Theor Appl Genet (1993) 87:233–237

THEORETICAL AND APPLIED GENETICS

Use of near-isogenic lines derived by backcrossing or selfing to map qualitative traits

S. M. Kaeppler*, R. L. Phillips, T. S. Kim **

Department of Agronomy and Plant Genetics and the Plant Molecular Genetics Institute, University of Minnesota, St. Paul, MN 55108, USA

Received: 27 August 1992 / Accepted: 11 February 1993

Abstract. Near-isogenic lines (NILs) are a valuable resource for detecting linkages between qualitative trait loci and molecular markers. Molecular marker studies are expensive and methods that require genotyping fewer individuals, such as the NIL-analysis method, are desirable. We present a theory for using sets of NILs to detect linkages between molecular markers and introgressed loci. The probability that a marker a specific distance from the introgressed gene will have a donor parent allele in a near-isogenic line is a function of the distance between the marker and the gene, and the number of backcrosses and/or selfs used in deriving the NIL. The binomial probability formula is used to calculate the probability of having a donor parent allele at a given marker when sets of NILs are used. The formulae given allow calculation of the probability that a marker is linked to the introgressed gene, as well as the probability that a gene will be successfully detected when using given numbers of NILs, backcrosses, and molecular markers.

Key words: Molecular markers – RFLPs – Introgression

Introduction

The genetic mapping of qualitative trait loci using segregating populations and conventional or molecular markers requires the analysis of large numbers of individuals e.g., 50–100. Methods that require analyzing fewer individuals would be beneficial in terms of time and resources. A procedure is available which in certain circumstances could contribute greatly to the efficiency of mapping. This procedure makes use of existing near-isogenic lines (NILs) produced by backcrossing or selfing (Allard 1960; Fehr 1987) to locate introgressed genes. Genetic mapping can potentially be carried out with as few as three individuals – the recurrent parent, the donor parent, and the NIL, although the power of the method is enhanced when multiple NILs are used.

Muehlbauer and coworkers (1988) provide a theory for using near-isogenic lines produced by backcrossing to facilitate the integration of conventional and molecular marker maps. They suggest that NILs can be used to identify putative linkages of molecular and conventional genetic markers. Segregating populations would then be used to confirm the linkages and to calculate map distances. This theory requires assumptions about both genome size and the size and number of individual chromosomes, and is affected by the location of the gene on a chromosome. It is most correctly applied when the location of the molecular markers is random and not evenly spaced throughout the genome. This approach has been shown to be useful by several researchers (Muehlbauer et al. 1989, 1991; Paterson et al. 1990).

We present an alternative theory for the use of NILs in mapping that does not require any assumptions of genome size and is not affected by the position of a gene on a chromosome. We demonstrate how this theory can be applied to sets of independently derived NILs containing the same introgressed gene. Finally, we show by example how to determine in which situations (e.g., number of backcrosses, number of NILs, number of markers, etc.) this method could be effectively used to determine the map position of the gene of interest.

Communicated by F. Salamini

* *Present address:* Department of Agronomy, University of Nebraska Lincoln, Lincoln, NE 68503, USA

** *Present address:* Molecular Genetics Division, Agricultural Biotechnology Institute, Rural Development Administration, Suweon, 441-707 South Korea

Correspondence to: R. L. Phillips

Theory

The theoretical probability of having a donor parent (DP) allele at marker positions linked ($r<0.5$) or unlinked ($r=0.5$) to the introgressed gene is a function of the genetic distance (r) and the number of backcrosses or selfs used in producing the NIL (b or s, respectively). We provide formulae to calculate the probability (d) of having a DP allele at marker positions any genetic distance from the introgressed gene for NILs produced by three different methods (Table 1). Formulae listed under Case 1 correspond to NILs produced using b recurrent backcrosses with (1b) or without (1a) selfing following the last backcross. Selfing after each backcross with selection for a homozygous genotype, a procedure often used when introgressing recessive alleles, does not change the expected probabilities as given in Case 1. The formula in Case 2 gives the probability that a marker any genetic distance from the gene of interest would have an allelic contrast in NILs produced by selfing (Allard 1960).

The formulae in Table 1 are presented in terms of the recombination fraction, r. Measurements in centimorgans (cM) must be converted to r for use of these formulae. This can be done using equations such as those presented by Haldane (1919) or Kosambi (1944).

Extension of theory to multiple independently derived backcross lines having common parents

The power of this method is enhanced when multiple independently derived NILs are analyzed. Independently derived lines are lines that have been handled separately throughout the entire NIL derivation. The binomial probability distribution is used to calculate the probability that a certain marker allele will be found in a given number of independently derived NILs by chance ($r=0.5$), or because of linkage ($r<0.5$).

For example, the probability (D) that the donor parent allele for a marker will be present in a given number of independently derived NILs which were derived as in Case 1b, Table 1 is:

$$D = p(\text{DP allele in a given number of NILs}) = \frac{T!}{X!\,Y!}(d)^x(1-d)^y$$

where X = number of NILs containing a donor parent allele for a given marker, Y = number of NILs containing a recurrent parent allele for a given marker, T = total number of near-isogenic lines in set (X + Y), d = p (donor parent allele at marker) $=(1-r)^{b+1}$.

Substitution of the appropriate probability formula for d (Table 1) allows probability calculations for sets of NILs produced by backcrossing and/or selfing.

Table 1. Probability of a donor parent allele (Case 1) or divergent alleles (Case 2) at a marker a given distance from the introgressed gene for NILs derived by two different schemes

Case	Method of NIL derivation	Probability of DP allele at marker (d)[a]
1a	Backcrossing only	$(1-r)^b$
b	Backcrossing followed by selfing to homozygosity	$(1-r)^{b+1}$
2	Selfing	$(1-r)^{s+1}$

[a] r, Recombination fraction; b, number of backcrosses; s, number of selfs

Application of theory

The formulae described above can be used to analyze data from a molecular marker analysis of NILs. The probability that a given situation (e.g. three out of three NILs have the DP allele for a marker) would occur by chance ($r=0.5$) is calculated. If this probability is less than a chosen significance level, it is determined that the situation has occurred because of linkage and not by chance.

Consider, for example, the situation where a set of three NILs produced after three backcrosses followed by selfing to homozygosity is being analyzed. The following probabilities can be calculated for each marker used, with the restriction $r=0.5$.

$$p(0 \text{ out of } 3 \text{ have DP allele}) = \frac{3!}{0!3!}((0.5)^4)^0((1-(0.5)^4)^3 = 0.824,$$

$$p(1 \text{ out of } 3 \text{ have DP allele}) = 0.1647,$$

$$p(2 \text{ out of } 3 \text{ have DP allele}) = 0.0109,$$

$$p(3 \text{ out of } 3 \text{ have DP allele}) = 0.00024.$$

Assume for a given probe that three out of three of the NILs have the DP allele. The probability that this has occurred by chance is 0.00024. If our chosen significance level is 0.00052 (corresponds to an approximate overall experiment error rate (α_e) of 0.05 when 100 markers are used) we would conclude that this probe is likely linked to the introgressed gene because 0.00024 is less than 0.00052.

The choice of a significance level is very important. Multiple tests on the same data set increase the probability that in one or more of the tests there will be a mistaken conclusion. We suggest that an experiment error rate (α_e) be chosen and that the error rate for each individual test (α_m) be calculated as follows (Weir 1990):

$$\alpha_m = 1 - 10^{[(\log(1-\alpha e))/n]}$$

Table 2. Individual test Type 1 error rates (α_m) which approximate the stated overall Type 1 error rate (α_e) for 6 situations

Markers	Maximum marker-gene distance for 20 M genome	α_e [a]	
		0.05	0.10
		α_m [b]	α_m
40	25 cM	0.013	0.0026
100	10 cM	0.00051	0.0011
200	5 cM	0.00026	0.00053

[a] α_e, Experiment error rate
[b] α_m, Individual test error rate

where

α_m = Type 1 error rate for each individual test
α_e = overall Type 1 error rate
n = number of markers used

This approximation approaches the desired overall experiment-wise rate. Table 2 shows the individual test significance level (α_m) appropriate for two experiment-wise error rates (α_e). Note that this approximation is based only on the number of markers used and the desired overall Type 1 error rate.

Results and Discussion

The preceding discussion outlines how this method can be used to determine if data indicate linkage between a marker and an introgressed gene. We now employ this theory to give examples of situations (e.g., number of NILs, markers, backcrosses) for which this approach will be successful. When a certain number of markers is used to analyze a genome of known size and the markers can be considered to be evenly distributed, a maximum distance between a marker and a gene (i.e., a maximum value for r) can be determined. Using this value we can apply this method to calculate a minimum probability of successfully detecting a linkage. Table 3 gives examples of some of these calculations. Information is provided for 40, 100, and 200 markers; 1, 3, 5, 7, 9, and 20 backcrosses (BC), and 1–10 backcross-derived lines (BDLs). These data are calculated using the formula from Table 1, Case 1 b, and the binomial probability formula.

The columns labelled * in Table 3 give the number of lines out of the NIL set that must contain a DP allele at a given marker for that marker to have a significant probability of being linked to the introgressed gene. In each case the probability that a given situation (e.g., 2 out of 10 lines produced by 3 backcrosses have the DP allele) would occur by chance (r = 0.5) was calculated and compared to a value corresponding to a 0.05 experiment-wise Type 1 error rate. The number given is the number of NILs out of the total for which the probability of occurring by chance was less than our chosen error rate; ns indicates that it was not possible to obtain a significant probability value for this combination using $\alpha_e = 0.05$. The minimum number of NILs that must contain the DP allele depends upon the number of markers, the number of backcrosses, and the chosen α_e.

The columns labelled *P* give the probability that the minimum number of lines out of a NIL set (corresponds to value in adjacent* column) actually would contain a DP allele for a marker the maximum distance from a gene. This can be considered the probability of successfully locating the introgressed gene using the given error rates and is an appropriate statistic to use in determining if given materials are appropriate for analysis by this method. These calculations were made assuming a 2,000 cM genome (approximately the size for maize) and equally spaced markers. For 40, 100, and 200 markers the maximum distance between a marker and the introgressed gene would be 25, 10, and 5 cM, respectively. The value used for α_m was 0.0013, 0.00052, and 0.00026 for these numbers of markers, respectively.

The data indicate that increasing the number of markers will increase the probability of detecting a linkage between a marker and a gene. When 40 markers are used the highest probability of detecting a linkage between a marker and a gene with less than 0.05 chance of error is 0.52 when using a 10 NIL set produced by 7 backcrosses. When 200 markers are used this probability is 0.95 or greater in 21 of the 60 situations depicted.

When mapping a gene in a segregating population, it would usually not be necessary to have a 10 cM interval marker saturation. In fact, a gene could be effectively mapped with markers distributed at 50-cM intervals. This would be 40 markers for the 2,000 cM genome example. Therefore, when mapping a specific gene, an analysis of a segregating population would be the most efficient approach since fewer markers would need to be scored.

There can be, however, a gain in efficiency due to scale. When a number of different genes need to be mapped, and NIL sets are available for each gene, the NIL analysis method would be more efficient. For each gene one would genotype the donor parent, the recurrent parent, and as few as 1 NIL. A number of such sets could be analyzed simultaneously with the same effort required to genotype the 75–100 progeny usually used in the analysis of a population segregating for a single or small number of genes. In our gel system, for example, once the DNA is extracted it is nearly as efficient to genotype 144 individuals as it is to genotype a single individual. These 144 individuals could be comprised of two F_2 populations of 72 individuals each or 48 NIL sets.

In most cases where this method might be used, the NILs will have been produced for another purpose. The

Table 3. Number of NILs with DP allele required for significance (*) and minimum probability of successfully detecting a linkage between a marker and an introgressed gene (*P*) for a number of situations

Number of markers	Number of BDLs	1BC		3BC		5BC		7BC		9BC		20BC	
		*	*P*	*	*P*	*	*P*	*	*P*	*	*P*	*	*P*
40	1	ns	–	ns	–	ns	–	ns	–	1	0.11	1	<0.01
	2	ns	–	ns	–	2	0.07	2	0.03	2	0.01	1	<0.01
	3	ns	–	3	0.07	2	0.17	2	0.08	2	0.03	1	0.03
	4	ns	–	3	0.19	3	0.06	2	0.13	2	0.06	1	0.04
	5	5	0.11	4	0.09	3	0.12	2	0.20	2	0.09	1	0.05
	6	6	0.07	4	0.19	3	0.19	2	0.27	2	0.13	1	0.05
	7	6	0.22	4	0.31	3	0.27	2	0.33	2	0.17	1	0.06
	8	7	0.15	4	0.43	3	0.35	2	0.40	2	0.21	1	0.07
	9	7	0.31	5	0.29	3	0.43	2	0.46	2	0.25	1	0.08
	10	8	0.24	5	0.39	3	0.51	2	0.52	2	0.29	1	0.09
100	1	ns	–	ns	–	ns	–	ns	–	ns	–	1	0.14
	2	ns	–	ns	–	2	0.32	2	0.22	2	0.15	1	0.26
	3	ns	–	3	0.32	3	0.18	2	0.46	2	0.34	1	0.36
	4	ns	–	4	0.22	3	0.42	2	0.64	2	0.51	1	0.45
	5	ns	–	4	0.50	3	0.63	2	0.77	2	0.64	1	0.52
	6	6	0.32	4	0.72	3	0.77	2	0.86	2	0.75	1	0.59
	7	7	0.27	4	0.85	3	0.87	2	0.92	2	0.83	1	0.65
	8	7	0.59	5	0.78	3	0.93	2	0.95	2	0.88	1	0.70
	9	8	0.53	5	0.88	3	0.96	3	0.88	2	0.92	1	0.74
	10	8	0.76	5	0.94	3	0.98	3	0.92	2	0.95	1	0.77
200	1	ns	–	ns	–	ns	–	ns	–	ns	–	1	0.34
	2	ns	–	ns	–	2	0.54	2	0.44	2	0.36	1	0.56
	3	ns	–	3	0.54	3	0.40	2	0.73	2	0.65	1	0.71
	4	ns	–	4	0.44	3	0.71	2	0.88	2	0.82	1	0.81
	5	ns	–	4	0.77	3	0.88	2	0.96	2	0.91	1	0.88
	6	6	0.54	4	0.92	3	0.96	2	0.98	2	0.96	1	0.92
	7	7	0.49	5	0.88	3	0.99	3	0.96	2	0.98	1	0.95
	8	8	0.44	5	0.95	3	0.99	3	0.97	2	0.99	1	0.96
	9	8	0.78	5	0.98	4	0.99	3	0.99	2	>0.99	1	0.98
	10	9	0.75	5	>0.99	4	>0.99	3	>0.99	2	0.99	1	0.98

ns no significance possible at the 0.05 level for this marker-NIL-BC combination; * number of lines out of total NIL set that must contain the DP allele for a marker to be declared significantly linked to the gene; *P* the probability that the DP allele for a given marker will be found in the minimum number of NILs in a set required for significance, or more, if the marker is the maximum distance from the gene based on the marker density

number of NILs and backcrosses will be a fixed value. The theory presented can then be used to determine (1) if the number of lines available are suitable for mapping a gene using this method, (2) the probability that given situations (e.g., 3 out of 3 NILs have the DP allele) indicate linkage, and (3) the number of markers to use to have a given probability of success.

NILs can be a valuable resource in the integration of molecular marker and genetic marker maps as recognized by Muehlbauer et al. (1988). The theory we provide will be useful in evaluating most situations encountered when using this method. It provides a way to evaluate the probability that regions containing DP DNA in 1 or more NILs are linked to the introgressed gene. In addition, the optimum number of NILs, backcrosses and/or selfs, and markers can be determined for most situations based on desired probability levels.

The use of NILs is one way to reduce the number of individuals genotyped when mapping a gene. In general, the NIL analysis method will be most useful when NILs are available and several loci can be mapped simultaneously. Other methods such as the bulk-segregant analysis proposed by Michelmore et al. (1991) may also be useful. The choice of method will depend on the materials available. Regardless of the method used, exact linkage values will likely need to be estimated in a segregating population.

Acknowledgements. Paper no. 20026, Scientific Journal Series, Minnesota Agricultural Experiment Station.

References

Allard RW (1960) Principles of plant breeding. John Wiley, New York

Fehr WR (1987) Principles of cultivar development. MacMillan, New York

Haldane JBS (1919) The combination of linkage values, and the calculation of distance between the loci of linked factors. J Genet 8:299–309

Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12:172–175

Michelmore RW, Paran I, Kesseli RV (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

Muehlbauer GJ, Specht JE, Thomas-Compton MA, Staswick PE, Bernard RL (1988) Near-isogenic lines – a potential resource in the integration of conventional and molecular marker linkage maps. Crop Sci 28:729–735

Muehlbauer GJ, Specht JE, Staswick PE, Graef GL, Thomas-Compton MA (1989) Application of near-isogenic line mapping technique to isozyme markers. Crop Sci 29:1548–1553

Muehlbauer GJ, Staswick PE, Specht JE, Graef GL, Shoemaker RD, Keim P (1991) RFLP mapping using near-isogenic lines in the soybean [*Glycine max* (L.) Merr.]. Theor Appl Genet 81:189–198

Paterson AH, Deverna JW, Lanini B, Tanksley SD (1990) Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes in an interspecies cross of tomato. Genetics 124:735–742

Weir BS (1990) Genetic data analysis. Sinauer, Sunderland, Mass.