Am. J. Hum. Genet. 47:957–967, 1990

Multipoint Analysis of Human Quantitative Genetic Variation

David E. Goldgar

Department of Medical Informatics, University of Utah, Salt Lake City

Summary

A unique method of partitioning human quantitative genetic variation into effects due to specific chromosomal regions is presented. This method is based on estimating the proportion of genetic material, *R*, shared identical by descent (IBD) by sibling pairs in a specified chromosomal region, on the basis of their marker genotypes at a set of marker loci spanning the region. The mean and variance of the distribution of *R* conditional on IBD status and recombination pattern between two marker loci are derived as a function of the distance between the two loci. The distribution of the estimates of *R* is exemplified using data on 22 loci on chromosome 7. A method of using the estimated *R* values and observed values of a quantitative trait in a set of sibships to estimate the proportion of total genetic variance explained by loci in the region of interest is presented. Monte Carlo simulation techniques are used to show that this method is more powerful than existing methods of quantitative linkage analysis based on sib pairs. It is also shown through simulation studies that the proposed method is sensitive to genetic variation arising from both a single locus of large effect as well as from several loosely linked loci of moderate phenotypic effect.

Introduction

Through its GENOME initiative, the U.S. Government has allocated significant scientific and budgetary resources to both genetic and physical mapping, with the ambitious goal of eventually sequencing the human genome. It is anticipated that in the near future this effort will yield dense genetic maps of every chromosome, which will make it possible to map—and eventually sequence—genes responsible for a large number of human genetic diseases. It is also likely that the identification of major genes influencing complex disease phenotypes will proceed as well, albeit more slowly. The focus of the present paper is on perhaps the most difficult task, i.e., identifying chromosomal regions containing specific genes or gene clusters responsible for significant polygenic variation of a quantitative trait.

Most methods developed thus far for locating genes influencing quantitative traits have been dependent on selection experiments or on making special crosses, and therefore they are not, in general, applicable to human genetic research. For example, Harrison and Mather (1950) and Gibson and Thoday (1962), by selection experiments, were able to locate on a particular chromosome the polygenes affecting bristle number in *Drosophila*. Lander and Botstein (1989) devised a unique method of searching for quantitative-trait loci having a significant phenotypic effect in lower organisms, given a large number of offspring from specific crosses of inbred strains and a complete RFLP map of the genome in question. Although not applicable to human quantitative traits, this multipoint approach was successfully applied to several quantitative traits (fruit mass, concentration of soluble solids, and fruit pH) in the domestic tomato, *Lycopersicon esculentum*, by Paterson et al. (1988); a number of significant quantitative-trait loci were detected for each quantitative trait examined, and each was mapped to a specific chromosomal region.

In humans, on the other hand, most previous efforts at linkage of quantitative traits have used sib-pair or sibship methods for pairwise analyses of a hypothesized trait locus and a single marker locus. The first person to address the problem of linkage with a quantitative trait in humans was Penrose (1983), who developed a method which used the interaction between the marker genotype and the quantitative phenotype in sib

Received March 15, 1990; final revision received August 15, 1990.

Address for correspondence and reprints: David E. Goldgar, Ph.D., Genetic Epidemiology, University of Utah, 420 Chipeta Way, Suite 180, Salt Lake City, UT 84108.

0002-9297/90/4706-0012$02.00

pairs from particular parental matings as a test for linkage. Hill (1975) used a nested analysis-of-variance design for detecting and estimating linkage between a quantitative trait and a marker locus. Haseman and Elston (1972) utilized the concept of identity by descent (IBD) to devise a method for linkage of quantitative traits by using sib pairs. At any given genetic locus, two siblings can share either zero, one, or two genes IBD. Haseman and Elston derived the joint distribution of the number of genes IBD at a marker locus and the number of genes IBD at an hypothesized locus by determining a quantitative trait in terms of the recombination fraction between the two loci. They used these results to develop a method based on the regression of the squared sib-pair difference for the trait on the sibs' estimated genetic correlation at the marker locus.

The strategy adopted here is, in a sense, an extension of the Haseman and Elston approach to multiple markers and multiple siblings. The basic approach is to estimate the expected proportion of genetic material on a particular chromosome (or arm) shared IBD for each pair of siblings in a given sibship, on the basis of their genotypes at a series of marker loci on that chromosome. These estimates will be used to statistically partition the genetic variance of a quantitative trait into effects of loci on specific chromosomes or chromosomal regions. Risch and Lange (1979) and Suarez et al. (1979) showed that, for the entire genome, the variance of sib-pair genetic identity is quite small. In previous work (Goldgar and Kimberling 1980; Goldgar 1981) my colleagues and I showed that, when any one chromosome arm is considered, this variance is relatively high and that sib pairs who are nearly identical or (nonidentical) are frequent. In the present paper I will demonstrate that the data contained in available family panels such as the Centre d'Etude du Polymorphisme Humain (CEPH) resource or the collection of disease families oriented toward specific chromosomes are useful to reliably estimate chromosome-specific sib-pair identity and that these data provide sufficient power for detecting chromosome-specific polygenic variation under a wide variety of underlying models. The proposed method will thus be able to detect three types of genetic variation: (1) variation due to a single major locus; (2) variation due to several loci located in the same region or chromosome; and (3) variation due to several unlinked loci, each having a moderate influence on the quantitative trait. Existing methodology is largely aimed at the first of these situations. In addition, no other method uses multiple marker loci simultaneously to obtain more precise information regarding the number of crossovers occurring on a particular chromosomal segment. I begin with the derivation of the distribution of the proportion of genetic material shared IBD by a sibling pair, conditional on their genotypes at a set of syntenic marker loci.

Methods

Derivation of Sib-Pair Identity Conditional on Marker Information

In the present paper I will assume the Haldane (1919) model of recombination which assumes no interference and that the number of crossover events follows a Poisson distribution, although the results can be generalized to other recombination models. Specifically I make the following assumptions: (1) the number of crossover events occurring in a distance of λ Morgans follows a Poisson distribution with parameter λ; (2) for any given number of crossovers in an interval, the locations of these crossover events are uniformly distributed across the genetic map of the region (this follows from assumption 1; e.g., see Feller [1968]); and (3) the number of crossovers occurring during gametogenesis for each sibling is independent.

I initially consider two informative loci located on a particular chromosome that are separated by genetic distance λ Morgans. If the location of every crossover event occurring during meiosis for each sib could be observed, then the proportion of genetic material shared IBD, R, could be determined (at least in theory) exactly. Observation of the marker genotypes flanking the region of interest allows me to make certain inferences about R, given the assumptions stated above. Specifically, it can be shown that the probability density function of R, conditional on identity and recombination status, is given by a mixture of beta distributions, with the mixture proportions and parameters of each beta density corresponding to each possible number of crossover events compatible with the marker types (see the Appendix). For a given sib pair, I can derive the mean and variance of R, as a function of λ and the pattern of recombination and identity at the marker loci. Seven cases can be distinguished: (1) the pair are IBD at both loci, and neither sib is recombinant between the two loci; (2) the pair are different at both loci, and neither sib is recombinant; (3) the pair are IBD at both loci, and both sibs are recombinant; (4) the pair are different at both loci, and both sibs are recombinant; (5) the pair are IBD at one of the two loci and are different at the other; (6) the pair are IBD

Table 1

E(*R*) and V(*R*) Shared IBD by Two Siblings (S1 and S2), Conditional on IBD and Recombination Patterns at Two Marker Loci (M1 and M2)

IBD Status[a]		Recombination Status[a]			
M1	M2	S1	S2	E (*R*)	V (*R*)
I	I	NR	NR.....	$\frac{2(1-\theta)^2\lambda+\theta(1-\theta)+\lambda(1-2\theta)}{4(1-\theta)^2\lambda}$	$\frac{2\theta(1-\theta)^3\lambda+2(1-\theta)^2\lambda^2(1-2\theta)-2\theta(1-\theta)(1-2\theta)\lambda-\theta^2(1-\theta)^2-(1-2\theta)^2\lambda^2}{16(1-\theta)^4\lambda^2}$
N	N	NR	NR.....	$\frac{2(1-\theta)^2\lambda-\theta(1-\theta)-\lambda(1-2\theta)}{4(1-\theta)^2\lambda}$	$\frac{2\theta(1-\theta)^3\lambda+2(1-\theta)^2\lambda^2(1-2\theta)-2\theta(1-\theta)(1-2\theta)\lambda-\theta^2(1-\theta)^2-(1-2\theta)^2\lambda^2}{16(1-\theta)^4\lambda^2}$
I	I	R	R.....	$\frac{2\theta^2\lambda+\theta(1-\theta)-\lambda(1-2\theta)}{4\theta^2\lambda}$	$\frac{2\theta^3(1-\theta)\lambda-2\theta^2\lambda^2(1-2\theta)+2\theta(1-\theta)(1-2\theta)\lambda-\theta^2(1-\theta)^2-(1-2\theta)^2\lambda^2}{16\theta^4\lambda^2}$
N	N	R	R.....	$\frac{2\theta^2\lambda-\theta(1-\theta)+\lambda(1-2\theta)}{4\theta^2\lambda}$	$\frac{2\theta^3(1-\theta)\lambda-2\theta^2\lambda^2(1-2\theta)+2\theta(1-\theta)(1-2\theta)\lambda-\theta^2(1-\theta)^2-(1-2\theta)^2\lambda^2}{16\theta^4\lambda^2}$
I	N		.....	½	$\frac{\theta^2\lambda+(1-\theta)^2\lambda-\theta(1-\theta)}{16\theta(1-\theta)\lambda^2}$
N	I				
I	U		.....	$\frac{\lambda+\theta(1-\theta)}{2\lambda}$	$\frac{\lambda-2\theta^2(1-\theta)^2+\theta(1-\theta)}{8\lambda^2}$
U	I				
N	U		.....	$\frac{\lambda-\theta(1-\theta)}{2\lambda}$	$\frac{\lambda-2\theta^2(1-\theta)^2+\theta(1-\theta)}{8\lambda^2}$
U	N				

NOTE.—Example of derivation is given in the Appendix.

[a] I = Siblings IBD for marker locus; N = sib pair not IBD for marker locus; U = sib pair uninformative with respect to IBD status at marker locus; R = sib recombinant between M1 and M2; NR = sib nonrecombinant between M1 and M2.

[b] θ = Recombination fraction between M1 and M2 = $\frac{1}{2}(1-e^{-2\lambda})$.

at one of the two loci and are not informative at the other; and (7) the pair are different at one of the two loci and are not informative at the other. The latter two cases correspond to the region from the centromere to the first marker and from the last marker to the telomere when an entire chromosome arm is examined. The derivation of the mean and variance of R for the first case is provided in the Appendix. The other cases are derived analogously. Table 1 contains these means and variances as a function of distance λ for each of the seven cases outlined above, while figure 1 exhibits these relationships graphically as a function of distance between markers. In practice, the region of interest will be made up of a number of these segments. The overall mean and variance of R for a segment containing n such intervals of total length L are obtained as

$$E(R) = \sum_{i=1}^{n} \lambda_i E(R_i)/L$$

and

$$V(R) = \sum_{i=1}^{n} \lambda_i^2 V(R_i)/L^2,$$

where $L = \sum_{i=1}^{n} \lambda_i$.

To examine the distributions of estimated proportion shared IBD in human multipoint data, I chose for analysis a set of loci located on chromosome 7 that were extracted from the CEPH published data base. These loci were selected from a total of 63 markers examined by Barker et al. (1987) in a study of a subset of 21 of the CEPH families. From these 63 markers, I selected 22 probe-enzyme combinations reflecting 17 distinct loci from the CEPH data base. Criteria for selection were (1) number of informative meioses, (2) number of phase-known meioses, and (3) presence of an unambiguous map position on the published map (Barker et al. 1987). These loci were bounded by CRIL1020 and CRIL281, spanning a distance of 225 cM in females and 120 cM in males. A modified version of the CRIMAP program (Barker et al. 1987; Donis-Keller et al. 1987) was used to produce gametic strings under the most probable phase choice with consistent representation of allele origins (grandmaternal or grandpaternal) and to calculate the probability of the most probable phase. Offspring with either no genotyping or minimal typing were removed from the analysis. These 21 families yielded a total of 565 sib pairs. For each of the 565 possible sib pairs contained

in the 21 families, the procedure described above was used to obtain the expected value and variance of the distribution of R. Separate male and female genetic map distances were used in this process.

Statistical Methodology for Partitioning Genetic Variance to a Specific Chromosome

For the moment I will assume that the phenotype under study is determined by the additive effects of one or more loci located within the test chromosomal region (G_c), the additive effects of all other loci (G_A), and random environmental effects (E). The trait loci not on C are assumed to be independent of each other and of the loci on C. Letting X represent the phenotype under investigation, I represent the model as

$$\begin{aligned} X &= G_c + G_A + E; \\ E(G_c) &= 0,\ V(G_C) = V_C; \\ E(G_A) &= 0,\ V(G_A) = V_A; \\ E &\sim \text{Normal } (0, V_E). \end{aligned}$$

For simplicity I will assume that $V_T = V_C + V_A + V_E = 1$, that $V_G = V_A + V_C$ is the total genetic variance, and that $V_G/V_T = h^2$ is the heritability of X. Let R_{ij} be the true proportion of chromosome C shared IBD by sibs i and j. The covariance between the values for the two siblings is given by

$$\text{COV}(X_i, X_j) = R_{ij}V_C + V_A/2 = [R_{ij}P+(1-P)/2]\,V_G, \quad (1)$$

where P is the proportion of genetic variance due to loci on C. R_{ij} is an unobservable random variable; however, as shown above, the mean and variance of R_{ij} can be derived conditional on the identity and recombination pattern at a set of marker loci on C. For the moment I will use $R^* = [E(R_m)+E(R_p)]/2$ as my estimate of R, where R_m and R_p are the maternal and paternal genetic correlations, respectively.

For a sibship of size s, I have the observed phenotype vector x with expected value 0 and estimated covariance matrix

$$\underset{s\times s}{\overset{\wedge}{\sum}} = \{S_{ij}\} = [R^*_{ij}P+(1-P)/2]V_G \quad i \neq j \quad (2)$$

$$V_T \qquad\qquad i = j\,.$$

For the purpose of this analysis I consider the total genetic variance V_G (or, alternatively, h^2) and the total phenotypic variance V_T to be known constants, although they could be estimated from the same data set. The likelihood of observing the sibship phenotypic values under the above model is given by the multivariate normal density

$$2\pi^{-n/2}|\textstyle\sum|^{-1/2} \exp\{-1/2\ \mathbf{x}'\ \sum^{-1}\mathbf{x}\}\,. \quad (3)$$

For n such families the likelihood is the product of n such multivariate normal densities. Note that there is no requirement that the sibship size be the same in all families. When numerical optimization techniques are used, the maximum likelihood estimate of P can be obtained, and the null hypothesis of P = 0 can be tested using the generalized likelihood-ratio principle.

Effects of dominance at the trait loci, common sibling environment, and separate maternal and paternal effects present no theoretical difficulty and can easily be incorporated into the covariance structure.

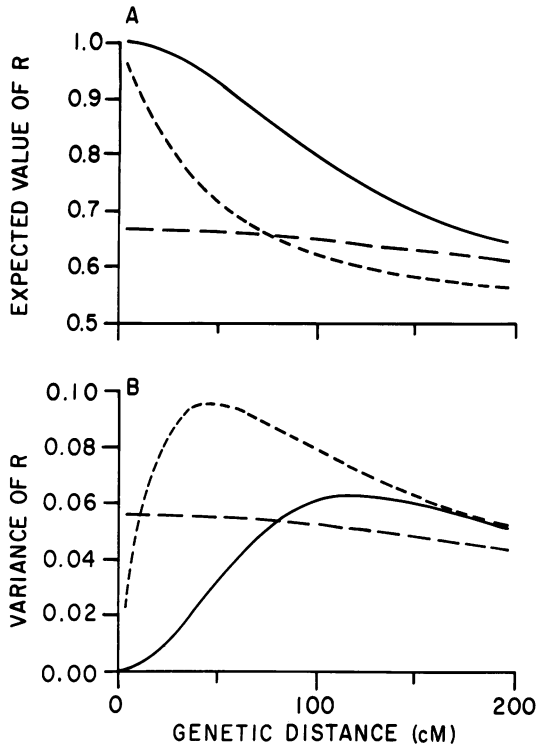

Figure 1 Graphs of (*A*) expected value and (*B*) variance of R, derived from a single parent, that is shared IBD by two sibs, as a function of genetic distance between two marker loci. I I - NR (—) represents the case where both sibs are identical by descent at both marker loci and neither sib shows observable recombination; I I - R (- -) represents the analogous case where both sibs are recombinant; and I U (- - - -) represents the case where the sibs are identical at one locus and are not informative at the other. The functions depicted in the graph are given in table 1.

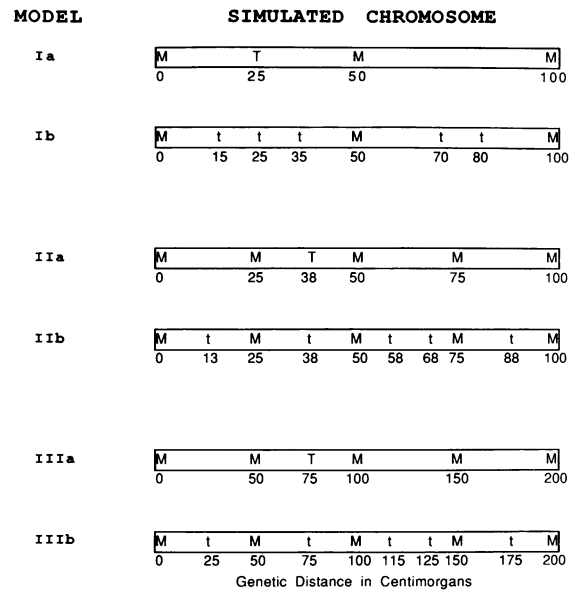

Figure 2 Estimated genetic correlations, i.e., E(R), for human multipoint data from chromosome 7. Data consist of 22 loci typed in 21 CEPH families and span a distance of 120 cM in males and 225 cM in females. A total of 565 sib pairs were analyzed. *A*, Estimated genetic correlation on maternally derived chromosome. *B*, Estimated genetic correlation on paternally derived chromosome. *C*, Combined maternal and paternal estimated genetic correlations.

Simulation Method

To demonstrate the overall utility of the method, a Monte Carlo simulation study was performed. Three combinations of marker density and chromosome length were examined in the simulation study: (1) a 100-cM length with a distance of 50 cM between markers (Model I); (2) a 100-cM arm with markers every 25 cM (Model II); and (3) a 200-cM length with 50 cM between markers (Model III). For each of these models of marker density and chromosome length, two models of the effect of loci on the test chromosome were examined: a single trait locus located midway between two markers (Models Ia, IIa, and IIIa) and a model in which five trait loci are scattered along the length of the test chromosome whose total effect is equal to that of the major locus (Models Ib, IIb, and IIIb). These six basic marker/trait-locus configurations are shown in detail in figure 2.

In all cases 10 additional loci were simulated independently both of each other and of the trait loci on the test chromosome C. Each trait locus is assumed to have two equally frequent alleles T and t, with genotypic effects given by 2a, 0, and $-2a$ for genotypes TT, Tt, and tt, respectively. The values of the trait-locus effects, a, are chosen so that the trait locus(loci) on C and the other trait loci reflect simulated values of P and h^2. All marker loci were chosen to be fully informative. Given the increasing numbers of VNTRs and other highly informative systems, and given the low density of the map required, this is not a restrictive assumption. Rather than estimate sample size for fixed power, I examined power for various values of P and h^2 by assuming a fixed sample of 40 families with eight offspring each. This number was chosen to represent large data sets, such as the CEPH or Venezuelan pedigrees, on which a large number of markers have already been typed. In accordance with this I have also assumed in the simulation that the marker loci are phase known. The simulation procedure is as follows:

1. Haplotypes at the marker and trait loci are simulated for the two parents, with the restriction that the marker loci are always informative.
2. Each offspring is generated by selecting at random either the grandmaternal or grandpaternal chromosome. Then recombination is simulated according to the recombination fraction between each successive pair of loci. The recombination fractions are converted from genetic distance by using Haldane's function. This process is repeated independently for each parent for each offspring.
3. Parental genotypes at the 10 additional trait loci are generated, and one allele is transmitted at random to each offspring from each parent.
4. A normal random deviate with mean 0 and variance $(1-h^2)$ is obtained using a normal random-number generator. The trait value x is computed for each offspring by adding both the appropriate genotypic effect at each trait locus and the individual-specific normal environmental component.
5. The simulated marker data are used to estimate the mean and variance of the proportion of the test chromosome shared IBD for each sib pair. The log likelihood for the sibship is calculated using equation (3).

This procedure is repeated for 40 such families. The likelihood is calculated at a grid of values of P, and the maximum likelihood estimate of P is obtained using quadratic interpolation of the likelihood (Ott 1985). The hypothesis of P = 0 is tested using the χ^2 approximation to the generalized likelihood ratio test with a nominal significance level of .05.

Three values of heritability (.75, .5, and .25) were simulated for five values of P (.0, .25, .5, .75, and 1.0).

For each combination of factors 100 data sets were replicated for each of the six chromosome models. For each, the mean and variance of the estimates of P and the number of times the hypothesis P = 0 was rejected were recorded.

A second simulation was performed to directly compare the method proposed here with an existing method—specifically, the Haseman and Elston (1972) sib-pair method which was modified to analyze multiple markers by using a multiple-regression approach as outlined below. The model simulated was a quantitative trait with a heritability of .75. The chromosome model was two fully informative marker loci separated by a distance of D = 30, 50, and 70 cM. A single trait locus accounting for 75% of the genetic variance (56% of the phenotypic variance) was located midway between the two markers. In addition, the case in which there were no loci affecting the trait in the region was analyzed to compare the empirical size of the two methods. Each data set consisted of 400 sib pairs and was analyzed by both methods. For each marker distance, 500 replications were performed. To implement the Haseman and Elston sib-pair procedure a multiple regression was performed using the IBD status of each marker as the independent variable and by using the squared sib-pair difference of the simulated quantitative trait as the dependent variable. The null hypothesis of no effect of loci between the two markers on the quantitative trait was rejected if either (or both) estimated regression coefficients was significantly less than zero when compared with its estimated standard error. The critical value for the test of each marker was chosen to reflect a one-sided P value of .0253, to achieve a nominal significance level for the overall test, when both markers were considered jointly, of .05.

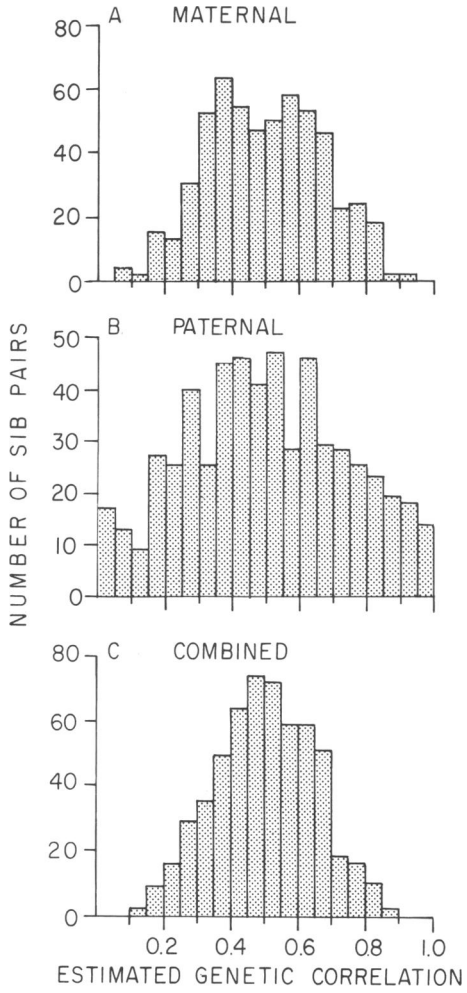

Figure 3 Descriptions of chromosome models used in simulation study. M = fully informative marker locus; T-trait locus of large effect; t = trait locus of small effect. The effects of trait loci on the quantitative phenotype are chosen so that the genetic variance attributable to five t loci is equal to that of the major locus T.

Results

The distribution of estimated proportions of maternal, paternal, and total identity for the region of chromosome 7 spanned by CRIL1020 to CRIL281 is shown in figures 3*A*, 3*B*, and 3*C*, respectively. As seen in figure 3*A*, when 22 polymorphic loci on chromosome 7 are used, the distribution of the maternal correlations is fairly narrow, with only 9% of the 565 sib-pair correlations being within the upper and lower quintiles; the paternal correlations (fig. 3*B*), as one would expect given the much shorter map (120 cM vs. 225 cM), showed a broader distribution, with 25% of the correlations being within these quintiles. Relative to the estimated variances, 23%, 41%, and 31% of maternal, paternal, and total correlations, respectively, were significantly different from .5.

Table 2 shows the results of the simulation study. For each of the six chromosome models, the average value of the estimate of P for 100 independent replications is given together with the number of these 100 replications in which the null hypothesis of P = 0 was rejected. The empirical size of the test, computed by averaging over all replications in which the true value of P was .0, was .042 (N = 1,800 total trials). This was not significantly lower than the nominal value of .05. Table 2 also shows that, for all models, reasonable empirical power (EP) is obtained when the proportion

Table 2

Results of Simulations

h^2 AND P	MODEL I				MODEL II				MODEL III			
	a		b		a		b		a		b	
	$\bar{\hat{P}}$	EP[a]	$\bar{\hat{P}}$	EP[a]	$\bar{\hat{P}}$	EP[a]	$\bar{\hat{P}}$	EP[a]	$\bar{\hat{P}}$	EP[a]	$\bar{\hat{P}}$	EP[a]
.75:												
.00.....	.06	4	.07	1	.05	3	.05	5	.09	4	.07	2
.25.....	.22	37	.23	38	.26	51	.26	64	.27	35	.25	32
.50.....	.45	89	.43	83	.50	97	.48	93	.46	73	.43	65
.75.....	.63	94	.68	99	.75	100	.75	100	.70	94	.69	95
1.00.....	.85	100	.88	100	.96	100	.93	100	.87	100	.85	100
.50:												
.00.....	.11	7	.12	6	.09	5	.10	3	.11	1	.15	6
.25.....	.26	20	.26	20	.28	25	.25	18	.29	17	.31	23
.50.....	.50	50	.52	53	.54	75	.47	56	.55	45	.46	36
.75.....	.59	70	.63	71	.78	94	.71	87	.68	66	.64	58
1.00.....	.82	90	.83	93	.88	98	.90	98	.81	81	.81	81
.25:												
.00.....	.26	7	.22	4	.25	6	.19	4	.25	6	.33	2
.25.....	.32	14	.29	10	.33	13	.34	13	.42	10	.32	6
.50.....	.37	12	.46	19	.52	27	.55	27	.46	13	.43	7
.75.....	.55	26	.67	35	.61	31	.61	35	.62	20	.62	23
1.00.....	.72	47	.70	50	.83	62	.76	53	.65	22	.71	26

NOTE.—Values in table are based on 100 replications. Models are described in the text and shown in fig. 3. $\bar{\hat{P}}$ = average estimate of P from 100 trials.

[a] % of 100 trials in which null hypothesis of P = 0 was rejected.

of phenotypic variance explained by trait loci located on the test chromosome is greater than about 35%. As expected, the accuracy of the estimates of P were much higher for the denser map (L = 100 cM; D = 25 cM), although, even in this case, estimates at the boundary were significantly biased as a result of the restriction on the parameter space.

Table 3 shows the results of the comparisons of the method described in the present paper with the Haseman and Elston sib-pair method. EP for the multipoint approach was, on average, about 50%–80% higher than that for the Haseman and Elston regression approach. The relative advantage of the proposed method over the Haseman and Elston method appeared to increase as a function of distance between the two markers (and, consequently, as a function of the distance between the trait major locus and each marker).

Discussion

The present paper outlines a new method for analyzing specific genetic components involved in the determination of human quantitative multifactorial traits. Previous methods, for the most part, have utilized sib pairs and examined only a single marker locus, although Hill (1975) described an analysis-of-variance approach

Table 3

Results of Simulations Comparing Proposed Method with Haseman and Elston Sib-Pair Procedure

D (cM)	EP (%)			
	Proposed Method		Haseman and Elston Method	
	P^a = .75	P^a = .0	P^a = .75	P^a = .0
30	97.4	5.4	71.8	4.8
50	84.8	4.6	51.4	3.8
70	60.8	5.0	33.6	4.4

NOTE.—Data are percentages of 500 replications in which the null hypothesis was rejected. The model and procedure used are described in the text.

[a] Simulated proportion of genetic variance due to major locus located midway between the two marker loci.

based on sibship data. Although interval-mapping approaches have been described for lower organisms (Lander and Botstein 1989), their general applicability to human data has not yet been demonstrated. The method described in the present paper has specifically been designed to take advantage of the pedigree structures and multilocus genotypic data inherent in collections of reference families used for general construction of genetic maps. Specifically, parental phase is often known with certainty for many loci, and, for a given sibship for multiple linked loci, even when many (or all) loci are phase unknown, the most probable phase typically has a high probability relative to alternative phase choices. Thus chromosome-specific genetic correlations for each sib pair can easily be computed using the methods provided in the present paper.

The simulation studies showed that adequate power was achieved for a variety of models of marker density and effects of the trait locus (loci). It is surprising that the number of trait loci on the test chromosome (one vs. five) did not have a great influence on either the power of the test or the accuracy of the estimates. Comparison of the results for the model having a 100-cM chromosomal segment and a marker density of 50 cM with that for the model having a 200-cM segment of equal density shows about a 10%–20% reduction in power for the longer segment versus the shorter segment. It is also apparent that for a sample of this size it may not be fruitful to search for linkage of a trait that has heritability much less than 50%. For a fixed proportion of total phenotypic variance accounted for by trait loci on the test chromosome, higher power was achieved when there was less environmental variation relative to additional genetic variation on other chromosomes. It should be emphasized that all the simulated quantitative traits analyzed as part of the simulation study were of an "ideal" nature. That is, each locus contributing to the trait was assumed to have two equally frequent, completely additive alleles, there was no epistasis, and the environmental effects were normally distributed. Thus, the EP figures shown in table 2 are perhaps somewhat higher than one might achieve for traits likely to be encountered in practice. Further work is necessary to examine the method's robustness with regard to departures from this idealized model.

The explicit comparisons with the Haseman and Elston sib-pair method show that the use of a true multipoint approach—rather than a multiple marker approach—provides higher power for detecting the effects that a major locus has on a quantitative phenotype. This was true even for the case of sib *pairs*, rather than the larger sibships for which this method was designed. The proposed method avoids the statistical problem of analyzing a sibship as a set of mutually independent sib pairs, an issue which has not been totally resolved (Amos et al. 1989; Demenais and Amos 1989). A more detailed comparison of the proposed method with the Haseman and Elston procedure and its extensions under a more complete set of models is underway and will be the subject of a subsequent paper.

In making comparisons to the sib-pair method I should point out that my method tests a more general hypothesis than does either the sib-pair method or the method discussed by Lander and Botstein (1989). While other methods frame their hypotheses in terms of linkage/recombination, my method is designed to detect the effect of all loci located in a specific chromosomal region on a quantitative phenotype. This region may be as large as an entire chromosome or a single 20-cM interval between marker loci. Thus it would be anticipated that, in a case in which loci which influencing a quantitative trait are spread out along a segment of chromosome, the method proposed here would succeed while others most likely would fail to detect this effect.

Clearly there is a trade-off between the size of the region studied and the power to detect effects of loci in that region. In the limiting case, one could analyze as a separate region each interval between adjacent loci; however, this presents a multiple-comparisons problem, since a large number of possibly nonindependent analyses would be performed. At the other extreme, if the region is too large, then power to detect the effect of trait loci located in a small subregion would be reduced. The results of the simulation study indicate that a reasonable choice for regions to be studied would be (*a*) individual chromosome arms, for the longer metacentric chromosomes, and (*b*) entire chromsomes, for acrocentric and shorter metacentric chromosomes. My experience with this method indicates that a relatively small number of highly polymorphic markers spaced 25–50 cM apart are preferable to a more dense map of less informative markers. In a manner analogous to a genomic search followed by fine mapping for a Mendelian disease, when the effects of loci on a quantitative phenotype in a particular region are detected, the chromosome or arm could be divided into two or more subregions bounded by marker loci. These subregions could be analyzed simultaneously to determine the region producing the largest effect on the trait under study. I am currently planning studies to directly compare my approach with existing methods, to test more complex

models by incorporating common sibling environment or dominance variation, and to incorporate the variance as well as expected value of sib-pair identity into the analysis.

Acknowledgments

This research was supported in part by National Institutes of Health grants CA-36362 and CA-48711. The author also wishes to thank Dr. Phil Green, for modifications to the CRIMAP program that were necessary for the analysis of the chromosome 7 data, and Drs. Timothy Bishop and Mark Skolnick, for their helpful comments.

Appendix

Derivation of Expected Value and Variance of Proportion of Genetic Material between Two Marker Loci Shared IBD by Two siblings, Conditional on IBD Status at Each Locus and on Recombination Pattern between the Two Loci

I shall assume the Haldane (1919) model of crossing-over as a Poisson process. This model implies that there is no interference and that, for a fixed number of crossovers in a given interval, the locations of crossover are uniformally distributed in the interval. I also assume that the number and locations of crossovers occurring during meiosis for a given offspring are independent of those for any other offspring. The following derivation is for the case in which the sibs are IBD at both marker loci and in which there is no observable recombination between the two loci. The other cases outlined in the text are similarly derived. I begin by showing that, for a fixed number of crossover events between the markers, the proportion shared IBD in the region for two sibs is distributed as a beta random variable.

Specifically, let k be the total number of crossovers occurring during both meioses (either maternal or paternal) for both sibs under consideration. Further, let x_i, $i = 1,2, \ldots, k$ be the locations of the k crossovers within the interval. It is easily demonstrated that the proportion shared IBD in the interval does not depend on which of the k occurred in which meiosis; for simplicity I can assume that all crossovers occurred in the first meiosis and that none occurred in the second. Assume that the sibs are IBD at the first locus; the derivation for the case in which the sibs are nonidentical at this locus is similar, with D representing the proportion shared by the sibs and with $R = D$ being substituted for $R = 1-D$. Let y_i, $i = 1,2, \ldots, k$ be the ordered positions of the $\{xi\}$. The proportion D, for which the sibs are not IBD, is given by

$$D = \begin{cases} 0 & \text{for } k = 0 ; \\ \sum_{i=1}^{k/2} (y_{2i} - y_{2i-1}) & \text{for } k \text{ even} ; \\ \sum_{i=1}^{(k-1)/2} (y_{2i} - y_{2i-1}) + 1 - y_k & \text{for } k \text{ odd} . \end{cases}$$

In the terminology and theorems of Wilks (1962), the difference between successive-order statistics is called a "sample block" and the sum of n such sample blocks is called a "coverage." When the sampling distribution of the x_i is independent uniform (0,1) then the probability density function of the coverage is given by a beta distribution with parameters n and $k-n+1$. For k odd we see that D is the sum of $(k+1)/2$ such blocks, and for k even D is the sum of $k/2$ blocks. Thus

$$\begin{aligned} &D = 0 \text{ with probability 1} && \text{for } k = 0 ; \\ &D \sim \text{Be}(k/2, k/2 + 1) && \text{for } k = 2, 4, \ldots ; \\ &D \sim \text{Be}((k+1)/2, (k+1)/2) && \text{for } k = 1, 3, \ldots . \end{aligned}$$

Since D is the proportion not shared, I let $R = 1 - D$ denote the proportion shared IBD by the two sibs. Since, if X is Be(a,b) then $Y = (1-X)$ is Be(b,a), the distribution of R, conditional on k, is given by

$$\begin{aligned} &f(R|k) = 1 \text{ with probability 1} && \text{for } k = 0 ; \\ &f(R|k) \sim \text{Be}((k+2)/2, k/2) && \text{for } k = 2, 4, \ldots ; \\ &f(R|k) \sim \text{Be}((k+1)/2, (k+1)/2) && \text{for } k = 1, 3, \ldots . \end{aligned}$$

For the derivation of the overall mean and variance of R for all crossover distributions compatible with the identity and recombination pattern, we will need the expected value of R and expected value of R^2 for a given k. This is obtained from the properties of the beta distribution, as

$$E(R|k) = \begin{cases} (k+2)/2(k+1) & \text{for } k = 0, 2, 4, \ldots \\ 1/2 & \text{for } k = 1, 3, 5, \ldots \end{cases} \quad \text{(A1)}$$

and

$$E(R^2|k) = \begin{cases} (k+4)/4(k+1) & \text{for } k = 0, 2, 4, \ldots \\ (k+3)/4(k+2) & \text{for } k = 1, 3, 5, \ldots . \end{cases} \quad \text{(A2)}$$

We want the expected value and variance of R, given

that the sibs are IBD at both markers and show no recombination between the marker loci (I I – NR).

$$E(R|\text{sibs I I} - \text{NR}) = \sum_k E(R|k \text{ crossovers}) \times \Pr(k \text{ crossovers } |\text{I I} - \text{NR}) \quad \text{(A3)}$$

and

$$E(R^2|\text{sibs I I} - \text{NR}) = \sum_k E(R^2|k \text{ crossovers}) \times \Pr(k \text{ crossovers}|\text{I I} - \text{NR}) . \quad \text{(A4)}$$

The fact that the sibs are IBD at both loci implies that the total number of crossover events for the two sibs from that parent was even. Moreover, the fact that there was no observable recombination between the loci for either sib implies that the number of crossovers occurring for each sib was even. Thus, for a distance of λ Morgans between the two marker loci,

$$\Pr(k|\text{I I} - \text{NR}) = \frac{\Pr(2i \text{ crossovers sib 1}) \Pr(2j \text{ crossovers sib 2})}{\Pr(\text{No recombination between markers in two meisoses})}$$

$$i, j = 0, 1, 2, \ldots \quad \text{(A5)}$$

$$= \frac{\lambda^{2i} e^{-\lambda}/(2i)! \; \lambda^{2j} e^{-\lambda}/(2j)!}{(1-\theta)^2} ,$$

where $k = 2i + 2j$ and where θ, the recombination fraction between the two markers, is $1/2\,(1-e^{-2\lambda})$. As shown earlier, $E(R|k) = (k+2)/2(k+1)$, and, replacing k by $2i + 2j$ and substituting the latter into equation (3), I have

$$E(R|\text{sibs I I} - \text{NR}) = \frac{1}{(1-\theta)^2} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{\lambda^{2i} e^{-\lambda}}{(2i)!} \frac{\lambda 2j \; e^{-\lambda}}{(2j)!} \frac{(2i+2j+2)}{2(2i+2j+1)}$$

$$= \frac{1}{(1-\theta)^2} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{\lambda^{2(i+j)} e^{-2\lambda}}{\{2(i+j)\}!} \frac{[2(i+j)]!}{(2i)!(2j)!} \frac{(2i+2j+2)}{2(2i+2j+1)} .$$

Letting $m = i + j$, I have

$$\frac{1}{(1-\theta)^2} \sum_{m=0}^{\infty} \sum_{j=0}^{\infty} \frac{\lambda^{2m} e^{-2\lambda}}{(2m)!} \binom{2m}{2j} \frac{(2m+2)}{2(2m+1)}$$

$$= \frac{1}{(1-\theta)^2} \sum_{m=0}^{\infty} \frac{(2\lambda)^{2m} e^{-2\lambda}}{(2m)!} \frac{(2m+2)}{2(2m+1)} \sum_{j=0}^{m} \binom{2m}{2j} (1/2)^{2m} .$$

The latter summation is equal to 1 for $m = 0$, and for $m > 0$ it is the first m terms of a binomial distribution with parameters $n = 2m$ and $p = 1/2$; thus, in this case, this sum is equal to 1/2. Using this relation and rewriting $(2m+2)/2(2m+1)$ as $[1+1/(2m+1)]/2$, I get

$$\frac{1}{4(1-\theta)^2} \sum_{m=0}^{\infty} \frac{(2\lambda)^{2m} e^{-2\lambda}}{(2m)!} + \sum_{k=2}^{\infty} \frac{(2\lambda)^{2m} e^{-2\lambda}}{(2m+1)\,(2m)!} + 2e^{-2\lambda} .$$

The first term in this summation is equal to the probability of an even number of crossover events in a distance of 2λ, which is equal to $1 - \theta(2\lambda) = (1+e^{-4\lambda})/2$. The second term is $1/2\lambda$ times the probability of an odd number of crossovers in a distance of $2\lambda = (1-e^{-4\lambda})/2$. The conditional expectation I want can now be written as

$$E(R|\text{sibs I I} - \text{NR}) = \frac{(1 + e^{-4\lambda})/2 + (1 - e^{-4\lambda})/4\lambda + 2e^{-2\lambda}}{4(1-\theta)^2} .$$

After applying a bit of algebra and substituting $\theta = 1/2\,(1-e^{-2\lambda})$ in the numerator, I get the result shown in table 1, namely,

$$E(R|\text{sibs I I} - \text{NR}) = \frac{2(1-\theta)^2\lambda + \theta(1-\theta) + \lambda(1-2\theta)}{4(1-\theta)^2\,\lambda} .$$

A similar process substituting $E(R^2|k) = (k+4)/4(k+1)$ for $E(R|k)$ allows me to obtain $E(R^2|\text{sibs I I} - \text{NR})$ as

$$E(R^2|\text{sibs I I} - \text{NR}) = \frac{(1+e^{-4\lambda})/2 + 3(1-e^{-4\lambda})/4\lambda + 4e^{-2\lambda}}{8(1-\theta)^2}$$

$$= \frac{2(1+\theta)^2\lambda + 3\theta(1-\theta) + 3\lambda(1-2\theta)}{8(1-\theta)^2\lambda} ,$$

and from this the variance can be obtained as

$$V(R|\text{sibs I I} - \text{NR}) = E(R^2|\text{sibs I I} - \text{NR}) - [E(R|\text{sibs I I} - \text{NR})]^2 .$$

References

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

Barker DF, Green P, Knowlton R, Schumm J, Lander E, Oliphat A, Willard H, et al (1987) Genetic linkage map of human chromosome 7 with 63 DNA markers. Proc Natl Acad Sci USA 84:8006–8010

Demenais FM, Amos CI (1989) Power of the sib-pair and lod-score methods for linkage analysis of quantitative traits. In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based upon affected pedigree members: Genetic Analysis Workshop 6. Alan R Liss, New York, pp 201–206

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, et al (1987) A genetic linkage map of the human genome. Cell 51:319–337

Feller W (1968) An introduction to probability theory and its applications, vol 1, 3d ed. John Wiley & Sons, New York

Gibson JB, Thoday JM (1962) Effect of disruptive selection. VI. A second chromosome polymorphism. Heredity 17:1–26

Goldgar DE (1981) Partitioning the genetic variance of a quantitative trait to specific chromosomes: an alternative approach to quantitative linkage analysis. PhD diss, University of Colorado, Boulder

Goldgar DE, Kimberling WJ (1980) The partitioning of genetic variance to specific linkage groups. Am J Hum Genet 32:143(A)

Haldane JBS (1919) The combination of linkage values and the calculation of distance between the loci of linked factors. J Genet 8:299–309

Harrison BJ, Mather K (1950) Polygenic variability in chromosomes of drosophila melanogaster obtained from the wild. Heredity 4:295–312

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

Hill AP (1975) Quantitative linkage: a statistical procedure for its detection and estimation. Ann Hum Genet 38: 443–449

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Ott J (1985) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335: 721–726

Penrose LS (1983) Genetic linkage in graded human characters. Ann Eugenics 9:133–138

Risch N, Lange K (1979) Application of a recombination model in calculating the variance of sib pair genetic identity. Ann Hum Genet 43:177–186

Suarez B, Reich T, Fishman PM (1979) Variability in sib pair genetic identity. Hum Heredity 29:37–41

Wilks SS (1962) Mathematical statistics. John Wiley & Sons, New York