POOLED-SAMPLING MAPPING MAKES HIGH RESOLUTION MAPPING PRACTICAL WITH DNA MARKERS

by

Gary A. Churchill,
James J. Giovannoni
and
Steven D. Tanksley

BU-1150-M February 1992

POOLED-SAMPLING MAPPING MAKES HIGH RESOLUTION MAPPING PRACTICAL WITH DNA MARKERS

Gary A. Churchill, James J. Giovannoni and Steven D. Tanksley

Department of Plant Breeding and Biometry
Cornell University
Ithaca, NY 14853

Key terms: RFLP, Map-based cloning, tomato, gene mapping

Abstract

A pooled sample approach to the construction of high resolution genetic maps is described. The strategy depends on the existence of an easily selectable target locus and the ability to produce large segregating populations. If these requirements are met, the pooled sample mapping approach allows tightly linked markers (e.g. RFLPs) to be mapped relative to the target with a great economy of effort. The recombination fractions among loci can be estimated by the maximum likelihood method and a simple approximate estimator is derived. The order of loci is deduced using a Bayesian statistical framework to yield posterior probabilities for all possible orderings of a marker set. Optimal pooling strategies and the effects of misclassification of selected individuals are discussed and studied by computer simulation. The feasibility of this method is demonstrated by the high resolution mapping of a region on chromosome 5 of tomato that contains a gene regulating fruit ripening.

Introduction

The idea of using restriction fragment length polymorphisms (RFLPS) for genetic mapping was introduced in 1980 (1). Since that time there has been rapid progress in the development of genetic maps in a variety of organisms, including human, mouse, and many crop species (2-5). One of the reasons for constructing DNA-based genetic maps is for use in chromosome walking. Currently many interesting and important genes are known only by their phenotype. Lack of knowledge of their gene products inhibits traditional methods of gene cloning; however, knowing the position of such genes on a DNA-based map opens the opportunity for walking to the gene from adjacent marker(s). In this manner, a number of genes have been cloned from humans and other higher eucaryotes in recent years (6).

Two requirements for chromosome walking are: 1) availability of tightly linked DNA marker(s) in the vicinity of the gene of interest and 2) knowledge of the position of these markers relative to the targeted gene. In the past, finding markers near a gene of interest proved to be very time consuming. However, increased efforts in genome mapping have led to the generation of RFLP-based maps for many organisms (2-5). In addition, there are now methods in place for rapidly identifying new DNA markers specific to any region of a genome (7-11).

Once a number of markers tightly linked to a gene of interest have been identified, the two markers that most closely flank the targeted gene must be identified since it is these two markers that provide the most efficient starting points for a walk. In cases where the markers are very tightly linked to the targeted gene (e.g., < 1 cM) analysis of hundreds or even thousands of segregating progeny may be required to determine the order of markers in the vicinity of the target gene (12). This can be costly and time consuming since it requires isolation and analysis of DNA from each individual in a segregating population.

In an effort to overcome the problem encountered with mapping large populations, we have devised and tested a pooled-sample method for high resolution mapping around genes targeted for cloning. The method exploits the fact that when mapping many markers in a small segment of a chromosome, very few individuals (from a segregating population) contain chromosomes with a crossover in the region of interest and thus most individuals provide little useful information. By pooling individuals for analysis, the effort required to construct a high resolution map can be reduced many fold. The steps for using this pooled-mapping technique in a segregating population (e.g., F2) are as follows: 1) Identify (by phenotype) those individuals that are homozygous (usually homozygous recessive) for the target gene; 2) divide these individuals into pools (see Results section for optimum pool sizes). Extract DNA *en masse* from each pool using approximately equal amounts of tissue from each individual. This bulked DNA is then probed with clones known to be located in the vicinity of the gene. The proportion of pools containing at least one crossover event is recorded and the resulting data used to construct a high resolution map. Using that method, the number of samples from which DNA must be isolated and analyzed can be reduced by a factor of 10 or even greater.

We present here basic theoretical considerations necessary for utilizing pooled-mapping and demonstrate the technique by mapping a gene regulating fruit ripening in tomato. This method leads to the identification of the two markers most closely flanking the targeted gene. It can also be used to determine the map order of other markers and an estimate of map distance from the targeted gene. While the method has been demonstrated in a plant, it can be used in any diploid, sexually reproducing organism for which large segregating populations can be obtained.

Material and Methods

Calculations and computer simulations. The algorithms described below have been implemented in the C programming language and tested on a SUN workstation.

Simulations and Monte Carlo integrations were implemented using the 48-bit congruential pseudo-random number generator drand48() supplied with the UNIX operating system.

Pooling. Individuals from an F2 population segregating for a target locus (T) are determined to be homozygous for the target locus (either T/T or t/t) and are divided into n pools of k individuals each. Since each individual contains two chromosomes, the number of independent meiotic events represented by each pool equals $2k$. If a backcross population is used, the number of meiotic events represented by each pool equals k, and the results below should be modified accordingly. DNA is isolated *en masse* from each pool and analyzed with markers thought to be linked to T but not necessarily ordered with respect to T or each other. It is assumed that a single recombinant chromosome can be detected in a pool of otherwise non-recombinant chromosomes, but that the exact number of recombinant chromosomes cannot be determined. Therefore, the results from probing a pooled DNA sample with a linked marker is classification of that pool as recombinant (i.e., contains at least one chromosome that is recombinant between the marker and the target gene) or nonrecombinant. The combined result of probing a pool with a set of m markers, $\mathbf{M} = \{M_1, M_2, \ldots, M_m\}$, is a classification of the pool into one of 2^m possible *pool types.* A pool type is defined by the set S of markers in **M** that are recombinant with respect to T for at least one chromosome in the pool. The observed number of pools of type S will be denoted by Y_s and the 2^m vector of observed counts will be denoted by **Y**.

Ordering of Markers. The goal of a pooled sample mapping experiment is to infer the map order among a set of markers relative to the target locus. In particular, we wish to identify the two markers which most closely flank the target locus. One widely used approach to the problem of inferring map order among a set of markers is to choose that order which has the highest maximized likelihood (15, 16). However, the maximum likelihood approach has several limitations. First, when markers are known to be tightly linked to the target locus, this prior knowledge should be incorporated into the linkage analysis. This will generally be the case when a pooled sample approach to mapping is being considered, although unlinked or loosely linked markers will be detected by this method. Second, the different order hypotheses are not nested and thus one cannot construct a likelihood ratio to formally test the best order. For these reasons we have employed a Bayesian approach to the marker ordering problem.

Let H denote the map order of markers in the set $\mathbf{M} \cup \mathrm{T}$ and let $\mathbf{R} = (r_1, r_2, \ldots, r_m)$ denote the recombination probabilities between adjacent markers in the ordered set. Inference of the map order will be based on the posterior probability of H which can be computed as

$$\begin{aligned}\Pr(\mathrm{H}|\mathrm{Y}) &\propto \Pr(\mathrm{H})\Pr(\mathrm{Y}|\mathrm{H}) \\ &= \Pr(\mathrm{H}) \int \Pr(\mathrm{Y}|\mathbf{R}, \mathrm{H}) \Pr(\mathbf{R}|\mathrm{H})\, d\mathbf{r}, \qquad (1)\end{aligned}$$

where Pr(H) is the prior distribution on map orders, Pr(**R**|H) is the prior distribution on recombination probabilities and Pr(Y|**R**, H) is the likelihood of the observed counts. The integral can be evaluated to desired numerical precision by the Monte Carlo method of composition (18).

The likelihood of the observed counts is multinomial on 2^m classes which are the pool types,

$$\Pr(\mathrm{Y} = \mathbf{y}|\mathbf{R}, \mathrm{H}) \propto \prod_{\mathrm{S} \subseteq \mathbf{M}} p_{\mathrm{S}}^{y_{\mathrm{S}}} \qquad (2)$$

where the pool type probabilities p_{S} are to be defined in RESULTS. The multinomial distribution follows from assumptions that individual pools contain identical numbers of chromosomes generated by independent meioses and that each pool is probed independently of others.

Prior Distributions. The prior knowledge available about map order and recombination probabilities will depend on the source of the markers. For example, we may consider markers selected at random from a fixed interval of known size (e.g., markers identified using nearly isogenic lines, ref. 9), markers known to segregate closely with the target locus in previous crosses, or markers selected in the vicinity of the target locus from a previously constructed high density genetic map. For purposes of this analysis, the markers are assumed to be uniformly placed in a region around T with known density *D* markers /100cM.

In the cases of unmapped or cosegregating markers, it is natural to assume that all possible orders are equally likely. For a set of m markers plus the target T, there are $(m+1)!/2$ distinct orders and the equally likely prior distribution is

$$\Pr(\mathrm{H=h}) = 2/(m+1)!, \tag{3}$$

for all orders h. In situations where order information is available, the posterior probabilities from previous experiments can be used as the prior distribution for the new experiment.

The prior distribution on recombination probabilities is taken to be a product of m – 1 independent Beta distributions (13),

$$\Pr(\mathrm{R=r|H}) \propto \prod_{i=1}^{m} r_i^{a-1} (1-r_i)^{b-1} . \tag{4}$$

The Beta distribution with parameters a and b has mass on the interval (0, 1) with mean a/(a+b). It is an analytically convenient choice and when the genetic distances involved are small, approximates the distribution of spacing between randomly placed markers. For unmapped markers selected from a region with local density *D* markers/100cM, the parameters a=1 and b=D – 1 are appropriate. Because recombination probabilities can vary significantly in different crosses it is not clear what form of prior information should be used when previous experimental data are available.

Decision Rules. The posterior probabilicies (eq. 1) are used to make decisions about the map order. The decision rule $\delta(\mathrm{Y})$ which chooses the order with highest posterior probability as an estimate of the true map order is optimal in the sense of being the Bayes rule for a $0-1$ loss function (ref. 13, p. 163). However, if the markers are to be used to attempt a chromosome walk, an incorrect decision could be very costly. We wish to ensure that the two nearest markers which span the target are correctly ordered with high probability. A more stringent criterion is to decide that an order is correct only if its posterior probability exceeds a specified critical value (e.g., 0.95 or 0.99). Otherwise, no decision is made and probing of additional pools will be required to resolve the map order.

Our criterion for comparing different pooling strategies and decision rules will be one minus the Bayes risk (ref. 13, p. 11) for assigning an order to triplets of adjacent loci (i.e., two markers plus the target). For a decision rule δ and a $0-1$ loss function the Bayes risk is just the probability of an incorrect (or no) decision averaged over all realizations of Y and over all possible orders H,

$$\rho(\delta) = \sum_{\mathrm{h}}\left[\Pr(\mathrm{H}=\mathrm{h})\left[\sum_{\mathbf{y}} 1(\delta(\mathbf{y}) \neq \mathrm{h}) \Pr(\mathbf{Y}=\mathbf{y}|\mathrm{H}=\mathrm{h})\right]\right] \tag{5}$$

where the summations run over all possible values of h and y.

Exact calculation of the Bayes risk involves extensive summations and integration over a high dimensional space. However, its value can be approximated by the following Monte Carlo algorithm:

1. Generate $\mathrm{h}^* \sim \Pr(\mathrm{H})$.
2. Generate $\mathbf{r}^* \sim \Pr(\mathbf{R}|\mathrm{H}=\mathrm{h}^*)$.
3. Generate $\mathbf{y}^* \sim \Pr(\mathbf{Y}|\mathbf{R}=\mathbf{r}^*, \mathrm{H}=\mathrm{h}^*)$.
4. If $\delta(\mathbf{y}^*)=\mathrm{h}^*$ increment # correct.
5. Repeat steps 1–4 N times.
6. The estimated Bayes risk is $\rho(\delta) = 1-\#\ \mathrm{correct}/\mathrm{N}$.

Results.

Estimating the genetic map distance between a marker and the target gene. Let T denote the large locus, A denote a marker locus near T and r the probability of recombination between A and T within a single chromosome. The probability that the i^{th} pool contains at least one recombinant is

$$p_A = 1 - (1\text{-}r)^{2k}. \tag{6}$$

When r is small, this probability is approximated by $1 - e^{-2kr}$. An approximate maximum likelihood estimator (MLE) for r is

$$\hat{r} = -\frac{1}{2k}\,\ell n(1 - \frac{y_A}{n}), \tag{7}$$

where y_A is the total number of recombinant pools. The large-sample variance of this estimator (the reciprocal of Fisher's information),

$$\text{Var }(\hat{r}) = \frac{1}{4k^2 n}\left(\frac{1 - e^{-2kr}}{e^{-2kr}}\right), \tag{8}$$

can be used to place approximate confidence limits on r.

An exact MLE for the pooled sample mapping problem can be obtained by an EM algorithm (Expectation-Maximization, see ref. 19). This iterative optimization procedure has a long history in genetic linkage studies (see for example 14, 15, 16). In the E-step, a current estimate of r is used to compute the expected composition of each pool, given the pool type. In the M-step, the expected compositions are treated as data, recombinants are counted and a new estimate of r is obtained. These steps are iterated until the estimate converges. This EM algorithm can be readily generalized to multipoint mapping for a given map order. Numerical comparisons of the exact and approximate MLEs show very close agreement when the true value of the recombination probability is small (less than 0.1). Because the EM algorithm becomes cumbersome to compute for large pool size, we prefer to use the approximation.

Ordering markers relative to the target gene. Two-point data can be used to estimate map distances from the target to each member of a set of markers but will not be sufficient to order the markers relative to the target. For this purpose three-point data (i.e., a target T and two segregating markers A and B) are required. It is well known that there is a significant information gain in standard (unpooled) genetic mapping when one considers triplets of markers rather than pairwise data (17). Thus we expect that three-point data will also be useful in resolving the order of markers which may be unresolved by two-point data.

For the three-point analysis we will focus on the ordering problem. There are three distinct orders to consider:

$$\begin{array}{l} H_1\text{: A--T--B} \\ H_2\text{: A--B--T} \\ H_3\text{: T--A--B} \end{array}$$

each with its own set of recombination parameters $\mathbf{R} = (r_1,\ r_2)$. The probability of a recombination between the "left" pair of markers is r_1 and the recombination probability for the "right" pair is r_2. Assuming a no-interference model, recombination between the outer pair of markers will occur with probability $(1-r_1)r_2 + r_1(1-r_2)$.

The individual chromosomes in a pool can be classified into four types (ϕ, A, B, and AB) indicating which markers are recombinant with respect to T. The chromosome type probabilities, denoted by q, depend on the order of markers as shown:

$$\begin{array}{lcccc} & & \mathrm{H}_1 & \mathrm{H}_2 & \mathrm{H}_3 \\ \hline q_\phi & = & (1\text{-}r_1)(1\text{-}r_2) & (1\text{-}r_1)(1\text{-}r_2) & (1\text{-}r_1)(1\text{-}r_2) \\ q_{\mathrm{A}} & = & r_1(1\text{-}r_2) & r_1(1\text{-}r_2) & r_1 r_2 \\ q_{\mathrm{B}} & = & (1\text{-}r_1)r_2 & r_1 r_2 & (1\text{-}r_1)r_2 \\ q_{\mathrm{AB}} & = & r_1 r_2 & (1\text{-}r_1)r_2 & r_1(1\text{-}r_2). \end{array} \tag{9}$$

The pool type is determined by the collection of chromosome types contained in the pool. For example, a pool of type A will contain at least one chromosome of type A and none of types B or AB. A pool of type AB may contain one or more AB chromosomes or at least one each of the types A and B. The outcome probabilities are

$$\begin{aligned} p_\phi &= q_\phi^{2k} \\ p_\mathrm{A} &= (q_\phi + q_\mathrm{A})^{2k} - q_\phi^{2k} \\ p_\mathrm{B} &= (q_\phi + q_\mathrm{B})^{2k} - q_\phi^{2k} \\ p_\mathrm{AB} &= 1 - (q_\phi + q_\mathrm{A})^{2k} - (q_\phi + q_\mathrm{B})^{2k} + q_\phi^{2k} \,. \end{aligned} \tag{10}$$

The probability p_ϕ that a pool contains no recombinant chromosomes decreases to zero and the probability p_AB increases to one as the pool size is increased. The probabilities p_A and p_B initially increase and then fall off toward zero (Figure 1). The pool type probabilities are substituted into the likelihood (equation 2) and together with the prior distributions (equations 3 and 4) define the posterior distribution over map orders (equation 1). Intuitively, the best results should be obtained when a moderate number of pools are expected to contain at least one recombinant chromosome but few pools are expected to contain more than one.

Multipoint Mapping. The results for three point mapping with pooled data are readily generalized to the problem of ordering a set of m markers relative to a target. Let $\mathbf{M} = \{\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_m\}$ denote the set of markers. An ordering on $\mathbf{M} \cup \mathrm{T}$ divides the set $\mathbf{M}$ into a "left set" $\mathbf{L} = \{\mathrm{L}_1, \mathrm{L}_2, \ldots, \mathrm{L}_{m_l}\}$ and a "right set" $\mathbf{R} = \{\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_{m_r}\}$ whose elements are ordered as they radiate outward from the target.

As before, our approach is to determine the probabilities of each chromosome-type and then express the pool-type probabilities in terms of these. Chromosome-type probabilities will be denoted by q_S where $\mathrm{S} \subseteq \mathbf{M}$ is the set of markers that are recombinant with respect to T. For a given order H and recombination probabilities $\mathbf{R}$, the calculation proceeds as follows. Let $\mathrm{L}_0 = \mathrm{R}_0 = \mathrm{T}$ and note that $\mathrm{T} \notin \mathrm{S}$. Consider first the left set $\mathbf{L}$. A recombination has occurred in the i$^{\text{th}}$ interval of the left set if exactly one of L_i and $\mathrm{L}_{\mathrm{i}-1}$ is an element of S. The corresponding recombination probability r_j is included as a factor in q_S. Otherwise, no recombination has occurred and a facto. $1-r_j$ is included. Factors for all intervals in the right set are similarly included. The pool–type probabilities are computed from the chromosome–type probabilities by the method of inclusion-exclusion

$$p_S = \left(\sum_{V \subseteq S} q_V\right)^{2k} + \sum_{V \subset S} (-1)^{\| S \backslash V \|} p_V \qquad (11)$$

where $\| \cdot \|$ denotes cardinality of a set. The recursion terminates at $p_\phi = q_\phi^k$.

Posterior probabilities can now be computed for each of the $(m+1)!/2$ possible orders (equation 1). These computations will become impractical for large collections of markers. However, in most cases the order will be resolved by the three-point analysis and the multipoint analysis will not be needed.

Optimum pool size. We consider two criteria for determining an optimum pool size. First, maximization of the information for mapping a single marker relative to the target yields a simple expression for optimum pool size. Second, we maximize the probability of inferring the correct order among triplets of adjacent loci (including the target). The latter criterion is motivated by the chromosome walking application and is studied by computing the Bayes risk (Equation 5) over a range of pool sizes. These two approaches to determining optimum pool size give similar results, with the latter criterion tending to give slightly smaller (by one or two F2 individuals) optimum pool sizes. Additional considerations for determining pool size are discussed below.

The optimum number of individuals to include in a pool is determined primarily by the local density of markers around the target locus. Consider the nearest adjacent markers to T. The probability that a chromosome is recombinant for either of these markers will decrease as the local density of markers is increased. If the pool size is small, most pools will contain no recombinant chromosomes and the amount of mapping information per pool will be very low. If the pool size is very large, most pools are likely to contain multiple recombinant chromosomes and mapping also becomes very inefficient. An optimum pool size is found between these two extremes and is larger for higher marker densities.

Exact inference for pooled binary experiments is discussed by Thompson (1962) who also considered the problem of choosing an optimal pool size. He showed that, for a fixed number of pools, a pool size of

$$k_{\text{opt}} \doteq \frac{1.594}{2r} \tag{12}$$

minimizes the variance of the estimator $\hat{r}$ (egn. 8). He also points out that k_{opt} tends to overestimate the true optimum when n is small. Note that when the pool size is near k_{opt}, the expected number of recombinations per pool in the interval is near 1.

It is clear that during the planning stages of an experiment r will be unknown. The purpose of pooling samples is to increase the efficiency of an experiment and even small pool sizes (2 to 5 plants) can greatly reduce the total effort required. However, in some circumstances, if the pool size is too large a very inefficient experiment can result. To protect against the possibility of overpooling, a value r_0 should be chosen which is an upper bound on the possible values of r. For example, if the target gene and the markers to be mapped are known to lie within a 10 cM interval, a pool size of $k \leq 8$ would protect against overpooling.

Maximum information yield per pool is a reasonable goal if the cost per plant is negligible compared to the cost per pool. If however, the cost per plant is substantial, one may wish to find the pool size which yields the maximum information per unit cost. The information yield per unit cost is proportional to

$$\frac{4k^2}{2ck+1} \; \frac{\mathrm{e}^{-2kr}}{1-\mathrm{e}^{-2kr}}, \tag{13}$$

where c is the ratio of the cost per plant to the cost per pool. Optimization of this quantity with respect to k for $r = 0.10$ and $c = 1.0$, 0.1 and 0.01 yields optimal pool sizes of 2.0, 4.6 and 7.2 respectively. These approach k_{opt} in the limit as c goes to zero.

To study the effect of pool size on the inference of map order we considered sets of three adjacent loci (one of which is the target locus) selected from a region with known marker density D markers/100cM. Simulations were carried out to estimate the probability of correct ordering (see METHODS for the algorithm) as a function of pool size for different marker densities, numbers of pools, critical values for the decision rule, and misclassification rates (see below). Selected results are summarized in Figure 2. An optimal pool size gives the highest probability of correct ordering.

The local density of markers has the greatest effect on optimal pool size. At low marker densities smaller optimal pool sizes are obtained and overpooling can result in a very inefficient experiment. Higher marker densities give larger optimal pool sizes and are more robust to overpooling. For the lowest marker density studied (D=16 markers/100cM in Figure 2A), the maximum probability of correct ordering is attained with a pool size of 3 individuals. This density corresponds to 3 loci in an interval of 18.75cM (r=3/D) and the optimal pool size (equation 12) is $k_{opt} = 4.25$. At this low density of markers, an experiment with more than about 8 pooled individuals becomes less efficient than probing single individuals. For higher densities (D=32 and 64 markers/100cM) the maximum ordering probabilities are attained with larger pool sizes and the range of efficient experiments is much broader. The maxima correspond well with the optimal sizes $k_{opt} = 8.5$ and $k_{opt} = 17$ respectively. For the highest marker density tested (D=128 markers/100cM) the maximum is not sharply defined by the simulation. The curve reaches its peak at some point beyond a pool size of 15 and does not drop noticeably out to a pool size of 40. The k_{opt} value corresponding to D=128 is 34 F2 individuals.

As more pools are sampled, the probability of selecting the correct order increases but the number of pools sampled has little effect on the optimum pool size. This is illustrated in Figure 2B for sample sizes of N = 20, 40 and 80. The shapes of the three curves are very similar suggesting that the probability of correct ordering increases uniformly for all pool sizes.

The critical value of the decision rule also effects the probability of choosing the correct order but has little effect on the optimum pool size. Figure 2C illustrates that optimal pool sizes are not altered by changing the cut off. We would like to note that most of the curves in Figure 2 are well below 1.0 because we have used a high critical value (c = .90 for most of the curves). Although the true order may have the highest posterior probability, if this probability does not exceed c, no decision is made. If the decision rule is very stringent (c=0.99) large sample sizes may be needed to assign order unambiguously but confidence in the chosen order is very high.

Misclassification. When forming pools of individuals of type t/t it is possible that one or more individuals of type T/t or T/T may be misclassified and included in a pool. Such a pool may appear to contain recombinant chromosomes. Misclassification will alter the outcome probabilities on which order inferences are based. If the misclassification rate is α per chromosome (i.e., q for T/t heterozygote and q^2 for T/T homozygote), the new chromosome type probabilities are

$$q_S^* = (1-\alpha)q_S + \alpha\, q_{M\backslash S} \, .$$

The pool type probabilities are as before. The misclassification rate can be estimated by extension of the EM algorithm but the expected precision of this estimate is very low.

We have studied the effects of misclassification by simulating data with misclassification and analyzing the data assuming none. If the misclassification rate is low (1 to 2 percent) correct ordering inferences are made with high probability. As misclassification increases, fewer correct inferences are made. The effect is more pronounced for large pool sizes ($2k > 20$) and for high marker densities. Figure 2D illustrates the effect of misclassification. At higher misclassification rates, the optimum pool size is slightly reduced and the loss of efficiency due to overpooling is more serious. The posterior probability of the true order is obtained assuming that there is no misclassification. The presence of a misclassified individual in a pool will generally reduce the posterior probability for the two markers flanking the target and thus the procedure is conservative. Pools which contain a misclassified individual will typically be recombinant for all markers and will inflate the estimated recombination fractions in the intervals immediately flanking the targe locus. Pools which are recombinant for markers flanking the target but are non-recombinant for more distant markers contain an obligate double crossover or a misclassification. The phenotypes of individuals in such pools should be rechecked if possible (see the example below).

Technical limits in detecting recombinants in a pooled sample. One of the assumptions of pooled mapping is that a single recombinant chromosome can be detected in a pool of otherwise non-recombinants. The degree to which this is technically possible depends on the organism being studied and the molecular detection techniques being employed. RFLPs represent the common type of molecular marker now being used in higher organism and they are normally detected on southern blots with single copy probes (1). Tomato (*Lycopersicon esculentum*), like most eucaryotes, has a genome that is relatively large and complex (haploid DNA content = 900 Mb). To test the limits of pooling, DNA from tomato plants of two different genotypes was mixed and subjected to southern analysis with a single copy DNA probe (REF). The results indicate that in a mixture as great as 40:1, one can still detect the rare allele. This is comparable to detecting a single recombinant chromosome in a pool of 20 plants. However, since the pooling is done before DNA extraction, and is based on utilizing approximately equal amounts of tissue from individuals within the pool, there is room for additional error. For this reason, for the purposes of testing pooled-mapping, we decided to use pools of five plants (see next section).

Pooled mapping of the rin locus. To test the pooled mapping strategy, a large F2 population (1840 plants), segregating for the *rin* (ripening inhibitor) gene, was planted in the field and grown to maturity. Fruit from plants homozygous for the recessive *rin* allele do not ripen and it is believed that rin represents an upstream regulatory switch for the ripening process (21). *rin* is on chromosome 5 and a number of DNA markers have been identified that are in the vicinity of the gene (22). 190 plants were unambiguously determined to be homozygous (*rin/rin*) and tissue from these plants was pooled into groups of 5 to form 38 pools. DNA was extracted from each pool and scored for six RFLP markers known from previous, experiments to be linked to the *rin* locus. After the pools were formed and scored, a misclassified plant was discovered in one pool. Thus, the analysis presented is based on the remaining 37 pools.

Previous analyses suggested that the seven loci (*rin* plus six RFLP markers) are located in approximately a 20 cM interval (22). Thus the local density is 35 markers/100cM. The prior distribution for recombination probabilities was taken to be Beta (a=1, b=15). All orders were assumed to be *a priori* equal likely.

Two point analyses were carried out to estimate distances between the marker and the target. Results are summarized in Table 1. The left and right groupings were readily established by three-point analysis. Posterior probabilities for ordering adjacent markers are shown in Table 2. The critical ordering inference involves markers **E** and **B** which appear to span the *rin* locus. The posterior probability of the three-point ordering ETB is 0.899, strong evidence in favor of this ordering. When the data of 38 pools, including the misclassified pool, are analyzed there is only one substantial change in the results (see Table 2). The three-point order ETB is still the most likely order of markers immediately flanking the target, however, the posterior probability of this order is reduced to 0.709.

Discussion.

We have presented the basic theory necessary to create high resolution genetic maps using pooled DNA samples. Our results suggest that this is a practical and highly efficient approach to high resolution mapping of DNA markers. A number of factors were shown to affect the probability of choosing the correct order for a set of markers. We summarize these findings:

The pool size can be optimized to yield maximal probability of choosing the correct order. The optimal pool size is determined primarily by the local density of markers. For a low marker density, small to moderate pool sizes are most likely to yield the correct order. Overpooling can result in an inefficient experiment. If the local density of markers is higher, larger pool sizes are optimal and the analysis is more robust to overpooling.

As more pools are sampled, the probability of the correct order increases. This consistency property may hold even when certain model assumptions are violated. In particular, Speed et al. (20) have shown that inferences of order assuming a no-interference model are consistent in the presence of interference. For the densities considered in our simulation studies, (up to 128 markers/100 cM) sample sizes of 40 to 80 pools can yield high posterior probabilities for the correct order. Increasing the stringency with which we will accept an order as being correct decreases the probability of making a correct decision (by increasing the probability of no decision).

Applications of pooled mapping. Map-based cloning represents one of the most promising strategies for isolating genes known only by the phenotype they impart (23). High resolution mapping is a prerequisite for map-based cloning and the pooled method described in this paper can facilitate this process and may therefore aid in the isolation of new genes from both plants and animals. In tomato alone, there are more than 1000 genes identified by the phenotype they impart to the plant (24). Included in the list are genes for resistance to a broad spectrum of plant pathogens as well as genes controlling-differentiation and plant architecture. Currently none of these genes have been cloned, but all are prime candidates for map-based cloning. The same situation exists for most crop species as well as model species for genetic and molecular research including

Arabidopsis, Drosophila and mouse. Pooled mapping could be utilized in these species in the same manner as demonstrated here for tomato. Moreover, in species with smaller genome (less DNA), larger numbers of individuals could be pooled making this strategy even more effective.

Finally, it should be noted that pooled-mapping results in the ordering of all markers in the vicinity a scorable locus. Therefore easily scorable loci can be used to develop region-specific, high resolution maps, even if the scorable loci themselves are not the target of map-based cloning. Thus high resolution maps might be constructed for genes that are not readily assayed, but which are linked to genes that are easily scored, including quantitative trait loci.

Acknowledgement

This work was supported in part by grants from the National Research Initiative Competitive Grants Program, USDA nos. 91-37300–6565 and 91-37300-6418 and by the Binational Agricultural Research and Development Fund (no. IS–1822–90C). Thanks to Walter Federer for helpful discussions and to April Denman and Colleen Bushnell for preparing the manuscirpt.

References

1. Botstein, D., White, R.L., and Skolnick, M.H. (1980) *Amer. J. Hum. Genet.* **32**: 314-331.
2. Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., Botstein, D., Akots, G., Rediker, K.S., Gravius, T., Brown, V.A., Rising, M.B., Parker, C., Powers, J.A., Watt, D.E., Kauffman, E.R., Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T.R., Ng, S., Schumm, J.W., Braman, J.C., Knowlton, R.G., Barker, D.F., Crooks, S.M., Lincoln, S.E., Daly, M., and Abrahamson, J. (1987) *Cell* **51**: 319-337.
3. Copeland, N.G. and Jenkins, N.A. (1991) *TIG* **7**: 113-118.
4. Chang, C., Bowmna, J.C., DeJohn, A.W., Lander, E.S. and Meyerowitz, E.S. (1988) *Proc. Natl. Acad. Sci. USA* **85**: 6856-6860.
5. Helentjaris, T. (1987) *Trends in Genet.* **3**: 217-221.
6. Davies, K. (1991) *Nature* **353**: 798-799.
7. Van Dilla, M.A., Deavan, L.L., Albright, K.L., Allen, N.A., Aubuchon, M.R., Bartholdi, M.F., Brown, N.C., Campbell, E.W., Carrano, A.V., Clark, L.M., Cram, L.S., Crawford, B.D., Fuscoe, J.C., Gray, J.W., Hildebrand, C.E., Jackson, P.J., Jett, J.H., Longmuire, J.L., Lozes, C.R., Luedemann, M.L., Martin, J.C., McNinch, J.S., Meincke, L.J., Mendelsohn, M.L., Meyne, J., Moyzis, R.K., Munk, A.C., Perlman, J., Peters, D.C., Silva, A.J., Task, B.J. (1986) *Bio/Technology* **4**: 537-552.
8. Saunders, R.D.C., Glover, D.M. Ashburner, M., Siden-Kiamos, I., Louis, C., Monastirioti, M., Savakis, C., and Kafatos, F. (1989) *Nucleic Acids Research* **17**: 9027-9037.
9. Martin, G.B., Williams, J.G.K., and Tanksley, S.D. (1991) *Proc. Natl. Acad. Sci. USA* **88**: 2336-2340.
10. Michelmore, R.W., Paran, I., and Kesseli, R.V. (1991) *Proc. Natl. Acad. Sci. USA*

(in press).

11. Giovannoni, J.J., Wing, R.A., Ganal, M.W., and Tanksley, S.D. *Nucleic Acid Research* **19**: 6553-6558.
12. Messeguer, R., Ganal, M., de Vicente, M.C., Young, N.D., Bolkan, H., and Tanksley, S.D. (1991) *Theor. Appl. Genet.* **82**: 529-536.
13. Berger, J.O. (1986) *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed. Springer, New York.
14. Smith, C.A.B. (1958) *Ann. Eugen.* **21**: 254-276.
15. Lander, E.S. and Green, P. (1987) *Proc. Nat. Acad. Sci. USA* **84**: 2363-2367.
16. Thompson, E.A. (1984) *IMA Journal of Math. Appl. in Med. and Biol.* **1**: 31-49.
17. Thompson, K.H. (1962) *Biometrics* **18**: 568-578.
18. Tanner, M.A. (1991) *Tools for Statistical Inference, Observed Data and Data Augmentation Methods.* Springer, New York.
19. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) *Journal R. Statist. Soc. B.* **39**: 1-22.
20. Speed, T., McPeek M.S. and Evans, S.N. (1991) Tech. report No. 308, Department of Statistics, University of California, Berkeley.
21. Grierson, D. (1985) CRC Crit. Rev. Plant Sci. **3**: 113-132.
22. Giovanonni and Tanksley (unpublished data)
23. Wicking, C., Williamson, B. (1991) From linked marker to gene. Trends in Genetics 7: 288-293.
24. Rick, C.M. (1975) The tomato. In (R.C. King, ed.) Handbook of Genetics. Plenum Press pp. 247-280.

Table 1: Estimated distances from target locus *rin*

marker	$\hat{r}$	standard error
CT93	0.0353	0.0107
TG96	0.0315	0.0100
ACC4	0.1046	0.0223
CD64	0.0903	0.0199
TG503	0.0056	0.0039
TG448	0.0838	0.0188

Approximate recombination probabilities (Equation 7) between markers and the target gene and their standard errors (Equation 8) were estimated from two-point data.

Table 2: Three point ordering probabilities

	posterior probability	
ordered markers	N = 37	N = 38
CD64–CT93–rin	0.9993	0.9996
CT93–TG503–rin	0.9237	0.9749
TG503–rin–TG96	0.8978	0.7227
rin–TG96–TG448	0.9991	0.9995
rin–TG448–ACC4	0.9994	0.9994

Posterior probabilities (equation 1) were computed for ordering all 15 pairs of markers relative to the target. The prior distribution for spacings was taken to be Beta (a=1, b=15). The integral in equation 1. was computed by the Monte-Carlo method using 10,000 random samples from Pr(R|H) for each order H. Only 5 of the 15 ordered triplets are needed to confirm the map intervals in Figure 3. The second column shows order probabilities using the data from 37 pools. The third column shows results when all 38 pools including the misclassification are analyzed.

Figure 1: Pool type probabilities.

The pool-type probabilities from probing a pooled sample of DNA with two markers A and B (Equation 10) are shown as a function of pool size (k = number of F2 individuals per pool). The assumed locus order is H_1: $A - T - B$ and the recombination probabilities are $r_1 = 0.05$ and $r_2 = 0.02$ in this example.

Figure 2: Probability of correct order.

The probability of determining the correct ordering of three adjacent loci (two markers and the target) is computed to be 1 minus the Bayes risk (Equation 5) and is plotted as a function of pool size. All orders are assumed to be equally likely. The effects of marker density D, sample size N, critical value for the decision rule c and misclassification rate α were studied. The curves shown are based on constant values of D = 0.32 markers per cM, N = 40 pools, c = 0.90 and $\alpha = 0.0\%$ except that one factor is varied as follows:

A. The marker density is varied over 0.16, 0.32, 0.64 and 1.28 markers per cM.

B. The sample size is varied over N = 20, 40 and 80 pools.

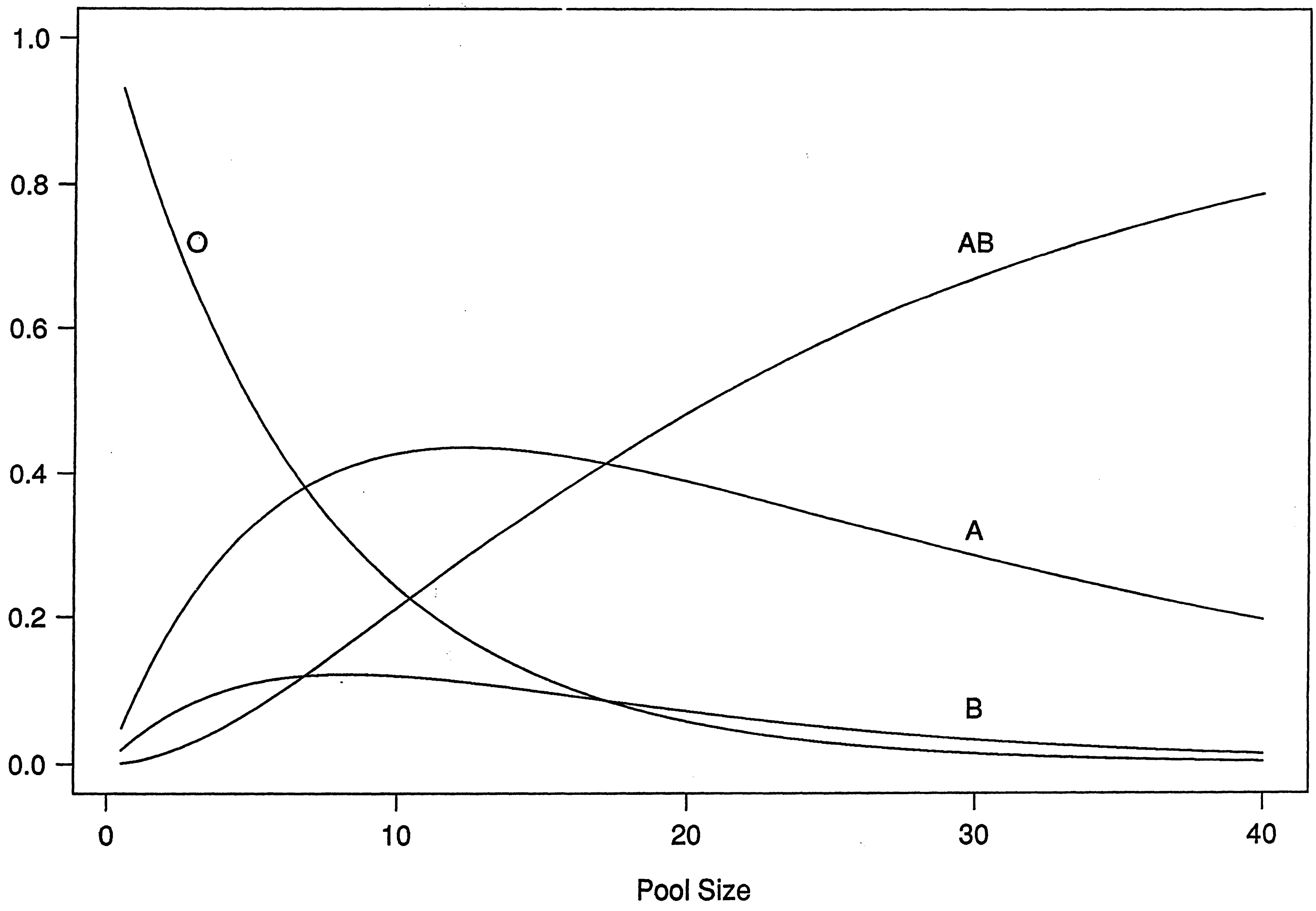
C. The critical level of posterior probability for calling a correct order is varied over 0.5, 0.90, 0.95 and 0.99.

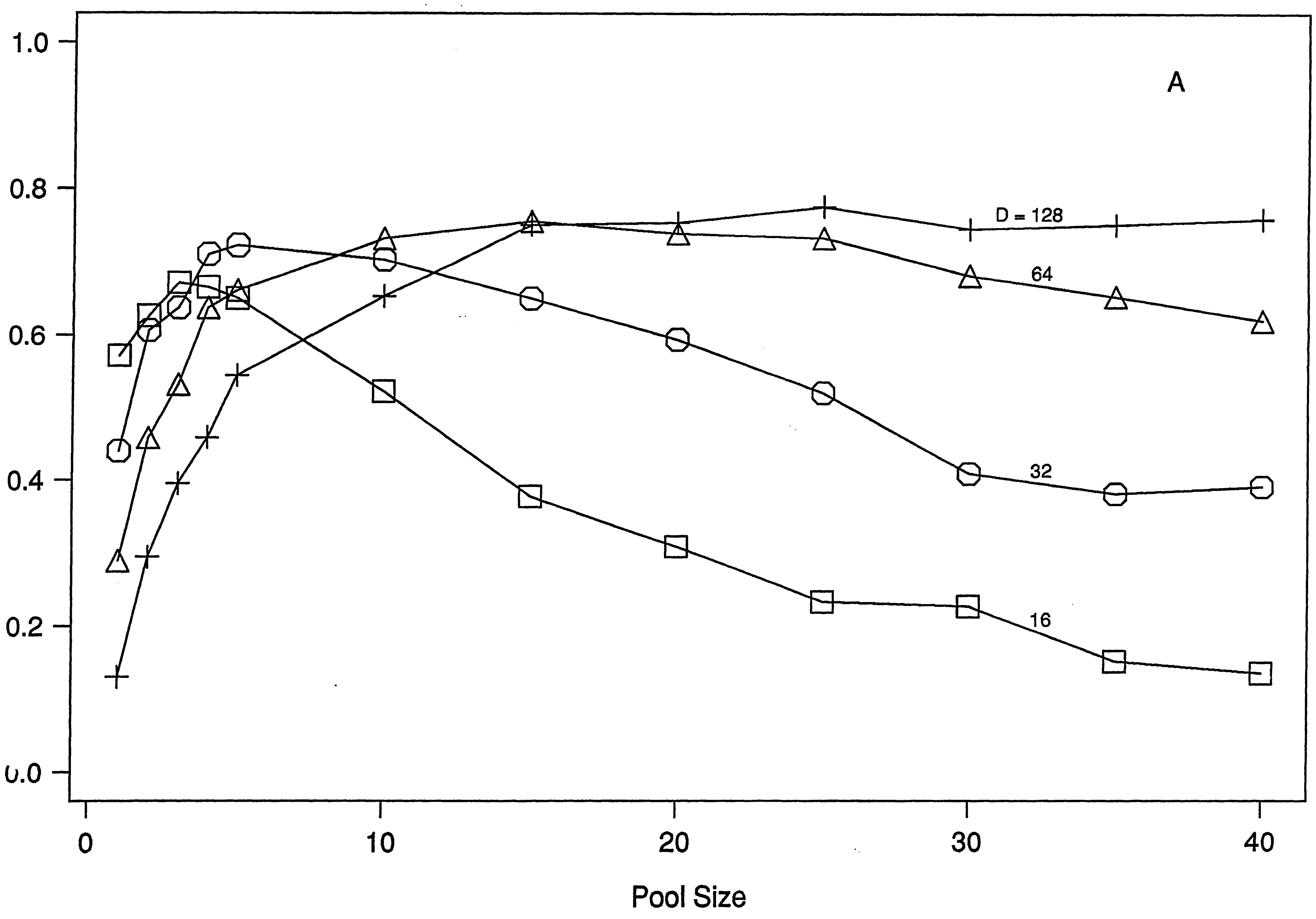
D. The proportion of misclassifed chromosomes is varied over 0%, 1%, 2% and 5%.

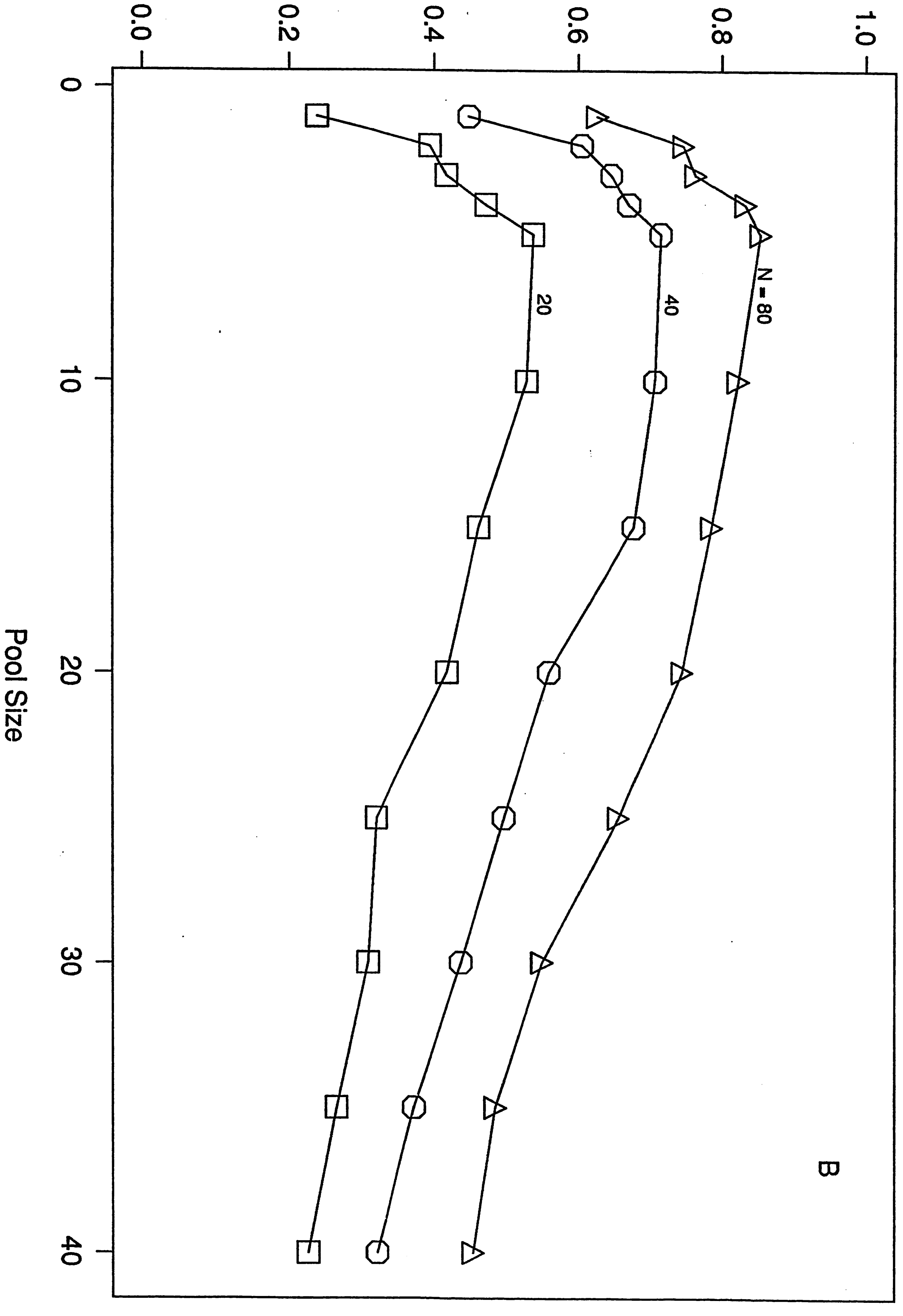
We wish to note that the sample size and critical values were chosen to give curves for which the correct order probability is well below one. Such curves are most informative regarding optimal pool sizes. In general, the true order is found to have the highest posterior probability and by reducing the stringency or increasing the sample size the correct order will be chosen with probability near one.

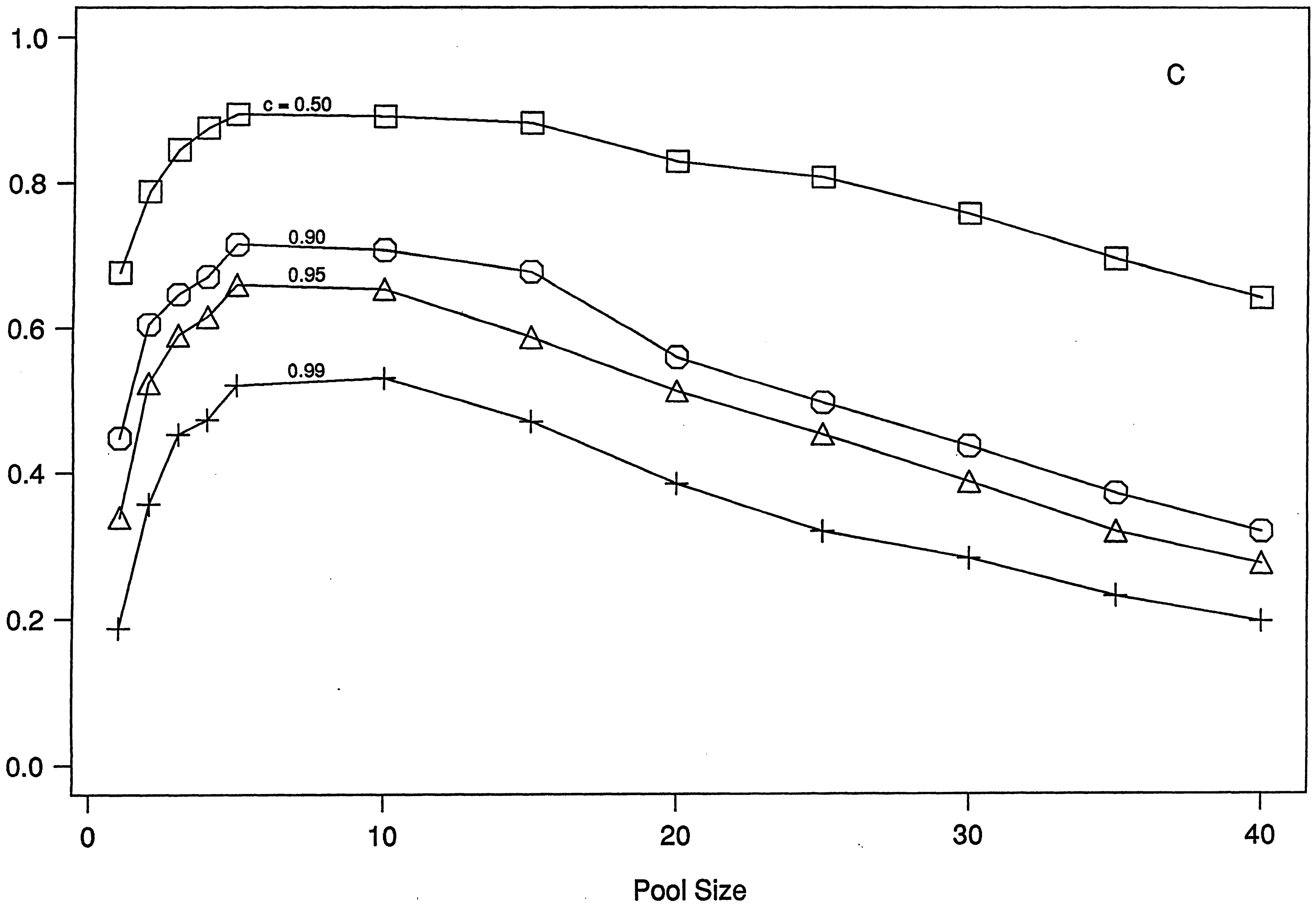
Figure 3. Map of the *rin* region.

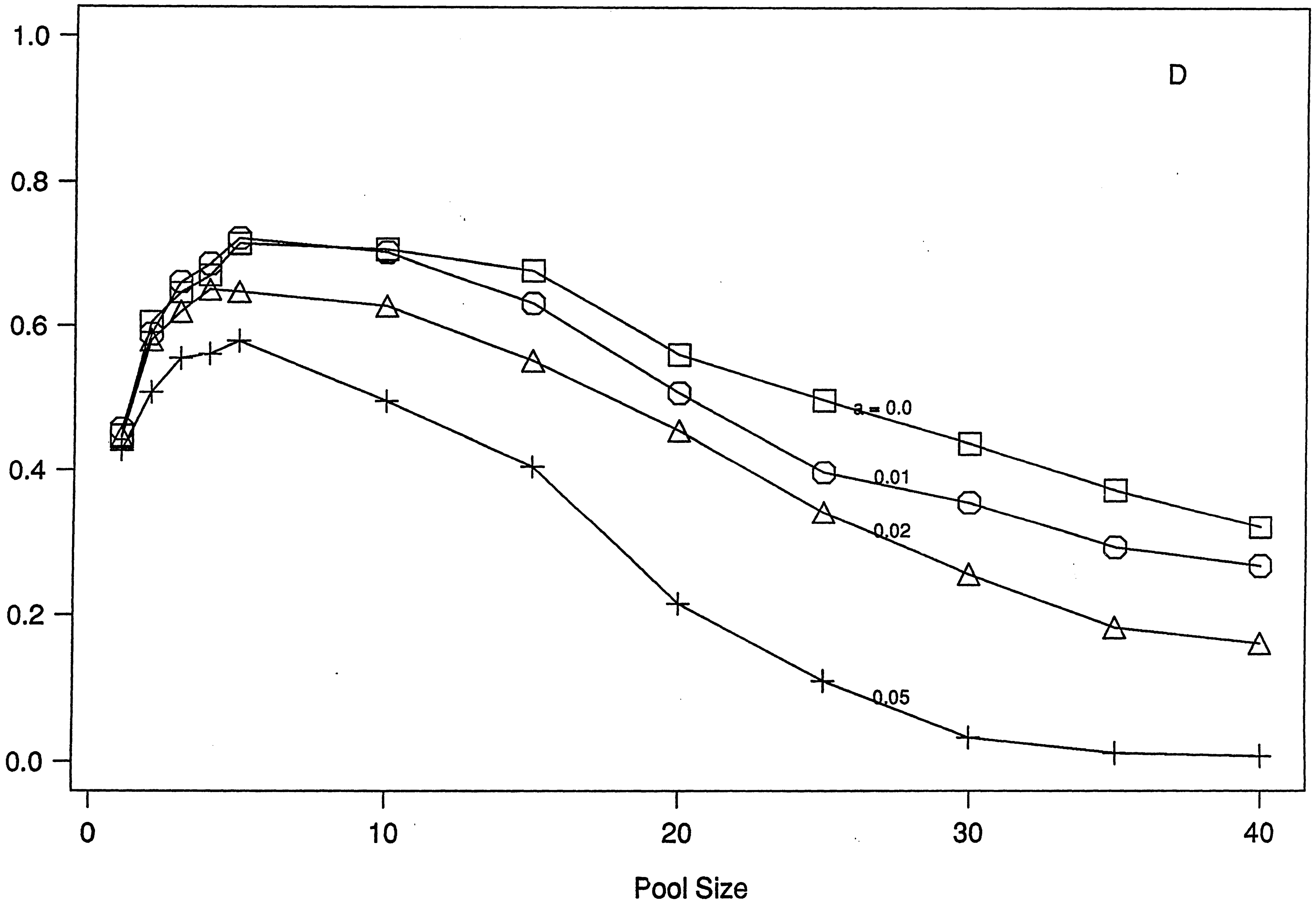
The positions of seven markers relative to the target locus *rin* are shown. The distances indicated on the figure are estimated percent recombination (≈ cM) between each marker and the target are shown. Precisions of the estimates are indicated by 95% confidence intervals $\left(\hat{r} \pm 2\sqrt{\mathrm{Var}(\hat{r})}\right)$, shown as vertical bars to the left of the figure. Note that the individual markers are not mapped independently and thus overlapping confidence intervals do not necessarily indicate uncertainty in the map order (see table 2).

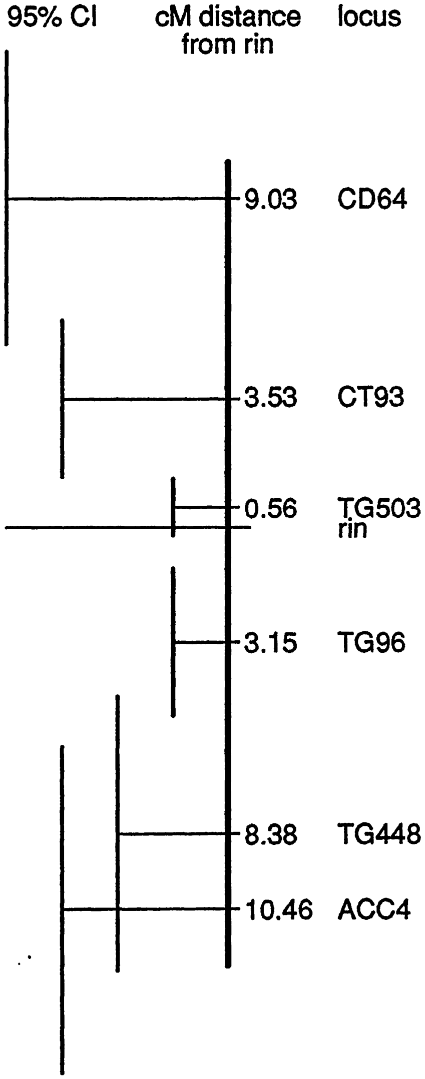

1.0
0.8
0.6
0.4
0.2
0.0
O
AB
A
B
0
10
20
30
40
Pool Size


A
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
Pool Size
D = 128
64
32
16


B
N = 80
40
20
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
Pool Size


C
c = 0.50
0.90
0.95
0.99
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
Pool Size


1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
Pool Size
D
a = 0.0
0.01
0.02
0.05


95% CI
cM distance from rin
locus
9.03
CD64
3.53
CT93
0.56
TG503
rin
3.15
TG96
8.38
TG448
10.46
ACC4