Am. J. Hum. Genet. 66:1298–1309, 2000

Linkage Analysis in the Presence of Errors III: Marker Loci and Their Map as Nuisance Parameters

Harald H. H. Göring[1,*] and Joseph D. Terwilliger[2,3,4]

[1]Department of Genetics and Development and [2]Department of Psychiatry and [3]Columbia Genome Center, Columbia University and [4]New York State Psychiatric Institute, New York

In linkage and linkage disequilibrium (LD) analysis of complex multifactorial phenotypes, various types of errors can greatly reduce the chance of successful gene localization. The power of such studies—even in the absence of errors—is quite low, and, accordingly, their robustness to errors can be poor, especially in multipoint analysis. For this reason, it is important to deal with the ramifications of errors up front, as part of the analytical strategy. In this study, errors in the characterization of marker-locus parameters—including allele frequencies, haplotype frequencies (i.e., LD between marker loci), recombination fractions, and locus order—are dealt with through the use of profile likelihoods maximized over such nuisance parameters. It is shown that the common practice of assuming fixed, erroneous values for such parameters can reduce the power and/or increase the probability of obtaining false positive results in a study. The effects of errors in assumed parameter values are generally more severe when a larger number of less informative marker loci, like the highly-touted single nucleotide polymorphisms (SNPs), are analyzed jointly than when fewer but more informative marker loci, such as microsatellites, are used. Rather than fixing inaccurate values for these parameters a priori, we propose to treat them as nuisance parameters through the use of profile likelihoods. It is demonstrated that the power of linkage and/or LD analysis can be increased through application of this technique in situations where parameter values cannot be specified with a high degree of certainty.

Introduction

In linkage and linkage-disequilibrium (LD) analysis, investigators are painfully aware of the consequences of misspecifying the mode of inheritance of the phenotype (see Risch and Giuffra 1992; Göring and Terwilliger 2000*a*) and of genotyping errors at the marker loci (see Smith 1937; Lathrop et al. 1983; Terwilliger et al. 1990; Buetow 1991; Göring and Terwilliger 2000*b*). It is also well known that incorrect specification of marker-locus allele frequencies can lead to a systematic increase in false-positive rates when pedigrees are ascertained on the basis of the phenotype under study (see Ott 1992; Terwilliger and Ott 1994, exercise 28). In general, whenever parameter values are misspecified, the properties of likelihood-based analysis are likely to suffer: power may be diminished, false positive rates may be increased, and parameter estimates may be biased and/or inconsistent.

Received February 5, 1999; accepted for publication August 20, 1999; electronically published March 23, 2000.

Address for correspondence and reprints: Dr. Joseph D. Terwilliger, Columbia University, 1150 St. Nicholas Avenue, Room 548 (Unit 109), New York, NY 10032. E-mail: jdt3@columbia.edu

* Present affiliation: Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX.

0002-9297/2000/6604-0014$02.00

Errors in the assignment of trait-locus genotypes can be a function of incorrect assumptions about the mode of inheritance, errors in characterization of the phenotypes, and biases resulting from the ascertainment scheme. To deal with some types of errors in the genotype assignment at the trait locus, we have elsewhere proposed the use of complex-valued recombination fractions (Göring and Terwilliger 2000*a*), which can be applied in both "model-based" and "model-free" analysis of linkage and/or LD (Göring and Terwilliger 2000*c*). Marker-locus genotype assignment errors can result from laboratory errors (such as genotyping errors, sample mislabeling, etc.), incorrect assumptions about inheritance parameters of the marker loci (such as genotype frequencies, genetic maps, etc.), and so on. We have elsewhere introduced hypercomplex-valued recombination fractions as a means of compensating for both random and systematic laboratory errors in marker-locus genotyping (Göring and Terwilliger 2000*b*). In this manuscript, we propose the use of profile likelihoods maximized over nuisance parameters of the marker loci to circumvent the pathologies that result from setting these parameters to fixed, but erroneous, values a priori.

To test for correlations between a set of observed marker-locus genotypes, $\mathbf{G}_M$, and disease phenotypes, **Ph**, on a data set of any size and structure, one can

compute the likelihood, $L \propto P(\mathbf{Ph},\mathbf{G_M})$, as a function of numerous biological parameters, by partitioning over all possible disease-locus genotypes for all individuals in a data set, $\mathbf{g_D}$, as

$$P(\mathbf{Ph},\mathbf{G_M}) = P(\mathbf{Ph}|\mathbf{G_M})P(\mathbf{G_M}) = \sum_{\mathbf{g_D}} P(\mathbf{Ph}|\mathbf{g_D})P(\mathbf{g_D}|\mathbf{G_M})P(\mathbf{G_M}) \ .$$

$P(\mathbf{Ph}|\mathbf{g_D})$ is a function of the mode-of-inheritance assumed for the analysis; $P(\mathbf{g_D}|\mathbf{G_M})$ is a function of the disease-locus genotype frequencies, as well as linkage and LD between the disease and marker loci; and $P(\mathbf{G_M})$ is a function of the marker-locus genotype frequencies. Table 1 provides a more complete enumeration of parameters that can affect each term in the likelihood formula above. As shown elsewhere (Göring and Terwilliger 2000*c*), this "model-based" likelihood formulation can be generalized to "model-free" analysis as well, so that the techniques proposed here are directly applicable to either situation. For simplicity, we derive the theory and practice for "model-based" analysis alone.

In principle, the likelihood can be maximized over any subset of these underlying parameters. The values of some parameters, in particular the recombination fraction between the disease and the marker locus (or, equivalently, the map position of the trait locus in relation to several marker loci analyzed jointly), may be the object of inference, whereas the values of other parameters, such as allele frequencies of the marker loci, are typically not of inferential relevance and can therefore be treated as nuisance parameters. In any case, the likelihood can be maximized over parameters of interest and nuisance parameters alike. A likelihood maximized over a set of nuisance parameters is referred to as a profile likelihood (see Kalbfleisch and Sprott 1970; Royall 1997). Many of the specific likelihoods to be discussed in this manuscript are summarized in table 2, which states explicitly, for each of the enumerated likelihoods, what is assumed about the underlying biological phenomena, which parameters are fixed, and over which parameters the likelihood is maximized. For statistical inference, likelihoods maximized under two nested hypotheses may be compared to each other by likelihood ratio tests. The application of profile likelihoods to marker-locus parameters in linkage and/or LD analysis will be the primary focus of this paper.

Marker-locus Genotype Frequencies—Theoretical Model

In the likelihood formulation above,

$$P(\mathbf{Ph},\mathbf{G_M}) = \sum_{\mathbf{g_D}} P(\mathbf{Ph}|\mathbf{g_D})P(\mathbf{g_D}|\mathbf{G_M})P(\mathbf{G_M}) \ ,$$

$\mathbf{G_M}$ represents the observed marker-locus genotypes. In reality, however, the actual genotypes are often not known precisely for all individuals, since ungenotyped individuals are present in most data sets (especially for late-onset diseases), phase is often unknown for at least some individuals, and genotyping errors are unavoidable in practice. If we define the vector $\mathbf{g_M}$ to represent a set of marker-locus genotypes (with phase) for all individuals in the data set, the likelihood can be partitioned over all admissible vectors, as

$$P(\mathbf{Ph},\mathbf{G_M}) = \sum_{\mathbf{g_M}} P(\mathbf{g_M},\mathbf{G_M}) \sum_{\mathbf{g_D}} P(\mathbf{Ph}|\mathbf{g_D})P(\mathbf{g_D}|\mathbf{g_M})$$

$P(\mathbf{g_M},\mathbf{G_M})$ is a function of the marker-locus genotype frequencies in the population (which could, for example,

Table 1

Overview of the Parameters Determining the Likelihood in Model-Based Linkage and/or LD Analysis

Probability Term	Parameters
$P(\mathbf{Ph}\|\mathbf{g_D})$	For a qualitative trait: P(Phenotype|Genotype) (i.e., penetrances)
	For a quantitative trait: $f(\mu,\sigma^2\|\text{Genotype})$, or, in "model-free" "pseudomarker" analysis, P(meiosis informative)
$P(\mathbf{g_D}\|\mathbf{g_M})$	Disease-locus allele frequencies, denoted by $\mathbf{p_D}$
	Linkage between marker and disease loci, denoted by θ (or x_D in multipoint analysis)
	LD between marker and disease loci, denoted by $\boldsymbol{\delta_D}$
	Disease-locus heterogeneity and epistasis, denoted by $\boldsymbol{\beta_i}$
	Deviations from Hardy-Weinberg equilibrium (e.g., inbreeding coefficient, population substructure), denoted by F_{is},F_{st}
$P(\mathbf{g_M},\mathbf{G_M})$	Marker-locus allele frequencies, denoted by $\mathbf{p_i}$
	Intermarker LD, denoted by $\boldsymbol{\delta_M}$
	Marker-locus map positions (including locus order), denoted by $\mathbf{x_i}$
	Deviations from Hardy-Weinberg equilibrium (e.g., inbreeding coefficient, population substructure), denoted by F_{is},F_{st}

NOTE.—The likelihood is formulated as $P(\mathbf{Ph},\mathbf{G_M}) = \Sigma_{\mathbf{g_M}} P(\mathbf{g_M},\mathbf{G_M}) \Sigma_{\mathbf{g_D}} P(\mathbf{Ph}|\mathbf{g_D}) P(\mathbf{g_D}|\mathbf{g_M})$, where **Ph** denotes *the* set of *observed* trait phenotypes, $\mathbf{g_D}$ *a* set of possible *underlying* trait-locus genotypes, $\mathbf{G_M}$ *the* set of *observed* marker-locus genotypes, and $\mathbf{g_M}$ *a* set of possible *underlying* marker-locus genotypes for all individuals in the data set jointly. Bold-faced symbols represent vectors.

Table 2

Likelihoods That Can Be Compared in Likelihood-Ratio Tests

A. Two-Point Analysis

Marker-Disease Correlations				
Linkage	LD	θ	$\boldsymbol{\delta}_D$	Likelihood
No	No	0.5	**0**	$\max_{\mathbf{p}_i} L(\theta = 0.5, \boldsymbol{\delta}_D = \mathbf{0}, \mathbf{p}_i)$
Yes	No	$\tilde{\theta}$	**0**	$\max_{\theta,\mathbf{p}_i} L(\theta, \boldsymbol{\delta}_D = \mathbf{0}, \mathbf{p}_i)$
No	Yes	0.5	$\tilde{\boldsymbol{\delta}}_D$	$\max_{\boldsymbol{\delta}_D,\mathbf{p}_i} L(\theta = 0.5, \boldsymbol{\delta}_D, \mathbf{p}_i)$
Yes	Yes	$\hat{\theta}$	$\hat{\boldsymbol{\delta}}_D$	$\max_{\theta,\boldsymbol{\delta}_D,\mathbf{p}_i} L(\theta, \boldsymbol{\delta}_D, \mathbf{p}_i)$

B. Multipoint Analysis: Additional Parameters

Marker-Marker Correlations				
Map	LD	$\mathbf{x}_i$	$\boldsymbol{\delta}_M$	Likelihood[a]
Known	No	Fixed	**0**	$\max_{x_D,\mathbf{p}_i} L(x_D, \mathbf{x}_i, \boldsymbol{\delta}_D = \mathbf{0}, \boldsymbol{\delta}_M = \mathbf{0}, \mathbf{p}_i)$
Known	Yes	Fixed	$\hat{\boldsymbol{\delta}}_M$	$\max_{x_D,\boldsymbol{\delta}_M,\mathbf{p}_i} L(x_D, \mathbf{x}_i, \boldsymbol{\delta}_D = \mathbf{0}, \boldsymbol{\delta}_M, \mathbf{p}_i)$
Unknown	No	$\tilde{\mathbf{x}}_i$	**0**	$\max_{x_D,\mathbf{x}_i,\mathbf{p}_i} L(x_D, \mathbf{x}_i, \boldsymbol{\delta}_D = \mathbf{0}, \boldsymbol{\delta}_M = \mathbf{0}, \mathbf{p}_i)$
Unknown	Yes	$\tilde{\mathbf{x}}_i$	$\hat{\boldsymbol{\delta}}_M$	$\max_{x_D,\mathbf{x}_i,\boldsymbol{\delta}_M,\mathbf{p}_i} L(x_D, \mathbf{x}_i, \boldsymbol{\delta}_D = \mathbf{0}, \boldsymbol{\delta}_M, \mathbf{p}_i)$

Note.—The top half of the table lists the possible hypotheses regarding linkage and/or LD in two-point analysis of disease phenotypes and genotypes of a single marker locus. In the bottom half of the table, the additional parameters involved in multipoint analysis are enumerated, for the case of linkage but no LD between the disease locus and the marker loci (the likelihood could, of course, be computed for each of the other hypotheses from the top half of this table as well). Note that all bold-faced symbols represent parameter vectors (e.g., $\mathbf{p}_i$ represents the allele frequency distributions for one or more marker loci). In all cases shown here, the likelihood is assumed to be maximized over the marker-locus allele frequencies.

[a] Disease and marker loci linked, $\boldsymbol{\delta}_D = \mathbf{0}$.

be parameterized as a function of the marker-locus allele frequencies and the inbreeding coefficient, F_{is} [Wright 1922]) and errors in the marker-locus genotype assignment (which could, for example, be parameterized using imaginary components of hypercomplex recombination fractions [Göring and Terwilliger 2000*b*]). The main focus of statistical inference, however, is whether $P(\mathbf{g}_D|\mathbf{g}_M) = P(\mathbf{g}_D)$, independent of $\mathbf{g}_M$ (i.e., is there linkage or LD between the disease and marker loci or not?). $P(\mathbf{Ph}|\mathbf{g}_D)$ and $P(\mathbf{g}_M,\mathbf{G}_M)$ serve as weighting functions for the many possible underlying marker- and disease-locus genotype combinations, between which one tests for correlations in linkage and/or LD analysis. If these weights are inaccurate, the correlations between the underlying disease and marker-locus genotypes, $P(\mathbf{g}_D|\mathbf{g}_M)$, will be inappropriately quantified in the likelihood calculation, leading to pathological statistical behavior.

Pedigrees generally are ascertained on the basis of the phenotype to be studied, in such a way that there is a preponderance of affected individuals in the bottom generation(s) (in affecteds-only analyses, by definition, the only individuals who are phenotyped are affected). An apparent segregation bias results (under any assumed genetic model) at the disease locus, since parents who are heterozygous at the disease locus (D/+) would appear to have preferentially transmitted disease-predisposing alleles (D) to their offspring, as most of them are affected as a consequence of the ascertainment scheme. If there were a similar segregation distortion at the marker locus for some reason independent of the trait, false-positive evidence of linkage and/or LD may result (in contrast to the claim of Ott [1999], p. 270). This is because the offspring would appear to have inherited the same marker-locus allele identical by descent more often than would be expected by chance, just as they appear to have preferentially inherited disease alleles identical by descent as a consequence of the ascertainment. This increased allele sharing at both disease and marker loci would likely be interpreted as evidence of linkage. Let us now assume that all children in a nuclear pedigree are homozygous for the same allele at a marker locus and that their parents have not been genotyped. Under the assumption that this allele is common, this marker locus would provide little linkage information, since the parents themselves would likely be homozygous for this allele, such that one cannot distinguish identity by state from identity by descent among the marker-locus alleles inherited by the offspring. However, if this allele were erroneously assumed to be rare, the parents would most likely be inferred to be heterozygous for this allele, which would then lead to the erroneous inference that the children received the allele identical by descent. One can see that incorrect marker-locus allele frequencies assumptions can lead to an apparent segregation bias at the marker locus. If the nuclear pedigree had been ascertained on the basis of disease, such that a preponderance of the offspring are affected, this would bias the results of a linkage analysis towards false-positive evidence of linkage, as explained above. Note that this will not generally lead to high rates of false positives in linkage analysis when pedigrees are randomly ascertained, independent of the phenotype. In general, errors leading to an apparent segregation bias in a data set do not lead to false positives when they occur at only one locus, but when they occur at both loci, between which correlations are being tested, problems are ubiquitous (Smith 1953; Clerget-Darpoux et al. 1986).

In likelihood analysis of pedigrees and/or singletons, it has been advised to estimate the marker-locus allele

frequencies directly from the data (e.g., Falk and Rubinstein 1987; Boehnke 1991). Because allele frequencies are parameters of the likelihood, the likelihood can be maximized over them. Since ascertainment of pedigrees is performed independent of the marker-locus genotypes, there is no bias in estimating these parameters on the same data set to be used in subsequent linkage analysis. However, because the marker-locus allele frequencies are used to weight the correlations between marker- and disease-locus genotypes, as described above, the marker-locus allele frequencies are not orthogonal to the parameters used to quantify linkage and LD (which correlate the marker- and disease-locus genotypes). Therefore, the marker-locus allele frequencies should be estimated jointly with the linkage and LD parameters, whenever possible. A general technique for dealing with such unknown parameter values would be to compute the profile likelihood, maximized over the marker-locus allele frequencies as nuisance parameters. The LOD-score statistic testing for linkage would be computed as

$$Z_p = \log_{10}\{[\max_{\theta,\mathbf{p}_i} L(\theta,\mathbf{p}_i)]\}/[\max_{\mathbf{p}_i} L(\theta = 0.5,\mathbf{p}_i)]\}$$

(see Terwilliger and Ott 1994, p. 186), where θ represents the correlation caused by linkage between marker- and disease-locus genotypes in pedigrees and $\mathbf{p}_i$ denotes the vector of marker-locus allele frequencies (see Eguchi [1991] and Chin [1992] for more details about the use of nuisance parameters in likelihood ratio tests). The likelihoods in numerator and denominator would be independently maximized over the frequencies of all marker-locus alleles. As verified below by simulation, $2\ln(10^{Z_p})$ asymptotically converges to a 50-50 mixture of point mass at 0 and $\chi^2_{(1)}$.

In populations with substantial substructure, individuals may have homozygous genotypes (at all loci) more often than would be expected under Hardy-Weinberg equilibrium (Hardy 1908, Weinberg 1908). If this is not taken into account, it can likewise lead to false-positive evidence of linkage. The reason is similar to the one given above for errors in assumed allele frequencies: too many meioses among the ungenotyped individuals in upper generations of a pedigree will be inferred to be informative, leading to an overestimate of the probability that two affected individuals share marker-locus alleles identical by descent. Since alleles of the disease locus are inferred to be shared identical by descent at inflated frequencies, because of the aforementioned ascertainment bias, false-positive evidence of correlations between trait and marker loci would obtain. One can formulate the likelihood as a function of the inbreeding coefficient, F_{is} (Wright 1922), which can also be treated as a nuisance parameter in the analyses (Agarwala et al. 1999; Hovatta et al. 1999).

In practice, maximizing the likelihood over a large number of parameters can be very slow and computationally inefficient when performed using the ILINK program (Lathrop et al. 1984; Cottingham et al. 1993), especially if the starting values are not close to their maximum-likelihood estimates. Particularly when the number of marker-locus alleles is large, crude allele-counting procedures should be used to select reasonable starting values for marker-locus allele-frequency estimation with ILINK. Simple ad-hoc algorithms include counting the occurrences of each allele among typed founders (i.e., "gene counting" [Smith 1957]), or, if a substantial portion of founders is not genotyped, counting the occurrences of each allele in all genotyped individuals in the data set as if they were unrelated (as done, for example, by the DOWNFREQ program [Terwilliger 1994]). These allele-frequency estimates should be close to the maximum-likelihood estimates obtained when no linkage or LD is assumed between trait and marker loci. A more precise, but computationally intensive, procedure would be to maximize the likelihood of the marker-locus data alone (leaving the disease out of the analysis for the moment), which would yield marker-locus allele -frequency estimates that are identical to those obtained in joint analysis of disease phenotypes and marker-locus genotypes in the absence of linkage (since, when $\theta = .5$, the marker- and disease-locus genotypes are inherited independently of one another). Using the same estimates to compute the pedigree likelihood under linkage would be conservative, since the numerator of the likelihood ratio would be less than or equal to that obtained when maximized over the marker-locus allele frequencies and recombination fraction jointly (see Boehnke 1991; Terwilliger and Ott 1994, p. 186). Note that estimation of the allele frequencies jointly with the recombination fraction under the alternative hypothesis of linkage and use of these estimates to also compute the null-hypothesis likelihood of no linkage would lead to an anticonservative statistic and is not advised.

Marker-Locus Genotype Frequencies—Practice

For an illustration of the importance of having reliable estimates of the marker-locus allele frequencies, consider the pedigree shown in figure 1, in which a dominant disease is segregating and only the two affected individuals in the bottom generation are genotyped at the marker locus. The linkage information in this pedigree is highly dependent on the assumed allele frequency of the 1 allele. If this allele were, in reality, common in the population, this pedigree would contain almost no information about linkage, since the probability of the two genotyped individuals inheriting the same 1 allele identical by descent—even under the assumption of tight

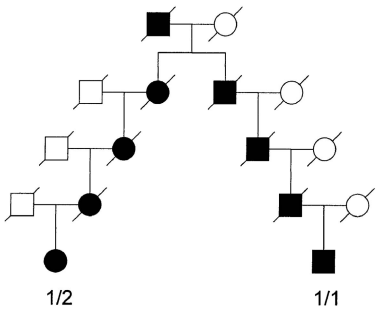

Figure 1 Example pedigree demonstrating the importance of marker-locus allele frequencies in linkage analysis. Assuming a rare dominant disease with no phenocopies, the maximum LOD score in this pedigree is 0 when the frequency of marker-locus allele 1 is set to its correct value of .5. If the frequency of the 1 allele were set erroneously to .0001, the maximum LOD score in this pedigree would be inflated to 1.8. In multipoint analysis, the marker-marker LD correlations also play a role (see text for details).

linkage—would be much smaller than the probability that they inherited the allele from different founders, since it may have entered the pedigree multiple times. If, however, the 1 allele were rare, the pedigree would provide a lot of linkage information, as the two genotyped individuals would be inferred to have inherited the same 1 allele identical by descent from a common ancestor. Under the assumption of a rare dominant disease with no phenocopies, if the frequency of marker-locus allele 1 were 0.5 in reality and were set to this correct value in the analysis, then the maximum LOD score in this pedigree would be 0, whereas, if the frequency of the allele were set erroneously to be .0001, then the maximum LOD score in this pedigree would be 1.8—a rather striking difference! At first glance, this example may seem exaggerated. It should be noted, however, that allele-frequency estimates obtained from the literature will often not be applicable to one's own data, since in most cases they are obtained from a population other than from which one's own pedigrees have been ascertained. Since the allele-frequency distributions for the vast majority of loci studied differ significantly between populations—indeed, the allele-frequency differences themselves can be used to reconstruct population history (see Cavalli-Sforza et al. 1994)—it is unwise to assume that any a priori estimates of allele frequencies not derived from the study population will be appropriate (see Hovatta et al. 1999 for an empirical example in which failure to allow for population substructure would have led to a spuriously high rate of false positives).

To demonstrate that the false-positive tendency caused by fixed and erroneous marker-locus allele frequencies can increase with the number of marker loci jointly analyzed in multipoint analysis, a data set of 350 affected sib-pairs with untyped parents was simulated, assuming absence of linkage between the marker and disease loci. A set of diallelic marker loci, separated serially by pairwise recombination fractions of .05, were simulated assuming true allele frequencies of .8 and .2 for the alleles of each marker locus. In the analysis, however, the alleles were incorrectly assumed to be equally frequent; that is, $P(1) = P(2) = 0.5$. In figure 2, the cumulative distribution functions are shown graphically for the LOD scores obtained using a "pseudo-marker" algorithm (analogous to "model-free" affected sib-pair analysis; see Göring and Terwilliger 2000*c*). While the LOD scores computed by treating the unknown marker-locus allele frequencies using profile likelihoods give a good fit to the theoretical null-hypothesis distribution of $2\ln(10^{Z_p}) \sim \{0.5(0) + 0.5\chi^2_{(1)}\}$ (Nordheim 1984; Tai and Chen 1989), the distribution of the LOD-score statistic shifts dramatically towards the right as the number of marker loci with fixed incorrect specifications of allele frequencies increases.

Correlations between Genotypes of Multiple Marker Loci—Theory

To describe the relationship between the genotypes of multiple marker loci, one needs to allow for correlations between them (see bottom of table 2). Normally, one

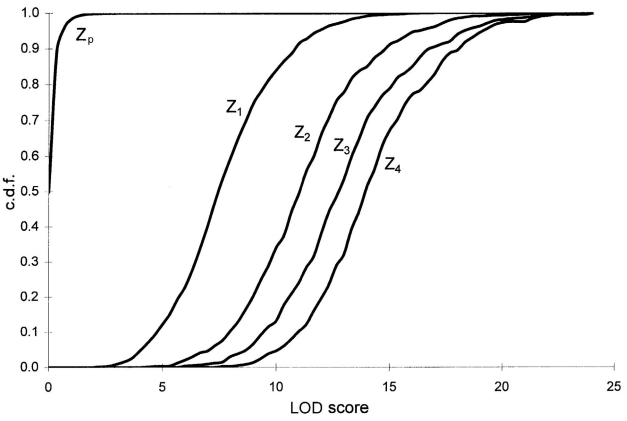

Figure 2 False-positive LOD-score distribution as a function of the number of marker loci with incorrect allele-frequency estimates analyzed jointly. In the simulation, the disease locus is unlinked to diallelic marker(s) having true allele frequencies of .2 and .8. Shown are the results using the profile-likelihood approach (Z_p) and when equal allele frequencies of .5 were falsely assumed for each of the 1 (Z_1), 2 (Z_2), 3 (Z_3), or 4 (Z_4) marker loci analyzed jointly. While the LOD-score distribution of the profile likelihood approach fits the predicted $0.5\chi^2$ distribution well, the false positive tendency clearly increases as the number of marker loci with fixed and erroneous assumed marker-locus allele frequencies considered jointly increases. The results are based on 100 simulated replicates of 350 affected sib-pairs with ungenotyped parents.

specifies parameter values that quantify these relationships, including marker-marker LD (as expressed in terms of the haplotype frequencies), and the marker-locus genetic map (i.e., marker-locus order and inter-marker recombination fractions).

Marker-Marker LD

Let us first focus on the effects of errors in specification of marker-marker LD (i.e., marker-locus haplotype frequencies), which can be understood by direct analogy to the effects of erroneous marker-locus allele frequency assumptions as discussed above. If one falsely assumes a common haplotype to be rare, it will likely be found more often than expected in the data set, leading to the false inference that this haplotype was inherited identical by descent by many pedigree members (many of whom will be affected with the disease as well, owing to nonrandom pedigree ascertainment). This may lead to false-positive evidence that this haplotype cosegregates with the disease and ultimately to false conclusions of linkage (by analogy to Ott 1992) and LD (by analogy to the well-known problem of "population stratification"—see Chase [1977]). The profile likelihood-ratio test for linkage between the disease locus and the marker loci can be written as

$$\log_{10}\frac{\max_{x_D,\mathbf{p}_i,\boldsymbol{\delta}_M} L(x_D,\mathbf{x}_i,\mathbf{p}_i,\boldsymbol{\delta}_D = 0,\boldsymbol{\delta}_M)}{\max_{\mathbf{p}_i,\boldsymbol{\delta}_M} L(x_D = -\infty,\mathbf{x}_i,\mathbf{p}_i,\boldsymbol{\delta}_D = 0,\boldsymbol{\delta}_M)} ,$$

where $\mathbf{x}_i$ represents the map of marker loci, x_D the map position of the disease locus relative to the marker loci ($x_D = -\infty$ if the disease locus is unlinked to the marker loci), $\boldsymbol{\delta}_M$ represents the LD relationships among the marker loci, $\boldsymbol{\delta}_D$ the LD relationships between the disease and marker loci, and $\mathbf{p}_i$ represents the marker-locus allele frequency distributions over all marker loci. The distribution of this statistic is similar to that of Z_p, with the exception that a multipoint LOD score has an intrinsic distribution that is a function of the length of the marker-locus map being analyzed (see Dupuis et al. 1995; Göring and Terwilliger 2000*a*; Göring et al. 1997). The distribution theory for an analogous statistical test based on "pseudomarkers" is covered by Göring and Terwilliger (2000*c*).

As in the case of single-marker-locus allele-frequency estimation, the ILINK program (Lathrop et al. 1984) can be used to maximize the likelihood over marker-marker haplotype frequencies, as described by Terwilliger and Ott (1994, chapter 23). ILINK has recently been extended to handle haplotype-frequency estimation conditional on fixed allele frequencies of the trait locus (in FASTLINK version 4.1P) (Cottingham et al. 1993). In practice, however, it may be difficult to obtain convergence to the maximum-likelihood estimates, unless the starting values for the haplotype frequencies are reasonably close to these values. By analogy to the approach suggested above for marker-locus allele-frequency estimation, one could obtain crude marker-marker haplotype-frequency estimates by a "gene-counting" procedure based on observed multiple marker-locus genotypes of all typed founders (as implemented, for example, in the EH program [Terwilliger and Ott 1994, chapters 23–24]). Alternatively, if there are a large number of ungenotyped founders, one could estimate haplotype frequencies on the basis of all genotyped individuals in the gene-counting procedure, as if they represented a random sample of unrelated individuals. While the likelihood-ratio tests performed by the EH program would in this situation be invalid for inference about marker-marker LD, the haplotype frequency estimates would provide reasonable starting values for the likelihood maximization by ILINK. (If one wanted to allow for LD not just between the marker loci themselves but also between the marker loci and the disease locus, one should hold the disease-locus allele frequencies constant, as described in Terwilliger and Ott 1992).

If, in the analysis, one wanted to use fixed marker-locus haplotype frequencies, obtained from a different source (e.g., Hellsten et al. 1993; Tienari et al. 1994), it is important that these estimates be representative of the genetic population from which one's own sample was obtained. Note that there is great variation between populations in the nature and strength of LD between loci, as a function of various population characteristics (see Clark et al. 1998; Terwilliger and Weiss 1998; Terwilliger et al. 1998).

Marker-Locus Maps

Let us focus now on inaccuracies of the marker-locus linkage map, which also affects the likelihood computation, through the term $P(\mathbf{g}_M,\mathbf{G}_M)$. A goal of the Human Genome Project has been to develop reliable genetic maps of marker loci, most frequently using a portion of the CEPH reference pedigree set (Dausset et al. 1990) for estimation of intermarker recombination fractions. However, the overall number of available meioses in the reference pedigrees is not sufficient for obtaining accurate estimates of the genetic distances between closely linked marker loci. In fact, it is often impossible to genetically order tightly linked marker loci in the available data sets, either because no recombination events may have taken place between two neighboring marker loci, or because the pair of marker loci are not both informative in the meioses where recombination did occur. The locus order can largely be resolved through physical mapping techniques, but this information is not always readily available to the analyst. Furthermore, physical distances cannot be simply converted into genetic dis-

tances, given the poor fine-scale correlations between physical and genetic distances (Chakravarti 1991; Jorde et al. 1994).

To alleviate the problems posed by an unknown marker-locus map, the map itself can be treated as a nuisance parameter by use of the profile-likelihood technique. In contrast to the classic locus-ordering problem where there is no definable null hypothesis (see Terwilliger and Ott 1994, chapter 14), the null hypothesis in this situation is that the disease locus is unlinked to the collection of marker loci, no matter what their order, and the alternative hypothesis is that the disease locus is linked to this collection of marker loci, irrespective of their order. The likelihood ratio can be written in such a way that these two hypotheses are nested, with a difference of one free parameter between them. If the position of marker locus 1 is arbitrarily defined to be $x_1 = 0$, the likelihood is maximized over all possible map positions of the other marker loci and the disease locus relative to this fixed position under the alternative hypothesis of linkage. Under the null hypothesis of no linkage, the likelihood is maximized over all possible positions of the other marker loci, conditional on the disease locus being unlinked to any of them ($x_D = -\infty$). In the case of, say, three marker loci and one disease locus, there are three free parameters to estimate under the alternative hypothesis (x_2, x_3 and x_D; $x_1 = 0$), while under the null hypothesis there are only two parameters to estimate (x_2 and x_3; $x_1 = 0$ and $x_D = -\infty$). The resulting multipoint LOD score can be written as

$$\Omega = \log_{10}\frac{\max_{x_D,x_2,x_3} L(x_D,x_1 = 0,x_2,x_3)}{\max_{x_2,x_3} L(x_D = -\infty,x_1 = 0,x_2,x_3)}$$

$$= \log_{10}\frac{\max_{x_D,\mathbf{x}_i} L(x_D,\mathbf{x}_i)}{\max_{\mathbf{x}_i} L(x_D = -\infty,\mathbf{x}_i)} \ .$$

Note that there is no restriction in this formulation on the order of the marker loci, as each of the x_i could assume any value on the range$(-\infty,\infty)$. This statistic is equally applicable to "model-free" and "model-based" analysis (see Göring and Terwilliger 2000*c*). By contrast, the conventional multipoint LOD score would be written as

$$Z = \log_{10}\frac{\max_{x_D} L(x_D,x_1 = 0,x_2 = c_2,x_3 = c_3)}{L(x_D = -\infty,x_1 = 0,x_2 = c_2,x_3 = c_3)}$$

$$= \log_{10}\frac{\max_{x_D} L(x_D,\mathbf{x}_i)}{L(x_D = -\infty,\mathbf{x}_i)} \ ,$$

where the map positions of marker loci 2 and 3 are fixed at $x_2 = c_2$ and $x_3 = c_3$ (on the basis of the best available marker-locus maps) and are assumed to be true in the analysis. Asymptotically, the behavior of these two statistics is predicted to be similar under the null hypothesis, whereas, under the alternative hypothesis, Ω should be more powerful when the marker-locus map is poorly characterized (see simulations below), as it will be in virtually all practical situations.

Sex-specific differences in recombination rates, when ignored, can have similar effects, since this is just another type of error in the specification of the intermarker recombination fractions (though the locus order is not sex-specific). It has been shown (Daw et al. 1998) that improper specification of these parameters can lead to errors in the conclusions of a multipoint linkage study. It has likewise been demonstrated (Sall and Bengtsson 1989; Terwilliger and Ott 1994, chapter 19.3) that the erroneous assumption of identical recombination rates in spermatogenesis and oogenesis can mimic chiasma interference in terms of the observed multilocus recombination rates, which can likewise lead to erroneous conclusions from multipoint analysis (Weeks et al. 1991). Solutions involving profile likelihoods analogous to those discussed above can be implemented. One could treat the recombination fractions as nuisance parameters in a sex-specific manner, or one could treat a constant ratio of the genetic distances in the two sexes as a nuisance parameter, jointly with the male (or female) recombination fractions (see Terwilliger and Ott 1992, chapter 18). Chiasma interference can be dealt with by an extension of this profile-likelihood protocol, as has been described and implemented in a specialized versions of the CILINK program by Weeks et al. (1991).

Correlations between Genotypes of Multiple Marker Loci—Practice

Marker-Marker LD

For an example of the effects of not allowing for marker-marker LD when it exists, let us consider an extreme situation where there are three tightly linked ($\theta \sim 0$) diallelic marker loci which are in complete LD, such that only the haplotypes 1 1 1 and 2 2 2 exist in the population under study. Let us return to the pedigree shown in figure 1—only now let us assume that the genotypes of the two individuals who are genotyped are 1 1 1/2 2 2 and 1 1 1/1 1 1. Let us assume, correctly, that the frequency of the 1 allele at each of the three marker loci was .2. (Since the marker loci are in complete LD, this would also be the frequency of haplotype 1 1 1.) When the correct allele-frequency estimates are used but absence of LD is incorrectly assumed, the maximum multipoint LOD score would be 1.5. However, if complete LD is allowed for (correctly) between the marker loci and the correct haplotype frequencies, .2 and .8, are used, the maximum multipoint LOD score would be reduced to only 0.4. In this example, a sub-

stantial inflation of the LOD score arises solely because of incorrectly specified marker-marker LD (parameterized in terms of the marker-marker haplotype frequencies). The explanation for this finding is that the frequency of the 1 1 1 haplotype is greatly underestimated as (0.2)(0.2)(0.2) = 0.008, when LD is not taken into account, versus 0.2 in reality, which inflates the probability that the shared haplotypes have been inherited identical by descent, as in the example of incorrect marker-locus allele frequencies above. In realistic situations, it can likewise be demonstrated that there is a systematic increase of the type I–error rate if the LD is not modeled accurately.

Marker-Locus Maps

To empirically examine the properties of treatment of the map of marker loci (marker-locus order and the intermarker genetic distances) as nuisance parameters by means of profile likelihoods, the following simulation study was performed: first, four marker loci were simulated in the entire panel of 64 CEPH reference pedigrees (Dausset et al. 1990). The true map of marker loci was M_1-(0.05)-M_2-(0.05)-M_3-(0.05)-M_4, and the marker loci had 75% heterozygosity. The same four marker loci, together with a disease locus positioned in the middle between the second and the third marker locus (order M_1-(0.05)-M_2-(~0.025)-D-(~0.025)-M_3-(0.05)-M_4) were also simulated in a set of 50 nuclear pedigrees with three to five affected children each and one affected parent. The disease penetrances were $f_{DD} = 0.75$, $f_{D+} = 0.25$; $f_{++} = 0.025$, with disease-locus allele frequency $\mathbf{p}_D = 0.1$. For the computation of the conventional multipoint LOD score, *Z,* on the data set of nuclear pedigrees, the recombination fractions between adjacent marker loci were first estimated on the CEPH pedigrees with CILINK (Lathrop et al. 1984), and the resulting maximum-likelihood estimates of the intermarker recombination fractions were fixed in the subsequent analysis of the 50 nuclear pedigrees. In the computation of the "map as nuisance parameter" statistic, Ω, the CEPH panel was not used at all, as the intermarker recombination fractions were treated as nuisance parameters in the actual analysis of the nuclear pedigrees. This procedure was repeated 100 times to get an estimate of the difference in the power of the two statistics (table 3). By design, the simulated situation was not very powerful, in order to illustrate more clearly the difference in performance of the two methods. Note that Ω is more powerful than *Z* for all LOD-score thresholds considered. In other examples (data not shown), the increase in expected LOD score from use of Ω instead of *Z* ranged from 2% to 30%, depending on the sample sizes, marker-locus density, and mode of inheritance of the disease. For more-distantly-spaced marker loci, the effect is attenuated, as the multilocus likelihood is less sensitive to small errors in the estimates of larger recombination fractions. When fewer reference pedigrees were used to estimate the marker-locus map—and, in practice, only 10 CEPH pedigrees are often typed—the gain in power was greater, as expected. With larger multigenerational pedigrees, the gains in power can also be greater, since phase is more often known, making the effect of errors in the intermarker recombination fractions potentially larger. The simulated example illustrates the effects of reliance on a poorly characterized genetic marker-locus map in even the simplest of situations.

Table 3

Comparison of Conventional Multipoint LOD Scores to Multipoint LOD Scores Treating the Marker-Locus Map as a Nuisance Parameter with Profile Likelihoods, in Situations where the Marker-Locus Map Is Incorrect

LOD-Score Threshold	Theoretical *P* Value	Observed Frequency			
		1-2-D-3-4		1-2-3-4-D	
		Z	Ω	*Z*	Ω
$Z > 0.5$	.0646	.80	.85	.06	.05
$Z > 1.0$	.0159	.44	.52	.02	.01
$Z > 2.0$	.0012	.16	.23	.00	.00
$Z > 3.0$	.0001	.02	.06	.00	.00
EMLOD	.1086	2.01	2.22	.123	.119

Note.—In the alternative-hypothesis simulations (100 replicates), the marker-locus order (given with interlocus recombination fractions in parentheses) is M_1-(.05)-M_2-(~.025)-D-(~.025)-M_3-(.05)-M_4, whereas in the null-hypothesis simulations (100 replicates), the marker-locus order is M_1-(.05)-M_2-(.05)-M_3-(.05)–M_4-(.5)-D. The bottom line of the table gives the expected maximum LOD score (EMLOD) for each statistic/locus-order combination.

To verify that under the null hypothesis there is no inflation of the LOD score when the intermarker distances were treated as nuisance parameters, the same simulation procedure was performed with the disease locus unlinked to the set of marker loci. As can be seen in the same table (table 1), the distribution of both statistics was very similar, suggesting that the nuisance-parameter statistic does not lead to an inflation of the type I error–rate relative to conventional LOD-score analysis.

Discussion

Likelihood analysis is a powerful and intuitive vehicle for statistical inference if the probability of a given set of observed data can be written as a function of a set of parameters. Whenever the null and alternative hypotheses can be fully described with such parameters, the likelihoods under each hypothesis can be compared to evaluate the evidence in support of one hypothesis versus another. One need not be interested in making inferences about all parameters. In fact, the values of

certain parameters are irrelevant and/or impossible to evaluate, in many cases. Either such parameters can be fixed, a priori, to some values that can be assumed for the purpose of the analysis, or the likelihood can be maximized with respect to those parameters by means of profile likelihoods. If the likelihood is maximized independently over these parameters under both hypotheses (i.e., nuisance-parameter analysis by means of profile likelihoods), then the difference in the two likelihoods would provide evidence about the parameters that are the basis of inference. In this article, we have discussed situations in which the use of profile-likelihood treatment of nuisance parameters may be preferential to assumption of some fixed values for these parameters from the outset. This technique can be useful in dealing with parameters that are not the primary focus of inference, such as allele frequencies of the marker loci, LD between them, marker-locus maps, and the sex-specificity thereof, as shown here. Other parameters related to marker loci and their map—including inbreeding coefficients (Agarwala et al. 1999; Hovatta et al. 1999) and parameters describing chiasma interference (Weeks et al. 1991) and genotyping errors (Göring and Terwilliger 2000*b*)—could be dealt with in similar fashion. However, parameters of the disease locus underlying the phenotype through which the sample is ascertained are typically not straightforward to deal with in this manner, because of the effects of ascertainment bias. For example, the ascertainment of multiplex pedigrees would cause estimates of the mode of inheritance and/or disease-locus allele frequencies to be strongly biased, unless ascertainment correction was made (which is typically impossible, as real-world ascertainment schemes are rarely mathematically tractable). If pedigrees are randomly ascertained, without regard to phenotype, or if pedigrees are ascertained on the basis of a phenotype which is uncorrelated with the phenotype being studied, then one could compute profile likelihoods over trait-locus parameters as well (e.g., means and variances of quantitative traits, or penetrances of qualitative traits [Almasy and Blangero 1999]). Other trait-locus parameters, such as those related to epistasis and oligogenic inheritance in the extended admixture test (Terwilliger, in press) can be treated as nuisance parameters, even when the data is ascertained on the basis of the phenotype.

Unfortunately, difficulties may arise when too many (often nonorthogonal) parameters are estimated jointly, especially if the size of the data set is small, because of the risk of overfitting some model to the data. In the case of marker-locus allele frequencies and marker-marker haplotype frequencies, there is often little difference between the estimates when the disease locus is linked and when it is not, so that, when the large number of nuisance parameters becomes a concern, one could estimate the allele or haplotype frequencies under the null hypothesis of no linkage to the disease locus and could fix those estimates for the likelihood computation under the alternative hypothesis of linkage (see Boehnke 1991). This would avoid, in a conservative manner, some of the distributional complexities of likelihood ratios with more nuisance parameters than data points (see Eguchi 1991). Simulations presented here and others not shown indicate that the statistics with marker-locus allele frequencies or marker-marker haplotype frequencies as nuisance parameters behave well in the moderately-sized simulated data set of 350 nuclear pedigrees. Similar results were obtained for the marker-locus map locations treated as nuisance parameters. If the likelihood is maximized over both allele/haplotype frequencies and map positions jointly, the distribution may become quite complicated—especially as many such parameters are not completely orthogonal. Nonorthogonality of nuisance parameters can lead to likelihood-based statistics that deviate from their predicted distributions (e.g., Terwilliger 1995, 1996), especially in "small" samples, and care must be taken to avoid the consequences of these problems in practice.

The field of human genetics is moving rapidly towards analysis of complex multifactorial phenotypes against dense maps of tightly linked single nucleotide polymorphisms (SNPs) (Terwilliger et al. 1992; Collins et al. 1997; Pennisi 1998). These marker loci are very similar in their statistical properties to the restriction fragment length polymorphisms (RFLPs) (Botstein et al. 1980) which were popular in the last decade (most RFLPs *are*, in fact, SNPs that were studied with different experimental techniques). At that time, it was widely appreciated that errors in the assumed marker-locus allele frequencies would lead to high rates of false positives (see Ott 1992). When microsatellite marker loci were introduced, this problem dissipated somewhat—at least in analysis of large pedigrees—because the marker loci were sufficiently informative that the situations in which marker-locus allele-frequency errors lead to high false-positive rates were encountered much less frequently (those situations being when parental genotypes could not be uniquely determined on the basis of their children's genotypes and common alleles were falsely assumed to be rare). Investigators may have forgotten the severity of the consequences of allele-frequency errors in the relatively uninformative RFLPs, especially when parents of affected individuals were unavailable for genotyping (an increasingly common situation as we move towards analysis of chronic diseases of old age). In LD analysis, this has been recognized throughout (Falk and Rubinstein 1987), leading to the common platitude that case-control studies suffer a risk of false positives when the cases and controls are not well matched (meaning that the marker-locus allele frequen-

cies were inaccurately estimated from the ascertained control sample [Chase 1977]).

In the days of RFLPs, the genetic maps (e.g., Morton and Collins 1990) derived from linkage analysis in the CEPH pedigrees (Dausset et al. 1990) were not very accurate either, because the RFLPs were individually rather uninformative and the total pool of informative meioses available for study was accordingly too small to accurately estimate even the relatively large recombination fractions between the RFLPs. When microsatellites were studied, a larger proportion of meioses was informative, increasing the accuracy of the resulting maps based on such small data sets, though for many chromosomal regions even these maps can be inconsistent (compare the maps of Murray et al. 1994; Dib et al. 1996; Broman et al. 1998, for example). Now, with SNPs, the accuracy of the genetic maps will likely be much worse, because the distances between them are much smaller (thus requiring more informative meioses to estimate the genetic maps), while, at the same time, the proportion of informative meioses in a given data set (e.g., the CEPH reference panel) will be much smaller. The effects of errors in these maps may become more significant as more and more marker loci are being analyzed jointly, despite the popular belief to the contrary (see Kruglyak 1997). While the SNPs may be physically ordered, and their physical map positions known accurately, this does not help us that much in linkage analysis because it is the interlocus recombination fractions that are critical to the accurate computation of multipoint pedigree likelihoods, and the correlations between physical distance and genetic distance are poor, even on a macroscopic scale (Chakravarti 1991; Jorde et al. 1994). There is substantial evidence that the correlation is even more unpredictable on the microscopic scale (e.g., Lichten and Goldman 1995; Ajioka et al. 1997; Clark et al. 1998; Mohrenweiser et al. 1998). Furthermore, the recombination rates are often quite different in spermatogenesis and oogenesis (Haldane 1922; Tanzi et al. 1992; Broman et al. 1998; Mohrenweiser et al. 1998), and this sex difference also needs to be allowed for in order for multipoint analysis to be efficient and powerful (see Terwilliger and Ott 1994, chapter 18, 19.3). Since a dense set of SNPs must be analyzed jointly, in a multipoint manner, if there is to be any power in linkage or LD analysis, problems may arise from correlations between genotypes of tightly linked marker loci. It is clear that when SNPs are very closely linked, there may be substantial LD between them (e.g., Clark et al. 1998), which must be allowed for in the analysis. Furthermore, chiasma interference may play an important role, and it may be necessary to allow for it as well. However, if one allows for either of these phenomena, the application of Markov models (Lander and Green 1987; Terwilliger et al. 1992) to the computation of multipoint likelihoods becomes impossible, since the requisite "memoryless" Markov property no longer applies as one moves from marker locus to marker locus along the chromosome.

In conclusion, it is a major drawback of most multipoint methods of analysis that marker-locus maps must be assumed to be known accurately, while in reality fine-scale intermarker recombination fractions cannot be accurately estimated without thousands of informative meioses. Most simulation studies of the power of multipoint methods simulate a fixed map of marker loci with known allele frequencies, and perform the analysis assuming that the marker-locus map and other parameters are known with 100% accuracy, significantly inflating the predicted power of the investigated approaches in practice. This is probably one factor contributing to the widespread, though gradually dissipating (see Pennisi 1998; Terwilliger and Weiss 1998; Terwilliger and Göring 2000), support for the hypothesis that a dense genome-spanning map of diallelic marker loci will be ideal for the mapping of genes predisposing to complex disease (see Terwilliger et al. 1992; Kruglyak 1997). When the incumbent errors in map distance estimates and marker-locus allele frequencies are allowed for, the predicted sample size requirements and requisite density of genotyped SNPs might increase dramatically (see Terwilliger et al. 1992). This should be investigated in more detail before one abandons the polymorphic microsatellite marker loci currently in use (see Terwilliger et al. [1998] and Terwilliger and Weiss [1998] for further reasons to stick with highly polymorphic microsatellite marker loci). It is hoped that the use of statistical approaches like those described here may alleviate a few of the basic criticisms of dense maps of diallelic marker loci.

Software

To obtain the following software (written for VMS using DEC Pascal), please contact the authors by e-mail (jdt3@columbia.edu, hgoring@darwin.sfbr.org). A Digital Unix version is available for some of the software and is expected to be available shortly for the remainder. a) DOWNFREQ is a utility program for gene counting to generate starting values for marker-locus allele frequency estimation with ILINK. b) Shell software is available for performing two-point linkage analysis with marker-locus allele frequencies as nuisance parameters, which calls DOWNFREQ and repeatedly restarts ILINK with varying sets of starting values until some specified convergence criteria are met. Note that this mimics the manual manipulations one normally has to do to get good convergence from ILINK when varying parameters other than the recombination fraction (Lalouel 1979). The same set of shell scripts also performs a similar

procedure allowing for LD between trait and marker loci in "pseudomarker" analysis (see Göring and Terwilliger 2000*c* for details). c) For multipoint analysis with the marker-locus map as a nuisance parameter, the MULTI-ILINK program (Terwilliger 1994) has been upgraded to allow for likelihood maximization over all possible marker-locus orders, as described in this text. Earlier versions maximized the likelihood only over the intermarker recombination fractions for a fixed marker-locus order.

Acknowledgments

A Hitchings-Elion Fellowship from the Burroughs-Wellcome Fund (to J.D.T.) is gratefully acknowledged, as is grant HG00008 from the National Institute of Health to Jürg Ott (thesis advisor of H.H.H.G.). Critical comments on an earlier version of this manuscript by Clyde C. Clark, Daniel E. Weeks, and two anonymous reviewers are gratefully appreciated.

References

Agarwala R, Biesecker LG, Schäffer AA. Inverse inbreeding coefficient problems with an application to linkage analysis of recessive diseases in inbred populations. Discrete Appl Math (in press)

Ajioka R, Jorde LB, Gruen JR, Yu P, Dimitrova D, Barrow J, Radisky E, et al (1997) Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. Am J Hum Genet 60:1439–1447

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62: 1198–1211

Boehnke M (1991) Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25

Botstein D, White RL, Skolnick MH, Davies RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314–331

Broman KW, Murray JC, Sheffield VS, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63:861–869

Buetow KH (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. Am J Hum Genet 49: 985–994

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Chakravarti A (1991) A graphical representation of genetic and physical maps: the Marey map. Genomics 11:219–222

Chase GA (1977) Genetic linkage, gene-locus assignment, and the association of alleles with diseases. Transplant Proc 9: 167–171

Chin DC (1992) On the connection between maximum likelihood sensitivity analysis and nuisance parameter analysis. IEEE Transact Aerosp Electron Syst 28:884–886

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård J, Salomaa V et al (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63: 595–612

Clerget-Darpoux F, Bonaïti-Peilié C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Collins FS, Guyer MS, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Cottingham RW Jr, Idury RM, Schaffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53: 252–263

Dausset J, Cann H, Cohen D, Lathrop GM, Lalouel JM, White R (1990) Centre d'Étude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575–577

Daw EW, Thompson EA, Wijsman EM (1998) Bias in multipoint linkage analysis arising from map misspecification. Am J Hum Genet Suppl 63:A17

Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, et al (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154

Dupuis J, Brown PO, Siegmund D (1995) Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 140:843–856

Eguchi S (1991) A geometric look at nuisance parameter effect of local powers in testing hypotheses. Ann Inst Statist Math 43:245–260

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Göring HHH, Terwilliger JD (2000*a*) Linkage analysis in the presence of errors I: complex valued recombination fractions and complex phenotypes. Am J Hum Genet 66:1095–1106

——— (2000*b*) Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. Am J Hum Genet 66:1107–1118

——— (2000*c*) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. Am J Hum Genet 66:1310–1327 (in this issue)

Göring HHH, Terwilliger JD, Ott J (1997) A likelihood-based approach to extended haplotype analysis of shared segments using a Markovian branching process. Am J Hum Genet Suppl 61:A277

Haldane JBS (1922) Sex ratio and unisexual sterility in hybrid animals. J Genet 12:101–109

Hardy GH (1908) Mendelian proportions in a mixed population. Science 28:49–50

Hellsten E, Vesa J, Speer MC, Mäkelä TP, Järvelä I, Alitalo K, Ott J et al (1993) Refined assignment of the infantile neuronal ceroid lipofuscinosis (INCL, CLN1) locus at 1p32: incorporation of linkage disequilibrium in multipoint analysis. Genomics 16:720–725

Hovatta IM, Varilo T, Suvisaari J, Terwilliger JD, Ollikainen V, Arajärvi R, Juovinen H, et al (1999) A genomewide screen

for schizophrenia genes in an isolated Finnish subpopulation suggesting multiple susceptibility loci. Am J Hum Genet 65: 1114–1124

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A, Leppert M (1994) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 54:884–898

Kalbfleisch JD, Sprott DA (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). J R Stat Soc (B) 32:175–208

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–24

Lalouel J-M (1979) GEMINI—a computer program for optimization of general nonlinear functions. Tech rep 14, University of Hawaii, Honolulu

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84: 2363–2367

Lathrop GM, Hooper AB, Huntsman JW, Ward RH (1983) Evaluating pedigree data: I. The estimation of pedigree error in the presence of marker mistyping. Am J Hum Genet 35: 241–262

Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446

Lichten M, Goldman AS (1995) Meiotic recombination hotspots. Annu Rev Genet 29:423–444

Mohrenweiser HW, Tsujimoto S, Gordon L, Olsen AS (1998) Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of human chromosome 19. Genomics 47:153–162

Morton NE, Collins A (1990) Standard maps of chromosome 10. Ann Hum Genet 54:235–251

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier Heddema T, Manion F, Quillen J, et al (1994) A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). Science 265: 2049–2054

Nordheim EV, O'Malley DM, Chow SC (1984) On the performance of a likelihood ratio test for genetic linkage. Biometrics 40:785–790

Ott J (1992) Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet 51: 283–290

Ott J (1999) Analysis of human genetic linkage. 3d ed. Johns Hopkins University Press, Baltimore

Pennisi E (1998) A closer look at SNPs suggests difficulties. Science 281: 1787–1789

Risch N, Giuffra L (1992) Model misspecification and multipoint linkage analysis. Hum Hered 42:77–92

Royall RM (1997) Statistical evidence: a likelihood paradigm. Chapman & Hall, London

Sall T, Bengtsson BO (1989) Apparent negative interference due to variation in recombination frequencies. Genetics 122: 935–942

Smith HF (1937) Test of significance for Mendelian ratios when classification is uncertain. Ann Eugen 8:94–95

Smith CAB (1953) The detection of linkage in human genetics. J R Stat Soc 15B:153–184

——— (1957) Counting methods in genetical statistics. Ann Hum Genet 21:254–276

Tai JJ, Chen CL (1989) Asymptotic distribution of the lod score for familial data. Proc Natl Sci Counc Repub China [B] 13: 38–41

Tanzi RE, Watkins PC, Stewart GD, Wexler NS, Gusella JF, Haines JL (1992) A genetic linkage map of human chromosome 21: analysis of recombination as a function of sex and age. Am J Hum Genet 50:551–558

Terwilliger JD (1994) The available possibilities to analyse data of polygenic disease statistically. Paper presented at the IVth workshop of the Nordic Genome Initiative, Helsinki, September 3–5

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56: 777–787

——— (1996) Likelihood ratio tests for linkage and linkage disequilibrium: Asymptotic distribution and power—reply. Am J Hum Genet 58:1095–1096

——— . A likelihood-based admixture model of oligogenic inheritance in "model-based" or "model-free," two-point or multi-point, linkage and/or LD analysis. Eur J Hum Genet (in press)

Terwilliger JD, Ding Y, Ott J (1992) On the relative importance of heterozygosity and intermarker distance in gene mapping. Genomics 13:951–956

Terwilliger JD, Göring HHH (2000) Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. Hum Biol 72:63–132

Terwilliger JD, Ott J (1992) A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337–346

——— (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Terwilliger JD, Weeks DE, Ott J (1990) Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross overestimation of the recombination fraction. Am J Hum Genet Suppl 47:A201

Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 9:578–594

Terwilliger JD, Zöllner S, Laan M, Pääbo S (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: "drift mapping" in small populations with no demographic expansion. Hum Hered 48:138–154

Tienari PJ, Terwilliger JD, Ott J, Palo J, Peltonen L (1994) Two-locus linkage analysis in multiple sclerosis. Genomics 19:320–325

Weeks DE, Ott J, Lathrop GM (1991) Multipoint mapping under different models of interference using the LINKAGE programs. Am J Hum Genet Suppl 49:A372

Weinberg W (1908) Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64:368–382

Wright SE (1922) Coefficients of inbreeding and relationship. Am Nat 56:330–338