Genetic Epidemiology 12:877–882 (1995)

Detection of Genome Similarity as an Exploratory Tool for Mapping Complex Traits

Sun-Wei Guo

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

For one- and two-trait-locus models, we show that the lod score based on affected relative pairs or trios is a monotonically increasing function of the genome similarity measured by the proportion of alleles shared identical by descent (IBD) conditional on observed marker data. These results can be generalized to multi-trait-locus models. Thus, we can use conditional probability of genes shared IBD as a tool to reveal chromosomal segments that are likely to harbor the genes underlying the complex traits. ©1995 Wiley-Liss, Inc.

Key words: affected relative method, exploratory analysis, genetic model, genome similarity, linkage, robust linkage methods

INTRODUCTION

One of the problems in the genetic analysis of complex genetic traits is to evaluate the statistical significance of a linkage result. The problem arises when one tests multiple genetic models and/or multiple markers. Both practices are fishing expeditions because of uncertainties. Because of its obvious importance in practice, the problem has drawn considerable attention. Thompson [1984], Ott [1985], Weeks et al. [1990], Risch [1991], and Kong et al. [1992] have considered the problem from various aspects.

In this paper, we consider two questions. First, given pedigree and marker data and without knowledge of the true genetic model underlying the trait, can we develop an exploratory method to reveal a region or regions in the genome that are likely to harbor trait-

Address reprint requests to Dr. Sun-Wei Guo, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029

causing gene(s)? Second, does simultaneous use of multiple affected relatives (as compared with pairwise relationships) give more information for gene mapping? The answers to both questions are positive, at least for affected relative pairs and trios under arbitrary genetic models. We will propose two measures of genetic similarity which may be used to detect a trait-causing locus or loci, as well as to suggest plausible genetic models underlying the trait. The method proposed in this paper, in conjunction with a recently proposed method for probabilistic determination of multi-locus IBD for pedigrees [Guo, unpublished], provides a practical tool for mapping complex genetic traits without knowledge of the underlying model.

ONE-LOCUS MODEL

Let $\mathcal{A}$ denote the event that a (prespecified) set of individuals in a given pedigree are all affected. Let $\mathcal{M}$ be the marker data observed on the pedigree. The marker data may be multilocus, and will be specified in the context. A generalization of Risch's [1990b] likelihood ratio for observing $\mathcal{M}$ given two affected relatives gives

$$\Lambda = \frac{P(\mathcal{M}|\mathcal{A})}{P(\mathcal{M})} = \frac{P(\mathcal{A}|\mathcal{M})}{P(\mathcal{A})} = \frac{\sum_{\pi_t} P(\mathcal{A}|\pi_t)P(\pi_t|\mathcal{M})}{\sum_{\pi_t} P(\mathcal{A}|\pi_t)P(\pi_t)} \tag{1}$$

where π_t is the proportion of alleles shared (IBD) at the trait locus t.

The above equation holds regardless of the true underlying genetic model and how many genes are involved. The base-10 logarithm of Λ is the lod score [Risch 1990b]. However, (1) is not very informative, because $P(\mathcal{A}|\pi_t)$ is in general very complicated, despite the fact that $P(\pi_t)$ and $P(\pi_t|\mathcal{M})$ can be calculated (see below). We now consider (a) a single-locus trait X, with $X = 1$ if an individual is affected with the trait or $X = 0$ if not, and (b) $\mathcal{A}$ consists of only two affected relatives, i and j. Following James [1971], for the ith member of a pedigree,

$$X_i = K + \alpha_{g_i} + \alpha_{h_i} + d_{g_i h_i} \tag{2}$$

where $K = E(X)$ is the population prevalence; g_i and h_i are the maternal and paternal alleles of the individual, respectively; α and d are the additive and dominance deviations of the penetrance function [see also Kempthorne, 1957].

Assuming Hardy-Weinberg equilibrium, James [1971] and Risch [1990a] showed that the risk ratio λ_R for a type R relative of an affected individual compared with population prevalence is

$$\lambda_R = \frac{P(\mathcal{A})}{K^2} = \frac{K_R}{K} = 1 + \frac{\operatorname{cov}(X_i, X_j)}{K^2} \tag{3}$$

where $K_R = E(X_j|X_i = 1)$ is the recurrence risk for a type R relative of an affected individual.

If we follow James's [1971] and Kempthorne's [1957] derivations closely, it can be easily shown [see Suarez et al., 1978] that, for two affected relatives,

$$P(\mathcal{A}|\pi_t) = \begin{cases} K^2 & \text{if } \pi_t = 0 \\ K^2 + \frac{1}{2}V_A & \text{if } \pi_t = \frac{1}{2} \\ K^2 + V_A + V_D & \text{if } \pi_t = 1 \end{cases} \tag{4}$$

where $V_A = 2\sum_i p_i \alpha_i^2$ is the additive genetic variance with p_i's being the allele frequencies, and $V_D = \sum_{i,j} p_i p_j d_{ij}^2$ is the dominance variance.

Substituting (4) into (1), we have

$$\Lambda = \frac{K^2 + V_A E(\pi_t|\mathcal{M}) + V_D P(\pi_t = 1|\mathcal{M})}{K^2 + V_A E(\pi_t) + V_D P(\pi_t = 1)} \tag{5}$$

In the absence of dominance, or if the dominance is negligibly small as in many cases, the above likelihood ratio for a sib pair can be written as $\Lambda = [1 + 2(\lambda_S - 1)E(\pi_t|\mathcal{M})] / \lambda_S$, where λ_S is the risk ratio to sib or parent/offspring. Thus, a genuine lod score can be computed once we know λ_S, which is often available from genetic-epidemiological studies. In general, however, since one does not know *a priori* K, V_A, and V_D, computation of the lod score is not possible without additional (and often *untestable*) assumptions. Nevertheless, it is important to point out that the lod score (5) is a monotonically increasing function of $E(\pi_t|\mathcal{M})$ and $P(\pi_t = 1|\mathcal{M})$ since $E(\pi_t)$ and $P(\pi_t = 1)$ are constants. This suggests that one may plot $E(\pi_t|\mathcal{M})$ and $P(\pi_t|\mathcal{M})$ across the genome based on data $\mathcal{M}$ observed on two or more flanking markers. In any chromosomal interval flanked by two markers which harbors no trait-causing gene, $E(\pi_t|\mathcal{M})$ would fluctuate around its mean $E(\pi_t)$, a constant depending only on the relationship between the two relatives. Thus there is no need to test for linkage in regions where $E(\pi_t|\mathcal{M})$ is low but one may wish to further examine regions with elevated values of $E(\pi_t|\mathcal{M})$ since elevated values of $E(\pi_t|\mathcal{M})$ may signal a possible location of the trait-causing gene (due to the sharing of a common gene). Of course, the elevated value $E(\pi_t|\mathcal{M})$ could also be due to chance. The empirical probability that the elevation is due to chance can be evaluated by, say, a randomization test.

Note that, if $\mathcal{M}$ denotes data observed on two flanking markers l and r, $E(\pi_t|\mathcal{M}) = \sum_{\pi_l,\pi_r} E(\pi_t|\pi_l, \pi_r)P(\pi_l, \pi_r|\mathcal{M})$.

$P(\pi_t|\pi_l, \pi_r)$ depends only on the pedigree structure and can be calculated (results not shown). $P(\pi_l, \pi_r|\mathcal{M})$ can be also calculated. Marker data on more than two flanking marker loci can be used by a telescopic relationship as suggested above, if non-interference is assumed.

We now consider *three* affected individuals. In a fashion similar to James [1971], if $\mathcal{A}$ denotes the event that three relatives 1, 2, and 3 are all affected, then, under model (2),

$$\begin{aligned} \frac{P(\mathcal{A})}{K^3} &= 1 + \frac{\sum_{i<j} \operatorname{cov}(X_i, X_j)}{K^2} + \frac{E(X_1 - K)(X_2 - K)(X_3 - K)}{K^3} \\ &= 1 + \frac{1}{K^2}\left[\sum_{i<j}\left(\frac{\phi_{ij} + \phi'_{ij}}{2}\right)V_A + \phi_{ij}\phi'_{ij}V_D\right] + \frac{1}{K^3}\left[\left(\frac{\phi_{123} + \phi'_{123}}{2}\right)V'_A\right. \\ &\quad \left. + \frac{\phi_{123}(\phi'_{12} + \phi'_{13} + \phi'_{23}) + \phi'_{123}(\phi_{12} + \phi_{13} + \phi_{23})}{2}V'_{AD} + \phi_{123}\phi'_{123}V'_D\right] \end{aligned}$$

where ϕ_{ijk} and ϕ'_{ijk} are the probabilities that individuals i, j, and k share maternal and paternal alleles IBD, respectively, $V_{A^3} = 8\sum_i p_i \alpha_i^3$, $V_{AD^2} = \sum_{i,j} p_i p_j \alpha_i d_{ij}^2$, $V_{A^2D} = 4\sum_{i,j} p_i p_j \alpha_i \alpha_j d_{ij}$, and $V_{D^3} = 2\sum_{i,j} p_i p_j d_{ij}^3$.

Hence, given π_t,

$$P(\mathcal{A}|\pi_t) = \begin{cases} K^3 + KV_A + \frac{1}{3}KV_D + \frac{1}{6}V_{A^2D} & \text{if } \pi_t = 0 \\ K^3 + 2KV_A + KV_D + \frac{1}{8}V_{A^3} + \frac{1}{2}(V_{A^2D} + V_{AD^2}) & \text{if } \pi_t = \frac{1}{2} \\ K^3 + 3K(V_A + V_D) + \frac{1}{4}V_{A^3} + \frac{3}{2}V_{A^2D} + 3V_{AD^2} + V_{D^3} & \text{if } \pi_t = 1 \end{cases} \tag{6}$$

Substituting (6) into (1), we have

$$\log_{10}\Lambda = \log_{10}\left[\frac{C_0 + C_1E(\pi_t|\mathcal{M}) + C_2P(\pi_t = 1|\mathcal{M})}{C_0 + C_1E(\pi_t) + C_2P(\pi_t = 1)}\right] \tag{7}$$

where $C_0 = K^3 + KV_A + \frac{1}{3}KV_D + \frac{1}{6}V_{A^2D}$, $C_1 = 2KV_A + \frac{4}{3}KV_D + \frac{2}{3}V_{A^2D} + \frac{1}{4}V_{A^3} + \frac{1}{4}V_{AD^2}$, and $C_2 = \frac{4}{3}KV_D + \frac{2}{3}V_{A^2D} + \frac{11}{4}V_{AD^2} + V_{D^3}$.

Again, the likelihood ratio is a monotonically increasing function of $E(\pi_t|\mathcal{M})$ and $P(\pi_t = 1|\mathcal{M})$. Since $E(\pi_t)$ and $P(\pi_t = 1)$ are now smaller as compared with the case when there are only two affected individuals, the above lod score would be higher, at the locus where the trait gene is located, than the lod score using the pairwise relationship only. This also removes the problem of dependency if the lod score is computed using all affected individuals simultaneously. Our preliminary results show that for a wide range of parameters and marker configurations, the lod scores using all affected individuals are higher than using pairwise relationships only. This is probably due to the fact that the change in posterior probability, from prior probability, that all three sibs share genes IBD is more dramatic than that in sib pairs.

TWO-LOCUS MODEL

Complex genetic traits are likely to involve multiple genes. We now consider *unlinked* two trait loci, A and B, each segregating with frequencies p_i $(i = 1, \ldots, m)$ and q_j $(j = 1, \ldots, n)$, respectively. Following James [1971] and Kempthorne [1957], a general two-locus model is now

$$\begin{array}{rll} f_{ijkl} = & K & \text{prevalence} \\ & +\alpha_i + \alpha_j & \text{additive deviation of } A_i \text{ and } A_j \\ & +d^A_{ij} & \text{dominance deviation of } A \text{ locus} \\ & +\beta_k + \beta_l & \text{additive deviation of } B_k \text{ and } B_l \\ & +d^B_{kl} & \text{dominance deviation of } B \text{ locus} \\ & +(\alpha\beta)_{ik} + (\alpha\beta)_{il} + (\alpha\beta)_{jk} + (\alpha\beta)_{jl} & \text{additive by additive interaction} \\ & +(d\beta)_{ijk} + (d\beta)_{ijl} & \text{dominance by additive interaction} \\ & +(\alpha d)_{ikl} + (\alpha d)_{jkl} & \text{additive by dominance interaction} \\ & +(dd)_{ijkl} & \text{dominance by dominance interaction} \end{array} \tag{8}$$

After tedious algebra, we can show using equation (3) that, for two affected relatives,

$$P(\mathcal{A}|\pi_A) = \begin{cases} C_0 & \text{if } \pi_A = 0 \\ C_0 + \frac{1}{2}C_1 & \text{if } \pi_A = \frac{1}{2} \\ C_0 + C_1 + C_2 & \text{if } \pi_A = 1 \end{cases} \tag{9}$$

where π_A is the proportion of alleles shared IBD at locus A by the two relatives, ϕ_{ij} and ϕ'_{ij} are the probabilities that individuals i and j share maternal and paternal alleles IBD, respectively,

$$C_0 = K^2 + \left(\frac{\phi_{ij} + \phi'_{ij}}{2}\right) V_{A_2} + \phi_{ij}\phi'_{ij} V_{D_2},$$

$$C_1 = V_{A_1} + \left(\frac{\phi_{ij} + \phi'_{ij}}{2}\right) V_{A_1 A_2} + \phi_{ij}\phi'_{ij} V_{A_1 D_2},$$

$$C_2 = V_{D_1} + \left(\frac{\phi_{ij} + \phi'_{ij}}{2}\right) V_{D_1 A_2} + \phi_{ij}\phi'_{ij} V_{D_1 D_2}$$

and V_{A_i}, V_{D_i}, $V_{A_1 A_2}$, $V_{A_i D_j}$, and $V_{D_1 D_2}$ are the additive, dominance, additive × additive, additive × dominance, and dominance × dominance variances.

Substituting (9) into (1), we have

$$\log_{10} \Lambda = \log_{10} \left[\frac{C_0 + C_1 E(\pi_A|\mathcal{M}) + C_2 P(\pi_A = 1|\mathcal{M})}{C_0 + C_1 E(\pi_A) + C_2 P(\pi_A = 1)}\right] \qquad \textbf{(10)}$$

Again, we see that the likelihood ratio is a monotonically increasing function of $E(\pi_A|\mathcal{M})$ and $P(\pi_A = 1|\mathcal{M})$. This conclusion is true if there are N unlinked trait loci. If two trait loci are *linked*, then eq. (10) still holds unless one relative is *not* a common ancestor of the other, such as half sibs, full sibs, and cousins [Cockerham, 1956]. If two relatives share a common ancestor, then C_i ($i = 0, 1$ and 2) will be affected. However, the result still holds.

For three affected individuals, results similar to (7) and (10) can be obtained, but the derivation is much more tedious (results not shown). This again suggests that one can compute $E(\pi_A|\mathcal{M})$ and $P(\pi_A = 1|\mathcal{M})$ along the genome for all pedigrees combined. Regions with elevated values of $E(\pi_A|\mathcal{M})$ need to be further examined. Plotting $P(\pi_A = 1|\mathcal{M})$ along the genome may shed light on the possible mechanism of the underlying genetic model.

DISCUSSION

Although exploratory analysis methods are popular in other fields, they have received scant attention in gene mapping. It is often the case in gene mapping that formal testing procedures are used without critical examination of the underlying assumptions, making the interpretation of results very difficult. The exploratory method proposed in this paper can avoid the problem of multiple testing, and may provide empirical perspective to develop insights into the possible mechanisms underlying the trait of interest.

Phenomenal successes of mapping simple genetic traits, coupled with availability of a dense map of highly polymorphic markers, have generated enormous interest in mapping complex traits. Due to the complexity of the traits, a genome scanning approach is often taken to localize the trait-causing genes. It seems fitting to consider exploratory methods as a screening device without formal testing, when multiple markers are typed.

ACKNOWLEDGMENTS

This research was supported by NIH grants GM52205 and HG00209 and by a Pilot/Feasibility grant from the University of Michigan Diabetes Research and Training Center.

REFERENCES

Cockerham CC (1956): Effects of linkage on the covariances between relatives. Genetics 41:138-141.

James JW (1971): Frequency in relatives for an all-or-none trait. Ann Hum Genet 35:47-48.

Kempthorne O (1957): "An Introduction to Genetic Statistics." New York: John Wiley & Sons.

Kong A, Frigge M, Irwin M, Cox N (1992): Importance sampling. I. Computing multimodel p values in linkage analysis. Am J Hum Genet 51:1413-1429.

Risch N (1990a): Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228.

Risch N (1990b): Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242-253.

Risch N (1991): A note on multiple testing procedures in linkage analysis. Am J Hum Genet 48:1058-1064.

Suarez BK, Rice J, Reich T (1978): The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87-94.

Thompson EA (1984): Interpretation of LOD scores with a set of marker loci. Genet Epidemiol 1:357-362.

Weeks DE, Lehner T, Squires-Wheeler E, et al. (1990): Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. Genet Epidemiol 7:237-243.