Theor Appl Genet (1995) 90: 699–706

Yang Da · Harris A. Lewin

Linkage information content and efficiency of full-sib and half-sib designs for gene mapping

Received: 2 November 1994 / Accepted: 22 November 1994

Abstract The accuracy of a genetic map depends on the amount of linkage information contained in the data set used for construction of the map. The amount of linkage information is related to the designs employed for linkage analysis. The purpose of this study was to provide general formulations for various genotyping schemes and family structures in order to evaluate the amount of linkage information in a data set. Linkage information content (LIC) was defined as the frequency of fully informative gametes, which are gametes from doubly heterozygous parents with known linkage phases. Depending on the design, LIC is based on two generations if the parental phases are determined statistically, or three generations if the parental phases are determined genetically. Different schemes were considered in deriving LIC: (1) genotyping of one parent or two parents, and (2) genotyping of two or three generation families. The LIC for a full-sib design was found to be generally greater than for a half-sib design but requires typing a large number of individuals when at least one locus has only two alleles. The efficiency of the full-sib design is reduced significantly if a sex-specific linkage map is sought.

Key words Linkage-information content · Full-sib design Half-sib design · Gene mapping

Introduction

Efficient designs for linkage studies are important to obtain maximum linkage information with finite resources. The analysis of pedigree data has been a major method to study linkage between genes through meiotic recombination observed in offspring of heterozygous parents (Murray et al. 1994). However, genotypes of offspring may not contain information useful for linkage analysis. The amount of linkage information is also related to gene frequencies, mating systems, and family structures. Full-sib families are common in humans and domestic-swine populations (Rohrer 1994), whereas large half-sib families are typical in cattle (Lewin et al. 1994) where artificial insemination has been widely implemented. The amount of linkage information derived from each design is important for conducting and planning linkage studies, including the selection of reference families.

Fisher's information, Edwards' equivalent number of meiosis, and the expected LOD score (ELOD) are often used to measure the informativeness of data (Ott 1991; van der Beek and van Arendonk 1993), but do not measure the percentage of individuals that contribute directly to the estimation of recombination frequencies. The polymorphism information content (PIC) of a locus (Botstein et al. 1980) is a widely used measure to determine the informativeness of linkage data. The single-locus PIC was recently applied to microsatellite-typing data in dairy cattle to determine sire-allele origin (Ron et al. 1993). The single-locus PIC can not be used directly to measure linkage information because demonstrating linkage requires at least two loci and sometimes three-generation data. The single-locus PIC can not be applied when the genotype of either parent is unknown. A two-locus PIC and the frequency of fully informative gametes were derived for the CEPH human pedigrees (Chakravarti 1991), but the results do not apply when one parent is not typed. Consequently, if one or more parents have an unknown genotype, in either the grandparental or parental generation, then the two-locus PIC is not applicable to either the full-sib design (FSD) or the half-sib design (HSD) for determining informativeness.

The purpose of the present study was to provide general formulations to measure linkage information for a random mating population under various genotyping schemes and to use the new measures to evaluate the efficiencies of FSD and HSD for generating genetic maps. These formu-

Communicated by L. D. Van Vleck

Y. Da · H. A. Lewin (✉)
Department of Animal Sciences, 206 Plant and Animal Biotechnology Laboratory, 1201 West Gregory Drive, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

Table 1 Frequency of informative genotypes for one locus from a mating between a heterozygous parent (A_iA_j) and a random parent

Random parent		Genotypic array in offspring	Informative genotypes		Frequency of informative genotypes	
Genotype[a]	Frequency		Two parents typed	One parent typed	Two parents typed	One parent typed
A_kA_l	$(1-p_i-p_j)^2$	$1/4\,A_iA_k+1/4\,A_iA_l+1/4\,A_jA_k+1/4\,A_jA_l$	All	All	1	1
A_iA_k	$2p_i(1-p_i-p_j)$	$1/4\,A_iA_i+1/4\,A_iA_j+1/4\,A_iA_k+1/4\,A_jA_k$	All	A_iA_i A_iA_k A_jA_k	1	3/4
A_jA_k	$2p_j(1-p_i-p_j)$	$1/4\,A_jA_j+1/4\,A_iA_j+1/4\,A_iA_k+1/4\,A_jA_k$	All	A_iA_i A_iA_k A_jA_k	1	3/4
A_iA_j	$2p_ip_j$	$1/4\,A_iA_i+1/2\,A_iA_j+1/4\,A_jA_j$	A_iA_i A_jA_j	A_iA_i A_jA_j	1/2	1/2
A_iA_i	p_i^2	$1/2\,A_iA_i+1/2\,A_iA_j$	All	A_iA_i	1	1/2
A_jA_j	p_j^2	$1/2\,A_jA_j+1/2\,A_iA_j$	All	A_jA_j	1	1/2

[a] Subscripts satisfy $i \neq j \neq k$, but k may be equal to l

lations will be useful for evaluating linkage designs and predicting the informativeness of data for gene-mapping studies.

Linkage information content (LIC)

Assumptions

Two loci with co-dominant alleles will be assumed. Locus 1 has n alleles and locus 2 has m alleles. Allele *i* at locus 1 (A_i) has a frequency of p_i , and allele *i* at locus 2 (B_i) has a frequency of t_i , where *i* for locus *A* ranges from 1 to n, and *i* for locus *B* ranges from 1 to m. All genotypes can be determined without error. Mating in the general population is assumed random with respect to the two loci to be analyzed for linkage. Under this assumption, all matings have an equal probability of producing offspring and all genotypes have an equal probability of surviving until they can be genotyped. Hence, the genotypic array of the population is given by $(\Sigma p_iA_i)^2\,(\Sigma t_iB_i)^2$.

Definitions

Linkage information content (LIC) will be defined as the frequency of fully informative gametes. A *fully informative gamete* (FIG) is a gamete with unequivocal identification of allele origin from a phase-known doubly heterozygous (DH) parent, and contributes directly to the calculation of recombination frequency. For example, if a phase-known DH sire (*AB/ab*) and a homozygous dam (*aabb*) have an offspring with the genotype *Aabb*, then that offspring received a recombinant haplotype *Ab* from the DH sire. The *Ab* gamete is fully informative because the phase and origin of alleles are known and can be counted for the calculation of recombination frequency. An *informative gamete* is a gamete with unequivocal identification of allele origin from a phase-unknown DH parent. For example, assume the phase of the DH sire is unknown (*AaBb*) and the dam's genotype is *aabb*. An *Ab* gamete in an offspring with the genotype *Aabb* is informative, but not fully informative, because whether the *Ab* haplotype is recombinant or non-recombinant is not known, although the origin of the alleles is. An *allele-informative gamete* is a gamete with an identifiable origin of the alleles from a parent of any genotype, where phase can be known or unknown. For example, if a sire with the genotype *AABB* and a dam with the genotype *aabb* have an offspring with the genotype *AaBb*, then that offspring must have received *AB* from the sire and *ab* from the dam. Both haplotypes are allele-informative because the parental origin of the alleles in each gamete can be unambiguously determined. However, the gamete in this example contains no linkage information, because neither of the parents is DH. Allele-informative gametes can be used to identify parental linkage phase but may not have linkage information (e.g., when the parent is homozygous at a locus). A *non-informative gamete* has no information about the parental origin or the phase of its alleles.

A *fully informative genotype* of an individual contains at least one fully informative parental gamete; an *informative genotype* contains at least one informative parental gamete; an *allele-informative* genotype contains at least one allele-informative parental gamete; and a *non-informative* genotype contains two non-informative parental gametes. A *fully informative mating* (consistent with the definitions described above) produces fully informative genotypes with a non-zero probability; *informative mating*

produces informative genotypes with a non-zero probability; *allele-informative mating* produces allele-informative genotypes with a non-zero probability; and *non-informative mating* produces non-informative genotypes with a probability of one. Fully informative gametes, genotypes, and matings are necessarily informative; informative gametes, genotypes, or matings are necessarily allele-informative, but the reverse is not always true. For example, the matings $AABB \times aabb$ and $Aabb \times aabb$ are allele-informative but are not informative by these definitions.

A full-sib family consists of two or more offspring with the same sire and dam, and a half-sib family has only a single common parent.

Rules to identify informative genotypes

For one locus, the rules to identify informative genotypes can be deduced from Table 1. Given a mating between a DH parent and a random parent, with both parents having known genotypes, then the offspring's genotype is informative if one of the following conditions is satisfied: (1) the parents do not have the same genotype; or (2) the genotype of the offspring is different from the parents when the parents have the same genotype. When only one parent has a known genotype, e.g., when one parent is typed and the other parent is not, the offspring's genotype is informative if the offspring and the typed parent have different genotypes. For two loci, a genotype is informative for linkage analysis if the genotype is informative at both loci. The same rules apply to allele-informative genotypes.

Probabilities to yield LIC

The LIC of a dataset has different components, depending on the methods used to determine the linkage phase of alleles in a parent and on the typing scheme. The parental linkage phase can be determined unequivocally or statistically. The unequivocal determination of parental linkage phase requires three-generation data (Ott 1991), and is based on the segregation of alleles from the grandparent to the parent and then to the offspring. Hence, this will be referred to as genetic phase determination (GPD). GPD is the most reliable method to determine parental phase but the probability may be small that allele transmissions can be identified in two-generation intervals. When GPD is not applicable (e.g., two-generation data), a statistical approach can be used to infer the most likely phase, such as the maximum-likelihood method for phase-unknown families (Ott 1991), and the maximum-likelihood method for sperm- and oocyte-typing data (Li et al. 1988; Arnheim et al. 1990; Cui et al. 1992; Lewin et al. 1992). Inference on linkage phase based on a statistical method will be referred to as statistical phase determination (SPD). SPD requires two generations for family data or one generation for single sperm- or oocyte-typing data (Arnheim et al. 1990; Lewin et al. 1992). SPD has the advantage that it may apply where GPD fails but has a chance of incorrect phase determination due to the nature of statistical inference. For GPD, LIC involves two probabilities, the probability that the phase of the DH parent can be determined, and the probability that the allele transmission from the parent to the offspring can be identified. For SPD, LIC involves only one probability, the probability that the transmission of alleles for two loci from the DH parent to the offspring can be identified. For each probability, two cases must be considered: typing two parents or typing one parent. Therefore, the following probabilities will be defined to obtain LIC: i_k=probability that a genotype is informative when k parents are typed for genetic markers (k=1 or 2), w_k=probability that the phase of the DH parent can be determined by genotypes of k grandparents, k=1 or 2.

Frequency of informative gametes when two parents are typed (i_2)

Based on the mating frequency and the probability of informative genotypes for each mating (Table 2), the frequency of informative gametes (i_2) from a DH ($A_iA_jB_iB_j$) parent when typing two parents per offspring can be obtained as:

$$i_2=1-0.5(f_3+f_6+f_{10})-0.5[1+2\theta(1-\theta)]f_8 \quad (1)$$

where f_3, f_6, f_8 and f_{10} are defined in Table 2, and θ=recombination frequency between the two loci in question.

Frequency of informative gametes when one parent is typed (i_1)

From Table 2, the frequency of informative gametes when typing one parent per offspring is:

$$\begin{aligned} i_1=1&-1/4f_2-1/2(f_3+f_5)-1/2[1-1/2(1-\theta)^2]f_4 \\ &-1/2[1+\theta(1-\theta)]f_6-1/2(1+1/2\theta)f_7 \\ &-1/2[1+2\theta(1-\theta)]f_8-1/2(1+\theta)f_9-3/4f_{10} \end{aligned} \quad (2)$$

where f_i is defined in Table 2, i=2, ..., 10.

Relationship between frequency of informative gametes and single-locus PIC

From Table 1, when the heterozygous parent, the random parent, and the offspring are typed, the expected frequency of informative genotypes in the offspring of the heterozygous parent is:

$$i_{2A}=1-p_ip_j\,. \quad (3)$$

When the heterozygous parent and the offspring (but not the random parent) are typed, the expected frequency of informative genotypes in the offspring for the heterozygous parent is:

$$i_{1A}=1-1/2(p_i+p_j)(1-p_i-p_j)-1/2(p_i+p_j)^2=1-1/2(p_i+p_j) \quad (4)$$

Table 2 Informative genotypes for two loci from a mating between a doubly heterozygous parent ($A_iA_jB_iB_j$) and a random parent

Random parent[a]		Frequency of informative genotypes	
Genotype	Frequency	Two parents typed	One parent typed
$A_kA_lB_kB_l$	$f_1 = (1-p_i-p_j)^2(1-t_i-t_j)^2$	f_1	f_1
$A_xA_kB_kB_l$ $A_kA_lB_xB_k$	$f_2 = 2(p_i+p_j)(1-p_i-p_j)(1-t_i-t_j)^2$ $+2(t_i+t_j)(1-t_i-t_j)(1-p_i-p_j)^2$	f_2	$3/4\,f_2$
$A_iA_jB_kB_l$ $A_kA_lB_iB_j$	$f_3 = 2p_ip_j(1-t_i-t_j)^2$ $+2t_it_j(1-p_i-p_j)^2$	$1/2\,f_3$	$1/2\,f_3$
$A_xA_kB_xB_k$	$f_4 = 4(p_i+p_j)(t_i+t_j)$ $\times(1-p_i-p_j)(1-t_i-t_j)$	f_4	$[1/2+1/4(1-\theta)^2]f_4$
$A_xA_xB_kB_l$ $A_kA_lB_xB_x$	$f_5 = (p_i^2+p_j^2)(1-t_i-t_j)^2$ $+(t_i^2+t_j^2)(1-p_i-p_j)^2$	f_5	$1/2\,f_5$
$A_iA_jB_xB_k$ $A_xA_kB_iB_j$	$f_6 = 4p_ip_j(t_i+t_j)(1-t_i-t_j)$ $+4t_it_j(p_i+p_j)(1-p_i-p_j)$	$1/2\,f_6$	$1/2[1-\theta(1-\theta)]f_6$
$A_xA_xB_xB_k$ $A_iA_xB_xB_x$	$f_7 = 2(p_i^2+p_j^2)(t_i+t_j)(1-t_i-t_j)$ $+2(t_i^2+t_j^2)(p_i+p_j)(1-p_i-p_j)$	f_7	$(1/2-1/4\theta)f_7$
$A_iA_jB_iB_j$	$f_8 = 4p_ip_jt_it_j$	$1/2[\theta^2+(1-\theta)^2]f_8$	$1/2[\theta^2+(1-\theta)^2]f_8$
$A_xA_xB_xB_x$	$f_9 = (p_i^2+p_j^2)(t_i^2+t_j^2)$	f_9	$1/2(1-\theta)f_9$
$A_xA_xB_iB_j$ $A_iA_jB_xB_x$	$f_{10} = 2t_it_j(p_i^2+p_j^2)$ $+2p_ip_j(t_i^2+t_j^2)$	$1/2\,f_{10}$	$1/2\,f_{10}$
Sum	1.0	i_2	i_1

[a] Subscripts for the random parent are defined as: $i\neq j\neq k$ but k may be equal to l; x=i or j. θ=recombination frequency, i_2=frequency of informative genotypes when typing two parents, i_1=frequency of informative genotypes when typing one parent

which is the same as in Ron et al. (1993). For a random sample of parents (sires or dams), the expected frequency of informative genotypes is obtained by summing over all heterozygous parents (sires or dams), i.e.,

$$I_{2A} = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_ip_j = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i^2p_j^2 \quad (5)$$

$$I_{1A} = 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_ip_j = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_ip_j\,(p_i+p_j). \quad (6)$$

The expected frequency of informative genotypes when typing two parents for a random sample of parents (I_{2A}) given by equation (5) equals the PIC of Botstein et al. (1980), i.e., the single-locus PIC is equal to the frequency of informative genotypes in the offspring. The I_{1A} of (6) could be considered as the PIC when typing one parent, because I_{1A} measures the same frequency as PIC measures, except that PIC requires typing both parents. Assuming no linkage (θ=1/2), numerical results revealed a simple relationship between the frequency of informative offspring for two loci considered jointly and for two loci considered separately , i.e.,

$$i_2 = i_{2A} \times i_{2B} \quad (7)$$

$$i_1 = i_{1A} \times i_{1B} \quad (8)$$

where i_{2A} is defined by equation (3), i_{2B} is also defined by equation (3) but p_i and p_j are replaced with t_i and t_j, i_{1A} is defined by equation (4), and i_{1B} is also defined by equation (4) but p_i and p_j are replaced with t_i and t_j. The relationships of (7) and (8) hold irrespective of the number and frequencies of alleles. The above results are intuitive, because the informativeness for two loci considered jointly should be the product of the informativeness of each locus if the two loci are independent.

Frequency of allele-informative gametes (w_k)

Typing three generations can determine the linkage phase unequivocally but has a chance of failing to do so. A formulation for the probability of failing to determine phase using three-generation data was given by Chakravarti (1991). The formulations to be presented here were derived using a different approach, taking into consideration various typing schemes. When a DH offspring is obtained, a parental genotype must contain at least one allele at each locus that appears in the offspring's genotype. Among all possible matings to produce a DH offspring, some matings are non-informative, i.e., for at least one locus, no information is available about which alleles were passed from each parent to the DH offspring. For example, a mating of *AaBb* × *AaBb* can produce a DH offspring (*AaBb*) but unequivocal identification of parental alleles is not possible. Therefore, the probability that the phases of the DH individual can be determined is the conditional probability of informative and allele-informative matings between the

Table 3 Linkage information content (LIC) under full-sib and half-sib designs for offspring of one doubly heterozygous parent. A θ of 0.20 was assumed in the calculation of i_k, k=1, 2. LIC_{f22} and LIC_{f21} are calculated using equation (13), i_2 using equation (1), i_1 using equation (2), and w_k using equation (12). When statistical phase determination is used, $LIC_{hk}=i_k$

Alleles at locus 1	Alleles at locus 2	LIC_{f22}	LIC_{f21}	i_2 (LIC_{h2})	i_1 (LIC_{h1})	w_2	w_1
2	2	0.30	0.18	0.59	0.31	0.51	0.31
2	3	0.49	0.32	0.73	0.39	0.67	0.44
2	4	0.53	0.38	0.77	0.42	0.69	0.49
2	5	0.55	0.40	0.78	0.44	0.70	0.51
3	3	0.70	0.50	0.84	0.50	0.83	0.60
4	4	0.84	0.65	0.91	0.61	0.92	0.71
5	5	0.89	0.73	0.94	0.67	0.95	0.78
6	6	0.93	0.78	0.96	0.72	0.97	0.82
7	7	0.95	0.82	0.97	0.76	0.98	0.85
8	8	0.96	0.84	0.97	0.78	0.98	0.87
9	9	0.97	0.86	0.98	0.80	0.99	0.88
10	10	0.97	0.88	0.98	0.82	0.99	0.89

parents, given all possible matings to produce the DH individual. For clarity, the following notation is used: $u_{il}=\text{Prob}(A_iA_l)=p_i^2+2p_i(1-p_i-p_j)$, $u_{jl}=\text{Prob}(A_jA_l)=p_j^2+2p_j(1-p_i-p_j)$, $u_{ij}=\text{Prob}(A_iA_j)=2p_ip_j$, $v_{il}=\text{Prob}(B_iB_l)=t_i^2+2t_i(1-t_i-t_j)$, $v_{jl}=\text{Prob}(B_jB_l)=t_j^2+2t_j(1-t_i-t_j)$, $v_{ij}=\text{Prob}(B_iB_j)=2t_it_j$, where subscript l may or may not equal i or j, and i and j are unequal. Then, under the assumption of random mating, the frequency of all possible matings that can produce a DH individual ($A_iA_jB_iB_j$) is:

$$S=(2S_A+u_{ij}^2)(2S_B+v_{ij}^2) \tag{9}$$

where $S_A=(u_{il}u_{jl}+u_{il}u_{ij}+u_{jl}u_{ij})$, and $S_B=(v_{il}v_{jl}+v_{il}v_{ij}+v_{jl}v_{ij})$.

When typing two parents, the offspring is not informative if the offspring and the parents have the same genotype at any locus, because in this case the transmission of alleles can not be identified unequivocally (see section on 'rules to identify informative genotypes'). Therefore, the frequency of non-informative matings between parents is:

$$Q_2=2(v_{ij}^2S_A+u_{ij}^2S_B)+u_{ij}^2v_{ij}^2 \tag{10}$$

When typing one parent, the offspring is not informative if the offspring and the typed parent have the same genotype at any locus. Therefore, the frequency of non-informative matings is:

$$Q_1=Q_2+u_{ij}(u_{il}+u_{jl})S_B+v_{ij}(v_{il}+v_{jl})S_A \tag{11}$$

From equations (6) through (11), the probability that the phase of an informative gamete can be determined is

$$w_k=1-Q_k/S \quad \text{for k=1, 2 .} \tag{12}$$

Frequency of fully informative gametes

When the phases of parental alleles are determined by three-generation data, the frequency of FIGs from a DH parent is the product of the probability that the phase of the DH parent can be determined and the frequency of informative offspring of the DH parent. The subscript "f" is used to indicate three-generation data and "h" to indicate two-generation data. Then, the frequency of FIGs for a DH parent based on three-generation data is:

$$LIC_{fkl}=i_kw_l \quad \text{for k, l=1 or 2} \tag{13}$$

where k=number of parents typed, and l=number of paternal or maternal grandparents typed. When parental phases can be determined statistically from two-generation data, the frequency of FIGs is simply the frequency of informative offspring, i.e.,

$$LIC_{hk}=i_k \quad \text{for k=1 or 2} \tag{14}$$

For a random sample of parents, the frequency of FIGs is obtained by summing LIC_{fkl} over all DH parents:

$$I_{fkl}=4\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}p_ip_j\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}t_it_jLIC_{fkl},$$

$$I_{hk}=4\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}p_ip_j\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}t_it_jLIC_{hk}.$$

Numerical illustration

Selected values for LIC given by equation (13) and for probabilities to generate LIC given by equations (1), (2) and (12) are shown in Table 3, including LIC values for two typical typing schemes for FSD: two grandparents or one grandparent are/is typed to determine the linkage phase of the DH parent. FSD, for typing both parents and both grandparents on each parental path, has a higher LIC (LIC_{f22}) than HSD for typing one parent (i_1), except when each locus has two alleles ($LIC_{f22}=0.30$ and $i_1=0.31$). If one grandparent has a known genotype and one has an unknown genotype, the probability that the parental linkage phase can be determined drops from $w_2=0.51$ to $w_1=0.31$, and the LIC for the FSD is reduced from $LIC_{f22}=0.30$ to $LIC_{f21}=0.18$, showing vulnerability of GPD to missing information for a grandparent. The probabilities to generate LIC in Table 3 can be used to calculate the LIC under various designs and typing schemes.

Approximation for unequal frequencies of alleles

When alleles have unequal frequencies, the results in Table 3 are not applicable but they can be used as approximations. One way to obtain an approximate estimate of

LIC using the results in Table 3 is to translate heterozygosity into an equivalent number of alleles, using n=1/(1–h), where n=equivalent number of alleles, and h=heterozygosity for the locus (Ott 1991). For example, if 20 alleles at a locus have h=0.80, then those 20 alleles would have LIC that is equivalent to five alleles with equal frequencies. Because the formulations used to derive Table 3 can consider unequal allele frequencies, the original formulations should be used rather than approximations.

Efficiencies of full-sib design (FSD) and half-sib design (HSD) for gene mapping

LIC is not the only factor that affects the efficiency of a mapping design, because the number of individuals that must be typed but do not contribute to the counting of recombination events must be taken into account. In fact, using LIC alone could give a distorted picture of the relative efficiencies of different typing schemes. For example, when both parents are typed, LIC with HSD is actually larger than with FSD, because i_2 is larger than $w_2 i_2$, whereas HSD almost always has lower efficiency than FSD if two parents are typed and the total number of individuals typed is taken into account. This section evaluates efficiencies for FSD and HSD when both LIC and the total number of individuals typed are considered.

The relative efficiency for FSD and HSD was compared using two measures: (1) the expected number of FIGs for a given number of individuals to be typed; and (2) the number of individuals required to detect a given recombination frequency. For FSD, typing three generations was considered necessary to determine the linkage phases of parental alleles. For HSD, the number of offspring was assumed to be large ($n \geq 50$), so that the linkage phase of a male parent could be determined statistically. A two-step typing scheme was assumed for both FSD and HSD. The first step is to determine the parental genotypes. For FSD, the second step is to type offspring and grandparents when a parent is DH. For HSD, the second step is to type offspring when the sire is DH.

Expected number of FIGs

The expected number of FIGs from a given sample is calculated based on the LIC and the number of offspring, parents and grandparents to be typed. Let T=total number of individuals to be genotyped, s_f=number of male parents for FSD, s_h=number of male parents for HSD, n_f=number of offspring per family for FSD, and n_h=number of offspring per male parent for HSD. The number of female parents is assumed to equal the number of male parents for FSD, and is equal to the number of offspring for HSD. Then the total number of individuals to be typed can be expressed as:

$$T=s_f[2+4H+Dn_f] \quad \text{for FSD, all individuals typed} \qquad (15)$$

$$T=s_h(1+2Hn_h) \quad \text{for HSD, female parent typed} \qquad (16)$$

$$T=s_h(1+Hn_h) \quad \text{for HSD, female parent not typed} \qquad (17)$$

where $H=(1-\Sigma p_i^2)(1-\Sigma t_i^2)$=heterozygosity for two loci, and $D=1-(1-H)^2$=probability that at least one parent is DH.

For FSD, the number of typed offspring that are potentially informative from one parent is $Hs_f n_f$, not $Ds_f n_f$, because only offspring of DH parents are potentially informative. For HSD, the number of typed offspring that are potentially informative from one parent (e.g., sire) is $Hs_h n_h$. Let g_{f1}=expected number of FIGs for one parent with FSD, g_{f2}=expected number of FIGs for two parents with FSD, g_{h1}=expected number of FIGs for HSD when female parents are not typed, and g_{h2}=expected number of FIGs for HSD when both parents are typed. Then, using equations (15) to (17), and (13) and (14), the numbers of fully informative genotypes are:

$$g_{fak}=a(LIC_{fkl})Hs_f n_f = a(LIC_{fkl})HTn_f/(2+4H+Dn_f) \quad \text{for a, k=1, 2} \qquad (18)$$

$$g_{hb}=(LIC_{hb})Hs_h n_h = (LIC_{hb})THn_h/(1+bHn_h) \quad \text{for b=1, 2} \qquad (19)$$

where a=number of parents per family counted for recombination events, k=number of parents typed using FSD, and b=number of parents typed using HSD.

Subject to the restrictions of equations (15) through (17), the design with the largest expected number of FIGs for the same numbers of typed individuals is the most efficient design. Based on equations (18) and (19), FSD has higher efficiency than HSD if the following equation is satisfied:

$$n_f > \left[\frac{(2+4H)(LIC_{hb})n_h}{a(LIC_{fkl})(1+bHn_h)}\right] \Big/ \left[1-\frac{D(LIC_{hb})n_h}{a(LIC_{fkl})(1+bHn_h)}\right] \qquad (20)$$

and HSD has higher efficiency than FSD if the following equation is satisfied:

$$n_h > \left[\frac{a(LIC_{fkl})n_f}{LIC_{hb}(2+4H+Dn_f)}\right] \Big/ \left[1-\frac{ab(LIC_{fkl})Hn_f}{LIC_{hb}(2+4H+Dn_f)}\right] \qquad (21)$$

Using inequality (20), full-sib family sizes that are equivalent to 50 and 100 half-sibs were calculated (Table 4), assuming 2–8 alleles at each locus and $\theta=0.20$. The same total number of typed individuals is assumed for all designs. The expected number of FIGs when counting gametes from one parent with FSD is always smaller than with HSD when only one parent is typed. When any locus has only two alleles and gametes from two parents are counted, FSD requires at least nine full-sibs per family to be more efficient than HSD when typing one parent. If at least one locus has four or more alleles and gametes from two parents are counted, FSD generally requires six or fewer full-sibs per family to be more efficient than HSD if one parent is typed and the number of offspring is 100 or less. Counting gametes from one parent, FSD requires at least nine fullsibs per family to outperform HSD when both parents are typed, and is never more efficient than HSD if any

Table 4 Family sizes required for full-sib design (FSD) to have equal LIC as half-sib design (HSD), assuming an equal total number of individuals genotyped and θ=0.20. n_h= number of half-sibs. f1=FSD when typing two parents per offspring and four grandparents and counting gametes from one parent, f2=FSD when typing two parents per offspring and four grandparents and counting gametes from both parents, h1=HSD when typing one parent per offspring, h2=HSD when typing two parents per offspring. Assumptions for FSD and HSD are described in the text. The number of families and the total number of offspring can be calculated using equations (15), (16), (17). The symbol "–" indicates that FSD cannot have more FIGs than HSD

Alleles at locus 1	Alleles at locus 2	n_h=50			n_h=100		
		f1/h2	f2/h1	f2/h2	f1/h2	f2/h1	f2/h2
2	2	–	36	32	–	47	36
2	4	–	9	8	–	9	8
2	6	–	9	7	–	9	7
2	8	–	9	6	–	6	7
3	3	57	6	5	65	6	5
4	4	18	5	3	18	6	5
5	5	13	5	3	13	5	3
6	6	10	5	3	11	5	3
7	7	9	5	2	10	5	3
8	8	9	5	2	9	5	2

of the two loci has two alleles. The family size required for HSD to outperform FSD can be calculated using equation (21). Counting gametes from one parent, FSD is less efficient than HSD if one parent is typed, indicating that FSD is always less efficient than HSD for generating a sex-specific map.

Total number of individuals to be typed to demonstrate linkage

Let T_{fak}=number of individuals (including sires, dams and offspring) to be typed for FSD, counting FIGs of a (a=1 or 2) parents when k parents (k=1 or 2) are typed; T_{hb}=number of individuals to be typed for HSD when b (b=1 or 2) parents are typed.

When females and males have the same recombination frequency, the total number of individuals to be typed for FSD is obtained by solving (18) for T:

$$T_{fak} = \frac{g_{fak}}{a(LIC_{fkl})H}\left[D + \frac{2+4H}{n_f}\right]. \quad (22)$$

For HSD, the total number of individuals to be typed is obtained by solving (19) for T:

$$T_{hb} = \frac{g_{hb}}{LIC_{hb}H}\left(bH + \frac{1}{n_h}\right) \quad \text{for } b = 1, 2\,. \quad (23)$$

From equations (22) and (23), the minimum value of g_{fak} and g_{hb} required to detect a given recombination frequency can be determined using equation (5.20) in Ott (1991). Assuming a LOD score of 3, and a power of 0.9, the minimum value of g_{fak} and g_{hb} required to detect a θ of 0.10, 0.20, or 0.30 is 28, 54, or 126, respectively. Using these numbers and equations (22) and (23), the numbers of typed individuals required for demonstrating linkage are given in Table 5. The results show that typing three generations under FSD requires typing the largest number of individuals when the numbers of alleles are small, and requires the least number of individuals when the numbers of alleles are large. For HSD, typing one parent always requires typing fewer individuals than typing two parents.

When females and males have different recombination frequencies, the numbers required for a sex-averaged map are obtained by substituting for $\bar{\theta}$ for θ (22), where $\bar{\theta}=1/2(\theta_f+\theta_m)$=the average recombination frequency of the two sexes, and θ_f=the recombination frequency for females, and θ_m=the recombination frequency for males. If there is a sex difference in θ (see Chapter 9 in Ott 1991), the numbers required to demonstrate linkage may be considerably larger in the sex with the larger θ. Assuming

Table 5 Total number of individuals to be typed to detect a given recombination frequency for full-sib design (FSD) and half-sib design (HSD). f2: FSD when typing two parents per offspring and four grandparents and counting gametes from two parents; h1: HSD when typing one parent per offspring; h2: HSD when typing two parents per offspring. Total number of typed individuals include offspring, parents and grandparents. For a sex-specific map with FSD, the total number of typed individuals required to detect a given recombination frequency is twice the number for f2. The number of full-sibs per family is nine, and is 50 for HSD. The exact number of families needed to achieve entries in this table can be obtained by solving equation (15), (16) and (17). A LOD score of 3 and a power of 0.9 are assumed.

Alleles at locus 1	Alleles at locus 2	θ=0.10			θ=0.20			θ=0.30		
		f2	h1	h2	f2	h1	h2	f2	h1	h2
2	2	158	96	106	323	210	218	789	546	533
2	4	78	73	82	156	153	164	374	383	395
2	6	72	69	81	144	142	161	344	351	385
2	8	71	67	81	140	138	161	337	338	385
3	3	56	61	75	111	128	150	266	321	359
4	4	43	51	70	85	105	138	204	259	331
5	5	38	46	67	75	94	134	180	230	320
6	6	35	44	66	70	88	131	167	214	315
7	7	33	42	65	67	84	130	159	203	311
8	8	32	40	65	64	81	129	154	195	309

θ_f=0.20 and θ_m=0.30, so that $\bar{\theta}$=0.25, then the number of individuals to be typed is larger than for the male map by 48% (eight alleles at each locus) to 86% (two alleles at each locus).

Discussion

Comparison with other methods

The expected number of FIGs (Chakravarti, 1991) and the LIC parameter developed here are designed to estimate the frequency of FIGs. The LIC parameter applies to various genotyping schemes in the parental and grandparental generations, and is applicable for arbitrary allele frequencies and values. The expected number of FIGs (Chakravarti 1991) considers unequal allele frequencies and can accommodate arbitrary θ values after slight modification to the formulation but applies to only one genotyping scheme, i.e., all individuals are required to have known genotypes. Calculations for the probability of being unable to detect linkage phase in Chakravarti (1991) and for the same probability presented here have two differences. In Chakravarti (1991, Table 3), the possible mating types between grandparents to produce the DH parent were conditional on the parent having a coupling or repulsion phase, and for some matings with two possibilities (e.g., the *AABB* or *AaBb* only one possibility was considered. In the calculations in the present study, conditioning was only on the parent being a DH individual and every possible mating to produce the DH parent was taken into account.

Implications for comparative mapping

A central goal of comparative mapping is to determine whether gene order and distance are phylogenetically conserved. Most genes on comparative maps are "type-I" markers, e.g., structural genes for proteins. The majority of these markers are diallelic. According to this study, HSD is more efficient than FSD for mapping such markers, even when the second locus is highly polymorphic, such as would be the case with a microsatellite (type-II) marker. Therefore, HSD may be more useful than FSD for tying together type-I and type-II marker maps (Beever et al. 1994).

Acknowledgements The authors thank J. E. Beever, J. I. Weller and the reviewers for helpful comments and suggestions. This study was supported in part by grants from the United States Department of Agriculture, grant No. IS-1939–91R from the Binational Agricultural Research and Development Fund, and grant No. 91-37205-6335 under the National Research Initiative.

References

Arnheim N, Li H, Cui X (1990) PCR analysis of DNA sequences in single cells: single sperm gene mapping and genetic disease diagnosis. Genomics. 8:415–419

Beek S van der, van Arendonk JAM (1993) Criteria to optimize designs for detection and estimation of linkage between marker loci from segregating populations containing several families. Theor. Appl. Genet. 86:269–280

Beever JE, Da Y, Ron M, Lewin HA (1994) A genetic map of nine loci on bovine chromosome 2. Mamm Genome 5:542–545

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

Chakravarti A (1991) Information content of the Centre d'Etude du Polymorphisme Humain (CEPH) family structure for linkage studies. Hum Genet 87:721–724

Cui X, Gerwin J, Navidi W, Li H, Kuehn M, Arnheim N (1992) Genecentromere linkage mapping by PCR analysis of individual oocytes. Genomics. 13:713–717

Lewin HA, Schmitt K, Hubert R, van Eijk MJT, Arnheim N (1992) Close linkage between bovine prolactin and BoLA-DRB3 genes: genetic mapping in cattle by single sperm typing. Genomics 13: 44–48

Lewin HA, Beever JE, Da Y, Faulkner DB, Hines HC (1994) The bovine B and C blood group systems are not likely to be the orthologues of human RH: an interesting twist in the comparative map. Anim Genet 25:13–18

Li H, Gyllensten UB, Cui X, Saiki RK, Erlich HA, Arnheim N (1988) Amplification and analysis of DNA sequences in single human sperm and diploid cells. Nature 335:414–417

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, Quillen J, Sheffield VC, Sunden S, Duyk GM, Weissenbach J, Gyapay G, Dib C, Morrissette J, Lathrop GM, Vignal A, White R, Matsunami N, Gerken S, Melis R, Albertsen H, Plaetke R, Odelberg S, Ward D, Dausset J, Cohen D, Cann H (1994) A comprehensive human linkage map with centimorgan density. Science 265:2049–2054

Ott J (1991) Analysis of human genetic linkage. Revised edition. The Johns Hopkins University Press, Baltimore London

Rohrer GA, Alexander LJ, Keele JW, Smith TP, Beattie CW (1994) A microsatellite linkage map of the porcine genome. Genetics 136:231–245

Ron M, Band M, Wyler A, Weller JI (1993) Unequivocal determination of sire allele origin for multiallelic microsatellites when only the sire and progeny are genotyped. Anim Genet 24: 171–176