**Analysis Notes**

Fifty-nine kindreds initially met the entry criteria for this study. Before conducting any statistical analysis, families were checked for pedigree and genotyping errors by using the RELTEST and MARKERINFO programs in the S.A.G.E. package. As a result, six families were eliminated from the analysis for misspecified relationships: in four cases the affected pairs were most likely half siblings, and in two families the affected pairs were monozygotic (MZ) twins. Ultimately, there were 53 families available for analysis (Figs. 3 and 4).

There are three different types of sibling pairs in the analysis: concordantly affected pairs, in which both siblings are affected with colon neoplasia; discordant pairs, in which one sibling is affected and the second sibling is unaffected; and concordantly unaffected pairs, in which both siblings are unaffected. For the purposes of this analysis individuals were classified as affected if they were diagnosed with invasive cancer, an adenoma with high-grade dysplasia (HGD), or an adenoma ≥1 cm. To be classified as unaffected, a sibling must have had a negative screening endoscopy of the colon; all other individuals, including those diagnosed after age 65, were considered to have unknown phenotype for the purpose of the analysis.

Allele frequency estimates were obtained by using the program FREQ in S.A.G.E.. FREQ obtains the maximum likelihood estimates of the allele frequencies among the founders of the families using all genotyped family members. There is one African American family in the sample, but both parents were genotyped and so the estimated number of alleles shared identical by descent (IBD) in this family was not affected by the allele frequency estimates. Multipoint estimates of IBD sharing were obtained at 2-cM intervals, incorporating information from all markers by using a Kosambi map function, as implemented in the S.A.G.E. program GENIBD. These estimates, which were robust to the allele frequencies used, were then used for the mean tests and Haseman-Elston regression as implemented in the SIBPAL program in S.A.G.E. The mean tests compare, separately for each type of sibling pair, the estimated proportion of alleles shared IBD

($\hat{\pi}$) with that expected under the null hypothesis of no linkage, which is ½, using the appropriate one-sided one-sample *t* tests. To identify regions of interest, we selected those regions in which the allele sharing for the concordantly affected siblings (the usual "mean test") had a *P* value ≤.016, which is asymptotically equivalent to a logarithm of odds score ≥1. This resulted in the identification of six regions of interest on chromosomes 1, 6, 9, 10, and 16 (two of the peaks are on chromosome 6). Modeling of a two-stage linkage design, by using the DESPAIR program in S.A.G.E., suggested that, in an analysis of a whole genome scan from concordantly affected sibling pairs, a cutoff at the first stage of $P \leq 0.016$ would provide 80% power for detecting a true linkage to a disease gene associated with a sibling relative recurrence risk $\lambda = 2.5$. The model also showed that this relatively nonstringent cutoff could be associated at the first stage with detecting six to seven false positives.

Further analysis of these linkage peaks was accordingly conducted using the original Haseman-Elston regression test, which uses the information on both concordant and discordant pairs. Thus, of the different options available in SIBPAL for the dependant variable, we used the squared sibling trait difference (multiplied by -½, so that at the trait locus the regression coefficient is the trait genetic variance), without loss of generality giving affected siblings the value 1 and unaffected siblings the value 0, which pools the concordantly affected and unaffected pairs and uses a *t* test to contrast their mean allele sharing with that of the discordant pairs. This resulted in a symmetrical distribution of 93 concordant pairs and 94 discordant pairs for the analysis. The Haseman-Elston method regresses the dependent variable, *y*, on the proportion of alleles a sibling shares identical by descent, $\hat{\pi}$, using a regression equation of the form $y = \alpha + \beta\hat{\pi}$, and the significance of the degree to which allele sharing is associated with concordance or discordance in affection status determines the significance of the regression coefficient β. As implemented in SIBPAL, the coefficient β at any point along the genome estimates the total locus-specific genetic variance of the trait, attenuated by the recombination fraction between that point and the trait locus. When the six linkage peaks identified as of interest by the affected pair mean test were examined with the Haseman-Elston regression method, significant linkage was obtained on chromosome 9, peaking at D9S1786 with a

*P* = 0.00055 (Table 2). Moreover, this P value is not dependent on our estimates of the marker allele frequencies (Table 3).

In considering the derivation of both the mean tests and the Haseman-Elston regression method, the authors point out the common observation that the pairs of sibling pairs within a given sibship are clearly not independent. Importantly, such independence is not required for the validity of the methods. First, as proved by Hodge (1), the allele sharings within a sibship are pairwise independent under the null hypothesis of no linkage. It follows that all of the covariances between the estimated allele sharing between pairs are zero, with the result that the variance of the mean allele sharing is correctly estimated even when the pairs are treated as although they are independent. This conclusion, that the variance of the mean allele sharing is correctly estimated even when the pairs are treated as although they are independent, was further validated by the demonstration by Blackwelder and Elston (2) (see appendix to that article) that when there is no linkage, the distribution of the mean allele sharing statistic is asymptotically normal with mean 0 and variance 1, even when all of the sibling pairs do not come from different sibships.

An appropriate permutation test is the gold standard, but is of no use in obtaining the empirical distribution of the test statistic when all pairs have the same phenotypes, all concordant or all discordant, as for example in the mean tests. It *can* be usefully used in Haseman-Elston regression, and results in valid empirical *P* values. In a permutation analysis, the significance of a statistic is judged without making any distributional assumptions by comparing it to its permutation distribution, i.e., the distribution of the statistic when calculated for every possible permutation of the data that would be expected to be exchangeable under the null hypothesis. Except for very small samples, the number of possible permutations is so large that in practice we only consider a random sample of all possible permutations, the size of this sample being determined by the desired accuracy with which the significance level is to be determined. In the S.A.G.E. program SIBPAL, the permutation sample replicates are obtained by shuffling the measure of allele sharing between sibling pairs (the estimated proportion of alleles shared identical by descent by each pair of siblings). This is equivalent to shuffling the phenotypic status

of each sibling pair (concordant or discordant), but has the advantage that, when covariates are included in the regression model, these covariate values remain attached to the corresponding sibling-pair phenotypic status. In this way, the shuffling is with respect to allele sharing alone, disrupting solely the linkage signal. The number of permutation replicates used is determined by the user specifying the confidence (here taken to be 95%) with which the estimated *P* value (i.e., the proportion of the replicate sample statistics exceeding in value that for the observed data) is to be within a given percentage of the true *P* value (here specified to be 5%) (3). The shuffling is performed both within sibships (which has no effect in the case of sibships of size 2) and across sibships; however, in the latter case the shuffling must be across sibships of the same size, to allow for the fact that sibships of different sizes have different correlational structures. In the instance of this study, the significance of the linkage peak identified by the Haseman-Elston regression method on chromosome 9 at D9S1796 was assessed by a permutation sample in which 216,583 permutations were created, sufficient to assure with 95% confidence that the estimated *P* value was within 5% of the true *P* value. This analysis yielded a *P* value for linkage of 0.00045, slightly more significant than the *P* value of 0.00055 derived from the asymptotic analysis above.

To further examine the linkage peak at D9S1786, we introduced as an independent variable in the regression equation the probability that the sibling pair shares two alleles IBD (estimated by GENIBD) to decompose the total genetic variance at D9S1786 into additive genetic and nonadditive genetic components (4). Table 4 gives the findings for the contributions of the additive and nonadditive components of variance.

The coefficient of sharing two alleles IBD, which measures the nonadditive component of variance attenuated by recombination, was not significant. Under a recessive mode of inheritance, or very common dominant allele(s), the nonadditive component of variance would be expected to be significant. Finding that only the additive component was significant is most consistent with any susceptibility alleles located in this region being most likely inherited in a dominant fashion.

On the assumption of rare dominant alleles, the proportion (λ) of the disease in the population that can be attributed to a locus on the basis of the excess allele sharing $\hat{\pi}$ between affected sibling pairs at that locus can be calculated from the formula [$\hat{\pi}$ = 0.50 + (0.25)(λ)] (5). This formula is derived by treating the population of sibling pairs as being a mixture of two groups. One group, of proportion (λ), is linked to a dominant disease allele, and thus shows average allele sharing of 0.75, and hence contributes (0.75) (λ) to the allele sharing of the population. The second group, of proportion (1 - λ) is unlinked, and thus shows average allele sharing of 0.50, and hence contributes (0.50) (1 - λ) to the allele sharing of the population. The sum of these two populations hence shows total allele sharing of (0.75) (λ) + (0.50) (1 - λ) = 0.50 + (0.25) (l).

As indicated in the article, the excess sharing of 0.59 [95% confidence limits: 0.5903 ± (1.96)(0.0341) = 0.5235 to 0.6571] for the affected pairs could indicate dominant alleles at this location accounting for 36% of the disease [95% confidence limits: 9–63%]. The value of 0.75 for allele sharing among affected sibling pairs arising due to a dominantly inherited disease allele is derived from such siblings having a 100% chance of sharing the jointly inherited dominant disease allele, and having a 50% chance of randomly inheriting a second identical parental marker allele, resulting in an average sharing of 75% of marker alleles in common.

Subsequent to the completion of this initial study, and indeed during the review process for this manuscript, a secondary study was performed to determine whether individuals with small adenomas that were not advanced were also likely to have linkage to D9S1786. To address this question we reanalyzed the 53 kindreds described in this study, classifying as affected with disease not only individuals with colon cancer and advanced adenomas, but also individuals with small adenomas of size <1 cm. The recalculation of linkage within our 53 kindreds after classifying as affected those with adenomas of any size increased the *P* value, and hence reduced the significance level, for linkage by a factor of 10 (nominal Haseman-Elston regression *P* value = 0.006 and *P* value from permutation testing = 0.004). Thus, the strength of linkage to D9S1786 demonstrated among individuals in these kindreds affected with colon cancer or advanced colon

adenomas is much diminished when individuals with small adenomas are admixed in, suggesting that in these kindreds the development of small adenomas is for the most part not linked to the putative disease allele(s) at D9S1786.

As demonstrated by this study, the sibling pair method of linkage analysis provides a robust approach for identifying genetic linkage for diseases such as colorectal neoplasia in which the lethality of the disease, and its onset in later life, limit the ability to obtain DNA from multigenerational kindreds. In this study, six candidate regions for linkage were initially identified by the mean tests among concordantly affected sibling pairs. One of these regions, defined by marker D951786, was confirmed as particularly significant by the mean test of discordant sibling pairs, and by Haseman-Elston regression analysis contrasting allele sharing in concordant versus discordant sibling pairs. However, these analytical methods are not able to exclude the possibility of linkage of colorectal neoplasia to other loci. Fuller statistical consideration of this point is provided in work by Dizier *et al.* (6).

As discussed in the article, the addition to this analysis of discordant sibling pairs provides several advantages. It increases the size of the sample available, and hence increases the power for detecting linkage, and also helps to distinguish true linkages from false positive linkages that may be obtained when only concordant sibling pairs are studied. Of course, we recognize the increased risk of misclassification in assigning individuals as being "unaffected." as this presumes they are less likely to develop future colon neoplasia. In the case of this study, unaffected individuals were those that had undergone a negative colon endoscopy. The risk of misclassification of these individuals as "unaffected" is mitigated by these individuals being as a group older than their affected siblings, and by a negative colon endoscopy putting individuals at low risk for development of advanced colon neoplasia for from 5 to 10 years after the endoscopy. (7–9). Moreover, the classification of these individuals as "unaffected" is a conservative one, as any misclassification will only reduce the power for obtaining linkage.

In summary, our analysis of an entire genome scan on 53 families that met the criteria for severe histopathology yielded evidence for linkage to chromosome 9 ($P = 0.00045$) with a peak at 95 cM. In addition, there is evidence to support an additive component of variance significant at 95 cM on chromosome 9, which, in the absence of a significant nonadditive component, is consistent with the mode of inheritance for one or more rare dominant alleles at this location. The excess sharing of 0.59 for the affected pairs could indicate dominant alleles at this location accounting for 36% of the disease.

1. Hodge, S. (1984) *Genet. Epidemiol.* **1,** 109–122.

2. Blackwelder, W. & Elston, R. C. (1985) *Genet Epidemiol.* **2,** 85–97.

3. Shete, S., Jacobs, K. B. & Elston, R. C. (2003) *Hum Hered* **55,** 79–85.

4. Elston, R. C., Buxbaum, S., Jacobs, K. B. & Olson, J. M. (2000) *Genet. Epidemiol.* **19,** 1–17.

5. Wiesner, G., Platzer, P., Buxbaum, S., Lewis, S., MacMillen, M., Olechnowicz, J., Willis, J., Chakravarti, A., Elston, R. & Markowitz, S. (2001) *J. Natl. Cancer Inst.* **93,** 635–639.

6. Dizier, M. H., Quesneville, H., Prum, B., Selinger-Leneman, H. & Clerget-Darpoux, F. (2000) *Ann. Hum. Genet.* **64,** 433–442.

7. Selby, J. V., Friedman, G. D., Quesenberry, C. P., Jr. & Weiss, N. S. (1992) *N. Engl. J. Med.* **326,** 653–657.

8. Rex, D. K., Cummings, O. W., Helper, D. J., Nowak, T. V., McGill, J. M., Chiao, G. Z., Kwo, P. Y., Gottlieb, K. T., Ikenberry, S. O., Gress, F. G., *et al.* (1996) *Gastroenterology* **111,** 1178–1181.

9. Rex, D. K., Johnson, D. A., Lieberman, D. A., Burt, R. W. & Sonnenberg, A. (2000) *Am. J. Gastroenterol.* **95,** 868–877.