ANIMAL BIOTECHNOLOGY, 9(3), 205-228 (1998)

DESIGNS OF REFERENCE FAMILIES FOR THE CONSTRUCTION OF GENETIC LINKAGE MAPS

by
Yang Da[1], Paul M.VanRaden[2], Ning Li[3], Craig W. Beattie[1],
Changxin Wu[3], and Lawrence B. Schook[1]
[1]Department of PathoBiology, Program in Comparative Genomics, University of Minnesota, Saint Paul, Minnesota 55108, USA
[2]Animal Improvement Programs Laboratory, USDA-ARS, Beltsville, Maryland 20705, USA
[3]China Agricultural University, Beijing 100094, P.R. China

ABSTRACT

The reference family panel is the foundation of a gene mapping program because it affects the cost and quality of the genetic linkage maps, and should be designed to yield reliable linkage detection and locus ordering at minimal gene mapping cost. A map cost function was defined as the number of genotypes required per marker per unit of genome coverage and was used to obtain optimal designs with respect to linkage detection. An ordering reliability function was defined as the likelihood ratio of the most likely order to the second most likely order of genetic markers and was used to find optimal designs with respect to locus ordering. Optimum levels of recombination frequency were found to be in the neighborhood of 0.11-0.15 for linkage detection and were in the region of 0.05-0.20 for locus ordering. Therefore, recombination frequencies optimal for linkage detection are also optimal for locus ordering. Based on the optimal detection levels, sample size (number of offspring) and map cost requirements were derived for six representative designs, assuming gender-specific linkage maps and two alleles with equal frequency for each marker. The sample size required for linkage detection ranged from 168 to 432 offspring for full-sib designs and ranged from 350 to 600 offspring for half-sib designs depending on the family size and the target LOD score, with corresponding minimal map costs of 10-20 genotypes per marker per centiMorgan map coverage. Locus ordering generally requires more genotypes than linkage detection. For full-sib designs, meioses from both genders should be used for locus ordering even when the maps are gender-specific. For half-sib designs, additional families may be needed for locus ordering. Sample size for ordering closely linked loci as required by positional cloning were provided. Effects of family size, grandparents, and marker polymorphism on design efficiency were analyzed.

Keywords: reference families, map cost, sample size, linkage detection, locus ordering

INTRODUCTION

Linkage maps of genetic markers are important tools for finding genes associated with economically important traits and disease traits in animal and plant species. Although linkage maps have been constructed for several domestic species (Bumstead et al., 1996; De Gortari, 1998; Kappes et al.,

1997; Rohrer et al., 1996), the construction of linkage maps is just being planned for some other species, e.g., a gene mapping program for several aquacultural species has recently been planned (Alcivar-Warren, 1997; Kocher et al., 1997; Gaffney et al., 1997; Waldbieser et al., 1997). Considering the rapid progress in gene mapping and the large number of species in existence, many more genetic linkage maps may be constructed in the future.

The construction of linkage maps requires analyzing genetic marker data generated from individuals in a set of families with designed structure that allows linkage detection with minimal genotyping requirement and sufficient DNA resource to generate the necessary number of genotypes for each individual. Such families are referred to as reference families, or a reference family panel (Hetzel, 1991). The design of reference families is the foundation of a gene mapping program because reference families affect the quality and cost of the linkage maps. Once established and put into usage, discarding an inefficient family would waste the genotypes already generated, in addition to costs associated with DNA collection for the family. Therefore, it is important to design reference families correctly at the beginning of the gene mapping program.

The construction of linkage maps consists of two tasks, linkage detection and locus ordering. Therefore, the design of reference families should yield sufficient resources for both of these tasks. Several aspects of designing reference families for linkage detection have been discussed in the literature. Botstein et al. (1980) estimated the number of offspring required to detect a given level of recombination frequency based on the polymorphism information content (PIC). White et al. (1985) suggested a three-generation structure for constructing the human genetic linkage maps. For domestic animals, Hetzel (1991) outlined principles of designing reference families for gene mapping. Van der Beek et al. (1993) analyzed optimal designs classified by parental gametic mating types based on the expected maximum LOD scores, Van der Beek and Van Arendonk (1993) analyzed the efficiency of designs classified by pedigree structures taking into account heterozygosity of the genetic markers. Elsen et al. (1994) studied optimum family structures, contributions of grandparents and dams for half-sib designs. Da and Lewin (1995) compared relative efficiencies of full-sib and half-sib designs taking into account joint marker informativeness for each pair of genetic markers. These studies provided certain solutions to the design

issue in terms of detecting given levels of recombination frequency but none of them addressed the issue in terms of optimum detection level and minimal map cost. Furthermore, none of these studies addressed the need for locus ordering.

Total map coverage is a function of the recombination frequency (or map distance) to be detected and the number of markers to be used for genome coverage. To detect a large recombination frequency would require genotyping a large number of individuals, whereas to detect a small recombination frequency would require genotyping a large number of markers, both translating into large gene mapping costs. Therefore, in designing reference families, optimal levels of recombination frequency for both linkage detection and locus ordering should be sought to minimize the gene mapping costs while yielding linkage maps with reliable linkage detection and locus ordering.

The purpose of this study was to investigate the designs of reference families necessary for linkage detection and locus ordering while minimizing the gene mapping cost, to study the relationship between optimal recombination frequencies for linkage detection and for locus ordering, to study factors that affect the map cost and design, and to derive resource requirements for various gene mapping designs applicable to species where mating systems and family structure can be designed.

THEORY AND METHODS

Different designs of reference families have their own advantages and disadvantages, e.g., the most efficient design could be the most difficult to implement. Therefore, a general description of applicable designs and the analysis of their advantages and disadvantages are necessary before selecting designs for further analysis. This section will start with the description and analysis of typical designs of reference families, followed by statistical formulations to obtain optimal designs with respect to linkage detection and locus ordering, and the estimation of resource requirements for selected designs. Animal species will be assumed, although the results are applicable to any species where mating system and family structure can be designed.

Basic designs of reference families According to the number of missing grandparents and family structure, designs of reference families applicable to gene mapping in animals can be divided into nine basic designs (Figure 1). Pedigree structures with more than three generations are not considered

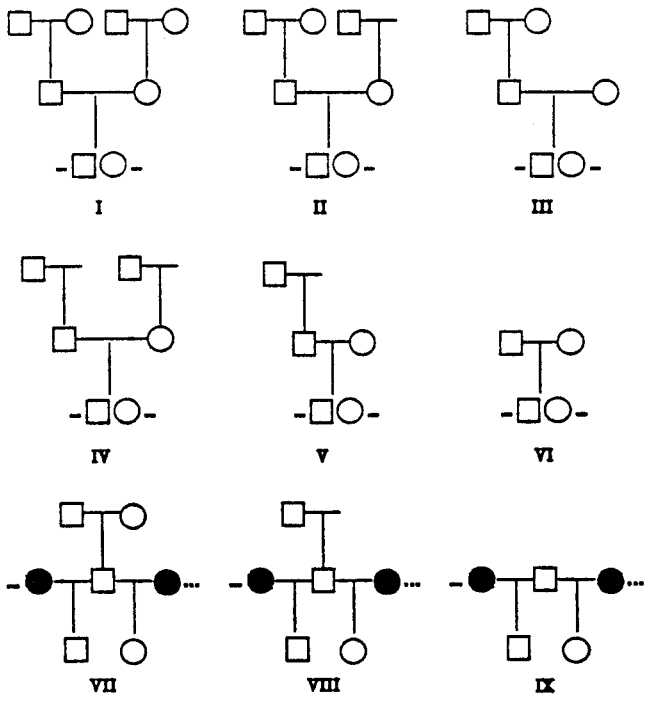

Figure 1. Typical designs to construct genetic linkage maps. A square denotes male, a circle denotes female, and a dark circle denotes an individual with unknown genotype. Designs I-VI are full-sib designs, and VII-IX are half-sib designs (a full-sib family with one missing parent can be one of VII-IX but is considered a non-typical design, see text). Designs VI and IX are two-generation designs, and the other seven are three-generation designs.

because more than three generations are not necessary for the construction of linkage maps of genetic markers. More complex designs such as a mixture of full-sib and half-sib designs can be treated as a combination of these nine designs.

Designs I-VI are full-sib designs and VII-IX are half-sib designs. Full-sib designs yield meioses for both genders and can be used to construct gender-average linkage maps as well as gender-specific maps. Half-sib designs yield meioses for one gender only and consequently can construct linkage maps only for one gender. Therefore, if linkage maps for both genders are required, full-sib designs are the only choices. It could be argued that half-sib designs could construct maps for both genders provided dams and maternal grandparents are genotyped. This practice, however, would be too costly to be practical

because the total number of genotypes can be up to four times as large as without genotyping dams and maternal grandparents. Therefore, the possibility of using half-sib designs to construct gender-average linkage maps will not be considered. Full-sib designs are more efficient than half-sib designs for linkage detection and locus ordering if both genders have the same recombination frequencies. If both genders do not have the same recombination frequencies, full-sib designs may not be more efficient than half-sib designs for linkage detection but are still about twice as efficient as half-sib designs for locus ordering. This is because both genders can be assumed to have the same locus order so that meioses from both genders can be used for locus ordering. Full-sib designs also yield more informative offspring per family than half-sib designs, because both parents are genotyped in full-sib designs whereas half-sib designs typically genotype only one parent for each family. The disadvantage of full-sib designs is that a large family size is difficult to obtain for uniparous species such as cattle and horses. For these species, half-sib designs (VII-IX) can have much larger family sizes. The half-sib designs result in a by-product of gene mapping, i.e, estimates of allele frequencies. This is because half-sib designs generally have many more dams than full-sib designs and dam alleles observed in the offspring can be used for the estimation of allele frequencies (Ron et al., 1995). Note that the dam of each offspring does not have to be genotyped to observe her alleles because a dam allele is identified if the allele transmitted from the sire to the offspring is identified. For an application-oriented gene mapping program, where gender differences in recombination rate are not of interest, a half-sib design could be a suitable choice for constructing a gender-specific linkage map and for the estimation of marker allele frequencies.

Designs VI and IX are two-generation designs, and the other seven designs are three-generation designs. Three-generation designs yield phase-known and phase-unknown data whereas two-generation designs yield only phase-unknown data. In terms of overall efficiency, design I is most efficient if the family size is sufficiently large (Da and Lewin, 1995). However, design I is also the most difficult to implement because it requires the collection of DNA for family members of three generations. For each missing grandparent, there is a loss of linkage information but the design becomes easier to implement.

A gene mapping design should have sufficient resource to yield gender-specific linkage maps even when gender-average maps are available, because a gender difference in recombination frequencies

may exist. For the construction of gender-specific linkage maps, designs III, V, VI, VII, VIII, and IX are representative designs, and will be studied in detail. Resources sufficient for gender-specific maps are also sufficient for gender-average maps, because gender-average maps generally require half of the genotypes required by gender-specific maps if equal recombination frequencies for both genders are assumed.

The mating system for the nine designs (Figure 1) can be a cross or crosses between divergent breeds to increase the heterozygosity of each marker. The disadvantage of making crosses is the time and animal maintenance cost involved, especially for three-generation designs and for species with longer generation intervals. An alternative to crosses would be to select families from the currently available families where efficient families may exist. For example, in the US dairy cattle population, several large full-sib families are available and some of the families have 25 to 46 offspring per family. In terms of family sizes, these families can be a powerful tool for gene mapping because large families are more efficient as will be demonstrated. Designing families with these family sizes can be costly. The disadvantage of using purebred families is the potentially lower heterozygosity. Therefore, the decision to select a mating system for the reference families would require a balance between the time and maintenance cost required by making crosses and the loss of information due to lower heterozygosity by using a purebred population. The relative advantage and disadvantage of each design can be evaluated using the statistical approach presented in this study.

Definitions and assumptions *Genome coverage* refers to the total map distance covered by the genetic markers on the linkage map in the unit of centiMorgans (cM) or Morgans (1 Morgan = 100 cM). *Detection level* will be defined as the average value of recombination frequency to be detected by the reference family panel. *Detection distance* (x) will be defined as the map units in cM corresponding to the specified detection level. Let x = detection distance in cM between adjacent genetic markers, k = number of chromosomes, m = number of markers per chromosome, then the genome coverage can be expressed as $C = k(m-1)x$. Therefore, genome coverage is a function of the detection distance and the number of markers. Large detection distance requires genotyping a large number of offspring and small detection distance requires genotyping a large number of markers, resulting in unnecessarily high gene

mapping cost. *Map cost* will be defined as the number of genotypes required per marker per unit (per cM) of genome coverage. The detection level that minimizes the map cost will be referred to as *optimum detection level. Ordering reliability* refers to the relative reliability of the most likely order of loci to the second most likely order, and will be defined as the likelihood ratio for the two orders. An *optimum ordering level* will be defined as the detection level that maximizes the ordering reliability. *Sample size* will be defined as the total number of offspring in the reference family panel. *Data Structure* refers to the division of sample size into families with certain family sizes. *An informative meiosis* refers to a meiosis that can be identified as a recombinant or nonrecombinant. *Gender orientation* of the linkage map refers to whether the linkage map is gender-specific or gender-average. *Gender-specific map* allows unequal recombination frequencies for two genders and the linkage map is constructed separately for each gender. *Gender-average map* assumes equal recombination frequency in both genders and is constructed using meioses from both genders. Generally, gender-specific maps require twice as many meioses as gender-average maps. A gene mapping goal generally consists of four factors, the detection level, the ordering reliability, genome coverage, and gender orientation of the genetic map. Each of the four factors affects the requirement for sample size map cost. Codominant markers with equal allele frequencies will be assumed. Chiasma and chromatid interference will be assumed absent.

Sample size Sample size is obtained by the product of family size and number of families, where family size is the number of offspring per family. In the following, family size will be assumed known, and the number of families required to detect a recombination frequency will be derived from the maximum LOD score for the given family size. Let θ= the true recombination frequency, q = number of informative meioses per family, $q\theta$ = expected number of recombinants, $q(1-\theta)$ = expected number of nonrecombinants, f = number of families with phase-known data, m = number of families with phase-unknown data. The maximum LOD score for q phase-known informative meioses is $(q)[\log_{10}(2) + (1-\theta)\log_{10}(1-\theta)+\theta\log_{10}(\theta)]$ (p.41, Ott, 1991). Assuming equal family has q phase-known informative meioses, the maximum LOD score for f families of phase-known data can be expressed as:

$$Z_{PK} = (f)(q)[\log_{10}(2) + (1-\theta)\log_{10}(1-\theta)+\theta\log_{10}(\theta)] \quad (1)$$

Given phase-known data, the relationship between the genotypes and the informative meioses is $q = n(i_j)$,

where n = number of offspring with known genotypes, i_j = frequency of informative offspring (equations 1-2 in Da and Lewin, 1995). Replacing q in equation (1) by $n(i_j)$, the maximum LOD score for f families of phase-known data becomes:

$$Z_{PK} = (f)(n)(i_j)[\log_{10}(2) + (1-\theta)\log_{10}(1-\theta)+\theta\log_{10}(\theta)] \quad (2)$$

For phase-unknown data, the likelihood of informative meioses need to consider two alternative parental linkage phases, the coupling phase and the repulsion phase. Assume that recombinants in the offspring are less frequent under coupling phase than under repulsion phase, and let L_1 = the likelihood of the informative meioses in coupling parental linkage phase, L_2 = the likelihood of the informative meioses in repulsion linkage phase. Then the likelihood function for phase-unknown data can be expressed as $L = (L_1 + L_2)/2$ (pp. 94-104, Ott, 1991). For q informative meioses of phase-unknown data, the maximum LOD score is $[(n)(i_j) - 1]\log_{10}(2) + \log_{10}(L_1 + L_2)$ (p.102, Ott, 1991). For m families of phase-unknown informative meioses, the maximum LOD score can be expressed as:

$$Z_{PU} = (m)\{[(n)(i_j) - 1]\log_{10}(2) + \log_{10}(L_1 + L_2)\} \quad (3)$$

where

$$L_1 = (1-\theta)^{n(i_j)(1-\theta)}\ \theta^{n(i_j)\theta}, \quad L_2 = (1-\theta)^{n(i_j)\theta}\ \theta^{n(i_j)(1-\theta)}$$

For s families of three-generation data, (s)(H) of the families are expected to be doubly heterozygous, where H is the double heterozygosity of two markers. Let u + v = s such that f = (u)(H) families are expected to be phase-known and m = (v)(H) families are expected to be phase-unknown. Then, from equations (2-3), the likelihood function for the s families can be written as:

$$L = (\prod_{i=1}^{f} L_{1i})[\prod_{i=1}^{m}(0.5)(L_{1i} + L_{2i})] = (\prod_{i=1}^{uH} L_{1i})[\prod_{i=1}^{vH}(0.5)(L_{1i} + L_{2i})]$$

and the corresponding maximum LOD for the s families of three-generation data is:

$$Z = \log_{10}(L)/\log_{10}[0.5^{(s\,q)}] = (s)(H)[w_k Z_{PK} + (1 - w_k)Z_{PU}] \quad (4)$$

where $w_k = (uH)/(sH) = u/s$ = frequency of phase-known families when genotyping k grandparents on the parental path (k = 1 or 2), and $q = n(i_j)$, as defined in equation (2). Note that the expected value of w_k

(equation 12 in Da and Lewin, 1995) is a function of allele frequencies. For three-generation designs, $0 < w_k \leq 1$; for two-generation designs, $w_k = 0$. Therefore, equation (4) can be considered as a general formulation for two-generation and three-generation designs, or a mixture of two-generation and three generation designs.

From equation (4), the number of families required to achieve the target LOD score Z is:

$$s = Z/\{H[w_k Z_{PK} + (1 - w_k)Z_{PU}]\} \quad (5)$$

and the sample size (total number of offspring) required to achieve the specified Z is:

$$N = (n)(s) = (n)(Z)/\{H[w_k Z_{PK} + (1 - w_k)Z_{PU}]\} \quad (6)$$

where n is the number of offspring per family in Z_{PK} (equation 2) and Z_{PU} (equation 3).

In equation (6), w_k and H are functions of allele frequencies, i_j in Z_{PK} and Z_{PU} is a function of allele frequencies and recombination frequency, and Z_{PK} and Z_{PU} are functions of family size (number of offspring per family), recombination frequency, and the frequency of informative offspring (i_j). Therefore, the sample size function given by equation (6) is a function of marker polymorphism, recombination frequency, the target LOD score, and the design of the reference families characterized by different w_k, i_j, and family size. However, the effect of the design on the sample size function is only partial, because not all members of the parents and grandparents in each design are considered by equation (6). In contrast, the effect of the design on the map cost function will be complete, because every individual in the design is taken into account by the map cost function, as shown below.

The map cost function Let G = the map cost function = number of genotypes required per map unit in cM per marker, y = number of individuals to be genotyped per family per marker, h = expected heterozygosity of the marker, and x = the map units in cM corresponding to a specified detection level obtained from the Haldane map function (Haldane, 1919; Ott, 1991). In gene mapping practice, the parents generally are screened first. If a parent is heterozygous, then the offspring and grandparents (if applicable) in the family are genotyped for the marker. Based on this practice, the number of individuals to be genotyped per family for a marker can be expressed as $y = 1 + (n+t)h$, where t = the number of parents and grandparents in each family = 3, 2, 1, 2, 1, and 0 for designs III, V, VI, VII, VIII, and IX

respectively, and n = family size. Then, by the definition of G given at the beginning of this section and noting equation (5), the map cost function can be expressed as:

$$G = (s)(y)/x = -[1-(n+t)h]Z/\{50[\log(1-2\theta)]H[w_k Z_{PK} + (1-w_k)Z_{PU}]\} \quad (7)$$

In comparison with the sample size function given by equation (6), The map cost function given by equation (7) is affected by the design of reference families because it involves the whole structure of the design, i.e., the number of families and all members in each family. In addition, the map cost function is also affected by marker polymorphism, recombination frequency, and the target LOD score (Z), the same factors that affect the sample size function. The map cost function is a more useful criterion than the sample size to evaluate designs because the map cost function considers the relationship between the number of genotypes required, all family members involved in the design, the map distance as well as factors considered by the sample size function.

Ordering reliability, and the sample size and map cost requirements for locus ordering

The likelihood ratio of the most likely order to the next most likely order is a widely accepted statistical evidence for locus ordering (Weeks, 1991). This likelihood ratio in $\log_{10}$ scale will be referred to as ordering reliability and will be used to locate the optimum ordering level and to derive the sample size and map cost requirements for locus ordering. Assume phase-known data and three loci with the true order of 1-2-3, then two incorrect orders are possible, 1-3-2 and 2-1-3. Assume equal recombination frequency for adjacent intervals, i.e., $\theta_{12} = \theta_{23} = \theta$, and assume no chiasma interference such that $\theta_{13} = 2\theta(1-\theta)$. Under these assumptions, the two incorrect orders have equal likelihood, and the likelihood analysis for ordering is reduced to the comparison between the correct order and one of the two incorrect orders. For three loci, four categories of haplotypes are possible, nonrecombinant for both intervals, nonrecombinant for interval 1-2 but recombinant for interval 2-3, recombinant for interval 1-2 but nonrecombinant for interval 2-3, and recombinant for both intervals with probabilities $p_1 = (1-\theta)^2$, $p_2 = (1-\theta)\theta$, $p_3 = \theta(1-\theta)$, and $p_4 = \theta^2$ respectively. Let R = ordering reliability, Q = total number of triply fully informative gametes required to obtain the specified reliability, and $n_i = Qp_i$ = the expected number of observations for haplotype category i, i = 1, ..., 4. Then the ordering reliability can be expressed as:

$$R = (Q)(R_1) \quad (8)$$

where

$$R_1 = p_1 \log[\frac{1\theta}{1 - 2\theta\ (1 - \theta)}] + p_2 \log(\frac{1}{2\theta}) + p_3 \log[\frac{1}{2(1 - \theta)}] + p_4 \log[\frac{\theta}{1 - 2\theta\ (1 - \theta)}]$$

From equation (8), the number of triply informative meioses required for a specified ordering reliability R is $Q = R/R_1$. Through the use of Q, the number of offspring and map cost required for a given R can be estimated.

The first step is to relate Q to the number of informative gametes for two loci (q) and this can be done by the relationship between the expected frequencies of informative gametes for two loci and for three loci. Let I_k = the frequency of fully informative gametes for two markers when genotyping k parents (Da and Lewin, 1995), J_k = the frequency of fully informative gametes for three markers when genotyping k parents. When two markers are unlinked, the joint frequency of informative offspring for two markers is the product between the two individual frequencies of the two markers considered separately (Da and Lewin, 1995). Based on this result, approximately, $J = I_k^{3/2}$. This is an approximation because the linkage between the third marker and the first two markers is ignored although the linkage between the first two markers is taken into account by I_k. Then, for the same number of offspring, I_k of these offspring are expected to be informative for two loci, whereas $I_k^{3/2}$ of the offspring are expected to be informative for three loci. Therefore, $(Q)/(I_k^{3/2}) = (q)/(I_k)$. From this relationship and noting equation (8), the number of fully informative gametes for two loci can be expressed in terms of the number of fully informative gametes for three loci by $q = Q/(I_k^2) = R/[(R_1)(I_k^2)]$.

The second step to estimate the number of offspring and map cost required for ordering is to relate Q to the number of offspring required for linkage detection. Substituting $q = R/[(R_1)(I_k^2)]$ into equation (1) yields the LOD score equivalent to the resource required by locus ordering if such resource is applied to linkage detection. Let this LOD score be denoted by Z_o, then $Z_o = R/[(I_k)(R_1)][\log_{10}(2) + (1-\theta)\log_{10}(1-\theta)+\theta\log_{10}(\theta)]$. The ratio of Z_o/Z is an indicator whether locus ordering would require a larger (or smaller) sample size than linkage detection, where Z is the LOD score required for linkage detection. Let N_o = the sample size requirement for locus ordering, G_o = the map cost requirement for locus

ordering, then the sample size and map cost requirements for locus ordering can be obtained by:

$$N_o = (Z_o/Z)N \quad (9)$$

$$G_o = (Z_o/Z)G \quad (10)$$

where N is given by equation (6), and G is given by equation (7).

Ordering closely linked loci is needed for special purposes, e.g., positional cloning would require the ordering of loci about 0.10 cM apart (Kruglyak and Lander, 1995). Locus ordering at such small map distances will require many more meioses than at optimal ordering levels. The mathematical formulations for locus ordering, however, can be greatly simplified when the map distances are so small. In such cases, grandparents become unimportant and family size can be relaxed. Therefore, the three full-sib designs (III, V, VI) and the three half-sib designs (VII, VIII, IX) for gender-specific maps can be referred to as the full-sib design and the half-sib design, ignoring the differences in the numbers of grandparents. Now the locus ordering problem can be reduced to Aone recombinant ordering@, i.e., the observation of a single recombinant type between any pair of markers would yield sufficient statistical evidence for locus ordering. In this case, the pair with one recombinant should be placed at the ends of the interval, and the locus without recombination with either of the pair should be placed in the middle of the interval. Using the one recombinant ordering, the number of meioses required to detect one recombinant is $N \approx 1/\theta$ because θ is the probability for the one recombinant to occur, and the corresponding number of offspring required is

$$T = N/i_k = 1/[(\theta)(i_k)] \quad (11)$$

where i_k = frequency of informative offspring for two markers. When θ is small, $\theta^2 \approx \theta$, and $1 - 2\theta \approx 1 - \theta$ Substituting these approximations into equation (8), the ordering reliability can be simplified to

$$R \approx 1/\theta \approx N \quad (12)$$

Equation (11-12) are convenient formulae to estimate total number of offspring required and the ordering reliability for closely linked loci.

RESULTS AND DISCUSSION

Optimum detection level and minimum map cost The optimum detection level and its corresponding minimum map cost are obtained by minimizing the map cost function (equation 7), and

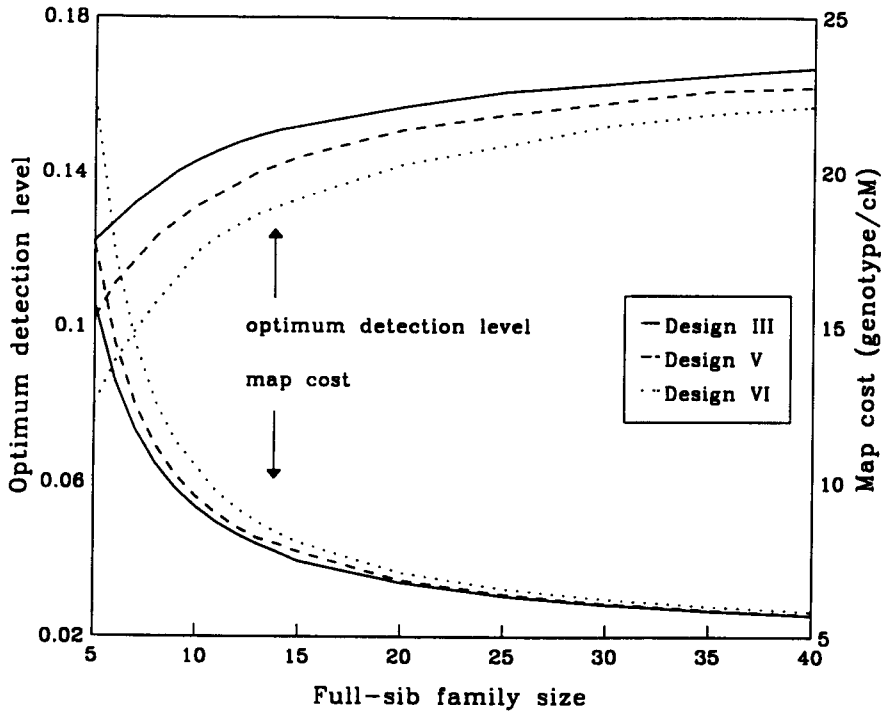

Figure 2. Optimum detection level and minimal map cost for three full-sib designs. Gender-specific linkage map, two alleles with equal frequency and a LOD score of 3.0 are assumed. Design III has two grandparents on the parental path, design III has one grandparent and design VI has no grandparent.

the results for various family sizes are shown for three full-sib designs (Figure 2) and three half-sib designs (Figure 3). The difference among the three designs in each figure is in the number of grandparents, i.e., designs III and VII have two grandparents, V and VIII have one grandparent, and VI and IX have no grandparents. For gender-specific linkage maps, a LOD score of 3.0 and two alleles with equal frequency for each marker are assumed. For the full-sib designs (Figure 2), the range of optimum detection levels is from $\theta = 0.08$ for design VI with 5 offspring per family to $\theta = 0.167$ for design III with 40 offspring per family, and the corresponding minimum map costs are 22.4 and 5.5 genotypes per cM genome coverage respectively. For half-sib designs, the range of optimum detection level is from $\theta = 0.11$ (design XI with 20 offspring per family) to $\theta = 0.156$ (design VII with 100 offspring per family) and the corresponding minimal map costs are 14.89 and 9.54 genotypes/cM per marker respectively. These

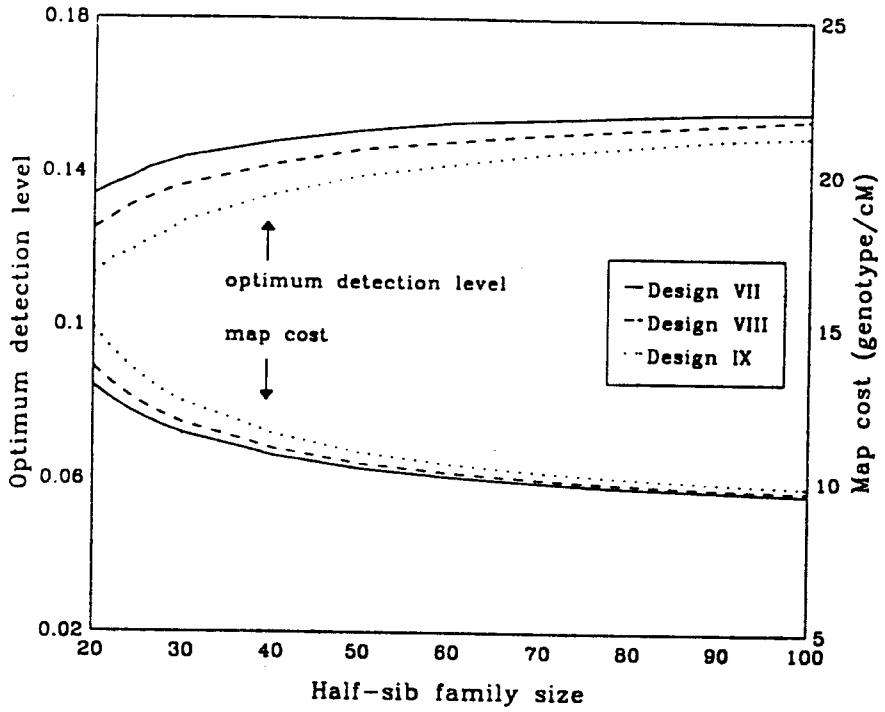

Figure 3. Optimum detection level and minimal map cost for three half-sib designs. Gender-specific linkage map, two alleles with equal frequency and a LOD score of 3.0 are assumed.

ranges of half-sib designs are narrower than those of the full-sib designs because the family sizes chosen for the half-sib designs are relatively large.

Effect of family size: As the family size increases, the optimum detection level increases and the minimal map cost decreases. These changes, however, mostly occur for family sizes below 15 for full-sib designs (Figure 2), and for family sizes below 40 for half-sib designs (Figure 3). When family sizes are below these threshold values, increases in family size may result in a significant reduction in map cost. For example, design III with a family size of 5 would cost 15.68 genotypes per cM genome coverage, whereas the same design with a family size of 15 would require only 7.95 genotypes for the same genome coverage. Assuming a genome coverage of 3000 cM, then the design with a family size of 5 would require 23,190 more genotypes than the same design with family size of 15.

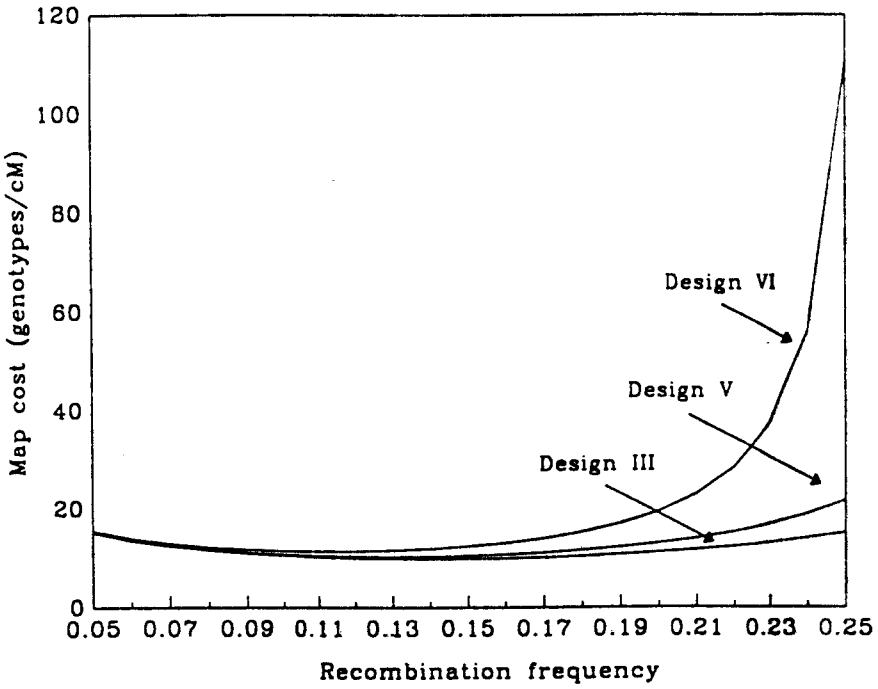

Figure 4. Map cost function for three full-sib designs. Gender-specific map and nine offspring per family were assumed. This figure shows that the map cost is minimal for recombination frequency in the region of 0.11-0.15. Within each design, larger family size is more capable of detecting larger recombination frequency than smaller family size but the effect of family size decreases as the detection level decreases.

Effect of grandparents: Grandparents are important for small family sizes. As shown in Figures 2 and 3, designs with two grandparents on each parental path are more efficient than designs with one or two missing grandparents but this difference is most pronounced when the family size is small. For example, for design III with 5 full sibs, the minimum map cost is 15.68 with two grandparents and is increased to 22.44 for the same family size without grandparents. Therefore, two missing grandparents would require 20,280 genotypes for a total genome coverage of 3000 cM for this example. In terms of linkage detection, the disadvantage of missing grandparents in small families is the inability of the design to detect a large recombination frequency, as shown in Figure 4, where design VI has two missing grandparents and has a rapidly increasing map cost when recombination frequency increases beyond 0.20.

In contrast, the designs with at least one grandparent have a map cost that increases slowly with increasing recombination frequency. This is because two missing grandparents result in phase-unknown data and the LOD score for phase-unknown data given by equation (3) decreases rapidly as the family size decreases and the number of families increases. When family size increases to a certain size (15 for full-sib design, 40 for half-sib design), grandparents become unimportant in the reduction of map cost. However, it should be noted that grandparents always have unique usage, i.e., unequivocal determination of parental linkage phase. As detection level decreases, grandparents also become less important (Figure 4). For detecting a large recombination frequency, small families without grandparents are costly, e.g., the map cost increases rapidly as the detection level increases over 0.20 (Figure 4).

Effect of marker polymorphism: Marker polymorphism has a significant effect on the sample size and map cost requirements. Figure 5 shows the effect of marker polymorphism on map cost for design III with two family sizes, 9 and 15, assuming equal allele frequencies. If the number of alleles is increased from 2 to 3, then the map cost is reduced by about 50%. However, further increase in the number of alleles beyond 4 does not result in significant further reduction in the map cost. Note that the polymorphism assumed in this study is the number of alleles with equal allele frequency. In practice, alleles generally do not have equal allele frequencies. To apply results in this study, effective number of alleles can be used be in place of the number of alleles. Effective number of alleles for the purpose of describing marker polymorphism was defined as the number of alleles with equal allele frequency equivalent to the observed marker polymorphism (Da et al., 1997). The polymorphism information content when genotyping two parents (PIC_2; Botstein et al., 1980) corresponding to 4 effective alleles is 70.3%. Therefore, genetic markers with a PIC_2 value of 70% or above are sufficiently effective for the construction of genetic linkage maps.

Optimum ordering level Figure 6 shows that the ordering reliability (equation 8) reaches its maximum value and the required number of triply fully informative gametes reaches its minimum value at $\theta = 0.12$. As θ changes in either direction, the ordering reliability changes rather quickly but the number of triply informative meioses required to achieve the specified ordering reliability remained relatively stable (below 40) in the region of $0.05 \leq \theta \leq 0.22$, which can be considered as the region of

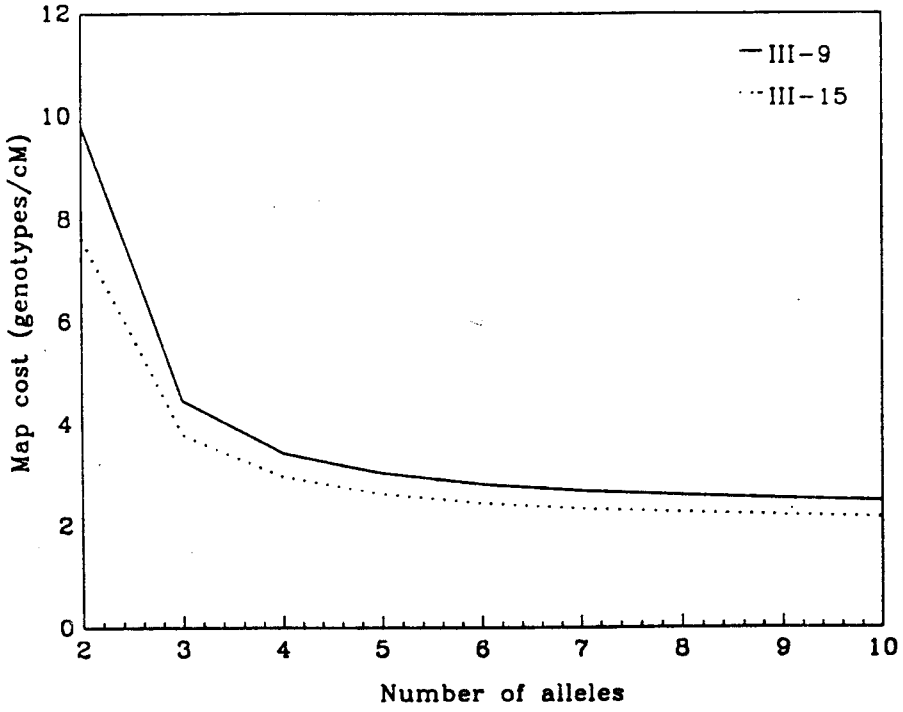

Figure 5. Effect of marker polymorphism on the map cost function. The figure shows the map cost of design III with family size of 9 offspring (III-9) and with family size of 15 offspring (III-15) for different number of alleles assuming equal allele frequency.

optimum ordering levels. Therefore, the region of optimum ordering levels contains the optimum detection levels, and designs optimal for linkage detection should also be optimal for locus ordering. Note that the optimum ordering levels were derived under the assumption of phase-known data because this assumption allowed simple mathematical results such as equations (8) and (12). The region of optimum ordering levels derived for phase-known data should contain the region of optimum ordering levels for phase-unknown data, because phase-unknown data generally are less capable than phase-known data either for locus ordering as well as for linkage detection, resulting in narrower region of optimum detection levels. This implies that the resource requirements for locus ordering based on phase-known data should be considered as the minimal requirements in the context that phase-unknown data would required more resources for the same ordering task.

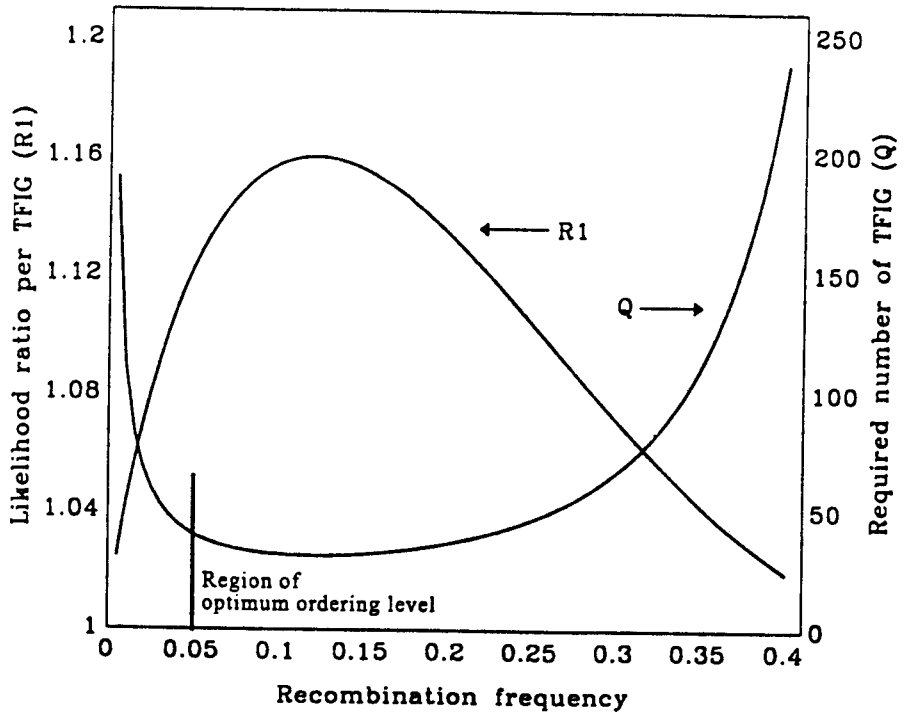

Figure 6. Likelihood ratio of the correct order to an incorrect order per triply fully informative gametes (TFIG) and total number of TFIG required to order three loci. A likelihood ratio of (correct order)/(incorrect order) = 100:1 and equal recombination frequency in adjacent marker interval were assumed. Ordering loci requires the least number of TFIG at recombination frequency of 0.12 for each interval. For recombination frequency in the range of 0.05-0.22, ordering loci requires 31-40 TFIG. As recombination frequency decreases from 0.05 to 0, or increases from 0.30 to 0.50, ordering loci becomes more difficult rapidly.

Sample size and map cost for linkage detection Sample size requirements at optimal detection level (Table 1) and their corresponding minimal map costs (Table 2) for the construction of gender-specific linkage maps were calculated for three full-sib and three half-sib designs. An optimum detection level of 0.13 was assumed for full-sib designs and an optimum detection level of 0.15 was assumed for half-sib designs. A larger optimum detection level was assumed for half-sib designs because larger family sizes were assumed. The number of offspring required ranged from 165 to 432 for the full-sib designs and ranged from 350 to 600 for the half-sib designs. The map cost ranged from 6.81 to 19.22

Table 1: Family size, number of families, and sample size of the reference family panel required by full-sib and sib designs to construct gender-specific linkage maps at optimum detection level [a]

Maximum	Full-sib designs			Half-sib designs		
LOD	III	V	VI	VII	VIII	IX
3.0	9[b] /189[c]	9/216	9/261	40/360	40/360	40/400
	15/165	15/195	15/195	50/350	50/350	50/350
	21/168	21/168	21/189	60/360	60/360	50/360
4.0	9/252	9/288	9/351	40/480	40/480	40/520
	15/285	15/195	15/270	50/450	50/450	50/500
	21/231	21/231	21/252	60/420	60/480	60/480
5.0	9/315	9/351	9/432	40/600	40/600	40/640
	15/285	15/300	15/330	50/550	50/600	50/600
	21/273	21/294	21/315	60/540	60/600	60/600

[a] Each marker is assumed to have two alleles with equal allele frequency. The optimum detection level is $\theta = 0.13$ for full-sib designs and is $\theta = 0.15$ for half-sib designs.
[b] Number to the left of each / is the family size assumed for the design.
[c] Number to the right of each / is the sample size required by the design and was calculated using equation (6), in which the number of families required was rounded up to the nearest integer.

genotypes/cM for full-sib designs and ranged from 10.75 to 19.26 genotypes/cM for half-sib designs. These results were derived based on the assumption of two alleles with equal allele frequencies. Deviation of marker polymorphism from this assumption could make the sample sizes more or less sufficient and make the map cost estimates more or less than the actual cost. For microsatellite marker linkage maps, the sample size and map cost requirements in Tables 1 and 2 may be overestimates. For example, of the 247 microsatellite markers in Ma et al. (1996), more than 80% had more than two effective alleles. When constructing linkage maps for new genetic markers, the map costs in Table 2 could be underestimates because marker spacing may not be optimal. In this case, design based on optimal spacing is still the best

Table 2: Expected map cost of designs of reference family panel [a]

Maximum	Full-sib designs			Half-sib designs		
LOD	III	V	VI	VII	VIII	IX
3.0	9[b] /9.82[c]	9/10.22	9/11.53	40/10.75	40/11.00	40/11.56
	15/7.63	15/7.70	15/8.05	50/10.34	50/10.51	50/10.91
	21/6.81	21/6.82	21/6.99	60/10.07	60/10.21	60/10.51
4.0	9/13.09	9/13.63	9/15.38	40/14.34	40/14.67	40/15.41
	15/10.17	15/10.26	15/10.73	50/13.78	50/14.02	50/14.54
	21/9.08	21/9.10	21/9.32	60/13.43	60/13.61	60/14.01
5.0	9/16.37	9/17.04	9/19.22	40/17.92	40/18.34	40/19.26
	15/12.71	15/12.83	15/13.41	50/17.23	50/17.52	50/18.18
	21/11.35	21/11.37	21/11.65	60/16.78	60/17.01	60/17.52

[a] Each marker is assumed to have two alleles with equal allele frequency. The optimum detection level is $\theta = 0.13$ for full-sib designs and is $\theta = 0.15$ for half-sib designs.
[b] Number to the left of each / is the family size assumed for the design.
[c] Number to the right of each / is the map cost of the design and is calculated using equation (7).

strategy. Markers not mapped due to non-optimal spacing can be mapped later by adding more families. Attempting to map all markers with various marker spacing can be costly and impractical. The total number of genotypes needed for a gene mapping program can be estimated by the map cost multiplied by the number of markers.

Sample size and map cost for locus ordering Assuming two alleles with equal allele frequency, I_k^2 in equations 9-10 equals 0.56 for large two-generation half-sib families and at most equals 0.55 for small three-generation full-sib families. Therefore, approximately, the number of offspring that generate one triply fully informative meioses is expected to generate about two doubly fully informative meioses. The number of meioses for locus ordering at $\theta = 0.12$ (optimum ordering level) with a likelihood

Table 3: Number of offspring required by ordering closely linked loci for positional cloning based on the one recombinant ordering strategy.

Map	Full-sib designs			Half-sib designs		
distance (cM)	2[a]	3	4	2	3	4
0.10	1603[b]	1247	1134	2672	1803	1545
0.11	1457	1135	1031	2429	1639	1404
0.12	1336	1040	945	2227	1503	1287
0.13	1232	960	873	2056	1388	1189
0.14	1145	892	810	1910	1289	1104
0.15	1070	833	757	1784	1204	1031

[a] Numbers in this row are numbers of alleles for each marker.

[b] Numbers in the body of the table are numbers of offspring required. The ordering reliability is approximately 1/x, where x is the map distance.

ratio of 100:1 would generate a LOD score of 7.28, and would generate a LOD score of 10.57 if the resource is sufficient for locus ordering with a likelihood ratio of 1000:1. Consequently, locus ordering would require twice as many meioses as linkage detection with a LOD score of 3.64 if the target likelihood ratio is 100:1, and a likelihood ratio of 1000:1 would require twice as many meioses as required by a LOD score of 5.3 . The solution to the resource requirements by locus ordering would not be to increase the sample size for linkage detection because this is a costly solution. Two options are available. For full-sib designs, meioses from both genders can be used for locus ordering even when gender-specific maps are required, because different genders have the same locus order although they may have different map distance. This approach doubles the meioses available for locus ordering and would be sufficient for most ordering needs. For half-sib designs, reserve families should be a suitable solution. These reserve families are genotyped only for those markers that have unreliable ordering or linkage detection. For ordering closely linked loci, such as the map distance required by positional cloning (Kruglyak and Lander, 1995), the number of offspring required is given in Table 3 for various

map distances and number of alleles. Assuming two alleles with equal frequency and 0.10 cM map distance between adjacent markers, about 1603 offspring are required to find one recombinant for full-sib designs, and 2672 offspring are required for half-sib designs. The approximate ordering reliability (equation 12) is $R \approx 1/0.10 = 1000:1$, and the exact ordering reliability (equation 8) is $R = 1007:1$. For full-sib designs, the number of offspring can be reduced by 50% if meioses from both genders are used for locus ordering. As the number of alleles increases, the number of offspring required decreases significantly (Table 3).

Comparisons with previous studies The optimum detection levels for various designs were found to be around $\theta = 0.11$-15, which is lower than some detection levels used in the literature , e.g., $\theta = 0.20$ was used in Van der Beek et al. (1993) and in Da and Lewin (1995), and 40 cM was suggested in White et al. (1985). This study agrees with the literature on the effect of family size, i.e., large families generally are more efficient than small families but does not necessarily agree with the literature in terms of the actual sample size required for a particular design. This study found that grandparents become unimportant when family size is large, about 15 offspring for full-sib designs and 40 offspring for half-sib designs. This is in disagreement with the conclusion that knowledge of parental phase does not improve the quality of estimation if the family size is 5 or more (Elsen et al., 1994). It should be noted that results obtained in this study should be interpreted as expected results, because the realized set of data are assumed to equal the expected set of data. Effects of variation, such as different family sizes, different numbers of families, are not described but can be evaluated by using assumed variations in the sample size and map cost formulations, with appropriate modification to the formulations. Variations in allele frequencies can be evaluated by using the actual allele frequencies in the linkage information measures H, w_k, and I_k according to formulations given in Da and Lewin (1995).

Conclusions The principle of designing reference families is to yield genetic linkage maps with maximum map coverage, reliable linkage detection and locus ordering at minimal genotyping costs. Optimal designs can be obtained by minimizing the map cost function, which is a function of the map distance, the target LOD score, marker polymorphism and data structure. Designs obtained by minimizing the map cost functions not only are optimal for linkage detection but also for locus ordering, because

optimal detection levels overlap with optimal ordering levels. However, locus ordering, especially ordering closely linked loci, is a more difficult task than linkage detection and requires more resource than linkage detection. Resource requirements for gene mapping can be derived by minimizing the map cost function for species where mating system and family structure can be designed. Highly polymorphic markers and large families can result in significant reductions in map costs. Grandparents reduce map costs significantly for small families.

ACKNOWLEDGMENTS

The authors would like to thank Drs. R. L. Fernando, B. Southey, J. A. M. Van Arendonk and two anonymous reviewers for helpful comments and suggestions. This research was supported in part by The United States Department of Agriculture, The Chinese Natural Science Foundation, and The China-Cornell Fellowship Program.

REFERENCES

Alcivar-Warren, A. 1997. Development of expressed sequence TAGs (ESTs) and microsatellite markers for mapping the shrimp genome. *Plant and Animal Genome V* (abstract), p.27, San Diego, CA.

Bumstead, N., H. Cheng, and L. Crittenden. 1996. Consensus chicken genetic linkage map, http://poultry.mph.msu.edu/.

Botstein, D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32:** 314-331.

Da, Y., P.M. VanRaden, M. Ron, J.E. Beever, J. Song, G.R. Wiggans, R.Z. Ma, J.I. Weller, and H.A. Lewin. 1997 Standardization of Marker Informativeness Measures for Different Gene Mapping Designs. *Plant and Animal Genome V* (abstract), p.148, San Diego, CA.

Da, Y. and H.A. Lewin. 1995 Linkage information content and efficiency of fullsib and halfsib designs for gene mapping. *Theor. Appl. Genet.* **90:** 699-706.

De Gortari, M.J., B.A. Freking, R.P. Cuthbertson, S.M. Kappes, J.W. Keele, R.T. Stone, K.A. Leymaster, K.G. Dodds, A.M. Crawford, C.W. Beattie. 1998. A Second Generation Linkage Map of the Sheep Genome. *Mammalian Genome* **9:**204-209.

Elsen, J.M., B. Mangin, B. Goffinet, and C. Chevalet. 1994 Optimal structure of protocol designs for building genetic maps in livestock. *Theor. Appl. Genet.* **88:**129-134.

Gaffney P, Allen SK Jr, Hedgecock D. 1997 Genetics of farmed oyster species: current status. *Plant and Animal Genome V* (abstract), p.27, San Diego, CA.

Haldane, J.B.S. 1919 The combination of linkage values, and the calculation of distance between the loci of linked factors. *J. Genet.* **8:** 299-309.

Hetzel, D.J.S. 1991 Reference families for genome mapping in domestic livestock. In: Schook LB, Lewin HA, McLaren DG (eds) Gene-mapping techniques and applications. Marcel Dekker, New York, pp 51-64.

Kappes, S. M., J. W. Keele, R. T. Stone, R. A. McGraw, T. S. Sonstegard, T. P. Smith, N. L Lopez-Corrales, C. W. Beattie, 1997 A second-generation linkage map of the bovine genome. *Genome Res.* 7:235-49.

Kocher, T. D., Lee W. Sobolewska H, Penman D, and McAndrew B. 1997 A genetic linkage map of the Tilapia (Oreochromis niloticus). *Plant and Animal Genome V* (abstract), p.27, San Diego, CA.

Kruglyak, L. and E.C. Lander. 1995 High-resolution genetic mapping of complex traits. *Am. J. Hum. Genet.* 56:1212-1223.

Ma, R.Z., J.E. Beever, Y. Da, C.A. Green, I. Russ, C. Park, D.W. Heyen, R.E. Everts, S.R. Fisher, K.M., Overton, A.J. Teale, S.J. Kemp, H.C. Hines, G. Guérin, and H.A. Lewin. 1996 A male linkage map of the cattle (*Bos taurus*) genome. *J. Heredity* **87: 261-271.**

Ott, J. 1991 Analysis of human genetic linkage. Revised Edition. The Johns Hopkins University Press, Baltimore and London.

Rohrer, G.A., L.J. Alexander, Z. Hu, T.P. Smith, J.W. Keele, C.W. Beattie. 1996 A comprehensive map of the porcine genome. *Genome Res.* **6:371-391.**

Ron, M., H.A. Lewin, Y. Da, M. Band, A. Yanai, Y. Blank, E. Feldmesser, and J.I. Weller. **1995.** Prediction of informativeness for microsatellite markers among progeny of sires used for detection of economic trait loci in dairy cattle. Anim. Genet. **26:** 439-441.

Van der Beek, S., J.A.M. Van Arendonk. 1993 Criteria to optimize designs for detection and estimation of linkage between marker loci from segregating populations containing several families Theor. Appl. Genet. **91:** 1115-1124.

Van der Beek, S., A.F. Groen, and J.A.M. Van Arendonk. 1993 Evaluation of designs for reference families for livestock linkage mapping experiments. Anim. Biotechnology. **86:** 269-280.

Waldbieser GC, Bosworth BG, W. R. Wolters. 1997 Development of a channel catfish genetic Genome V (abstract), p.28, map. Plant and Animal San Diego, CA.

Weeks, D. E. 1991 Human linkage analysis: Strategies for locus ordering. pp. 297-330 in *Advanced Techniques in Chromosome Research*, edited by K. W. Adolph. Marcel Dekker, New Yor

White, R., M. Leppert, D.T. Bishop, D. Barker, J. Berkowitz, C. Brown, P. Callahan, T. Holm, and L. Jerominski. 1985 Construction of linkage maps with DNA markers for human chromosomes. Nature **313:** 101-105.