Heredity **80** (1998) 101–109 *Received 25 February 1997*

Pooling DNA in the identification of parents

ROBERT N. CURNOW* & ANDREW P. MORRIS

Department of Applied Statistics, The University of Reading, PO Box 240, Earley Gate, Reading RG6 6FN, U.K.

The preliminary pooling of DNA samples to reduce the number of tests required to identify which of a set of potential parents could be the parent of an individual progeny plant is discussed. The progeny plant may be homozygous or heterozygous at the marker locus involved and the marker alleles of the other parent plant may or may not be known. The expected number of tests is derived for varying pool sizes when the DNA of each parent plant is in one and only one DNA pool. Optimal pool sizes are calculated for a range of number of potential parents and of marker allele frequencies. Designs with each parent in more than one pool and the sequential use of up to three independent marker loci are also discussed.

Keywords: experimental design, group screening, parentage, pooling DNA.

Introduction

Molecular markers are increasingly being used to decide which of a group of animals or plants could possibly be a parent of a particular individual animal or plant and which can definitely be excluded as a parent (Ellstrand, 1984; Tammisola *et al.*, 1994). Although equally applicable to animal populations, we shall, for ease of presentation, assume that we are dealing with a plant species which is outbreeding and diploid. In this paper we shall assume that only the male parent needs identification. We may or may not have information about the marker alleles of the female parent. A future paper will discuss the identification of both parents. Often the parentage of several different plants will be of interest and compromises will then be needed because, as we shall see, the best way to pool the DNA of the potential parents does depend on the genotype of the progeny plant.

Assuming no errors in the typing of the molecular markers and no germ-line mutations, a plant can be excluded as a possible parent if it possesses neither of the alleles of the progeny plant. Knowing the marker alleles of the female parent may provide further evidence leading to exclusion. All the plants to be tested will be referred to as potential parents, and all the plants that cannot be excluded on the basis of their marker alleles will be referred to as possible parents. We shall assume throughout that one of the potential parents is the real parent. This can clearly only be a reasonable assumption in controlled environments or with closed, isolated populations.

Determining the molecular markers at a locus for a large number of potential parents can be time-consuming and expensive. If there are many potential parents and the probability of exclusion on the basis of the marker alleles at the locus is high, then a preliminary screen based on pooling the DNA of groups of plants may substantially reduce the number of tests required. If the pooled DNA does not contain the required allele or alleles, then all the parents contributing to the pool can be excluded. If the pooled DNA does contain the required allele or alleles then we know that the pool contains at least one possible parent. We shall assume that a pool being positive, i.e. containing the required allele or alleles, does not provide any information about the number of possible parents in the pool but simply that there is at least one.

The pooling of the DNA is almost equivalent to the use of group screening for defective items described originally by Dorfman (1943) and Sterrett (1957), with more recent work by Balding *et al.* (1994) and Bruno *et al.* (1995) on the screening of clone libraries for rare 'positives'. Our problem differs in that we know that there is at least one possible parent, the real parent, among the potential parents.

By pooling the DNA, we hope to reduce the number of tests required to identify all the possible parents. Until the penultimate section, only two stages of testing will be allowed and so the parents in pools not excluded by the first stage of testing will be tested individually. More complicated schemes

*Correspondence: E-mail: r.n.curnow@reading.ac.uk

will often be too costly in time and organization. We shall attempt to minimize, by choice of the size of the pool, the expected number of tests required. However, cost may depend on the number of tests possible on each electrophoretic gel as well as on the total number of tests. This may be a consideration in choosing between similar schemes.

In the next section, we derive an expression for the expected number of tests required using designs, allocations of potential parents to pools, in which each parent is allocated to a single pool. The optimal pool size will be derived in terms of the number of potential parents and the probability, q, that a potential parent, not the real parent, does not have the required allele or alleles to be a possible parent. Then, the exclusion probability, q, will be derived as a function of the genotype of the progeny and the frequencies of the marker alleles in the population. When the progeny plant is heterozygous at the marker locus and the marker alleles of the female parent are not known, the expected number of tests has to be averaged over the probabilities of the marker type of the female parent. Following this the single-replicate designs will be compared with designs including each parent in two, three or four different pools. Finally the use of sequences of molecular markers will be considered.

Derivation of the expected number of tests

If the number of potential parents, N, can be factorized as $N = nk$, then we can form n pools of k parents each. The probability that a pool will contain at least one possible, i.e. nonexcludable, parent will be $(1-q^k)$, where q is the probability that an individual parent can be excluded. All parents in such a pool will be tested individually and so, recalling that all the parents in the pool containing the real parent are bound to be tested, the expected number of tests is N if $k = 1$, and $N/k + k + (N-k)(1-q^k)$ if $k \geq 2$. The expected number of tests per potential parent is therefore $E = 1$ if $k = 1$, and:

$$E = 1 + \frac{1}{k} - \left(1 - \frac{k}{N}\right) q^k, \quad \text{if } k \geq 2. \tag{1}$$

The values of E for these single-replicate designs for a range of values of N and q are given in Table 1. The possible larger pool sizes omitted from the table

Table 1 Expected number of tests per potential parent required in single-replicate designs for a range of values of number of potential parents, N, pool size, k, number of pools, n, and probability of exclusion, q

		q						
k	n	0.0	0.1	0.5	0.8	0.9	0.95	1.0
$N = 36$								
1	36	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	18	1.500	1.491	1.264	0.896	0.735	0.648	0.556
3	12	1.333	1.332	1.219	0.864	0.665	0.547	0.417
4	9	1.250	1.250	1.194	0.886	0.667	0.526	0.361
6	6	1.167	1.167	1.154	0.948	0.724	0.554	0.333
$N = 64$								
1	64	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	32	1.500	1.490	1.258	0.880	0.715	0.579	0.531
4	16	1.250	1.250	1.191	0.866	0.635	0.403	0.312
8	8	1.125	1.125	1.122	0.978	0.748	0.410	0.250
$N = 144$								
1	144	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	72	1.500	1.490	1.254	0.869	0.701	0.563	0.514
3	48	1.333	1.732	1.211	0.832	0.620	0.426	0.354
4	36	1.250	1.250	1.189	0.852	0.612	0.371	0.278
6	24	1.167	1.167	1.152	0.915	0.657	0.343	0.208
8	18	1.125	1.125	1.121	0.966	0.718	0.354	0.181
9	16	1.111	1.111	1.109	0.985	0.748	0.365	0.174
12	12	1.083	1.083	1.083	1.020	0.824	0.407	0.167

were never optimal. Clearly, for these values of N, the testing of each individual parent ($k = 1$) is to be preferred if the probability of exclusion, q, is less than $q = 0.5$. When q approaches 1, the expected number of tests per potential parent approaches

$$E = \frac{1}{k} + \frac{k}{N},$$

which is minimized by $n = k = \sqrt{N}$. Therefore, if q is likely to be much above $q = 0.5$, a reasonable choice for n and k is $n = k = \sqrt{N}$. There will be substantial savings in the number of tests if q is near $q = 1$ with the proportional saving increasing with the number of potential parents, N.

Calculating the probability of exclusion

The probability, q, that a particular plant can be excluded on the basis of an individual test depends on whether the progeny plant is homozygous or heterozygous at the marker locus, and also on the frequencies of the marker alleles in the population. Write $f_1, f_2, \ldots, f_m$, with

$$\sum_{i=1}^{m} f_i = 1,$$

for the frequencies of the m marker alleles M_1, $M_2, \ldots, M_m$ in the population.

Consider first a homozygous progeny with marker genotype M_1M_1. The only potential parents that can be excluded by their genotype at the marker locus are those with no M_1 allele and so $q = (1-f_1)^2$, whatever the marker genotype of the female parent. Equation 1 and Table 1 can therefore be used with $q = (1-f_1)^2$.

Consider now a heterozygous progeny, M_1M_2. All the other alleles, M_3, M_4, $\ldots$, M_m, can be classified as a single allele, $\overline{M}$, with frequency $\bar{f} = f_3 + f_4 + \ldots + f_m = 1 - f_1 - f_2$. The marker genotype of the female parent, which may be known, must be M_1M_1, M_1M_2, M_2M_2, $M_1\overline{M}$ or $M_2\overline{M}$. Bayes' Theorem can be applied to the genotypes of the female parent and the progeny plant to obtain the probability of the genotype of the female parent given the genotype of the progeny plant as:

$$\mathrm{P}(\text{Female Parent Genotype} \mid \text{Progeny Genotype})$$

$$= \frac{\mathrm{P}(\text{Progeny Genotype} \mid \text{Female Parent Genotype})}{\mathrm{P}(\text{Progeny Genotype})} \times \mathrm{P}(\text{Female Parent Genotype}).$$

Assuming random mating with respect to the marker locus, the probabilities of the five genotypes for the female parent are shown in the second column of Table 2. The third column shows the male parent genotypes that would be excluded if we knew the female genotype and the final column the probabilities of these genotypes in the population.

If the marker genotype of the female parent is known, then the last column of Table 2 can be used to provide the values of q for (1) and hence for Table 1.

If the marker genotype of the female parent is not known, then the expected number of tests (eqn 1) must be averaged over the distribution of q shown in Table 2. So (1) becomes:

$$E = 1 \qquad \text{if } k = 1$$

$$E = 1 + \frac{1}{k} - \left(1 - \frac{k}{N}\right)\mathrm{E}(q^k) \qquad \text{if } k \geq 2, \tag{2}$$

where

$$\mathrm{E}(q^k) = \tfrac{1}{2}(1-f_2)^{2k+1} + \tfrac{1}{2}(1-f_1)^{2k+1} + \tfrac{1}{2}(1-\bar{f})\bar{f}^{2k}.$$

Table 3 shows the expected number of tests for $N = 36$, 64 and 144 when the pool sizes are $k = \frac{1}{2}N^{\frac{1}{2}}$ and $N^{\frac{1}{2}}$ for some possible sets of values for the marker allele frequencies f_1, f_2 and $\bar{f}$. The only gains from pooling, $E < 1$, occur when both of the alleles of the progeny plant are rare, $f_1 = f_2 = 0.05$. In this case, a pool size of $k = \frac{1}{2}N^{\frac{1}{2}}$ is slightly more efficient than $k = N^{\frac{1}{2}}$.

Table 2 Probability calculations for a heterozygous progeny, M_1M_2

Female parent marker genotype	Frequency	Excluded male parent genotype	Probability exclusion, q
M_1M_1	$\frac{1}{2}f_1$	M_1M_1, $M_1\overline{M}$	$(1-f_2)^2$
M_1M_2	$\frac{1}{2}(f_1+f_2)$	—	$\bar{f}^2$
M_2M_2	$\frac{1}{2}f_2$	M_2M_2, $M_2\overline{M}$	$(1-f_1)^2$
$M_1\overline{M}$	$\frac{1}{2}\bar{f}$	M_1M_1, $M_1\overline{M}$	$(1-f_2)^2$
$M_2\overline{M}$	$\frac{1}{2}\bar{f}$	M_2M_2, $M_2\overline{M}$	$(1-f_1)^2$

Table 3 Expected number of tests per potential parent for a heterozygous progeny plant for varying number of potential parents N, pool sizes k, and marker allele frequencies $f_1, f_2, \bar{f}$. Single-replicate with $k = \frac{1}{2}N^{\frac{1}{2}}$ and $N^{\frac{1}{2}}$

				E	
N	f_1	f_2	$\bar{f}$	$k = \frac{1}{2}N^{\frac{1}{2}}$	$k = N^{\frac{1}{2}}$
36	0.10	0.90	0.00	1.114	1.061
36	0.05	0.05	0.90	0.669	0.727
64	0.10	0.90	0.00	1.068	1.052
64	0.05	0.05	0.90	0.639	0.751
144	0.10	0.90	0.00	1.045	1.050
144	0.05	0.05	0.90	0.661	0.825

Similar calculations are possible when more than one progeny plant is available but the results will then depend on the individual genotypes of all the progeny.

Pools with parents replicated

So far, only pool designs with each potential parent in one and only one pool have been considered. The calculation of the expected number of tests, E, for designs with replicates, is complicated by the need to allow for the possibility that in a nonexcludable pool all but one of the parents can be excluded on the basis of information from other pools. The probability of this occurring depends on high-order association patterns of parents in pools and can only be expressed algebraically if there is so much replication that there will almost certainly be more testing than with individual testing of the potential parents.

Ignoring this complication, we shall assume that each parent is replicated equally often and does not occur in the same pool with any other parent more than once. As before, N is the number of potential parents, k the pool size, and q the probability of exclusion of a parent on the basis of an individual test. The number of replicates of each parent will be r, so that the number of pools is:

$$n = Nr/k.$$

The real parent and all other nonexcludable parents will require further individual testing and the expected number of these further tests is:

$$1 + (N-1)(1-q). \quad (3)$$

Any of the excludable parents in a pool containing the real parent will be further tested if all of the other $(r-1)$ pools containing it include at least one possible parent. This has probability:

$$(1-q^{k-1})^{r-1}$$

and contributes an expected number of further tests:

$$qr(k-1)(1-q^{k-1})^{r-1}. \quad (4)$$

Similarly, the expected number of further tests for excludable parents not in a pool with the real parent, is:

$$q[N-1-r(k-1)](1-q^{k-1})^{r}. \quad (5)$$

Adding eqns 3, 4 and 5 and the first-stage testing of the n pools and dividing by N, gives the expected number of tests per potential parent as:

$$E = 1, \qquad k = 1$$

$$E = [n + 1 + (N-1)(1-q) + qr(k-1)(1-q^{k-1})^{r-1} + q(N-1-r(k-1))(1-q^{k-1})^{r}]/N \quad k \geq 2. \quad (6)$$

Setting $r = 1$ reproduces, as it must, formula 1. As before, the expected number of tests (eqn 6) will need to be averaged over the distribution of q in Table 2 if the progeny plant is heterozygous and the marker genotype of the female plant is not known.

Designs do not exist for all combinations of N, r and k. The simplest two-replicate designs, $(r = 2)$, occur when N is a perfect square. As a simple example, when $N = 9$ we can write the number of each potential parent in a square array:

1 2 3
4 5 6
7 8 9

and form six pools of three parents each by using the rows and columns of the square. Three-replicate designs can similarly be formed if N is a perfect cube. These designs with $k = N^{\frac{1}{2}}$ and $k = N^{\frac{1}{3}}$ are special cases of lattice designs (Cochran & Cox, 1957) and larger numbers of replicates can be obtained. In Table 4 all these designs are listed for $N = 36$, 64 and 144 with $r \leq 4$. With $r > 4$, the number of pools is greater than $4N^{\frac{1}{2}}$ or $4N^{\frac{2}{3}}$ and little would be gained over individual testing. Table 4 includes the one other design available for $N = 36$ which has $r = 2$, $k = 8$ (Bose *et al.*, 1954).

Table 4 shows the expected number of tests per potential parent for the various designs when the progeny plant is homozygous, $q = (1-f_1)^2$, or is heterozygous and the female parent's marker genotype is known so that q is one of the values in the final column of Table 2. Table 4 shows, as expected, that replication is only worthwhile compared with single-replicate designs in terms of reducing the number of further tests required if the probability of individual exclusion, q, is large, for example greater

Table 4 Approximate expected number of tests per potential parent for designs with replication compared with the 'best' single-replicate designs and designs with no pooling

				q						
N	*k*	*n*	*r*	0	0.1	0.5	0.8	0.9	0.95	1.0
36	1	36	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
36	6	6	1	1.167	1.167	1.154	0.948	0.724	0.554	0.333
36	6	12	2	1.333	1.333	1.308	0.956	0.666	0.503	0.361
36	8	9	2	1.250	1.250	1.244	1.010	0.700	0.488	0.278
36	6	18	3	1.500	1.500	1.462	1.008	0.722	0.603	0.528
36	6	24	4	1.667	1.667	1.617	1.092	0.836	0.750	0.694
64	1	64	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
64	8	8	1	1.125	1.125	1.122	0.978	0.748	0.544	0.250
64	8	16	2	1.250	1.250	1.243	0.983	0.654	0.444	0.266
64	4	48	3	1.750	1.750	1.594	1.068	0.888	0.820	0.766
64	8	24	3	1.375	1.375	1.365	1.011	0.653	0.485	0.391
64	8	32	4	1.500	1.500	1.486	1.056	0.706	0.581	0.516
144	1	144	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
144	12	12	1	1.083	1.083	1.083	1.020	0.824	0.588	0.167
144	12	24	2	1.167	1.167	1.166	1.046	0.723	0.434	0.174
144	12	36	3	1.250	1.250	1.249	1.076	0.676	0.405	0.257
144	12	48	4	1.333	1.333	1.332	1.110	0.666	0.436	0.340

than 0.8. Even then the savings are only appreciable, about 20 per cent, with the largest value of N studied, when three- or four-replicate designs are best. The designs with replicates may have an advantage in that they may provide some check on the occurrence of errors in classifying pools as excludable or not.

Table 5 gives a few examples of the expected number of tests when the progeny plant is heterozygous and the marker type of the female parent is unknown. Pooling is still preferable to no pooling, $E<1$, only when both alleles in the heterozygous progeny are rare, $f_1=f_2=0.05$. Additional replication does not reduce the number of tests per parent by a worthwhile amount.

Some checks can be made on the importance of ignoring the possibility of identifying a possible parent because all other parents in a pool containing it have been excluded on the basis of evidence from other pools. If the only possible parent is the real parent, then, for a single-replicate design, the expected number of tests per potential parent is:

$$E = 1/k + k/N \tag{7}$$

and for a two-replicate design ($k=N^{\frac{1}{2}}$, $n=2N^{\frac{1}{2}}$), the true parent will be identified without further testing and so:

$$E = 2/k. \tag{8}$$

If there is just one possible parent in addition to the real parent, then, with a single-replicate design, the probability that the two possible parents are in the same pool is $(k-1)/(N-1)$ leading to:

Table 5 Expected number of tests per potential parent for a heterozygous progeny plant for varying number of potential parents, N, pool sizes, k, and marker allele frequencies $f_1, f_2, \bar{f}$. Female parent marker alleles unknown. Varying number of replicates, r, with $k=N^{\frac{1}{2}}$

N	*k*	*r*	*n*	f_1	f_2	$\bar{f}$	*E*
36	6	1	6	0.10	0.90	0.00	1.061
36	6	1	6	0.05	0.05	0.90	0.727
36	6	2	12	0.10	0.90	0.00	1.152
36	6	2	12	0.05	0.05	0.90	0.671
36	6	4	24	0.10	0.90	0.00	1.396
36	6	4	24	0.05	0.05	0.90	0.843
144	12	1	12	0.10	0.90	0.00	1.050
144	12	1	12	0.05	0.05	0.90	0.825
144	12	2	24	0.10	0.90	0.00	1.104
144	12	2	24	0.05	0.05	0.90	0.727
144	12	4	48	0.10	0.90	0.00	1.219
144	12	4	48	0.05	0.05	0.90	0.674

$$E = 1/k + (2N-k-1)\,k/[N(N-1)]. \quad (9)$$

With a two-replicate design four individual tests are needed if the two possible parents are never in the same pool, leading to:

$$E = 2/k + 4(N-2k+1)/[N(N-1)]. \quad (10)$$

Equations (7) and (9) show that, at least for $N \leq 100$, E is generally minimized for single-replicate designs by taking $n = k = N^{\frac{1}{2}}$. The only exception is when there is an additional possible parent and $N = 36$; then $n = 9$, $k = 4$ is slightly preferable to $n = k = 6$, with $E = 16.7$ compared with $E = 17.1$. With $k = N^{\frac{1}{2}}$, single- and two-replicate designs have the same E-value when the only possible parent is the real parent.

Table 6 shows the values of E from eqns (7), (9) and (10) for a range of values of N with $k = N^{\frac{1}{2}}$. The two-replicate designs are more efficient than the single-replicate designs when there is a possible parent in addition to the real parent. The advantage increases to 20 per cent when $N = 100$.

Using several molecular markers

An alternative to replication is to test those pools not excluded using one molecular marker by using a second molecular marker, and those pools not excluded by the second marker by a test using a third marker, and so on. We shall assume a single-replicate design for the N potential parents with n pools of k plants each, $N = nk$. We shall assume that there are no correlations in the occurrence of particular alleles at different molecular marker loci. As potential parents are eliminated, the optimal pool size k will change. The resulting changes in the number of tests required would probably not be sufficient to compensate for the extra work involved in reconstructing the pools. The pools will therefore be kept intact.

Table 6 Expected number of tests per potential parent, E, one- and two-replicate designs with $k = N^{\frac{1}{2}}$

		Real parent + one possible parent	
N	Real parent only $r = 1, 2$	$r = 1$	$r = 2$
4	1.00	1.33	1.33
9	0.67	0.91	0.89
16	0.50	0.70	0.65
25	0.40	0.57	0.51
36	0.33	0.48	0.41
49	0.29	0.41	0.35
64	0.25	0.36	0.30
81	0.22	0.32	0.26
100	0.20	0.29	0.23
144	0.17	0.24	0.19

Writing P_i as the probability that a particular pool not containing the real parent would be excluded by the ith molecular marker, the probability that the pool would be excluded by at least one of the first l markers is (Feller, 1968):

$$P_{[l]} = P_1 + P_2 + \ldots + P_l - (P_1P_2 + P_1P_3 + \ldots + P_{l-1}P_l) + \ldots + (-1)^{l-1}P_1P_2\ldots P_l. \quad (11)$$

The expected number of tests of the pool using up to M molecular markers will be:

$$1 + (1-P_{[1]}) + (1-P_{[2]}) + \ldots + (1-P_{[M-1]}) = M - [P_{[1]} + P_{[2]} + \ldots + P_{[M-1]}).$$

If all individuals in the pools not excluded by the M markers are tested individually using all M markers, the expected total number of tests for the $(n-1)$ pools not containing the true parent is:

$$(n-1)[M - \{P_{[1]} + P_{[2]} + \ldots + P_{[M-1]}\} + kM(1-P_{[M]})].$$

The pool containing the true parent will require $(M + kM)$ tests. Thus the expected total number of tests per potential parent is:

$$[Mn(1+k) - (n-1)[P_{[1]} + P_{[2]} + \ldots + P_{[M-1]}] - (n-1)kMP_{[M]}]/N. \quad (12)$$

With no pooling, $k = 1$ and the expected total number of tests per potential parent is:

$$\{M + (N-1)[1 - P_{[1]} - P_{[2]} - P_{[M-1]}]\}/N.$$

The expected number of possible parents at the conclusion of the testing is:

$$1 + (N-1)\prod_{i=1}^{M}(1-q_i), \quad (13)$$

where q_i is the probability of a potential parent being excluded on the basis of alleles at the marker i locus.

If the ith marker locus in the progeny plant is homozygous, P_i in (11) is:

$$P_i = q_i^k = (1-f_{i1})^{2k},$$

where f_{i1} and, later, f_{i2} and $\overline{f}_i$ now refer to the frequencies of the marker alleles M_1, M_2 and $\overline{M}$ at the ith marker locus. If the ith marker locus in the progeny plant is heterozygous:

$$P_i = \mathrm{E}(q_i^k),$$

where as in eqn (2):

$$E(q_i^k) = \tfrac{1}{2}(1-f_{i2})^{2k+1} + \tfrac{1}{2}(1-f_{i1})^{2k+1} + \tfrac{1}{2}(1-\bar{f}_i)\bar{f}_i^{2k}$$

if the marker genotype of the female parent is unknown. The value of q_i can be taken from the last column of Table 2 if the maternal marker type is known.

Table 7 shows the performance of a three-locus system with 64 possible parents for a variety of frequencies for the alleles of the progeny plant; for example, 0.05 represents a homozygote with allele frequency 0.05 and 0.05/0.2 a heterozygote with allele frequencies 0.05 and 0.2. The genotypes of the maternal parent are assumed not known. The order of testing and k, the size of the pool, have been chosen to minimize the expected number of tests. The fifth column shows the expected number of tests per parent; the sixth column the expected number of tests with no pooling, $k = 1$, and the seventh column the expected number of possible parents after the testing. Table 8 shows comparable values for single-locus and two-loci testing.

Table 7 shows that pooling has considerable advantages in terms of number of tests required unless the allele frequencies are relatively high. The overall, and unsurprising, advantage of rare alleles is clear in terms of both the expected number of tests and the expected number of possible parents

Table 7 Performance of three-marker loci. Test loci in optimal order and pool size optimal, $N = 64$

Progeny allele frequencies			Optimal pool size, k	E (optimal k)	E ($k = 1$)	Expected remaining possible parents
Locus 1	Locus 2	Locus 3				
0.05	0.05	0.05	4	0.907	2.025	1.058
0.05	0.05	0.05/0.05	4	0.911	2.025	1.061
0.05	0.05/0.05	0.05/0.05	4	0.915	2.025	1.064
0.05/0.05	0.05/0.05	0.05/0.05	4	0.922	2.026	1.067
0.05	0.05	0.2/0.05	4	0.992	2.025	1.147
0.05	0.05/0.05	0.2/0.05	4	1.000	2.025	1.154
0.05/0.05	0.05/0.05	0.2/0.05	4	1.010	2.026	1.161
0.05	0.05	0.2	4	1.065	2.025	1.216
0.05	0.05	0.2/0.2	4	1.074	2.025	1.249
0.05	0.05/0.05	0.2	4	1.075	2.025	1.226
0.05	0.05/0.05	0.2/0.2	4	1.085	2.025	1.261
0.05/0.05	0.05/0.05	0.2	4	1.087	2.026	1.237
0.05/0.05	0.05/0.05	0.2/0.2	4	1.097	2.026	1.273
0.05	0.2/0.05	0.2/0.05	4	1.160	2.039	1.369
0.05/0.05	0.2/0.05	0.2/0.05	4	1.176	2.040	1.386
0.05	0.2/0.05	0.2	4	1.291	2.039	1.542
0.05	0.2/0.05	0.2/0.2	4	1.308	2.039	1.626
0.05/0.05	0.2/0.05	0.2	4	1.310	2.040	1.567
0.05/0.05	0.2/0.05	0.2/0.2	4	1.328	2.040	1.656
0.05	0.2	0.2	2	1.430	2.050	1.796
0.05/0.05	0.2	0.2	2	1.444	2.052	1.834
0.05	0.2	0.2/0.2	2	1.448	2.020	1.920
0.05/0.05	0.2	0.2/0.2	2	1.462	2.052	1.964
0.05	0.2/0.2	0.2/0.2	2	1.472	2.056	2.063
0.05/0.05	0.2/0.2	0.2/0.2	2	1.487	2.057	2.113
0.2/0.05	0.2/0.05	0.2/0.05	4	1.502	2.075	1.926
0.2/0.05	0.2/0.05	0.2	2	1.644	2.075	2.361
0.2/0.05	0.2/0.05	0.2/0.2	2	1.671	2.075	2.573
0.2/0.05	0.2	0.2	2	1.792	2.102	3.000
0.2/0.05	0.2	0.2/0.2	2	1.831	2.102	3.312
0.2/0.05	0.2/0.2	0.2/0.2	2	1.881	2.116	3.671
0.2	0.2	0.2	2	2.062	2.143	3.939
0.2	0.2	0.2/0.2	2	2.119	2.143	4.397
0.2	0.2/0.2	0.2/0.2	1	2.163	2.163	4.924
0.2/0.2	0.2/0.2	0.2/0.2	1	2.186	2.186	5.535

Table 8 Performance of one- and two-marker loci. Loci in optimal order and pool size optimal, $N = 64$

Progeny allele frequencies		Optimal pool size, k	E (optimal k)	E ($k = 1$)	Expected remaining possible parents
Locus 1	Locus 2				
0.05	—	4	0.628	1.000	7.143
0.05/0.05	—	4	0.639	1.000	7.434
0.2/0.05	—	4	0.880	1.000	16.435
0.2	—	1	1.000	1.000	23.680
0.2/0.2	—	1	1.000	1.000	27.208
0.05	0.05	4	0.682	1.112	1.599
0.05	0.05/0.05	4	0.689	1.112	1.627
0.05/0.05	0.05/0.05	4	0.700	1.116	1.657
0.05	0.2/0.05	2	0.815	1.112	2.505
0.05/0.05	0.2/0.05	2	0.825	1.116	2.576
0.05	0.2	2	0.880	1.112	3.211
0.05/0.05	0.2	2	0.893	1.116	3.316
0.05	0.2/0.2	2	0.900	1.112	3.555
0.05/0.05	0.2/0.2	2	0.914	1.116	3.677
0.2/0.05	0.2/0.05	2	1.102	1.257	4.782
0.2/0.05	0.2	2	1.245	1.257	4.557
0.2/0.05	0.2/0.2	1	1.257	1.257	7.421
0.2	0.2	1	1.370	1.370	9.165
0.2	0.2/0.2	1	1.370	1.370	10.435
0.2/0.2	0.2/0.2	1	1.425	1.425	11.903

remaining after the tests. The latter is independent of the number of parents in each pool.

Table 8 shows the same general features noted for Table 7. Comparing the results of Tables 7 and 8, the two-locus tests incur more tests per parent than a single-locus test but at considerable savings in the number of possible parents remaining at the end of the tests. The same is true for the comparison of three-loci tests in place of two. The loci for which the progeny plant is homozygous and the loci for which the progeny plant has the rarest allele should be tested before the other loci. The optimal size of the pool for a single locus is four for rare alleles and one for the commoner alleles. For two loci, the optimal pool size can also be two for intermediate or mixed-allele frequencies. The optimal size of pool with three loci is often four, and only one when all the alleles possessed by the progeny plant are relatively common.

Discussion

There are clear advantages in pooling the DNA of the potential parents if the number of such parents is large and the alleles found in the progeny are rare. A good rule, whether the parent of a single progeny plant or animal or the parents of several different progeny are sought, is to choose a pool size close to $\frac{1}{2}N^{\frac{1}{2}}$, where N is the number of potential parents. There are considerable advantages in the sequential use of different markers in terms of reducing the number of possible parents remaining at the conclusion of the tests. Unless the number of potential parents is very large, there is little advantage in including each potential parent in more than one pool.

Acknowledgements

We are grateful to the referees for helpful comments on an earlier version of this paper and to the EPSRC for an Advanced Course Studentship (A.P.M.).

References

BALDING, D. J., BRUNO, W. J., KNILL, E. AND TORNEY, D. C. 1994. A comparative survey of non-adaptive pooling designs. In: Speed, T. and Waterman, M. S. (eds) *Proceedings of the Mathematics and Molecular Biology Meeting of the Institute for Mathematics and its Applications, University of Minnesota*, pp. 133–154. *July 1994*, Springer Verlag, New York.

BOSE, R. C., CLATWORTHY, W. H. AND SHRIKHANDE, S. S. 1954. Tables of Partially Balanced Designs with Two

Associate Classes. *North Carolina Agr. Exp. Sta. Tech. Bull*., **107**.

BRUNO, W. J., KNILL, E., BALDING, D. J., BRUCE, D. C., DOGGETT, N. A., SAWHILL, W. W. *ET AL*. 1995. Efficient pooling designs for library screening. *Genomics*, **26**, 21–30.

COCHRAN, W. G. AND COX, G. M. 1957. *Experimental Designs*, 2nd edn. John Wiley & Sons, New York.

DORFMAN, R. 1943. The detection of defective members of large populations. *Ann. Math. Statist.*, **14**, 436–440.

ELLSTRAND, N. C. 1984. Multiple paternity within the fruits of the wild radish, *Raphanus sativus*. *Am. Nat.*, **123**, 819–828.

FELLER, W. 1968. *An Introduction to Probability Theory and its Applications*, vol. 1, 3rd edn. John Wiley & Sons, New York.

STERRETT, A. 1957. On the detection of defective members of large populations. *Ann. Math. Statist.*, **28**, 1033–1036.

TAMMISOLA, J., AKERMAN, R. M., LAPINJOKI, S. AND KAUPPINEN, V. 1994. Strategies of pooling for parentage analyses applying DNA markers. In: Van Oijen, J. W. and Jansen, J. (eds) *Biometrics in Plant Breeding: Applications of Molecular Markers*, pp. 186–194. Proc. 9th meeting EUCARPIA Section. *Biometrics in Plant Breeding*, July 1994. CPRO-DLO, Wageningen, The Netherlands.