High-Resolution Mapping of Quantitative Trait Loci by Selective Recombinant Genotyping

Y. Ronin,* A. Korol,*,[1] M. Shtemberg,* E. Nevo* and M. Soller†

**Institute of Evolution, University of Haifa, Mount Carmel, 31095 Haifa, Israel and †Department of Genetics, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel*

Manuscript received November 10, 2002
Accepted for publication April 21, 2003

ABSTRACT

Selective recombinant genotyping (SRG) is a three-stage procedure for high-resolution mapping of a QTL that has previously been mapped to a known confidence interval (target C.I.). In stage 1, a large mapping population is accessed and phenotyped, and a proportion, P, of the high and low tails is selected. In stage 2, the selected individuals are genotyped for a pair of markers flanking the target C.I., and a group of R individuals carrying recombinant chromosomes in the target interval are identified. In stage 3, the recombinant individuals are genotyped for a set of M markers spanning the target C.I. Extensive simulations showed that: (1) Standard error of QTL location (SEQTL) decreased when QTL effect (d) or population size (N) increased, but was constant for given "power factor" ($PF = d^2N$); (2) increasing the proportion selected in the tails beyond 0.25 had only a negligible effect on SEQTL; and (3) marker spacing in the target interval had a remarkably powerful effect on SEQTL, yielding a reduction of up to 10-fold in going from highest (24 cM) to lowest (0.29 cM) spacing at given population size and QTL effect. At the densest marker spacing, SEQTL of 1.0–0.06 cM were obtained at PF = 500–16,000. Two new genotyping procedures, the half-section algorithm and the golden section/half-section algorithm, allow the equivalent of complete haplotyping of the target C.I. in the recombinant individuals to be achieved with many fewer data points than would be required by complete individual genotyping.

LOW resolution of the estimated chromosomal location of quantitative trait loci (QTL) is a major obstacle in application of QTL linkage mapping results for marker-assisted selection and comparative positional cloning of the gene corresponding to the QTL. Up to a certain point, mapping resolution (defined as the standard deviation of estimated QTL location, or SEQTL) can be improved by increasing marker density (Darvasi *et al.* 1993). However, for given sample size and standardized QTL substitution effect, ultimate map resolution is fixed and cannot be further improved even with infinite marker density (Darvasi *et al.* 1993; Darvasi and Soller 1997). Consequently, approaches to improving QTL map resolution primarily involve increasing the standardized QTL substitution effect, *e.g.*, by using replicated progenies (Soller and Beckmann 1990; Weller *et al.* 1990), by including the effects of cosegregating QTL as regression cofactors (Zeng 1994; Jansen and Stam 1994), or by employing multiple-trait analysis (Jiang and Zeng 1995; Korol *et al.* 1995, 2001). More complex approaches, termed "genetic chromosome dissection," involve producing or identifying recombinants in the chromosomal intervals shown to carry significant QTL and evaluating the recombinant chromosomes through progeny testing (Darvasi 1997a, 1998; Hill 1998; Soller and Andersson 1998). Effective sample size can also be increased by accumulating recombinants in advanced generations (Darvasi and Soller 1995).

The most straightforward method for increasing mapping resolution, however, is simply to increase the size of the mapping population, in this way accumulating recombinants in the interval of interest. When this strategy is employed, a useful tactic for reducing genotyping costs has been to produce a mapping population with easily scorable morphological markers flanking the interval containing the target locus. A few hundred recombinant individuals for these markers are identified, and only these individuals are genotyped for the set of closely spaced molecular markers spanning the target interval (*e.g.*, Klein-Lankhorst *et al.* 1991; see also Rhodes *et al.* 1998 and review in Darvasi 1998).

Here we propose a similar procedure, selective recombinant genotyping (SRG), to be applied when the target locus is a QTL with a moderate or even small substitution effect. In this case, mapping resolution is expressed as the C.I. or as the SEQTL. SRG would ordinarily be implemented following an initial total or partial genome scan that has detected a QTL in a backcross (BC), F_2, or half-sib sire-family design. It presupposes the possibility of forming or accessing a very large mapping population. In addition, we present two new genotyping procedures, the half-section algorithm and the golden section/half-section algorithm, which allow the equiva-

[1] *Corresponding author:* Institute of Evolution, University of Haifa, Mount Carmel, 31095 Haifa, Israel. E-mail: korol@esti.haifa.ac.il

lent of complete haplotyping of the target C.I. in the recombinant individuals to be achieved with many fewer data points than would be required by complete individual genotyping.

The procedure described here is similar in conception to the "contrast mapping" procedure of Thaller and Hoeschele (2000). The present study generalizes and extends their results by considering BC and F_2 populations and the effects of selective genotyping and marker spacing on the accuracy of QTL location. The results are also presented in a form somewhat different from that used by Thaller and Hoeschele (2000), namely, as SEQTL rather than as the proportion of QTL located to the true QTL interval. However, the present study amply supports the bottom line conclusion of Thaller and Hoeschele (2000), namely, with large family sizes "it is feasible to map a QTL to a region of 2 to 4 cM" (p. 103).

THEORY

Selective recombinant genotyping: We assume a situation in which QTL mapping by any of the customary procedures (complete individual genotyping, selective genotyping, selective DNA pooling) has detected a QTL in a confidence interval defined by a pair of flanking markers, M_1 and M_k. It is further assumed that a set of additional evenly spaced ordered markers (denoted M_2, . . . , M_i, . . . M_{k-1}) spanning the interval from M_1 to M_k are available and that haplotypes of the parental lines or individual sires have been determined with respect to the entire set of markers. In the proposed scheme, high-resolution mapping is based on genotyping the markers M_2–M_{k-1} only for those individuals from the high and low population tails that are recombinant for the flanking markers. Thus, if the parental F_1 or sire chromosomes are M_1M_k/m_1m_k, the progeny individuals chosen for further genotyping will be those that carry recombinant parental chromosomes M_1m_k and m_1M_k. The main question concerns the degree to which the SEQTL depends on the standardized QTL allele substitution effect d, on the total size of the mapping population (N), and on marker spacing ($\boldsymbol{c}$ in centimorgans) in the interval M_1–M_k. In addition, Darvasi (1997b) has shown that most of the information for QTL map location is found in the high and low tails of the mapping population. To explore this possibility of reducing genotyping costs, we also studied the effect of genotyping only the high and low proportions (P) of the population for the initial recombinants.

To address these questions, a Monte Carlo analysis was employed. Standard interval maximum-likelihood (ML) analysis was used combined with selective genotyping that uses trait values of both genotyped and nongenotyped individuals to provide ML estimates of the QTL effect and position (Lander and Botstein 1989; Ronin *et al.* 1998). The interval analysis was unconditional, with no prior assumption of the QTL location. The simulated QTL was located at the center of a chromosome of a total length of 480 cM, so that end effects did not limit the SEQTL. Each of the tails was composed of t individuals, so that $P = t/N$. Then, for each marker subinterval $M_i - M_{i+1}$ ($i = 1, \ldots, k - 1$) from the interval M_1–M_k, the conditional LOD score was calculated, assuming that the QTL resides in this subinterval. The estimates of the QTL effect and residual variance obtained in the initial analysis for the entire M_1–M_k interval were used as coordinates of the starting point in the optimization procedure for each subinterval. It was assumed that all individuals in the high and low selected groups had been genotyped for markers M_1 and M_k and that the M_1–m_k and m_1–M_k recombinants had been identified.

Width of the M_1–M_k interval was taken as 24 cM; QTL location was at the midpoint of the interval. It was assumed that mapping takes place within a backcross or half-sib population, so that contrast values in SD units are d (or α, in the case of a half-sib population). The following parameter combinations were investigated in the main body of the simulations: $d = 0.25$, 0.50, 1.00; $N = 1000$, 2000, 4000, 8000, 16,000; $P = 0.05$, 0.10, 0.20, 0.25, 0.50; $c = 24$, 8, 2.66, 0.88, 0.29 (marker spacing was chosen to ensure that in no instance did a marker position coincide with a QTL position). For a BC population, $d = 0.25$, and 0.50 and 1.00 correspond to QTL variances of 0.015, 0.0625, and 0.25, respectively. For each combination of parameters, 1000 Monte Carlo runs were conducted. The direct empirical value of the SEQTL was calculated on the basis of the estimated values of QTL location.

Genotyping requirements: Genotyping requirements will differ somewhat depending on whether the SRG procedure is implemented in a BC, F_2, or half-sib family (half-sib) designs. For clarity, a complete analysis is provided first for the BC design, and modifications required by F_2 and half-sib designs are then discussed. It is convenient to organize the genotyping requirements according to the three steps of the SRG fine-mapping procedure. Table 1 provides a summary of genotyping requirements for the three designs, according to these steps, and for total genotyping.

BC design

Step I. Identifying recombinant offspring: The proposed procedure is based upon individual genotyping of the entire selected sample to identify recombinant individuals in the region M_1–M_k. This will involve $4NP$ data points = $2NP$ individuals × 2 data points/individual (data point: the genotype of a single individual with respect to a single marker) and will identify $R = r(2NP)$ recombinants, where r is the proportion of

TABLE 1

Summary of genotyping requirements for application of the SRG procedure within BC, F_2, and half-sib designs

	Genotyping requirements			
Design	Parental haplotyping	Identifying recombinants	Genotyping recombinants[a]	Total
BC	$2\ M_L$	$4PN$	$2LPN\ M/100$	$2PN\ (2 + LM/100) + 2M_L$
F_2	$2\ M_L$	$2PN$	$2LPN\ M/100$	$2PN\ (1 + LM/100) + 2M_L$
Half-sib	3–13 M_L	$8PN$	$4LPN\ M/100$	$2PN\ (4 + 2LM/100) + (3\text{–}13)M_L$

L, target interval in centimorgans; assuming small L, proportion of recombination across target interval, $\sim L/100$; M_L, number of internal markers; N, total population size; P, proportion of each tail taken for selective genotyping.

[a] For complete genotyping, $M = M_L$; for HS genotyping, $M = 4$ or 5; for GS-HS genotyping, $M = 3$.

recombination between markers M_1 and M_k. For small target intervals of length L cM, $r \sim L/100$. Parental haplotypes for the flanking markers are obtained in the course of identifying recombinant individuals.

Step II. Determining the parental haplotypes with respect to the internal markers: This step is needed to identify the complete marker genotype for each individual as required for the interval mapping procedure. Given a segment of length L, the number of additional markers needed within the segment to provide marker spacing c, is given by $M_L = (L/c) - 1$. Determining parental haplotypes for F_2 or BC designs is simply achieved by genotyping the parental lines. Thus, the number of genotyping data points required in this case will be $2M_L$.

Step III. Genotyping recombinant individuals for the markers within the target segment: Once parental haplotypes are known, each recombinant individual is genotyped for all internal markers. The total genotyping data points for the recombinant individuals will thus equal $LNPM_L/100$.

F_2 design

Since an F_2 individual can receive a recombinant haplotype from either of the two parents, the proportion of F_2 recombinant individuals is twice that of a comparable BC population. Most F_2 progeny will carry only a single recombinant chromosome. For these individuals, analysis is the same as for a BC design. Some of the F_2 progeny will be double recombinants. These will be of two sorts: (1) double recombinants involving opposite-phase haplotypes (*i.e.*, $M_1 \ldots m_k/m_1 \ldots M_k$), which will not be recognized as recombinants in the initial screen for recombinant progeny and will not be included among the recombinant progeny and (2) double recombinants involving same-phase haplotypes (*i.e.*, $M_1 \ldots m_i \ldots m_j \ldots m_k/M_1 \ldots M_i \ldots m_j \ldots m_k$). These will be included among the recombinant progeny and will carry twice as much information as a single recombinant. Thus, in an F_2 design overall, the total number of progeny genotyped, and hence requiring genotyping data points to identify the recombinant individuals, will be half that for a BC design. Once the recombinant individuals are identified, however, genotyping requirements are more or less the same as for the BC design, although double recombinants will require some additional data points to establish both points of recombination.

Half-sib design

In principle, a half-sib design is the exact equivalent of a BC design, in that any individual progeny will receive a recombinant chromosome from only one parent (the sire). However, they differ in that, in a BC design, all markers are fully informative, because the allele derived from the F_1 parent can be identified unequivocally, and hence the recombinant individuals and their haplotypes at each marker are determined by genotyping that marker. This is not the case for the half-sib design, because of the incomplete informativity of the individual markers in an outcrossing population. That is, when an individual has the same (heterozygous) genotype as its sire, it is not possible to determine the marker allele transmitted to the individual from its sire. In this case, the genotyping data point will not be informative for determining recombinant status of the haplotype transmitted from the sire to the individual. The same will hold when the dam is genotyped, if individual sire and dam all share the same (heterozygous) genotype. This applies both to the initial step of identifying progeny that received recombinant haplotypes from their sire and to the step of identifying the full haplotype of the recombinant individual. The easiest way around this is to genotype additional markers close to the initially chosen marker. Assuming conservatively that only 50% of genotypings are informative, it is easy to see that the total number of genotypings required to identify and haplotype the recombinant progeny is double that required for the BC or the F_2 situation.

In addition to the above, obtaining the sire haplotype is also affected by incomplete informativity of the markers. In this case, haplotype of the sire for the flanking markers will be obtained from the many progeny that are genotyped in the screen for recombinant progeny. With respect to the internal markers, DNA will often be available for one or both parents of the sire. In this case, genotyping the sire, his sire, and his dam for the internal markers, *i.e.*, 3 M_L genotyping data points, will provide the sire haplotypes for all markers except those for which the sire and his parent(s) are heterozygous for the same pair of alleles. For these markers, it will be necessary to genotype progeny of the sire. For this, it will be efficient to use the nonrecombinant progeny, already identified as described in the preceding screen for recombinants. Because nonrecombinant progeny are used and the phase of the flanking markers is known, a single individual will provide a sire haplotype for all sire markers, except those for which sire and progeny are heterozygous for the same pair of alleles. Since maximum heterozygosity for the same pair of alleles is 0.5, 10 nonrecombinant individuals should easily be sufficient for haplotyping a sire. These individuals will need to be genotyped only for M_L internal markers. Thus, haplotyping a sire will require 3–13 M_L genotyping data points.

The half-section algorithm

The total number of genotyping data points can be reduced greatly by assuming that all M_1–m_k, m_1–M_k recombinants represent single recombination events in the interval M_1–M_k. This is plausible since double recombinants are not included among the observed M_1–m_k, m_1–M_k recombinants, and triple recombinants are exceedingly rare. Consequently, the marker genotype of each recombinant individual is determined completely by the single point of recombination within the target segment for that individual. The location of the point of recombination within the target segment can be progressively narrowed by noting further that, once a subinterval spanning several markers within the segment is shown to be nonrecombinant, it is no longer necessary to further genotype any of the markers in this subinterval. Clearly, by genotyping a single marker in the center of the recombinant subinterval, the size of the subinterval containing the point of recombination is progressively reduced by one-half. Thus, if a total of $\boldsymbol{M}$ markers are taken to span the target segment (including the two flanking markers), the number of markers genotyped per individual that are required to identify the point of recombination will be between n and $n + 1$, where n = integer part of $\log_2 \boldsymbol{M}$. A small number of worked examples show that the average n is closely approximated by $n = \log_2 \boldsymbol{M}$.

Application of this principle leads to a procedure that we term the "half-section (HS) algorithm," illustrated in Figure 1. For the HS algorithm, the genotype of each individual is determined independently of all others. Thus, the total number of genotyping points for the entire set of recombinants, T, is simply the average number of genotyping points per individual multiplied by the total number of recombinant individuals, R, *i.e.*, $T = Rn$.

Application of the HS algorithm involves sequential splitting of the recombinant progeny into progressively smaller subgroups. Each subgroup is genotyped for a different marker and split further. Thus, the early markers are used on subgroups with many members, the later markers on subgroups with only a few members. As the genotyping progressed, more and more markers were used in each round, but each marker was set up and used only once on a specific subgroup. For example, consider genotyping 400 recombinants for 31 markers. In complete genotyping, each individual is genotyped for all 31 markers: a total of 12,400 genotyping data points. When using the HS algorithm, all individuals are genotyped for marker 1; 200 individuals are genotyped for markers 2 and 3; 100 individuals each are genotyped for markers 4–7; 50 individuals are genotyped for markers 8–15; and 25 individuals are genotyped for markers 16–31—overall, a total of 1600 genotyping data points. Only four rounds are required for the entire HS genotyping procedure. With the negligible exception of three-point recombination within the target interval, the genotyping results given by the HS algorithm are exactly equivalent to those given by complete genotyping. Set-up costs for markers are the same as for complete genotyping; the only additional cost is for sorting the progeny for genotyping, according to the results of the previous round.

The golden section algorithm

The number of required genotyping data points can be reduced even more by noting that, within the target segment, the complete genotype of all individuals is required only across the subinterval that contains the QTL. If mapping analysis is carried out concurrently with genotyping, it is possible to progressively narrow the interval within the segment within which the QTL is found. It is then necessary to genotype only recombinants in this QTL-containing interval to further narrow the QTL location. Recombinants outside of this interval do not contribute information for QTL map location within the interval. Since we consider a situation with a single QTL in the target chromosomal region, it can be assumed that the expected LOD function (ELOD) will be a unimodal function (Hyne and Kearsey 1995; Ronin *et al.* 1999). This is so, even though other data sets of a comparable mapping population will have a LOD score function whose maximum is at a different location. Therefore, in applying this principle, we can use the golden section (GS) algorithm (Gill *et al.* 1981) to choose the markers for genotyping to progressively

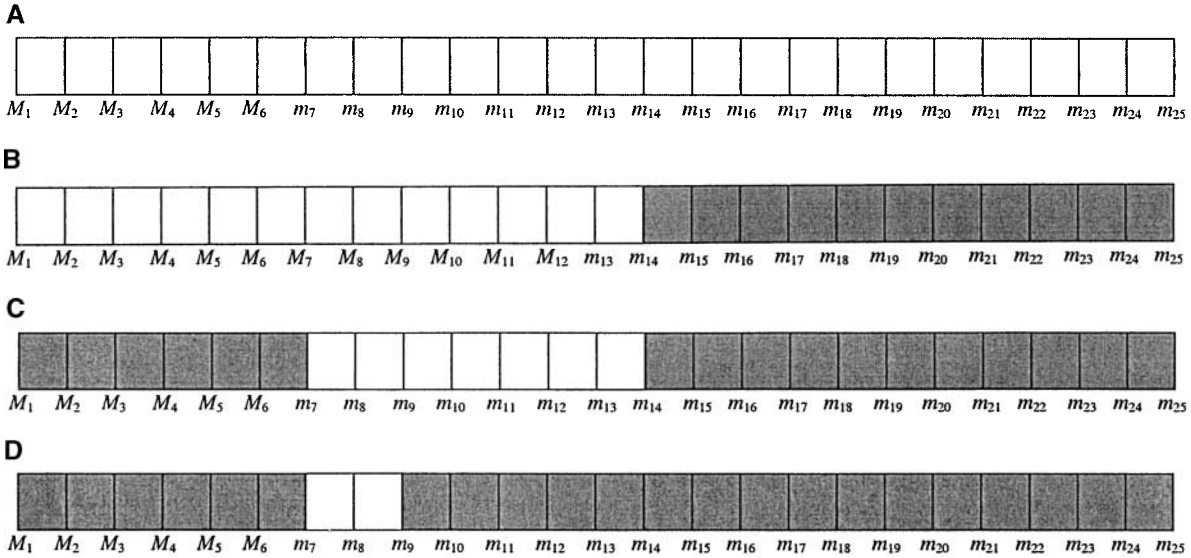
A
M1 M2 M3 M4 M5 M6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 m17 m18 m19 m20 m21 m22 m23 m24 m25
B
M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 m13 m14 m15 m16 m17 m18 m19 m20 m21 m22 m23 m24 m25
C
M1 M2 M3 M4 M5 M6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 m17 m18 m19 m20 m21 m22 m23 m24 m25
D
M1 M2 M3 M4 M5 M6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 m17 m18 m19 m20 m21 m22 m23 m24 m25

FIGURE 1.—Illustration of the "half-section" algorithm for determining marker genotype of a recombinant individual. (A) In the example, the point of recombination is between M_6 and M_7, so that in the initial scan for recombinants the individual was identified as an M_1–m_{25} recombinant. Notation: standard letters, *e.g.*, M_{16}, denote a marker; italicized letters, *e.g.*, M_{16} and m_{16}, denote alleles at the marker.

Step 1 (B): Genotype M_{13}, the central marker in the interval M_1–M_{25}. The allele found is m_{13}, showing that the point of recombination is between M_1 and M_{13}. From M_{13} to M_{25}, the genotype of the individual is now known to be $m_{13} \ldots m_{25}$.

Step 2 (C): Genotype M_6. The allele found is M_6, showing that the point of recombination is in the interval M_6–M_{13}. From M_1 to M_6, the genotype of the individual is now known to be $M_1 \ldots M_6$.

Step 3 (D): Genotype M_9: The allele found is m_9, showing that the point of recombination is in the interval M_6–M_9. From M_9 to M_{13}, the genotype of the individual is now known to be $m_9 \ldots m_{13}$.

Step 4: There are two options:

i. Genotype M_7. In this case, the allele found is m_7, showing that the point of recombination is between M_6 and M_7. The complete genotype of the individual is known.

ii. Genotype M_8. In this case, the allele found is m_8, showing that the point of recombination is between M_6 and M_8. It is still necessary to genotype M_7. The allele found is m_7, showing that the point of recombination is between M_6 and M_7, and the complete genotype of the individual is now known.

Total genotyping points for this individual will be 4 if option (i) is chosen and 5 if option (ii) is chosen.

narrow the subinterval within which the QTL is found. The GS algorithm is commonly used in numerical analysis for efficiently finding the maximum of a function with a single maximum (or minimum) measured without errors. As applied to QTL mapping, the GS algorithm basically involves identifying two flanking points between which the maximum of the mapping criterion (LOD function) is known to reside and evaluating the LOD function at these two points. The chosen points are, respectively, F and $1 - F$ of the distance between the two flanking points [where F is the golden section parameter equal to the Fibonacci constant, $F = 1/(1 + \sqrt{5}) \approx 0.38$]. The point for which the value of the LOD function is less defines a new flanking point. The process is now reiterated. It can be shown that the number of individuals genotyped at each successive step will be R, $(1 - F)R$, $(1 - F)^2R$, etc. Thus, application of this procedure will require only $[1 + (1 - F) + (1 - F)^2 + \ldots]R \approx 2.62R$ genotyping data points, irrespective of the number of markers in the target segment. Figure 2 illustrates the application of the GS algorithm.

In practice, due to finite population size, the LOD values will deviate slightly from the ELOD values. That is, there always will be some small fluctuations from monotonic behavior of the LOD function to both sides of the final estimate of QTL position on the chromosome implicit in the given data set (see HYNE and KEARSEY 1995). Consequently, there is a nonzero (albeit a very small) probability of placing the QTL in a wrong subinterval (and of following up the wrong recombinant individuals) using the GS criterion. Under such a situation, the final steps in the application of the GS method (which is an efficient tool for optimization of deterministic unimodal functions) become inefficient. Therefore, we propose employing the optimal properties of GS in producing the first $2.62R$ data points. Then, using an internal pair of already genotyped markers, M_i and M_j, which flank the last location of the maximum LOD, we continue with complete genotyping of all remaining markers (*i.e.*, residing between M_i and M_j) for individuals that are recombinants M_im_j and m_iM_j. This complete genotyping is conducted on the basis of the high-saving HS algorithm. Total genotyping data points for the internal segment under this combined GS-HS procedure

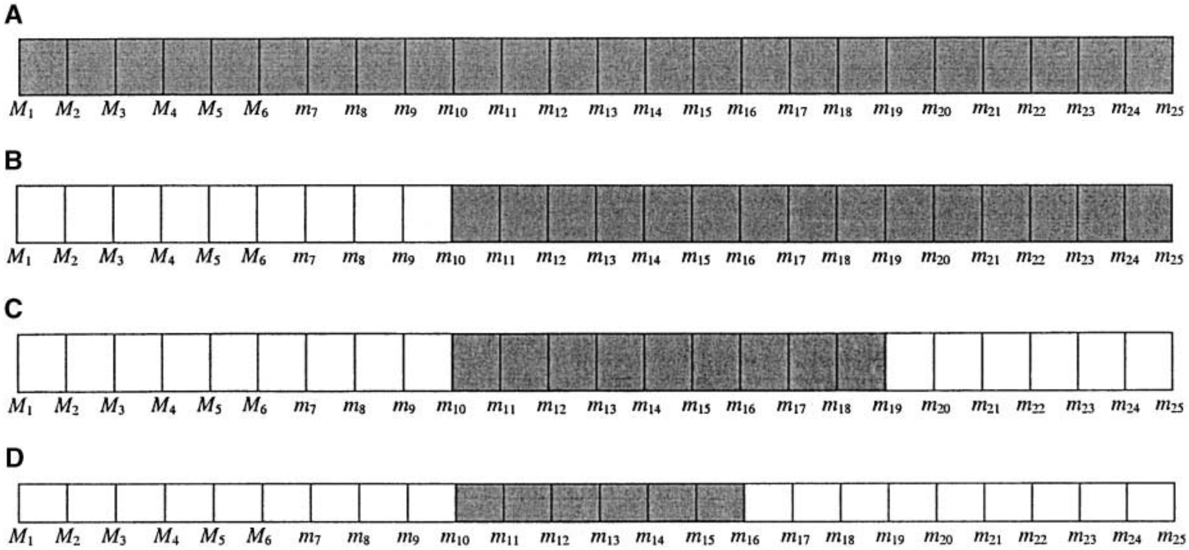
A
M_1 M_2 M_3 M_4 M_5 M_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_{20} m_{21} m_{22} m_{23} m_{24} m_{25}
B
M_1 M_2 M_3 M_4 M_5 M_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_{20} m_{21} m_{22} m_{23} m_{24} m_{25}
C
M_1 M_2 M_3 M_4 M_5 M_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_{20} m_{21} m_{22} m_{23} m_{24} m_{25}
D
M_1 M_2 M_3 M_4 M_5 M_6 m_7 m_8 m_9 m_{10} m_{11} m_{12} m_{13} m_{14} m_{15} m_{16} m_{17} m_{18} m_{19} m_{20} m_{21} m_{22} m_{23} m_{24} m_{25}

FIGURE 2.—Illustration of the "golden section" algorithm for determining marker genotype of a recombinant individual. (A) As in Figure 1, the point of recombination of the tracked individual is between M$_6$ and M$_7$, so that in the initial scan for recombinants the individual was identified as an M_1–m_{25} recombinant. The maximum LOD score (MLS) for the given data set is assumed to be located near M$_{14}$.

Step 1: Genotype M$_{16}$, located 0.62 of the distance from M$_1$ to M$_{25}$. Since all recombinant individuals are genotyped, the number of genotyping points for this step is *R*. We assume that points of recombination are distributed uniformly throughout the interval M$_1$–M$_{25}$. Consequently, of the total number of recombinant individuals, 0.62*R* will have allele m_{16}, and 0.38*R* will have allele M_{16}. For the individuals with allele m_{16}, the genotype at all markers in the interval M$_{16}$–M$_{25}$ is known: $m_{16} \ldots m_{25}$. For the individuals with allele M_{16}, the genotype at all markers in the interval M$_1$–M$_{16}$ is known: $M_1 \ldots M_{16}$.

Step 2 (B): Genotype M$_{10}$, located 0.38 of the distance from M$_1$ to M$_{25}$. Only the 0.62 of recombinant individuals that were m_{16} need to be genotyped. Individuals that were M_{16} will not be genotyped further. Thus, the total number of genotyping points for this step is 0.62*R*. Of the individuals genotyped for marker M$_{10}$, 0.62 will have allele M_{10}, and 0.38 will have allele m_{10}. Mapping analysis is now carried out with markers M$_1$, M$_{10}$, M$_{16}$, and M$_{25}$. According to the assumed location of the MLS, M$_{16}$ will have a higher LOD score than M$_{10}$. Thus, M$_{10}$ now becomes the new left flank marker. At this point markers M$_2$–M$_9$ are no longer of interest from the point of view of high-resolution mapping and will not be genotyped in any individuals.

Step 3 (C): Genotype M$_{19}$, located 0.62 of the distance from M$_{10}$ to M$_{25}$. For the 0.62 of individuals that were m_{16}, genotype at M$_{19}$ is known to be m_{19}. Thus, genotyping needs to be done only for the $0.38R = (0.62)^2R$ individuals that were M_{16}. Note that the individual with point of recombination between M$_6$ and M$_7$ is included in the group that need not be genotyped for M$_{19}$. Mapping analysis is now carried out with added marker M$_{19}$. M$_{16}$ will have a higher LOD score than M$_{19}$. Thus, M$_{19}$ now becomes the new right flank marker. Markers M$_{21}$–M$_{25}$ are no longer of interest.

Step 4 (D): Genotype M$_{13}$, located 0.62 of the distance from M$_{19}$ to M$_{10}$. It is necessary to genotype, only for M$_{13}$, those individuals that were m_{16} and $m_{13} = 0.24R = (0.62)^3R$ individuals. Mapping analysis is now carried out with added marker M$_{13}$. M$_{13}$ will have a higher LOD score than M$_{16}$. Thus, M$_{16}$ becomes the new right hand flanking marker, and the QTL has been located between M$_{10}$ and M$_{16}$. Markers M$_{17}$ and M$_{18}$ are no longer of interest.

Step 5: At this point it is convenient to determine genotypes for all markers in the interval M$_{10}$–M$_{16}$ (hence the entire algorithm can be referred to as GS-HS rather than as GS). For M$_{11}$ and M$_{12}$ genotypes need to be determined only for the 0.125 of individuals that were M_{10}–m_{13}–m_{16} (0.25*R* genotyping points), and for M$_{14}$ and M$_{15}$, for the 0.125 of individuals that were M_{10}–M_{13}–m_{16} (0.25*R* genotyping points). For all other individuals, genotypes for markers M$_{11}$, M$_{12}$, M$_{14}$, and M$_{15}$ can be inferred as above from the previously determined genotypes for M$_{10}$, M$_{13}$, and M$_{16}$. For example, the individual we are tracking would have been m_{10}–m_{16}, so that genotypes for markers M$_{11}$, M$_{12}$, M$_{14}$, and M$_{15}$ must be m_{11}, m_{12}, m_{14}, and m_{15}.

will be $3\boldsymbol{M}$ or less. Clearly, $3R < R \log_2 \boldsymbol{M}$, for $\boldsymbol{M} > 8$. In principle, therefore, the GS algorithm will generally require fewer data points than the HS algorithm. However, both represent major savings relative to complete genotyping. If, in the data set obtained in an actual experiment, the ELOD function was bimodal, the GS algorithm would not be applicable, and the HS algorithm would be used.

RESULTS AND DISCUSSION

The complete set of simulation results (data not shown) gave the SEQTL according to proportion selected in each tail (P), allele substitution effect at the QTL (d), size of mapping population (N), and marker spacing (c). A very wide spectrum of SEQTL values was obtained, ranging from 77.7 cM for the least powerful parameter combination ($P = 0.05$, $d = 0.25$, $N = 1000$, $c = 24$) to 0.05 cM for the most powerful combination ($P = 0.50$, $d = 1.00$, $N = 16{,}000$, $c = 0.29$). In an attempt to condense and simplify the total data set, nonlinear regression analysis was used to express the SEQTL as a power function of the simulation parameters. While the prediction equation obtained in this way explained much of the variation in SEQTL, many individual points were quite far from their predicted values. Consequently, the regression equation could not be used as a substitute for the tabulated values. However, the regression analysis did show a tight relationship between effects of N and d on SEQTL. This accorded with the well-known fact that test statistics for determining linkage between markers and QTL stand in proportion to d^2N (Song *et al.* 1999). Indeed, within a given combination of P and c, SEQTL were more or less the same for parameter combinations of d and N, for which d^2N was the same. For example, within the parameter combination $P = 0.05$, $c = 0.29$; SEQTL for $d = 0.25$, $N = 16{,}000$; $d = 0.5$, $N = 4000$ and $d = 1.0$, $N = 1000$ ($d^2N = 1000$ in each case) were 0.54, 0.51, and 0.59, respectively. Because of its powerful effect on SEQTL, the parameter d^2N is termed the "power factor" or *PF*. Examination of Table 5 of Thaller and Hoeschele (2000) shows the same dependence of accuracy of inferring QTL location on d^2N; compare, *e.g.*, in their Table 5, the "power" values for QTL effect 0.5, $N = 100$, 500, 2500 to those for QTL effect 0.25 and $N = 400$, 2000, and 10,000.

On the basis of the above relationship, a second table was prepared, giving SEQTL according to P, c, and d^2N (data not shown). Where there were two or more combinations of d and N with the same value of d^2N, these were averaged. The effect of proportion selected, P, was now examined. When this was done, with increase in P there was a consistent reduction in SEQTL at given *PF* and c, with the exception of the transition from $P = 0.25$ to $P = 0.50$, which was accompanied by only a very slight overall reduction in SEQTL (SEQTL at $P = 0.50$ was on average 0.96 of SEQTL at $P = 0.25$). This is expected, since virtually all of the information for QTL map location is found in the high and low 25% of the population (Darvasi 1997b). When the reduction in SEQTL in going from $P_j = 0.05$, 0.10, and 0.20 to $P = 0.25$ was calculated for given *PF* and c, there was much fluctuation within the individual cells of the table, but for given P_j, overall trends were not found, and the reduction in SEQTL appeared to be consistent across the entire table of values (data not shown). The average reduction in SEQTL relative to $P = 0.25$ in going from $P_j = 0.05$, 0.10, and 0.20 to $P = 0.25$ was 0.46, 0.69, and 0.94, respectively. SEQTL for $P = 0.05$, 0.10, 0.20, and 0.50 were therefore transformed to a $P = 0.25$ basis by multiplying by the appropriate average factor (0.46, 0.69, 0.94, and 1.04, respectively). The results were averaged and are given in Table 2. It is of interest that the factors for $P = 0.05$, 0.10, and 0.20 appear to stand in close proportion to $(P/0.5)^{0.5}$, indicating a massive reduction in information content of the marginal data points in each case.

Examining the effect of marker spacing in Table 2 shows that the phenomenon of maximum achievable resolution for given *PF* noted by Darvasi *et al.* (1993) is found only for the lowest power factor, $PF = 62.5$. At all other power factors, with each step decrease in c there was a consistent, albeit often small, reduction in SEQTL. The reduction in SEQTL with successive step decreases in c (*i.e.*, from $c = 24$ to $c = 8$, $c = 8$ to $c = 2.66$, $c = 2.66$ to $c = 0.88$, and $c = 0.88$ to $c = 0.29$) differed in a nonlinear manner depending on the power factor and on the specific step. In general, the reduction in SEQTL per step decrease in c was greater for the initial steps and smaller for the final steps and was greater for large *PF* and smaller for small *PF* (Table 1). It is noteworthy that an increase in marker spacing alone can increase map resolution by as much as eightfold, depending on the power factor. This finding is potentially of major importance. It tells us that when *PF* is high, saturation of the genomic interval carrying the detected QTL by additional markers is justified. Furthermore, in many cases, by the use of multiple-trait analysis (Korol *et al.* 2001) the scaled multiple-trait allele substitution effect of a QTL (D) is much greater than the single trait effect (d). Since the *PF* stands in proportion to D^2, this will markedly increase the *PF* at the same N. This increase in *PF*, in turn, will enable a further major decrease in SEQTL by adding even more markers to the genomic interval carrying the detected QTL. Thus by combining multiple-trait analysis with marker saturation, map resolution for given N can be increased manifold. The possibility of multiple-trait interval mapping analysis for selective genotyping design was already shown by Ronin *et al.* (1998).

Along similar lines, there was a consistent reduction in SEQTL with an increase in *PF* at all levels of c. However, the reduction did not stand in simple proportion either to the *PF* itself or to the square root of the *PF*. Thus, a further simple reduction of Table 1 with respect to c or *PF* was not possible. Table 2 can therefore be taken as the final condensed representation of the data.

The actual SEQTL for given d, N, P, and c can be approximated closely by going to the corresponding value of *PF* and c in Table 2 and multiplying by the inverse of the P_j to $P = 0.25$ reduction factor. For example, the SEQTL for $d = 0.5$, $N = 4000$ ($PF = 1000$), $c = 2.66$, $P = 0.2$ in the initial data simulation was 1.14. To reconstruct this value from Table 1, go to $PF = 1000$, $c = 2.66$ in Table 2 to find the value 0.785. This is multiplied by the factor 1/0.69 to give SEQTL = 1.14, which in this case happens to equal exactly the value found by simulation (data not shown). Not all equivalents were this exact, but most were very close.

Darvasi and Soller (1997) showed by simulation that the 95% confidence interval of QTL map location with a backcross or half-sib design, using a completely saturated map, can be closely approximated by the expression 95% C.I. = $3000/d^2N$. On this approximation, the expected SEQTL with a fully saturated map can be approximated as SEQTL = 95% C.I./4 = 750/*PF*. These values are also shown in Table 2 and should be compared to those obtained for $c = 0.29$, which are the limit values of the present simulation. The values obtained in the present study for $PF = 62.5$ and $PF = 125$ were much greater than the Darvasi and Soller (DS; Darvasi

TABLE 2

SEQTL according to marker spacing (*c*) and power factor (d^2N)

	d^2N								
c (cM)	62.5	125	250	500	1,000	2,000	4,000	8,000	16,000
24	42.19	18.46	6.05	2.78	1.70	1.24	0.87	0.70	0.51
8	33.52	14.47	4.04	1.76	1.06	0.69	0.47	0.37	0.26
2.66	34.15	13.89	3.13	1.39	0.79	0.46	0.28	0.21	0.15
0.88	31.59	12.28	2.85	1.17	0.64	0.36	0.21	0.13	0.08
0.29	33.08	11.73	2.69	1.14	0.60	0.31	0.17	0.10	0.06
R	0.78	0.64	0.44	0.41	0.35	0.25	0.20	0.14	0.12
DS	12.00	6.00	3.00	1.50	0.75	0.38	0.19	0.09	0.05

For proportion selected, $P = 0.25$, calculated as the average value obtained by converting values for $P = 0.05$, $P = 0.10$, $P = 0.20$, and $P = 0.50$ to a $P = 0.25$ basis, and the obtained value for $P = 0.25$. *R*, the proportion of reduction in SEQTL in going from $c = 24$ cM to $c = 0.29$ cM. DS, the approximate values obtained by the Darvasi and Soller (1997) simulation.

and Soller 1997) values. This is due to the fact that the DS simulation assumed that the QTL was within the simulated target region and hence gives smaller values than the present simulation gives when the SEQTL is large and when some estimated QTL positions are outside the target region. The values obtained in the present simulation for $PF = 250$–4000 were somewhat less than the DS values. The reason for this is not clear. Finally, the present study gave values equivalent to those of the DS approximation for $PF = 8000$ and 16,000. In general, therefore, the values given by the DS approximation are consistent with those of the present simulation.

Figures 1 and 2 illustrate the HS and GS genotyping procedures. An example of the relative efficiency of the HS, GS, and combined GS-HS algorithms alone on mapping resolution is given in Table 3, which explores these relationships by simulation for the cases $d = 1$; $N = 4000, 8000$; $P = 0.10, 0.20$; $c = 0.125$; and an initial interval of 24 cM, so that total number of markers = (24/0.125) + 1 = 193. Total genotyping data points required by HS, GS, and GS-HS algorithms are 7.58*R*, 2.62*R*, and 3*R*, respectively. At this very dense spacing, SEQTL obtained by use of the GS algorithm alone are two- to threefold greater than SEQTL obtained by use of the HS algorithm. SEQTL obtained by the use of the combined GS-HS algorithm, however, are essentially equal to those obtained by the HS algorithm. Since genotyping results obtained by the HS algorithm are exactly the same as those provided by complete genotyping, the latter procedure was not simulated separately.

TABLE 3

SEQTL based on the HS, GS, and combined GS-HS algorithms applied to selective genotyping design

N	*P*	HS[a]	GS	GS-HS
4000	0.1	0.34	0.50	0.36
	0.2	0.18	0.29	0.17
8000	0.1	0.13	0.29	0.12
	0.2	0.10	0.28	0.11

The results correspond to $d = 1$, averaged over 1000 Monte Carlo runs; marker spacing $c = 0.125$ cM, $L = 24$ cM. *N*, total size of the mapping population; *P*, proportion of individuals selected for genotyping.

[a] Results obtained by full genotyping of all recombinant individuals were exactly the same as those obtained by HS genotyping.

Clearly, the need for a small additional genotyping "investment" caused by moving from the GS to complete genotyping (2.62*R* → 3*R*) is due to fluctuations caused by finite sample size. The estimates in Table 3 demonstrate that this small investment provides the same resolution as given by HS at a higher cost (note the close results for HS and GS-HS obtained by 7.58*R* and 3*R* genotyping data points, respectively).

PRACTICAL FEASIBILITY AND IMPLEMENTATION

The results of this study show that when large mapping populations are available, SEQTL can be reduced to subcentimorgan levels, even for QTL of moderate effect ($d = 0.25$). This gives 95% confidence intervals of QTL location in the range of 1–5 cM. Confidence intervals of this magnitude provide tightly linked markers for marker-assisted selection, a strong basis for a search for population-wide linkage disequilibrium in outcrossing populations, and a platform for a search for the actual gene corresponding to the QTL.

By careful consideration of Table 2, the trade-off between population size, proportion selected, and marker spacing can be calculated, so as to obtain maximum return for the research investment. If large families are available and samples can easily be accessed, it will be more cost effective to use a small *P* with largest possible family size and wider marker spacing. If families are relatively small, or if it is difficult to access samples, it will be more cost effective to use a large *P* and closer marker spacing.

The major requirement for application of these procedures is availability of a population of required size and sufficient density of informative markers. The common dinucleotide microsatellite markers are generally not available at a spacing of <1–2 cM. However, with the introduction of single nucleotide polymorphism markers an increase by one or two orders of magnitude in the number of markers and a decrease of an order of magnitude in costs of genotyping are confidently expected for the near future.

With respect to population size, F_2 and BC populations of 10,000 or more can readily be produced in many species of agricultural plants. Thus, these species are excellent candidates for SRG. In agricultural animal species, the enormous sire half-sib families, consisting of 10,000 or more daughters that are routinely produced through artificial insemination in dairy and in some beef cattle populations, have the requisite family structure for QTL mapping, and phenotypic information is available on each individual. For application of SRG to poultry and swine breeding nuclei, progeny can be collected across a number of sires heterozygous for the same QTL to provide the desired total number of progeny for high-resolution mapping. This would require a preliminary step in which many sires are analyzed to identify sires heterozygous at the QTL. To reduce genotyping costs, screening of sires for heterozygosity could be achieved by selective DNA pooling (Darvasi and Soller 1994; Lipkin *et al.* 1998).

Given the required population size, the genotyping load is not overly great when the GS-HS or HS algorithms are used. For example, for a QTL mapped to a target interval of 20 cM, and with a mapping population of $N = 10{,}000$ for BC or half-sib designs or of 5000 for F_2 designs, application of SRG at $P = 0.20$ and $c = 1$ cM would require ~7500, 11,500, or 23,000 genotyping data points for F_2, BC, or half-sib designs, respectively, plus a small number for haplotyping the parents. This comes out to only a little more than one or two data points per daughter!

Although a given SRG mapping population will allow high-resolution mapping of all QTL segregating in the population, each QTL will have to be analyzed separately. Thus, high-resolution mapping of 10 QTL in the above mapping population would require a total of 100,000–200,000 data points. However, this is still only 10–20 data points per individual, 250-fold less than would be required for high-resolution mapping of the entire genome at a marker spacing of $c = 1.0$ cM.

An important aspect of the considered procedure is the assumption that the target QTL was correctly assigned to the segment bounded by the flanking markers M_1–M_k. Depending on the choice of C.I. stringency, the possibility will always exist that the true QTL position is not within the target segment, but in the adjacent segment, to the right or left. Thus, if SRG analysis indicates that the QTL is located to the extreme end of the target segment, one would go on to identify recombinants in the adjacent segment (at a cost of $2NP$ data points) and conduct an SRG analysis across both segments. Setting the initial target interval with much wider limits than the 95% C.I. would not be as useful, because in most instances the QTL will map within its 95% C.I. so that the additional effort is not needed, and, with a very wide target interval, double and triple recombinants will play more of a spoiling role.

The present results relate to expected SEQTL under various assumed design and parameter combinations. The question arises as to the relevance of the SEQTL of the present study, obtained across many simulations, to the C.I. of map location as it might be estimated from the one-time data of an actual experiment. In this case, bootstrap and information matrix methods are available to obtain approximate confidence intervals for QTL map location. However, it would also be possible to use the estimate of QTL effect obtained from the actual experiment to obtain a SEQTL estimate by interpolation in Table 2. We believe that C.I. obtained by the two approaches will be similar, but this remains to be explored in detail.

In addition, once an estimate of QTL effect has been obtained, the results of this study are relevant to deciding whether and to what degree further marker density in the C.I. could reduce the SEQTL and C.I. of map location.

Constructive comments of two referees are acknowledged with thanks. This study was supported by the Israeli Ministry of Absorption and German-Israel Cooperation project (DIP project founded by the Inernationales Deutsch-Israelische des BMBF Projektkooperation), by the Framework 5 Program of the E.U., and by the U.S.-Israel Binational Science Foundation.

LITERATURE CITED

Darvasi, A., 1997a Interval specific congenic strains (ISCS): an experimental design for mapping a QTL into a 1-centimorgan interval. Mamm. Genome **8:** 163–167.

Darvasi, A., 1997b The effect of selective genotyping on QTL mapping accuracy. Mamm. Genome **8:** 67–68.

Darvasi, A., 1998 Experimental strategies for the genetic dissection of complex traits in animal models. Nat. Genet. **18** (1): 19–24.

Darvasi, A., and M. Soller, 1994 Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Genetics **138:** 1365–1373.

Darvasi, A., and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141:** 1199–1207.

Darvasi, A., and M. Soller, 1997 A simple method to calculate resolving power and confidence interval of QTL map location. Behav. Genet. **27** (2): 125–132.

Darvasi, A., A. Vinreb, V. Minke, J. I. Weller and M. Soller, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

Gill, P. E., W. Murray and M. H. Wright, 1981 *Practical Optimization.* Academic Press, New York.

Hill, W. G., 1998 Selection with recurrent backcrossing to develop congenic lines for quantitative trait loci analysis. Genetics **148:** 1341–1352.

Hyne, V., and M. J. Kearsey, 1995 QTL analysis: further uses of 'marker regression'. Theor. Appl. Genet. **91:** 471–476.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136:** 1447–1455.

Jiang, C., and Z-B. Zeng, 1995 Multiple trait analysis and genetic mapping for quantitative trait loci. Genetics **140:** 1111–1127.

Klein-Lankhorst, R. M., A. Vermun, R. Weide, T. Liharska and P. Zabel, 1991 Isolation of molecular markers for tomato (*L. esculantum*) using random amplified polymorphic DNA (RAPD). Theor. Appl. Genet. **83:** 108–114.

Korol, A. B., Y. I. Ronin and V. M. Kirzhner, 1995 Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics **140:** 1137–1147.

Korol, A., Y. Ronin, A. Itzcovich and E. Nevo, 2001 Enhanced efficiency of QTL mapping analysis based on multivariate complexes of quantitative traits. Genetics **157:** 1789–1803.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lipkin, E., M. O. Mosig, A. Darvasi, E. Ezra, A. Shalom *et al.*, 1998 Mapping loci controlling milk protein percentage in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers. Genetics **149:** 1557–1567.

Rhodes, M., R. Straw, S. Fernando, A. Evans, T. Lacey *et al.*, 1998 A high-resolution microsatellite map of the mouse genome. Genome Res. **8:** 531–542.

Ronin, Y. I., A. B. Korol and J. I. Weller, 1998 Selective genotyping to detect quantitative trait loci affecting multiple traits: interval mapping analysis. Theor. Appl. Genet. **97:** 1169–1178.

Ronin, Y., A. Korol and E. Nevo, 1999 Single- and multiple-trait analysis of linked QTLs: some asymptotic analytical approximation. Genetics **151:** 387–396.

Soller, M., and J. S. Beckmann, 1990 Marker-based mapping of quantitative trait loci using replicated progeny. Theor. Appl. Genet. **80:** 205–208.

Soller, M., and L. Andersson, 1998 Genomic approaches to improvement of disease resistance in farm animals. Rev. Sci. Tech. **17:** 329–345.

Song, J. Z., M. Soller and A. Genizi, 1999 The full-sib intercross line (FSIL) design: a QTL mapping design for outcrossing species. Genet. Res. **73:** 61–73.

Thaller, G., and I. Hoeschele, 2000 Fine-mapping of quantitative trait loci in half-sib families using current recombinations. Genet. Res. **76** (1): 87–104.

Weller, J. I., Y. Kashi and M. Soller, 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.

Zeng, Z-B., 1994 Precise mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: J. B. Walsh