

Multipoint Linkage Methods for Localizing Susceptibility Genes of Complex Diseases Based on Affected Sibling Pairs

DV Glidden. Division of Biostatistics, UCSF; K-Y Liang. Department of Biostatistics, Johns Hopkins; Y-F Chiu. Department of Biostatistics, University of North Carolina; AE Pulver. Department of Psychiatry, Johns Hopkins

Problem

Multiple markers from affected siblings indicate a region is linked to a disease gene. How to estimate the location of the putative disease locus? Can covariate information improve localization?

Multipoint Locus Estimation

Liang *et al.* [2001] proposed a general multipoint method, which estimates the position of a susceptibility locus, denoted τ , based on allele sharing among affected siblings (ASP). Let R denote a chromosomal region of length T cM with M typed markers at loci

$0 \leq t_1, \dots, t_M \leq T$. Let $S_i(t)$ be the number of alleles shared identical-by-descent (IBD) at locus t $0 \leq t \leq T$ for the i th sibling pair. With a single locus in R , Liang *et al.* showed

$$\mu(t) = E\{S(t)\} = 1 + \{1 - 2\theta(t, \tau)\}^2 [E\{S(\tau)\} - 1],$$

where $\theta(t, \tau)$ is the recombination fraction between t and τ . Using the Haldane function, yields

$$\mu(t; \tau, C) = 1 + \exp(-0.041 t - \tau) C,$$

where $(1+C)/2$ is the fraction of allele shared IBD by ASPs. A C value of zero means no linkage of region R .

The allele sharing on depends on τ and C . Locus location is given by τ and C measures the degree of overall linkage to R and determines the precision with which t is estimated. C is also the parameter that incorporates covariate data.

Incorporating Covariates

In complex diseases multiple factors interplay to cause disease. This can give to weak linkage signals and contribute to inconsistent findings for linkage. When covariates are predictive of linkage, the value of C varies according to subgroups of pairs defined by external information — age at onset or other covariates.

Let X denote a set of covariates. Under the assumptions of Liang and assuming recombination doesn't depend on X ,

$$\mu(t; \tau, C | X=x) = 1 + \exp(-0.041 t - \tau) C(x),$$

where $C(x) = E\{S(\tau) | X=x\}$. This is true even for sampling that depends on the covariates (e.g., oversampling of early-onset sib pairs).

By fitting the model we obtain estimates of τ and a form for C . When heterogeneity exists, exploiting the formulation will permit more precise estimates of τ . This uses all the available data yet allows the subjects with the strongest 'genetic signal' to provide the bulk of the information about τ .

Multiple parametric model could be used for $C(x)$. Estimation for such a model can be based on generalized estimating equations. The method only requires that the form for μ is correctly specified. One approach is to define L categories based on the covariates, each with C_i ($i=1, \dots, L$). This makes no assumption about the dependence of C on the covariates. However, the covariate groupings may be arbitrary and the number of groups should be kept small.

Data Example

We applied multipoint methods to a 64 ASP schizophrenia linkage study [Blouin *et al.* 1998] that found strong evidence of linkage to 13q32. We examine estimates of susceptibility loci on chromosome 13, incorporating the age at onset information.

Subjects were considered to have early-onset if the age at onset of schizophrenia ≤ 21 years, the median onset age among affected, and late-onset otherwise. Based on this classification, 20 siblings pairs were both early-onset subjects (EE), 27 pairs were both late-onset subjects (LL) and 17 pairs had one early-onset and one late-onset sibling (EL). We incorporated age of onset by allowing the above formulation with C_i ($i=1,2,3$) for EE, LL and EL pairs respectively.

The estimate of t and its standard error (SE) were compared with the approach of Liang *et al.* [2001]. Both estimates will be unbiased, differing only in their precision.

Applying the method of Liang *et al.* [2001] estimates tau equal to 113.4cM with a SE of 2.40. The estimated $C=0.34$ with a SE of 0.10. Incorporating age-at-onset yields an estimated $\tau=112.9$ cM on a standard error of 1.60. The gain in precision is the equivalent of a doubling of the sample size. The gain comes from the heterogeneity in the estimated allele sharing at t which are $C_1=0.66$, $C_2=0.18$, $C_3=0.21$ with SEs of 0.17, 0.13 and 0.17, respectively. There appears to be a very strong genetic signal for early onset pairs (EE) with an estimated allele sharing equal to $(1+0.66)/2 = 0.83$ at τ .

Simulation Studies

We generated data with 11 fully polymorphic markers spaced every 10cM from 0 to 100 cM with $\tau=45$. ASPs were grouped by age at onset of disease into EE, EL and LL. We simulated data with sample sizes of 100, 250, and 500 ASPs. We compared the standard GENEFINDER estimate (τ_2) to the approach with separate C for age at onset groups (τ_1). The results in *Table 1* confirm the efficiency gains of our method.

Table 1: Mean Standard Error and Relative Efficiency of New Method (τ_1) vs. Standard GENEFINDER (τ_2) with Age at Onset Effect Similar to Schizophrenia Data Example

No Sib Pairs	Mean	SE	Rel. Efft of τ_1 to τ_2
100	τ_1 45.0	4.06	
	τ_2 45.0	4.47	1.21
250	τ_1 45.0	1.91	
	τ_2 45.0	2.41	1.58
500	τ_1 45.0	1.23	
	τ_2 45.0	1.50	1.49

To examine efficiency with is no age-at-onset heterogeneity, we simulated data as above but with IBD sharing at τ identical for all ages at onset. The analysis of 50,000 datasets, shown in *Table 2*, suggests minimal cost from applying the extended GENEFINDER method when there is no age-at-onset effect.

Table 2: Mean Standard Error and Relative Efficiency of New Method (τ_1) vs. Standard GENEFINDER (τ_2) with No Age at Onset Effect

No Sib Pairs	Mean	SE	Rel. Efft of τ_1 to τ_2
100	τ_1 45.0	4.75	
	τ_2 45.1	4.56	0.92
250	τ_1 45.0	2.44	
	τ_2 45.0	2.43	0.99
500	τ_1 45.0	1.51	
	τ_2 45.0	1.51	1.01

Conclusions

We developed a family of approaches for exploring, testing and exploiting covariates in multipoint locus estimation and saw increased efficiency from their use. The method is computationally simple and uses all available data. Our example demonstrates even elementary modeling may increase precision. Efficiency loss is negligible for non-predictive covariate.

Email: dave@biostat.ucsf.edu

URL: <http://www.epibiostat.ucsf.edu/dave>

GENEFINDER: <http://www.biostat.jhsph.edu/biostat/research/genefinder.shtml>

Genetic Epidemiology, in press.