Theor Appl Genet (1992) 85:293–302

THEORETICAL AND APPLIED GENETICS

Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild *Sorghum bicolor*

P. R. Aldrich and J. Doebley

Department of Plant Biology, University of Minnesota, St. Paul, MN 55108, USA

Received September 30, 1991; Accepted February 5, 1992
Communicated by A. R. Hallauer

Summary. Fifty-six accessions of cultivated and wild sorghum were surveyed for genetic diversity using 50 low-copy-number nuclear DNA sequence probes to detect restriction fragment length polymorphisms (RFLPs). These probes revealed greater genetic diversity in wild sorghum than in cultivated sorghum, including a larger number of alleles per locus and a greater portion of polymorphic loci in wild sorghum. In comparison to previously published isozyme analyses of the same accessions, RFLP analysis reveals a greater number of alleles per locus. Furthermore, many RFLP alleles have frequencies between 0.25–0.75, while the vast majority of isozyme alleles are either rare (<0.25) or near fixation (>0.75). Correlations between genetic and geographic distances among the accessions were stronger when calculated with RFLP than with isozyme data. Systematic relationships revealed by nuclear and chloroplast restriction site analysis indicate that cultivated sorghum is derived from the wild ssp. *arundinaceum*. The portion of the wild gene pool most genetically similar to the cultivars is from central-northeastern Africa. Previous published data also suggested that this is most likely the principal area of domestication of sorghum. Introgression between wild and cultivated sorghum was inferred from disconcordant relationships shown by nuclear and chloroplast DNA markers. Introgression apparently occurs infrequently enough that the crop and its wild relatives maintain distinct genetic constitutions.

Key words: Sorghum – RFLP – Genetic diversity – Domestication – Introgression

Correspondence to: J. Doebley

Introduction

Molecular markers are valuable tools for studying the evolution of crops and their wild relatives. Isozyme markers have proven useful to this end (reviewed in Doebley 1989); however, they are restricted to a limited number of coding regions of the genome and they can detect only those mutations that cause changes in protein mobility. In contrast, restriction fragment length polymorphisms (RFLPs) reveal variability at the level of the primary DNA sequence, which is the most basic material for addressing issues of evolutionary descent (Crawford 1990). Restriction analysis can be used to study all classes of mutations in any portion of the genome, thus allowing the resolution of a potentially unlimited number of loci (Clegg 1990). Although the process of RFLP analysis is generally more expensive and labor intensive than isozyme analysis, thereby sacrificing statistical strength to small sample sizes, the extensive nature of RFLP variation makes it an attractive option for some studies.

Restriction site analysis of the nuclear genome can aid in the study of several topics of interest regarding crop/progenitor relationships. Estimates of overall levels of genetic diversity may be more reliable when measured with restriction site analysis than by isozymes because of the greater number of loci available and the ability to survey both coding and noncoding regions of the genome (Doebley and Wendel 1989; Clegg 1990). In addition, taxonomic relationships may be better resolved since restriction analysis appears to reveal higher levels of variation than isozymes (Bernatzky and Tanksley 1989). This degree of resolution should allow a more thorough documentation of phylogenetic relationships below the species level (Doebley and Wendel 1989). A greater number of marker alleles resolved in the gene pool will also facilitate the identification of unique genetic profiles for indi-

viduals and populations. This identification is needed to test associations that may exist between geographic proximity and genetic composition in the crop's gene pool.

Restriction fragment variation in the chloroplast genome (cpDNA) may also be utilized to study a crop's gene pool. Although conservative in its evolution and generally offering limited variation below the species level, very little homoplasy is encountered for cpDNA restriction site mutations (Palmer et al. 1988). Combined analysis of cpDNA and nuclear markers may allow one to detect introgression between a crop and its relatives. If a lack of concordance between phylogenies produced from the two data sets exists, it is probable that the chloroplast genome of one taxon has introgressed into a nuclear background of another. It is also possible to use cpDNA and nuclear markers to distinguish seed dispersal from pollen flow. CpDNA is inherited maternally in most angiosperms, therefore pollen dispersal often can be ruled out as a causal factor in the resulting spatial distributions of chloroplast variants.

Levels of diversity in the primary gene pool of sorghum [*Sorghum bicolor* (L.) Moench] have been studied previously with isozymes (Morden et al. 1989; 1990; Aldrich et al. accepted). In this paper, we offer another perspective utilizing evidence from restriction fragment variation in the nuclear and chloroplast genomes to address four issues concerning evolution in the gene pool of this crop and its wild relatives. (1) What levels of diversity are present in the nuclear and chloroplast genomes as revealed by RFLP analysis, and how does this compare with isozyme variation? 2) What are the systematic relationships in the primary gene pool of sorghum based on nuclear and chloroplast RFLP data? 3) Is genetic composition strongly associated with geographic proximity in the wild or cultivated taxa, and if so, which method of analysis better detects this relationship, isozymes or RFLPs? 4) What role has introgression played in the evolution of *Sorghum bicolor*?

Materials and methods

Fifty-six accessions of *Sorghum bicolor* were selected to maximize geographic and racial diversity in the data set (Table 1). For cultivated sorghum, 3 accessions from each of ten countries in Africa and Asia were surveyed with the exception of Sudan (4 accessions). In a similar fashion, wild sorghum was sampled for 3 accessions from each of eight countries in Africa with the exception of Ethiopia (4 accessions). A single plant was assayed from each of the 56 accessions. The 56 accessions represent the following taxonomic groups: *Sorghum bicolor* ssp. *arundinaceum* (Desv.) de Wet et Harlan race aethiopicum (4 accessions), race arundinaceum (8 accessions), race verticilliflorum (8 accessions), race virgatum (4 accessions); S. *bicolor* ssp. *bicolor* race bicolor (7 accessions), race caudatum (6 accessions), race durra (6 accessions), race guinea (6 accessions), race kafir (6 accessions); S. *bicolor* ssp. *drummondii* (Steud) de Wet (1 accession).

Table 1. Country of origin, race, identification number, and sample number (in parentheses) of the *Sorghum bicolor* accessions included in this study. Seed from accessions with IS and PI prefixes were obtained from the International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) and USDA Regional Plant Introduction Station, Experiment, Georgia (RPIS), respectively

CHINA-Bicolor: IS12711 (1). Caudatum: IS3392 (2). Kafir: IS259 (3)

EGYPT-Aethiopicum: IS18820 (32). Caudatum: IS20857 (4). Durra: IS2870 (5), IS2874 (6). Virgatum: IS18809 (33), IS18810 (34)

ETHIOPIA-Aethiopicum: PI302105 (35). Arundinaceum: PI302118 (36). Caudatum: IS12608 (7). Guinea: IS10018 (8). Kafir: IS8539 (9). Verticilliflorum: IS14583 (37), IS14719 (38)

INDIA-Durra: IS1022 (10). Guinea: IS3955 (11). Kafir IS21963 (12)

IVORY COAST-Arundinaceum: IS18824 (39), IS18830 (40), IS18881 (41)

KENYA-Arundinaceum: IS21340 (42). Bicolor: IS8822 (13). Durra: IS12577 (14). Guinea: IS14541 (15). Verticilliflorum: IS14569 (43), IS14572 (44)

NIGERIA-Arundinaceum: IS18878 (45), IS18879 (46). Bicolor: IS7542 (16). Guinea: IS3614 (17). Kafir: IS2901 (18). ssp. Drummondii: PI186570 (47)

SENEGAL-Bicolor: IS20073 (22). Caudatum: IS19973 (23). Durra: IS3425 (24)

SOUTH AFRICA-Arundinaceum: IS14301 (48). Bicolor: IS1252 (19). Caudatum: IS2377 (20). Guinea: IS3137 (21). Verticilliflorum: IS14277 (49), IS18870 (50)

SUDAN-Aethiopicum: IS14564 (51), IS14485 (52). Bicolor: IS2482 (25). Caudatum: IS12568 (26). Durra: IS12570 (27). Kafir: IS9618 (28). Verticilliflorum: IS18908 (53)

UGANDA-Bicolor: IS2668 (29). Guinea: IS2724 (30). Kafir: IS10400 (31). Verticilliflorum: IS14505 (54). Virgatum: IS18803 (55), IS18806 (56)

Four grams of fresh leaf material for each plant were ground in liquid nitrogen. Total cellular DNA was extracted using the protocol of Saghai-Maroof et al. (1984) with a slightly modified extraction buffer (100 m*M* TRIS-HCl, 2% mixed alkytrimethylammonium bromide, 700 m*M* NaCl, 20 m*M* EDTA, 1% 2-mercaptoethanol, 1% sodium bisulfite, pH 8.0). DNAs were stored in buffer solution (10 m*M* TRIS-HCl, 1 m*M* EDTA pH 8) at 1 μg/μl concentrations at 5 °C.

For analysis of genetic diversity in the nuclear genome, approximately 7 μg of each DNA preparation were digested separately with each of three restriction enzymes (*Eco*RI, *Eco*RV, *Hin*dIII – Bethesda Research Laboratories) according to manufacturer's instructions. For cpDNA analysis, restriction digests were performed with 1–2 μg of total cellular DNA using six restriction enzymes: *Bam*HI, *Dra*I, *Eco*RI, *Sac*II, *Ssp*I, *Stu*I (Bethesda Research Laboratories). DNA digests were electrophoresed in 0.8% agarose gels with a running buffer of 100 m*M* TRIS-acetate, 1 m*M* EDTA (pH 8.1). After the DNA fragments were denatured and neutralized, fragments were transferred to Magna (MSI) nylon membranes without HCl nicking (Maniatis et al. 1982).

Membranes used for studies of diversity in the nuclear genome were probed separately with plasmid clones of 50 low-copy-number nuclear DNA sequences of maize acquired from Brookhaven National Laboratory (BNL), Native Plant Indus-

tries (NPI), and University of Missouri-Columbia (UMC). Cloned inserts were separated from the plasmid in low-melting-point agarose electrophoretic gels, labelled with [^{32}P]dCTP (Feinberg and Vogelstein 1983), and hybridized to the nylon membranes (Helentjaris et al. 1985). Membranes used for chloroplast DNA analysis were probed in a similar fashion using non-overlapping cloned portions of the plastid genome including: λ-5 (14.7 kb), λ-9 (16.4 kb), and λ-11 (18.7 kb) of maize cpDNA (Larrinua et al. 1983). These cpDNA probes were selected because it was known that they would reveal polymorphisms in *Sorghum bicolor* cpDNA (Duvall and Doebley 1990).

Of the 50 low-copy-number nuclear probes 47 each hybridized to a separate, single region of the genome. In most cases, these probes revealed a single band in each individual. In cases where individuals possessed two bands, the bands showed a reduced intensity. Furthermore, each of the two bands occurred independently in other accessions, indicating a heterozygous state in the double-banded individuals. Three nuclear probes (*UMC7, UMC42, UMC88*) appeared to hybridize to two separate regions of the genome, representing two putative loci. Genetic analyses (Whitkus and Doebley, unpublished) support this interpretation. Because the alleles at each locus migrated to different portions of the gel, it was possible to score two loci for these three probes.

Levels of genetic diversity and genetic relationships among the accessions were estimated using nuclear probes. Genetic diversity in the wild and cultivated gene pools of sorghum were estimated from a data set containing allele frequencies averaged over all wild and weedy accessions (ssp. *arundinaceum* and ssp. *drummondii*) and all cultivated accessions (ssp. *bicolor*), respectively. The percentage of loci that were polymorphic (PLP) and total panmictic heterozygosity were calculated from these averages. Phenetic analyses were conducted on the individual accessions in order to elucidate taxonomic relationships. Principal component analysis was conducted using the variance-covariance matrix on all 56 wild and cultivated accessions and separately on the 25 wild and weedy accessions of sorghum. Average linkage cluster analysis was also conducted on all 56 accessions based on a matrix of modified Rogers' distances (Wright 1978).

A nonparametric test, Kendall's Tau (Dietz 1983), of correlation between genetic and geographic distance was conducted for the wild and the cultivated sorghum using both nuclear RFLP and previously published isozyme data. All but 7 (IS3137, IS12570, IS14277, IS14572, IS18824, IS18830, and IS18870) of the geographically identified accessions surveyed for nuclear RFLP diversity in this study had been assayed previously for diversity at 30 isozyme loci (Morden et al. 1989; 1990; Aldrich et al. accepted). For these 49 accessions, RFLP and isozyme allele frequency data were transformed, separately, into matrices of genetic distances (modified Rogers'), keeping the wild and cultivated data sets separate. Seed bank collection information indicates only the country of origin for each accession, so geographic distances between collections were estimated by measuring between the approximate centers of each pair of countries on a map. The Mantel test (Mantel and Valand 1970) was used to determine the significance of the resulting correlations. This is a distribution-free method that tests the significance of departure from randomness in the associations between two distance matrices. Two thousand permutations were run for each of the four tests: cultivar-isozyme, cultivar-RFLP, wild-isozyme, and wild-RFLP.

Results and discussion

The nuclear genome

Genetic diversity. Greater levels of genetic diversity were found in the wild gene pool of sorghum than in the cultivated gene pool. A total of 201 different alleles are distributed among the 53 loci examined, with an average of 3.45 alleles per locus in wild sorghum and 2.28 in cultivated sorghum (Table 2). Estimates of the number of loci polymorphic, based on the 99% criterion, reveal that 87% of the loci in the wild are polymorphic compared to 75% in the cultivars. Seven loci are monomorphic in both wild and cultivated sorghum, and an additional 6 loci are monomorphic only in the cultivated gene pool. Total panmictic heterozygosity is also higher in the wild gene pool (0.39) than in the cultivated gene pool (0.28). These results agree with those based on isozyme analysis (Morden et al. 1990; Aldrich et al. accepted), indicating a loss of genetic diversity in the cultivated sorghum as compared to its wild relatives.

Genetic differences between the wild and cultivated sorghum can be attributed to the existence of both different most common alleles predominating at a locus and low frequency alleles that are unique to one gene pool or the other. An analysis of restriction fragment variation revealed that the wild and cultivated gene pools in sorghum share the same most common allele at 41 (77%) of the loci examined. At the remaining 12 loci, the highest frequency allele is different in the wild and cultivated gene pools. Alleles that are unique to one gene pool, primarily occurring at low frequencies, accounted for much of the remaining differences. Of the 183 alleles identified in wild sorghum 80 (44%) were found only in the wild gene pool with an average frequency of 0.10 and a range of 0.02–0.82. Only 18 (15%) of the 121 alleles found in the cultivated samples were unique to the cultivated gene pool with an average frequency of 0.10 and a range of 0.02–0.26. Thus, a high proportion of the alleles found in the cultivars (85%) is also present in the wild sorghum, as would be expected if cultivated sorghum was derived from ssp. *arundinaceum* over the past 8,000 years.

This study also compared diversity in the nuclear genome of sorghum as resolved by RFLP analysis with that resolved by isozyme analysis (Morden et al. 1989, 1990; Aldrich et al. accepted). Three principal features are noteworthy. First, RFLP analysis reveals a greater number of alleles per locus. In the 49 accessions surveyed for diversity by both methods, RFLP analysis revealed an average of 3.66 alleles per locus whereas isozyme analysis revealed only 2.37. This is particularly remarkable given that 1 plant per accession was analyzed for RFLPs while 3–8 plants per accession were assayed for isozymes. Second, many mid-frequency alleles (0.25–0.75) were found using RFLP analysis (Table 2), whereas isozyme analysis generally resolved a single predominant

Table 2. List of nuclear RFLP alleles observed in *Sorghum* and their frequency within the cultivated and wild gene pools. "*n*" is the number of accessions in which the allele was present among the 56 accessions assayed

Locus-allele	Cultivated		Wild	
	n	Frequency	*n*	Frequency
BNL5.09- A	0	0.000	2	0.080
B	4	0.129	2	0.060
C	4	0.129	2	0.080
D	0	0.000	2	0.060
E	1	0.032	1	0.040
F	22	0.710	16	0.640
G	0	0.000	1	0.040
BNL5.71- A	0	0.000	1	0.020
B	31	1.000	11	0.420
C	0	0.000	1	0.040
D	0	0.000	3	0.120
E	0	0.000	1	0.020
F	0	0.000	2	0.080
G	0	0.000	1	0.020
H	0	0.000	7	0.280
BNL9.44- A	1	0.032	0	0.000
B	3	0.097	0	0.000
C	25	0.806	21	0.840
D	2	0.065	4	0.160
BNL10.24-A	0	0.000	1	0.040
B	0	0.000	21	0.820
C	8	0.258	0	0.000
D	23	0.742	3	0.120
E	0	0.000	1	0.020
BNL14.07-A	11	0.339	11	0.420
B	20	0.645	15	0.580
C	1	0.016	0	0.000
BNL14.28-A	1	0.032	0	0.000
B	1	0.032	0	0.000
C	26	0.834	23	0.920
D	0	0.000	1	0.040
E	3	0.097	1	0.040
NPI350- A	5	0.161	11	0.420
B	26	0.839	9	0.360
C	0	0.000	6	0.220
UMC005- A	0	0.000	3	0.100
B	4	0.129	0	0.000
C	3	0.097	9	0.340
D	24	0.774	12	0.460
E	0	0.000	3	0.100
UMC006- A	0	0.000	1	0.020
B	0	0.000	3	0.100
C	31	1.000	22	0.880
UMC07a- A	0	0.000	1	0.040
B	25	0.806	4	0.160
C	4	0.129	20	0.800
D	2	0.065	0	0.000
UMC07b- A	20	0.645	25	1.000
B	11	0.355	0	0.000
UMC008- A	0	0.000	3	0.100
B	25	0.806	21	0.840
C	6	0.194	2	0.060

Table 2. (continued)

Locus-allele[a]	Cultivated		Wild	
	n	Frequency	*n*	Frequency
UMC012- A	0	0.000	10	0.400
B	8	0.258	9	0.340
C	7	0.226	0	0.000
D	0	0.000	1	0.020
E	13	0.419	6	0.240
F	3	0.097	0	0.000
UMC021- A	20	0.645	13	0.520
B	0	0.000	9	0.340
C	11	0.355	4	0.140
UMC027- A	0	0.000	2	0.060
B	6	0.194	3	0.100
C	15	0.484	12	0.480
D	10	0.323	8	0.300
E	0	0.000	1	0.040
F	0	0.000	1	0.040
UMC029- A	31	1.000	25	1.000
UMC031- A	31	1.000	25	1.000
UMC032- A	0	0.000	1	0.020
B	25	0.806	7	0.280
C	6	0.194	15	0.620
D	0	0.000	2	0.080
UMC037- A	31	1.000	25	1.000
UMC42a- A	0	0.000	1	0.040
B	0	0.000	1	0.040
C	0	0.000	1	0.040
D	0	0.000	5	0.180
E	30	0.968	15	0.600
F	0	0.000	1	0.040
G	1	0.032	2	0.060
UMC42b- A	2	0.065	2	0.060
B	7	0.226	2	0.080
C	0	0.000	3	0.120
D	22	0.710	19	0.740
UMC044- A	1	0.032	10	0.380
B	1	0.032	1	0.040
C	10	0.323	2	0.060
D	0	0.000	4	0.160
E	19	0.613	9	0.360
UMC049- A	0	0.000	4	0.160
B	31	1.000	19	0.760
C	0	0.000	2	0.080
UMC050- A	0	0.000	1	0.040
B	31	1.000	24	0.960
UMC053- A	3	0.097	1	0.040
B	0	0.000	1	0.040
C	20	0.645	21	0.840
D	0	0.000	2	0.060
E	2	0.065	1	0.020
F	1	0.032	0	0.000
G	5	0.161	0	0.000
UMC054- A	31	1.000	25	1.000
UMC055- A	0	0.000	2	0.080
B	0	0.000	6	0.240
C	12	0.387	6	0.240

Table 2. (continued)

Locus-allele[a]	Cultivated		Wild	
	n	Frequency	*n*	Frequency
D	0	0.000	2	0.080
E	17	0.548	8	0.320
F	2	0.065	1	0.040
UMC061- A	2	0.645	7	0.260
B	1	0.032	5	0.200
C	28	0.903	14	0.540
UMC068- A	31	1.000	25	1.000
UMC072- A	18	0.581	24	0.940
B	7	0.226	2	0.060
C	6	0.194	0	0.000
UMC085- A	7	0.226	5	0.180
B	19	0.613	15	0.600
C	5	0.161	6	0.220
UMC086- A	30	0.968	16	0.640
B	1	0.032	9	0.360
UMC88a- A	0	0.000	2	0.080
B	31	1.000	19	0.760
C	0	0.000	1	0.040
D	0	0.000	3	0.120
UMC88b- A	20	0.645	1	0.020
B	0	0.000	1	0.040
C	11	0.355	16	0.620
D	0	0.000	6	0.240
E	0	0.000	1	0.040
F	0	0.000	1	0.040
UMC093- A	1	0.032	2	0.080
B	14	0.452	19	0.760
C	0	0.000	1	0.020
D	16	0.516	4	0.140
UMC094- *A*	30	0.968	22	0.860
B	1	0.032	4	0.140
UMC097- A	31	1.000	25	1.000
UMC106- A	10	0.306	16	0.620
B	22	0.694	9	0.340
C	0	0.000	1	0.040
UMC107- A	0	0.000	1	0.040
B	26	0.839	18	0.720
C	0	0.000	1	0.040
D	5	0.161	1	0.020
E	0	0.000	5	0.180
UMC108- A	0	0.000	3	0.120
B	0	0.000	1	0.040
C	0	0.000	1	0.040
D	31	1.000	18	0.720
E	0	0.000	2	0.080
UMC109- A	0	0.000	1	0.040
B	15	0.484	2	0.060
C	0	0.000	5	0.180
D	16	0.516	16	0.620
E	0	0.000	2	0.080
F	0	0.000	1	0.020

Table 2. (continued)

Locus-allele[a]	Cultivated		Wild	
	n	Frequency	*n*	Frequency
UMC117- A	1	0.032	0	0.000
B	1	0.032	5	0.180
C	28	0.903	15	0.600
D	1	0.032	6	0.220
UMC122- A	16	0.516	12	0.480
B	15	0.484	13	0.520
UMC124- A	3	0.097	2	0.080
B	26	0.839	16	0.620
C	2	0.065	8	0.300
UMC130- A	0	0.000	4	0.160
B	7	0.226	7	0.280
C	24	0.774	14	0.560
UMC132- A	31	1.000	25	1.000
UMC135- A	0	0.000	8	0.300
B	31	1.000	11	0.440
C	0	0.000	3	0.100
D	0	0.000	3	0.100
E	0	0.000	1	0.020
F	0	0.000	1	0.040
UMC136- A	1	0.032	0	0.000
B	24	0.774	7	0.280
C	6	0.194	17	0.680
D	0	0.000	1	0.040
UMC149- A	0	0.000	9	0.340
B	1	0.032	0	0.000
C	17	0.548	14	0.560
D	13	0.419	3	0.100
UMC156- A	0	0.000	1	0.020
B	0	0.000	2	0.060
C	16	0.516	1	0.040
D	15	0.484	22	0.880
UMC167- A	1	0.032	1	0.040
B	0	0.000	1	0.040
C	24	0.774	18	0.720
D	6	0.194	5	0.200
UMC168- A	0	0.000	1	0.040
B	0	0.000	2	0.080
C	0	0.000	1	0.040
D	31	1.000	20	0.780
E	0	0.000	2	0.060
UMC177- A	0	0.000	1	0.020
B	27	0.871	2	0.080
C	1	0.032	0	0.000
D	3	0.097	23	0.900

[a] Locus identification numbers are abbreviated as follows: BNL, Brookhaven National Laboratories; NPI, Native Plant Industries; UMC, University of Missouri-Columbia

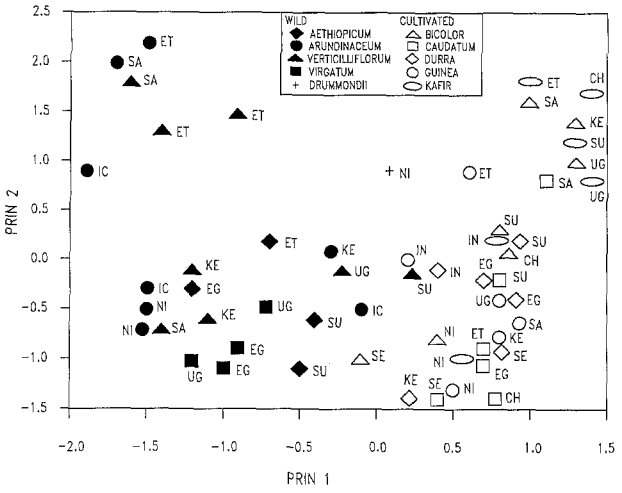

Fig. 1. Graph of the first two components from a principal component analysis based on RFLP allele frequency data from 56 wild and cultivated sorghum accessions. Countries of origin are abbreviated as follows: *CH* China, *EG* Egypt, *ET* Ethiopia, *IN* India, *IC* Ivory Coast, *KE* Kenya, *NI* Nigeria, *SA* South Africa, *SE* Senegal, *SU* Sudan, *UG* Uganda

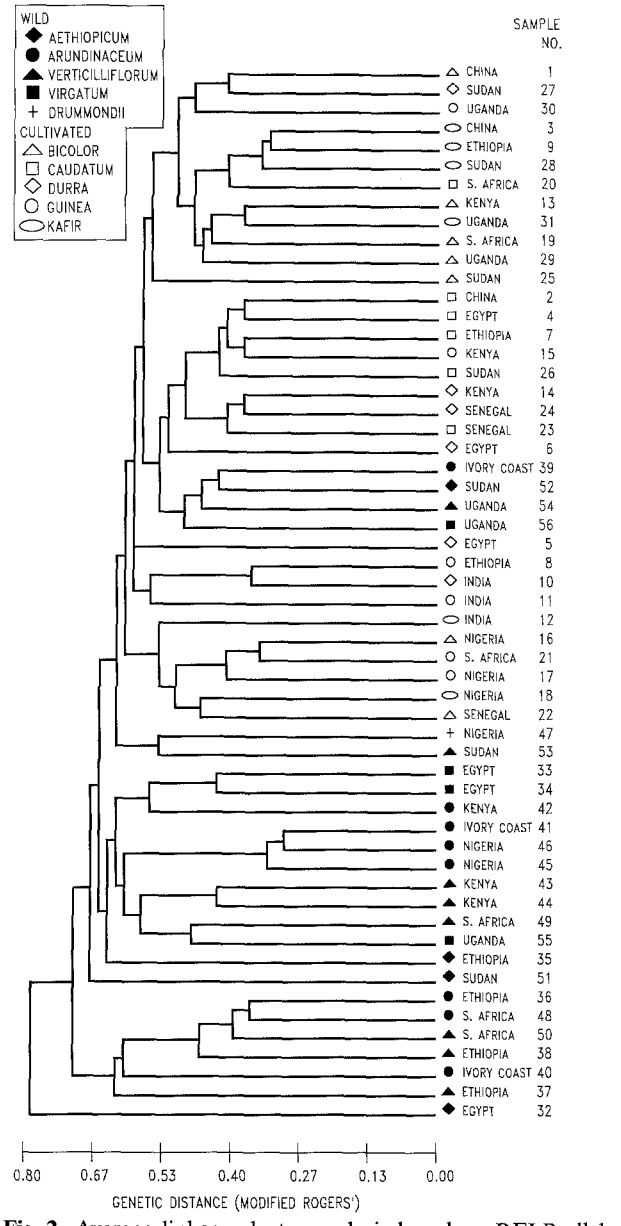

Fig. 2. Average linkage cluster analysis based on RFLP allele frequency data from 56 wild and cultivated sorghum accessions using modified Rogers' distance (Wright 1978)

allele and a few low frequency alleles at most loci. For example, 47 of the 183 (26%) RFLP alleles in ssp. *arundinaceum* have frequencies between 0.25 and 0.75, while none of the 96 isozyme alleles in this subspecies were mid-frequency (Morden et al. 1990). This suggests either that a different evolutionary dynamic (selection-mutation balance) operates on RFLP and isozyme variants or that the common isozyme allele is a synthetic allele consisting of several alleles that do not differ in charge. Third, subspecific differences in sorghum are better resolved using RFLPs. Of the RFLP loci examined, 22.6% (12 loci) carried a different most common allele in the wild and cultivated gene pools, whereas this occurred at only 3.3% (1 locus) of the isozyme loci (Aldrich et al. accepted). Thus, nuclear RFLP analysis using low-copy-number sequence probes detects a greater evenness and richness of allelic diversity at the subspecific level in *Sorghum bicolor* than does isozyme analysis.

Genetic relationships. Genetic relationships among individual accessions of wild and cultivated sorghum were estimated by principal component analysis (Fig. 1) and average linkage cluster analysis (Fig. 2). The wild and cultivated sorghum are separated along the first axis (16% of the variation) of the principal component plot. Most wild collections also fail to cluster with the cultivars in the dendrogram (Fig. 2). This indicates that ssp. *arundinaceum* and ssp. *bicolor* represent fairly distinct germ plasm pools, as previously shown with isozyme analysis (Aldrich et al. accepted).

Some geographic portions of the wild gene pool are genetically more similar to the cultivars than others. Collections of wild sorghum from Uganda, Sudan, and the Ivory Coast exhibit the highest genetic similarity with the cultivars (Figs. 1, 2). Although this latter collection from the Ivory Coast (sample 39) is quite similar to the cultivars, the other 4 wild collections from northwest Africa (Ivory Coast and Nigeria) do not share this relationship and are isolated in the principal component plot at the end opposite the cultivars (Fig. 1). Similarly, wild sorghum from southern Africa is quite distinct from the cultivars. Most accessions from this region occur in the upper left of the PCA plot and at the base of the phenogram. In contrast, the majority of wild collections from northeast Africa (Egypt, Ethiopia, and Sudan) and cen-

tral Africa (Kenya and Uganda) are fairly similar to the cultivars (Fig. 1). Thus, the majority of wild sorghum from both southern and northwest Africa share less resemblance with the cultivars of the same region than does wild sorghum of central and northeast Africa. Therefore, most of the cultivated sorghum was probably domesticated from wild progenitors of the northeast and central African regions. This observation is supported by studies of isozyme variation (Aldrich et al. accepted) and morphology and biogeography (Harlan and Stemler 1976).

Principal component analysis was conducted on all accessions of ssp. *arundinaceum* and ssp. *drummondii* in order to characterize genetic relationships in the wild gene pool (Fig. 3). Accessions generally fail to cluster according to racial status, supporting the conclusion that the wild races are poorly differentiated genetically (Morden et al. 1990). Geographic relationships are more evident with three primary geographic groups resolved in Fig. 3: (1) northeast-central Africa (lower right), (2) northwest Africa (upper right), and (3) South Africa/Ethiopia (lower left). These same three groups were revealed by cluster analysis using isozyme data (Aldrich et al. accepted).

Correlation of geographic and genetic distance. Nonparametric tests of correlation (Table 3) indicate an association between genetic and geographic distance in the wild gene pool when calculated from RFLP data (0.529, $P=0.001$) and isozyme data (0.284, $P=0.067$). Two conclusions may be drawn from these results. First, the higher correlation obtained with RFLPs indicates that RFLPs may be better predictors of geographic relationships than isozymes. This may result from the greater number of available loci and resolvable alleles using RFLP analysis. Second, the genetic constitution is more closely associated with geographic distance in wild sorghum than it is in cultivated sorghum. Causal factors might include (1) obliteration of regional differences in cultivated sorghum by long-distance dispersal through human migration and trade, and (2) a shorter evolutionary history of cultivated sorghum allowing insufficient time for substantial genetic differentiation of its populations along geographic lines.

The chloroplast genome

Levels of diversity. Several mutations in the chloroplast genome of sorghum had been identified previously (Duvall and Doebley 1990). Our study characterizes the frequencies of these mutations in the wild and cultivated gene pools by using a larger sample (56 wild and cultivated accessions). Nine mutations were resolved and identified as restriction site loss/gains (mutations 16, 31, 39, 45, 46, 48, and 66) or as deletions (D1 and D2) (see Duvall and Doebley 1990). A total of six unique chloroplast genome types (I-VI) were identified in the *Sorghum bicolor* gene pool. The most parsimonious tree constructed from the data set requires 11 steps and is depicted (Fig. 4) along with the geographic and racial identities of the collections.

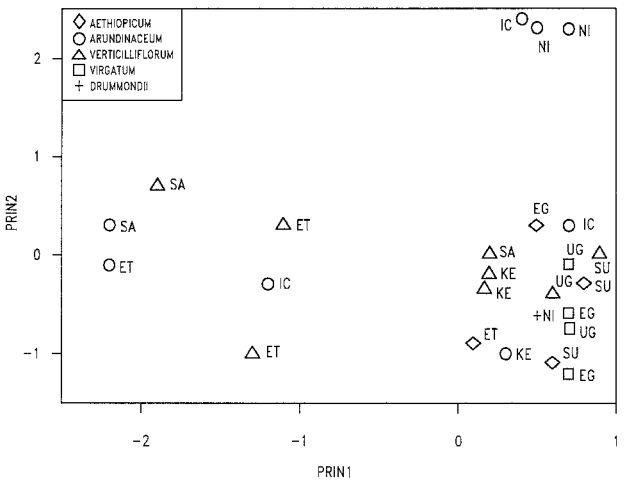

Fig. 3. Graph of the first two components of a principal component analysis based on RFLP allele frequency data from individual accessions of ssp. *arundinaceum* and ssp. *drummondii*. Countries of origin are abbreviated as follows: *EG* Egypt, *ET* Ethiopia, *IC* Ivory Coast, *KE* Kenya, *NI* Nigeria, *SA* South Africa, *SE* Senegal, *SU*, Sudan, *UG* Uganda

Table 3. Correlation coefficients (Kendall's Tau) and significance of association (Mantel Test) between geographic distance and genetic distance (isozyme and RFLP data) in wild and cultivated gene pools of sorghum

Taxon	Method of analysis	Correlation (Kendall's Tau)	*P* (Mantel Test)
Cultivar	Isozyme	0.086	0.3955
Cultivar	RFLP	0.109	0.2350
Wild	Isozyme	0.284	0.0670
Wild	RFLP	0.529	0.0010

Geography and dispersal. Collections from the same country often share the same cpDNA type, although little regional homogeneity is evident in the cladogram overall (Fig. 4). This indicates the occurrence of long-distance seed dispersal in the gene pool. The fact that two distinct cpDNA types (I and IV) are found in Asia and that both types are common in Africa suggests there have been minimally two introductions of sorghum into Asia. It is therefore unlikely that Asian cultivated sorghum has arisen as a monophyletic lineage from the introduction of a single seed type into the region.

Ancestry of domesticated sorghum. The wild and cultivated gene pools are not well-differentiated from one anoth-

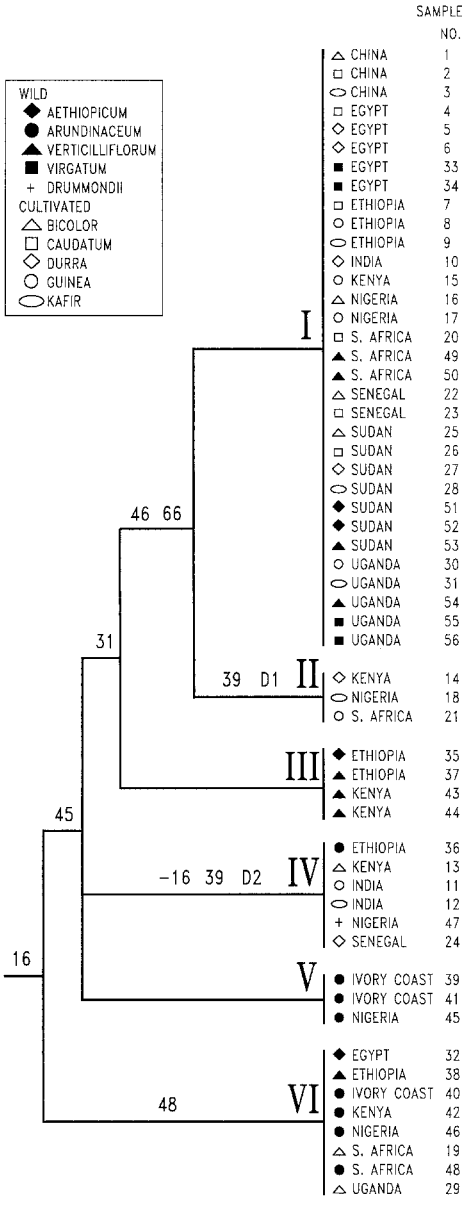

Fig. 4. Phylogenetic tree for the 56 accessions of wild and cultivated sorghum. The tree was constructed from seven cpDNA restriction site mutations and two deletion mutations. Identification numbers for the mutations appear *above* the branches. *Roman numerals* indicate the different cpDNA types that were distinguished in this sample. The tree is rooted by comparing it with an equivalent tree (Duvall and Doebley 1990) produced using fewer accessions but resolving the same mutations and using *Sorghum macrospermum* as and outgroup

er on the basis of cpDNA variation (Fig. 4), which is consistent with the hypothesis that domesticated sorghum arose from ssp. *arundinaceum*. Four cpDNA types (I, II, IV, and VI) were found in the cultivated gene pool and five (I, III-VI) were found in the wild gene pool. Only a single cpDNA type (II) present in the cultivars was not found in the wild sorghum. The mutations that define cpDNA type II may have arisen in the cultivated gene pool after domestication (Fig. 4).

Some geographic portions of the wild gene pool are genetically distinct from the majority of cultivars in their chloroplast genome. CpDNA types V and VI are common in wild sorghum of northwest Africa (Ivory Coast and Nigeria), although uncommon in the cultivated gene pool (Fig. 4). CpDNA type V was found only in the wild sorghum of northwest Africa, while cpDNA type VI is shared by a few other wild collections and only 2 cultivated collections (South Africa and Uganda). In contrast, the majority of cultivated accessions (25 of 30) carry cpDNAs that possess mutation 31 (types I-III), and they share these cpDNA types with wild sorghum from northeast Africa (Egypt, Ethiopia, Sudan), central Africa (Kenya and Uganda), and southern Africa (South Africa). This demonstrates that the chloroplast genome of most cultivars does not originate from the wild sorghum of northwest Africa but instead most likely comes from the wild sorghum of northeast and central Africa, or possibly southern Africa. This outcome would be expected if northwest Africa has contributed little to the germ plasm of domesticated sorghum, as previously suggested (Harlan and Stemler 1976; Aldrich et al. accepted).

Introgression/multiple domestications. Wild and cultivated sorghums share several cpDNA types, i.e., I, IV, and VI (Fig. 4). The presence of these three cpDNA varieties in both wild and cultivated collections could be explained two ways: (1) at least three independent domestication events have occurred, and/or (2) introgression of the chloroplast genome has taken place between the wild and cultivated taxa. It is possible to distinguish between these two explanations by studying the genetic relatedness of wild and cultivated accessions as shown by both nuclear and chloroplast genetic markers. Analyses of the two types of data are likely to be disconcordant if introgression has transferred the chloroplast genome of one taxon into a nuclear genetic background of another.

Chloroplast DNA type I is the most common type for both wild and cultivated sorghum (Fig. 4). Of the 24 collections of ssp. *arundinaceum* 10 carry this type, as do 22 of the 31 collections of ssp. *bicolor*. Information from nuclear RFLP analysis also indicates that these collections are genetically similar to one another (Figs. 1, 2). Thus, this cpDNA is probably shared as a result of common ancestry rather than introgression.

Two other cpDNA types (IV and VI) are more likely shared by both wild and cultivated accessions as a result of introgression. A single wild accession from Ethiopia (sample 36) shares chloroplast type IV with 4 cultivars (samples 11–13 and 24), but is genetically unlike these cultivars in its nuclear genome (Figs. 1, 2). Similarly, the cultivated accessions from South Africa (sample 19) and

Uganda (sample 29) share chloroplast type VI with 6 wild accessions (Fig. 4). However, these 2 cultivated accessions are unlike these wild sorghum in their nuclear genetic profile (Fig. 2). These results suggest that a chloroplast genome from either cultivated or wild sorghum has been transferred via introgression into a nuclear genetic background of the other taxa. Overall, molecular data presented here and elsewhere (Aldrich et al. accepted) indicate that introgression between cultivated and wild sorghum may be quite common. However, it seems also to have a relatively small effect, and the wild and cultivated gene pools remain quite distinct (cf. Figs. 1, 2).

Concluding remarks

The use of both isozyme analysis (Morden et al. 1989, 1990; Aldrich et al. accepted) and RFLP analysis (this paper) for studying diversity in *Sorghum bicolor* provided the opportunity to compare the nature of the variation resolved by the two methods. It is apparent that RFLP analysis resolves a greater richness and evenness of allelic diversity than does isozyme analysis. RFLP markers also appear to be better predictors of geographic proximity, at least in the wild gene pool, and may be a useful alternative as molecular markers when insufficient resolution is attained with isozyme analysis.

The ancestry of domesticated sorghum was also addressed by this study. Our data support the theory that cultivated sorghum was derived from ssp. *arundinaceum* since (1) cultivated sorghum and some forms of ssp. *arundinaceum* are closely related on the basis of nuclear and chloroplast restriction sites, and (2) the cultivars generally contain a subset of the alleles found in ssp. *arundinaceum*. Studies of wild sorghum indicate that the wild races are not well-differentiated genetically from one another, although geographic sections of their range were found to be distinct. The portion of the wild gene pool that is genetically most alike the cultivars is from northeast and central Africa. Wild sorghum races from these regions share a high degree of similarity with the cultivars in both their nuclear and chloroplast genomes, and these regions probably represent the area of primary domestication. In contrast, the wild sorghum races of northwest and southern Africa possess nuclear and chloroplast genotypes that show less similarity to the cultivars.

Introgression and long-distance seed dispersal also appear to have been factors influencing the distribution of diversity in sorghum's primary gene pool. Nuclear and chloroplast markers indicate conflicting genetic relationships between some wild and cultivated collections, suggesting that introgression has occurred between wild and cultivated sorghum. Furthermore, a lack of correlation between genetic and geographic distances in the cultivated gene pool points to evolutionary processes affecting the spatial distribution of genetic diversity in the cultivated gene pool that may be less prevalent in the wild gene pool where genetic and geographic distances were found to be correlated. Causal factors may include long-distance seed dispersal through agricultural trade and a short evolutionary history of the crop allowing insufficient time for differentiation of its populations.

Acknowledgements. Grateful acknowledgement is given to Amy Bacigalupo for her assistance with the chloroplast analysis, to Keith Schertz for supplying the necessary germ plasm, to Lawrence Bogorad, Jack Gardiner, and Scott Wright for supplying the RFLP probes, and to Glenn Furnier, Ed Cushing, and an anonymous reviewer for helpful comments on the manuscript. This research was supported in part by a grant from Pioneer Hi-Bred of Johnston, Iowa.

References

Aldrich PR, Doebley J, Schertz KF, Stec A (submitted) Patterns of allozyme variation in cultivated and wild *Sorghum bicolor*. Theor Appl Genet (1992) (accepted)

Bernatzky R, Tanksley SD (1989) Restriction fragments as molecular markers for germplasm evaluation and utilization. In: Brown AHD, Frankel OH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, New York, pp 353–362

Clegg MT (1990) Molecular diversity in plant populations. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) Plant population genetics, breeding, and genetic resources. Sinauer Assoc, Mass., pp 98–115

Crawford DJ (1990) Plant molecular systematics: macromolecular approaches. John Wiley and Sons, New York

Dietz EJ (1983) Permutation test for association between two distance matrices. Syst Zool 32:21–26

Doebley JF (1989) Isozymic evidence and the evolution of crop plants. In: Soltis DE, Soltis PS (eds) Isozymes in plant biology. Dioscorides Press, Portland, Ore., pp 165–191

Doebley JF, Wendel JF (1989) Application of RFLPs to plant systematics. (Current communications in molecular biology series. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Duvall MR, Doebley JF (1990) Restriction site variation in the chloroplast genome of *Sorghum* (Poaceae). Syst Bot 15:472–480

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

Harlan JR, Stemler ABL (1976) The races of sorghum in Africa. In: Harlan JR, de Wet JMJ, Stemler ABL (eds) Origins of African plant domestication. Mouton Press, The Hague, pp 465–478

Helentjaris T, King G, Slocum M, Siedenstrang C, Wegman S (1985) Restriction fragment polymorphisms as probes for plant diversity and their development as tools for applied plant breeding. Plant Mol Biol 5:109–118

Larrinua IM, Muskavitch EJ, Gubbins EJ, Bogorad L (1983) A detailed restriction endonuclease site map of the *Zea mays* plastid genome. Plant Mol Biol 2:129–140

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

Mantel N, Valand RS (1970) A technique of nonparametric multivariate analysis. Biometrics 26:547–558

Morden CW, Doebley JF, Schertz KF (1989) Allozyme variation in old world races of *Sorghum bicolor* (Poaceae). Am J Bot 76:247–255

Morden CW, Doebley JF, Schertz KF (1990) Allozyme variation among the spontaneous species of *Sorghum* section *Sorghum* (Poaceae). Theor Appl Genet 80:296–304

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Mo Bot Gard 75:1180–1206

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA 81:8014–8018

Wright S (1978) Evolution and the genetics of populations. (Variability within and among natural populations vol 4). University of Chicago Press, Chicago