Copyright © 1994 by the Genetics Society of America

Selective DNA Pooling for Determination of Linkage Between a Molecular Marker and a Quantitative Trait Locus

A. Darvasi and M. Soller

Department of Genetics, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

Manuscript received May 19, 1994
Accepted for publication August 24, 1994

ABSTRACT

Selective genotyping is a method to reduce costs in marker-quantitative trait locus (QTL) linkage determination by genotyping only those individuals with extreme, and hence most informative, quantitative trait values. The DNA pooling strategy (termed: "selective DNA pooling") takes this one step further by pooling DNA from the selected individuals at each of the two phenotypic extremes, and basing the test for linkage on marker allele frequencies as estimated from the pooled samples only. This can reduce genotyping costs of marker-QTL linkage determination by up to two orders of magnitude. Theoretical analysis of selective DNA pooling shows that for experiments involving backcross, F_2 and half-sib designs, the power of selective DNA pooling for detecting genes with large effect, can be the same as that obtained by individual selective genotyping. Power for detecting genes with small effect, however, was found to decrease strongly with increase in the technical error of estimating allele frequencies in the pooled samples. The effect of technical error, however, can be markedly reduced by replication of technical procedures. It is also shown that a proportion selected of 0.1 at each tail will be appropriate for a wide range of experimental conditions.

WITH the development of methods for uncovering genetic variation at the DNA-level, it was evident that this would allow the widespread application to agricultural and experimental species of experimental designs aimed at mapping individual loci affecting quantitative traits (SOLLER and BECKMANN 1982, 1983; BECKMANN and SOLLER 1983, 1986). Genetic analyses of quantitative traits based on DNA-level markers have indeed been carried out successfully (HILBERT *et al.* 1991; NIENHUIS *et al.* 1987; PATERSON *et al.* 1988). However, identifying linkage between marker loci and quantitative trait loci (QTL) having effects of low or moderate size requires large samples for useful power (DARVASI *et al.* 1993; KASHI *et al.* 1990; SOLLER *et al.* 1976; SOLLER and GENIZI 1978; WELLER *et al.* 1990). Thus costs of marker genotyping for these applications are relatively high, limiting their utilization for genetic analysis and genetic improvement (KASHI *et al.* 1990).

In some cases it is possible to reduce the amount of genotyping, by genotyping only individuals at the phenotypic extremes of the population (DARVASI and SOLLER 1992; LANDER and BOTSTEIN 1989; LEBOWITZ *et al.* 1987), a procedure termed "selective genotyping." HILLEL *et al.* (1990) and PLOTSKY *et al.* (1990, 1993) used DNA pooling procedures to identify a DNA fingerprint band in linkage to a QTL affecting abdominal fat pad weight. MICHELMORE *et al.* (1991) used DNA pooling to locate a genetic marker in close linkage to a resistance gene. They crossed a resistant and susceptible inbred line, and then pooled the DNA of resistant and susceptible F_2 segregants. Markers that were in linkage with the resistance gene showed a marked difference in frequency in the two alternative DNA pools. PACEK *et al.* (1993) and KHATIB *et al.* (1994) have shown that relative frequencies of microsatellite alleles can be determined by quantitative densitometry analysis of DNA amplified from pooled samples. Similarly, highly accurate densitometry methods for quantifying radioactive patterns on gels and blots, based on phosphor storage technology, have been developed (JOHNSTON *et al.* 1990). These can readily be applied to estimation of marker allele frequencies in genotyping procedures based on Southern blot analysis. Thus, in principle it would appear feasible to identify linkage between marker loci and QTL by combining DNA pooling and selective genotyping (henceforth: selective DNA pooling). In this procedure, two DNA pools are formed. One pool would include individuals with high phenotypic value for the trait of interest; the other would include individuals with low phenotypic value. Quantitative densitometry of allelic bands is then employed to provide estimates of relative allele frequencies in the two pools. Thus, irrespective of the total number of individuals actually included in the selected tails, when using selective DNA pooling each marker is genotyped only twice, once in each pooled sample (although in practice some replication might be required to reduce technical error). Consequently, selective DNA pooling can provide a substantial reduction in the total amount of genotyping required for marker-QTL linkage determination, as compared to selective genotyping with

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

separate genotyping of each individual in the selected tails (henceforth: "standard selective genotyping").

The present study provides a statistical test for marker-QTL linkage, based on selective DNA pooling, and a detailed examination of the statistical attributes of this procedure.

THEORY

The theoretical analysis of selective DNA pooling differs from standard selective genotyping in the following aspects: (i) In pooled DNA samples only marker allele frequencies can be estimated, whereas in standard selective genotyping the individual marker genotypes are obtained, (ii) In pooled DNA samples, marker allele frequencies will normally be estimated with some degree of technical error, and (iii) When analyzing data obtained through selective DNA pooling, the quantitative trait value of each individual cannot be individually assigned to a particular marker genotype, since information is not available on individual genotypes. Nevertheless, by using the allelic frequencies at each tail to estimate the respective genotype frequencies and by assigning the sample average at each tail to every individual at that tail, the problems raised by the above differences can be overcome.

We consider the case where a molecular marker, with alleles M and m, is in absolute linkage to a QTL with alleles Q and q. When the marker is not in absolute linkage to the QTL, experimental power decreases and the estimate of QTL effect is biased downward, but this does not affect the mode of analysis (Darvasi and Soller 1992). Standardized values of the QTL genotypes QQ, Qq and qq are assumed normally distributed with means d, h and $-d$, respectively.

Selective DNA pooling provides estimates of marker allele frequencies in the selected tails. In principle and in practice, therefore, a test for marker-QTL linkage determination can simply be based on the differences in estimated marker allele frequencies in the two pools, relative to their standard error (Khatib *et al.* 1994). It turns out, however, that the different experimental designs (backcross, F_2, half-sib) affect marker allele frequencies in the two tails differently. Consequently, it proved cumbersome to develop expressions for the power of selective DNA pooling based on significance tests of the difference in marker allele frequencies in the two tails of the experimental population. Instead, the theory is developed according to the usual approach to power calculations in simple tests for marker-QTL linkage determination, which are often carried out by defining two marker genotypic groups and comparing mean trait value of these groups. For example, in a backcross design, marker genotypes Mm and mm are compared, while in an F_2 design, marker genotypes MM and mm are compared (Soller *et al.* 1976). Similarly, in the present instance, power calculations are also developed in terms of two marker genotypic groups that are relevant to QTL mapping, and whose frequencies in the pooled tails can be derived from estimates of marker allele frequencies in the pools. For generality, the two marker genotypic groups are designated A and B. Using this model, a uniform and general test for marker-QTL linkage based on marker allele estimates obtained from pooled DNA samples is derived, and its statistical attributes explored. Specific application of the general model to particular experimental designs is deferred to a later section.

As will be detailed in the section on specific designs, the genotypic groups A and B can be defined: (i) as marker genotypes *per se* (*e.g.*, in a backcross population, genotypic group A represents marker genotype Mm, and genotypic group B represents marker genotype mm), or (ii) through their marker allele composition (*e.g.*, in a half-sib population, genotypic groups A and B represents progeny that received marker allele M and m, respectively, from their common sire). For the specific designs considered, the designated genotypic groups, A and B, have equal expected frequency in the experimental population, and their respective quantitative trait values are assumed normally distributed with means and variances μ_A, μ_B and $\sigma_A^2 = \sigma_B^2$. In many instances, the gene effect at the QTL, d, is small, so that the QTL in question contributes only a small component to the overall population variance. In this case $(\mu_A - \mu_B)/\sigma_{A \text{ or } B}$ is also small so that σ_A^2 and σ_B^2 are approximately equal to σ^2, the overall population variance. This ensures that the approximations used are appropriate. For theoretical analysis, it is convenient to standardize trait values, setting the population variance to 1 and the expectations of the genotypic groups A and B to δ (termed: the genotypic group effect) and $-\delta$ respectively, so that the difference between them is equal to 2δ.

Testing for marker-QTL linkage: Although, as mentioned above, a test for marker-QTL linkage can be based directly on the densitometric estimates of marker allele frequencies in the two tails, in order to maintain consistency with the power calculations the test developed here will be based on the marker genotypic group frequencies at the tails, as estimated from the marker allele frequencies. The exact procedure for obtaining such estimates depends on the specific experimental design employed for linkage analysis, and will be detailed later for a number of specific designs. Given such estimates, however, the test for marker-QTL linkage is based on rejecting the null hypothesis that the expected relative frequency of genotypic group A in the upper tail and of genotypic group B in the lower tail, denoted π, equals 0.5. Using the Normal distribution approximation, the rejection criteria will be: reject the null hypothesis with type I error, α, if:

$$\hat{\pi} - 0.5 > \sqrt{\text{Var}(\hat{\pi})} Z_{1-\alpha/2} \qquad (1)$$

where, $\hat{\pi}$ is the estimate of π, $\text{Var}(\hat{\pi})$ is the variance of

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

$\hat{\pi}$, and $Z_{1-\alpha/2}$ is the ordinate of the standard normal distribution such that the area from $-\infty$ to $Z_{1-\alpha/2}$ equals $1 - \alpha/2$.

Depending on the specific experimental design, the estimate of π is obtained as a function of allelic band intensity in a specific gel. In principle, estimates of π can be obtained by separately estimating the frequencies of the alleles M and m in each tail relative to a standard. This would provide four independent estimates of π (two from each tail). However, in practice, it was found that quantifying band intensities using an internal control is more accurate than through an external standard (NEDELMAN *et al.* 1992). Therefore we envision that estimating the frequencies of M and m will be carried out relative to one another at each DNA pool. This will provide two independent estimates, one from the M allele at the upper tail and one from the m allele at the lower tail, and their average is used to estimate π. Since at each tail alternative alleles are used for estimating π, any biases in the expected allelic frequencies due to differential amplification, or unknown PCR bands or PCR "shadow" bands, are eliminated. In the following derivations we assume that this procedure is followed.

The variance of $\hat{\pi}$ includes two components, one being the variance due to binomial sampling of the genotypic groups in each tail; the other being the variance generated from the technical procedure of estimating π from band intensity in a gel. Thus, where p is the proportion selected over both tails ($p/2$ at each tail); N is the total population size; and V_π is the technical component of the variance of $\hat{\pi}$ in one of the tails,

$$\mathrm{Var}(\hat{\pi}) = \frac{0.25}{pN} + \frac{V_\pi}{2} \tag{2}$$

henceforth denoted *experimental error variance*. V_π is a simple function (depending on experimental design) of the *technical error variance*, V_T, of estimating the frequency of a particular allele in one of the tails. As previously described, two independent estimates of π are available (one from each tail) and their average is used to estimate π. Consequently, both binomial and experimental error variance are reduced by a factor of 2.

Power of the test: The expected genotypic group frequency, π, can be approximated as:

$$\pi = \frac{\Phi(Z_{p/2} + \delta)}{p} \tag{3}$$

where $\Phi(\ )$ is the cumulative normal distribution function (DARVASI and SOLLER 1992). The power of the test can then be obtained from the properties of the normal distribution,

$$Z_{1-\beta} = \frac{\pi - 0.5}{\sqrt{\mathrm{Var}(\hat{\pi})}} - Z_{1-\alpha/2} \tag{4}$$

where α is the type I error and $1 - \beta$ is the power. Power calculations can be obtained by incorporating Equations 2 and 3 into Equation 4:

$$Z_{1-\beta} = \frac{\Phi(Z_{p/2} + \delta)/p - 0.5}{\sqrt{0.25/pN + V_\pi/2}} - Z_{1-\alpha/2} \tag{5}$$

This power can be compared with the power of standard selective genotyping as presented by DARVASI and SOLLER (1992).

The optimal proportion selected: If DNA is available for all individuals in the total sample population, any proportion, p, of the sample can be chosen for pooling and experimental costs will not depend on the proportion chosen. Therefore the optimal proportion selected will be the one that maximizes the power of the experiment, as given by Equation 5. Maximization of Equation 5 as a function of p must be carried out numerically. Nevertheless, it can be seen directly from Equation 5 that the only parameters that can influence the optimal proportion are δ and the product $NV\pi$.

Estimating the genotypic group effect: Let A_s and B_s represent groups comprising all individuals from *both selected tails* with genotypes A and B, respectively; and $\bar{A}_s$ and $\bar{B}_s$ represent the estimated mean value of groups A_s and B_s, respectively. These are computed as follows:

$$\bar{A}_s = \hat{\pi}\bar{X}_U + (1 - \hat{\pi})\bar{X}_L \tag{6}$$

$$\bar{B}_s = (1 - \hat{\pi})\bar{X}_U + \hat{\pi}\bar{X}_L \tag{7}$$

where $\bar{X}_U$ and $\bar{X}_L$ are the sample trait averages of the upper and lower selected tails, respectively. Let D_T be the expected difference of the genotypic groups means. Then, since $E(\bar{X}_U) = -E(\bar{X}_L) = i_{p/2}$ (FALCONER 1989),

$$D_T = E(\bar{A}_s) - E(\bar{B}_s) = 2(2\pi - 1)i_{p/2} \tag{8}$$

where $i_{p/2} = 2x_{p/2}/p$ is the mean of an upper tail of a standard normal distribution, and $x_{p/2}$ is the ordinate of the standard normal distribution at the point $Z_{p/2}$. For small values of δ, D_T was found to be a linear function of δ (DARVASI and SOLLER 1992). Therefore, Taylor's expansion of the first order was used on Equation 8 with π represented by Equation 3, to provide an accurate approximation of D_T as a function of δ:

$$D_T = 2\delta(i_{p/2})^2. \tag{9}$$

Thus, under this procedure the actual *genotypic group effect* can be estimated as:

$$\hat{\delta} = \frac{(\bar{A}_s - \bar{B}_s)}{2(i_{p/2})^2}. \tag{10}$$

Specific expressions for estimating the *QTL effect* (*i.e.*, d and h) from the *genotypic group effect* depend on the specific experimental design, and are given in the following section.

Application to specific experimental designs: The general procedure will now be applied to backcross (BC), F_2 and half-sib designs. Backcross and F_2 designs are appropriate for analysis of data originating in a cross between two inbred lines with marker-QTL genotypes

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

MQ/MQ and *mq/mq*, respectively; the half-sib design is appropriate for analyzing data originating from a single sire, heterozygous at the QTL and at marker loci, in an outcrossing population, segregating for marker and QTL alleles. The half-sib design is particularly attractive for analysis of dairy cattle and other livestock populations, where very large half-sib families are often produced through artificial insemination.

In applying the general theory developed above to specific experimental designs it is necessary to define the following four elements for each design: (i) The genotypic groups corresponding to *A* and *B*, (ii) a mode of obtaining independent estimates of the genotypic group frequency, π, at each tail, from the estimated marker allele frequencies: M_U and m_U in the upper tail, and M_L and m_L in the lower tail, (iii) an expression for V_π in terms of V_T, and (iv) a mode of expressing the genotypic group effect δ, in terms of the QTL effects, *d* and *h*. Elements (i) to (iii) are required to carry out the statistical test for linkage. Element (iv) is required for power calculations and for QTL gene effect estimation.

In a backcross population: (i) The two genotypic groups, *A* and *B*, are represented by the two marker genotypes, *Mm* and *mm*, respectively. (ii) Independent estimates of π, the frequency of *Mm* in the upper tail and *mm* in the lower tail, are provided respectively, by $2M_U$ in the upper tail and by $2m_L - 1$ in the lower tail. (iii) Note that, in estimating π, the allelic frequency estimates are doubled. Hence V_π, the *experimental variance* of the estimate of π, is four times the *technical error variance* of the estimate of the marker allele frequency. (iv) Since groups *A* and *B* represent genotypes *Mm* and *mm* in the backcross design, the genotypic group effect, δ, represents $(d + h)/2$.

In an F_2 population: The model developed here applies exactly for a QTL showing co-dominance (i.e., $h = 0$). Nevertheless, as argued later, using this model for the analysis of a QTL with $h \neq 0$ in an F_2 population does not significantly alter the results. In considering the situation where $h = 0$, (i) the heterozygous individuals at the QTL (QTL genotype *Qq*, and marker genotype *Mm*) will be symmetrically distributed at both extremes and so are not informative in the analysis. Consequently, the two genotypic groups *A* and *B*, can be represented by the homozygous marker genotypes *MM* and *mm*, respectively. Note that since marker heterozygotes are not included in the analysis, the effective population size is half the total sample size and $N/2$ should be used instead of N in Equations 2 and 5. (ii) For small gene effects and with co-dominance at the QTL ($h = 0$), the frequency of the genotype *Mm* is approximately ½ at each tail. For example, for gene effect of $2d = 0.25$ and analyzing a proportion of 0.25 or 0.05 at each tail, the proportion of the heterozygotes at each tail will be 0.498 or 0.488, respectively. Consequently, independent estimates of π, the frequency of *MM* in the upper tail and *mm* in the lower tail, are provided respectively, by $2M_U - ½$ in the upper tail and by $2m_L - ½$ in the lower tail. (iii) As in the BC design, V_π is fourfold the technical error variance of the estimate of the marker allele frequency. (iv) Since groups *A* and *B* represent genotypes *MM* and *mm* in the F_2 design, the genotypic group effect, δ, represents *d*.

In the F_2 design a further increment, $0.5/pN$, in the variance of $\hat{\pi}$, as given in Equation 2, is present due to the binomial variance of the heterozygous individuals.

As noted above, when dealing with a QTL showing partial or complete dominance ($h \neq 0$) the model presented in this study does not provide exact estimates of power gene effect for an F_2 design. This is due to the fact that the two tails are no longer symmetrical with respect to marker allele frequencies. Nevertheless the effect of dominance at the QTL on marker allelic frequencies in the two tails are at least partially compensated. Dominance at the QTL causes an increase in the frequency of *m* at the upper tail (the marker allele in linkage to the QTL allele associated with low trait value), but a decrease in the frequency of marker allele *M* in the lower tail. Consequently, the influence of partial or even complete dominance at the QTL on power and estimated gene effect in an F_2 population is expected to be moderate. For example, numerical calculations show that for a gene effect of $2d = 0.25$ and analyzing a proportion of 0.25 at each tail, complete dominance does not have a noticeable influence on either the power of the test for linkage or the estimate of the QTL effect, *d*. It should be emphasized that the presence of dominance at the QTL does not affect the general utility of DNA pooling as a mean of testing marker-QTL linkage. The statistical test for linkage can be conducted as usual, and statistical significance remains an indication of linkage. Only the theoretical power and the estimation of gene effect will not be precise.

In the half-sib design: A single sire, assumed to be heterozygous at the QTL and the linked marker, is crossed with several different dams to produce a half-sib population. For purposes of analysis, it is assumed that the marker-QTL phase of the sire is *MQ/mq* and that the dams are in independent Hardy-Weinberg equilibrium for the QTL and the marker. Let *t* and $1 - t$ be the frequencies of QTL alleles *Q* and *q*, respectively, in the dam population. Since the marker locus in the dam population can have more than the two alleles, *M* and *m*, that are present in the sire, let *p*, *q* and *r* be the frequencies of marker alleles *M*, *m* and *m** in the dam population, where *m**, denotes any marker allele other than *M* or *m*. The expected frequencies of the various genotypes in the half-sib population are presented in Table 1.

For this case: (i) genotypic group *A* comprises all individuals that received the *M* marker allele from the sire; genotypic group *B*, all individuals that received the *m*

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

TABLE 1

Proportion of marker and QTL genotypes in a half-sib design

Marker genotype	QTL genotype		
	QQ	Qq	qq
MM	$pt/2$	$p(1-t)/2$	0
Mm	$qt/2$	$[tp + q(1-t)]/2$	$p(1-t)/2$
mm	0	$qt/2$	$q(1-t)/2$
Mm^*	$rt/2$	$r(1-t)/2$	0
mm^*	0	$rt/2$	$r(1-t)/2$

The expected proportion of marker and QTL genotypes in half-sib progeny of an heterozygous sire (genotype Mm), mated to randomly chosen dams. M and m denote the sire marker alleles; m^* denotes all other marker alleles found in the dams; p, q and r are the frequencies of M, m and m^* in the dam population; Q and q denote QTL alleles in the sire and in the dam population; t is the frequency of Q in the dam population.

TABLE 2

Adjustment of the design-dependent elements of the general theory to specific experimental designs

	Backcross	F_2	Half-sib
Genotypic group			
A	MM	MM	M (from sire)
B	Mm	mm	m (from sire)
Estimate of π			
Upper	$2M_U$	$2M_U - ½$	$M_U(2-r) - p$
Lower	$2m_L - 1$	$2m_L - ½$	$m_L(2-r) - p$
V_π	$4V_T$	$4V_T$	V_T to $5V_T$
δ	$(d + h)/2$	d	$\dfrac{\alpha(1 - r/2)}{4pq - 2r + 3}$

Genotypic groups, A and B; estimate of π from the frequencies of marker allele M in the upper and lower tail, M_U and M_L respectively; V_π in terms of V_T; and δ in terms of d and h (see text for details).

marker allele from the sire. (ii) In order to estimate the genotypic group frequency, π, the marker allele frequencies of the dams, p, q and r, must be known or estimated from a pooled DNA sample representative of the dams. Simple calculations then show that π can be estimated by $M_U(2 - r) - p$ in the upper tail, and by $m_L(2 - r) - q$ in the lower tail. (iii) It should be noted that in this design, experimental error variance, V_π, will range from equal to the technical error variance (when $r = 1$ and the allele frequencies of the dams are known), to as much as five times the technical error variance (when $r = 0$ and the marker allele frequencies of the dams are estimated through a DNA pooling procedure). (iv) In the half-sib design, the genotypic group effect, δ, represents a rather complex function (denoted δ^*) which is calculated from the expectation of the marker and QTL genotype frequencies given in Table 1, as:

$$\delta^* = \frac{\alpha(1 - r/2)}{4pq - 2r + 3} \qquad (11)$$

where $\alpha = d - (1 - 2t)h$ is the average effect of an allele substitution (FALCONER 1989). Expression (11) reaches a maximum value of $\alpha/2$, when neither of the marker alleles of the sire are present in the dams (*e.g.*, $p = q = 0$ and $r = 1$), and a minimum value of $\alpha/4$, when $p = q = ½$.

Table 2 presents a summary of the four elements needed for the adjustment of the general theory to the specific population designs.

NUMERICAL RESULTS

The optimal proportion selected at each tail in order to achieve maximum power, was obtained by numerically maximizing Equation 5 with respect to p for various values of NV_π (Figure 1). The results show that when the allelic frequencies in the pooled samples can be estimated without error ($V_\pi = 0$), the optimal total proportion selected over both tails is 0.48, 0.41 and 0.29, for genotypic group effects of $\delta = 0.125$, 0.25 and 0.5, respectively. Thus, the optimal proportion selected is lower for larger gene effects. This effect, however, becomes insignificant for values of NV_π greater than 10.

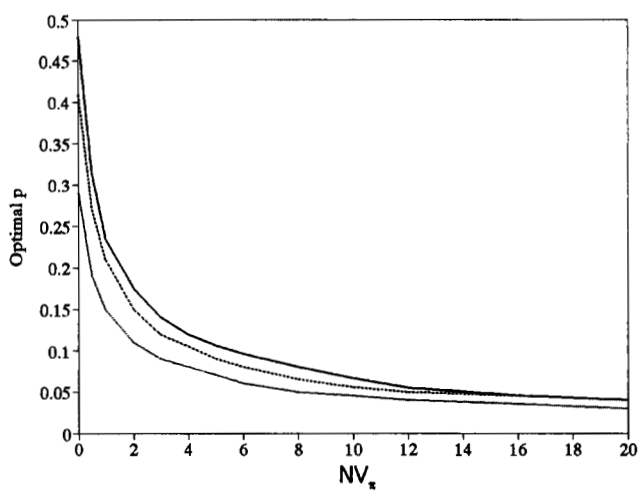

FIGURE 1.—Optimal proportion for selective DNA pooling. The optimal total proportion of the population, p, that should be taken in order to maximize experimental power for marker-QTL linkage determination is presented as a function of the product NV_π (N = population size, V_π = experimental error variance of the genotypic group frequency estimate). Proportions for three genotypic group effect, δ, are presented: solid line, $\delta = 0.125$; dashed line $\delta = 0.25$; dotted line $\delta = 0.5$. The two DNA pools are made up of the extreme $p/2$ of the population at each phenotypic tail, respectively.

The optimal proportion selected decreased exponentially as either N or V_π increased. The rationale for this decrease is as follows. When N increases, the error due to sampling decreases. Consequently, for given sampling error a smaller proportion can be taken at the extremes. This increases power by increasing the expected genotypic group frequency, π. The effect of increasing V_π on optimal proportion selected is more subtle, but can be explained by noting that reducing p will tend to increase both π (as above) and $\mathrm{Var}(\hat{\pi})$ (by increasing sample error), but the effect of these two changes on power is not the same (see Equations 4 and 5).

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

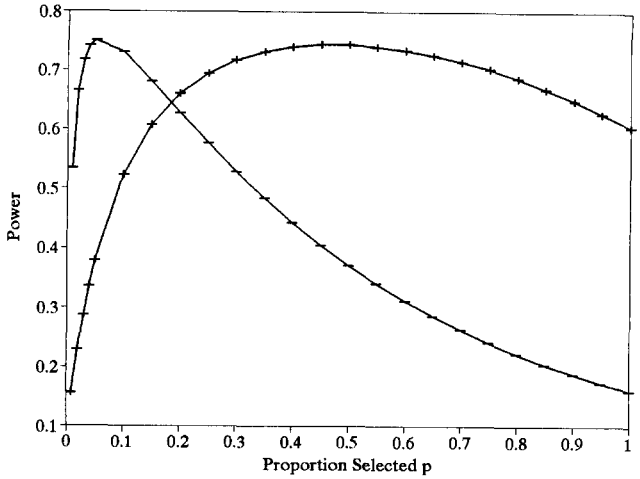

FIGURE 2.—Experimental power of marker-QTL linkage determination using selective DNA pooling as a function of the proportion selected, p. Two representative cases are presented. The first (dashes), with an optimal proportion of 0.06 (obtained with, $\delta = 0.25$ and $V_\pi = 0.02$), and the second (crosses), with an optimal proportion of 0.48 (obtained with, $\delta = 0.125$ and $V_\pi = 0$). In both cases $N = 500$ and type I error, $\alpha = 0.05$.

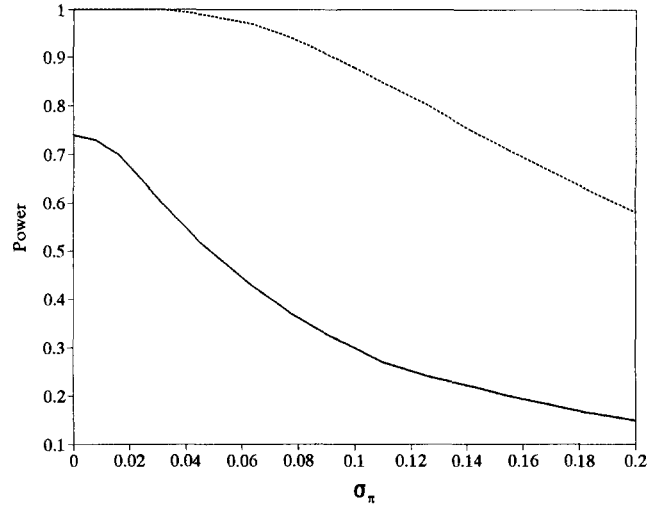

FIGURE 3.—Experimental power of marker-QTL linkage determination using selective DNA pooling. Results are presented for a trait with standardized genotypic group effect, $\delta = 0.125$ (solid line) and 0.25 (dashed line); population size, $N = 500$; and type I error, $\alpha = 0.05$, as a function of experimental error, σ_π. At each value of σ_π, the optimal proportion selected for that value was taken for the power calculation.

Figure 2 presents the power of marker-QTL linkage determination as a function of the proportion selected, p, for two representative cases. The first, with an optimal proportion of 0.06 (obtained with $\delta = 0.25$ and $V_\pi = 0.02$), and the second, with an optimal proportion of 0.48 (obtained with $\delta = 0.125$ and $V_\pi = 0$). In both cases $N = 500$ and type I error $\alpha = 0.05$. Values of δ were chosen to provide similar maximum power but otherwise do not affect the shape of the curves, which are determined primarily by V_π. It can be seen that when the optimal proportion is relatively high (*i.e.*, $p = 0.48$, in this example), the optimum is quite robust, and choosing a value close to the optimum will not have a significant influence on power. However, when the optimal proportion is very small (*i.e.*, $p = 0.06$, in this example), deviating from the optimal proportion can cause a marked reduction in power. Considering Figures 1 and 2, the following recommendations can be made, according to the value of NV_π, the product of experimental error variance and total experimental size. (i) For $NV_\pi > 10$, a proportion selected of $p = 0.1$ (0.05 at each tail) seems reasonable. Although, as presented in Figure 1, the optimum proportion might be even smaller than this, as shown in Figure 2 the effect on power of taking $p = 0.1$ instead of a smaller value will be slight. Furthermore, using a smaller proportion than 0.05 at each tail might introduce a disproportionate influence of individuals having extreme phenotypic values due to nongenetic factors or technical errors. (ii) For $2 < NV_\pi < 10$, a proportion selected of $p = 0.15$ (0.075 at each tail) is suggested. (iii) For $NV_\pi < 2$, the optimum proportion selected is more sensitive both to changes in NV_π and to genotypic group effect, δ. Nevertheless, at this range the optimal proportion selected is relatively high, and as shown in Figure 2 quite robust. Therefore the use of an intermediate value such as $p = 0.3$ will be appropriate for this case.

When an estimate of V_π is not available, choosing a proportion of $p = 0.2$ will not decrease power significantly, as shown in Figure 2, even if the optimal proportion is much lower or much higher than 0.2. Therefore, a proportion of $p = 0.2$ (0.1 at each tail) seems appropriate for an overall *a priori* proportion selected. On the other hand, if an accurate estimate of V_π is available and the range of the gene effect relevant to the experiment is known, the optimum proportion selected can be determined more exactly using Figure 1.

Figure 3 presents the power of detecting a QTL as a function of $\sigma_\pi = (V_\pi)^{1/2}$ for two representative cases having total population size, $N = 500$; type I error, $\alpha = 0.05$; and genotypic group effects $\delta = 0.125$ and 0.25, respectively. The values in this figure were calculated using Equation 5 with optimal p at each point. The decrease in power is presented as a function of σ_π rather than V_π, so as to be in direct proportion to experimental error. It can be seen that when power at $\sigma_\pi = 0$ is high (*i.e.*, when $\delta = 0.25$ in this example) power remains high even when the experimental error is fairly large. Indeed, for $\delta = 0.5$ (data not shown), the power remains practically 1.0 for the entire range presented in Figure 2. However, when initial power at $\sigma_\pi = 0$ is only moderate (*i.e.*, at $\delta = 0.125$ in this example) experimental error of even a small magnitude reduces experimental power substantially.

Table 3 compares the relative amount of genotyping and power for complete population genotyping, standard selective genotyping and selective DNA pooling, for

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

TABLE 3

Example of experimental power and number of genotypings required according to experimental procedure

Experimental procedure	Power	Genotypings (no.)
Complete genotyping	0.80[a]	500
Selective genotyping[b]	0.77[a]	250
Selective DNA pooling[c]		
$\sigma_\pi = 0.0$	0.74	2
$\sigma_\pi = 0.04$	0.50	2
$\sigma_\pi = 0.04$[d]	0.67	8

Presented for the case of total population size $N = 500$, genotypic group effect $\delta = 0.125$ and type I error $\alpha = 0.05$.

[a] Taken from DARVASI and SOLLER (1992).

[b] Twenty-five percent at each tail.

[c] σ_π, degree of experimental error standard deviation.

[d] Analyzed with four replications per marker.

the case $N = 500$, $\delta = 0.125$ and $\alpha = 0.05$. As compared to complete genotyping standard selective genotyping reduces the number of genotypings from 500 to 250, while decrease in power is only slight (from 0.80 to 0.77). It can be seen that when the allelic frequencies in the pooled DNA samples are estimated without error ($\sigma_\pi = 0$), selective DNA pooling provides a further very marked reduction in the number of genotypings per marker scored (from 250 to 2), while again decreasing power only slightly (from 0.77 to 0.74). Standard selective genotyping for this case will provide a power of 0.77, requiring 250 genotypings per marker. We parenthetically note that this means that practically all the power for QTL detection with selective genotyping, derives from the allelic frequency difference at the extremes. With experimental error $\sigma_\pi = 0.04$, however, power with selective DNA pooling drops to 0.5; but in this case power can be increased by replicating the genotyping so as to reduce the effective experimental error. For example, fourfold replication of allele frequency estimates in the above case reduces experimental error by a factor of two, increasing power to 0.67 while increasing the number of genotypings per marker from 2 to 8, still much less than the 250 required by standard selective genotyping.

Summary of procedure for carrying out the test for marker-QTL linkage: In carrying out a marker-QTL linkage analysis by way of selective DNA pooling the following procedure should be followed. In a preliminary step, an experimental design is chosen (backcross, F_2, half-sib) and an estimate of *technical error variance*, V_T, is obtained, either on the basis of previous experience or experimentally [*e.g.*, by genotyping DNA pools having known proportions of individuals of various marker genotypes for a representative set of the markers to be used in the analysis, see *e.g.*, KHATIB *et al.* (1994)]. Expected *experimental error variance*, V_π, is then estimated using the appropriate expression according to the experimental design, as given in this study, and taking into account the degree of replication to be implemented. Then total experimental size, N, will be determined as a function of genotyping replication so as to provide desired power for QTL effects of interesting magnitude, taking into consideration overall experimental costs. The optimum proportion of selection can now be determined.

The experimental population is produced and phenotyped for all traits of interest. For each trait, DNA pools are then constructed from the high and low tails of the population trait value, using the optimal proportion. Pooled DNA samples are genotyped, and marker allele frequencies in the upper and lower tails, M_U, m_U, M_L and m_L are estimated by appropriate densitometry. Using these estimates, the genotypic group frequency, π, is calculated according to the experimental design. The statistical test, as previously defined, is implemented in order to determine the statistical significance of marker-QTL linkage. The trait value averages at the upper and lower selected tails, $\bar{X}_U$ and $\bar{X}_L$, are then calculated in order to estimate the QTL gene effect.

DISCUSSION

The results of this study show that for given population size, and depending on experimental error variance, selective DNA pooling can provide statistical power that is only marginally reduced compared to that provided by complete genotyping or by standard selective genotyping. The great advantage of selective DNA pooling is in the marked reduction in the amount of genotyping required. In the case analyzed in the results section, for example, genotyping is reduced from 500 genotypings per marker for complete genotyping, or 250 genotypings for standard selective genotyping, to only two genotypings per marker for selective DNA pooling without replication, or to eight genotypings per marker when a fourfold replication of genotyping is carried out. That is, considered in the context of a 150 marker experiment, the number of genotypings is reduced from 75,000 to 1,200.

Thus, the DNA pooling strategy dramatically reduces genotyping requirements, and hence genotyping costs in marker-QTL linkage determination. This can make it possible to analyze extremely large populations and consequently detect QTL having moderate to low effects that were previously below experimental power. In addition, reduced genotyping requirements can allow cost effective utilization of marker-assisted selection in situations, such as elite sire marker-QTL evaluation, that were previously problematic (KASHI *et al.* 1990).

The most important limitation of selective DNA pooling as compared to complete genotyping or standard selective genotyping, is the strong dependence of experimental power on technical error variance (*i.e.*, the accuracy with which allelic frequencies are estimated in the DNA pools). Technical error variance has a number of components, including: difficulty in sampling exactly

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

equal amounts of DNA from each individual included in the pool; lack of quantitative accuracy of the molecular procedures to detect the marker allele in the pooled DNA; and errors in quantifying the primary experimental results as numerical values, even when using a densitometer or a computerized scanner. At every stage, however, the sources of experimental error appear to be unbiased. Consequently, by replicating any of the above procedures the magnitude of technical error variance can be controlled. Furthermore, as shown in Figure 3, QTL with large effects can be detected with little loss of power using selective DNA pooling even when technical error variance is relatively large. At the least, selective DNA pooling will be useful for initial screening of a population for marker-QTL linkage. In this case a relatively high type I error will be set and only those markers showing significant effects will be genotyped individually. It can also be expected that the great advantages of selective DNA pooling will encourage the development of accurate technical methods to estimate allelic frequencies in pooled DNA samples. This will make it possible to utilize selective DNA pooling to its maximal extent.

With selective DNA pooling, genotyping costs are influenced by total population size only to the extent that this increases the number of DNA samples that are collected and handled to make up the pools. Consequently, this mode of analysis is particularly useful when applied to existing large agricultural populations, which are reared and phenotyped for existing breeding or commercial purposes. A case in point in this regard are dairy cattle, where large half-sib families are common and genotyping can be based on somatic cells in individual milk samples routinely obtained for herd management purposes (Lipkin *et al.* 1993). By means of selective DNA pooling, these populations can be analyzed for marker-QTL linkage determination using a large number of markers, yet with relatively little effort.

A further advantage of selective DNA pooling, as compared to standard selective genotyping, is its suitability for multi-trait marker-QTL linkage analysis. In standard selective genotyping, all individuals with extreme values for a given trait are genotyped. Consequently, when simultaneously analyzing a number of traits, most of the individuals in a population will be classified as extreme for one or other of the traits, and almost the entire population will be genotyped (Darvasi and Soller 1992; Lebowitz *et al.* 1987). Thus, in this case selective genotyping will not be useful. In contrast, selective DNA pooling provides a marked reduction in overall genotyping requirements when several traits are analyzed, even though for each trait a different subset of the entire population is pooled. For example, extending the numerical example given in Table 3 to the case where 10 traits are analyzed, selective DNA pooling will still only require 20 genotypings per marker (or 80 when fourfold replications are carried out). This remains a major reduction as compared to the 500 genotypings required per marker with complete genotyping.

The main contribution of selective DNA pooling is in establishing marker-QTL linkage. This often is the first step in a program aimed at mapping QTL with respect to flanking markers on the chromosome. Such mapping will ordinarily require individual genotyping in order to identify recombinants. In some cases, however, selective DNA pooling may also provide a means for estimating QTL map location because the closer a marker is to a QTL, the greater the QTL effect associated with that marker. Thus, if a saturated genetic map is available and a relative small technical error can be achieved, a large number of markers can be scored in the region of the marker that detected the QTL, and the QTL map location will be estimated at the marker that showed the greatest associated quantitative effect. Since with selective DNA pooling, genotyping a large number of markers is relatively simple, this may provide an alternative to QTL mapping procedures based on flanking markers.

This research was supported by the joint biotechnology program of the Israel Ministry of Science and Technology and the German Bundesministerium fuer Forschung und Technologie.

LITERATURE CITED

Beckmann, J. S., and M. Soller, 1983 Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. Theor. Appl. Genet. **67:** 35–43.

Beckmann, J. S., and M. Soller, 1986 Restriction fragment length polymorphisms in plant genetic improvement. Oxf. Surv. Plant Mol. Cell Biol. **3:** 196–250.

Darvasi, A., and M. Soller, 1992 Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. **85:** 353–359.

Darvasi, A., A. Weinreb, 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.

Falconer, D. S., 1989 *Introduction to Quantitative Genetics*, Ed. 3. Longman, New York.

Hilbert, P., K. Lindpainter, J. S. Beckmann, T. Serikawa, F. Soubrier *et al.*, 1991 Chromosomal mapping of two genetic loci associated with blood-pressure regulation in hereditary hypertensive rats. Nature **353:** 521–529.

Hillel, J., R. Avner, C. Baxter-Jones, E. A. Dunnington, A. Cahaner *et al.*, 1990 DNA fingerprints from blood mixes in chickens and in turkeys. Anim. Biotechnol. **1:** 201–204.

Johnson, R. F., S. C. Pickett and D. L. Barker, 1990 Autoradiography using storage phosphor technology. Electrophoresis **11:** 355–360.

Kashi, Y., E. M. Hallerman and M. Soller, 1990 Marker-assisted selection of candidate bulls for progeny testing programs. Anim. Prod. **51:** 63–74.

Khatib, H., A. Darvasi, Y. Plotsky and M. Soller, 1994 Determining relative microsatellite allele frequencies in pooled DNA samples. PCR Methods Appl. **4:** 13–18.

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

Lebowitz, R. J., M. Soller and J. S. Beckmann, 1987 Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet. **73:** 556–562.

Lipkin, E., A. Shalom, H. Khatib, M. Soller and A. Friedmann, 1993 Milk as a source of deoxyribonucleic acid and substrate for the polymerase chain reaction. J. Dairy Sci. **76:** 2025–2032.

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024

MICHELMORE, R. W., I. PARAN and R. V. KESSELI, 1991 Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci. USA **88:** 9828–9832.

NEDELMAN, J., P. HEAGERTY and C. LAWRENCE, 1992 Quantitative PCR: Procedures and precisions. Bull. Math. Biol. **54:** 477–502.

NIENHUIS, J. T., T. HELENTJARIS, M. SLOCUM, B. RUGGERO and A. SCHAEFFER, 1987 Restriction fragment length polymorphism analysis of loci associated with insect resistance in tomato. Crop Sci. **27:** 797–803.

PACEK, P., A. SAJANTILA and A.-C. SYVANEN, 1993 Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. PCR Methods Appl. **2:** 313–317.

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN *et al.*, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335:** 721–726.

PLOTSKY, Y., A. CAHANER, A. HABERFELD, U. LAVI and J. HILLEL, 1990 Analysis of genetic association between DNA fingerprint bands and quantitative traits using DNA mixes. Proc. 4th World Congr. Genet. Appl. Livestock Prod. **13:** 133–136.

PLOTSKY, Y., A. CAHANER, A. HABERFELD, U. LAVI, S. J. LAMONT *et al.*, 1993 DNA fingerprint bands applied to linkage analysis with quantitative trait loci in chickens. Anim. Genet. **24:** 105–110.

SOLLER, M., and J. S. BECKMANN, 1982 Restriction fragment length polymorphisms and genetic improvement. Proc. 2nd World Congr. Genet. Appl. Livestock Prod. **6:** 396–404.

SOLLER, M., and J. S. BECKMANN, 1983 Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. **67:** 25–33.

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

SOLLER, M., A. GENIZI and T. BRODY, 1976 On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47:** 35–39.

WELLER, J. I., Y. KASHI and M. SOLLER, 1990 Daughter and granddaughter designs for mapping of quantitative trait loci in dairy cattle. J. Dairy Sci. **73:** 2525–2537.

Communicating editor: W. G. HILL

Downloaded from https://academic.oup.com/genetics/article/138/4/1365/6012663 by guest on 19 April 2024