J. M. Elsen · B. Mangin · B. Goffinet · C. Chevalet

Optimal structure of protocol designs for building genetic linkage maps in livestock

Received: 26 April 1993 / Accepted: 23 July 1993

Abstract We investigated protocol designs for gene mapping in livestock. The optimization of the population structure was based on the empirical variance of the recombination rate estimator. We concluded that a mixture of half-sib and full-sib families is preferred to half-sib families; a knowledge of parental phases does not improve the quality of the estimation for typical livestock families with five offspring or more; and measurements of the genotype of the mates in half-sib families are not useful. Graphs and algebraic approximations for the practical choice of family size and structure are given.

Key words Linkage map · Livestock · Simulation

Introduction

Evidence of polymorphisms at the DNA level has greatly stimulated interest in building genetic maps for livestock. Among the possible applications of these maps the most significant one is the detection of QTL (quantitative trait loci). While this idea is not new (Neimann-Sorensen and Robertson 1961), it has been recently developed, in particular by Soller and Genezi (1978), Weller et al. (1990) and Knott and Haley (1992).

While not valid for the human species, gene mapping of livestock may be done using experimental populations such that both the number of animals to be measured and the required population structure may be chosen. In practice, the experimental population will consist of a set of sire families, and its structure will be optimized by choosing the numbers of sires, dams/sire and progeny/dam. More precisely, the following questions will have to be answered:

1) Should the families be half sib or full sib (with possibly more than one mate per sire)?
2) Is it necessary to measure the grandparents?
3) In the case of half-sib families, is it necessary to measure the genotypes of the dams?

In this paper, we attempt to address these questions. The precision of the recombination rate estimator was the primary criterion used for a comparison of possible population structures.

Cases studied

Experimental designs compared

The population is a collection of F sire families, with D dams per sire and P progeny per dam. Two extreme cases should be noted:

1) The nuclear family case where $D = 1$
2) The half-sib family case where $P = 1$. This is the most frequent situation in cattle, as full sibs are mainly obtained by embryo transfer techniques.

Two parameters characterize the experimental designs:

1) The total number of measured individuals M being $M = F + FD + FDP$ if the dams are measured, and $M = F + FDP$ if the dams are not measured. This number M is directly linked to the laboratory cost of mapping. The protocol designs will thus be compared for a given M. Note that the costs of measuring the F sires are negligible compared to the total costs and that the extreme situation where neither the dams nor the sires are measured has little practical interest.
2) The total number of progeny $N = FDP$, which is more or less linked to the precision of the recombination rate estimator r.

It must be emphasized that the general structure we consider is a mixture of half-sib and full-sib families, with each sire mated to more than one dam, each giving more than one progeny.

Communicated by J. S. Beckmann

J. M. Elsen (✉) · B. Mangin · B. Goffinet · C. Chevalet
Institut National de la Recherche Agronomique, Centre de Recherches de Toulouse, BP 27, 31326 Castanet-Tolosan, France

Markers studied

Loci with 2 or 6 isofrequent alleles or with uniformly distributed frequencies were considered. The recombination rates between markers were either given the values 5%, 10%, 15% or 20% or were randomly chosen from a uniformly distributed frequency.

Methods for the choice of a family structure

The family structures will be compared here using the variance $var(\hat{r})$ of the maximum likelihood estimation $\hat{r}$ of the recombination rate r between two loci. In the very special backcross case, this variance is $r(1-r)/N$. For any precision $var(\hat{r})$, and thus for any size N of the theoretical backcross design, the cost of the practical design, which is proportional to M, may be minimized giving the optimal practical design.

Notations

The data are the marker genotypes of progeny (G_{ijk}), of their sire (G_i), and, in some cases, of their dam (G_{ij}). Alleles are always assumed to be codominant. The genotype G of an individual is given by the list of allele pairs at the first, A (a_1, a_2) and second, B (b_1, b_2), locus, respectively.

The likelihood of a progeny genotype depends on the phases of the parents. We thus distinguish from the genotype $G=(a_1\,a_2\,b_1\,b_2)$ of an individual, its haplotype combination H, which is either $[a_1 b_1/a_2 b_2]$ or $[a_1 b_2/a_2 b_1]$.

Likelihood expression

The population likelihood $L(r)$ is given by:

$$L(r)=\prod_{i=1}^{F}\sum_{H_i}P(H_i/G_i)\prod_{j=1}^{D}\sum_{H_{ij}}P(H_{ij}/G_{ij})\prod_{k=1}^{P}P(G_{ijk}/H_i,H_{ij}) \qquad (1)$$

We are here interested in progeny observation (G_{ijk}) likelihood given the parent observations, which are either their genotypes (G_i, G_{ij}) or, when the phases are known from the ancestors, their haplotype combinations (H_i, H_{ij}). In the latter case, likelihood $L(r)$ is simply:

$$L(r)=\prod_{i=1}^{F}\prod_{j=1}^{D}\prod_{k=1}^{P}P(G_{ijk}/H_i,H_{ij}) \qquad (2)$$

The probabilities of haplotype combinations $P(H/G)$ (omitting temporarily indices i and j) are given the values 1/2 or 0 when the unordered list of H alleles is either identical or not identical to the list of G alleles. The way the progeny genotype probabilities $P(G_{ijk}/H_i, H_{ij})$ are calculated may be found in Ott (1991).

Comparison methods

The variance of the recombination rate estimator $\hat{r}$ was obtained using simulations. Each case studied was defined by the population structure (F, D, P), the measurements (dams may be measured or not, and grandparents may or may not be measured in order to get the parent phases) and marker characteristics (true recombination rate r, allele number and frequency). Each case was repeated 500 times, and the variance of the recombination rate estimator $var(\hat{r})$ was estimated by the variance of the empirical distribution of $\hat{r}$.

Another criterion for the protocol design comparison was the probability that the markers were correctly ranked on the genome. This probability was estimated for the three-locus case, with either 6, 6 and 2 or 6, 2 and 6 isofrequent alleles. Here, the recombination rates were randomly chosen in a uniform distribution between 0.1 and 0.3, given the constraint that $r_{12}+r_{23}-2r_{12}r_{23}$ was less than 50%.

For each simulation, the maximum likelihood rates r_{12} and r_{23} were obtained assuming that the parental phases were known, either with a 2-point analysis, repeated three times for the three possible pairs of markers (in which only markers on each side of the interval are taken into account) or with a 3-point analysis (where the two rates r_{12} and r_{23} are simultaneously estimated by maximizing the likelihood of the three-marker's genotypes).

Results

Comparison of the structures after the recombination rate estimator variance

Figure 1 shows the marker type effect on $var(\hat{r})$ for two population sizes with ten sires mated to a variable but entire number of dams, each dam giving ten progeny. As expected, the variance increased with the recombination rate and when the number of alleles per locus decreased. Results concerning two loci with 2 and 6 alleles, respectively, were the mean and were probably the most frequent situation. Subsequently, we only present this case with a 20% recombination rate; the other situations studied gave similar results.

Do we need full-sib or half-sib families?

Table 1 gives the standard deviation of $\hat{r}$ distribution for different values of M (number of measurements), F (number of sires) and P (number of progeny per dam). It is quite clear that (1) strict half-sib families are much less informative than a mixture of half- and full-sib families – thereafter called "full-sib families" – (and have a standard deviation $\sqrt{var(\hat{r})}$ twice as high); and (2) concerning full-sib families, the $\hat{r}$ variance is not very sensitive to the family structure defined by F, D and P(it should be noted that it is best to measure a small number of large families).

Fig. 1 Effect of the type of gene on $var(\hat{r})$

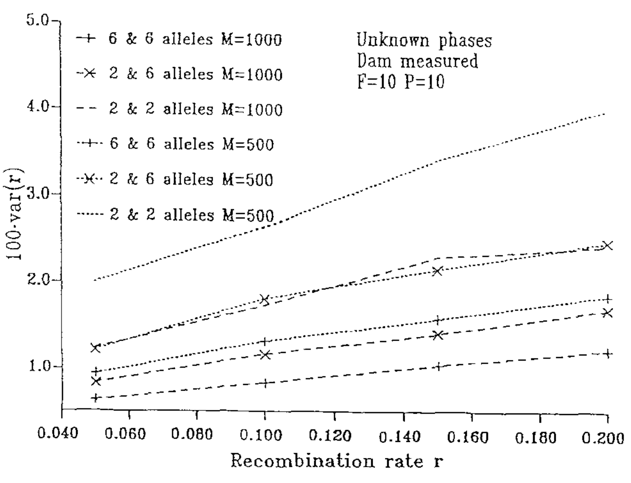

Table 1 Effect of the family structure on the precision $\sqrt{var(\hat{r})}$ of the recombination rate estimator (M Number of measurements, F number of sires, P number of progeny per dam)

M	F	P				M	F	P			
		1	5	10	15			1	5	10	15
	5	4.87	2.56	2.41	2.34		5	2.94	1.51	1.34	1.39
		100[a]	*20*	*10*	*7*			*300*	*60*	*30*	*20*
500	10	5.30	2.58	2.47	2.41	1500	10	2.88	1.47	1.29	1.34
		50	*10*	*5*	*3*			*150*	*30*	*15*	*10*
	20	5.06	2.62	2.54	2.74		20	2.77	1.45	1.40	1.38
		25	*5*	*3*	*1*			*75*	*15*	*7*	*5*
	5	3.43	1.73	1.78	1.97		5	2.51	1.43	1.27	1.34
		200	*40*	*20*	*13*			*400*	*80*	*40*	*27*
1000	10	3.67	1.69	1.69	1.52	2000	10	2.69	1.33	1.22	1.28
		100	*20*	*10*	*7*			*200*	*40*	*20*	*13*
	20	3.51	1.80	1.79	1.72		20	2.26	1.25	1.14	1.18
		50	*10*	*5*	*3*			*100*	*20*	*10*	*7*

[a] The number of dams is given in italics for each combination

Table 2 Effect of the family structure on the probability of a correct ranking of three loci linked according to the locus polymorphisms and classification criteria (FS Full-sib family, HS half-sib family)

	6, 6 and 2 alleles				
Analysis	Type of family				
	8 FS	19 HS	29 HS	34 HS	39 HS
Two-point	0.981	0.947	0.9735	0.986	0.9875
Three-point	0.992	0.9675	0.985	0.9945	0.997
	6, 2 and 6 alleles				
Analysis	Type of family				
	8 FS	19 HS	29 HS	34 HS	39 HS
Two-point	0.983			0.9835	0.99
Three-point	0.9945			0.9855	0.994

These results are partly due to our constraint of a fixed number of measurements M. Indeed, variance of $\hat{r}$ was approximately inversely proportional to the number of measured progeny, which was close to $M(P/P+1)$ with $M \approx FD$ $(1+P)$. Compared to an infinitely large full-sib family, variance of $\hat{r}$ increased by a factor of 2 for a half-sib family, this factor decreasing rapidly towards 1 when P increased (1.2, 1.1 and 1.07 with $P = 5, 10$ and 15 progeny per dam). Another explanation of $\hat{r}$ variance in relation to the type of family is that with full-sib families, dam meiosis is usable for estimation of the recombination rate. This possibility would increase the amount of available information by about 50% (Ott 1991).

These results are confirmed by the probability of a correct ordering for three loci. We looked for the equivalence between a design with $M = 300$ individuals in a population of 30 full-sib families of size 8 and a half-sib protocol giving the same ranking probability. The results are given in Table 2 for the 2-allele configurations and two ranking criteria. Three times more measurements were needed in the half-sib protocol than in the full-sib protocol.

Is it useful to know the parental phases?

The way in which the precision of the recombination rate estimator increased when the parental phases were known *a priori* is illustrated in Fig. 2 for a population of about $M = 1000$ (the total number of dam D is adjusted to give the total number M closest to 1000) and various levels for F and P. The increase was large for half-sib families, but negligible for full-sib families. Thus, measurement of grandparent genotypes seems to be useless for this type of structure, partly because the phase of the parents cannot always be deduced from these observations.

Is it useful to measure the genotype of the dam for half-sib families?

In these families there are as many dams as there are progeny. Thus, avoiding dam measurements increases by a factor of 2 the number of progeny observed. Another argument against dam measurements is that unless the grandparents are measured (and this is very expensive), the phases of the dams cannot be ascertained and the dams' meioses cannot be used for linkage estimation.

The effect of dam genotype measurement on the precision of the recombination rate estimator is given in Fig. 3 in a design with ten sires. When the loci were highly polymorphic (with 6 alleles for each locus) the dam genotype should not be measured, the sire gamete being mostly identified from progeny genotypes. When there were only 2 alleles at each locus, it was better to measure the dam genotype as this information greatly increased the number of progeny for which the origin of the genes is known. On the whole, however, there were only small differences between the two options.

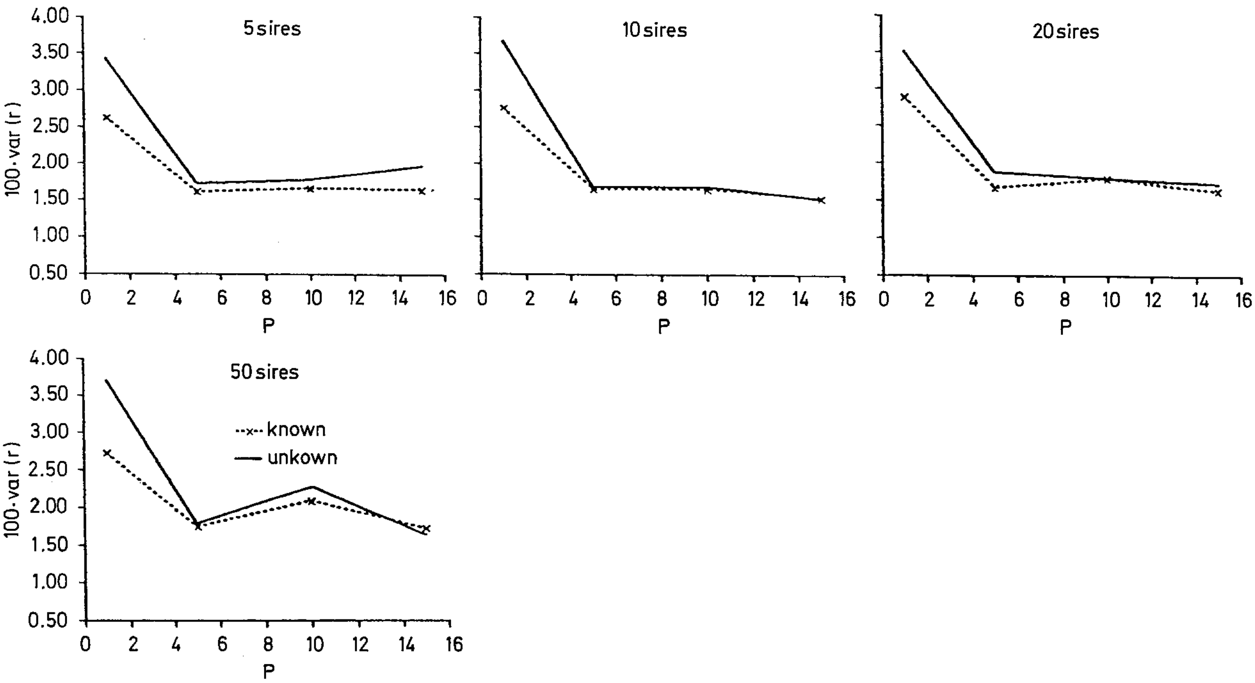

Fig. 2 Effect of the knowledge of phases on $var(\hat{r})$: dam measured, 2 and 6 alleles, $r = 0.20$

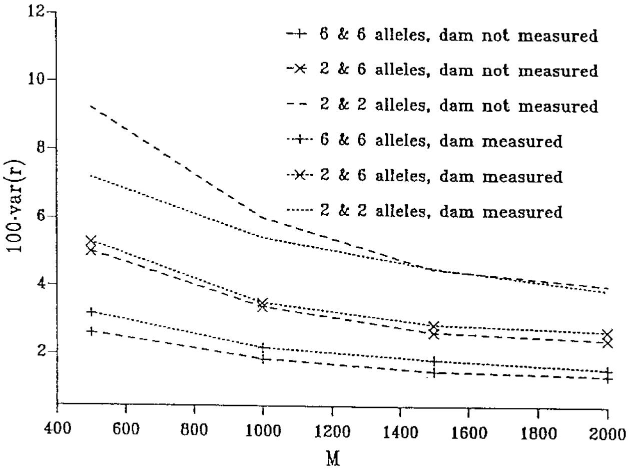

Fig. 3 Effect of the measurement of dams on $var(\hat{r})$: $F = 10$, $P = 1$, $r = 0.20$ unknown phases

Approximation of the variance of the recombination rate estimator

The previous results were obtained using simulations, with probably a small accuracy in spite of intensive computations. In order to generalize the results and to offer a rapid computation technique, we tested an approximation method. It is based on the observation of the very limited effect of knowledge of parental phases on the accuracy of the recombination rate estimation. We suggest to assume that the sire phase is known and to approach the $\hat{r}$ variance by the sample amount of information. In practice, let $\mathscr{G}_{ij} = \{H_i, G_{ij}\}$ be the vector of the genotype of parents ij, and D_{ij}, the vector of the genotypes of their progeny. Then, the variance of $\hat{r}$ will be:

$$var(\hat{r}) \approx \frac{1}{FD} var(\tilde{r}) = \frac{1}{FD} \sum_{\mathscr{G}_{ij}} p(\mathscr{G}_{ij}) var(\tilde{r}/\mathscr{G}_{ij}) \qquad (3)$$

where $\tilde{r}$ is the recombination rate estimation based on a nuclear family:

$$var(\tilde{r}/\mathscr{G}_{ij}) \approx -1/E\left[\frac{\partial^2 \log L(r)}{\partial r^2} / \mathscr{G}_{ij}\right] = 1/E\left[\left(\frac{\partial \log L(r)}{\partial r}\right)^2 / \mathscr{G}_{ij}\right]$$
$$= 1/I_{\mathscr{G}_{ij}}. \qquad (4)$$

The amount of information is (e.g. Green 1981, appendix 3)

$$I_{\mathscr{G}_{ij}} = \sum_{\mathscr{D}_{ij}} \frac{\left(\dfrac{\partial\, Prob(\mathscr{D}_{ij}/\mathscr{G}_{ij})}{\partial r}\right)^2}{Prob(\mathscr{D}_{ij}/\mathscr{G}_{ij})} \qquad (5)$$

Since the dam's phase is unknown, we get:

$$Prob(\mathscr{D}_{ij}/\mathscr{G}_{ij}) = \sum_{h} Prob(H_{ij} = h/G_{ij})\, Prob(\mathscr{D}_{ij}/H_i, H_{ij} = h) \quad (6)$$

The summation in h concerns only double heterozygous dams. We assumed linkage equilibrium, thereby fixing $Prob(H_{ij} = h/G_{ij}) = 1/2$ for the two possible haplotype combinations H_{ij}.

The genotypes of the progeny with an equal probability may be pooled in classes following the Ott approach (1991, pp 90–91). The probability of the genotypes of the progeny given the parental haplotypes' combination is given by:

$$Prob(\mathscr{D}_{ij}/H_i, H_{ij} = h) = \left(\prod_{c=1}^{C} p_{hc}^{D_C}\right) \cdot \frac{D!}{D_1! \ldots D_C!} \qquad (7)$$

where h is the haplotype combination of the dam, c is the index of a genotype class with identical probability ($c = 1, \ldots, C$),

Table 3 Components needed for approximation of the recombination rate estimator variance ($\mathscr{G}_{ij}$ Parents' genotype, H_i sire haplotype combination, $n_1 (n_2)$ number of alleles at the first (second) locus, G_{ij} dam genotype, C number of classes of progeny genotype, h_i phasis of the dam, r recombination rate. Alleles differing by their indices (a'_1 and a'_2) or quotes (a'_1 and a''_1) are different. Alleles without quotes are indifferent)

Parents' genotype $\mathscr{G}_{ij}$		$prob(\mathscr{G}_{ij}) \Big/ \dfrac{4(n_1-1)(n_2-1)}{n_1^3 n_2^3}$	h	Class probability ($p_{hc}, c = 1, \ldots, C$)					
H_i	G_{ij}			p_{h1}	p_{h2}	p_{h3}	p_{h4}	p_{h5}	p_{h6}
$[a'_1 b'_1/a'_2 b'_2]$	$(a''_1 b''_1 a''_2 b''_2)$	$\left(\dfrac{n_1(n_1-1)}{2}-1\right)\left(\dfrac{n_2(n_2-1)}{2}-1\right)$	1	$(1-r)^2$	$r(1-r)$	$r(1-r)$	r^2		
			2	$r(1-r)$	$(1-r)^2$	r^2	$r(1-r)$		
$[a'_1 b'_1/a'_2 b'_2]$	$(a'_1 b''_1 a'_2 b''_2)$	$\dfrac{n_1(n_1-1)}{2}+\dfrac{n_2(n_2-1)}{2}-2$	1	$\dfrac{(1-r)^2}{2}$	$\dfrac{r(1-r)}{2}$	$r(1-r)$	$\dfrac{r^2+(1-r)^2}{2}$	$\dfrac{r(1-r)}{2}$	$\dfrac{r^2}{2}$
			2	$\dfrac{r(1-r)}{2}$	$\dfrac{(1-r)^2}{2}$	$\dfrac{r^2+(1-r)^2}{2}$	$r(1-r)$	$\dfrac{r^2}{2}$	$\dfrac{r(1-r)}{2}$
$[a'_1 b'_1/a'_2 b'_2]$	$(a'_1 b'_1 a'_2 b'_2)$	1	1	$\dfrac{(1-r)^2}{2}$	$2r(1-r)$	$\dfrac{r^2+(1-r)^2}{2}$	$\dfrac{r^2}{2}$		
			2	$\dfrac{r(1-r)}{2}$	$r^2+(1-r)^2$	$r(1-r)$	$\dfrac{r(1-r)}{2}$		
$[a'_1 b'_1/a'_1 b'_2]$	$(a''_1 b''_1 a''_2 b''_2)$	$\dfrac{n_2 n_1(n_1-1)+n_1 n_2(n_2-1)}{4}-1$	1	$1-r$	r				
			2	r	$1-r$				
$[a'_1 b'_1/a'_1 b'_2]$	$(a_1 b''_1 a_1 b''_2)$	$\dfrac{n_2 n_1(n_1-1)+n_1 n_2(n_2-1)}{4}-1$	1	$1-r$	r				
$[a_1 b'_1/a_1 b'_2]$	$(a''_1 b'_1 a''_2 b'_2)$	$\dfrac{n_2+n_1}{2}$	1	$\dfrac{1-r}{2}$	$\dfrac{r}{2}$	$\dfrac{1}{2}$			
			2	$\dfrac{r}{2}$	$\dfrac{1-r}{2}$	$\dfrac{1}{2}$			
$[a'_1 b'_1/a'_2 b'_2]$	$(a_1 b'_1 a_1 b'_2)$	$\dfrac{n_2+n_1}{2}$	1	$\dfrac{1-r}{2}$	$\dfrac{r}{2}$	$\dfrac{1}{2}$			
$[a_1 b_1/a_1 b_1]$	$(a''_1 b''_1 a''_2 b''_2)$	$\dfrac{n_2 n_1}{4}$	1	$1-r$	r				
			2	r	$1-r$				
$[a'_1 b'_1/a'_2 b'_2]$	$(a_1 b_1 a_1 b_1)$	$\dfrac{n_2 n_1}{4}$	1	$1-r$	r				

D_c is the number of observed progeny in this genotype class ($\Sigma_c D_c = D$) and p_{hc} is the probability of this genotype class, given the haplotype combination h as a function of the recombination rate r.

The first derivation follows:

$$\frac{\partial Prob(\mathscr{D}_{ij}/\mathscr{G}_{ij})}{\partial r} = \frac{1}{2}\left(\sum_h Prob(H_{ij} = h/G_{ij}) \cdot Prob(\mathscr{D}_{ij}/H_i, H_{ij} = h) \cdot \sum_{c=1}^{C} D_c \cdot q_{hc} \right) \quad (8)$$

with $q_{hc} = (\partial p_{hc}/\partial r)/p_{hc}$.

All information required for the computation of $var(\hat{r})$ when the alleles are equiprobable is given in Table 3. The nine different subparts of the table describe all the informative combinations of parental genotypes $\mathscr{G}_{ij}$. The $\mathscr{G}_{ij}$ summation in Eq. 3 will thus include nine differents parts. The list of classes is not given explicitly in Table 3 but can be found easily from the corresponding probability p_{hc}. For instance, in $[a'_1\, b'_1/a'_2\, b'_2] \times (a_1\, b''_1 a_1\, b''_2)$ matings, eight progeny types are possible, given two classes: non-recombinant, with $p_{hc} = 1 - r$: $(a_1 b'_1 a_1 b''_1)$, $(a_1 b'_1 a_1\ b''_2), (a_2 b'_2 a_1 b''_1), (a_2 b'_2 a_1 b''_2)$; and recombinant, with $p_{hc} = r$: $(a_1 b'_2 a_1 b''_1), (a_1 b'_2 a_1 b''_2), (a_2 b'_1 a_1 b''_1), (a_2 b'_1 a_1 b''_2)$.

The variance of the recombination rate was calculated with our approximation method and by simulations (10 000 samples per case) for a population with $F = 20$ sire families each with D dams and with $P = 1, 2, \ldots, 10$ progeny per dam, the number D being adjusted to a maximum total number of measurements $M = 500$. Two recombination rates 5% and 20% were considered for two loci with 2 and 6 iso-frequent alleles. Figure 4 shows the similarity of the two estimations of $\hat{r}$ variance, which were only significantly different for half-sib families ($P = 1$) with a 20% recombination rate. It must be noted that the fluctuations shown in the figure were derived from the variation of the total number of measurements M due to the adjustment of D to a whole number.

As an example, Fig. 5 gives the standard error of the recombination rate for families of 2–20 progeny per dam with a rate r ranging from 5% to 20%. The standard error for a population of F sires with D dams can be obtained by dividing the result by $\sqrt{FD}$.

Discussion – conclusion

We found that the population is optimumly structured as sire families with subsets of "full-sib families" and that measuring

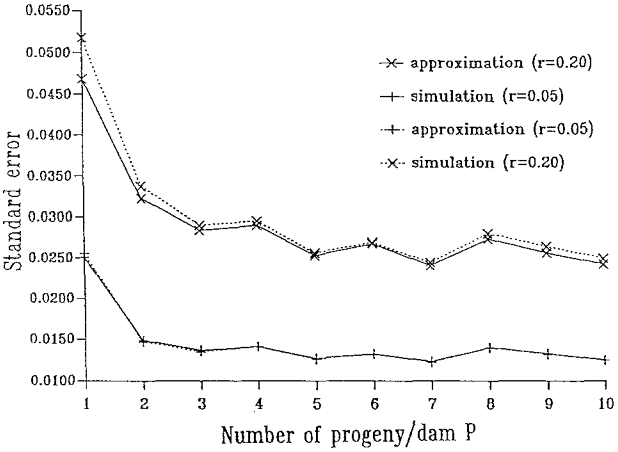

Fig. 4 Standard deviation of the recombination rate. Comparison between simulations and proposed approximation: 2 and 6 alleles, $F = 20$

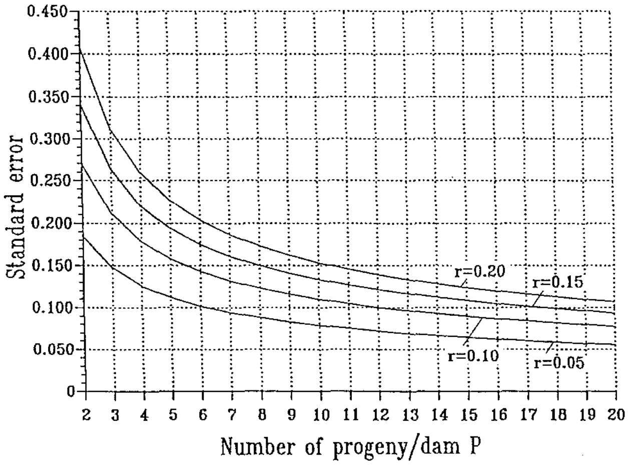

Fig. 5 Approximate standard deviation of the recombination rate estimator: 2 and 6 alleles

the phenotypes of the grandparents is not necessary for sibship of about five offspring or more. The approximate formulae for recombination rate variance estimator we gave allows a rapid comparison of possible structures.

The criteria chosen (recombination rate variance estimator) is questionable. The quality of a design should be measured in relation to our objective of building the genetic map. Different purposes may be assigned (localization and identification of major genes, tools for keeping genetic variability or for introgression). As mentioned in the introduction, the detection of QTLs is probably the most important. With respect to this objective, the optimal size N of the reference population used for the building of the genetic map may be estimated in a simple backcross situation. In practice, for livestock populations, a design of similar value will have to be based on a larger number of animals. We suggest using the variance of the recombination rate estimate as an equivalence criteria. Thus, an optimal backcross population of 100 progeny gives a variance of $\sqrt{0.2 \times 0.8/100} = 4\%$ for a linkage at a 20% recombination rate. Using our approximate formulae for recombination rate variance from Fig. 5, we could propose as an optimal design a population of $F = 5$ sire families, with $D = 5$ dams/sire and $P = 6$ progeny/dam ($N = 150$ and $M = 180$) or a population of $F = 8$, $D = 5$, $P = 4 (N = 160, M = 208)$.

Acknowledgments We thank the French "MMM" (Mathematics and molecular markers) group for helpful discussions about this paper.

References

Green EL (1981) Genetics and probability in animal breeding experiments. MacMillan Publ, London Basingstole

Knott SA, Haley CS (1992) Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet Res 60:139–161

Niemann-Sorensen A, Robertson A (1961) The association between blood groups and several production characteristics in three Danish cattle breeds. Acta Agric Scan 11:163–196

Ott J (1991) Analysis of human genetic linkage. The John Hopkins University Press, Baltimore

Soller M, Genizi A (1978) The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. Biometrics 34:47–55

Weller JL, Kashi Y, Soller M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J Dairy Sci 73:2525–2537