Theor Appl Genet (1992) 85:205-212

THEORETICAL AND APPLIED GENETICS

Evaluation of soybean RFLP marker diversity in adapted germ plasm

P. Keim [1], W. Beavis [2], J. Schupp [1], and R. Freestone [3]

[1] Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ 86011-5640, USA
[2] Pioneer HI-Bred International, Johnston, IA 50131, USA
[3] Pioneer HI-Bred International, Waterloo, IA 50703, USA

Received September 10, 1992; Accepted April 7, 1992
Communicated by A. L. Kahler

Summary. Soybean RFLP markers have been primarily developed and genetically mapped using wide crosses between exotic and adapted genotypes. We have screened 38 soybean lines at 128 RFLP marker loci primarily to characterize germ plasm structure but also to evaluate the utility of RFLP markers identified in unadapted populations. Of these DNA probes 70% detected RFLPs in this set of soybean lines with an average polymorphism index of 0.30. This means that only 1 out of 5 marker loci was informative between any particular pair of adapted soybean lines. The variance associated with the estimation of RFLP genetic distance (GD_R) was determined, and the value obtained suggested that the use of more than 65-90 marker loci for germ plasm surveys will add little precision. Cluster analysis and principal coordinate analysis of the GD_R matrix revealed the relative lack of diversity in adapted germ plasm. Within the cultivated lines, several lines adapted to Southern US maturity zones also appeared as a separate group. GD_R data was compared to the genetic distance estimates obtained from pedigree analysis (GD_P). These two measures were correlated with $r = 0.54$ for all 38 lines, but the correlation increased to $r = 0.73$ when only adapted lines were analyzed.

Key words: *Glycine max* - DNA - Genetic distance - Pedigree analysis

Introduction

Protocols for revealing restriction fragment length polymorphisms (RFLPs) have been developed for many crops in order to identify quantitative trait loci (QTLs), to characterize genetic diversity in breeding populations, and to discriminate between varieties for legal purposes. Maize has been most intensively studied (Helentjaris 1987; Burr et al. 1988; Smith et al. 1990), but other crops such as tomato, lettuce, *Brassica*, and soybean have well-developed RFLP maps, and in many cases these markers have been associated with QTLs (Patterson et al. 1988; Landry et al. 1987; Neinhuis et al. 1987; Keim et al. 1990a, b). The genetic diversity present in breeding populations can be characterized with RFLPs and, hence, can be used to its maximum potential for crop improvement. RFLPs have not been used in the judicial system for legal purposes with respect to crops as yet, but their use in human forensic work (Jefferys et al. 1985) illustrates their future value.

In several crops little RFLP diversity has been observed within adapted germ plasm, prompting researchers to use exotic accessions for the initial mapping studies. In general, self-fertilizing crops have shown less diversity than out-crossing species. It would seem that open-pollinated crops are more "tolerant" of molecular changes that create RFLPs. Hence, unadapted germ plasm with greater genetic diversity has been used in autogamous crops for RFLP mapping. In tomato and soybean, this strategy has been successful for the construction of RFLP genetic maps and for the identification of QTLs (Neinhuis et al. 1987; Patterson et al. 1988; Keim et al. 1990a). Most of the effort being put into soybean breeding is concentrated on adapted germ plasm. Because soybean RFLP markers have only been developed using exotic populations, it is not known how useful RFLP markers will be in evaluating adapted germ plasm. Markers identified in very wide crosses will generally only be useful if they are able to reveal variation in adapted germ plasm.

Correspondence to: P. Keim

We have screened 132 public RFLP probes in a collection of adapted and ancestral soybean lines. Our goal was to estimate the usefulness of such markers in revealing variation with adapted germ plasm and to estimate genetic diversity in soybean breeding populations.

Materials and methods

The 38 soybean [*Glycine max* (L.) Merr.] lines surveyed in this study were selected because they are currently used in North American commercial soybean breeding programs or because they have contributed to the current breeding lines as ancestral parental material (see Table 1). Of these 38 lines 20 are considered to "adapted" germ plasm. Two of the adapted lines are Pioneer be proprietary breeding lines and are coded P-1 and P-2. Two unadapted plant introductions (PI 88.788 and PI 437.654) were included in this study because they have been used in Pioneer's breeding program for the introgression of specific agronomic traits. RFLP genotypes were determined from DNA extracted from a sample of at least 25 individuals representing each line.

RFLP analysis of soybean lines involved the Southern transfer technique and molecular hybridization with radioactive DNA probes. Recombinant DNA probes used to detect RFLP markers were derived from a random *Pst*I library (Keim and Shoemaker 1988; Keim et al. 1990a). These probes were obtained from Drs. K. G. Lark (Department of Biology, University of Utah) and R. C. Shoemaker (USDA, Agronomy Hall, Iowa State University). All probes were used in combination with a single restriction enzyme (*Eco*RI, *Eco*RV, *Hin*dIII, *Dra*I, or *Taq*I), which in earlier studies detected polymorphisms (Apuya et al. 1988; Keim et al. 1990a). The probes obtained from the University of Utah had previously detected variation in a *G. max* × *G. max* population, while the USDA/ISU probes had previously detected variation in a *G. max* × *G. soja* Seib. and Zucc. population. DNA from each line was isolated from leaves according to Keim et al. (1988). Radioactive DNA probes were synthesized by random priming (Feinberg and Vogelstein 1983).

Table 1. Soybean lines surveyed with RFLP markers

Ancestral lines	*Adapted lines*
1. Mukden	19. Northup-King S1346
2. Manchu	20. Grant
3. A.K. Harrow	21. Essex
4. Mandarin (Ott.)	22. Keller
5. Tokyo	23. Dare
6. Dunfield	24. Pioneer 5482
7. Illini	25. Forrest
8. PI 88.306	26. Bragg
9. PI 84.686	27. Lee
10. PI 200.593	28. P-1
11. PI 92.567	29. P-2
12. PI 248.398	30. Midwest Oilseeds 30421
13. Bavender Special	31. Pride B216
14. PI 88.788	32. Williams
15. PI 437.654	33. N.A.P.B. HP2530
16. Richland	34. Asgrow 3127
17. Seneca	35. Pioneer 9271
18. Roanoke	36. Pioneer 2981
	37. Corsoy
	38. Asgrow 1564

DNA hybridization was in an aqueous cocktail (0.6 *M* NaCl, 0.12 *M* TRIS pH 8, 0.008 *M* EDTA, 0.1% saturated sodium pyrophosphate, 0.1% SDS, 20 μg/mL denatured salmon sperm DNA, 1 × Denhart's solution, and 6% polyethyleneglycol) at 65 °C. Membranes were washed 3 times at 65 °C in 0.2 × SSC, 0.1% SDS and then subjected to fluorography with intensifier screens (Maniatis et al. 1982).

Multiple RFLP loci frequently are detected with a single probe in soybean germ plasm. Distinguishing allelic from nonallelic fragments can be accomplished in soybean because mutually exclusive banding patterns are revealed among inbred lines (Keim et al. 1989; Keim et al. 1990a, b). In the present study, the term "probe" refers to the recombinant DNA clone that detects complementary restriction fragments. It is not synonymous with the term "marker" because a single probe may detect multiple fragments at different loci. Probes that hybridize with polymorphic fragments at different loci are useful for revealing polymorphisms at more than one genetic locus.

The utility of a marker can be judged by its ability to distinguish among lines, i.e., the number of informative comparisons it will provide. The number of informative comparisons is a function of the number of alleles detected with each probe and their frequencies. The most informative markers are those that have a large number of alleles at an equal frequency. A measure of this can be obtained by subtracting from unity the sum of the squared allele frequencies. This measure has been referred to as a polymorphism index (Marshall and Allard 1970) and as gene diversity (Nei 1973; Weir 1990). On a marker basis, we have calculated that the polymorphism index $=1-\sum p_i^2$ for alleles $i=1, 2, 3, \ldots n$. On a probe basis, the polymorphic index $=1-\sum\sum p_i^2$ where different polymorphic loci are summed as well.

We estimated genetic distance between all pairs of inbreds with both pedigree information (GD_P) and RFLP information (GD_R). GD_P was calculated as 1 − (the coefficient of parentage), where the coefficient of parentage was estimated using available pedigree information (Delannay et al. 1983). The proportion of similar RFLP loci, S_{XY}, between pairs of varieties was estimated as $2\,N_{XY}/(N_X+N_Y)$, where N_{XY} is the number of RFLP loci for which varieties X and Y possess the same allele, N_X is the number of alleles identified in variety X, and N_Y is the number of alleles identified in variety Y. This is algebraically equivalent to Nei and Li's (1979) estimate of the proportion of similar sized fragments generated by restriction sites within a locus, however we applied the calculation to numerous marker loci. GD_R was then calculated as $1-S_{XY}$.

Relationships among varieties based upon both pedigree and RFLP information were investigated using principal coordinate and cluster analyses. The principal coordinates were found by centering the genetic distance matrices and conducting an eigenvector analysis on the centered distance matrices (Chatfield and Collins 1980). Cluster diagrams were constructed using the average linkage clustering algorithm (Statistical Analysis Systems, Cary N.C.) on the distance matrices.

Results

Probes and markers

We evaluated recombinant DNA probes that identified polymorphisms (markers) in very diverse soybean germ plasm originally for their ability to distinguish among cultivars and ancestral genetic lines. We found that 69% of the 132 probes detected variation among the 38 lines (Table 2). The probe polymorphism frequency was af-

Table 2. Information from markers identified in different populations

Probe source	Probes screened	Number of polymorphic marker[c]		Average polymorphism index per marker
		All lines		
G. max × *G. soja*				
Population[a]	104	72 (69%)	97	0.30
G. max × *G. max*				
Population[b]	28	20 (71%)	31	0.30
Total	132	92 (69%)[d]	128	0.30
		Adapted lines only		
G. max × *G. soja*				
Population[a]	104	55 (53%)[e]	74	0.32
G. may × *G. max*				
Population[b]	28	18 (64%)[e]	28	0.30
Total	132	73 (55%)[d]	102	0.32

[a] *G. max* × *G. soja* (Keim et al. 1990 a, b)

[b] *G. max* × *G. max* (Apuya et al. 1988)

[c] Percentage of probes detecting RFLPs is enclosed by parentheses

[d] Highly significant difference ($Z = 2.4$; $P < 0.01$)

[e] No significant difference ($Z = 1.1$; $P < 0.15$)

fected by both the source (original screening population) of the probes and the type of soybean germ plasm being evaluated. When only adapted germ plasm was considered, the polymorphism frequency per probe was 0.55 versus 0.69 when all of the genotypes were considered. This is a significant improvement in the probe polymorphism frequency ($Z = 2.4$; $P < 0.01$). While the source population of the probes did not affect polymorphism frequency when all of the lines were evaluated, the probe polymorphism frequency was lower for the interspecific (0.53) source than for the intraspecific (0.64) source when only adapted germ plasm was considered. This difference was not statistically significant ($Z = 1.1$; $P < 0.15$), possibly due to the limited number of probes evaluated from the *G. max* × *G. max* population (28). However, it would not be surprising that RFLP variation found in very diverse soybeans is not always present in adapted germ plasm. These data suggest that the *G. max* × *G. max* population would be a better source for identifying RFLP probes with which to evaluate adapted germplasm.

How well a probe distinguishes among soybean genotypes is determined by the number of polymorphic loci as well as the number and frequency of alleles per locus detected. Random genomic probes frequently detect multiple loci in soybean (Keim et al. 1990 a). In previous studies these multiple polymorphic fragments were observed to segregate independently (Apuya et al. 1988; Keim et al. 1990 a). In the present study multiple RFLP loci were also detected with individual probes, although most of the probes (66%) detected only a single polymorphic locus (Fig. 1). As previously reported (Keim et al. 1990 a), probes revealing 1 RFLP marker usually detected additional monomorphic restriction fragments; probes detecting 2 polymorphic loci were approximately one-third less prevalent than single-locus probes; and probes revealing 3 polymorphic loci were one-third again less frequent than 2-loci probes. A similar one-third reduction in frequency was observed for probes revealing 3 and 4 polymorphic loci.

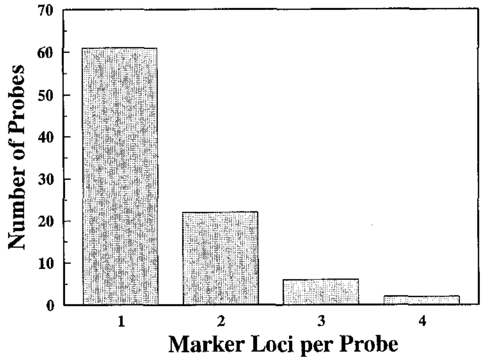

Fig. 1. Frequency of RFLP markers per probe. The number of probes detecting variation are categorized by the number of polymorphic loci they detect

The polymorphic loci observed in the present study were primarily represented by two alleles; only 3 loci had three alleles. The polymorphism index (p-i) is based upon the number and frequencies of alleles; this is equivalent to the frequency of pair-wise comparisons among lines that would be polymorphic for a particular genetic marker. Theoretically, the most useful marker in this study (1 with three alleles) would have a maximum p-i = 0.67, while the maximum for most markers (those with 2 alleles) would be p-i = 0.50. The average p-i value for all of the markers in this study was 0.30 (Table II). This average value was not noticeably affected by either the probe source or the type of germ plasm. The p-i distribution of markers (Fig. 2), however, does not represent a normal distribution around the mean (Kolmogorov-Smirnov test, $P > 0.001$ for all lines), but rather a distribution skewed toward the more polymorphic classes.

Estimates of the genetic distance (GD_R) between all pairs of the 38 lines used in this study were obtained using 128 polymorphic markers (Table 3). Of practical concern is the relationship between the number of markers and the precision of GD_R. Because alleles can be identified from their mutually exclusive banding patterns (Keim et al. 1989), we assumed that the marker loci represent independent samples of the genome. Thus, S_{xy} is distrib-

Table 3. Genetic distance matrix determined by RFLP and pedigree analysis

$\mathrm{GD_R}$ (×100)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1. A1564	**	32	20	22	37	30	13	33	25	37	35	33	24	19	28	27	35	24	37	30	50	23	24	37	30	34	35	33	33	23	35	23	31	31	32	33	33	30
2. A3127	95	**	19	26	30	24	30	25	18	20	24	21	20	20	27	21	29	36	20	23	50	16	15	44	31	32	30	32	35	24	24	21	30	35	35	28	33	14
3. AK Harr	91	94	**	23	32	21	19	25	13	26	26	21	10	01	22	14	32	35	17	22	51	16	15	43	28	32	32	28	37	17	26	16	35	33	35	28	33	18
4. B216	85	80	82	**	32	32	23	26	23	33	32	32	19	22	28	26	30	32	14	31	47	15	14	43	31	33	34	34	36	20	29	21	27	32	26	32	35	16
5. Bav. Sp.	100	100	100	100	**	32	33	26	25	37	39	32	34	32	27	40	28	36	28	38	43	33	34	39	35	17	28	29	35	35	33	35	35	35	28	42	37	31
6. Bragg	98	91	100	97	100	**	31	25	27	27	09	28	25	22	29	20	33	31	25	34	47	28	27	40	36	32	32	28	35	28	21	34	42	24	37	33	29	30
7. Corsoy	79	98	63	51	100	100	**	33	25	34	34	31	21	18	28	27	32	22	26	21	48	25	26	46	32	35	27	29	32	14	34	21	30	35	28	31	31	26
8. Dare	97	93	100	99	100	89	100	**	20	24	26	27	24	24	21	27	33	37	24	30	45	28	27	39	28	31	29	25	33	29	14	25	31	18	26	30	33	26
9. Dunfield	100	94	100	100	100	100	100	88	**	27	30	22	17	14	18	24	31	32	17	25	43	22	21	40	25	26	28	25	36	24	26	22	30	26	28	27	30	22
10. Essex	99	49	100	96	100	85	100	89	100	**	29	26	29	26	30	15	38	34	35	28	53	29	29	44	36	37	39	33	40	33	27	29	34	33	39	24	31	32
11. Forrest	99	88	100	97	100	43	100	80	94	79	**	32	25	25	35	20	36	37	27	35	46	26	25	40	38	37	32	33	38	29	16	35	44	27	37	33	29	29
12. Grant	94	88	100	85	100	100	100	100	100	100	100	**	22	21	28	23	30	31	27	24	52	23	22	43	28	32	32	27	32	26	31	24	34	26	34	14	30	22
13. HP2530	78	78	97	68	100	95	93	95	100	96	95	78	**	09	23	16	33	35	13	25	50	18	17	48	32	34	35	32	38	12	23	13	33	31	32	27	35	12
14. Illini	91	94	02	82	100	100	63	100	100	100	100	100	97	**	23	13	32	35	16	21	52	15	14	43	27	31	31	27	36	17	25	15	34	32	34	27	32	17
15. Keller	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	31	34	32	26	26	42	30	29	41	31	30	32	28	35	32	29	28	33	27	27	31	33	32
16. Lee	100	71	100	94	100	75	100	88	100	48	63	100	94	100	100	**	36	31	22	25	52	20	20	45	34	40	38	31	37	21	21	21	38	34	39	24	28	23
17. Manchu	94	88	100	85	100	100	100	100	100	100	100	75	78	100	100	100	**	29	30	34	47	28	27	45	24	32	24	33	21	34	37	37	27	37	34	39	38	30
18. Mandarin	72	100	100	82	100	100	63	100	100	100	100	100	91	100	100	100	100	**	40	32	47	35	35	41	31	38	35	35	24	32	39	36	33	32	33	32	30	40
19. MO30421	91	86	91	79	75	99	91	99	94	98	98	68	79	91	100	97	86	91	**	27	50	17	16	47	33	28	32	33	38	17	27	20	29	31	29	32	35	13
20. Mukden	82	100	100	100	100	100	100	100	100	100	100	100	94	100	100	100	100	100	100	**	47	27	27	45	33	38	30	33	36	24	33	29	32	37	32	17	35	27
21. PI437654	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	50	50	32	52	47	43	49	47	53	45	59	51	53	47	57	50	51
22. P2	90	41	88	41	100	94	74	96	97	73	92	86	73	88	100	82	86	91	83	100	100	**	01	41	28	32	33	32	32	22	29	19	33	36	35	29	36	15
23. P1	90	41	88	41	100	94	74	96	97	73	92	86	73	88	100	82	86	91	83	100	100	03	**	42	27	32	32	32	32	21	28	18	32	35	35	28	35	14
24. PI88.788	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	43	39	46	42	42	49	38	47	45	39	44	45	41	46
25. PI20059	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	37	23	28	33	31	36	32	29	32	32	33	39	33
26. PI24839	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	32	31	39	38	39	36	37	34	28	39	38	34
27. PI84686	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	28	31	32	34	36	35	30	28	37	32	32
28. PI88306	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	38	34	29	32	40	28	31	30	36	34
29. PI92567	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	37	38	37	30	37	34	38	33	37
30. PIO2981	62	88	91	75	100	97	79	98	100	98	98	89	75	91	100	97	89	72	85	88	100	81	81	100	100	100	100	100	100	**	32	14	26	33	30	29	31	16
31. PIO5482	98	90	100	98	100	66	100	41	91	84	41	100	95	100	100	75	100	100	98	100	100	94	94	100	100	100	100	100	100	98	**	27	42	29	35	31	34	25
32. PIO9271	87	74	85	68	100	99	69	99	94	98	98	88	77	85	100	97	88	79	84	100	100	71	71	100	100	100	100	100	100	80	98	**	32	35	32	31	32	16
33. Richland	82	94	100	94	100	100	100	100	100	100	100	100	88	100	100	100	100	100	97	100	100	94	94	100	100	100	100	100	100	85	100	94	**	33	26	33	37	27
34. Roanoke	100	94	100	100	100	100	100	75	100	88	100	100	100	100	100	100	100	100	100	100	100	97	97	100	100	100	100	100	100	100	88	100	100	**	28	29	22	36
35. S1346	92	98	100	99	100	100	100	94	100	97	100	100	98	100	100	100	100	100	100	88	100	99	99	100	100	100	100	100	100	94	97	99	88	75	**	35	34	30
36. Seneca	100	100	100	100	100	100	100	100	100	100	100	51	100	100	100	100	100	100	88	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	**	33	30
37. Tokyo	94	99	100	100	100	82	100	88	100	97	91	100	94	100	100	100	100	100	100	100	100	99	99	100	100	100	100	100	100	97	89	100	100	100	100	100	**	36
38. Williams	90	49	88	65	100	97	95	97	88	96	96	75	60	88	100	94	75	100	74	100	100	57	57	100	100	100	100	100	100	78	96	50	88	100	99	100	100	**

$\mathrm{GD_P}$ (×100)

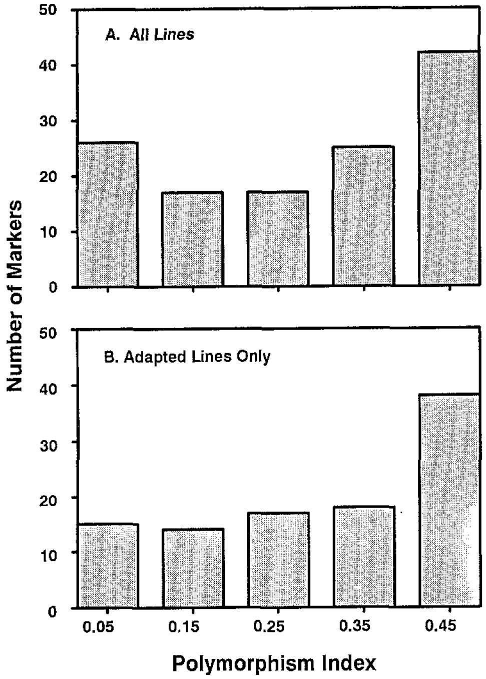

Fig. 2 A, B. Polymorphism index for RFLP markers. **A** Analysis using all lines from Table 1; **B** analysis using only adapted lines from Table

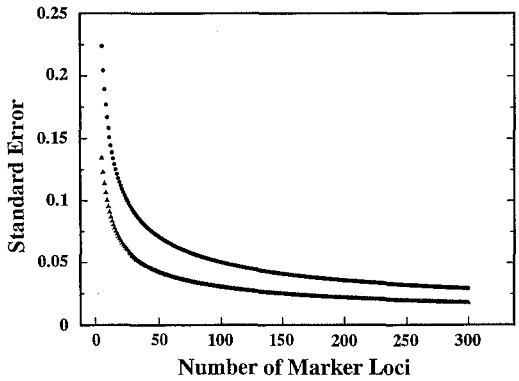

Fig. 3. Binomial estimation of the standard error associated with GD_R determination. The binomial proportion was used to investigate the effect of increasing the number of marker loci in estimating GD_R. Because the variance is not constant (see text), the modelling was done for the extreme cases where GDR = 0.5 [$var(GD_R) = 0.25$ (●)] and GDR = 0.1 or 0.9 [$var(GD_R) = 0.09$ (▲)]

uted as a binomial proportion with the variance = $S_{xy}(1 - S_{xy})/n$, where n is the number of marker loci. The precision of the GD_R estimate is not constant over the range of values (0–1.0): it is minimized for pairs of varieties that have half of their alleles in common ($GD_R = 0.5$) and maximized between pairs of varieties that have either all, or none, of their alleles in common (GD_R approaches 0 or 1). Figure 3 represents the effect of adding marker loci to the GD_R estimation using the variance associated with the least precise measure of GD_R [$GD_R = 0.5$ and $var(GD_R) = 0.25$], as well as the more precise measure when GD_R approaches 1.0 or 0 [$GD_R = 0.1$ or 0.9: $var(GD_R) = 0.09$]. We have used selected subsets of varieties and the Jackknife procedure (Quenouille 1956; Weir 1990) to empirically confirm these estimates of variance (data not presented). Very little decrease in standard error occurs with more than 90 marker loci for the higher $var(GD_R)$ or 65 marker loci for the lower $var(GD_R)$. We conclude that in soybean studies estimating GD_R, a minimum of 65 marker loci should be used, but more than 90 independent markers will provide little improvement in precision. Our use of 128 marker loci exceeded these lower limits.

Genetic structure of soybean

Estimates of genetic distances among the lines as revealed by RFLPs and pedigree information are given in Table 3. The average GD_R among these lines is 0.31, whereas the average GD_P is 0.95. This discrepancy may result from a lack of detailed pedigree information among many of these soybeans lines. Consequently, it is incorrectly assumed in the calculation of GD_P that lines showing no pedigree relationship are not related ($GD_P = 1.0$). GD_R, on the other hand, is a direct measure of the proportion of RFLP loci that are different, and these data are equally available for all varieties in this study. None-the-less, there are several instances in which the pedigree distances are in good agreement with the RFLP distances. 'Illini' is a cultivar that was selected out of the heterogenous collection'A.K. Harrow'. These two lines had near-identical RFLP patterns (1 out of 129 loci differed). The breeding lines P-1 and P-2 also had identical pedigrees and near-identical RFLP patterns. Overall, however, GD_R are poorly correlated with GD_P ($r = 0.54$). The correlation improved when only adapted line values were considered ($r = 0.73$). The more detailed pedigree information available for adapted lines is probably responsible for the greater correlation. The two plant introductions (PI 88.788 and PI 437.654) included in this study proved to be very different from all the other lines (Table 3). The average GD_R among "Northern" lines known to be adapted to maturity zones 0, I, II, and III was 0.22; the average GD_R among "Southern" lines adapted to zones V, VI, VII, and VIII was 0.21. In contrast, the average GD_R between the North and South groups was higher, 0.33. Soybean breeders can maximize genetic diversity in segregating populations by crossing between these maturity groups.

The actual genetic distance values are useful indicators of relationships but the identification of genetic relationships among lines was more easily accomplished following an analysis of principal coordinates (PCA, Fig. 4) or clustering (Fig. 5). Both analyses revealed that these

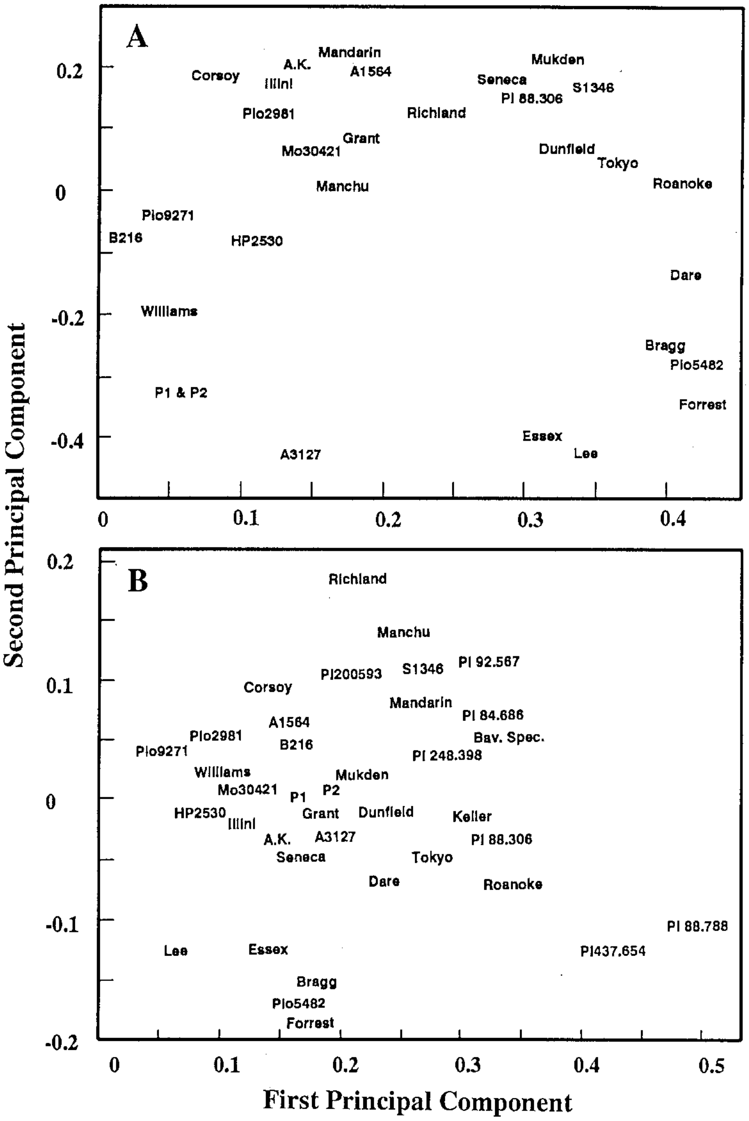

Fig. 4A, B. Principal coordinate analysis. **A** Pedigree relationships were used to calculate genetic distances. These were then analyzed by principal coordinate analysis. The first and second principal components explained 12% and 8% of the variation, respectively. Soybean lines lacking pedigree information were not included in this analysis; **B** RFLP marker data was used to calculated GD_R among soybean lines. This distance matrix was analyzed by principal coordinate analysis. The first two principal coordinates explained 21% and 13% of the variation, respectively

soybean lines are easily distinguished using RFLPs, but neither technique alone fully revealed the multidimensional genetic relationships that exist among these lines. PCA (Fig. 4b) revealed five varieties ('Lee', 'Essex', 'Bragg', 'Pio5482', and 'Forrest') that are disjunct from the others. These represent "Southern" germ plasm, which may account for their separation from the other varieties. This grouping is not as clearly revealed in the cluster analysis (Fig. 5b). However, the cluster analysis clearly showed the large GD_R that separates the plant introductions PI 88,788 and PI 437,654 from each other and from all other genotypes. While this separation is observed with PCA, the magnitude of the separation is more apparent with cluster analysis. Neither analysis is incorrect, rather each is revealing different aspects of the genetic relationships.

PCA and cluster analysis results using GD_P values are similar to the GD_R results only when detailed pedigree information exists. For example, Figs. 4a and 5a both reveal an association among the 5 adapted Southern varieties discussed above. However, overall many of the relationships observed with GD_R data are not seen with GD_P results. The association between PI 248.398 and Bavender Special observed with RFLP data (Figs. 4b and 5b) was not evident in the pedigree analysis. These differences can largely be attributed to the lack of pedigree information in the ancestral lines.

Discussion

The RFLP diversity observed in this study may be greater than that present in soybean cultivars used by producers. Soybean lines in this study were chosen to represent the "breadth" of genetic lines used for breeding cultivated soybeans adapted to the US. Therefore, they

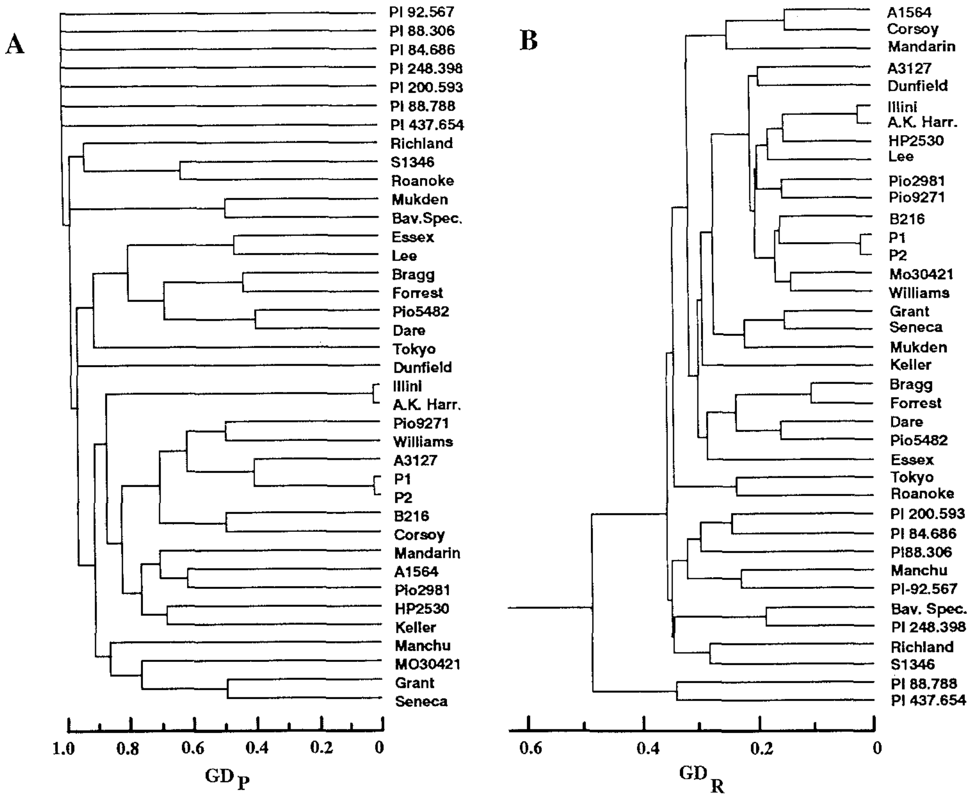

Fig. 5A, B. Cluster analysis. **A** Genetic distances calculated from pedigree relationships were analyzed by average-linkage clustering; **B** RFLP marker data were used to calculate GD_R. This distance matrix was analyzed using average-linkage cluster analysis

are not representative of the elite commercial or public lines used for the bulk of US soybean production. In a previous study (Keim et al. 1989) it was observed that 7 adapted soybean lines were identical when examined with 17 RFLP markers. The data from the study presented here is consistent with that result, although the adapted germ plasm studied here was not identical, with two exceptions (Fig. 5B). A comprehensive study of soybean production germ plasm is needed to determine the actual relatedness of currently used cultivars.

The majority of the soybeans in our study can be easily distinguished with RFLP markers. The exceptions have identical pedigrees and, therefore, should have very similar genotypes. In both cluster and principal component analysis, there are only a few examples of "tight" grouping or clustering, such as the Southern germ plasm example discussed above. Many of the soybean lines used in this study were equally distant to other lines. The striking separation of plant introductions PI 88.788 and PI 437.654 illustrates the relatively low diversity present in adapted germ plasm when contrasted with unadapted germ plasm. Plant introductions can provide the genetic diversity presently lacking in soybean breeding programs. For example, exotic germ plasm was used to construct Northrup-King var. 'S1346', and our RFLP data indicate that this genotype is one of the more diverse types of the adapted varieties observed in this study (Fig. 5b).

The large GD_R values that we observed separating soybean lines seem to contradict previous reports of low diversity. Indeed, these values are comparable to those observed in maize studies (Smith et al. 1990). Several important differences exist in how diversity is measured and in the type of genetic lines being examined. First, GD_Rs in this study are calculated only from polymorphic probes and polymorphic restriction fragments. Maize studies typically have variation in all DNA fragments with all probes. The soybean distances reported here do not include 30% of the probes that were monomorphic (Table 2), nor do they include multiple loci detected with each probe that were monomorphic. Our 128 RFLP markers probably represent a survey of over 300 loci, but only 128 were polymorphic. If monomorphic loci were included, the GD_R values could be reduced to as little as one-third of the values reported here. Secondly, corn studies (e.g., Smith et al. 1990) generally examine germ plasm involved in crop production and do not include all of the land races used to construct the current elite lines. Hence, the corn studies are only examining a fraction of the breadth of the germ plasm. As mentioned above, this soybean study has included plant introductions and not concentrated on the elite production cultivars. In summary, the direct comparison of corn genetic distances and soybean genetic distances can be misleading.

We have analyzed public recombinant DNA probes in cultivated soybean in order to characterize their general usefulness to soybean breeders. As the public RFLP map develops it is essential that a database on cultivated varieties be developed concurrently. Detailed information such as DNA fragment sizes, DNA fragment allelism, genetic locations, and discrimination power (i.e., polymorphism index) associated with each RFLP probe will make them more valuable to researchers. Of the 132 probes used in this study 70% detected variation with an average polymorphism index of 0.3. these numbers translate to approximately 1 in 5 probes ($0.7 \times 0.3 = 0.21$) being useful between any particular pair of the lines. Therefore, a map of 400 markers developed in unadapted germ plasm will provide only approximately 80 markers to researchers working with any one particular cultivated soybean population. For many studies 80 markers may be sufficient, but for precise identification of QTLs this will not be adequate. If future RFLP markers are to be developed in unadapted populations, researchers must develop fairly saturated maps. Alternatively, markers might best be developed in cultivated populations. The higher polymorphism frequency observed in adapted germ plasm for "intraspecific" probes supports this approach.

Acknowledgements. This paper is dedicated to Drs. J. H. Burkholder and G. J. Ikenberry, Jr. (McPherson College, McPherson Kan.), who have used their careers to teach students about genetics and plants. We would like to thank Drs. K. G. Lark and R. C. Shoemaker for providing probes for this study. Drs. A. Diehl, R. G. Palmer, J. S. C. Smith, J. E. Specht, and D. M. Webb provided helpful comments concerning the manuscript. This study was supported by a grant from Pioneer Hi-Bred, International to P.K.

References

Apuya N, Frazier BL, Keim P, Roth EJ, Lark KG (1988) Restriction length polymorphisms as genetic markers in soybean, *Glycine max* (L.) Merr. Theor Appl Genet 75:889–901

Burr B, Burr FA, Thompson KA, Albertson MC, Stuber CW (1988) Gene mapping with recombinant inbreds in maize. Genetics 118:519–526

Chatfield C, Collins AJ (1980) Introduction to multivariate analysis. Chapman and Hall, New York, pp 189–210

Delannay X, Rodgers DM, Palmer RG (1983) Relative genetic contribution among ancestral lines to North American soybean cultivars. Crop Sci 23:944–949

Feinberg AP, Vogelstein B (1983) A technique for radiolabelling DNA restriction endonuclease fragments to a high specific activity. Anal Biochem 132:6–13

Helentjaris T (1987) A genetic linkage map for maize base upon RFLPs. Trends Genet 3:217–221

Jeffreys AJ, Wilson V, Thein SL (1985) Individual specific fingerprints of human DNA. Nature 316:76–79

Keim P, Shoemaker RC (1988) construction of a random recombinant DNA library that is primarily single copy sequence. Soybean Genet Newsl 15:147–148

Keim P, Olson TC, Shoemaker RC (1988) A rapid protocol for isolating soybean DNA. Soybean Genet Newsl 15:150–152

Keim P, Shoemaker RC, Palmer RG (1989) RFLP diversity in soybean. Theor Appl Genet 77:786–792

Keim P, Diers BW, Olson TC, Shoemaker RC (1990a) RFLP mapping in soybean: Association between marker loci and variation in quantitative traits. Genetics 126:735–742

Keim P, Diers BW, Shoemaker RC (1990b) Genetic analysis of soybean hard seededness with molecular markers. Theor Appl Genet 79:465–469

Landry BS, Kesseli RV, Farrara B, Michelmore RW (1987) A genetic map of lettuce (*Lactuca sativa* L.) with restriction fragment length polymorphism, isozyme, disease resistance and morphological markers. Genetics 116:331–337

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Marshall DR, Allard RW (1970) Isozyme polymorphisms in natural populations of *Avena fatua* and *A. barbata*. Heredity 25:373–382

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Nei M, Lei W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5256-5273

Nienhuis J, Helentjaris T, Slocum M, Ruggero B, Schaefer A (1987) Restriction fragment length polymorphism analysis of loci associated with insect resistance in tomato. Crop Sci 27:797–803

Patterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors, using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721–726

Quenouille M (1956) Notes on bias in estimation. Biometrika 43:253–260

Smith OS, Smith JSC, Bowen SL, Tenborg RA, Wall SJ (1990) Similarities among a group of elite maize inbreds as measured by pedigree, F_1 grain yield, heterosis, and RFLPs. Theor Appl Genet 80:833–840

Weir B (1990) Genetic data analysis: methods for discrete population genetic data. Sinauer Assoc, Sunderland, Mass., pp 124–134