Phylogeographic Evidence of Crop Neodiversity in Sorghum

L. F. de Alencar Figueiredo,^{*,†} C. Calatayud,^{*} C. Dupuits,[‡] C. Billot,^{*} J.-F. Rami,^{*} D. Brunel,[‡]
X. Perrier,^{*} B. Courtois,^{*} M. Deu^{*} and J.-C. Glaszmann^{*,1}

^{*}Centre de Coopération Internationale en Recherche Agronomique pour le Développement, UMR Développement et Amélioration des Plantes, Montpellier F-34398, France, [†]Universidade Católica de Brasília, Brasília 70790-160, Brazil and [‡]Institut National de la Recherche Agronomique, UR Etude du Polymorphisme des Génomes Végétaux, Commissariat à l'Energie Atomique–Institut de Génétique–Centre National de Génotypage, Evry F-91057, France

Manuscript received January 19, 2008
Accepted for publication April 6, 2008

ABSTRACT

Sorghum has shown the adaptability necessary to sustain its improvement during time and geographical extension despite a genetic foundation constricted by domestication bottlenecks. Initially domesticated in the northeastern part of sub-Saharan Africa several millenia ago, sorghum quickly spread throughout Africa, and to Asia. We performed phylogeographic analysis of sequence diversity for six candidate genes for grain quality (*Shrunken2*, *Brittle2*, *Soluble starch synthase1*, *Waxy*, *Amylose extender1*, and *Opaque2*) in a representative sample of sorghum cultivars. Haplotypes along 1-kb segments appeared little affected by recombination. Sequence similarity enabled clustering of closely related alleles and discrimination of two or three distantly related groups depending on the gene. This scheme indicated that sorghum domestication involved structured founder populations, while confirming a specific status for the guinea margaritifera subrace. Allele rooted genealogy revealed derivation relationships by mutation or, less frequently, by recombination. Comparison of germplasm compartments revealed contrasts between genes. *Sh2*, *Bt2*, and *Sss1* displayed a loss of diversity outside the area of origin of sorghum, whereas *O2* and, to some extent, *Wx* and *Ae1* displayed novel variation, derived from postdomestication mutations. These are likely to have been conserved under the effect of human selection, thus releasing valuable neodiversity whose extent will influence germplasm management strategies.

WHAT is the genetic basis of crop success? Domestication, that is the outcome of a selection process leading to increased adaptation of plants to cultivation and utilization by humans, can be viewed as a long-term selection experiment (GEPTS 2004). It is generally considered as (i) driven by the selection of the most favorable alleles at genes involved in important and visible traits, and (ii) likely accompanied by a significant loss of diversity in the rest of the genome, due to genetic drift by random sampling among preexisting diversity. The genetic architecture of those traits that are part of the “domestication syndrome” is essential in making crop plant selection efficient and crop domestication possible, or not. Yet crop success is also determined by the potential for steady and diversified progress in terms of adaptation to new environments, making the crop able to accompany man in his early migrations and agricultural colonization of new regions. In many cases, continual spontaneous, then breeder-induced, introgression from wild relatives represents a major source of diversity among modern cultivars. Some crops, such as

rice, have probably been domesticated several times (SECOND 1985), thus providing a basis for cultivar diversification through introgression among contrasting early domesticates. Other crop species, however, do not seem to have the same opportunities for a broadening of their genetic basis; recent allopolyploids are in this category, such as bread wheat or groundnut for example. Therefore the process of crop evolution encompasses spontaneous variation that could expand the initial genetic basis and lead to further adaptability.

The rate and magnitude of mutations and their role in the course of domestication is still a matter of conjecture (GEPTS 2004). The recent progress of physiological trait understanding and of molecular investigation methods makes it now possible to focus diversity surveys on individual genes involved in traits of interest in agriculture and to resolve fine sequence variation. Allele sequence polymorphism generally enables the deciphering of allele genealogies, which opens the way to studying processes in time, such as domestication. OLSEN and PURUGGANAN (2002) presented a case of very fruitful phylogeographic analysis of genealogies among the various alleles found at the *Wx* locus, where the “glutinous” phenotype in rice is encoded.

We used diverse molecular approaches to understand diversity in sorghum. Sorghum (*Sorghum bicolor* L. Moench)

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos: EU387138–EU390699.

¹Corresponding author: CIRAD, UMR Développement et Amélioration des Plantes, TA A96/3, Ave. Agropolis, Montpellier F-34398, France. E-mail: glaszmann@cirad.fr

is an annual, predominantly inbreeding, cereal of African origin with five recognized races within cultivated forms (ssp. *bicolor*), namely *bicolor*, *caudatum*, *durra*, *guinea*, and *kafir*, as well as 10 intermediate types. This species has been studied with various types of molecular markers (OLLITRAULT *et al.* 1989; ALDRICH *et al.* 1992; DEU *et al.* 1994, 1995; CUI *et al.* 1995; DE OLIVEIRA *et al.* 1996; MENKIR *et al.* 1997; DJÈ *et al.* 2000; GRENIER *et al.* 2000; CASA *et al.* 2005) and has recently undergone detailed sequence diversity analysis using small representative panels of diverse accessions (HAMBLIN *et al.* 2004, 2006), which enabled investigation of linkage disequilibrium (LD) and patterns of selection among several hundred loci. This resulted in a fine understanding of ecogeographic patterns of variation in sorghum, in particular with reference to its geographic spreading out of the area of origin in the northeastern part of sub-Saharan Africa. For integrated characterization, we have decided to focus on a core sample of 210 accessions of diverse geographic origin from International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) and Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) germplasm banks, established to represent cultivated sorghum landraces from around the world, with sampling on the basis of race, as per the scheme of HARLAN and DE WET (1972), geographical origin, response to day length, and production system. As per DEU *et al.* (2006), RFLP diversity among those accessions led to identification of 10 clusters that appeared to feature combinations of race and geographical origin. They distinguished: guinea accessions from western Africa (cluster 1), guinea *margaritifera* subrace from western Africa (cluster 2), *durra* accessions from central and eastern Africa and from Asia (cluster 3), *bicolor* and *caudatum* accessions from China (cluster 4), *caudatum* accessions from Africa (cluster 5), a group of transplanted *caudatum* and *durra* accessions from Lake Chad region (cluster 6), accessions of the *kafir* race from southern Africa (cluster 7), guinea accessions from southern Africa (cluster 8) and from Asia (cluster 9), and accessions of the *caudatum* race from the African Great Lakes region (cluster 10). The accessions that did not fall into one of those clusters were most frequent in central and eastern Africa and usually classified as intermediate races (*e.g.*, *durra-caudatum*) or *bicolor*. The latter race *bicolor* is recognized as a diverse set of primitive forms that are closer to wild sorghum than the other four races (HARLAN 1995) and relates to the early domesticates prior to other race differentiation. Most of the *bicolor* accessions were intermediate; they did not fall into clear molecular-marker-based clusters, and they commonly displayed rare alleles (DEU *et al.* 1994, 2006). Cluster 2 was more differentiated from the others. It corresponded to the guinea *margaritifera* types, which were also differentiated from the rest of the *S. bicolor* ssp.

bicolor races for cytoplasmic markers (DEU *et al.* 1995). The differences between the other clusters were based on contrasted frequencies of shared alleles, rather than on diagnostic alleles that discriminated one group from all the others.

It is commonly agreed that early domestication of sorghum took place in the northeastern part of the distribution in Africa between Lake Chad and Ethiopia, giving rise to early *bicolor* types (DE WET *et al.* 1976; HARLAN 1995). Then migrations, starting >3 millennia ago, led to the emergence of the guinea types westward and to the *kafir* types southward. The *caudatum* types emerged in the center of origin and later spread in both directions. The *durra* types are predominant in South Asia and northeastern Africa; it is unclear whether they appeared first in Africa (DE WET and HUCKABAY 1967; DOGGETT 1988) or in Asia (HARLAN 1995). Exceptions to the global race geographical pattern, such as the guinea types in South Africa, might be due to more recent germplasm movement or to a more complex origin of diverse types classified under that race (OLLITRAULT 1987; OLLITRAULT *et al.* 1989; DÉGREMONT 1992; DEU *et al.* 1994; FOLKERTSMA *et al.* 2005). In this scheme, varietal groups localized in the most external regions could be considered of secondary origin, such as the guinea types of western Africa (clusters 1 and 2), the *kafir* and the guinea types of southern Africa (clusters 7 and 8), as well as all the forms found in Asia (clusters 4 and 9).

Besides the survey of presumably neutral polymorphisms (DEU *et al.* 2006), we analyzed sequence diversity in portions of six candidate genes involved in starch biosynthesis and protein content regulation. Grain characteristics are likely to be subject to natural selection, since grain reserves, such as carbohydrates or proteins, play a key role in life cycles, as well as to human selection during domestication, given their impact on “grain quality” in human uses. The patterns of diversity displayed by such candidate genes are thus expected to exhibit traces of selection. In this article we describe sequence diversity at those loci and we apply the unique properties of that type of data to further investigate sorghum evolution.

MATERIALS AND METHODS

Plant material: We analyzed 194 accessions representative of the diversity of cultivated sorghum that were part of the core sample of 210 varieties described above. The accession list with indication of the race, country of origin, and RFLP-based cluster can be found in supplemental Table S1.

The distribution of our material is shown in Table 1 and is schematized in Figure 1, together with current hypotheses on sorghum domestication. The regions are considered from west to east and north to south and include extreme West Africa (A), northcentral Africa (B), northeast Africa (C), southeast Africa (D), southern Africa (E), south Asia (SA), East Asia (EA), and others (O). We emphasized the comparison between the margins represented by clusters 1 and 2 in

TABLE 1
Distribution of accessions, according to region of origin and classification into RFLP clusters (DEU *et al.* 2006)

Region ^a	Cluster ^b										Unclustered accessions
	2	1	6	3	5	10	8	7	9	4	
Extreme West Africa (A)	10	17		1			1				3
Northcentral Africa (B)	3	5	11	2	11						6
Northeast Africa (C)				8	4	2					7
Southeast Africa (D)				1	2	5					4
Southern Africa (E)				1	6		12	27			4
South Asia (SA)				10					7		2
East Asia (EA)									1	10	3
Others				1	1						6

^a Countries included in the regions. Extreme West Africa: Benin, Burkina Faso, Gambia, Ghana, Mali, Senegal, Sierra Leone; northcentral Africa: Cameroon, Central African Republic, Chad, Niger, Nigeria; northeast Africa: Ethiopia, Somalia, Sudan, Yemen; southeast Africa: Burundi, Democratic Republic of Congo, Kenya, Rwanda, Tanzania, Uganda; southern Africa: Botswana, Lesotho, Malawi, Republic of South Africa, Swaziland, Zambia, Zimbabwe; south Asia: India, Nepal, Sri Lanka; East Asia: China, Korea; others: Algeria, Turkey, USA.

^b Nomenclature according to DEU *et al.* (2006): 2, guinea margaritifera accessions from western Africa; 1, guinea accessions from western Africa; 6, transplanted caudatum accessions from Lake Chad region; 3, durra accessions from central and eastern Africa and from Asia; 5, caudatum accessions from Africa; 10, caudatum and durra accessions from the African Great Lakes region; 8, guinea accessions from southern Africa; 7, kafir accessions from southern Africa; 9, guinea accessions from Asia; 4, bicolor and caudatum accessions from China.

West Africa, clusters 7 and 8 in southern Africa, clusters 4 and 9 in Asia, which we qualified as *secondary* units in the crop history, and the center of the distribution represented by clusters 6, 3, 5, and 10, which we qualified as *primary* units.

Genes and primer design: Six genes known to be important in the genetic control of grain quality in cereals, notably in maize, were selected. They were directly involved either in starch synthesis [*Shrunken2* (*Sh2*), *Brittle2* (*Bt2*), *Soluble starch synthase1* (*Sss1*), *Amylose extender1* (*Ae1*), and *Waxy* (*Wx*)] or as transcriptional activators controlling endosperm protein storage genes [*Opaque2* (*O2*); PIROVANO *et al.* 1994]. *Sh2* and *Bt2* encode, respectively, for the large and small subunits of a major enzyme involved in endosperm starch biosynthesis,

ADP-glucose pyrophosphorylase (AGPase) (SCHULTZ and JUVIK 2004). AGPase catalyzes the first step of starch synthesis in plants, *i.e.*, the production of ADP-glucose. ADPG is further used to build amylose, a linear glucosyl chain and amylopectin, a highly branched glucan, both constituting starch. Amylose synthesis is catalysed by granule-bound starch synthase, encoded by the *Waxy* (*Wx*) locus. Amylopectin synthesis is catalysed by starch branching enzymes (SBEI, SBEIIa, and SBEIIb encoded by *Sbe2b*, now called *Ae1* gene) and debranching enzymes (DBE) (WILSON *et al.* 2004 and references herein).

The sequence accessions from sorghum used for primer design were AF488412 (*Wx*), AF010283 (*Sh2*), CD426561 (*Bt2*), AY304540 (*Ae1*), AF168786 (*Sss1*), and X71636 (*O2*). For *Bt2*, *Ae1*, and *Sss1*, for which only the cDNA sequences from sorghum were available, we also used sequences from maize (AF334959 for *Bt2* and AF072725 for *Ae1*) and rice (AB026295 for *Sss1*). Primers were designed using the PRIMER3 program (<http://fokker.wi.mit.edu/primer3/input.htm>). The primers are listed in supplemental Table S2. Several segments were targeted per gene. For *Wx* and *O2*, primers were designed to cover the major part of the gene and for *O2* of the promoter. For the other genes, primers were selected to amplify two segments of at least 0.5 kb in size and separated by at least 1.5 kb (Figure 2). The segment positions were chosen preferentially within coding regions, in the protein motif when possible, with maximum intragene distance between segments to enable analysis of linkage disequilibrium.

PCR and DNA sequencing: Genomic DNA was extracted from fresh leaves harvested from a single 3-week seedling per accession following a cetyltrimethylammonium bromide (CTAB) protocol previously described (DEU *et al.* 1995). SP6 (5'-GATTAGGTGACACTATAG-3') and T7 (5'-TAATACGACTCACTATAGGGC-3') tails were added to the 5' ends of primers to facilitate direct sequencing of the PCR products. DNA amplifications were performed in 50 µl containing 25 ng of genomic DNA, 0.2 µM of each primer, 2 mM MgCl₂, 0.2 mM of dNTP, and 1 unit Taq DNA polymerase. Reactions followed the following cycling conditions: 94° for 4 min; 10 cycles [30 sec

FIGURE 1.—Geographical distribution of RFLP-based clusters (1–10) of sorghum varieties in relation to the different regions of Africa (as in Table 1). The black area is a schematic of the area of initial sorghum domestication and the arrows show the main migrations according to HARLAN (1995). The global pattern opposes primary units (clusters 3, 5, 6, and 10) to three secondary units (clusters 1 and 2, clusters 7 and 8, and clusters 4 and 9).

Sh2 *Bt2* *SssI* *AeI* *Wx* *O2*

FIGURE 2.—Gene representation showing exons (shaded) and sequenced segments (solid). *Sh2*, *Shrunken2* (sorghum sequence); *Bt2*, *Brittle2* (maize sequence); *SssI*, *Soluble starch synthaseI* (rice sequence); *AeI*, *Amylose extender1* (maize sequence); *Wx*, *Waxy* (sorghum sequence); *O2*, *Opaque2* (sorghum sequence). The positions of the beginning and the end of the sequenced segments are in base pairs and the number of the fragment are given at the bottom of the representation.

at 94°, 60 sec at melting temperature (TM) + 5°, 60 sec at 72°, the annealing temperature was reduced by 0.5°/cycle; 25 cycles (30 sec at 94°, 60 sec at TM, 60 sec at 72°); and a final extension step of 8 min at 72°. The annealing temperature was 55° except for *Wx* segment 1 (60°) and *Wx* segment 2 (65°).

Sequencing from PCR products was performed using an Applied Biosystems Prism 3100 DNA analyzer (Applied Biosystems, Foster City, CA) in only one direction, by Centre National de Genotypage, Evry, France (<http://www.cng.fr>) for *O2*, and by GATC Biotech (<http://www.gatc-biotech.com>) and Genome Express (<http://www.cogenics.com/>) for the other genes. Sequence data have been deposited with the EMBL/GenBank Data Libraries under accession nos: EU388245–EU388607 for *Sh2*, EU388985–EU389363 for *Bt2*, EU388608–EU388984 for *SssI*, EU387881–EU388244 for *AeI*, EU387138–387880 for *Wx*, and EU389364–EU390699 for *O2*.

Data analysis: Sequence quality control and clipping were performed using Sequencher 4.0 (Gene Codes, Ann Arbor, MI) with minimum Phred scores set to 20. All sequences, including the reference sequence, were aligned using Sequencher and base substitution among sequences single nucleotide polymorphisms (SNPs) and insertion or deletion polymorphisms (IDPs) were detected. Artemis version 7 (RUTHERFORD *et al.*

2000) was used to verify the position of splicing sites and determine the correct reading frame. Conservative and non-conservative amino acid substitutions were defined by the Blossum matrix (HENIKOFF and HENIKOFF 1992) and by calculating hydrophobicity (KYTE and DOOLITTLE 1982).

Only accessions with sequences for all segments of a given gene were kept for further data analyses. Thus, global SNP and IDP were assessed on slightly different samples for each gene: 184 accessions for *Bt2*, 166 for *SssI*, 153 for *Sh2*, 154 for *AeI*, and 129 and 146 for *Wx* and *O2*, respectively, the two genes with the largest number of segments. Further use was made of a core set of 53 accessions (supplemental Table S1) with no missing data across all genes and well distributed within the pattern of diversity revealed by RFLPs (DEU *et al.* 2006).

Nucleotide diversity was estimated using θ , Watterson's estimator of $4N_e\mu$ per base pair, on the basis of the number of segregating sites (where N_e is the effective population size and μ the mutation rate) (WATTERSON 1975), and π , the average number of pairwise differences per nucleotide between sequences (TAJIMA 1983). Tajima's *D* test (TAJIMA 1989) was used to test for deviations from neutral mutation-drift equilibrium. All those estimators were calculated using DnaSP 4.10 (ROZAS *et al.* 2003). Analyses were conducted separately

TABLE 2
Nucleotide diversity among a subset of sorghum accessions with complete data on a gene-per-gene basis,
for the six genes under survey

Species	Gene	N	Portion	Polymorphism				Summary statistics ^a		
				Length (bp)	No. IDP	No. polymorphic sites ^b	SNP frequency	Nucleotide diversity		Tajima's <i>D</i> test ^c
								π (10 ⁻³)	θ (10 ⁻³)	
Sorghum	<i>Sh2</i>	153	NC	849	2	19		0.98	4.16	-2.385**
			C	482	0	2s:1c		0.78	1.64	-1.063
			Total	1331	2	22	1/60	0.91	3.31	-2.314**
Maize (WHITT <i>et al.</i> 2002)	<i>Sh2</i>	30	—	6754	NA	NA	NA	5.0 ^d	NA	NA
Maize (MANICACCI <i>et al.</i> 2006)	<i>Sh2</i>	50	—	4669	18	69	1/68	3.6 ^d	3.9	-0.596
Sorghum	<i>Bt2</i>	184	NC	472	2	6		0.74	1.41	-0.974
			C	536	0	0		0.00	0.00	—
			Total	1008	2	6	1/168	0.34	0.66	-0.974
Maize (WHITT <i>et al.</i> 2002)	<i>Bt2</i>	30	—	6098	NA	NA	NA	2.3 ^d	NA	NA
Sorghum	<i>SssI</i>	166	NC	996	3	7		1.10	1.58	-0.790
			C	529	0	2s		0.47	0.85	-0.792
			Total	1525	3	9	1/169	0.93	1.34	-0.834
Sorghum	<i>AeI</i>	154	NC	955	8	14		4.42	2.41	2.391*
			C	422	0	5s:3c:3nc		3.83	1.91	2.071*
			Total	1377	8	25	1/55	4.26	2.27	2.625*
Maize (WHITT <i>et al.</i> 2002)	<i>AeI</i>	30	—	6781	NA	NA	NA	2.9 ^d	NA	NS
Sorghum	<i>Wx</i>	129	NC	1162	5	21		6.57	3.63	2.532*
			C	1105	0	5s:1c:1nc		1.87	1.34	1.006
			Total	2267	5	28	1/81	4.21	2.48	2.268*
Maize (WHITT <i>et al.</i> 2002)	<i>Wx</i>	30	—	2978	NA	NA	NA	11.5 ^d	NA	NA
Barley (KILIAN <i>et al.</i> 2006)	<i>Wx</i>	45	—	816	NA	10	1/82	3.32 ^e	7.87 ^d	-0.95 NS
Rice (OLSEN <i>et al.</i> 2006)	<i>Wx</i>	73	—	5300	NA	101	1/52	5.6 ^d	na	0.499
Sorghum	<i>O2</i>	146	NC	2707	9	38		1.92	3.11	-1.271
			C	1064	0	1s:7c:4nc:1ns		1.00	2.41	-1.683
			Total	3771	9	51	1/74	1.65	2.91	-1.461
Maize (HENRY and DAMERVAL 1997)	<i>O2</i>	21	—	1800	26	258	1/11	NA	52.2 ^d	NA
Maize (HENRY <i>et al.</i> 2005)	<i>O2</i>	33	—	1976	72	106	1/18	23.3 ^d	27.1 ^d	-0.657

N, number of accessions; C, coding portion; NC, noncoding portion; NA, not available. **P* < 0.05; ***P* < 0.01.

^a On a subset of 53 representative accessions with complete data for all gene segments.

^b s, synonymous; c, nonsynonymous conservative; nc, nonsynonymous nonconservative; ns, nonsense (CAG/TAG).

^c NS, nonsignificant.

^d Based on silent sites.

^e All sites considered.

for coding and noncoding regions and for the entire concatenated sequence of all segments of the same gene.

SNPs and IDPs were extracted from the sequences and further analyzed using Tassel version 1.9.5 (<http://www.maizegenetics.net/>). Both SNPs and IDPs were considered in haplotype analysis; however, singletons were excluded to prevent excessive impact of sequencing errors. Haplotype diversity was analyzed for each segment separately and for each gene separately (concatenating all segments). Relationships between haplotypes were analyzed according to the median-joining network method described by BANDELT *et al.* (1995, 1999) with Network 4.2 software (<http://www.fluxus-technology.com/sharenet.htm>), using the median-joining option, an equal weight for all sites, and an ϵ -parameter of 0, which

retained only frequent haplotypes for describing cycles in the network, *i.e.*, a representation of alternate paths between two haplotypes due to homoplasies (recurrent mutations, recombination, but also sequence errors).

RESULTS

Sequence diversity: The polymorphism revealed among all the accessions is presented in detail for each gene in a file entitled *sorghum_SNP.pdf* downloadable from <http://tropgenedb.cirad.fr/sorghum/>. Its global features are described in Table 2. A total of

TABLE 3
Haplotype organization for the six genes under survey

H	<i>Sh2</i>			<i>Bt2</i>			<i>SssI</i>			<i>Ael</i>			<i>Wx</i>			<i>O2</i>
	Frequency (%)	H		Frequency (%)	H		Frequency (%)	H		Frequency (%)	H		Frequency (%)	H		Frequency (%)
		S1	S2		S1	S2		S1	S2		S1	S2		S1	S2	
H1	71.9	a	a	63.6	a	a	82.6	a	a	24.0	a	a	27.9	a	a	20.5
H2	9.2	b	a	7.1	b	a	1.2	b	a	8.4	a	b	0.8	a	b	24.0
H3	7.2	a	b	14.1	b	b	7.8	c	b	15.6	b	b	7.8	a	c	17.1
H4	1.3	c	a	1.1	a	b	5.4	d	b	5.2	b	c	1.6	a	d	1.4
H5	2.6	d	a	0.5	d	b	1.8	c	c	1.3	b	d	2.3	a	e	1.4
H6	7.8	d	c	0.5	d	a	1.2	e	d	1.9	c	e	1.6	b	e	3.4
H7				3.3	c	a				20.8	c	d	4.7	b	a	1.4
H8				0.5	e	c				0.6	c	c	0.8	c	c	1.4
H9				5.4	f	c				14.9	d	c	0.8	c	a	1.4
H10				2.7	g	d				0.6	d	b	3.9	c	d	6.8
H11				0.5	g	a				0.6	e	c	0.8	c	f	3.4
H12				0.5	f	a				5.8	f	f	0.8	b	g	2.1
H13													0.8	b	h	2.7
H14													22.5	b	f	0.7
H15													0.8	c	i	1.4
H16													0.8	b	i	1.4
H17													0.8	a	i	2.1
H18													0.8	a	j	2.1
H19													7.8	a	k	0.7
H20													10.9	a	f	2.1
H21													0.8	a	g	2.7
H22													0.8	a	h	

There are two separate segments per genes (S1 and S2) except for *O2* (7 concatenated segments). The haplotypes observed for each segment were named alphabetically (a–k), as shown in supplemental Tables S3, S5, S7, S9 and S11.

170 polymorphisms, including 141 SNPs and 29 IDPs, were recorded within a total of 11.3 kbp scored. That resulted in an average of one SNP every 80 bp. The IDPs (14 involved 1 bp only) were confined to introns, promoter or 3'-UTR regions and accounted for 17% of all polymorphism. Only 36 of the SNPs were located in coding regions; among them 9 were nonconservative and 1 was a nonsense mutation. *O2* was the gene with the highest proportion of nonsynonymous SNPs.

For the sake of comparison among genes and with other studies, we computed global parameters for a subset of 53 diverse accessions (supplemental Table S1) that had complete data for all genes. The frequency of polymorphic sites per gene, considering each IDP as a unique event, varied between 0.3% (*Bt2*) and 1.4% (*Sh2*). It was generally twice as high for noncoding regions as for coding regions, except for *Bt2* for which all polymorphisms were located in noncoding regions. The amount of diversity, measured by π , ranged from 0.34/kbp for *Bt2* to 4.26/kbp for *Ael*. The frequency of

polymorphic sites, measured by θ , was highest for *Sh2* (3.31/kbp), due to the large number of singletons for that gene, intermediate for *Ael*, *Wx*, and *O2*, and low for *Bt2* and *SssI* (0.66 and 1.34/kbp, respectively). That resulted in contrasting Tajima's *D* estimates, with *Sh2* showing a significant negative *D* value and *Ael* and *Wx* showing significant positive *D* values.

Haplotype structure and distribution: *Shrunken2* (*Sh2*): A total of six haplotypes were observed for *Sh2* (Table 3; supplemental Tables S3 and S4 for details). Segment 1 displayed two groups of haplotypes, one comprising a predominant type (a, H1 + H3) and two unrelated minor types (b in H2, with one difference with a, and c in H4, with 2 differences), the other comprising a single type (d in H5 + H6). Segment 2 displayed two groups of haplotypes, differentiated by at least 10 of the 11 polymorphic sites: with one predominant (a in H1 + H2 + H4 + H5, b in H3 with one difference with a) and one minority (c in H6, 8%). For both segments, it was possible to represent the haplotype relationships in a

FIGURE 3.—Haplotype network for each of the six genes under study, showing the distribution of the various sorghum accessions classified according to their origin, as described in Figure 1. Circled areas are proportional to haplotype frequencies. White is used for accessions belonging to clusters 3, 5, 6, and 10. Red is used for accessions belonging to clusters 1 and 2, all from western Africa. Yellow is used for accessions belonging to clusters 7 and 8, all but one from southern Africa. Blue is used for accessions belonging to clusters 9 and 4, all from Asia. Gray is used for accessions that are unclustered and/or from other regions of the world.

simple manner (Figure 3). The pattern of association between the two segments was very strong, highlighting two groups, of uneven frequency, of full-length haplotypes and a minor recombinant-like type (H5) found in only four accessions (2.6%). On that basis, it can be concluded that there is strong LD along the whole gene and a simple interpretation can be attempted to describe the genealogical relationships among the various alleles. The minority group, and to some extent the recombinant haplotype, were typical of cluster 2 of the guinea margaritifera subrace in West Africa, but were occasionally found in other backgrounds (supplemental Table S4).

Brittle2 (Bt2): A total of 12 haplotypes were recorded for *Bt2* (Table 3; supplemental Table S5 and S6 for details), of which six were very rare (<2%). Both segments displayed little differentiation, with almost continuous variation and random combination of polymorphisms. There was only a weak LD along the gene and no simple inter-haplotype genealogical hypothesis was possible (Figure 3). Here again, the guinea margaritifera subrace cluster was highlighted; it encompassed all the accessions that displayed haplotype H9 (supplemental Table S6).

Soluble starch synthase1 (SssI): A total of six haplotypes were observed for *SssI* (Table 3; supplemental Tables S7 and S8 for details). Segment 1 displayed two groups of haplotypes differentiated by at least four of the seven polymorphic sites, one made up of the predominant type and one minor type (a and b in H1 and H2), the other made up of three types (c, d, and e in H3–H6). Segment 2 also displayed two groups of haplotypes, differentiated by at least three of the five polymorphic sites, one made up of the predominant type (a), the other made up of three types (b, c, and d in H3–H6). The pattern of association between the two segments was very strong, highlighting strong LD along the gene. A simple interpretation can be attempted to describe the genealogical relationships among the various alleles (Figure 3). One minority haplotype (H3) was typical of cluster 2 of the guinea margaritifera subrace in West Africa but was occasionally found in other backgrounds (supplemental Table S8). The predominant group was found in all other varietal clusters and regions.

Amylose extender1 (Ae1): A total of 12 haplotypes were observed for *Ae1* (Table 3; supplemental Tables S9 and S10 for details). Segment 1 displayed three groups of haplotypes, one made up of a predominant type (a in H1 + H2), another made up of one haplotype (f in H12), the third one made up of four related types differentiated at one to three polymorphic sites (b in H3 + H4 + H5, e in H11, c in H6 + H7 + H8, and d in H9 + H10). Within the latter group, simple genealogical relationships can be inferred between the various haplotypes (Figure 3). Segment 2 displayed six haplotypes distributed in two groups differentiated by at least eight of the 13 polymorphic sites, contrasting H1 and

H12 (a and f) on one side to H2–H11 (b–e) on the other; it was also possible to consider H12 as a recombinant type between a in H1 and b in (H2 + H3 + H10). Among H2–H11, four types existed (b, c, d, and e), among which simple genealogical hypotheses can be formulated. Segments 1 and 2 displayed strong associations; only H2 (13 accessions, 8.4%) was an exception to association between segment-specific groups, and may thus be interpreted as the result of recombination. Considering the rarer variants within each group, H12 was the result of markedly differentiated haplotypes on both segments. It is noteworthy that the most recent haplotypes based on our simple genealogical hypotheses, *i.e.*, H9 + H10 for segment 1 and H6 for segment 2, were most frequent in secondary clusters (supplemental Table S10).

Waxy (Wx): A total of 22 haplotypes were recorded for *Wx* (Table 3; supplemental Tables S11 and S12 for details). The main feature of *Wx* gene diversity was the existence of two gene portions that displayed total or near total internal LD and little or no LD between portions. The first portion consisted of segment 1; it featured one predominant haplotype (a) and two other haplotypes (b and c) that differed from it by 13 or 14 polymorphic sites. The second portion consisted of segment 2, which displayed two groups of haplotypes that were differentiated from one another by at least 5 of the 8 recorded polymorphic sites, as well as segment 3 and segment 4, which displayed a series of infrequent (<10%) haplotypes differentiated by 1–4 polymorphic sites from the predominant haplotype (Figure 3). Many of those minor haplotypes seemed confined to secondary varietal groups or secondary regions, suggesting that they might have been derived from recent mutations (supplemental Table S12). Interestingly, there was no LD between that portion and the first portion. This suggested that intragenic recombination had occurred at the same pace as point mutations and had had a significant impact on allele diversification in the course of domestication.

Opaque2 (O2): The *O2* gene displayed a total of 21 haplotypes (Table 3; supplemental Tables S13 and S14 for details). They fell into two groups of haplotypes characterized by a uniformly distributed contrast over the full 4 kb length: H1–H19 on one side and H20 + H21 on the other differed by at least 31 of the 55 polymorphic sites. Within a smaller magnitude of variation, the first group, which was also markedly predominant, displayed a remarkable tendency toward multidirectional radiation (Figure 3). H1, the haplotype that corresponded to the root, had a global frequency of 20%; the various branches displayed between one and up to seven mutations. Interestingly, most of the branches expanded in a specific geographical direction and the most recent alleles tended to be specific to secondary varietal clusters and regions: haplotypes H11, H12, and H13 in cluster 2 (guinea margaritifera) and H10 in

TABLE 4

Haplotype number and type, estimated number of mutations, and recombinations per kb observed in the sample of 194 accessions for the six genes under survey

	<i>Sh2</i>	<i>Bt2</i>	<i>SssI</i>	<i>Ae1</i>	<i>Wx</i>	<i>O2</i>
Sequence length (kb)	1.331	1.008	1.525	1.377	2.267	3.771
Total length covered (kb)	3.952	2.313 ^a	3.933 ^b	16.474 ^a	3.062	4.251
Number of haplotypes	6	12	6	12	22	21
Number of major haplotypes	1	2	1	4	3	3
Number of minor haplotypes ^c	5	10	5	8	19	18
Number of mutations/kb ^d	16.5	6.0	5.9	18.2	12.3	13.5
Number of indels/kb ^d	1.5	2.0	2.0	5.8	2.2	2.4
Estimated minimal number of recombination ^d	1	0	0	2	2	0
Estimated minimal number of recombination events/total length covered (kb)	0.25	0	0	0.12	0.65	0

^aLength covered based on the maize sequence accession.

^bLength covered based on the rice sequence accession.

^cHaplotype with frequency <10%

^dBased on concatenated sequences, using DnaSP (ROZAS *et al.* 2003).

cluster 1 in western Africa; H14–H17 in cluster 8 and H2 and H3 in cluster 7 in southern Africa (supplemental Table S14); H4 in two unclustered accessions from southern Africa.

DISCUSSION

Sequence diversity in sorghum: Our study was based on 1.7 Mbp sequence data enabling a comparison between 129 to 184 accessions of sorghum for 11.3 kbp covering six genes and 1.0–3.8 kbp per gene. The best reference for comparison in sorghum is provided by HAMBLIN *et al.* (2004), who described 27 sorghum accessions, including three wild ones, through 95 loci and a total of 29.2 kbp. Our sample of 53 accessions used to quantify variation across the genes was close to double in size, and gave more weight to western Africa, southern Africa, central Africa, and Asia, and less to eastern Africa. Whereas our study revealed a higher frequency of polymorphisms (one SNP every 80 bp, compared to one SNP every 123 bp), probably largely due to sample size, the average diversity (π) was remarkably similar in both studies (2.05 *vs.* 2.25/kbp).

Five of the genes we studied in sorghum (*Sh2*, *Bt2*, *Ae1*, *Wx*, and *O2*) have been studied in maize by HENRY and DAMERVAL (1997), WHITT *et al.* (2002), HENRY *et al.* (2005), and MANICACCI *et al.* (2007). Compared to that crop, the diversity appeared slightly higher in sorghum for *Ae1*, but lower for *Wx*, *Bt2*, *Sh2* and *O2* (Table 2). The most accurate comparison was possible for *Sh2* and *O2* where most of the coding sequence was studied (HENRY *et al.* 2005; MANICACCI *et al.* 2007); the diversity in sorghum was found to be three and seven times lower than in cultivated maize. The *Wx* gene enabled a comparison with cultivated barley and with Asian domesticated rice (KILIAN *et al.* 2006; OLSEN *et al.* 2006); the diversity in sorghum (4.21/kbp) was close to that in

barley (3.32/kbp) and in rice (5.6/kbp), despite their well-documented high intraspecific diversity, particularly the indica-japonica contrast in rice (Table 2).

A remarkable feature in sorghum was the existence of strongly differentiated haplotypes at most genes (Table 3). The diversity level for a particular gene was very much related to the relative frequency of the most contrasting haplotypes. For three of the six loci (*Sh2*, *Bt2*, and *SssI*), the three most common haplotypes bore over 80% of allelic diversity.

On the basis of Tajima's *D* test, three loci (*Sh2*, *Ae1* and *Wx*) exhibited significant deviations from neutral mutation-drift equilibrium (Table 2). That might be explained by population size variations, including population expansions or bottlenecks, as well as selection. A negative *D* value (*Sh2*) revealed an excess of very rare alleles, which might indicate a purifying selection pattern, while positive values (*Ae1* and *Wx*) might be explained by balancing selection. Pushing further this interpretation, the fact that coding regions appeared less affected than noncoding regions for *Sh2* and *Wx* (Table 2) might be related to background selection or hitchhiking (OTTO, 2000). Those three genes have been shown to be targets of selection in other cereals, in particular *Sh2* in maize (WHITT *et al.* 2002; MANICACCI *et al.* 2007), *Ae1* in maize (WHITT *et al.* 2002; WILSON *et al.* 2004) and *Wx* in rice (YAMANAKA *et al.* 2004; OLSEN *et al.* 2006). Explanations other than selection might yet be proposed, especially in a cultivated species such as sorghum, which displays clusters of contrasting haplotypes. Ancestral population structure, multiple domestications, or introgression from wild relatives have recently been highlighted by HAMBLIN *et al.* (2006), also in sorghum, as potentially confusing phenomena when the objective is to detect an episode of selection.

The level of within-gene LD was remarkably variable between genes. There was near-complete LD over >4 kb

of the whole-gene length in *O2*, whereas *Wx* displayed LD breakage between segments 1 and 2 (Table 4). If some haplotypes in *Wx* portion 2 (segments 2 + 3 + 4) derived from mutations that occurred after the start of domestication, this break in LD highlights intensive recombination within the 244 bp that separate sites within intron 6 from those located in intron 7 (supplemental Table S11). There was also strong LD across *Ae1* (which represents over 16 kb in maize), induced by the existence of markedly differentiated haplotypes that apparently seldom recombined (Tables 3 and 4). However, for the same gene, there were also minor variants, which probably appeared more recently and were not in LD. This illustrates that recombination might occur frequently, but that it affects only some combinations of alleles. Many other allele combinations may be lacking among the heterozygous forms, as a result of a departure from random mating. Such a bias can lead to conservation of major haplotype groups and maintain strong LD among those SNPs that discriminate them. In that case LD relates to population structure. The heterogeneity of LD among genes is well documented for several species (GUPTA *et al.* 2005); our study highlighted *Wx* as a spot with high recombination activity, as has been observed in maize and rice (OKAGAKI and WEIL 1997; INUKAI *et al.* 2000), but it also highlighted complete LD in *O2* whereas that gene displays traces of intensive recombination activity in maize (HENRY and DAMERVAL 1997).

Insights into sorghum domestication: LD patterns and predominant haplotypes can help trace crop history and crop domestication. Across the six genes, there were five instances of occurrence of two groups of haplotypes exhibiting moderate (*Bt2*) to high (*Sh2*, *SssI*, *Wx*, and *O2*) mutual contrast. In three cases, one of those groups was in the minority and clearly related to cluster 2 of guinea margaritifera, although it was not exclusive to it (supplemental Table S9–S14). In the other two cases, however, the main haplotype differentiation did not relate clearly to any known structure or geographical distribution. In the sixth case (*Ae1*), there were three strongly differentiated groups of haplotypes. It is noteworthy that two or more of those groups were observed in most varietal clusters. At the foundation of a crop, the diversity among the early domesticates depends both on the initial wild progenitor diversity structure and the distribution of the domestication process. If the size of the wild populations is large enough, it is expected that, for most loci, multiple alleles coexist, which may display complex derivation relationships and reveal no particular genealogical pattern. Any heterogeneity may result in a multiple foundation with discrete groups of early cultivated forms. Our study in sorghum revealed a marked contrast between a relatively loose structure when whole-genome markers were used and the existence of at least two, occasionally three, strongly differentiated haplotype groups at the individual gene level.

One interpretation is that the foundation was from a set of discrete lineages and that present variation was the result of profuse recombination after that foundation. The possible confusion between initial differentiation among founders, which impacted on differentiation among the early domesticates, and the later introgression from local wild forms, which might induce a localized appearance of highly differentiated alleles, cannot be settled without accurate geographical analysis. An extension of this type of work to more genes and gene segments, while including representatives of the wild progenitor across its geographical distribution, will provide firm evidence of the domestication process.

The most striking feature of our results is probably the observation of what seemed to be novel recent diversity. For several genes, primary varietal groups displayed more allele diversity, whereas secondary varietal groups in western Africa and southern Africa usually displayed the most frequent allele, which was generally the most ancient one (supplemental Tables S9–S14). Such patterns are typical of founder effects where secondary types only kept the predominant allele, as expected through drift at neutral loci. In contrast, *Ae1*, *Wx*, and especially *O2* displayed novel variation outside the area of origin of sorghum. There were many instances of alleles that were present in the varietal groups of secondary origin and absent, or very rare, in the area of origin. There might be various explanations for those situations, such as incomplete coverage of the survey in the area of origin, disappearance of some alleles in the area of origin, or introgression of new alleles in cultivated forms in varietal groups of secondary origin. However, those alleles that were specific to secondary groups were also the most recent ones in allele genealogies, suggesting they appeared in the course of the domestication process and concomitant varietal migrations. Such cases of potential novel alleles are more convincing than others. For *Ae1* the meaning of the patterns observed is unclear. One of the possible “novel” haplotypes (H9) was observed both in western Africa and in Asia, which are not known to have exchanged much sorghum germplasm, and it was not completely absent in region B and C samples; the other (H6) was found in southern Africa in only three varieties. In both cases the frequency of the haplotype might have increased through drift from a preexisting allele present at low frequency in primary regions and not be detected in our study. For *Wx* portion (2 + 3 + 4), one novel type (c) found only in western Africa and one (e) found only in southern Africa displayed only one variant base, whereas another (d), present in half of west African cluster 1 and totally specific to that cluster, exhibited three specific variant sites. It is noteworthy that those haplotypes for portion (2 + 3 + 4) were randomly associated with segment 1 haplotypes, showing that intragenic recombination occurred at the same pace as single-base substitutions. For *O2*, several novel haplotype branches

were observed, which displayed between one and five mutations (Figure 3). In western Africa, one branch (haplotypes 11, 12, and 13 in cluster 2) displayed one or two haplotype-specific variants and another (H10 in cluster 1) displayed three specific variants. In southern Africa one branch (H2, H3, and H4 essentially in cluster 7) represented a series with one, two, and three variants, respectively, which displayed increasing specificity, and another (haplotypes H14–H17 in cluster 8) had a distal part that was specific and displayed two to four more variants than the proximal part. Altogether, this suggests that multiple mutations occurred and were retained during the geographical spread of cultivars outside the area of early domestication and it points at *Wx* and especially *O2* as the cases with clearest evidence of such novel diversity. OLSEN and PURUGGANAN (2002) documented a similar situation with the *Wx* gene in rice, where several mutations were observed, which appeared posterior to the “glutinous-phenotype” mutation, itself occurring after the start of domestication.

Crop neodiversity: Mutation is generally considered to have little impact over domestication given the short time span. However, our results for sorghum indicated several instances where both mutation and intragenic recombination had generated novel alleles since the beginning of domestication. It is likely that the emergence of novel alleles at high frequency was made possible by positive selection. Cultivated forms are subjected to both natural and human selection. Selection for crop traits may screen among new recombinants and new mutants and differentially impact adaptive *vs.* neutral genes. It can foster rapid changes in gene diversity by selecting new alleles derived from recent mutations. This generates higher substitution rates in those genes that are subjected to selection. Our study suggests that certainly *O2*, and possibly *Wx*, were directly subjected to selection.

On a finer scale, our results and interpretation suggest that all mutations that contributed to the emergence of novel haplotypes, or at least many of them, were individually subjected to positive selection. For *O2*, those variants were localized in exon 1, where they were all nonsynonymous, but also in the promoter region and in intron 1; for *Wx*, they were found in exon 9 but most numerous in the 3'-UTR. The high frequency of those variants in noncoding regions is noteworthy, as polymorphisms that are subjected to selection are expected to be localized primarily in regulatory regions or in exons. Yet the scarce data available on noncoding regions in mammals suggest that the substitution rates in 5' regions and 3'-UTR are 2.5 times as high as those at nondegenerate exon sites but similar to those at twofold degenerate exon sites and only 0.6 times as high as those at fourfold degenerate exon sites, introns, or pseudogenes (LI 1997).

What we are observing can be called “crop neodiversity”: it is novel diversity that is directly the result

of human action through the selection of favorable mutants in the crop. This relates to the observations of RASMUSSEN and PHILLIPS (1997) who advocated the implication of *de novo* variation as the substrate for sustained barley improvement from an initial narrow basis. Among the phenomena that could contribute to this process, the authors mentioned single-base mutations, possibly favored by DNA methylation (COULONDRE *et al.* 1978), as well as intragenic recombination, which can be much more frequent (*e.g.*, 1 cM for 10–50 kb in maize BROWN and SUNDARESEN 1992; DOONER 1986) than recombination along the whole genome. These are two phenomena that are depicted in our data.

Crop phylogeographic analysis: Almost all crops have rapidly spread through migration out of their area of origin. This is likely to have been accompanied by strong selection by man, both for his usage of the products and for the adaptation of the varieties to cultivation in new environments. We have proposed an interpretation of our data in terms of sorghum evolution since domestication. Including wild progenitors as a representation of the reservoir of initial allelic diversity and of potential contributors through introgression will strengthen the interpretation. Germplasm collections of the most important crops are often very large and are becoming accurately characterized with neutral markers for large numbers of accessions (<http://www.generationcp.org>); this represents a wealth of materials and information for developing such pattern analyses.

The research described here was supported by Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) and Universidade Católica de Brasília and was funded by a grant of the National Council for Scientific and Technological Development of Brazil to L.F.A.F., a grant of Génoplatte to D.B., B.C., and M.D., and a grant of the Generation Challenge Programme to J.C.G.

LITERATURE CITED

- ALDRICH, P. R., J. DOEBLEY, K. F. SCHERTZ and A. STEC, 1992 Patterns of allozyme variation in cultivated and wild *Sorghum bicolor*. *Theor. Appl. Genet.* **85**: 451–460.
- BANDELT, H.-J., P. FORSTER, B. C. SYKES and M. B. RICHARDS, 1995 Mitochondrial portraits of human populations using median network. *Genetics* **141**: 743–753.
- BANDELT, H.-J., P. FORSTER and A. RÖHL, 1999 Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**: 37–48.
- BROWN, J., and V. SUNDARESAN, 1992 Genetic study of the loss and restoration of *Mutator* transposon activity in maize: evidence against dominant-negative regulator associated with loss of activity. *Genetics* **130**: 889–898.
- CASA, A. M., S. E. MITCHELL, M. T. HAMBLIN, H. SUN, J. E. BOWERS *et al.*, 2005 Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. *Theor. Appl. Genet.* **111**: 23–30.
- COULONDRE, C., J. H. MILLER, P. J. FARABAUGH and W. GILBERT, 1978 Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**: 775–780.
- CUI, Y. X., G. W. XU, C. W. MAGIL, K. F. SCHERTZ and G. E. HART, 1995 RFLP-based assay of *Sorghum bicolor* (L.) Moench genetic diversity. *Theor. Appl. Genet.* **90**: 787–796.

- DE OLIVEIRA, A. C., T. RICHTER and J. L. BENNETZEN, 1996 Regional and racial specificities in sorghum germplasm assessed with DNA markers. *Genome* **39**: 579–587.
- DE WET, J. M. J., and J. P. HUCKABAY, 1967 The origin of *Sorghum bicolor*. II. Distribution and domestication. *Evolution* **21**: 782–802.
- DE WET, J. M. J., J. R. HARLAN and E. G. PRICE, 1976 Variability in *Sorghum bicolor*, pp. 453–464 in *Origins of African Plant Domestication*, edited by J. R. HARLAN, J. M. J. DE WET and A. B. L. STEMLER. Mouton Press, The Hague, The Netherlands.
- DÉGREMONT, I., 1992 Evaluation de la diversité génétique et du comportement en croisement des sorghos (*Sorghum bicolor* L. Moench) de race guinea au moyen de marqueurs enzymatiques et morphophysiologiques. Ph.D. Thesis, University Paris XI, Orsay, France.
- DEU, M., D. GONZÁLEZ-DE-LEON, J.-C. GLASZMANN, I. DÉGREMONT, J. CHANTEREAU *et al.*, 1994 RFLP diversity in cultivated sorghum in relation to racial differentiation. *Theor. Appl. Genet.* **88**: 838–844.
- DEU, M., P. HAMON, J. CHANTEREAU, P. DUFOUR, A. D'HONT *et al.*, 1995 Mitochondrial DNA diversity in wild and cultivated sorghum. *Genome* **38**: 635–645.
- DEU, M., F. RATTUNDE and J. CHANTEREAU, 2006 A global view of genetic diversity in cultivated sorghums using a core collection. *Genome* **49**: 168–180.
- DJÈ, Y., M. HEUERTZ, C. LEFÈVRE and X. VEKEMANS, 2000 Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. *Theor. Appl. Genet.* **100**: 918–925.
- DOGGETT, H., 1988 *Sorghum*, Ed. 2. Longman Scientific and Technical, London.
- DOONER, H. K., 1986 Genetic fine structure of the *bronze* locus in maize. *Genetics* **113**: 1021–1036.
- FOLKERTSMA, R. T., F. H. RATTUNDE, S. CHANDRA, G. S. RAJU and C. T. HASH, 2005 The pattern of genetic diversity of guinea-race *Sorghum bicolor* (L.) Moench landraces as revealed with SSR markers. *Theor. Appl. Genet.* **111**: 399–409.
- GEPTS, P., 2004 Crop domestication as a long-term selection experiment, pp. 1–44 in *Plant Breeding Reviews*, Vol. 24, Pt. 2, edited by J. JANNICK. John Wiley & Sons, West Sussex, UK.
- GRENIER, C., M. DEU, S. KRESOVICH, P. J. BRAMEL-COX and P. HAMON, 2000 Assessment of genetic diversity in three subsets constituted from the ICRISAT sorghum collection using random vs non-random sampling procedures. B. Using molecular markers. *Theor. Appl. Genet.* **101**: 197–202.
- GUPTA, P. K., S. RUSTGI and P. L. KULWAL, 2005 Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol. Biol.* **57**: 461–485.
- HAMBLIN, M. T., A. M. CASA, H. SUN, S. C. MURRAY, A. H. PATERSON *et al.*, 2006 Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. *Genetics* **173**: 953–964.
- HAMBLIN, M. T., S. E. MITCHELL, G. M. WHITE, J. GALLEGO J. R. KUKATLA *et al.*, 2004 Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium, and selection in a diverse sample of *Sorghum bicolor*. *Genetics* **167**: 471–483.
- HARLAN, J. R., 1995 *The Living Fields, Our Agricultural Heritage*. Cambridge University Press, Cambridge, UK.
- HARLAN, J. R., and J. M. J. DE WET, 1972 A simplified classification of cultivated sorghum. *Crop Sci.* **12**: 172–176.
- HENIKOFF, S., and J. G. HENIKOFF, 1992 Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**: 10915–10919.
- HENRY, A. M., and C. DAMERVAL, 1997 High rates of polymorphism and recombination at the *Opaque-2* locus in cultivated maize. *Mol. Gen. Genet.* **256**: 147–157.
- HENRY, A. M., D. MANICACCI, M. FALQUE and C. DAMERVAL, 2005 Molecular evolution of the *Opaque-2* gene in *Zea mays* L. *J. Mol. Evol.* **61**: 551–558.
- INUKAI, T., A. SAKO, H.-Y. HIRANO and Y. SANO, 2000 Analysis of intragenic recombination at *wx* in rice: correlation between the molecular and genetic maps within the locus. *Genome* **43**: 589–596.
- KILIAN, B., H. OZKAN, J. KOHL, A. VON HAESELER, F. BARALE *et al.*, 2006 Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. *Mol. Genet. Genomics* **276**: 230–241.
- KYTE, J., and R. F. DOOLITTLE, 1982 A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**: 105–132.
- LI, W. H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- MANICACCI, D., M. FALQUE, S. LE GUILLLOU, B. PIEGU, A.-M. HENRY *et al.*, 2007 Maize *Sh2* gene is constrained by natural selection but escaped domestication. *J. Evol. Biol.* **20**: 503–516.
- MENKIR, A., P. GOLDSBROUGH and G. EJETA, 1997 RAPD based assessment of genetic diversity in cultivated races of sorghum. *Crop Sci.* **37**: 564–569.
- OKAGAKI, R. J., and C. F. WEIL, 1997 Analysis of recombination sites within the maize *waxy* locus. *Genetics* **147**: 815–821.
- OLLITRAULT, P., 1987 Evaluation génétique des sorghos cultivés (*Sorghum bicolor* L. Moench) par l'analyse conjointe des diversités enzymatique et morphophysiologique. Relations avec les sorghos sauvages. Ph.D. Thesis, University Paris XI, Orsay, France.
- OLLITRAULT, P., M. ARNAUD and J. CHANTEREAU, 1989 Polymorphisme enzymatique des sorghos. II. Organisation génétique et évolutive des sorghos cultivés. *Agron. Trop.* **44**: 211–222.
- OLSEN, K. M., and M. D. PURUGGANAN, 2002 Molecular evidence on the origin and evolution of glutinous rice. *Genetics* **162**: 941–950.
- OLSEN, K. M., A. L. CAICEDO, N. POLATO, A. MCCLUNG, S. MCCOUCH *et al.*, 2006 Selection under domestication: evidence for a sweep in the rice *Waxy* genomic region. *Genetics* **173**: 975–983.
- OTTO, S., 2000 Detecting the form of selection from DNA sequence data. *Trends Genet.* **16**: 526–529.
- PIROVANO, L., S. LANZINI, H. HARTINGS, N. LAZZARONI, V. ROSSI *et al.*, 1994 Structural and functional analysis of an *Opaque-2*-related gene from sorghum. *Plant Mol. Biol.* **24**: 515–523.
- RASMUSSEN, D. C., and R. L. PHILLIPS, 1997 Plant breeding progress and genetic diversity from de novo variation and elevated epistasis. *Crop Sci.* **37**: 303–308.
- ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.
- RUTHERFORD, K., J. PARKHILL, J. CROOK, T. HORSNELL, P. RICE *et al.*, 2000 Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944–945.
- SCHULTZ, J. A., and J. A. JUVIK, 2004 Current models for starch synthesis and the *sugary enhancer1* (*se1*) mutation in *Zea mays*. *Plant Physiol. Biochem.* **42**: 457–464.
- SECOND, G., 1985 Evolutionary relationships in the *Sativa* group of *Oryza* based on isozyme data. *Genet. Sel. Evol.* **17**(1): 89–114.
- TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**: 256–276.
- WHITT, S. R., L. M. WILSON, M. I. TENAILLON, B. S. GAUT and E. S. BUCKLER, 2002 Genetic diversity and selection in the maize starch pathway. *Proc. Natl. Acad. Sci. USA* **99**: 12959–12962.
- WILSON, L. M., S. R. WHITT, A. M. IBANEZ, T. R. ROCHEFORD, M. M. GOODMAN *et al.*, 2004 Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* **16**: 2719–2733.
- YAMANAKA, S., I. NAKAMURA, K. N. WATANABE and Y. SATO, 2004 Identification of SNPs in the *waxy* gene among glutinous rice cultivars and their evolutionary significance during the domestication process of rice. *Theor. Appl. Genet.* **108**: 1200–1204.