The use of $\bar{\pi}$ makes the statistic independent of any difference between $\bar{\pi}$ and 1/2 and thus creates a correlation-based statistic. The use of 1/2 creates a combination statistic, by drawing additional power from any deviation between $\bar{\pi}$ and 1/2. For a population sample, $\bar{\pi}$ should be ~1/2, so the two statistics are essentially the same, but, for discordant pairs, the difference is substantial (Szatkiewicz et al. 2003).

The third statistic that performed well in the studies by Szatkiewicz et al. (2003) is a variant of the "composite statistic" originally proposed by Forrest and Feingold (2000). The composite statistic is a weighted average of the Haseman-Elston regression statistic (Haseman and Elston 1972) and an IBD-sharing statistic. The Haseman-Elston statistic is based on the squared trait difference, $(X_{i1} - X_{i2})^2$, which does not involve any population trait parameters. Because the composite statistic does not depend on trait parameters, it avoids the parameter misspecification problems of the score statistic. However, for the composite statistic to have power as high as that of the score statistic, the weights for the two components must be intelligently chosen on the basis of some knowledge of the ascertainment scheme. Forrest and Feingold (2000) showed that it is not too difficult to find good weights if the ascertainment scheme is known, but it would be better if the weighting issue could be eliminated completely, so that it would not be necessary to know the ascertainment scheme.

We show here that the score statistic can be algebraically decomposed into a correlation-based statistic plus an IBD-sharing statistic, with weights that are data dependent. This helps to explain the relationship between the score statistic and the composite statistic and also suggests a statistic that combines the best properties of both. The decomposition is as follows.

$$\begin{aligned}
\text{Score statistic} &= \frac{\sum A_i\left(\pi_i - \frac{1}{2}\right)}{\sqrt{\frac{1}{n}(\sum A_i^2)\left[\sum\left(\pi_i - \frac{1}{2}\right)^2\right]}} \\
&= \frac{\sum A_i(\pi_i - \bar{\pi}) + \sum A_i\left(\bar{\pi} - \frac{1}{2}\right)}{\sqrt{\frac{1}{n}(\sum A_i^2)\left[\sum\left(\pi_i - \frac{1}{2}\right)^2\right]}} \\
&= \frac{\sum A_i(\pi_i - \bar{\pi})}{\sqrt{\frac{1}{n}(\sum A_i^2)\left[\sum\left(\pi_i - \frac{1}{2}\right)^2\right]}} \\
&\quad + \left(\frac{\sum A_i}{\sqrt{n\sum A_i^2}}\right)\frac{\bar{\pi} - \frac{1}{2}}{\sqrt{\left[\frac{1}{n^2}\sum\left(\pi_i - \frac{1}{2}\right)^2\right]}} .
\end{aligned}$$

The first term of the decomposed score statistic is essentially HE-COM-correlation, with an empirical variance estimate rather than a regression-based variance estimate in the denominator. The second term is an IBD-sharing statistic, again with an empirical variance estimate in the denominator. The two components are weighted by a factor that is large when the A_is have a large absolute value (as in discordant pair samples) and that is zero when the A_is have a mean of zero (as in population samples).

Viewed through the lens of this decomposition, the score statistic and the composite statistic have only two differences between them. The first difference is that the composite statistic uses arbitrary fixed weights for the two components, whereas the score statistic uses data-dependent weights that automatically adapt to the sampling scheme. For example, if both statistics were applied to a population sample, the composite statistic would perform quite poorly, because it would incorporate random noise from the IBD-sharing statistic, whereas the score statistic would automatically adjust the weight on

the IBD-sharing statistic to zero and, thus, would still have maximal power. If both statistics were applied to a sample of discordant pairs, the score statistic would again automatically adjust the weight to be optimal, no matter what the ascertainment rule was, whereas the composite statistic would only perform well if we had chosen a good weight beforehand.

The second difference between the score statistic and the composite statistic is that the score statistic is based on the trait function A_i, which involves the trait difference, the trait sum, and the three population trait parameters. The composite statistic, by contrast, is based on the original Haseman-Elston regression procedure and uses the trait function $(X_{i1} - X_{i2})^2$, which does not involve the trait sum or the trait parameters. For most types of samples, the function A_i contains more information than the squared trait difference alone (Sham and Purcell 2001), but our previous simulation results showed that, for discordant pairs, the information is almost equal (Forrest and Feingold 2000; Szatkiewicz et al. 2003). This is also clear in considering the expression A_i at a purely arithmetic level; for discordant pairs, the standardized trait sum is very small, and A_i is really determined by the trait difference. For example, for model 1, we averaged the ratio of the squared sum term (i.e., the first term) to the whole of A_i for different kinds of samples. For an extreme discordant sibling pair (EDSP) sample, the average ratio was .006. For a moderately discordant sibling pair (MDSP), the average ratio was 0.04. By contrast, for a sample of concordant pairs, the average ratio was 0.96, indicating that the trait sum is the dominant term for concordant pairs.

If we combine the data-dependent weights of the score statistic with the parameter-independent trait function of the composite statistic, then we should get a new statistic with the best features of both. We propose the following statistic, which we term the "robust discordant pair" (RDP) statistic:

$$\frac{-\sum (X_{i1} - X_{i2})^2 \left(\pi_i - \frac{1}{2}\right)}{\sqrt{\frac{1}{n}\left[\sum (X_{i1} - X_{i2})^4\right]\left[\sum \left(\pi_i - \frac{1}{2}\right)^2\right]}},$$

which is the same as the score statistic but with $-(X_{i1} - X_{i2})^2$ substituted for A_i. We have included the negative sign so that positive values of the statistic correspond to the alternative hypothesis. Note that the RDP statistic has the same numerator as the Haseman-Elston statistic, except with 1/2 instead of $\bar{\pi}$. The relationship between the RDP statistic and the Haseman-Elston statistic is analogous to the relationship between HE-COM-combination and HE-COM-correlation.

We updated the simulation studies done by Szatkiewicz et al. (2003), to include the RDP statistic. The methods were described in detail in that article (Szatkiewicz et al. 2003). In the present study, table 1 reports power (based on 1,000 replicates) for selected statistics applied to EDSPs—defined as pairs with one sibling in the top 10% of the trait distribution and the other sibling in the bottom 10%. Table 2 shows the analogous results for MDSPs—defined as pairs with one sibling in the top 35% of the trait distribution and the other sibling in the bottom 35%. Trait models 1–5 are simple mixture-of-normals models, with "increaser" allele frequencies of 0.1 (see Szatkiewicz et al. [2003] for model details). Models 1 and 4 are additive, models 2 and 5 are dominant, and model 3 is a recessive model that has substantial skewness and kurtosis. Models 1′ and 2′ are non-Gaussian models that were created by applying a signed square transformation to models 1 and 2. All statistics shown have correct type I error rate of 0.01 by our simulations (results not shown).

Several important features of the statistics are evident from table 1. First, the score statistic with correct parameter estimates, the composite statistic with extreme weights, and the RDP statistic all have essentially equal power for the somewhat Gaussian models (models 1, 2, 4, and 5) (see table 1). The composite statistic with equal

Table 1

Power for EDSPs at $\alpha = 0.01$

	Power (%) under Model						
Statistic	1	2	3	4	5	1′	2′
Score:							
Correct parameters	87	93	18	94	81	78	79
Mean estimate low by 1 SD	86	93	18	93	80	75	74
Mean estimate high by 1 SD	88	94	21	93	80	76	77
Composite[a]:							
Equal weights	79	78	22	81	62	66	71
Extreme weights	86	94	12	94	82	84	87
RDP	87	93	19	93	81	78	79

[a] Forrest and Feingold (2000) recommended equal weights for MDSPs and the weights 0.259 and 0.966 for EDSPs.

Table 2

Power for MDSPs at $\alpha = 0.01$

	Power (%) under Model						
Statistic	1	2	3	4	5	1′	2′
Score:							
Correct parameters	72	77	12	84	79	42	64
Mean estimate low by 1 SD	59	65	12	79	77	17	28
Mean estimate high by 1 SD	74	77	15	78	67	37	42
Composite[a]:							
Equal weights	73	77	11	83	78	49	74
Extreme weights	58	67	4	79	72	53	66
RDP	71	76	12	84	78	42	67

[a] Forrest and Feingold (2000) recommended equal weights for MDSPs and the weights 0.259 and 0.966 for EDSPs.

weights has much less power, underscoring the importance of the weights. For EDSPs, the score statistic does *not* lose much power if the mean is misspecified, because most of the linkage information comes from the IBD-sharing. In table 2, the results for MDSPs show essentially the same features of the statistics as those described for EDSPs (provided that, for MDSPs, equal weights are used for the composite statistic), except that the score statistic *does* lose a substantial amount of power when the mean is misspecified. We have not shown results for the misspecification of the other parameters. The effects of misspecifying other parameters follow similar patterns to the effect of a misspecified mean, but are smaller overall (see Szatkiewicz et al. [2003] for partial results).

The performance of the statistics for the non-Gaussian models requires separate comments. The score statistic is based on the likelihood of the data, under the assumption that the trait model is normal (not even a mixture of normals), so it should not necessarily perform well for non-Gaussian models. This is also somewhat true for the RDP statistic, whose form follows that of the score statistic. For models 1′ and 2′, the composite statistic, which does not depend on the normality assumption, does have higher power than the score statistic and the RDP statistic. This is not true, however, for model 3, which is also substantially skewed. We feel that the behavior of the statistics for non-Gaussian trait models still requires further study. It is not clear what types of non-Gaussian trait models are the most realistic and important, and it is also not clear how various features of the models and statistics interact to determine which statistic is the most powerful.

Overall, we recommend our new RDP statistic as the best choice for studies of discordant sibling pairs, in almost any situation. It has power equal to that of the score and composite statistics but is robust to parameter misspecification and does not depend on arbitrary weights. For EDSPs, it is probably fine to use the score statistic or even the IBD-sharing statistic instead, but using the RDP statistic adds an extra measure of robustness at no cost in power. For more moderately selected samples, the RDP statistic is clearly preferable.

A few caveats are in order. First, further study is required before we can make recommendations about statistics for substantially non-Gaussian trait models; however, on the basis of our results to date, the RDP statistic and the composite statistic (with appropriate weights) seem to be the best choices. Second, we should note that neither the RDP statistic nor the composite statistic is appropriate for combinations of discordant and concordant pairs, though the score statistic is appropriate. Statistics for mixtures of discordant and concordant pairs will be addressed in future work (Szatkiewicz and Feingold, in press).

Acknowledgment

This work was supported by the National Institutes of Health grant R01 HG02374-01.

References

Forrest WF, Feingold E (2000) Composite statistics for QTL mapping with moderately discordant sibling pairs. Am J Hum Genet 66:1642–1660

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

Sham PC, Purcell S (2001) Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet 68:1527–1532

Szatkiewicz JP, Feingold E. QTL mapping with discordant and concordant sibling pairs—new statistics and new design strategies. Genet Epidemiol, in press

Szatkiewicz JP, T.Cuenco K, Feingold E (2003) Recent advances in human quantitative-trait–locus mapping: comparison of methods for discordant sibling pairs. Am J Hum Genet 73:874–885

Tang H-K, Siegmund D (2001) Mapping quantitative trait loci in oligogenic models. Biostatistics 2:147–162