Am. J. Hum. Genet. 63:1517–1530, 1998

Disequilibrium Likelihoods for Fine-Scale Mapping of a Rare Allele

Jinko Graham[1] and Elizabeth A. Thompson[1,2]

Departments of [1]Biostatistics and [2]Statistics, University of Washington, Seattle

Summary

Genetic linkage studies based on pedigree data have limited resolution, because of the relatively small number of segregations. Disequilibrium mapping, which uses population associations to infer the location of a disease mutation, provides one possible strategy for narrowing the candidate region. The coalescent process provides a model for the ancestry of a sample of disease alleles, and recombination events between disease locus and marker may be placed on this ancestral phylogeny. These events define the recombinant classes, the sets of sampled disease copies descending from the meiosis at which a given recombination occurred. We show how Monte Carlo generation of the recombinant classes leads to a linkage likelihood for fine-scale mapping from disease haplotypes. We compare single-marker disequilibrium mapping with interval-disequilibrium mapping and discuss how the approach may be extended to multipoint-disequilibrium mapping. The method and its properties are illustrated with an example of simulated data, constructed to be typical of fine-scale mapping of a rare disease in the Japanese population. The method can take into account known features of population history, such as changing patterns of population growth.

Introduction

Genetic linkage studies based on pedigree data have limited resolution, because of the relatively small number of segregations (Boehnke 1994). Disequilibrium mapping, which uses population associations to infer the location of a disease mutation, provides one possible strategy for narrowing the candidate region. In this article, we develop a method for obtaining fine-scale mapping linkage likelihoods based on data on sampled disease haplotypes and knowledge of marker-allele frequencies in the general population.

Since the initial success of the disequilibrium mapping of cystic fibrosis (Cox et al. 1989), Huntington disease (Snell et al. 1989; Theilmann et al. 1989), and diastrophic dysplasia (DTD) (Hästbacka et al. 1992), disequilibrium mapping has attracted much attention from practitioners, and a number of authors have formulated inference approaches to the problem. Several of these approaches involve simulation of disease-haplotype histories or ancestries. Kaplan et al. (1995) provided the first likelihood approach to the problem. They used forward simulation of disease history, with rejection sampling of histories that do not lead to current disease-allele counts within a specified range. Rannala and Slatkin (1998) used simulation of ancestries of disease haplotypes, conditional on the current disease-allele counts and the sampling of a specified number of disease haplotypes. Xiong and Guo (1997) developed an approximate likelihood approach based on a population genetic model for the evolution of marker-allele frequencies in the disease population but did not condition on current disease-allele counts. Van der Meulen and te Meerman (1997) used simulation of haplotypes under a genetic drift model to investigate properties of disequilibrium linkage–detection methods based on haplotype sharing.

Since we base our method on first realizing the coalescent ancestry of a sample of current disease alleles, our approach is closest in spirit to that of Rannala and Slatkin (1998). Our method for realizing such ancestries is outlined below; details are given by Graham (1998). Then, given an ancestry, recombination events on disease-bearing haplotypes may be placed (by simulation) onto this ancestral coalescent. We define the notion of a "recombinant class." This is the set of current sampled disease haplotypes that descend from the meiosis at which a given recombination event occurred, without fragmentation by subsequent recombination events. This underlying coancestry induces dependence among the disease haplotypes in the sample, with those disease haplotypes within any recombinant class necessarily having the same marker haplotype. By use of our methods, large Monte Carlo samples of recombinant classes for various hypothesized trait gene locations are readily obtained. We combine these Monte Carlo samples with an analytic method for the computation of the probability of ob-

Received April 27, 1998; accepted for publication September 9, 1998; electronically published October 9, 1998.

Address for correspondence and reprints: Dr. Elizabeth A. Thompson, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195. E-mail: thompson@stat.washington.edu

0002-9297/98/6305/0030$02.00

served disease haplotypes conditional on latent recombinant classes, to obtain a likelihood for fine-scale mapping.

We first describe a hypothetical but realistic example of a rare gene in the Japanese population. We next outline briefly our approach to obtaining Monte Carlo realizations of disease ancestry and recombinant classes, and we then show how these can be used to obtain linkage likelihoods for a single marker. We extend the single-marker approach to interval mapping and discuss the relative precision of interval versus single-marker mapping, as well as further extensions to multipoint mapping. Throughout, the methods are illustrated with the Japanese example, and confidence bounds on likelihood estimates of trait gene location are obtained via the parametric bootstrap. We also consider the real example of DTD in Finland (Hästbacka et al. 1992), to provide a comparison with other methods.

A Japanese Variant

We introduce an example typical of fine-scale mapping of a disease allele in the Japanese population: this example is motivated by the recent mapping and positional cloning of the Werner syndrome gene in the Japanese (Yu et al. 1996). Werner syndrome is a rare autosomal recessive disease characterized by premature onset of a number of age-related traits. Although there are at least eight distinct mutations at the Werner syndrome locus in present-day Japanese, we consider a single monophyletic variant such as (probably) *WRN4*. *WRN4* is the most frequent of the eight mutations and represents ~51% of mutants (Matsumoto et al. 1997). In the Japanese, the estimated allele frequency of all Werner syndrome mutations combined is .002–.004 (Goddard et al. 1996). The higher estimated frequency of .004 and a current Japanese population size of at least 120,000,000 people, or 240,000,000 alleles (ISEI 1998; JIN 1998), gives a *WRN4* copy number of ~500,000.

The Japanese population has a well-documented history (Benedict 1989), and data on population size are available (Koyama 1979; ISEI 1998). Under the replacement hypothesis of Japanese origins (Hanihara 1991; Rose 1996), the modern Japanese population was founded ~94 generations ago, by small numbers of rice-growing immigrants arriving from the mainland. For this example, we assume a founding population of 1,000 individuals, or 2,000 copies. Population data are sparse until the Edo period of Japanese history (1603–1867 A.D.), during which the feudal government pursued a policy of almost total seclusion from the outside world. During this time, population size remained approximately constant (Benedict 1989), at ~60,000,000 copies (ISEI 1998), despite 11 generations of peace. Following the Meiji reform of 1867, the feudal system was abolished, and Japan underwent a period of rapid transformation and population growth that continues today (Benedict 1989).

The median age of a selectively neutral allele, given its current copy number, may be estimated by use of the subtree coalescent methods outlined by Griffiths and Tavaré (1998). For the Japanese population history described and the current number of *WRN4* copies, the median age of the *WRN4* allele is approximately the 94 generations since the population was founded (Graham 1998). We therefore assume a single copy of the allele in a founding Japanese population of 2,000 genes in 350 B.C. (94 generations ago).

Coalescent of the Disease Sample

In this section we present an outline of the generation of the ancestry of the sample of disease alleles. Details of the stochastic processes and probability distributions involved have been developed by Graham (1998) and will be given in another report. The historical pattern of population growth is assumed to be known. This total population is modeled as a continuous-time Moran process (Moran 1962) with additional birth events. Time t is measured backward from the present, in generations (i.e., generations before present [gbp]). The rate of the Moran process is chosen to be $N(t)/2$ events per generation, when the population is $N(t)$ genes in size. This rate gives the same inbreeding effective size (Crow and Kimura 1970) and coalescent rate (Kingman 1982*b*) as a Wright-Fisher population with the same number of genes (Felsenstein 1971). The superimposed birth process has rate $\lambda(t)$, reflecting the rate of population increase at any time t. Note that $\lambda(t)$ does not have to be constant over time.

The disease allele is assumed to be present as a single copy in the population, at a known time T gbp. For the Japanese example described in this article, $T = 94$, the time of the founding of the population (see above). Within the total population, the size, $N_D(t)$, $0 \leqslant t \leqslant T$, of the disease population fluctuates. For a Moran + birth total population and a rare disease allele, the disease-allele count very closely follows a birth-and-death process, with birth rate $1/2 + \lambda(t)$ and death rate 1/2. The advantage of this approximation is that the moment-generating function (and hence the moments) of past numbers can be found, conditional on the number $N_D(t)$ available for a more recent t, and on $N_D(T) = 1$.

In particular, given a single copy of the disease mutation at T and the number of disease copies, the mean, variance, and higher moments of past numbers of disease copies one generation previous may be computed by use of methods analogous to those of Thompson et al. (1992). Thus, the conditional distribution of copy numbers one generation ago may be approximated to arbi-

trary precision by a distribution matched on an appropriate number of moments. For example, given $N_D(0)$ and $N_D(T) = 1$, $N_D(1)$ may be realized. Similarly, given $N_D(1)$ and $N_D(T) = 1$, $N_D(2)$ may be realized. Continuing this process back in time, from 0 to T, gives a sample path for disease-population sizes, $[N_D(t) \mid N_D(0), N_D(T) = 1]$, at one-generation intervals. Figure 1 shows example realizations for the Japanese disease allele, on a $\log_{10}$ scale. For the more recent generations, variability over realizations in disease-allele copy numbers is dominated by temporal changes in the numbers themselves, giving the appearance of deterministic growth (fig. 1*A*). However, figure 1*B* shows that there indeed has been variability in more-recent generations.

Given a current sample of K disease alleles, the ancestral coalescent then can be realized, conditional on the past copy-number realization. Let $k(t)$ be the number of ancestors of the sample at time t, so that $k(0) = K$. The rate of coalescence at time t is

$$\frac{k(t)\,[k(t) - 1]}{2\,[N_D(t) - 1]}\,.$$

This rate differs slightly from the usual coalescent rate of $k(t)\,[k(t) - 1]/[2N(t)]$ for $k(t)$ lineages in a population of size $N(t)$ copies (Felsenstein 1971; Kingman 1982*a*). The difference derives from consideration of the coalescent within a subpopulation (Graham 1998).

Once the coalescent of the disease sample has been realized, recombination events between the disease locus and the marker locus may be placed on the ancestry. A branch of the ancestry of length G generations represents G meioses and, therefore, G opportunities for recombination between the disease locus and the marker. Thus, the probability of at least one recombination event on the branch is $1 - (1 - r)^G \approx 1 - e^{-Gr}$, where r is the recombination frequency between the disease locus and the marker. This overall approach of realizing the coalescent of the sample of disease alleles and then placing recombination events on it is the same as that used by Thompson and Neel (1997). This historical perspective on disequilibrium was first considered by Edwards (1981; also see Arnason et al. 1977; Thompson 1978). The details differ because of the overall population model and the more accurate analysis of the disease-allele population process and also because of the fact that Thompson and Neel (1997) conditioned on only age $[N_D(T) = 1]$ and survival $[N_D(0) > 0]$, rather than on the value of $N_D(0)$.

At a marker locus, we define a "recombinant class" to be a subset of the current sample that is descended from a given recombination event. As shown in figure 2, the recombinant classes form a partition of the sample, with all members within a class being identical by descent at the marker locus. Thus, all members of each recombinant class share a marker allele. Different recombinant classes carry independent alleles at the marker locus. Note that, when recombinant classes are generated, only the presence or absence of recombination events on a given branch of the ancestral tree needs to be considered. For a single marker, multiple recombination events on the branch have the same effect on recombinant classes as a single recombination event. Of course, recombinant classes are latent variables: they cannot be observed.

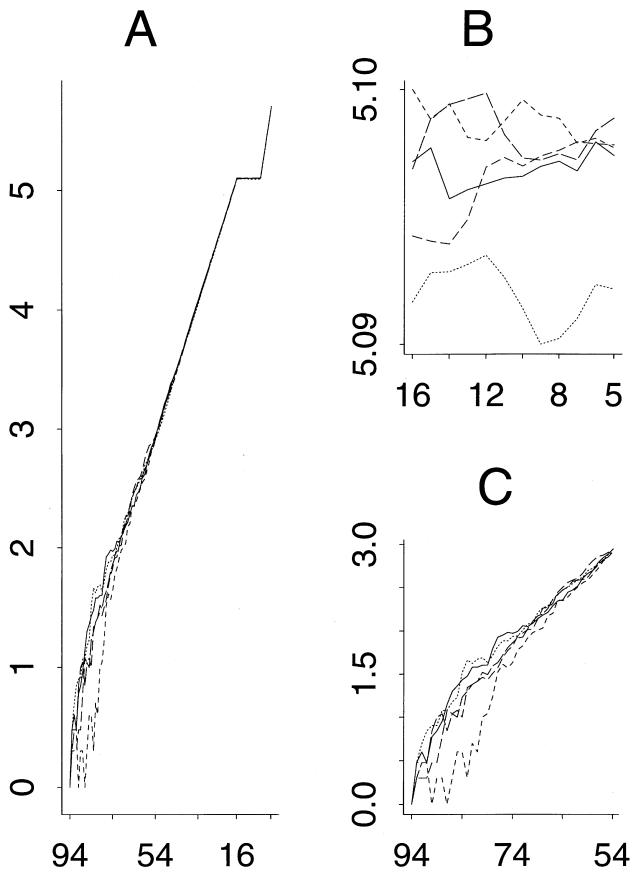

Figure 1 Realizations of past disease-allele copy numbers, for the Japanese example. The vertical axis indicates the $\log_{10}$ scale; the hoizontal axis indicates the generations before present (gbp). *A*, Five sample paths for disease-allele copy number, over all $T = 94$ generations since founding. *B*, Paths during the Edo period, 5–16 gbp. *C*, Paths for the first 40 generations after founding, 94–54 gbp.

Likelihood (Single-Marker Case)

Suppose K haplotypes carrying the disease mutation are sampled from a population for which growth is described by known demographic parameters, $\Delta = \{T, [\lambda(t){:}0 < t \leqslant T]\}$, where time T and population-growth rate $\lambda(t)$ are as defined previously. The recom-

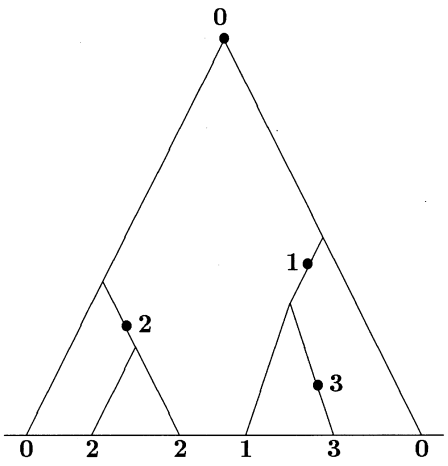

Figure 2 Definition of recombinant classes of the sampled disease haplotypes, with reference to a single linked marker. The ancestral marker allele is denoted by "0," and the bullets labeled "1," "2," and "3" denote subsequent recombination events. There are $K = 6$ disease haplotypes, labeled "0," "2," "2," "1," "3," and "0." Thus, there are four recombinant classes, two of size 1 and two of size 2, and $\boldsymbol{x} = (2, 2)$.

binant-class identifiers on sampled haplotypes can be summarized as $\boldsymbol{X}$, a vector of recombinant-class counts that is indexed by the size of the recombinant class (fig. 2). The element X_i of $\boldsymbol{X}$ is the number of recombinant classes of size i, $1 \leqslant i \leqslant K$.

Consider an m-allele marker at recombination frequency r from the disease locus. The population allele frequencies at the marker locus, $\boldsymbol{q} = (q_1, q_2, \ldots, q_m)$, are assumed to have remained constant over time. Let $\boldsymbol{Y}$ be the vector of marker-allele counts for the sample; Y_j is the number of sampled disease copies carrying marker allele j, $j = 1, \ldots, m$.

Then, the probability for the data $\boldsymbol{Y}$, or the likelihood for r, can be written as

$$L(r) = P_{q,r,\Delta}(\boldsymbol{Y}) = \sum_{\boldsymbol{X}} P_q(\boldsymbol{Y} \mid \boldsymbol{X}) P_{r,\Delta}(\boldsymbol{X}) \ , \qquad (1)$$

where $P_q(\boldsymbol{Y} \mid \boldsymbol{X})$ is the conditional probability of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and where $P_{r,\Delta}(\boldsymbol{X})$ is the probability of $\boldsymbol{X}$. More specifically, $\boldsymbol{X}$ is a function of the coalescent ancestry, $\boldsymbol{A}$, and the recombination events, $\boldsymbol{R}$, on its branches:

$$L(r) = P_{q,r,\Delta}(\boldsymbol{Y}) = \sum_{(\boldsymbol{A},\boldsymbol{R})} P_q[\boldsymbol{Y} \mid \boldsymbol{X}(\boldsymbol{A}, \boldsymbol{R})] P_r(\boldsymbol{R} \mid \boldsymbol{A}) P_\Delta(\boldsymbol{A}) \ . \qquad (2)$$

Equation (1) suggests Monte Carlo evaluation of the likelihood for r, given observed allelic classes $\boldsymbol{Y} = \boldsymbol{y}$, by sampling recombinant classes $\boldsymbol{x}$ from $P_{r,\Delta}(\boldsymbol{X})$ and averaging $P_q(\boldsymbol{y} \mid \boldsymbol{x})$ over the realized values $\boldsymbol{x}$ of $\boldsymbol{X}$. Realizations of $\boldsymbol{X}$ are obtained by simulation, as outlined in the previous section.

Evaluation of the likelihood requires

$$P_q(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{\boldsymbol{C}} P(\boldsymbol{C} \mid \boldsymbol{x}) \ , \qquad (3)$$

where $\boldsymbol{C} = \{c_{ij}\}$ denotes a configuration of recombinant classes consistent with $\boldsymbol{x}$ and $\boldsymbol{y}$ such that c_{ij} recombinant classes of size i are assigned to allele j. Figure 3 shows the notation. The row totals for $\boldsymbol{C}$ are given by $\boldsymbol{x}$, since $x_i = \sum_j c_{ij}$. The column totals are given by $\boldsymbol{n}$, where n_j is the number of recombinant classes assigned to marker allele j. The number y_j of sampled disease haplotypes carrying marker allele j is $y_j = \sum_{i=1}^{K} i c_{ij}$, the inner product of the recombinant-class sizes and the jth column of the matrix $\boldsymbol{C}$. The setup for determining possible configurations $\boldsymbol{C}$ is thus analogous to that for determining possible tables by use of Fisher's exact test (Fisher 1970), except that conditioning is on the row totals $\boldsymbol{x}$ and column inner products $\boldsymbol{y}$ rather than on the row totals $\boldsymbol{x}$ and column totals $\boldsymbol{n}$.

The probability $P(\boldsymbol{C} \mid \boldsymbol{x})$ is obtained as follows. Given $\boldsymbol{x}$, the ith row of $\boldsymbol{C}$ is multinomial, with parameters x_i and $\boldsymbol{q}$, and, thus, has probability

$$\frac{x_i!}{\prod_{j=1}^{m} c_{ij}!} \times \prod_{j=1}^{m} q_j^{c_{ij}} \ .$$

Therefore, the product of these independent multinomial distributions, over the rows (recombinant-class sizes), gives

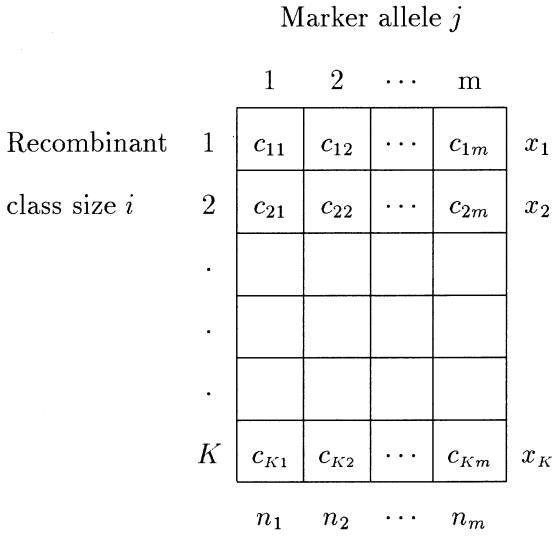

Figure 3 Notation for configuration $C = \{c_{ij}\}$ of recombinant classes. Note that $y_j = \sum_{i=1}^{K} i c_{ij}$.

$$P(\boldsymbol{C} \mid \boldsymbol{x}) = \frac{(\prod_{i=1}^{K} x_i!) \times (\prod_{j=1}^{m} q_j^{n_j})}{\prod_{i=1}^{K} \prod_{j=1}^{m} c_{ij}!} ,$$

so that equation (3) becomes

$$P_q(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{\boldsymbol{C}} P(\boldsymbol{C} \mid \boldsymbol{x}) = \left(\prod_{i=1}^{K} x_i!\right) \times \sum_{\boldsymbol{C}} \frac{\prod_{j=1}^{m} q_j^{n_j}}{\prod_{i=1}^{K} \prod_{j=1}^{m} c_{ij}!} . \quad (4)$$

To evaluate $P(\boldsymbol{y} \mid \boldsymbol{x})$, all tables or configurations $\boldsymbol{C}$ of recombinant classes consistent with $\boldsymbol{x}$ and $\boldsymbol{y}$ are enumerated. A variant of the network algorithm of Mehta and Patel (1983), in which a path through a network represents a consistent table, is used. Details are given in Appendix A.

Statistical uncertainty is measured by variance of estimators or curvature of the observed or expected likelihood surface. However, here there is no explicit form for the likelihood. Instead, a bootstrap approach may be applied, but this also is nontrivial, since there is a single realization of the coalescent process of disease ancestry and, hence, of the underlying recombinant classes. Thus, a nonparametric bootstrap approach, by sampling with replacement from the current disease sample, amounts to bootstrapping one observation and does not reflect the variation in recombinant classes across replicate populations. However, the uncertainty of inferences about the recombination fraction must include all sources of variance in the recombinant classes. We, therefore, adopt a parametric bootstrap approach and generate realizations of $\boldsymbol{Y}$ under the maximum-likelihood estimate (MLE), by reestimating r for each such realization. Confidence intervals for r are defined by dropping down from the maximum of the original likelihood surface. The interval capturing the appropriate percentage of the (parametric) bootstrapped MLEs is selected.

We illustrate the approach with simulated data from the Japanese example, for a sample of $K = 50$ disease haplotypes. Two markers, M_2 and M_3, separated by a recombination frequency of .01, flank the disease locus, which is located at recombination frequency .006 from M_2 and at .01–.006 = .004 from M_3. (For fine-scale mapping such as this, recombination frequencies are effectively additive.) Each marker has four equifrequent alleles. Estimated single-marker likelihood surfaces, calculated over a grid of recombination fractions spaced at 0.1% (.001), are based on 10,000 Monte Carlo replicates. With one user on a Pentium 133-MHz computer, construction of the likelihood surface for 20 recombination frequencies within the range .001–.020 takes ~19 min. Bootstrap confidence intervals are based on 200 bootstrap samples.

The simulated data are $\boldsymbol{y}_2 = (38, 6, 4, 2)$ for M_2 and $\boldsymbol{y}_3 = (6, 2, 41, 1)$ for M_3. LOD scores ($\log_{10}$ likelihood ratios) for M_2 are shown in figure 4. The MLE is $\hat{r} = .005$, with an associated 95% confidence interval of .002–.011. The confidence interval corresponds to an ~1–LOD-unit support interval, which is just over what would be computed on the assumption of a χ^2 approximation to the distribution of minus twice the $\log_e$ likelihood ratio. The likelihood surface for M_3 (results not shown) has more curvature for its MLE of $\hat{r} = .003$ than the likelihood surface for M_2. Increased curvature is expected at lower recombination fractions, because ancestral recombinant classes tend to be larger and sampled marker alleles more dependent. For M_3, the bootstrap 95% confidence interval is .0009–.0085 and corresponds to a LOD-score difference of ~1.3 units. The confidence interval for M_3 thus represents a drop in the LOD that is slightly larger than the confidence interval for M_2, as is expected with increased dependence under the smaller recombination fraction. As the recombination fraction gets smaller, confidence intervals based on the χ^2 approximation are expected to become increasingly too narrow. The χ^2 approximation assumes independence among marker alleles, but positive correlation increases as the recombination frequency decreases.

Interval and Multipoint Mapping

Suppose the disease locus lies in an interval of known length, $s \times 100$ cM, defined by two markers. Recombination events on either side of the disease locus occur independently on the coalescent ancestry $\boldsymbol{A}$. (Interference

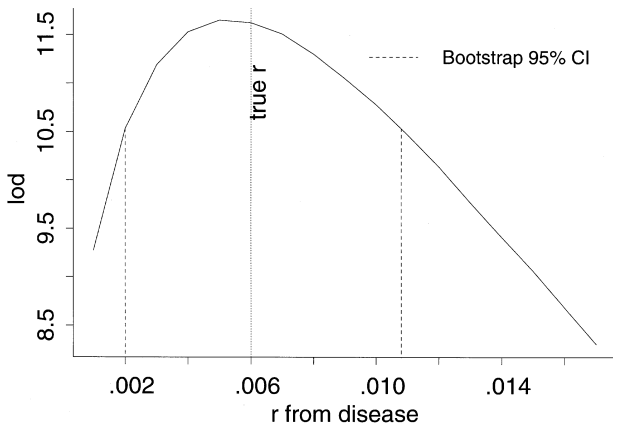

Figure 4 Single-marker LOD-score curve and bootstrap confidence interval for r. There are $K = 50$ disease haplotypes, and LOD scores were estimated from 10,000 Monte Carlo realizations.

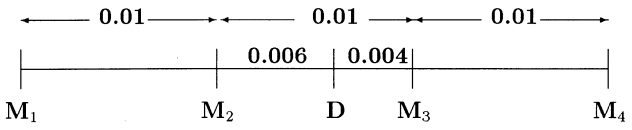

Figure 5 Map of four equispaced markers, M_1–M_4, and disease marker D, for the interval-mapping example.

is effectively irrelevant here, since the chance of recombination with markers on both sides of the disease locus, in a single meiosis, is negligible.) Let r and r^* denote the recombination frequencies between the disease locus and the first and second flanking markers, respectively. Under the scale typical of disequilibrium fine mapping, recombination fractions may be considered additive, so that $r + r^* = s$. Furthermore, under the assumption of an absence of allelic associations between the markers in the total population at the time of the most recent common ancestor of the sample, allelic types at the two markers are independent. Equation (2) becomes

$$L(r,\ r^* = s - r) = P_{q,r,r^*,\Delta}(\boldsymbol{Y},\ \boldsymbol{Y}^*)$$

$$= \sum_{\boldsymbol{A}} \left\{ \sum_{\boldsymbol{R}} P_q[\boldsymbol{Y} \mid \boldsymbol{X}(\boldsymbol{A},\ \boldsymbol{R})] P_r(\boldsymbol{R} \mid \boldsymbol{A}) \right\}$$

$$\times \left\{ \sum_{\boldsymbol{R}^*} P_q[\boldsymbol{Y}^* \mid \boldsymbol{X}^*(\boldsymbol{A},\ \boldsymbol{R}^*)] P_{r^*}(\boldsymbol{R}^* \mid \boldsymbol{A}) \right\} P_\Delta(\boldsymbol{A}) \ , \tag{5}$$

where $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$ are the data at the two markers, $\boldsymbol{X}$ and $\boldsymbol{X}^*$ are the underlying recombinant classes, and $\boldsymbol{R}$ and $\boldsymbol{R}^*$ are the two sets of recombinant events. Thus, simulation and computation may be performed as for the single-marker case.

We illustrate interval mapping with simulated data from the Japanese example. We assume four equispaced markers M_1–M_4, at recombination frequencies $s = .01$ apart, defining three intervals (fig. 5). The disease locus is in the second interval, at recombination frequency $r = .006$ from M_2 and at $r^* = s - r = .004$ from M_3. Each of the four markers has four equifrequent alleles. As before, estimated likelihoods are based on 10,000 Monte Carlo replicates, and bootstrap distributions are based on 200 bootstrap samples. For the $K = 50$ sampled disease haplotypes, the simulated (marginal) data for M_1–M_4 are $\boldsymbol{y}_1 = (9, 12, 24, 5)$, $\boldsymbol{y}_2 = (38, 6, 4, 2)$, $\boldsymbol{y}_3 = (6, 2, 41, 1)$, and $\boldsymbol{y}_4 = (5, 28, 7, 10)$, respectively. Under the assumption of no association between markers in the ancestral population, only the marginal allelic counts for each marker are required for interval mapping.

The procedure correctly identifies the disease interval. The maximum LOD score is at least 5 units above the maxima in adjacent intervals; the correct interval is at least 100,000 × more likely. The disease location is estimated at $\hat{r} = .000$ from its true location, with an associated bootstrap 95% confidence interval of $-.0028$ to .0025. The confidence interval is narrower than those obtained from separate consideration of flanking markers in single-marker mapping, indicating the greater power of interval mapping. The greater power also is reflected in the curvature of the likelihood surfaces at the MLE. Figure 6 shows the LOD-score surface relative to its maximum. For comparison, the single-marker log likelihood with the most curvature at its MLE—that is, the log likelihood for M_3—is also plotted relative to its maximum. The greater power of interval mapping, compared with the single-marker LOD score, is not surprising, given that interval mapping uses information at two markers that flank the disease locus. LOD scores are not plotted at the markers; at each marker, the log likelihood for 0% recombination is $-\infty$, since more than one allelic class is observed.

When flanking markers are not highly polymorphic, additional information on historical recombination events between the disease locus and flanking markers may be gained by joint consideration of several markers on each side of the disease locus. The more polymorphic a marker, the more closely the allelic classes determine the recombinant classes $\boldsymbol{x}$. In the limit, with infinite marker polymorphism, $\boldsymbol{x}$ is observed. Interval mapping is then fully efficient, and multipoint mapping is unnecessary. With regard to multipoint mapping, recombination events in disjointed segments of the chromosome

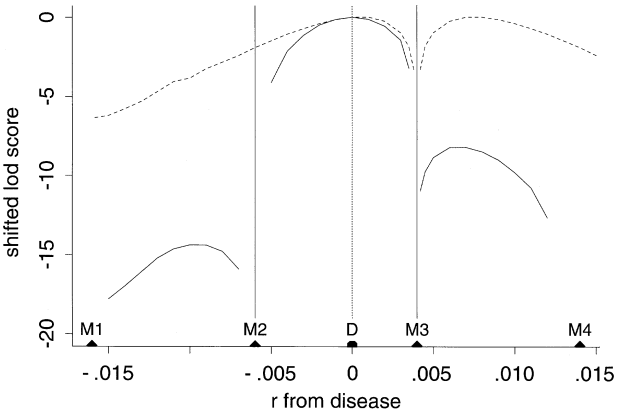

Figure 6 Information for interval versus single-marker mapping. By use of the marker map of figure 5, the interval-mapping LOD score is shown relative to its maximum (*solid line*). For comparison, the single-marker LOD score for the disease with marker M_3 is also shown relative to its maximum (*dashed line*). There are $K = 50$ disease haplotypes, and LOD scores were estimated from 10,000 Monte Carlo realizations.

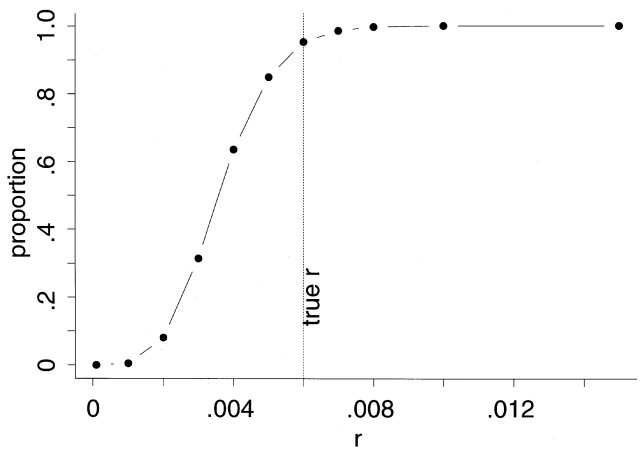

Figure 7 Proportion of realized $\boldsymbol{x}$ compatible with observed $\boldsymbol{y}$ = (38, 6, 4, 2), as a function of hypothesized r. There are $K = 50$ disease haplotypes, the true r value is .006, and the marker has four equifrequent alleles. The curve is based on 10,000 realizations of $\boldsymbol{x}$ at each value of r.

may be placed independently on the ancestral coalescent. However, in contrast to interval mapping (eq. [5]), the sets of single-marker recombinant classes are now dependent, given the ancestral coalescent $\boldsymbol{A}$. Recombination events between the disease locus and a flanking marker are a subset of those between the disease locus and a farther marker, and thus the recombinant classes for the farther marker partition those for a closer marker.

Multipoint mapping thus requires extension of the concept of a recombinant class. The defining principle is analogous: disease haplotypes in the same multimarker recombinant class share a common ancestor at all the markers. As for single-marker mapping, recombinant classes are obtained by placing recombination events on the realized ancestral tree, for each disjointed chromosome segment defined by the putative disease locus and the markers. The approach is outlined in Appendix B. Assignment of marker alleles to the resulting recombinant classes includes consideration of haplotype frequencies. As for a single marker, we make the simplifying assumption that population marker-allele and haplotype frequencies have remained constant since the disease mutation was introduced into the population.

Monte Carlo Properties

For the method presented in this article, in which coalescents of the disease sample and then the recombinant classes are realized independently, the Monte Carlo error is assessed readily and is primarily a function of Monte Carlo sample size. Note, however, that the effective Monte Carlo sample size for likelihood $L(r)$ varies with the value of r.

If the hypothesized location of the disease gene is much closer to a marker than is the true location, many of the recombinant-class realizations simulated at the hypothesized location will be inconsistent with the data $\boldsymbol{y}$, because of an ancestral recombinant class that is larger than the largest allelic class. Thus, for example, if 10,000 realizations of $\boldsymbol{x}$ are used at each location at which a likelihood is to be estimated, sample sizes close to 10,000 will be realized only for locations in the neighborhood of the true recombination fraction or for locations with recombination frequencies that are larger than those that are true (fig. 7). Figure 7 shows the number of compatible realized pairs $(\boldsymbol{x}, \boldsymbol{y})$, as a function of the hypothesized r, for one such Monte Carlo simulation for the Japanese example. The data $\boldsymbol{y}$ = (38, 6, 4, 2) are those considered previously (fig. 4). For values of r that are substantially smaller than the true value, only a small fraction of the 10,000 Monte Carlo realizations result in $\boldsymbol{x}$ values compatible with data $\boldsymbol{y}$.

However, the Monte Carlo error in estimated likelihoods is of most concern in the neighborhood of the maximum, where there are few incompatibilities, rather than for the smallest values of r, where the likelihood is small. Figure 8 shows the Monte Carlo standard error of the estimated likelihood, in comparison with the estimated likelihood value, for the above example. The estimated $\log_{10}$ likelihood is also shown. The relative Monte Carlo error is minimized in the neighborhood of the MLE at $\hat{r} = .005$. In absolute terms, the error in the likelihood at the maximum is $\sim 5 \times 10^{-6}$. The MLE at

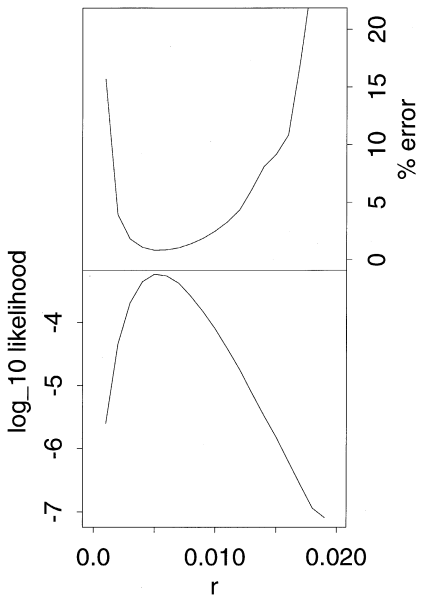

Figure 8 Relative Monte Carlo error as a percentage of the estimated likelihood, and, for comparison, the estimated $\log_{10}$ likelihood.

$\hat{r} = .005$ differs by 1.4×10^{-4} from the likelihood at the adjacent $r = .004$ on the search grid and by 3.5×10^{-5} from the estimated likelihood at $r = .006$. The Monte Carlo error in the MLE is, therefore, an order of magnitude smaller than the differences between estimated likelihood values in the neighborhood of the maximum. These results indicate that 10,000 Monte Carlo replicates are sufficient to locate the MLE to resolution .001 used in the grid search.

Another aspect of the precision and efficiency of Monte Carlo is the choice of the Monte Carlo paradigm. We have chosen to realize recombinant classes $\boldsymbol{x}$ and to compute $P(\boldsymbol{y} \mid \boldsymbol{x})$ analytically. Alternatively, allelic classes $\boldsymbol{Y}$, at present or in the past, can be realized and scored when they are consistent with observed data (Rannala and Slatkin 1998). In general, the more that can be computed analytically, the smaller the Monte Carlo error. To illustrate, we compare Monte Carlo sampling of recombinant classes to scoring of present allelic classes. The standard error of an estimator of a Monte Carlo likelihood based on scoring is $E = [P^*(1 - P^*)/S]^{1/2}$, where P^* is the likelihood value and S is the Monte Carlo sample size. For a given degree of desired accuracy, E, this may be solved for the required number of allelic-class replicates, S. When the observed data are $\boldsymbol{y} = (38, 6, 4, 2)$, $\sim 2.5 \times 10^7$ allelic-class replicates are required to achieve the same degree of accuracy in the neighborhood of the MLE as 10,000 recombinant-class replicates. Standard errors and estimated likelihoods from Monte Carlo sampling of recombinant classes are used in this calculation. When both Monte Carlo sampling schemes are calibrated to have the same degree of accuracy, the run based on scoring takes ~3,150 × longer. The scoring method used for this comparison generates allelic classes $\boldsymbol{y}$ by randomly assigning the underlying recombinant classes to alleles. As marker polymorphism increases, Monte Carlo sampling of recombinant classes approaches rejection sampling of allelic classes.

Replicates in our Monte Carlo approach correspond to a given coalescent ancestry, $\boldsymbol{A}$, that is recycled over recombination frequencies, r, in the search grid. Recycling is possible because the recombination process does not enter into the generation of ancestral coalescents (eq. [2]). Computation time is saved by recycling, and it may be possible to economize further by elimination of the generation of tree topologies, F, for which $\boldsymbol{A} = (F, \boldsymbol{t})$. Sampling of coalescent times, $\boldsymbol{t}$, and averaging of conditional probabilities of observed allelic classes $\boldsymbol{Y}$, given only these times, would be a Monte Carlo paradigm that would require fewer replicates than one sampling either recombinant classes $\boldsymbol{X}$ or allelic classes $\boldsymbol{Y}$. The form

$$L(r) = P(\boldsymbol{Y}) = \int_t P(\boldsymbol{Y} \mid \boldsymbol{t}) P(\boldsymbol{t}) d\boldsymbol{t}$$

may be contrasted both with our equation (2) and with the final equation on page 463 in the report by Rannala and Slatkin (1998). Since each pair of extant ancestral lineages has the same probability of coalescing, the topology F of the ancestral tree has a parameter-free distribution. The results reported by Harding (1971) on the probability distributions of unlabeled tree topologies could provide a basis for analytic evaluation of the conditional probability of observed allelic classes, given coalescent times. In principle, these probabilities of unlabeled tree topologies would be used to reweight the probability of allelic classes, given a tree (topology and coalescent times) in a sum over all possible topologies:

$$P(\boldsymbol{Y} \mid \boldsymbol{t}) = \sum_F P[\boldsymbol{Y} \mid \boldsymbol{A} = (F, \boldsymbol{t})] P(F) \ .$$

However, exact calculation of $P(\boldsymbol{Y} \mid \boldsymbol{t})$ currently seems impractical, since there are many unlabeled topologies F over which to sum, even for a disease sample of moderate size. Moreover, evaluation of allelic-class probabilities for any given ancestry $\boldsymbol{A}$ is computer intensive, since summation over recombination events $\boldsymbol{R}$ would be required.

Comparison with Other Methods: An Example

We use the data on DTD, reported by Hästbacka et al. (1992), to compare our method to other disequilibrium-mapping methods. DTD is an autosomal recessive disease with an allele frequency of ~.8% in the Finnish population of $\sim 10^7$ chromosomes. A single ancestral mutation is thought to account for >95% of the DTD chromosomes in Finland. The disease gene initially was localized to a 2-cM region between *RPS14* and *D5S372* on chromosome 5q, by use of data from pedigree studies. Hästbacka et al. (1990, 1992) used markers within the candidate region, for disequilibrium fine mapping. Striking population associations were observed for *Sty*I and *Eco*RI, two restriction-fragment–length polymorphisms in the *CSF1R* gene. Of 152 DTD haplotypes, 144 had both a *Sty*I and an *Eco*RI restriction site (the 1-1 haplotype), 1 had the *Sty*I restriction site only (the 1-2 haplotype), 0 had the *Eco*RI site only (the 2-1 haplotype), and 7 had neither site (the 2-2 haplotype). Estimated population frequencies for these *CSF1R* haplotypes were $q = (.03, .23, .06, .68)$, respectively.

Following Hästbacka et al. (1992), we assume that the Finnish population was founded by ~1,000 individuals, $t_f = \sim 100$ gbp, and that the population has grown exponentially, at constant rate $\lambda = .085$. We also assume

a single copy of the disease allele at population founding. The two restriction sites are treated as a single marker with four alleles. The marker allelic data therefore are $\mathbf{y} = (144, 1, 0, 7)$. Single-marker mapping yields an MLE of disease location at $\hat{r} = .00085$ from *CSF1R*. The support interval for $\hat{r}$, defined as the values of r with likelihoods within 2 $\log_e$ units of the maximum, is .0002–.0018. Estimates and support intervals are robust to the disease mutation being introduced into the population before founding. For instance, when a single copy of the disease mutation is assumed to be present at 115 gbp, rather than at 100 gbp, the estimated disease location is at $r = .00080$ from *CSF1R,* and the support interval is unchanged. In fact, 115 generations is the estimated median age of the mutation, on the basis of the current number of disease alleles (Griffiths and Tavaré 1998) and on the assumption of a constant-sized population of 2,000 copies, prior to 100 gbp. The estimated probability that the DTD mutation is older than 100 generations is .90.

Rannala and Slatkin (1998) collapsed haplotype categories for the restriction sites, to obtain diallelic data, and used these data to estimate the disease location at $\hat{r} = .00080$ from *CSF1R,* with a support interval of .0002–.0016. Kaplan et al. (1995) reported a support interval with an upper limit at $r = .0021$ from the combined restriction sites. Hästbacka et al. (1992) estimated the disease location at $r = .00064$ from *CSF1R*. All estimators are based on single-marker mapping, and all are in good agreement with the actual physical distance, ~70 kb from *CSF1R*.

We also applied interval mapping to DTD, combining the data on *CSF1R* with data on the diallelic flanking markers *D5S372* and *RPS14,* reported by Hästbacka et al. (1994). The genomic region is ~2 cM, or $r = .02$, in length, with *D5S372* and *CSF1R* separated by a recombination fraction of $r =\sim .00825$ and with *CSF1R* and *RPS14* separated by $r =\sim .01175$. The disease locus lies in the interval defined by *D5S372* and *CSF1R,* that is, $r =\sim .00075$ from *CSF1R*. The maximum LOD score for the interval containing the disease gene is 40 units higher than that for the other interval. Within the region *D5S372–CSF1R,* the disease location is estimated to be $r = .00075$ from *CSF1R,* and the associated support interval is .0002–.0017. In this case, there is little gain, over single-marker mapping, in the precision of the location estimate. The real gain is in the knowledge that the disease locus is much more likely to lie within the region *D5S372–CSF1R,* rather than within *CSF1R–RPS14*. Less precision is gained in the estimate of the disease location because the disease locus is associated much more strongly with *CSF1R* than with *D5S372*. The distance between the disease locus and *CSF1R* is <10% of the total intermarker distance between *D5S372* and *CSF1R*. Moreover, *CSF1R* is more polymorphic than *D5S372,* and the disease mutation appears to have been introduced into the population on a haplotype with an uncommon *CSF1R* allele.

Discussion

In this article, we present a likelihood-based method for fine-scale mapping of a rare monophyletic disease. The method is based on a coalescent model of the ancestry of a sample of disease alleles. Information on the demographic history of the population—including varying growth rates, such as those for the Japanese—may be incorporated. Generations on this ancestral coalescent correspond to meioses at which recombination events may occur. Recombination events on disease-bearing ancestral haplotypes determine the recombinant classes, which specify marker identity by descent in the sample. By use of our methods, large Monte Carlo samples of recombinant classes for various hypothesized trait gene locations are readily obtained. We combine these Monte Carlo samples with an analytic method for computation of the probability of observed disease haplotypes, conditional on latent recombinant classes, to obtain a likelihood for fine-scale mapping. The method extends easily from single-marker mapping to interval mapping. Interval mapping has been shown to perform substantially better than single-marker mapping, by correctly identifying both the interval containing the disease gene and the disease location, for our example of simulated data. Further extensions to multipoint mapping also are discussed. Throughout, confidence bounds on likelihood estimates of disease location are obtained by use of the parametric bootstrap.

Of the methods described in three recent articles, that of Xiong and Guo (1997) does not condition on the current disease-allele count, and that of Kaplan et al. (1995) realizes this count and rejects the realizations that are not within an acceptable range. Our method uses the approach, shared by Rannala and Slatkin (1998), of realizing the ancestry of a sample by conditioning on a current disease-allele count. Our method of realizing coalescent ancestry results in more of an approximation than that of Rannala and Slatkin (1998), since it relies on realizing in single-generation steps, on the basis of the conditional moments of the ancestral disease population. However, these moments can be computed for any assumed ancestral demographic growth pattern; thus, our demographic model can be more flexible. Our method does assume a single ancestral lineage at a given time *T*. However, Rannala and Slatkin (1998) suggested that disequilibrium likelihoods are robust to assumptions regarding the age of the disease allele. Also, the

probability distribution of T may be investigated, again given the flexible demographic assumptions.

Once coalescent ancestry is realized, our approach differs significantly from that of Rannala and Slatkin (1998). They sampled the marker allelic classes existing immediately after the most recent coalescent event, which is closer in spirit to sampling and scoring present allelic classes than is our approach of realizing underlying recombinant classes and computing analytically the probability of observed data, given the recombinant-class configuration. The realization of recombinant classes makes the use of multiallelic markers straightforward. Indeed, the performance and feasibility of our method do not depend significantly on marker polymorphism, whereas Rannala and Slatkin (1998) considered only diallelic markers. Also, Rannala and Slatkin (1998) considered only likelihoods for a single marker. Xiong and Guo (1997) used multiple markers in their approximate likelihoods but treated the information from each marker independently, ignoring haplotype information. Kaplan et al. (1995) realized marker-haplotype frequencies in the total disease population and then computed the probabilities of the observed sample of disease haplotypes, given these frequencies. In principle, their approach permits likelihoods to be based on multiple polymorphic markers, although the Monte Carlo efficiency will decrease rapidly as the possible diversity of haplotypes increases. Our coalescent-based method also extends readily to interval mapping and, by extension of the definition of the latent recombinant classes, also will permit multipoint mapping. However, the statistical gains and computational costs of multipoint disequilibrium mapping are uncertain and remain to be investigated. All these reports, including ours, discuss only disequilibrium mapping for a rare allele of monophyletic origin. Linkage detection for complex traits, by use of disequilibrium methods, is also an issue that will require future investigation.

Acknowledgments

This research was supported, in part, by National Institutes of Health grants DK-42654 (to J.G.) and GM-46255 (to E.A.T.). We are grateful to Ellen Wijsman and Katrina Goddard for their discussions about the Werner syndrome example and to Joe Felsenstein for his many helpful comments.

Appendix A

Evaluation of $P_q(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})$

This appendix gives details of the network algorithm used to evaluate $P_q(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})$. Mehta and Patel (1983) developed a network algorithm to compute P values for Fisher's exact test by use of contingency tables. Their application requires enumeration of all tables consistent with given marginal totals; these consistent tables are efficiently represented as paths through the network. Our application is similar, but the conditioning is on row totals x_i and column inner products $y_j = \sum_{i=1}^{K} ic_{ij}$, rather than on row totals x_i and column totals n_j (fig. 3). An additional difference is that, with our application, the sampled recombinant classes $\boldsymbol{x}$ may be impossible to assign to alleles in a manner compatible with the observed allelic classes $\boldsymbol{y}$. For example, the size of the largest recombinant class could be larger than the size of the largest allelic class, in which case $P_q(\boldsymbol{y} \mid \boldsymbol{x}) = 0$.

The number of paths through a network can be much larger than the number of nodes. For instance, one network for a diallelic marker had 11,774,790 paths but only 524 nodes. Thus, a network is an efficient representation of the often large number of tables consistent with marginal quantities. The nodes represent updated column margins; specifically, these are the recombinant classes remaining to be assigned to an allelic type. The edges represent the columns of cells in the table—that is, the counts c_{ij} of recombinant classes of size i that are assigned to a given allelic type j. A convenient feature of the network representation is that tables are easily extracted by traversing paths. As a path is traversed, the probability of the table is determined by multiplication of predetermined edge weights (see eq. [4]).

Consider a simple example for a three-allele marker with allelic classes of size $y_1 = 9$, $y_2 = 7$, and $y_3 = 6$ and with corresponding allele frequencies $\boldsymbol{q} = (q_1, q_2, q_3)$. Suppose that there are $x_1 = 2$ recombinant classes of size 1, $x_2 = 1$ recombinant class of size 2, and $x_3 = 6$ recombinant classes of size 3. Thus, $\boldsymbol{x} = (2, 1, 6)$ and $\boldsymbol{y} = (9, 7, 6)$. The three possible tables associated with $\boldsymbol{x}$ and $\boldsymbol{y}$ are shown in figure A1.

To obtain the possible tables, recombinant classes are assigned to small allelic classes before they are assigned to large allelic classes. Within any allelic class, the largest recombinant classes are assigned first. On the assumption that allelic classes in a table are listed in decreasing order of size, the network is built by starting at the bottom right-hand corner of a table and working up and to the left. Figure A2 shows the network for the example.

Steps of the Algorithm

Let $\boldsymbol{x}^*$ be the current-row margin and $\boldsymbol{y}^*$ the current-column margin. The network is built from the right (or root), starting with $\boldsymbol{x}^* = \boldsymbol{x}$ and $\boldsymbol{y}^* = \boldsymbol{y}$. Place $\boldsymbol{x}^*$ in the network as the initial or root node. For the example, $\boldsymbol{x}^* = (2, 1, 6)$ is the root node, and $\boldsymbol{y}^* = (9, 7, 6)$. Suppose that the largest recombinant class is of size I and that the marker has m alleles. For the example, $I = 3$

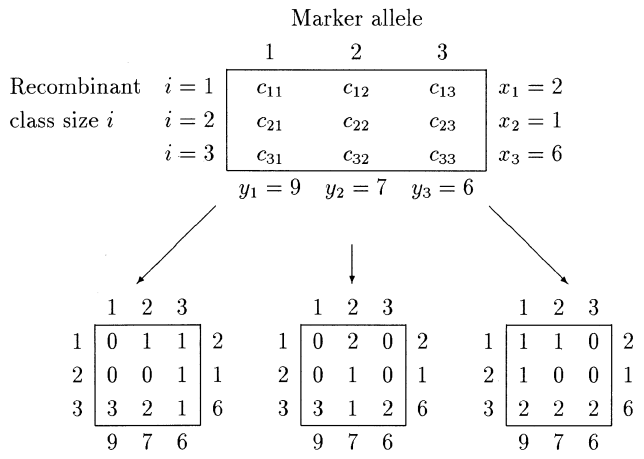

Figure A1 General configuration table, and three possible configurations for the example with $\boldsymbol{x} = (2, 1, 6)$ and $\boldsymbol{y} = (9, 7, 6)$. The network is shown in figure A2.

and $m = 3$. Starting with $i = I$ and $j = m$, the steps are as follows:

1. Compute a feasible range for c_{ij}. For the example, the feasible range for c_{33} is (1, 2). Details on computing the feasible range are given below.
2. Select a possibility from within this range for c_{ij}. For instance, suppose that $c_{33} = 1$ is selected.
3. Update $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$: $x_i^* \rightarrow x_i^* - c_{ij}$, and $y_j^* \rightarrow y_j^* - i \times c_{ij}$. For instance, after $c_{33} = 1$ is selected, $\boldsymbol{x}^*$ changes from (2, 1, 6) to (2, 1, 5), and $\boldsymbol{y}^*$ changes from (9, 7, 6) to (9, 7, 3).
4. Move up the column: $i \rightarrow i - 1$. Repeat steps 1, 2, and 3 for the new (i, j). For instance, after selection of c_{33}, move up the third column, to select c_{23}.
5. When the top of a column of the table has been reached ($i = 1$), place $\boldsymbol{x}^*$ in the network as a node, and connect it to the previous node that gave rise to it. If $\boldsymbol{x}^*$ is already in the network, do not add it again; just make the connection to the previous node. In the example, the choice $c_{33} = 1$ gives rise to unique possibilities, $c_{23} = c_{13} = 1$, with an associated network node $\boldsymbol{x}^* = (1, 0, 5)$ after completion of the column $i = 3$ (fig. A2).
6. Go to the next column of the table: $i \rightarrow I$ and $j \rightarrow j - 1$. For instance, after the top of the column is reached and c_{13} has been selected, move left in the table to the bottom of the second column and select c_{32}.
7. Repeat steps 1–5 for the new column.
8. When the table is completed ($i = j = 1$; $\boldsymbol{x}^* = [0, 0, 0]$) or when the selected path terminates prematurely, return to the last cell in the table at which unexplored values in the feasible range exist and repeat the steps, beginning at step 2.

In the example, the choice $c_{33} = 1$ determines a single feasible path and configuration. Returning to $i = j = 3$, the alternate choice, $c_{33} = 2$ also gives rise to unique possibilities $c_{23} = c_{13} = 0$ and to the third-column associated node $\boldsymbol{x}^* = (2, 1, 4)$. There then are two possible choices for c_{32}, each providing a feasible configuration (fig. A2).

Once a network has been constructed, the difference in elements of nodes flanking each edge defines the appropriate column of a table. The three paths in figure A2, from the root to the terminal node, correspond to the three possible tables, for which columns can be read off the edges. The third column has two possible assignments, (1, 1, 1) and (0, 0, 2). If the third column is (1, 1, 1), then the second column is (1, 0, 2). If the third column is (0, 0, 2), then the second column is either (2, 1, 1) or (1, 0, 2). If the third column is (1, 1, 1) and the second column is (1, 0, 2) or if the third column is (0, 0, 2) and the second column is (2, 1, 1), then the first column is (0, 0, 3). If the third column is (0, 0, 2) and the second column is (1, 0, 2), the first column is (1, 1, 2). The probability of a table, up to the multiplicative constant $\prod_{i=1}^{K} x_i! = \prod_{i=1}^{I} x_i!$, is determined by the product of the edge weights along the path, for which the weight for an edge corresponding to the jth column is $q_j^{n_j} \Big/ \prod_{i=1}^{K} c_{ij}! = q_j^{n_j} \Big/ \prod_{i=1}^{I} c_{ij}!$ (eq. [4]). Summation of these probabilities, over all paths, results in $P_q(\boldsymbol{y} \mid \boldsymbol{x})$ (eq. [3]).

A Feasible Range of Values for c_{ij}

The number of recombinant classes of size i assigned to the jth marker allele has four restrictions:

1. The number of disease copies $i \times c_{ij}$ defined by the cell can be no more than the size y_j^* of the appropriate allelic class; that is, $i \times c_{ij} \leqslant y_j^*$.
2. The number of recombinant classes c_{ij} of a given size i assigned to marker allele j can be no more than the number of recombinant classes x_i^* of that size; that is, $c_{ij} \leqslant x_i^*$.
3. The number of disease copies $i \times x_i^* - i \times c_{ij}$ remaining for the rest of row i to the left of cell c_{ij} (i.e., for columns $1 \ldots j - 1$ of row i) can be no more than the pooled size $\sum_{l=1}^{j-1} y_l^*$ of the corresponding allelic classes; that is, $i \times x_i^* - i \times c_{ij} \leqslant \sum_{l=1}^{j-1} y_l^*$.
4. The number of disease copies $y_j^* - i \times c_{ij}$ remaining for the rest of column j above cell c_{ij} (i.e., for rows $1 \ldots i - 1$ of column j) can be no more than the number of copies $\sum_{l=1}^{i-1} l \times x_l^*$ defined by the corresponding recombinant class sizes; that is, $y_j^* - i \times c_{ij} \leqslant \sum_{l=1}^{i-1} l \times x_l^*$.

The first two restrictions imply that the upper bound of the feasible range for c_{ij}, given above, is $c_{ij} \leqslant \min(y_j^*/i, x_i^*)$, whereas the second two restrictions imply that the lower bound is

$$c_{ij} \geqslant \max\left[x_i^* - \frac{1}{i}\sum_{l=1}^{i-1} y_l^*,\ \frac{1}{i}\left(y_j^* - \sum_{l=1}^{i-1} l \times x_l^*\right)\right].$$

Thus, in the example, the feasible range for c_{33} is (1, 2), since

$$c_{33} \leqslant \min\left(\frac{6}{3},\ 6\right) = 2$$

and

$$c_{33} \geqslant \max\left\{6 - \frac{1}{3} \times (9 + 7),\ \frac{1}{3} \times [6 - (1 \times 2 + 2 \times 1)]\right\} = \max\left(\frac{2}{3},\ \frac{2}{3}\right) = \frac{2}{3}\ .$$

Provided that $\boldsymbol{x}$ and $\boldsymbol{y}$ are compatible, at least one path in the network will terminate.

Assignment to cells of values within the feasible range, in order of largest to smallest recombinant class and smallest to largest allelic class, limits but does not eliminate nonterminating paths in the network. A small example, with $\boldsymbol{x} = (2, 1, 2)$ and $\boldsymbol{y} = (4, 4, 2)$, is shown in figure A3. It is necessary that $c_{33} = 0$, but then the feasible range for c_{23} is (0, 1). The choice $c_{23} = 1$ leads to a path reaching the terminal node and a feasible configuration (fig. A3*A*), but $c_{23} = 0$ leads to an empty feasible range at $i = j = 2$ (fig. A3*B*). In contrast, all paths of a network in the method of Mehta and Patel (1983) reach the terminal node, because the application requires that singletons, rather than recombinant classes, be assigned to cells.

(A) $c_{23} = 1$:

		Marker allele 1	2	3	
Recombinant class size i	$i = 1$	$c_{11} = 1$	$c_{12} = 1$	$c_{13} = 0$	$x_1 = 2$
	$i = 2$	$c_{21} = 0$	$c_{22} = 0$	$c_{23} = 1$	$x_2 = 1$
	$i = 3$	$c_{31} = 1$	$c_{32} = 1$	$c_{33} = 0$	$x_3 = 2$
		$y_1 = 4$	$y_2 = 4$	$y_3 = 2$	

(B) $c_{23} = 0$:

		Marker allele 1	2	3	
Recombinant class size i	$i = 1$	$c_{11} = -$	$c_{12} = -$	$c_{13} = 2$	$x_1 = 2$
	$i = 2$	$c_{21} = -$	$c_{22} = \emptyset$	$c_{23} = 0$	$x_2 = 1$
	$i = 3$	$c_{31} = -$	$c_{32} = 1$	$c_{33} = 0$	$x_3 = 2$
		$y_1 = 4$	$y_2 = 4$	$y_3 = 2$	

Figure A3 Configuration tables showing a case for which there is (*A*) one nonterminating path ($c_{23} = 1$) and (*B*) one terminating path ($c_{23} = 0$). The data are $\boldsymbol{y} = (4, 4, 2)$, and the recombinant-class vector is $\boldsymbol{x} = (2, 1, 2)$. A zero with a slash (Ø) indicates an empty feasible range, and a minus sign (−) indicates an undefined feasible range.

Appendix B

Multipoint Recombinant Classes

For simplicity, we consider the case of two markers, M_1 and M_2, separated by a known recombination fraction s and located to one side of the disease locus (fig. B1). However, this approach extends to any number of markers. Let M_1 be the marker that is closest to the disease locus. Suppose that the disease locus and M_1 are separated by the unknown recombination fraction r.

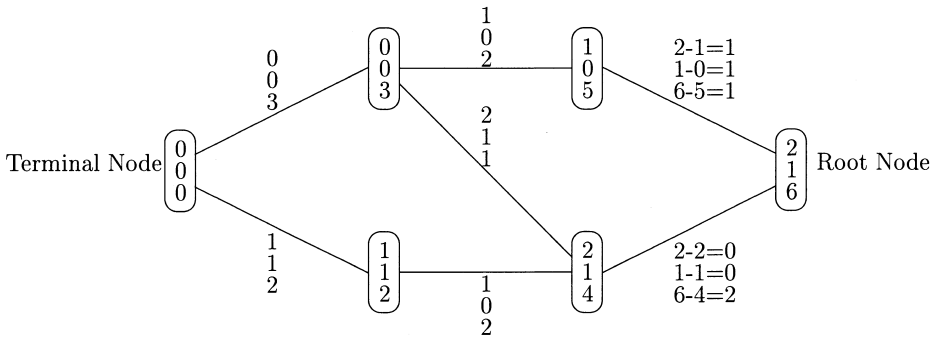

Figure A2 Network diagram for the example with $\boldsymbol{x} = (2, 1, 6)$ and $\boldsymbol{y} = (9, 7, 6)$. The corresponding configurations are shown in figure A1.

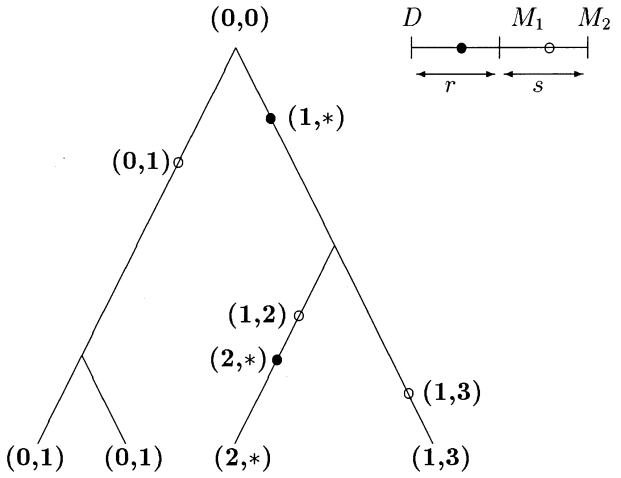

Figure B1 Definition of multipoint recombinant classes, for the case of two markers, M_1 and M_2, on one side of the disease locus D. Recombination events between the disease locus and M_1 are indicated by blackened circles, and those between M_1 and M_2 are indicated by unblackened circles. There are $K = 4$ sampled disease haplotypes, with recombinant-class identifiers (0, 1), (0, 1), (2, *), and (1, 3).

Since M_2 is farther from the disease locus than M_1, recombinant classes for M_2 partition those for M_1. M_2 thus can provide information about r when the underlying M_1 recombinant classes are poorly defined. When the M_1 recombinant classes are well defined, such as when M_1 is highly polymorphic, M_2 provides little or no information about r. For the extreme case in which M_1 is infinitely polymorphic, each recombinant class is of a different allelic type. The recombinant-class sizes for M_1 are then observed, and no extra information about r is gained from the M_2 recombinant classes.

To obtain a realization z of the joint recombinant-class information $\boldsymbol{Z}$, for M_1 and M_2, recombination events between both the disease locus and M_1 and between M_1 and M_2 are placed on the ancestral tree, as shown in figure B1. Each type of event is indexed in the order in which it occurs as the ancestral tree is traversed forward in time, from the root to the tips. The most recent common ancestral haplotype of the K sampled copies is denoted by the haplotype vector (0, 0). The first element of the vector corresponds to M_1 and the second element to M_2. Subsequent haplotypes, formed by recombination events on the ancestral tree, are recorded in haplotype vectors and are coded in the order of the recombination events defining them. When a recombination event occurs between the disease allele and M_1, a new (M_1, M_2) haplotype joins the disease allele. To indicate this, the M_2 element of the haplotype vector is coded with the wild-card symbol (*). The resulting joint recombinant-class information Z_i for each sampled copy i is given at the tips of the ancestral tree. In the example shown in figure B1, there are two joint recombinant classes, (2, *) and (1, 3), which are of size 1, and one recombinant class, (0, 1), which is of size 2. The recombinant class (2, *) is assigned alleles based on the (M_1, M_2) haplotype frequencies. On the other hand, since recombination events between M_1 and M_2 imply independent allele status at each marker locus, allelic assignment for classes (0, 1) and (1, 3) is based on the marginal allele frequencies at each marker. The two-marker likelihood for r,

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{z} P(\boldsymbol{y} \mid z)P(z) \ ,$$

is analogous to the single-marker likelihood, for which $\boldsymbol{y}$ is now the observed table of haplotype counts $\boldsymbol{Y}$, with $P(\boldsymbol{y} \mid z)$ evaluated analytically.

References

Arnason A, Larsen B, Marshall WH, Edwards JH, MacIntosh P, Olaisen B, Teisberg P (1977) Very close linkage between HLA-B and Bf inferred from allelic association. Nature 268: 527–528

Benedict R (1989) The chrysanthemum and the sword. Houghton Mifflin, Boston

Boehnke M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 55:379–390

Cox T, Kerem B, Rommens J, Iannuzzi M, Drumm M, Collins F, Dean M, et al (1989) Mapping of the cystic fibrosis gene using putative ancestral recombinants. Am J Hum Genet Suppl 45:A136

Crow J, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York

Edwards J (1981) Allelic association in man. In: Eriksson AW (ed) Population structure and genetic disorders: proceedings of the 7th Sigfred Juselius Foundation Symposia. Academic Press, New York, pp 239–256

Felsenstein J (1971) The rate of loss of multiple alleles in finite haploid populations. Theor Popul Biol 2:391–403

Fisher R (1970) Statistical methods for research workers, 14th ed. Oliver & Boyd, Edinburgh

Goddard KA, Yu CE, Oshima J, Miki T, Nakura J, Piussan C, Martin GM, et al (1996) Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1–21.1 markers. Am J Hum Genet 58: 1286–1302

Graham J (1998) Disequilibrium fine-mapping of a rare allele via coalescent models of gene ancestry. PhD thesis, Department of Biostatistics, University of Washington

Griffiths R, Tavaré S (1998) The age of a mutation in a general coalescent tree. Stoch Models 14:273–295

Hanihara K (1991) Dual structure model for the population history of the Japanese. Jpn Rev 2:1–33

Harding E (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Prob 3:44–77

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Hästbacka J, de la Chapelle A, Mahtani M, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, et al (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

Hästbacka J, Kaitila I, Sistonen P, de la Chapelle A (1990) Diastrophic dysplasia gene maps to the distal long arm of chromosome 5. Proc Natl Acad Sci USA 87:8056–8059

International Society for Educational Information (ISEI) (1998) Teachers' and textbook writers' handbook on Japan, http://www.isei.or.jp/books/66/isei_66_contents.html

Japan Information Network (JIN) (1998) Census of Japan, http://www.jinjapan.org/stat/index-f.html

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kingman JFC (1982*a*) The coalescent. Stoch Process Appl 13: 235–248

——— (1982*b*) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in probability and statistics. North-Holland, Amsterdam, pp 97–112

Koyama S (1979) Jomon subsistence and populations. Senri Ethnological Stud 2:1–65

Matsumoto T, Imamura O, Yamabe Y, Kuromitsu J, Tokutake Y, Shimamoto A, Suzuki N, et al (1997) Mutation and haplotype analyses of the Werner's syndrome gene based on its genomic structure: genetic epidemiology in the Japanese population. Hum Genet 100:123–130

Mehta CR, Patel NR (1983) A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J Am Stat Assoc 78:427–434

Moran P (1962) The statistical processes of evolutionary theory. Clarendon Press, Oxford

Rannala B, Slatkin M (1998) Likelihood analysis of disequilibrium mapping, and related problems. Am J Hum Genet 62:459–473

Rose M (1996) The peopling of Japan. Archaeology 48:43

Snell R, Lazarou L, Youngman S, Quarrell O, Wasmuth J, Shaw D, Harper P (1989) Linkage disequilibrium in Huntington's disease: an improved localisation for the gene. J Med Genet 26:673–675

Theilmann J, Kanani S, Shiang R, Robbins C, Quarrell O, Huggins M, Hedrick A, et al (1989) Nonrandom association between alleles detected at D4S95 and D4S98 and the Huntington's disease gene. J Med Genet 26:676–681

Thompson EA (1978) The number of ancestral haplotypes contributing to a sample of B8 alleles. Nature 272:288

Thompson EA, Neel JV (1997) Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am J Hum Genet 60:197–204

Thompson EA, Neel JV, Smouse PE, Barrantes R (1992) Microevolution of the Chibcha-speaking peoples of lower Central America: rare genes in an Amerindian complex. Am J Hum Genet 51:609–626

Van der Meulen M, te Meerman GJ (1997) Association and haplotype sharing due to identity by descent, with an application to genetic mapping. In: Pawlowitzki I, Edwards J, Thompson E (eds) Genetic mapping of disease genes. Academic Press, London, pp 115–136

Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531

Yu CE, Oshima J, Fu YH, Wijsman EM, Hisama F, Alisch R, Matthews S, et al (1996) Positional cloning of the Werner's syndrome gene. Science 272:258–262