



ARTICLE

A likelihood-based extended admixture model of oligogenic inheritance in 'model-based' and 'model-free' analysis

Joseph D Terwilliger

Columbia University, Department of Psychiatry and Columbia Genome Center, New York State Psychiatric Institute, Department of Neuroscience, New York, USA

The admixture test of linkage heterogeneity is the most often and most successfully applied oligogenic-model linkage and/or LD analysis method. Full two-locus model linkage analysis is possible, but can be computationally intensive and difficult to interpret because of the need to specify so many indeterminate parameters. A novel, computationally efficient method is proposed for combining single locus lod scores which can allow for varying degrees of epistatic interaction. This method can be applied to two-point or multipoint (using complex-valued recombination fractions) linkage and/or linkage disequilibrium analysis to jointly test for multiple unlinked disease loci. Unlike the traditional admixture test, this algorithm permits joint analysis of multiple disease loci with different modes of inheritance for each, and can be applied to 'model-free' analysis as well through the use of 'pseudomarkers'. Software is available for computation of the various likelihood ratio tests described, for comparison of a variety of possible hypotheses regarding locus homogeneity, locus heterogeneity, and epistasis. *European Journal of Human Genetics* (2000) 8, 399–406.

Keywords: oligogenic inheritance; admixture test; linkage heterogeneity; two-locus models; linkage disequilibrium analysis; complex disease

Introduction

Traditional lod score analysis of monogenic disease has been plagued by sensitivity to locus heterogeneity. Such diseases as retinitis pigmentosa¹ and nonsyndromic hereditary deafness² can be caused deterministically by genotypes of any of a large number of different loci, often with different modes of inheritance for each.³ To this end, methods have been developed for multi-locus linkage analysis which stratify families according to which disease gene is segregating in each family, assuming one and only one locus per family has risk genotypes segregating. In the case of simple diseases with a deterministic monogenic etiology, this can be a very powerful approach. However, for more complex disorders, there is likely to be a substantially larger number of loci with genotypes that influence disease phenotypes, with the

marginal effects of any single genotype at any single locus having minimal effect in the population as a whole.⁴ It may be that only in combination with specific constellations of environmental factors and risk genotypes of other loci will there be an influence of a given locus on the phenotype.^{3,5–8} In such cases, the admixture model described above may not capture any possible existing evidence of linkage or LD very effectively.⁹

Parametric two-locus model analysis of linkage and linkage disequilibrium (LD) jointly has been applied successfully in a study of multiple sclerosis on a set of multigenerational pedigrees from Finland,¹⁰ in which a full set of two-locus genotype/phenotype relationships (ie penetrances) was fully specified and the likelihood of different hypotheses about linkage and LD jointly were computed. This approach is practical and efficient only when there is a specific and well characterized parametric penetrance model of hypothesized epistasis¹¹ as this approach is computationally very intensive. In this manuscript, an extension of the admixture test¹² is proposed, which captures much of the information about

Correspondence: Joseph D Terwilliger PhD, Columbia University, 60 Haven Avenue No. 15-C, New York, NY 10032, USA.
Fax: +1 212 304 5515; E-mail: jdt3@columbia.edu
Received 28 July 1999; revised 17 December 1999; accepted 4 January 2000

epistatic interactions when they exist, with greatly reduced computational and theoretical burdens. The proposed method can be easily extended to joint analysis of numerous mutually unlinked disease-predisposing loci under a wide variety of models of epistasis and heterogeneity alike. In the case of model-free analysis, it will further be shown to be equivalent to a full multi-locus pseudomarker analysis.^{9,13}

Likelihood model for linkage and/or LD analysis

In a traditional single-locus analysis (where single locus here refers to the assumed mode of inheritance of the disease), one is interested in computing the joint probability of the observed marker locus genotypes, \mathbf{G}_M , and the observed trait phenotypes, \mathbf{Ph} , for all individuals in a dataset, as a function of some hypotheses about linkage and/or LD. $P(\mathbf{G}_M, \mathbf{Ph})$ is proportional to the likelihood, such that the lod score

$$Z = \log_{10} \frac{P(\mathbf{G}_M, \mathbf{Ph})}{P(\mathbf{G}_M)P(\mathbf{Ph})} = \log_{10} \frac{\max_{\theta} L(\theta)}{L(\theta = 0.5)}$$

as a function of the recombination fraction, θ , and a linkage disequilibrium χ^2 statistic

$$\Lambda = 2 \ln \frac{P(\mathbf{G}_M, \mathbf{Ph})}{P(\mathbf{G}_M)P(\mathbf{Ph})} = 2 \ln \frac{\max_{\delta} L(\delta)}{L(\delta = 0)}$$

as a function of the vector of linkage disequilibrium coefficients, δ , correlating disease and marker loci.^{3,13,14}

In a model-based linkage analysis, one decomposes the likelihood as

$$P(\mathbf{G}_M, \mathbf{Ph}) = P(\mathbf{Ph}|\mathbf{G}_M)P(\mathbf{G}_M) = P(\mathbf{G}_M) \sum_{\mathbf{G}_D} P(\mathbf{Ph}|\mathbf{G}_D)P(\mathbf{G}_D|\mathbf{G}_M),$$

taking the sum over all possible vectors of disease locus genotypes, \mathbf{G}_D , for all individuals in the dataset. $P(\mathbf{G}_M)$ is a function of the assumed model for the marker-locus genotype frequencies in the population, $P(\mathbf{Ph}|\mathbf{G}_D)$ is a function of the assumed penetrance model (or means and variances in the case of a quantitative trait), and $P(\mathbf{G}_D|\mathbf{G}_M)$ is a function of linkage and/or LD between disease and marker loci. For more precise details of the parameterization of the likelihood as a function of these probabilities, see Göring and Terwilliger (2000C,D).^{13,14}

Admixture test, two disease loci, one marker locus

In the simplest model of locus heterogeneity, some individuals are assumed to be affected with the disease because of genotypes of one particular gene, and other individuals are affected for independent reasons. In practice, one can formally describe this situation in terms of the genotype-phenotype relationships (ie penetrances) at a single disease locus. If we assume a dominant mode of inheritance,

$$P(\text{Affected}|\text{DD}) = P(\text{Affected}|\text{D+}) = kf; P(\text{Affected}|\text{++}) = f,$$

where f is the probability of being affected for some reason other than the disease locus under study, and

$$k = \frac{P(\text{Affected}|\text{DD or D+})}{P(\text{Affected}|\text{++})}$$

is the relative risk of being affected given the presence of at least one D allele at this disease locus. Heterogeneity due to a mixture of independent etiological factors is implicit in the analysis. It is further assumed that the causes of disease other than risk genotypes of the disease gene being modeled are not familially correlated.

If the penetrances of genotypes of a single locus are substantial, and the frequencies of the risk genotypes are small, one can allow for heterogeneity between multiple familial causes of disease by explicitly modeling a mixture of family types – one family type (with proportion α) in which the disease is caused by genotypes of one disease gene, and a second family type (with proportion $1 - \alpha$) where the disease is caused by some completely different familial risk factor, as is the case in many forms of retinitis pigmentosa.³ The likelihood can then be written as

$$L \propto P(\mathbf{Ph}, \mathbf{G}_M) = \alpha \sum_{\mathbf{G}_D} P(\mathbf{Ph}|\mathbf{G}_D)P(\mathbf{G}_D|\mathbf{G}_M)P(\mathbf{G}_M) + (1-\alpha) \sum_{\mathbf{G}_D} P(\mathbf{Ph}|\mathbf{G}_D)P(\mathbf{G}_D)P(\mathbf{G}_M).$$

In the latter term it is implicit that there is some familially transmitted risk factor that is independent of the marker locus in a proportion $1 - \alpha$ of families, be it genetic or environmental.

If we generalize this to two disease genes, either of which may be potentially linked to a given marker locus, we can compute the likelihood by partitioning over all possible genotypes of both disease loci jointly as

$$P(\mathbf{G}_M, \mathbf{Ph}) = P(\mathbf{G}_M) \sum_{\mathbf{G}_{D_1}, \mathbf{G}_{D_2}} P(\mathbf{Ph}|\mathbf{G}_{D_1}, \mathbf{G}_{D_2})P(\mathbf{G}_{D_1}, \mathbf{G}_{D_2}|\mathbf{G}_M).$$

If affected individuals in any single family have risk genotypes at either D_1 or D_2 , but not both (note that this is a simplifying approximation, since it is assumed that the risk genotypes of either locus are sufficiently rare that only one is segregating per family), then this can be rewritten as

$$P(\mathbf{G}_M) \left(P(D_1 \text{ segregating}) \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph}|\mathbf{G}_{D_1})P(\mathbf{G}_{D_1}|\mathbf{G}_M) + P(D_2 \text{ segregating}) \sum_{\mathbf{G}_{D_2}} P(\mathbf{Ph}|\mathbf{G}_{D_2})P(\mathbf{G}_{D_2}|\mathbf{G}_M) \right),$$

which is the standard heterogeneity likelihood formation in the admixture test.^{12,15} If $\alpha_1 = P(D_1 \text{ segregating})$ and

$$L_1(\theta_1) = \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph}|\mathbf{G}_{D_1})P(\mathbf{G}_{D_1}|\mathbf{G}_M),$$

then setting $\alpha_2 = 1 - \alpha_1$, reduces the equation to

$$P(\mathbf{G}_M, \mathbf{Ph}) \propto L(\theta_1, \theta_2, \alpha_1) = \alpha_1 L_1(\theta_1) + (1 - \alpha_1) L_2(\theta_2)$$

Note that the marker locus could be linked to either D_1 , D_2 , or both in this formulation.

Various tests of linkage and/or heterogeneity can be conducted by comparing the likelihoods under different hypotheses, as enumerated in Table 1a. The first row gives the likelihood when the marker loci are linked to both D_1 and D_2 ,

as was the case with X-linked retinitis pigmentosa.¹⁶ The next three lines enumerate the likelihood of the admissible hypotheses when the marker locus is unlinked to D_2 , and the last three rows assume the marker locus is unlinked to D_1 . If we assume the same mode of inheritance parameters at both D_1 and D_2 , then, without loss of generality, we can focus on tests of linkage and/or LD between the marker and D_1 , as outlined in Table 1b.

This is the simplest two-locus model of disease that can be incorporated in linkage and/or LD analysis, representing the most extreme statistical interaction possible. This may sound counterintuitive, so consider what the assumptions imply, in the context of a full two-locus model. Define the (2-locus) genotype-phenotype relationship and the prior (2-locus) genotype probabilities for random individuals in the population. In the case of the heterogeneity model above, assuming the same mode of inheritance for each of the two loci, $P(D$ allele at either locus) = p , and (assuming a dominant model as above), $P(\text{affected}|D_1D_1) = P(\text{affected}|D_1+)$ = $P(\text{affected}|D_2D_2) = P(\text{affected}|D_2+)$ = kf , and $P(\text{affected}|++)$ = $P(\text{affected}|+2+)$ = f . Because there is an implicit assumption that the D alleles are individually very rare, heterogeneity analysis assumes that in any given individual, only risk alleles at either D_1 or D_2 can be present, consistent with the penetrance and genotype frequency matrices shown in Table 2a (note that α is a function of the difference in frequencies of the disease-predisposing alleles of the two loci, not the genotype-phenotype relationship, as illustrated in Table 2a). The posterior genotype probabilities for a single

affected individual are shown in Table 2b.⁹ Because the disease-predisposing alleles at either locus are assumed to be extremely rare, other affected individuals in the same pedigree as a proband are inferred (a second level of hand-waving approximation) to have either the same disease-predisposing allele at the same locus, or none at all. Under these restrictive simplifying assumptions, the heterogeneity analysis is computationally efficient and can lead to increased power to find linkage with heterogeneous disorders. Note that the effects of locus heterogeneity on the recombination fraction estimates are analogous to those due to errors in the prediction of underlying disease-locus genotypes conditional on observed phenotypes.^{17,18}

Admixture model: two disease loci, two unlinked marker loci

One can generalize the admixture model to include markers linked to each of the two disease loci. If G_M consists of a set of two unlinked marker loci (M_1 and M_2), it is possible that D_1 is linked to M_1 and D_2 is linked to M_2 . In this case, the likelihoods shown in Table 1a and the tests outlined in Table 1b are directly applicable, where M would represent the set (M_1, M_2), and the statistics outlined would refer to the model where D_1 and D_2 are assumed to be linked to independent markers in the set M . Multiple test corrections would be indicated when many marker loci are tested individually, as in a genome scan experiment, however.^{9,19,20} Nevertheless, it is clear that one can extend these admixture models to any number of marker loci and any number of

Table 1a Admixture test hypotheses outlined (one locus maximum per family)

Locus D1		Locus D2		Likelihood
Segregating?	Linked?	Segregating?	Linked?	
Yes	Yes	Yes	Yes	$L(\alpha_1, \theta_1, \theta_2)$ $\alpha_1 L_1(\theta_1) + (1-\alpha_1)L_2(\theta_2)$
Yes	Yes	Yes	No	$L(\alpha_1, \theta_1, \theta_2 = 1/2)$ $\alpha_1 L_1(\theta_1) + (1-\alpha_1)L_2(\theta_2 = 0.5)$
Yes	Yes	No	No	$L(1, \theta_1)$ $L_1(\theta_1); (\alpha_1 = 1)$
Yes	No	No	No	$L(1, \theta_1 = 1/2)$ $L_1(\theta_1 = 0.5)$
Yes	No	Yes	Yes	$L(\alpha_1, \theta_1 = 1/2, \theta_2)$ $\alpha_1 L_1(\theta_1 = 0.5) + (1-\alpha_1)L_2(\theta_2)$
No		Yes	Yes	$L(0, \theta_2)$ $L_2(\theta_2); (\alpha_1 = 0)$
No		Yes	No	$L(0, \theta_2 = 1/2)$ $L_2(\theta_2 = 0.5)$

Table 1b Possible likelihood ratio tests of linkage between marker and trait locus 1

	LRT statistic	Approx. distribution
Linkage and homogeneity	$\Lambda = 2 \ln \frac{L(\hat{\alpha}_1 = 1, \hat{\theta}_1)}{L(\hat{\alpha}_1 = 1, \hat{\theta}_1 = 1/2)}$	$0.5\chi^2_{(1)}$
Linkage allowing for heterogeneity D2 unlinked to M	$\Lambda = 2 \ln \frac{L(\hat{\alpha}_1, \hat{\theta}_1, \hat{\theta}_2 = 1/2)}{L(\hat{\alpha}_1 = 1, \hat{\theta}_1 = 1/2, \hat{\theta}_2 = 1/2)}$	$0.5\chi^2_{(1)} + 0.25\chi^2_{(2)}$
Linkage allowing for heterogeneity D2 linked to M	$\Lambda = 2 \ln \frac{L(\hat{\alpha}_1, \hat{\theta}_1, \hat{\theta}_2)}{L(\hat{\alpha}_1, \hat{\theta}_1 = 1/2, \hat{\theta}_2)}$	$0.5\chi^2_{(1)}$

disease loci, and formulate likelihood ratio test statistics through generalization of this model. See Ott²¹ for an exhaustive enumeration of statistical tests based on the likelihood,

$$P(\mathbf{Ph}, \mathbf{G}_M) \propto L = \sum_i \alpha_i L_i(\theta_i).$$

Common alleles and complex traits: oligogenic models

In the simple admixture test, the assumption was made that alleles at each disease locus are rare and sufficiently penetrant that an affected individual was likely to carry disease-predisposing alleles at one and only one of the disease loci, which would imply that the same allele causes the disease in affected relatives as well. However, for common multifactorial diseases, this may be an irrational and unjustified assumption.²²⁻²⁵ The simple admixture test of locus homogeneity still has a meaningful application, but the subdivision of families into two types – the first carrying risk alleles of one locus and the second carrying risk alleles of a second locus – does not. When risk alleles are common, the assumption that families segregate disease-predisposing alleles at one of two loci, but not both, is the extreme statistical interaction one can hypothesize – equivalent to the ‘XOR’ model of Lucek and Ott²⁶ – a mode of inheritance that may be biologically inappropriate for most traits.

If there is epistasis, genotypes of one disease locus may differentially influence the probability of an observed phenotype conditional on the genotypes of a second locus. The most extreme case would be a model where only individuals with risk genotypes at both loci can be affected. Such models imply that families in which risk alleles are segregating at D_1 would also have risk alleles segregating at D_2 . If there are other potential causes of disease, but D_1 and D_2 can only affect the disease when risk genotypes of both loci are present, then those individuals without risk genotypes of D_1 would not be likely to have risk genotypes at D_2 either. If we

Table 2a Two-locus penetrance and genotype frequency model assumed in heterogeneity admixture test (dominant model with penetrance ratio k (see text))

	Penetrances			Genotype frequencies			
	D_1D_1	D_1+	$++$	D_1D_1	D_1+	$++$	
D_2D_2	?	?	kf	D_2D_2	0	0	$(1-\alpha)p^2$
D_2+	?	?	kf	D_2+	0	0	$(1-\alpha)2p(1-p)$
$++$	kf	kf	f	$++$	αp^2	$\alpha 2p(1-p)$	$(1-p)^2$

Table 2b Probabilities of each two-locus genotype conditional on affection status and the mode of inheritance described in Table 2a, where $\omega = k + (1-k)(1-p)^2$

	Genotype probabilities for affected proband		
	D_1D_1	D_1+	$++$
D_2D_2	0	0	$(1-\alpha)kp^2/\omega$
D_2+	0	0	$2(1-\alpha)kp(1-p)/\omega$
$++$	$\alpha kp^2/\omega$	$2\alpha kp(1-p)/\omega$	$(1-p)^2/\omega$

restrict our set, M , of marker loci to two markers, M_1 and M_2 , where M_1 is linked to D_1 , and M_2 is linked to D_2 , with D_1 and D_2 unlinked to each other, then we can write the likelihood as

$$P(\mathbf{Ph}, \mathbf{G}_M) = P(\mathbf{Ph}, \mathbf{G}_{M_1}, \mathbf{G}_{M_2}) \\ = \sum_{\mathbf{G}_{D_1}} \sum_{\mathbf{G}_{D_2}} P(\mathbf{Ph}|\mathbf{G}_{D_1}, \mathbf{G}_{D_2}) P(\mathbf{G}_{D_1}, \mathbf{G}_{D_2}|\mathbf{G}_{M_1}, \mathbf{G}_{M_2}) P(\mathbf{G}_{M_1}) P(\mathbf{G}_{M_2}).$$

Let us assume a multiplicative dominant mode of inheritance such that $P(\mathbf{Ph}|\mathbf{G}_{D_1}, \mathbf{G}_{D_2}) = \Gamma_{D_1} \Gamma_{D_2}$, where $\Gamma_{DD} = \Gamma_{D+} = \kappa \Gamma_{++}$; and $P(\mathbf{G}_{D_i}|\mathbf{G}_{M_i})$ is a function of linkage and/or LD between loci D_i and M_i . Without loss of generality, we can set $\Gamma_{++} = cf$, where $f = P(\text{Affected}|++)$ as in the single locus dominant model described above, such that $\Gamma_{DD} = \Gamma_{D+} = \kappa cf$, where c is some constant of proportionality. The full two-locus penetrance matrix in Table 3 would obtain. Note that under this multiplicative model, $P(\text{Affected}|++ \text{ at locus 1}) P(\text{Affected}|++ \text{ at locus 2}) = c^2(f)(f) = c^2 P(\text{Affected}|++ \text{ at locus 1}) P(\text{Affected}|++ \text{ at locus 2})$.

If we define the single locus marginal penetrance model such that $P(\text{Affected}|DD) = P(\text{Affected}|D+) = \kappa f$, then $P(\text{Affected}|D_1D_1, D_2+) = c^2(\kappa f)(\kappa f) = c^2 P(\text{Affected}|DD \text{ at locus 1}) P(\text{Affected}|D+ \text{ at locus 2})$.

In each such two-locus penetrance, the identical multiplicative factor c^2 occurs, which can be factored out of the likelihood as a constant of proportionality (see below).

We can rewrite the likelihood from above as follows, where $P(\mathbf{Ph}|\mathbf{G}_{D_i})$ is a function of the single locus marginal penetrances, and $C(c^2)$ is the constant of proportionality which disappears in the likelihood ratio:

$$L \propto P(\mathbf{Ph}, \mathbf{G}_{M_1}, \mathbf{G}_{M_2}) = C(c^2) \sum_{\mathbf{G}_{D_1}} \sum_{\mathbf{G}_{D_2}} P(\mathbf{Ph}|\mathbf{G}_{D_1}) \\ P(\mathbf{Ph}|\mathbf{G}_{D_2}) P(\mathbf{G}_{D_1}, \mathbf{G}_{D_2}|\mathbf{G}_{M_1}, \mathbf{G}_{M_2}) P(\mathbf{G}_{M_1}) P(\mathbf{G}_{M_2}).$$

Thus, the likelihood can be computed as a function of the single locus marginal penetrance models. Note that it is not necessary for the disease loci to have the same marginal mode of inheritance. In this formulation, unaffected individuals will have somewhat different genotype-phenotype relationships than the pure multiplicative model would dictate, but for complex traits, the information unaffected individuals provide about the underlying genotypes, \mathbf{G}_D , is minimal, since κ and f are both typically small for multifactorial phenotypes. Allowance for a mixture of family types will alleviate this to some extent, and will allow for a variety of models that are not strictly multiplicative.

Table 3 Multiplicative penetrance model in which $P(\text{Affected}|D_1D_1, D_2D_2) = \Gamma_{DD}\Gamma_{DD}$; $\Gamma_{DD}\Gamma_{D+} = \kappa \Gamma_{++}$, and $\Gamma_{++} = cP(\text{Affected}|DD \text{ in a single locus model}) = cf$

	Penetrances		
	D_1D_1	D_1+	$++$
D_2D_2	$\kappa^2(cf)^2$	$\kappa^2(cf)^2$	$\kappa(cf)^2$
D_2+	$\kappa^2(cf)^2$	$\kappa^2(cf)^2$	$\kappa(cf)^2$
$++$	$\kappa(cf)^2$	$\kappa(cf)^2$	$(cf)^2$

Common alleles and complex traits: extended admixture models

In the formulation of the two-locus likelihood in terms of the marginal penetrances, the term $P(\mathbf{G}_{D_1}, \mathbf{G}_{D_2} | \mathbf{G}_{M_1}, \mathbf{G}_{M_2})$ is not straightforward to compute. However, one can approximate a two-locus analysis by an extension of the admixture test in which some proportion of families would segregate both D_1 and D_2 (β_{12}), some proportion would segregate D_1 but not D_2 (β_1), some proportion would segregate D_2 but not D_1 (β_2) and some proportion would segregate neither D_1 nor D_2 ($\beta_0 = 1 - \beta_{12} - \beta_1 - \beta_2$). It is proposed that the likelihood be computed by partitioning over the possible family types, weighted by the β_i , which is valid when the marker loci (M_1 and M_2) are unlinked to each other, as $P(\mathbf{Ph}, \mathbf{G}_{M_1}, \mathbf{G}_{M_2}) =$

$$C(c^2) \left\{ \begin{aligned} &\beta_{12} \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph} | \mathbf{G}_{D_1}) P(\mathbf{G}_{D_1}, \mathbf{G}_{M_1}) \sum_{\mathbf{G}_{D_2}} P(\mathbf{Ph} | \mathbf{G}_{D_2}) P(\mathbf{G}_{D_2}, \mathbf{G}_{M_2}) \\ &+ \beta_1 \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph} | \mathbf{G}_{D_1}) P(\mathbf{G}_{D_1}, \mathbf{G}_{M_1}) \sum_{\mathbf{G}_{D_2}} P(\mathbf{Ph} | \mathbf{G}_{D_2}) P(\mathbf{G}_{D_2}) P(\mathbf{G}_{M_2}) \\ &+ \beta_2 \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph} | \mathbf{G}_{D_1}) P(\mathbf{G}_{D_1}) P(\mathbf{G}_{M_1}) \sum_{\mathbf{G}_{D_2}} P(\mathbf{Ph} | \mathbf{G}_{D_2}) P(\mathbf{G}_{D_2}, \mathbf{G}_{M_2}) \\ &+ \beta_0 \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph} | \mathbf{G}_{D_1}) P(\mathbf{G}_{D_1}) P(\mathbf{G}_{M_1}) \sum_{\mathbf{G}_{D_2}} P(\mathbf{Ph} | \mathbf{G}_{D_2}) P(\mathbf{G}_{D_2}) P(\mathbf{G}_{M_2}) \end{aligned} \right\}$$

Since the likelihood,

$$L_1(\theta_1) \propto \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph} | \mathbf{G}_{D_1}) P(\mathbf{G}_{D_1}, \mathbf{G}_{M_1}), \text{ and } L_1(\theta_1 = 1/2) \propto \sum_{\mathbf{G}_{D_1}} P(\mathbf{Ph} | \mathbf{G}_{D_1}) P(\mathbf{G}_{D_1}) P(\mathbf{G}_{M_1}),$$

the overall likelihood, as a function of the β_i , reduces to

$$L(\beta_{12}, \beta_1, \beta_2, \theta_1, \theta_2) = \beta_{12} L_1(\theta_2) L_2(\theta_2) + \beta_1 L_1(\theta_1) L_2(\theta_2 = 1/2) + \beta_2 L_1(\theta_1 = 1/2) L_2(\theta_2) + (1 - \beta_{12} - \beta_1 - \beta_2) L_1(\theta_1 = 1/2) L_2(\theta_2 = 1/2),$$

a function of the single-locus model homogeneity likelihoods, $L_1(\theta_1)$ and $L_2(\theta_2)$, and the β_i . Varying the β_i can cover a wide range of possible two-locus models based on the fixed marginal mode of inheritance assumptions used in the single locus likelihood computations. If one sets $\beta_{12} = 0$ and $\beta_0 (= 1 - \beta_1 - \beta_2 - \beta_{12}) = 0$, the likelihood is proportional to that computed in the conventional admixture test of locus homogeneity. Furthermore, if one hypothesized a model with two disease loci, each linked to one of the marker loci, and a third class of families not linked to either of the loci, that could be achieved by setting $\beta_{12} = 0$ alone (see Table 4), and letting β_1 , and β_2 vary freely (see Table 4).

One can formulate the likelihood in terms of the marginal heterogeneity parameters α_1 and α_2 , which were defined above in context of the conventional admixture test, by removing the restriction that each family can have no more than one of the two disease loci segregating. If the segregation of the two loci were independent, conditional on the ascertainment, then $\beta_{12} = \alpha_1 \alpha_2$, $\beta_1 = \alpha_1 (1 - \alpha_2)$, etc, as outlined in Table 4 (independent segregation model). Finally, one can add an interactive parameter ξ , as shown in the last column of Table 4 (general model) which allows the β_i to vary in an unconstrained manner. For example, if parameters were chosen such that $\alpha_1 = \alpha_2$, and $\xi = \alpha_1 (1 - \alpha_2)$ then there would be only two classes of families – those with disease-predisposing alleles at both D_1 and D_2 segregating, and those with disease-predisposing alleles at neither D_1 nor D_2 segregating – very strong epistasis. If the presence of risk alleles at one locus has no influence on the probability of risk alleles at the second locus, a pure heterogeneity model results (ie $\xi = 0$). Note that if the risk alleles are common, the probability of both loci segregating in the same family cannot reasonably be set to zero, as they are in the conventional admixture test.

Table 4 Set of possible hypotheses that can be tested with the 2-locus extended admixture test

Locus 1 Linked?	Locus 2 Linked?	Other Locus?	df	Hypothesis	β_{12}	β_1	β_2	β_0
Null hypothesis								
No	No	No	0	No linkage	0	0	0	1
Single locus linkage with homogeneity								
Yes	No	No	1	D_1 : Homogeneity	0	1	0	0
No	Yes	No	1	D_2 : Homogeneity	0	0	1	0
Single locus linkage with admixture								
Yes	No	Yes	2	D_1 : Heterogeneity	0	α_1	0	$1 - \alpha_1$
No	Yes	Yes	2	D_2 : Heterogeneity	0	0	α_2	$1 - \alpha_2$
Two-locus model linkage with extended admixture								
Yes	Yes	No	3	D_1 and D_2 Admixture	0	α_1	$1 - \alpha_1$	0
Yes	Yes	Yes	4	D_1 and D_2 Admixture and Heterogeneity	0	α_1	α_2	$1 - \alpha_1 - \alpha_2$
Yes	Yes	Yes	4	D_1 and D_2 Independent	$\alpha_1 \alpha_2$	$\alpha_1 (1 - \alpha_2)$	$(1 - \alpha_1) \alpha_2$	$(1 - \alpha_1) (1 - \alpha_2)$
Yes	Yes	Yes	5	D_1 and D_2 : General Model	$\alpha_1 \alpha_2 + \xi$	$\alpha_1 (1 - \alpha_2) - \xi$	$(1 - \alpha_1) \alpha_2 - \xi$	$(1 - \alpha_1) (1 - \alpha_2) + \xi$

Whilst a wide variety of intermediate levels of epistasis are allowed for in this model, it must be remembered that they represent an approximation to a complete two-locus parametric analysis. It can capture most, if not all, of the correlative information which exists, when (as is typically the case) one cannot accurately specify the mode of inheritance with sufficient accuracy to even conceptualize an appropriate complete two-locus penetrance matrix. Furthermore, the computational time is dramatically reduced from a full two-locus analysis, and can be performed in a matter of seconds after the initial two-point likelihoods have been computed. Table 4 outlines the statistical framework for a range of hypothesis testing.

Extension of this approach to more than two trait loci is immediate and straightforward, allowing for as many *unlinked* disease loci as one desires. For example, in the case of three disease loci, the likelihood could be computed as

$$\begin{aligned}
 L(\beta, \theta) = & \beta_{123}L_1(\theta_1)L_2(\theta_2)L_3(\theta_3) + \beta_{12}L_1(\theta_1)L_2(\theta_2)L_3(\theta_3 = 1/2) \\
 & + \beta_{13}L_1(\theta_1)L_2(\theta_2 = 1/2)L_3(\theta_3) + \beta_{23}L_1(\theta_1 = 1/2)L_2(\theta_2)L_3(\theta_3) \\
 & + \beta_1L_1(\theta_1)L_2(\theta_2 = 1/2)L_3(\theta_3 = 1/2) + \beta_2L_1(\theta_1 = 1/2)L_2(\theta_2)L_3(\theta_3 = 1/2) \\
 & + \beta_3L_1(\theta_1 = 1/2)L_2(\theta_2 = 1/2)L_3(\theta_3) + \beta_0L_1(\theta_1 = 1/2)L_2(\theta_2 = 1/2)L_3(\theta_3 = 1/2).
 \end{aligned}$$

Extension to more trait loci in a single analysis can be made, by direct and straightforward induction, in which all possible pairs of nested hypotheses can be compared using likelihood ratio tests, computing profile likelihoods over the nuisance parameters where appropriate¹⁴. As the number of loci increases, the number of admissible models of epistatic interaction correspondingly increases, such that some *a priori* formulation of sets of hypotheses based on biological pathways may be advised to target the analyses towards more likely hypotheses, in order to reduce the deleterious effects of multiple testing.

Unlike traditional single locus admixture models, one can use different mode of inheritance assumptions for each of the loci being analyzed, since the multiplicative penetrance model assumption does not restrict anything about the underlying single locus penetrance models, and in this framework, one could jointly analyze, say, a dominant locus on chromosome 5 and a recessive locus on chromosome 6, together with a sex-linked susceptibility locus, so long as the likelihood under the several models are combined multiplicatively under null and alternative hypotheses alike. These analyses can be performed using either two-point or multipoint likelihoods, so long as appropriate corrections for multiple testing are taken into account.

Complex-valued recombination fractions and complex traits

Let us extend this model to allow for errors in the mode-of-inheritance assumptions in an analysis. If one knows the genotypes of some trait locus unambiguously, then the likelihood $L \propto P(\mathbf{G}_M, \mathbf{G}_D) = P(\mathbf{G}_D | \mathbf{G}_M)P(\mathbf{G}_M)$.

However, when the trait locus genotypes cannot be accurately determined, one computes a weighted average of the likelihoods as a function of some assumed mode of inheritance parameters as

$$L \propto P(\mathbf{G}_M, \mathbf{Ph}) = \sum_{\mathbf{G}_D} P(\mathbf{Ph} | \mathbf{G}_D) P(\mathbf{G}_D, \mathbf{G}_M).$$

When the model is inaccurate, the weights will be misspecified and there will be substantial probabilities for misclassifying genotypes of the trait locus. One way to deal with this, as described by Göring and Terwilliger¹⁷ would be to explicitly allow for such errors in the analysis.

If we assume that misclassifying the genotype of the trait locus causes a recombinant to be misclassified as a non-recombinant and vice versa, then one can assume that there is a mixture of meiotic types, where the probability ε , the recombination status is misclassified. The probability of an observed recombinant would then be $\varepsilon(1-\theta) + (1-\varepsilon)\theta$.¹⁷ Note that this is analogous to the admixture model for heterogeneity among families in which $\alpha = P(\text{disease gene is segregating in the family})$, and the likelihood of a family is $\alpha P(\text{family} | \text{linkage}) + (1-\alpha)P(\text{family} | \text{no linkage})$. The only difference is that the heterogeneity is within families, across meioses, and is solely a function of how well the assumed disease model fits the observed segregation pattern. One should note that ε is analogous to a recombination fraction in a direction orthogonal to the chromosome, ie it estimates the frequency with which assumed trait locus genotypes do not cosegregate with the chromosomal position at which the trait locus resides. It is convenient to think of the recombination fractions with misclassification as having two components, the 'real' recombination fraction, θ , between the actual genotypes of the trait and marker loci, and the 'imaginary' recombination fraction, ε , between the actual and observed trait locus genotypes, which can be expressed as a complex recombination fraction $\Theta = \theta + \varepsilon i$, to demonstrate the orthogonality of the two components, and their interpretations as probabilities of 'real' and 'imaginary' recombination.

Göring and Terwilliger¹⁷ proposed using such a model to minimize the effects of inaccurate mode of inheritance assumptions in multipoint parametric linkage analysis, and it has been demonstrated¹³ that lod scores of the form

$$Z(x, \varepsilon) = \frac{\max_{\varepsilon \in (0, 0.5)} L(x, \varepsilon)}{L(x, \varepsilon = 0.5)}$$

have similar genome-wide behavior as model-free statistical tests. As an interesting note, it was shown⁹ that in the context of this model, it is always better to assume a mode of inheritance model that is too strong (ie in which genotypes predict phenotypes more deterministically than they actually do) in terms of both power and accuracy of the estimates of location of the underlying disease locus. In conjunction with the general model of intrafamilial heterogeneity and epistasis

presented here, even more robustness to errors is likely, though there may be enormous support intervals for gene location as is unavoidable in analysis of complex traits with many (often confounded) parameters in the model.

Model-free extended admixture analysis of multiple trait loci

In model-free analysis, one computes the likelihood $L \propto P(\mathbf{Ph}, \mathbf{G}_M) = P(\mathbf{G}_M | \mathbf{Ph}) P(\mathbf{Ph})$, where, in the likelihood ratio, $P(\mathbf{Ph})$ cancels out, as the ascertainment scheme makes the restriction that one exclusively samples pedigrees (or individuals) of a single structure (ie affected sib pairs, triads, or singletons). In this case, one can directly estimate all possible multinomial proportions $P(\mathbf{G}_M | \mathbf{Ph})$.¹³ In the case of two-locus analysis, one can compute the likelihood of two unlinked markers (or sets of markers in multipoint analysis) jointly as $P(\mathbf{G}_{M_1}, \mathbf{G}_{M_2} | \mathbf{Ph})$. One can likewise stratify this by the probability that disease loci linked to either or both marker loci are segregating in a given pedigree as $P(\mathbf{G}_{M_1}, \mathbf{G}_{M_2} | \mathbf{Ph}) = \beta_{12} P(\mathbf{G}_{M_1}, \mathbf{G}_{M_2} | \mathbf{Ph}) + \beta_1 P(\mathbf{G}_{M_1} | \mathbf{Ph}) P(\mathbf{G}_{M_2}) + \beta_2 P(\mathbf{G}_{M_2} | \mathbf{Ph}) P(\mathbf{G}_{M_1}) + \beta_0 P(\mathbf{G}_{M_1}) P(\mathbf{G}_{M_2})$.

As in model-based linkage analysis, one can consider all possible models outlined in Table 4, though the number of df depends on the number of parameters involved in $P(\mathbf{G}_{M_i} | \mathbf{Ph})$ and $P(\mathbf{G}_{M_i})$. Such analysis is a generalization of the common technique of stratifying a dataset based on whether there is linkage (or association) with one locus, in an attempt to identify additional loci for some disorder.²⁷

In many cases, one has not ascertained a single data structure, yet 'model-free' analysis is desired because of uncertainty about the true mode of inheritance. In most of these cases, it is assumed that there will be multiple disease-predisposing loci as well, so it would be advisable to allow for their existence. One technique for robust combination of data structures in a single 'model-free' analysis would be to convert the observed disease phenotypes, \mathbf{Ph} , into 'pseudo-marker' genotypes, \mathbf{G}_p , followed by likelihood-based linkage and/or LD analysis.^{9,13,28} Pseudomarker analysis involves computing pedigree likelihoods under the assumption that all meioses connecting affected individuals in a pedigree are informative for linkage with equal probability. It has been demonstrated that this is statistically equivalent to the affected sib-pair mean test on sibpair data structures,²⁹ and to affected relative pair methods on more distantly related pairs of relatives,^{13,28,30} an observation which has been generalized to multipoint analysis when complex recombination fractions are admitted, as described above.

The extended admixture test is described in Table 4 leads to an exact computation of the complete multi-locus pseudomarker likelihood. The approximations which were required in 'model-based' likelihood analysis are not necessary in the pseudomarker analysis, as the latter deterministically assigns genotypes to all disease loci in all individuals. The implicit inference is that all disease loci have the same fully informative genotype for all individuals, though the imagi-

nary components, ε_i , of the complex recombination fraction¹⁷ will be estimated independently for each of the putative disease loci. In this sense, a full multilocus 'model-free' analysis with pseudomarkers would be equivalent to the extended admixture test with pseudomarkers.⁹

Discussion

In this paper, a simple parameterization of the pedigree likelihood, $P(\mathbf{G}_M, \mathbf{Ph})$ in the presence of multiple disease-predisposing loci is proposed, by extension of the conventional A-test of linkage homogeneity.¹² Through this generalization, we can perform approximate oligogenic 'model-based' analysis, and exact oligogenic 'model-free' analysis with pseudomarkers. It has earlier been demonstrated that power is not substantially increased for linkage and/or LD detection in oligogenic disease through the use of multi-locus model linkage analyses,³¹ though sometimes it can help to delineate between marginal true and false positives, and improve parameter estimation.¹⁰ If there is strong epistasis, ascertaining pedigrees with multiple affected individuals would increase the chance that multiple loci have alleles segregating to the affected individuals in those pedigrees. If two loci are both segregating in a given family, linkage of marker loci to either one individually would be detected with reasonably high probability, while allowance for epistasis would not increase the power substantially, especially when a preponderance of affected individuals are ascertained in the genotyped dataset. On the other hand, in the presence of strict admixture, where only one disease locus can have risk alleles segregating in any given pedigree – a very extreme deviation from the hypothesis of independent segregation of the disease loci across pedigrees – substantial power can be gained from joint analysis of multiple loci, as seen in monogenic diseases like retinitis pigmentosa.^{3,15} Full two-locus model joint linkage and LD analysis can lead to recovery of appropriate recombination fraction estimates, but the power to detect linkage is not, in general, increased substantially for epistatic models, especially if one tries all possible pairs of loci sequentially, as this leads to an explosion of the type I error rate due to multiple testing. Stratification on a known locus, on the other hand, can lead to substantial improvements in power to detect and estimate linkage and/or LD.^{32,33}

Given the substantial cost of a full two-locus linkage analysis in terms of computational intensity and number of analyses required, a simple approximation which is computationally trivial can be computed as a weighted sum of the products of single-locus likelihoods, which can be calculated using the LINKAGE programs,³⁴ for example. Software, MULTILOC, for implementation of this statistical analysis model is available from the author (jdt3@columbia.edu) for analysis of as many as four disease loci (each against markers on unlinked chromosomal regions) jointly, using a slightly modified version of the input data structures expected by the

HOMOG programs of Ott.¹⁵ This software will compute all the relevant likelihood ratios for the traditional lod scores under homogeneity, the traditional admixture tests,¹² as well as the extended admixture tests proposed in this manuscript, and runs under VMS (with translation available for Digital Unix upon request), and is written in DEC Pascal.

Acknowledgements

Support from the Burroughs-Wellcome Foundation and the Columbia Genome Center is gratefully acknowledged. The models proposed in this paper were developed and implemented in collaboration with Drs Satu Kuokkanen, Pentti Tienari, and Leena Peltonen at the Kansanterveyslaitos in Helsinki, and their constructive comments and suggestion for the development of this type of multi-locus analysis method is acknowledged gratefully. Thanks also to Dr Harald HH Göring for critical reading of the manuscript and many useful discussions relating to this model.

References

- 1 Inglehearn CF: Molecular genetics of human retinal dystrophies. *Eye* 1988; **12**: 571-579.
- 2 Van Camp G, Willems PJ, Smith RJ: Nonsyndromic hearing impairment: unparalleled heterogeneity. *Am J Hum Genet* 1997; **60**: 758-764.
- 3 Terwilliger JD, Göring HHH: Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000; **72**: 63-132.
- 4 Weiss KM: Is there a paradigm shift in human genetics? Lessons from the study of human diseases. *Mol Phylogenet Evol* 1996; **5**: 259-265.
- 5 Prud'homme D, Bouchard C, Leblanc C, Landry F: Sensitivity of maximal aerobic power to training is genotype dependent. *Med Sci Sports Exerc* 1984; **16**: 489-493.
- 6 Bouchard C: Genetics of aerobic power and capacity. In: Malina RM, Bouchard C (eds) *Sports and Human Genetics*. Human Kinetics: Champaign, IL, 1986.
- 7 Shephard RJ, Rode A: The health consequences of 'modernization'. Cambridge University Press: Cambridge, 1996.
- 8 Bouchard C, Malina RM, Perusse L: *Genetics of Fitness and Physical Performance*. Human Kinetics: Champaign, IL, 1997.
- 9 Terwilliger JD: On the resolution and feasibility of genome scanning approaches to unraveling the genetic components of multifactorial phenotypes. In: Rao DC (ed.). *Genetic Dissection of Complex Traits*. Academic Press: San Diego (in press), 2000.
- 10 Tienari PJ, Terwilliger JD, Ott J, Palo J, Peltonen L: Two-locus linkage analysis in multiple sclerosis. *Genomics* 1994; **19**: 320-325.
- 11 Schork NJ, Boehnke M, Terwilliger JD, Ott J: Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am J Hum Genet* 1993; **53**: 1127-1136.
- 12 Smith CAB: Homogeneity test for linkage data. *Proceedings, Second International Congress on Human Genetics* 1961; **1**: 212-213.
- 13 Göring HHH, Terwilliger JD: Linkage analysis in the presence of errors. IV: Joint pseudomarker analysis of linkage and/or LD on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* 2000: (in press).
- 14 Göring HHH, Terwilliger JD: Linkage analysis in the presence of errors. III: Marker loci and their map as nuisance parameters. *Am J Hum Genet* 2000: (in press).
- 15 Ott J: *Analysis of Human Genetic Linkage*, 1st edn. Johns Hopkins University Press: Baltimore, 1985.
- 16 Ott J, Bhattacharya S, Chen JD *et al*: Localizing multiple X chromosome-linked retinitis pigmentosa loci using multilocus homogeneity tests. *Proc Natl Acad Sci USA* 1999; **87**: 701-704.
- 17 Göring HHH, Terwilliger JD: Linkage analysis in the presence of errors. I: Complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet* 2000: (in press).
- 18 Göring HHH, Terwilliger JD: Linkage analysis in the presence of errors. II: Marker-locus genotyping errors modeled with hyper-complex recombination fractions. *Am J Hum Genet* 2000: (in press).
- 19 Dupuis J, Brown PO, Siegmund D: Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics* 1995; **140**: 853-856.
- 20 Lander ES, Kruglyak L: Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241-247.
- 21 Ott J: *Analysis of Human Genetic Linkage*, 3rd edn. Johns Hopkins University Press: Baltimore, 1999.
- 22 Weiss KM: *Genetic Variation and Human Disease*. Cambridge University Press: Cambridge, 1995.
- 23 Sing CF, Havilland MB, Reilly SL: Genetic architecture of common multifactorial diseases. Ciba Foundation Symposium, 1996; **197**: 211-232.
- 24 Mackay TFC: The nature of quantitative genetic variation revisited: lessons from *Drosophila* bristles. *BioEssay* 1996; **18**: 113-121.
- 25 Terwilliger JD, Weiss KM: Linkage disequilibrium mapping of complex disease: Fantasy or reality? *Curr Opin Biotechnol* 1998; **9**: 578-594.
- 26 Lucek PR, Ott J: Neural network analysis of complex traits. *Genet Epidemiol* 1997; **14**: 1101-1106.
- 27 Lie BA, Todd JA, Pociot F *et al*: The predisposition to type 1 diabetes linked to the human leukocyte antigen complex includes at least one non-class II genes. *Am J Hum Genet* 1999; **64**: 793-800.
- 28 Terwilliger JD: Linkage analysis model based. In: Armitage P, Colton T (eds). *Encyclopedia of Biostatistics*. John Wiley and Sons: Chichester, 1998, pp 2279-2291.
- 29 Knapp M, Seuchter SA, Baur MP: Linkage analysis in nuclear families: relationship between affected sib-pair tests and lod score analysis. *Hum Hered* 1994; **44**: 44-51.
- 30 Trembath RC, Clough RL, Rosbotham JL *et al*: Identification of a major susceptibility locus on chromosome 6p and evidence for further disease loci revealed by a two-stage genome-wide search in psoriasis. *Hum Mol Genet* 1997; **6**: 813-820.
- 31 Terwilliger JD, Ott J: On the interpretation of two-trait-locus/two-marker-locus lod scores. *Psychiatr Genet* 1993; **3**: 136.
- 32 Ott J, Falk CT: Epistatic association and linkage analysis in human families. *Hum Genet* 1982; **62**: 296-300.
- 33 Mulcahy B, Waldron Lynch F, McDermott MF *et al*: Genetic variability in the tumor necrosis factor-lymphotoxin region influences susceptibility to rheumatoid arthritis. *Am J Hum Genet* 1996; **59**: 676-683.
- 34 Lathrop GM, Ott J: Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. *Am J Hum Genet* 1990; **47**: 188.