Theor Appl Genet (2011) 123:1231–1246
DOI 10.1007/s00122-011-1662-0

ORIGINAL PAPER

Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in a western African region

Fabrice Sagnard · Monique Deu · Dékoro Dembélé · Raphaël Leblois · Lassana Touré · Mohamed Diakité · Caroline Calatayud · Michel Vaksmann · Sophie Bouchet · Yaya Mallé · Sabine Togola · Pierre C. Sibiry Traoré

Received: 24 March 2011 / Accepted: 9 July 2011 / Published online: 3 August 2011

Abstract Gene flow between domesticated plants and their wild relatives is one of the major evolutionary processes acting to shape their structure of genetic diversity. Earlier literature, in the 1970s, reported on the interfertility and the sympatry of wild, weedy and cultivated sorghum belonging to the species *Sorghum bicolor* in most regions of sub-Saharan Africa. However, only a few recent surveys have addressed the geographical and ecological distribution of sorghum wild relatives and their genetic structure. These features are poorly documented, especially in western Africa, a centre of diversity for this crop. We report here on an exhaustive in situ collection of wild, weedy and cultivated sorghum assembled in Mali and in Guinea. The extent and pattern of genetic diversity were assessed with 15 SSRs within the cultivated pool (455 accessions), the wild pool (91 wild and weedy forms) and between them. F_{ST} and R_{ST} statistics, distance-based trees, Bayesian clustering methods, as well as isolation by distance models, were used to infer evolutionary relationships within the wild–weedy–crop complex. Firstly, our analyses highlighted a strong racial structure of genetic diversity within cultivated sorghum ($F_{ST} = 0.40$). Secondly, clustering analyses highlighted the introgressed nature of most of the wild and weedy sorghum and grouped them into two eco-geographical groups. Such closeness between wild and crop sorghum could be the result of both sorghum's domestication history and preferential post-domestication crop-to-wild gene flow enhanced by farmers' practices. Finally, isolation by distance analyses showed strong spatial genetic structure within each pool, due to spatially limited dispersal, and suggested consequent gene flow between the wild and the crop pools, also supported by R_{ST} analyses. Our findings thus revealed important features for the collection, conservation and biosafety of domesticated and wild sorghum in their centre of diversity.

Communicated by H. H. Geiger.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1662-0) contains supplementary material, which is available to authorized users.

F. Sagnard · M. Deu (✉) · C. Calatayud · M. Vaksmann · S. Bouchet
CIRAD, UMR AGAP, Avenue d'Agropolis,
34398 Montpellier, France
e-mail: monique.deu@cirad.fr

F. Sagnard
ICRISAT, co/ILRI, PO Box 39063, Nairobi, Kenya

D. Dembélé · Y. Mallé · S. Togola · P. C. S. Traoré
ICRISAT, Station de Samanko, PB 320, Bamako, Mali

R. Leblois
INRA, UMR CBGP (INRA-IRD-CIRAD-Montpellier SupAgro),
Campus International de Baillarguet, 34988 Montferrier-sur-Lez
Cedex, France

L. Touré · M. Vaksmann
IER-Sotuba, P.O. Box 1704, Bamako, Mali

M. Diakité
IRAG, CRA Bordo- Programme Céréales, Kankan, Guinea

Introduction

Some domesticated crops belong to the same biological species as their wild progenitors, are fully compatible with them and, in some cases, are sympatric with their relatives, which may be in the form of wild or weedy plants (Harlan and de Wet 1971; Papa and Gepts 2003). In a literature review, Ellstrand et al. (1999) reported that 12 of the world's major 13 food crops naturally hybridise with their wild relatives. The phenomenon is widespread and has probably occurred since the origins of agriculture, 10,000 years ago (Gepts and Papa 2003). Gene flow is a continuous source of useful variation in landraces and many authors suggest that natural wild-to-crop introgression continues to be a factor in increasing the genetic diversity of modern crops (Jarvis and Hodgkin 1999; Arnold 2004).

Most wild progenitors of crops contain more diversity than their respective crops, due to domestication processes that have induced tight bottlenecks (Gepts and Papa 2003; Papa et al. 2005). The wild relatives of a crop therefore constitute a reservoir of useful genes for plant breeders (Gepts and Papa 2003; Kameswara Rao et al. 2003). Genes from wild relatives have been used to provide resistance against specific pests and diseases, improve tolerance of abiotic stress, and increase nutritional qualities (Jarvis et al. 2008).

Nowadays, wild relatives are threatened by the loss of natural habitats due to human population growth and the ensuing associated anthropogenic factors. Conservation of wild species in nature and, on-farm of domesticated varieties, has been identified as two major issues by the Convention of Biological Diversity (CBD 1992). For more utilitarian conservation devoted to food and agriculture, it is first necessary to define the genetic relationships existing between a crop and its wild relatives, to set priorities, especially when conservation resources are limited (Maxted et al. 2006).

However, gene flow from crop to wild relatives has also been implicated in the evolution of increased weediness in seven major crop species (Ellstrand et al. 1999). Finally, massive gene flow from crop to wild can also lead to the displacement of genetic diversity in wild populations (Papa and Gepts 2003) and, in extreme cases, to the extinction of wild populations (Ellstrand et al. 1999).

With the biotechnological advances in GM technology, gene flow from cultivated crops to their wild and weedy relatives has been of increasing scientific interest due to its possible threat to wild populations (Ellstrand 2003). Knowledge of the geographical distribution of wild relatives and landrace diversity, particularly in centres of domestication, is a prerequisite for assessing the possibilities of transgene transfers via gene flow and for proposing natural population conservation strategies. More recently, molecular markers have been successfully used in various species for further evidence of secondary centres of domestication by analysing the structure and diversity within the wild–weedy–domesticated complex, and the extent and direction of gene flow. They have shed light on these major issues for common bean in Mesoamerica (Papa and Gepts 2003; Papa et al. 2005; Zizumbo-Villarreal et al. 2005), pearl millet in Niger (Mariac et al. 2006a, b) and maize in North and South America (Matsuoka et al. 2002; Vigouroux et al. 2005).

Sorghum is a staple cereal in sub-Saharan Africa, where it was domesticated. The primary gene pool of the genus *Sorghum* includes only two diploid species, *Sorghum propinquum* and *S. bicolor* (Harlan and de Wet 1971). The species *Sorghum bicolor* has been subdivided into three sub-species: ssp. *bicolor* (the domesticated forms), ssp. *verticilliflorum* (the closest wild relatives) and ssp. *drumondii* (the weedy forms, which are stabilized hybrids between wild relatives and cultivated forms). Based on spikelet morphology, Harlan and de Wet (1972) distinguished five main races within the cultivated forms (bicolor, caudatum, durra, guinea and kafir) and four races within the wild relatives (aethiopicum, arundinaceum, verticilliflorum and virgatum). These wild forms, which are fully interfertile with cultivated forms, grow sympatrically with landraces in many of the sorghum-growing regions of sub-Saharan Africa (de Wet et al. 1970; de Wet 1978; Tesso et al. 2008).

Surveys on the potential for crop–wild gene flow are still scarce for sorghum, especially in traditional agro-ecosystems in Africa. They have mainly concerned geographical and ecological distribution of wild and weedy sorghum in Kenya (Mutegi et al. 2010) and in Ethiopia and Niger (Tesso et al. 2008). In both cases, they concluded that crop-to-wild gene flow was likely to occur in many agro-ecosystems. Barnaud et al. (2009) first investigated the dynamics of genetic diversity within the weedy–crop complex in a traditional farming system in a village of northern Cameroon. Morphological traits and molecular markers confirmed the introgressed status of weedy forms in that village.

We report here on a large-scale genetic analysis of wild, weedy and cultivated sorghum in a western African region. We first collected wild, weedy and cultivated sorghum in the main agro-ecological zones of sorghum cultivation in Mali to enhance our knowledge of wild and weedy distribution. We then increased the geographical range and climate diversity of our collection by including accessions from Guinea. We genotyped all the collected material with a set of 15 SSR markers. We finally inferred evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in this western African region.

The specific objectives of this study were:

(1) to assess the pattern of genetic diversity revealed by SSR markers within a large collection of cultivated sorghum gathered in Mali in relation to botanical and climatic factors,

(2) to compare the extent and pattern of diversity within and between cultivated and wild Malian and North Guinean gene pools,

(3) to investigate gene flow between these cultivated and wild sorghum.

Materials and methods

Collection of wild, weedy and cultivated sorghum

Collection in Mali

Cultivated, wild and weedy sorghum accessions were collected in 63 villages in 2004 and 2005. The villages were selected as being representative of both the rainfall gradient (mean annual rainfall calculated over the 1971–2000 period ranging from 350 to 1,200 mm) and the main agro-ecological zones of sorghum cultivation in Mali. The collection period was selected in each agro-ecological, zone both to match the maturity of cultivated sorghum and to allow seed harvesting from wild and weedy sorghum before they shattered. Collective interviews were conducted in each village to gather information on crop systems and uses, the presence and abundance of wild and weedy sorghum, the origin and dynamics of each of the cultivated varieties and to establish the list of varieties grown in the village. In a few villages, different morphotypes (i.e., loose and compact panicle, or different glume colour) were grouped under a single folk name used by farmers. However, they were recognised as different varieties by farmers during the collective interview. Lastly, as in Niger (Deu et al. 2010), we found that folk names used to distinguish between varieties were mostly morphology-related traits (panicle type, seed or glume colour, cycle duration, etc.) or use-related traits (sugary sorghum). Thus, we collected samples both from different morphotypes sharing the same name and from different folk names. Cultivated varieties and wild or weedy forms were then collected with individual farmers. Each variety present in a village was collected and represented by one accession, provided by one farmer. Our objective was to sample ten panicles per accession, but we ultimately obtained an average number of five. For wild and weedy sorghum, we were able to collect different samples per village when different morphotypes were observed. The geographical coordinates associated with each village were recorded using a handheld Trimble GeoXT GPS unit.

Collection in Guinea

To extend collection to the Guinean climatic zone, which is more humid and forested and consequently suspected of hosting wild sorghum belonging to the ssp. *verticilliflorum* race arundinaceum, we collected wild and cultivated sorghum in four villages located in the Haute-Guinée region in November 2007. These villages were selected for their rainfall and latitudinal gradients: from Setiguiya (mean annual rainfall 1,200 mm) to Tokounou (1,700 mm). The latter village was located on the edge of the primary rainforest of Guinea.

The methods we used for interviews and collections were similar to those previously described for Mali.

Additional wild sorghum from the genebank

To assess the status of wild sorghum collected in Mali and Guinea, we included in the analysis a panel of 25 allopatric wild accessions, originating from the ICRISAT genebank and also stored at CIRAD. The four eco-geographical races of *Sorghum bicolor* ssp. *verticilliflorum* were sampled (23 accessions). One accession of *Sorghum lancelolatum* Stapf, now classified within *S. bicolor* ssp. *verticilliflorum* (de Wet 1978) was also included in this study. These accessions were listed in a previous study (Deu et al. 1995).

Racial characterisation of cultivated accessions

The Malian collection was grown in 2006 at the ICRISAT research station (Samanko) for agromorphological characterisation (not presented in this paper). Racial characterisation of cultivated accessions, based on panicle and spikelet morphology, was carried out in accordance with Harlan and de Wet's classification (1972). Within guinea sorghum, we used Snowden's classification (1936) to distinguish between the different sub-races.

Cultivated sorghum collected in Guinea were botanically determined during collection.

DNA extractions and SSR genotyping

Seeds from each accession of the Malian collection were germinated at room temperature. DNA was extracted from one seedling (2–3 weeks old) at the University of Bamako, from freshly harvested leaves using the modified MATAB method (Risterucci et al. 2000). Before genotyping, the quality and concentration of the DNA samples were checked at the CIRAD laboratory. Samples with low DNA quality or quantity, as well as samples from the Guinean and the genebank collections, were extracted in Montpellier from seedlings germinated in a greenhouse.

A set of 15 SSR markers (Supplementary Electronic File 1) was used to genotype the different accessions, as previously described (Barnaud et al. 2007; Deu et al. 2008). They were spread across the sorghum genome. They were selected for their reliability and scoring accuracy among SSR markers formerly used to assess in situ sorghum genetic diversity in Niger and Burkina Faso (Deu et al. 2008; Barro-Kondombo et al. 2010). Genotyping was carried out at the Languedoc Roussillon Génopole platform located on the CIRAD Campus (Montpellier, France) following methods described previously (Barnaud et al. 2007).

Our genotyping strategy (one individual per accession without taking into account within-accession variability) enabled us to maximise the number of accessions that could be genotyped on the large scale of this study. This strategy was formerly applied to analyse structure and diversity within the wild–crop complex of out-crossing species and the extent of gene flow (Matsuoka et al. 2002; Mariac et al. 2006a, b). It has shed light on large-scale evolutionary trends in sorghum in western and eastern African countries (Deu et al. 2008; Barro-Kondombo et al. 2010; Mutegi et al. 2011).

Genetic data analyses

In our study, we did not attempt to differentiate between wild subspecies or races. Races from the wild subspecies *verticilliflorum* "do not deserve formal taxonomic status as they grade morphologically and ecologically so completely into one another" (de Wet, 1978). The subspecies *drumondii* includes a "morphologically variable group in which derivatives of hybridisation between subspecies *verticilliflorum* and subspecies *bicolor* are not always easy to distinguish on the basis of morphology from wild sorghums" (de Wet et al. 1976) Consequently, the term "wild" refers to both wild and weedy types in the subsequent genetic analyses.

Genetic diversity parameters were calculated for each gene pool (wild and cultivated) with FSTAT software (Goudet 2002): Nei's unbiased gene diversity or expected heterozygosity (H_e), observed heterozygosity (H_o), total number of alleles (A^t), number of rare alleles (A^r, freq < 5%) and mean allelic richness across loci. The number of private alleles (A^p) was detected by GDA software version 1.1 (Lewis and Zaykin 2001). Allelic richness (R_s), and private allelic richness ($\prod^s$), were both corrected for sample size differences and estimated by using the rarefaction method implemented in HP-Rare 1.2 software (Kalinowski 2005). The significance of differences in H_e, R_s and $\prod^s$ between cultivated and wild sorghum was tested using Wilcoxon paired-rank tests.

To assess whether crop-to-wild gene flow might be preferential in Mali within some regions, or with some races (botanical types) of cultivated sorghum, we first clustered cultivated as well as wild sorghum according to climatic zones. Three climatic zones were defined based on average annual rainfall in Mali (1971–2001); the data sources were Agrhymet-IRD, FAO-LocClim, 2005. They roughly corresponded to the Sudano–Sahelian zone (under 600 mm of average annual rainfall), the Sudanian zone (from 600 to 900 mm) and the Sudano–Guinean zone (over 900 mm). We then clustered cultivated accessions only, according to racial characterisation, and compared them with the whole wild Malian pool. We used genetic diversity parameters described above and *F* statistics to explore population structure within and between cultivated and wild sorghum, in relation to the different factors investigated (botanical and climatic factors). F_{ST} were computed with GENETIX software and tested for their significance with 10,000 permutations (Belkhir et al. 2002).

To assess the genetic relationships within and between cultivated and wild sorghum, we used two complementary approaches, a distance-based method and a Bayesian model-based clustering method.

A dissimilarity matrix between all pairs of individuals was first computed using the shared allele distance. A dendrogram was then generated on the dissimilarity matrix applying the neighbour joining algorithm implemented in DARwin v5 software (Perrier and Jacquemoud-Collet 2006).

We then used the Bayesian model-based clustering method developed by Pritchard et al. (2000), implemented in STRUCTURE v2.2 software. We used an admixture model, with correlated allele frequencies, without prior population information. This model assumes that the genome of each individual is a mixture of genes originating from *K* unknown ancestral populations. It is therefore useful for identifying gene flow events. We ran 20 replicate analyses for each *K* value ranging from 1 to 10, with a burn-in period of 500,000 followed by 1.10^6 iterations. Analyses were conducted in five different data sets:

(1) data set 1 consisted of cultivated sorghum from Mali,
(2) data set 2 grouped all cultivated sorghum from Mali and Guinea,
(3) data set 3 included all Malian and Guinean wild and cultivated sorghum,
(4) data set 4 merged data set 3 and the wild accessions from the genebank,
(5) data set 5 comprised only the wild complex (wild sorghum from Mali, Guinea and genebank accessions).

We computed pairwise R_{ST} (differentiation based on allele size differences, Slatkin 1995, estimated as per Michalakis and Excoffier 1996) between all pairs of clusters defined by STRUCTURE analyses. We only kept individuals attributed to a cluster (ancestry >90%) within the data set 2 (cultivated from Mali and Guinea) and data set 5 (all wild forms). We compared R_{ST} with R_{ST} obtained after 10,000 allele size permutations (pR_{ST}) to provide insights into the genealogical history of the crop and wild groups (i.e., migration vs. divergence). R_{ST} is expected to be significantly higher than the mean permuted (pR_{ST}) under a phylogeographical pattern, i.e., populations have diverged for a long time and exchanged migrants at a low rate compared to the mutation rate (Hardy et al. 2003). Conversely, R_{ST} and pR_{ST} would not be strongly different where mutations do not contribute to the differentiation, i.e., if gene flow is large compared to the mutation rate. These analyses were performed with SPAGeDi version 1.2 (Hardy and Vekemans 2002).

Finally, isolation by distance patterns were assessed by regressing genetic distances between individuals ($\hat{a}_r$, Rousset 2000) over the logarithm of geographical distances and tested using Mantel tests with 30,000 permutations. Firstly, we considered only pairs of individuals taken from two different groups and discarded pairs of individuals taken from a single group (the "between-group" analysis). Secondly, we considered pairs of individuals taken from a single group and discarded pairs of individuals taken from different groups (the "within-group" analysis). Those analyses were performed using GENEPOP 4.0 (Rousset 2008) and an R script (R Development Core Team 2007), available upon request, which modified the Mantel test to calculate rank correlation coefficients and to permute the pairwise distances within groups only. Such "within and between-group" analyses enabled to test for the presence of gene flow between different groups of individuals (e.g., different habitats, hosts or any categories) in the context of isolation by distance (Rousset 1999, see Martel et al. 2003 for an example of such analysis). These analyses were performed within and between the wild and the cultivated pool to test for potential gene flow between them.

Results

Description of the Malian collection

We collected 420 cultivated sorghum from 60 villages in Mali (Supplementary Electronic File 2). The number of cultivated accessions collected per village ranged from 1 to 13 (average 7.0). Wild and weedy sorghum were found in all but three of the villages. Three neighbouring villages were visited to collect wild forms. A total of 83 wild and weedy sorghum were collected, with a maximum number of four per village. Wild sorghum were identified in different types of habitats. They were abundant, in decreasing order of prevalence, in cereal fields, fallows, home gardens and threshing zones. They were rarely found in most humid zones (temporary streams). Although there is great variability within wild and weedy sorghum for some agro-morphological traits (height, panicle size and compactness), Malian farmers do not name these forms differently, in contrast with farmers in North Cameroon (Barnaud et al. 2009). Vernacular names for wild sorghum generally indicate a relation with domesticated animals, in reference to a dispersal vector effect, or as feed for them (*tio douo*: horse sorghum). They also refer to their "dead" (*nio sou*) or useless (*baba*) nature.

All basic sorghum races except kafir were found in Mali. Guinea gambicum–guineense (52.9%) was the most prevalent type. Guinea margaritiferum (16.7%), characterised by smaller and vitreous grains, and bicolor (12.1%), which are sweet sorghum, were relatively abundant. Each of the other two races (caudatum and durra) amounted to <5% of all accessions. All the intermediate races accounted for 10% of the collection.

Racial distribution was not random (Fig. 1a). Durra and caudatum were mainly found in the Sudano–Sahelian zone (northern Mali); guinea margaritiferum were mainly cultivated in the Sudano–Guinean zone (southern Mali, area bordering Guinea and Ivory Coast). Bicolor were found in 70% of the villages. Although guinea gambicum–guineense accessions were grown in all the villages we visited, they were more frequent in the Sudanian cotton zone. Farmers relied on a broader racial diversity in the northern zone.

Description of the Guinean collection

In the four villages visited, we collected 35 cultivated and 8 wild or weedy sorghum accessions. In Guinea, unlike in Mali, the majority of cultivated accessions were guinea margaritiferum (57.1%). Guinea gambicum–guineeense accessions amounted to 25.7% of the collection. Sweet sorghum (bicolor or guinea-caudatum) were grown in all the villages except the one located in the most southern area (Tokounou).

Only weedy types were found in the three villages located in the tree savannah zone. In Tokounou, we finally discovered a true wild sorghum, growing in a pond margin. Its morphology corresponded to the arundinaceum race (de Wet and Harlan 1971). The farmers called it "Soulala kende" (Monkey sorghum) and considered this plant a natural grass.

Extent and structure of genetic diversity within the Malian collection according to various structuring factors

Genetic diversity parameters for the cultivated and the wild pool are presented in Table 1. In total, 165 alleles were detected with 15 SSR markers on the overall Malian collection comprising 420 cultivated accessions and 83 wild or weedy sorghum. The number of alleles detected in the cultivated pool (CP) was 144. Of them, 58 (40.3%) were private. In the wild pool (WP), 21 alleles (19.6%) were private out of the 107 alleles detected. Observed heterozygosity was significantly higher in the WP ($H_o = 0.1$) compared to the CP ($H_o = 0.051$, $P = 0.0045$). The

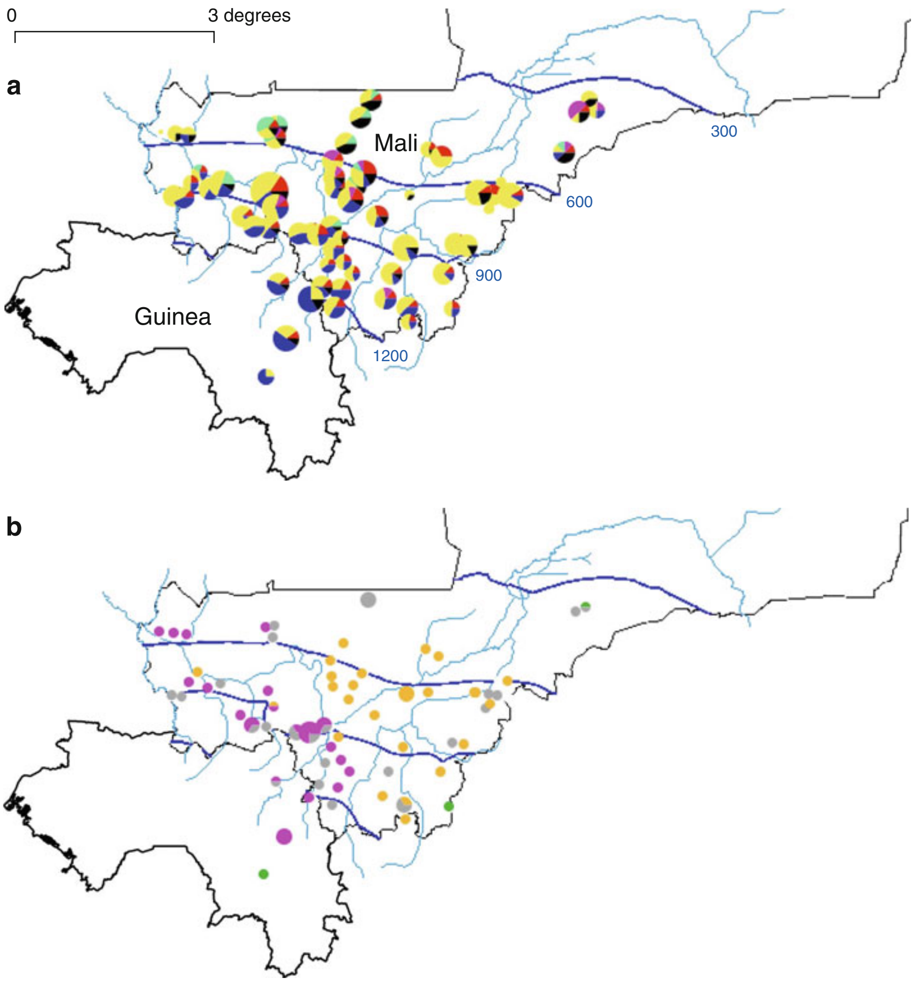

Fig. 1 Geographical distribution of cultivated and wild sorghum. **a** Geographical distribution of sorghum racial types per village. Each *colour* represents a racial type, *red* bicolor, *pink* caudatum, *green* durra, *yellow* guinea gambicum–guineense, *blue* guinea margaritiferum, *black* intermediate type. The size of the pies is proportional to the number of cultivated accessions collected per village. **b** Projection of the three groups obtained within the wild pool with STRUCTURE software. Only accessions with more than 90% ancestry were attributed to a cluster. *Orange*, *pink* and *green* indicate accessions attributed to clusters W1, W2 and W3, respectively. Unattributed accessions are shown in *grey*. The size of the pies is proportional to the number of wild or weedy sorghum collected per village

cultivated and wild pools displayed similar levels of gene diversity ($H_e = 0.569$ for the CP, $H_e = 0.588$ for the WP, $P = 0.65$). Allelic richness ($R_s = 7.48$ alleles per locus in the CP, $R_s = 7.02$ in the WP) and private allelic richness ($\prod^s = 1.42$ in the CP, $\prod^s = 0.98$ in the WP), both corrected for sample size and calculated on 70 accessions, were not significantly different between CP and WP ($P = 0.09$ for R_s, $P = 0.07$ for $\prod^s$). The cultivated and wild pools were moderately differentiated ($F_{ST} = 0.14$, $P < 0.01$).

To investigate the structure of genetic diversity within and between the cultivated and wild pools collected in Mali on more localized scales (climatic zones), or between some racial types of cultivated sorghum and the wild pool, we calculated genetic diversity and differentiation parameters for the various structuring factors (shown in Table 2).

Within the cultivated pool, sorghum from the Sudano–Sahelian zone had a significantly higher gene diversity ($H_e = 0.606$) than sorghum from the Sudanian zone ($H_e = 0.538$, $P = 0.046$). The Sudano–Sahelian cultivated pool also exhibited a significantly higher allelic richness ($R_s = 5.71$) compared to the Sudano–Guinean pool ($R_s = 4.32$, $P = 0.0026$). Private alleles were present in each zone. Although mean private allelic richness was higher in the Sudano–Sahelian zone ($\prod^s = 0.89$) compared to the other zones ($\prod^s < 0.4$), the differences were not significant. Cultivated sorghum from the different climatic zones were poorly differentiated ($F_{ST} = 0.037$).

Within the wild pool, no significant difference was observed between zones for gene diversity and private allelic richness. The wild pool from the Sudano–Guinean zone displayed a significantly lower allelic richness ($R_s = 4.18$) than the pool from the Sudanian zone ($R_s = 5.14$, $P = 0.03$). Wild sorghum were also poorly differentiated between climatic zones ($F_{ST} = 0.037$).

Differentiations, based on F_{ST}, between the cultivated and wild pools within climatic zones, were moderate. They varied from 0.14 in the Sudano–Guinean zone to 0.18 in the Sudano–Sahelian zone (data not shown).

Within the cultivated pool, gene diversity varied between racial types. It was significantly higher in intermediate ($H_e = 0.614$), caudatum ($H_e = 0.608$) and durra accessions ($H_e = 0.576$) compared to bicolor ($H_e = 0.350$), guinea gambicum ($H_e = 0.355$) and guinea

Table 1 Genetic diversity parameters for the different gene pools

Gene pool	N	H_e	H_o	A^t	$A^p_{(w)}$	$R_{s\ (70)}$	$\prod_{(70)}$	$A^P_{(a)}$	$R_{s\ (8)}$	$\prod_{(8)}$	F_{ST}
Mali											
Cultivated pool	420	0.569	0.051	144	58	7.48	1.42	36	4.03	0.59	0.142
Wild pool	83	0.588	0.100	107	21	7.02	0.98	11	4.06	0.56	
Guinea											
Cultivated pool	35	0.549	0.058	63	32	–	–	0	3.41	0.19	nc
Wild pool	8	0.536	0.117	47	16	–	–	4	3.13	0.44	
Wild genebank	25	0.737	0.073	103		–	–	20	5.59	1.91	

N, total number of accessions in each pool

H_e and $H_{o,}$ unbiased gene diversity and observed heterozygosity

A^t and A^p, number of total and private alleles

$A^p_{(w)}$, number of private alleles calculated separately within each of the two collections (i.e., in Mali, 58 and 21 private alleles were found only in the cultivated pool, and in the wild pool, respectively; in Guinea, 32 and 16 alleles were found only in the cultivated pool and wild pool, respectively)

$A^p_{(a)}$, private alleles calculated for each pool, taking into account both the Malian, Guinean and wild genebank collections (i.e., the cultivated pool from Guinea had no private alleles compared to the other four pools, whilst the wild pool from Mali had 11 private alleles)

R_s and $\prod$, allelic richness and private allelic richness estimated on a sample of balanced size; the number in brackets indicates the number of accessions selected to estimate these parameters

F_{ST}, genetic differentiation between cultivated and wild pool

Table 2 Genetic diversity parameters within the Malian collection according to the structuring factors

Malian collection	N	H_e	H_o	A^t	$A^p_{(w)}$	R_s	$\prod$	F_{ST}
Climatic zones								
Cultivated pool						(19)	(19)	0.037
Z1 (<600 mm)	112	0.606	0.057	116	19	5.71	0.89	
Z2 (600–900 mm)	201	0.538	0.051	116	13	5.00	0.39	
Z3 (≥900 mm)	107	0.534	0.043	91	6	4.32	0.31	
Wild pool						(19)	(19)	0.037
Z1 (<600 mm)	22	0.583	0.091	72	10	4.69	0.39	
Z2 (600–900 mm)	41	0.590	0.113	87	19	5.14	0.39	
Z3 (≥900 mm)	20	0.530	0.084	63	7	4.18	0.34	
Racial types						(12)	(12)	0.400
Bicolor	51	0.350	0.053	68	8	3.30	0.50	
Caudatum	14	0.608	0.033	60	4	3.97	0.41	
Durra	19	0.576	0.061	62	4	3.89	0.41	
Guinea g.	222	0.355	0.056	97	13	3.12	0.26	
Guinea m.	70	0.325	0.029	61	10	2.87	0.67	
Intermediate	44	0.614	0.055	90	5	4.74	0.47	

Z1, Z2 and Z3 are the Sudano–Sahelian, Sudanian and Sudano–Guinean zones, respectively. Guinea g. and Guinea m. designate guinea gambicum–guineense and guinea margaritiferum, respectively

R_s and $\prod$, allelic richness and private allelic richness were estimated on a sample of 19 accessions for climatic zones and on a sample of 12 accessions for racial types

$A^p_{(w)}$, number of private alleles calculated within each pool (i.e., for the cultivated pool, we defined the number of private alleles found in each climatic zone and in each type; for the wild pool, the number of private alleles was calculated in each climatic zone)

F_{ST}, average genetic differentiation between the different groups calculated for the two structuring factors

margaritiferum ($H_e = 0.325$) accessions (P values from 0.0015 to 0.015). Allelic richness was also significantly greater in the intermediate accessions (P values from 0.0035 to 0.035). Private allelic richness was significantly greater in guinea margaritiferum (0.67) compared to other guinea (0.26, $P = 0.017$) and durra (0.41, $P = 0.048$)

accessions. Cultivated sorghum from the different racial types were strongly differentiated ($F_{ST} = 0.400$), at more than ten times greater than climatic differentiation.

Pairwise F_{ST} between the different types varied from 0.07 to 0.55 (Supplementary Electronic File 3). Guinea margaritiferum and bicolor types showed higher differentiation compared to the other types. Pairwise F_{ST} between racial types and the Malian wild pool varied from 0.17 to 0.37). Guinea margaritiferum appeared closer to this wild pool ($F_{ST} = 0.17$), while guinea gambicum and bicolor were the most differentiated from this pool ($F_{ST} = 0.32$ and 0.37, respectively).

Genetic structure within the wild–weedy–crop complex

We identified the number of clusters obtained with the Bayesian model-based clustering method implemented in STRUCTURE software using different criteria: log likelihood, the congruence between runs for the different K values and the different parameters proposed by Evanno et al. (2005), including the change in the second order of likelihood (ΔK).

In both data set 1 (Malian cultivated pool) and data set 2 (cultivated sorghum from Mali and Guinea), the log likelihood increased up to $K = 4$ and to a lesser extent up to $K = 10$ (as shown in Supplementary Electronic File 4A for data set 2). At higher K values, the algorithm converged to distinct clustering schemes and large differences in the assignment of individuals were detected across runs for a given value of K. ΔK showed a clear peak at $K = 4$, which indicated that four was the most likely number of genetic clusters. These clusters roughly corresponded to racial types (Fig. 2). In data set 2, the cultivated accessions from Guinea were intermingled with the Malian accessions and assigned according to their botanical types. Guinea gambicum–guineense and guinea margartitiferum accessions showed very high average ancestry (more than 90%) in cluster A and cluster B, respectively (Table 3a). Bicolor accessions had an average ancestry of 81.1% in cluster C, while caudatum, durra and intermediate accessions showed average ancestries of 69.9, 92.9 and 72.5%, respectively, in the fourth cluster (cluster D). Accessions displayed low admixture: 87.9% of them had an ancestry over 80% in one cluster and 80% still showed an ancestry exceeding 90%. Guinea gambicum–guineense accessions showed higher admixture (21.1% were characterised by <90% of their genome in a cluster) than guinea margaritiferum (11.2%). If we considered that an accession was “pure cluster X” when its ancestry in cluster X exceeded 90%, cluster A mainly comprised guinea gambicum–guineense (181 out of 184 accessions), cluster B mainly comprised guinea margaritiferum (77 out of 80 accessions), cluster C (43 accessions) mainly comprised bicolor or intermediates with that race (41 of them were identified as sweet types) and finally, durra, caudatum, intermediate races and some bicolor made up cluster D (58 accessions). Note that a few cases of “wrong” attribution according to racial type were detected. They concerned two guinea gambicum–guineense accessions and one bicolor accession that showed ancestry exceeding 90% in the guinea margaritiferum cluster (cluster B), while two guinea margaritiferum accessions and one caudatum accession were clearly attributed to the guinea gambicum cluster (cluster A). This may have been due to gene flow between accessions or to difficulty in botanically classifying some admixed or hybrid accessions based on panicle and spikelet morphology.

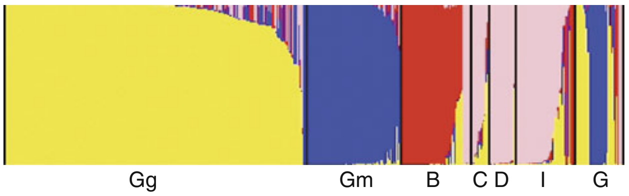

Fig. 2 Estimated population structure using the Bayesian model analysis implemented in STRUCTURE software within the whole cultivated pool (Malian and Guinean collection). Of the 20 runs, 17 performed gave similar memberships for each accession. We present here the run with the highest posterior probability of data. Cultivated sorghum were sorted according to the country of origin. Then, Malian cultivated sorghum were sorted according to the racial characterisation: *Gg* guinea gambicum–guineense, *Gm* guinea margaritiferum, *B* bicolor, *C* caudatum, *D* durra, *I* intermediate types. Sorghum from Guinea were also sorted according to the race and represented in the last cluster, called G. Each individual is represented by a *thin vertical line*, divided into *K* coloured fragments. Each *colour* represents one cluster and each fragment represents the membership fraction for one individual in *K* clusters. *Bold vertical lines*, shown in *black*, separate individuals of different groups

Within the joined set constituted by the WP and the CP from Mali and Guinea (data set 3), the different criteria (log likelihood, the congruence between runs, and ΔK) indicated two possible K values. The change in the second order of the log likelihood (ΔK) revealed two clear peaks for $K = 2$ and $K = 4$ (Supplementary Electronic File 4B). At $K = 2$, STRUCTURE analysis (Fig. 3a) showed that the cultivated and wild sorghum did not belong to distinct gene pools. Even at $K = 4$, wild sorghum did not constitute a separate cluster. Wild and guinea margaritiferum sorghum shared a very high average ancestry in cluster II at 67.2 and 91.0% for the wild and the guinea margaritiferum types, respectively (Fig. 3b; Table 3a). Wild sorghum also had a share of their ancestry (20.9%) in cluster IV, which also mainly grouped with cultivated sorghum belonging to the durra, caudatum and intermediate races. Finally, wild sorghum displayed low ancestry (9.6%) in cluster I, mainly comprising guinea gambicum–guineense accessions. We found that 31 wild sorghum (35.2% of them) and 74 guinea margaritiferum sorghum (83.1%) had an ancestry

Table 3 Percentage of average ancestry in a defined cluster for each sorghum type

	Data set 2				Data set 3			
	Clusters				Clusters			
	A	B	C	D	cI	cII	cIII	cIV
a) Within data set 2 (cultivated pool from Mali and Guinea) and data set 3 (cultivated and wild pools from Mali and Guinea)								
Bicolor	5.74	1.36	81.15	11.76	5.69	1.55	80.96	11.77
Caudatum	22.01	3.39	4.67	69.94	25.81	4.07	4.75	65.36
Durra	6.32	0.32	0.44	92.89	7.52	0.44	0.74	91.29
Guinea g.	93.07	2.97	1.41	2.55	93.01	2.83	1.62	2.54
Guinea m.	6.26	91.11	1.62	1.01	6.40	91.00	1.74	0.84
Intermediate type	8.70	4.42	14.34	72.55	9.50	4.83	14.62	71.04
Wild pool	–	–	–	–	9.60	67.19	2.27	20.93

Guinea g. and Guinea m. designate guinea gambicum–guineense and guinea margaritiferum, respectively

	Data set 5		
	Clusters		
	W1	W2	W3
b) Within data set 5: Wild from Mali, Guinea and genebank			
Wild Mali	47.37	41.57	11.05
Wild Guinea	2.64	68.34	29.04
Wild Genebank	9.58	3.02	87.38

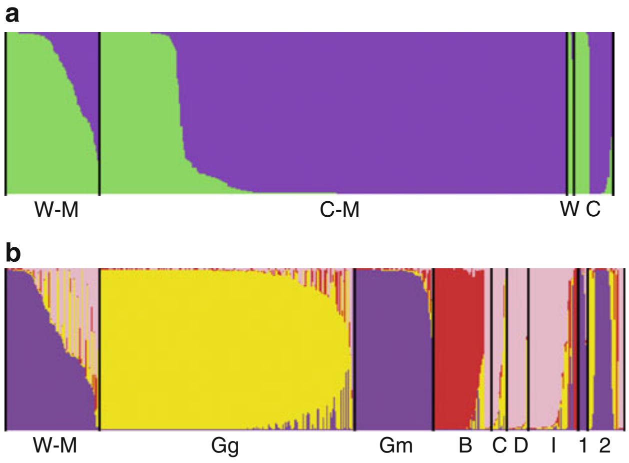

Fig. 3 Estimated population structure using the Bayesian model analysis implemented in STRUCTURE software within and between the cultivated and the wild pools, including the Malian and the Guinean collections. **a** Estimated population structure at $K = 2$. Within each country, wild and cultivated sorghum were sorted. *W–M*, *C–M*, *W* and *C* designate wild from Mali, cultivated from Mali, wild from Guinea and cultivated from Guinea, respectively. **b** Estimated population structure at $K = 4$. *W–M* designates wild sorghums from Mali. Malian cultivated sorghum were sorted according to the racial type: *Gg* guinea gambicum–guineense, *Gm* guinea margaritiferum, *B* bicolor, *C* caudatum, *D* durra, *I* intermediate types. Sorghum from Guinea were also sorted and represented in the last two clusters: *1* and *2* indicate wild and cultivated sorghum, respectively

exceeding 90% in cluster II. At this threshold of 90%, only one wild sorghum could be attributed to cluster I (guinea gambicum–guineense) and none could be attributed to clusters III or IV.

Within data set 4 (cultivated and wild sorghum from Mali and Guinea, and wild accessions from the genebank), four clusters were also identified (data not shown). Three clusters (guinea gambicum–guineense; guinea margaritiferum and wild sorghum; and bicolor accessions) remained unchanged. The wild accessions from the genebank had a high average ancestry (67.6%) in the fourth cluster, which mainly grouped with caudatum, durra and intermediate accessions (data not shown). These analyses confirmed the results obtained with F_{ST}: the wild accessions from the genebank shared greater ancestry with the durra, intermediate and caudatum races.

Wild accessions from the genebank, classified in the subspecies *verticilliflorum*, were considered as true wild. They exhibited greater diversity than wild and cultivated sorghum collected in the two countries (Table 1). They showed significantly higher gene diversity ($H_e = 0.737$) than wild types from Mali ($H_e = 0.588$) and Guinea ($H_e = 0.536$). The same trend was observed for allelic richness and private allelic richness. The distance-based method implemented in DARwin software was applied to the whole collection merging the cultivated and wild

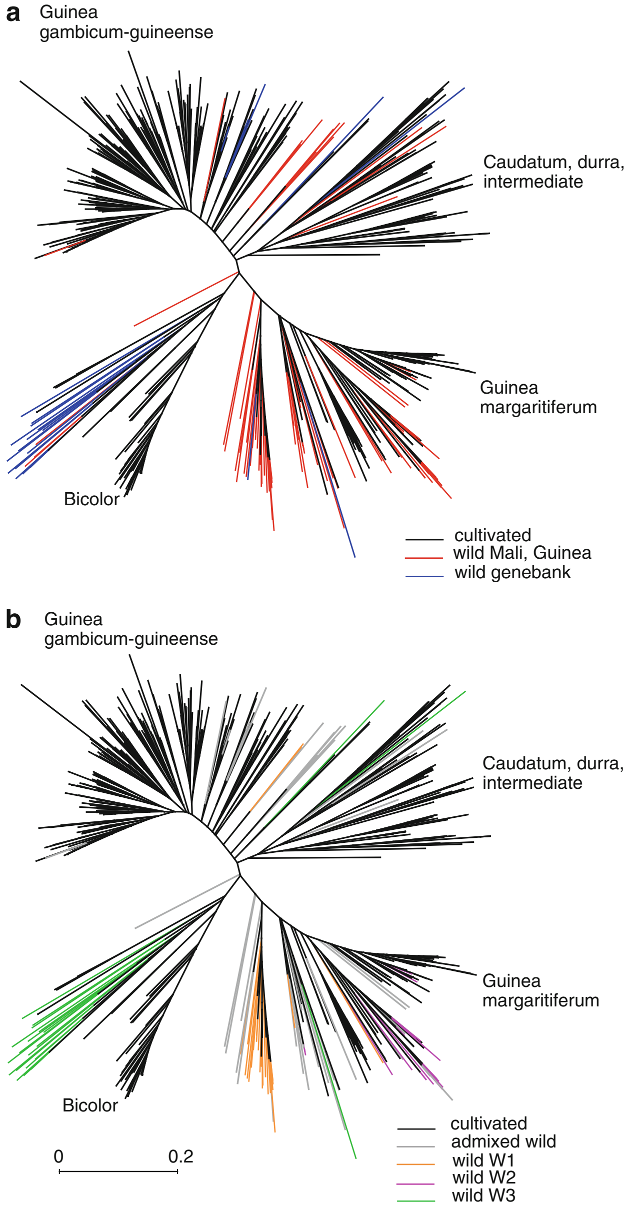

Fig. 4 Unrooted neighbour joining tree based on allelic data from 15 SSR loci among the total collection (471 accessions). **a** Projection of the different pools and origins. *Red* and *blue* represent wild forms from Mali or Guinea, and wild from the genebank, respectively. The cultivated pool is represented in *black*. **b** Projection of the STRUCTURE results obtained within the wild pool. Three groups were defined. An accession was attributed to a cluster when its ancestry in that cluster exceeded 90%. Admixed wild accessions (ancestry <90%) are shown in *grey*, whilst accessions attributed to cluster W1, W2 and W3 are shown in *orange*, *pink* and *green*, respectively

sorghum from Mali and Guinea and the wild sorghum from the genebank (Fig. 4a, and Supplementary Electronic File 5). This analysis confirmed the main results obtained with the Bayesian method and the pairwise F_{ST}: a strong racial structure within cultivated sorghum, a close proximity between wild types and guinea margaritiferum accessions, and the virtual absence of wild sorghum in the guinea gambicum–guineense cluster. Most wild accessions from the genebank also formed a separate group. However, this analysis may indicate slight differentiation into two groups within the Malian–Guinean wild pool: one located inside the guinea margaritiferum group, while the other was more distant from that group of cultivated sorghum.

We finally ran STRUCTURE analyses within the entire wild pool (data set 5 including wild sorghum collected in Mali and Guinea, and wild accessions from the genebank). We selected the smallest value of K(3). which captured the major structure in the data, assured congruence between the different runs and made it possible to assign a large share of individuals to distinct groups (Supplementary Electronic File 4C). At higher K values, we observed different clustering schemes and discrepancies between assignments of individuals. The three clusters showed a large congruence with the groups revealed by the distance-based method described above (Fig. 4b). At the 90% ancestry threshold, most of the wild genebank accessions were grouped in cluster W3 (shown in green in Fig. 4b; Table 3b), while the wild sorghum from Mali and Guinea were mainly clustered in two other groups. Cluster W2 (shown in pink) comprised wild types with a close relationship with guinea margaritiferum accessions, whilst cluster W1 (shown in orange) comprised wild sorghum that were more distant from these cultivated accessions. This structure in three genetic clusters within the wild pool was also confirmed by the principal coordinate analysis (PCoA) performed on the shared allele distance calculated for all pairs of wild forms (Supplementary Electronic File 6). Comparisons of mean values for the first axis (12.5% of the variance) showed significant differences between the three groups ($P < 0.0001$, Kruskal–Wallis test).

A Mann–Whitney test performed on longitudes indicated that wild types attributed to cluster W2 appeared to be located more in the west (average longitude −9.31) compared to cluster W1 (average longitude: −6.82, $P < 0.001$). Projection of individual estimated ancestries onto the map confirmed that the spatial distribution of these three clusters was uneven in Mali and Guinea. Wild types attributed to cluster W2 were mainly found in the southwestern zone where guinea margaritiferum were prevalent (Fig. 1b).

Pairwise pR_{ST} and R_{ST} estimates calculated among the genetic clusters identified by STRUCTURE analyses within the wild and the cultivated pools are presented in Table 4. Within the cultivated pool, all R_{ST} values that were significantly higher than pR_{ST} values involved cluster B (guinea margaritiferum). Between the two pools, all tests were significant between W2 and the clusters of cultivated sorghum, except for cluster B. These results suggested historical divergence with limited gene flow between

Table 4 Genetic differentiation among clusters defined by STRUCTURE analyses

	Cultivated				Wild		
	A	B	C	D	W1	W2	W3
A	–	0.430	0.588	0.397	0.426	0.476	0.503
B	**0.710**	–	0.606	0.421	0.299	0.232	0.478
C	0.682	**0.867**	–	0.429	0.622	0.600	0.495
D	0.487	**0.612**	0.394	–	0.229	0.349	0.168
W1	0.573	0.348	0.731	0.322	–	0.319	0.324
W2	**0.760**	0.256	**0.877**	**0.544**	0.297	–	0.270
W3	**0.692**	0.564	0.620	**0.322**	0.288	0.373	–

Above the diagonal are the permuted pR_{ST}. Below are the R_{ST} values with values significantly different from the permuted pR_{ST} indicated in bold. All calculations were performed with SPAGeDi

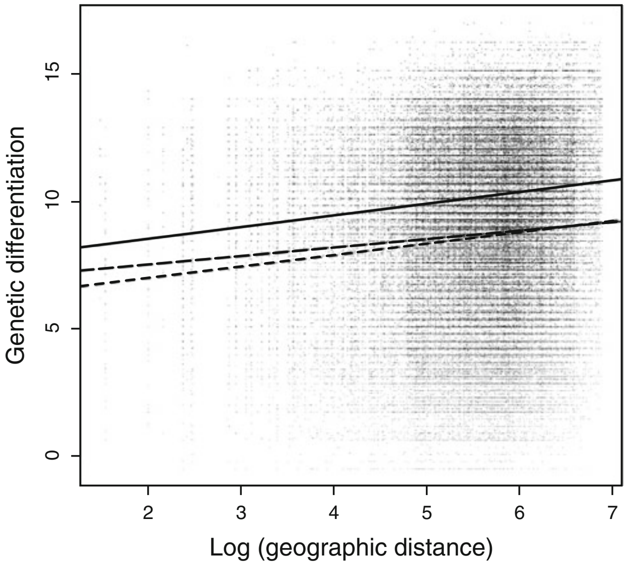

Fig. 5 Isolation by distance within and between cultivated and wild groups. Plots of the regressions of pairwise genetic differentiation between individuals ($\hat{a_r}$) against ln (*geographical distance*) within cultivated sorghum only (*short-dashed line*), within wild sorghum only (*long-dashed line*) and between cultivated and wild sorghum only (*solid line*)

cluster B and other cultivated sorghum, as well as between the latter ones and W2. Contrastingly, a gene flow between B and W2 was suggested. Likewise, other tests were not significant and indicated that migration rates could be greater than the mutation rate for all other comparisons.

The results of isolation by distance (IBD) within cultivated sorghum (all pairs of cultivated sorghum), within wild sorghum and between all pairs of cultivated and wild sorghum are presented in Fig. 5. A strong and significant IBD pattern ($P < 0.05$, Mantel test) was found for the different comparisons with regression slopes of 0.45 within cultivated sorghum, 0.21 within wild sorghum and 0.46 between them. Wild types showed lower levels of spatial genetic structure than cultivated sorghum, probably due to wider spatial dispersal (or smaller sample size and thus greater variance in slope estimate). Interestingly, the "between wild and cultivated" analysis also showed significant isolation by distance patterns, with a greater mean differentiation level than for the "within" analyses. A similar trend was observed for comparisons performed within cluster B (guinea margaritiferum), within wild cluster W2 and between them (data not shown).

Discussion

A strong botanical structure of genetic diversity within cultivated sorghum

We distinguished between three sorghum-growing regions in Mali according to a North–South rainfall gradient: (1) a Sudano–Sahelian zone (mean annual rainfall under 600 mm/year) where sorghum cultivation was in competition with pearl millet. Durra varieties, along with intermediate types with that race, were predominant. Some caudatum and guinea gambicum–guineense varieties were also found, but were rare and cultivated on small areas; (2) an intermediate zone, subjected to the Sudanian climate (from 600 to 900 mm), in which guinea gambicum–guineense were predominant and other types rare; (3) a southern zone (the Sudano–Guinean zone), characterised by rather extensive cropping and by the existence of forest habitats, with a higher mean annual rainfall (over 900 mm). Guinea gambicum–guineense varieties were still predominant, but guinea margaritiferum and bicolor sorghum were numerous and frequent, although grown on small areas.

Cultivated sorghum were poorly differentiated between climatic zones ($F_{ST} = 0.037$). The same trend was observed in Niger, Burkina Faso and Kenya (Deu et al. 2008; Barro-Kondombo et al. 2010; Mutegi et al. 2011). However, greater diversity was found in the Sudano–Sahelian zone, as shown by higher allelic richness, gene diversity and diversity of botanical types. A similar trend was previously observed in Burkina Faso within the guinea race (Barro-Kondombo et al. 2010). Various non-exclusive explanations can be put forward. Firstly, in drier areas, farmers rely on a large panel of varieties to make profitable all the available soils and niches with their cereals, which are the only crops that they can grow. It is a risk management strategy commonly adopted in traditional agricultural systems to increase resilience in the face of harsh heterogeneous environments (Teshome et al. 1999; Barro-Kondombo et al. 2010). Kouressy et al. (2008) highlighted the presence of both late and early maturing sorghum varieties in this zone, with late maturing varieties dedicated to low-lying areas of the landscape. Conversely, southwards, the lowlands are devoted to other crops (rice, etc.). Secondly, conservation programmes have been promoted in some villages of this zone, which has led to the development of a seed bank that is still active today. Lastly, this zone spans different ethnic groups, and ethnic factors are likely to explain a large share of the genetic structure in sorghum (Deu et al. 2008; Sagnard et al. 2008).

Our results obtained both with F_{ST} and the clustering methods confirmed that genetic diversity was chiefly associated with sorghum racial identity in Mali. The differentiation was especially marked (1) between the guinea and bicolor (sweet sorghum) races and a heterogeneous cluster of durra, caudatum and intermediate accessions, and (2) within the guinea race, between the guinea margaritiferum and guinea gambicum–guineense sub-races. A similar pattern of differentiation between races and within the guinea race had already been highlighted on a world scale using ex situ collections (Folkertsma et al. 2005; Deu et al. 2006; Brown et al. 2011), on a country scale in Niger (Deu et al. 2008) and on a local scale (Barnaud et al. 2007). The strong racial structure of genetic diversity probably reflected different evolutionary processes. Firstly, the history of domestication and human migrations has strongly influenced the pattern of genetic structure still observed today. Botanical races of cultivated sorghum have been intimately associated with human linguistic groups (Doggett 1988). Secondly, the spatial pattern of sorghum races in Mali and Niger related to different agricultural adaptation and local cropping systems might contribute to the limited admixture observed between these botanical types. Thirdly, the suspected existence of biological barriers to gene flow among landraces (differences in phenology and mating system, panicle morphology, etc.) might also contribute to the limited gene flow observed between the botanical types (Barnaud et al. 2007). Lastly, farmers' practices could be more important for maintaining landrace identity. When several varieties from different botanical races are planted closely together in fields, favouring a high potential gene flow between them, farmers in northern Cameroon have been found to exert strong selection against intermediate types and to preserve the genetic distinctiveness of their local landraces (Barnaud et al. 2007). Studies on the extent and direction of gene flow between landraces, linked to farmers' practices, are currently in progress on a local scale in Mali.

Guinea margaritiferum and bicolor accessions harboured low genetic diversity. Molecular markers highlighted that varieties defined as different, based on their vernacular name or collection site, were in fact very close. Some varieties present in the neighbouring villages probably had the same origin and might have been released by an informal seed system. Indeed, the expansion of bicolor varieties, which are mainly sweet sorghum, ought to be recent and based on a few varieties. A similar recent expansion was observed in Niger for sweet sorghum (Bezançon et al. 2009). Moreover, in Mali, these types are only cultivated on small areas and mainly cut before maturity when their sweet stem can be chewed. Farmers just need to select a few panicles and, in extreme cases, a single panicle to constitute their seed batch for the next growing season. These specific practices, as well as associated genetic drift and a narrow genetic base, probably explain their strong differentiation from each of the other racial types. Lastly, a phylogeographical pattern was revealed by R_{ST} analyses for guinea margaritiferum compared to each one of the cultivated clusters indicating that the former have long since diverged from the other clusters and exchanged migrants at a low rate.

Distribution, structure and nature of wild and weedy sorghum and conservation strategy

Wild and weedy sorghum were widespread throughout the sorghum agricultural zones in Mali. They were found in most of the villages visited, at relatively short distances from sorghum fields, enabling pollen-mediated gene flow. However, their abundance seemed variable depending on the regions. Farmers considered wild and weedy sorghum as major weeds, mainly in the Sudano–Guinean zone, the most humid and forested part of the country. In the southern and southwestern zones, some farmers declared that the invasion of their fields by wild sorghum could be a reason for field abandonment. In the Sudanian zone, they were not so abundant and principally restricted to manured gardens and threshing zones. In the Sudano–Sahelian zone, although abundant, manual weeding is enough to control them, except in some households where labour is in short supply. Consequently, in this zone, farmers did not generally consider them as a threat to cereal agriculture.

Based on the geographical distribution of the wild sorghum races (de Wet et al. 1970), we expected to find wild sorghum belonging to the arundinaceum race in the southern and southwestern parts of the country and to the verticilliflorum and aethiopicum races in northern Mali. However, "wild races grade morphologically and ecologically" (de Wet 1978) and therefore are difficult to distinguish between on the basis of their morphology alone (de Wet et al. 1976). It is also true that weedy forms (ssp. *drumondii*) can be very close to the wild races, based on phenotypic traits. Quantitative morphological traits do not allow any clear distinction to be made between the different races or subspecies within wild and weedy Kenyan sorghum (Mutegi et al. 2010). De Wet et al. (1976) also pinpointed that weedy sorghum never occupies stable habitats, but are mainly found in cultivated fields and other human-disturbed areas. Most of the wild and weedy sorghum we collected came from this type of habitat. Their agromorphological characterisation indicated that they mainly exhibited intermediate traits between wild and cultivated sorghum (data not shown). We thus suspected an introgressed status of most of the wild and weedy sorghum collected in the region. Our molecular data supported this hypothesis and indicated that the wild and weedy forms we collected would be better classified as weedy types belonging to the subspecies *drumondii*. Firstly, they were characterised by reduced genetic diversity compared to the wild genebank accessions, as measured with H_e, R_s and private allelic richness. They did not harbour greater genetic diversity compared to the cultivated pool, a finding which was not congruent with previous comparisons between wild and cultivated sorghum conducted in Kenya or on a larger scale (see Mutegi et al. 2011 for a review). However, our findings were congruent with those of two studies (Mutegi et al. 2011; Barnaud et al. 2009). In the former, similar gene diversity was harboured by cultivated and wild sorghum within a region of Kenya where more putative crop–wild hybrids were found. In the latter, the intermediate weedy morphotypes also showed higher observed heterozygosity compared to the domesticated forms in a village in northern Cameroon. Moreover, the higher observed heterozygosity found in the wild pool collected in Mali and Guinea hinted at a higher outcrossing rate facilitating crop-to-wild gene flow. Secondly, Bayesian and PCoA analyses indicated a genetic structure within the total wild pool and a differentiation between the wild genebank accessions, grouped in the W3 cluster, and most of the wild sorghum collected in Mali and Guinea.

Western African wild sorghum are still poorly represented in international genebanks, although it may be a valuable source of useful genes for crop improvement. They could actually be threatened by habitat modifications and especially the extension of cultivated lands. Nowadays, few populations of wild sorghum can be found in some regions of Ethiopia (Ejeta and Grenier 2005). The depletion of wild sorghum populations, reported by these authors, has been attributed to major human changes that occurred in recent years. The growing human population pressure, political decision-making and recurrent drought have forced farmers to cultivate the most marginal lands formerly occupied by wild sorghum. In addition, farmers' practices in relation to their management of sorghum wild relatives (strong selective rouging) largely contributed to the decrease in wild populations in the Hararghe region of Ethiopia (Tesso et al. 2008). Conversely, in other regions of Ethiopia and in northern Sudan, where human population pressure is lower, wild sorghum populations have been conserved (Ejeta and Grenier 2005). Habitat loss for wild sorghum in Ethiopia could have significantly reduced the apparent potential gene flow between cultivated and wild sorghum and led to genetic erosion in both types. In our study we were able to identify some Malian and Guinean wild sorghum that were attributed to the wild genebank accessions cluster. Of them, two were "forest grasses", found in natural habitats in Guinea, corresponding to the description of Doggett (1988) for the arundinaceum race, and one originating from Pays Dogon could be assigned to the aethiopicum race. These wild types, collected in the two extreme regions, probably constitute populations that are more "wild" and have not been subjected to introgressions from cultivated forms over various generations. Therefore, our first results indicated that the northern and southern margins of sorghum-growing areas in western Africa should be considered as priority targets for future collections and the development of in situ conservation programmes. Lastly, identifying endangered populations of wild sorghum could be investigated using statistical methods and molecular tools that are valuable for detecting recent bottlenecks in the absence of ancient historical data. Recently, this method was successfully applied to two Kenyan wild sorghum populations (Muraya et al. 2010). Conservation planning for wild species should be considered on a case-by-case basis in the light of social changes and agricultural practices.

Genetic relationships between cultivated and wild sorghum

Firstly, a closer genetic relationship was detected between wild sorghum and the guinea margaritiferum type ($F_{ST} = 0.17$) compared to the strong differentiation detected between wild and guinea gambicum–guineense types ($F_{ST} = 0.32$), which were largely dominant among the number of landraces in cultivated areas. This closeness was confirmed by both the Bayesian and the distance methods. Of the wild forms collected in Mali and Guinea, 35% were clearly attributed to the guinea margaritiferum cluster when four populations were defined with

STRUCTURE software. These weedy forms, which were further attributed to the wild cluster W2, were predominant in the south/southwestern region of Mali where guinea margaritiferum are grown. Three explanations concerning the nature of these weedy forms, which are not necessarily exclusive, could be put forward: (1) These weedy forms have been produced by endoferality from the domesticated forms of guinea margaritiferum. Endoferality is a process of dedomestication, which does not imply gene flow (Gressel 2005). Even though endoferality has been reported in various crop species, it was not really suspected in sorghum (Ejeta and Grenier 2005). The fact that selfing management of margaritiferum varieties did not produce spontaneous wild forms (Chantereau, personal communication) makes this hypothesis unlikely. (2) These weedy types could be derived from hybridisation between guinea margaritiferum varieties and the relict forms of local wild sorghum, which have differentially evolved under natural selection to adapt to local environments. Hybridisation events could be regular and repeated or more accidental and scarce as true wild sorghum populations seem to have been severely reduced and could be actually considered as a rarity in Mali. This hypothesis would imply the existence of subsequent post-domestication crop-to-wild gene flow. (3) The gene flow observed today may only reflect an historical process and indicate possible separate domestication of guinea margaritiferum varieties from the local wild pool in western Africa. In this last case, the common assumption of a monophyletic origin of cultivated sorghum that was domesticated in the northeastern quadrant of Africa around 8,000 years ago could be revisited. R_{ST} analyses favoured the hypothesis of a preferential gene flow between the type guinea margaritiferum and these weedy forms versus the hypothesis of an historical divergence without gene flow. Spatial analysis between these two clusters also revealed a significant pattern of isolation by distance, which is indicative of ongoing gene flow (at least at short geographical distance). Potential gene flow between the guinea margaritiferum types and these weedy types was therefore highlighted. Moreover, post-domestication crop-to-wild and wild-to-crop gene flow may have really been enhanced by farmers' practices. In an intensive study conducted in one agro-ecosystem typical of southern Mali (Mande region), farmers showed great difficulties in discriminating between weedy sorghum and guinea margaritiferum varieties before flowering due to a few shared vegetative traits (unpublished information). Weeding and selective uprooting could therefore be less efficient in or near margaritiferum fields, enabling a potential crop-to-weed gene flow with this group of cultivated sorghum. Several farmers' practices may also favour gene flow in the other direction, through the integration of hybrid seeds resulting from weed-to-crop gene flow in the traditional seed system. Farmers of this surveyed village pay no particular attention when selecting their seed for the next growing season in fields with low density of weedy forms. Weedy types and cultivated varieties have exhibited flowering synchronicity in many fields (unpublished data). Weedy types have also displayed a large flowering window. Both our genetic study and surveys thus suggest that dissemination of GM in traditional agro-ecosystems of southern Mali, where, weedy sorghum dynamics is not efficiently controlled by farmers, could lead to the escape of transgenes into weedy sorghum. Further studies on mating system and phenology of wild and weedy types in different farming systems could be a benefit for biosafety regulators.

Pairwise R_{ST} indicated a potential gene flow between the W1 wild types, and all cultivated forms as well as with the W2 wild types (Table 4). Thus, these W1 forms seem to have a complex nature that needs more investigation. A local survey on an agro-ecosystem level could help us to pinpoint the extent and direction of gene flow and the contribution of each racial cultivated type to that gene flow, and to evaluate the different rates of introgression within the wild forms. Finally, was the spatial isolation of the W1 and W2 wild forms the result of local ecological adaptation or/and different farmers' practices? This stresses the need for the integration of different disciplines such as social and GIS sciences to gain a clearer understanding of the processes and factors underlying the distribution of genetic diversity in wild and/or cultivated pools.

Overall, our results strongly suggested gene flow to some extent between the guinea margaritiferum type and the wild pool, and probably a more limited contribution of bicolor, guinea gambicum–guineense, and durra and caudatum types to crop-to-wild gene flow. The genetic relatedness between wild types and guinea margaritiferum accessions could be the result of both sorghum domestication history and preferential post-domestication crop-to-wild gene flow. Maternally inherited molecular markers (chloroplastic or mitochondrial) should help us to estimate the relative contributions of these two evolutionary factors. Further research is also needed to precisely evaluate the extent and direction of gene flow in different agro-ecosystems. This would allow for the scaling-up of some processes driving gene flow from a village to country scale.

Acknowledgments This paper is dedicated by the co-authors to Dr. Fabrice Sagnard, the principal investigator of the project "Environmental Risk Assessment of Genetically Engineered Sorghum in Mali and Kenya", who passed away in November 2008. We are grateful to the Biotechnology and Biodiversity Interface and Plant Biosafety Systems program funded by USAID for supporting the research activities reported here. We also acknowledge the Fonds Français pour l'Environnement Mondial (FFEM), which supported part of this project. We thank all the farmers from Mali and Guinea who shared their varieties and knowledge. The collection was performed in full compliance with the Convention on Biological Diversity. The authors thank the IER (Institut d'Economie Rurale, Mali) for the storage of all collected samples in view of future utilisation. We are also grateful to Dr. Ousmane Köita (professor at the University of Bamako) for hosting the master's student in charge of DNA extractions in his laboratory. The authors thank J. Chantereau for interesting discussions and comments on the manuscript and the two anonymous reviewers for valuable suggestions. A part of this work was carried out using the resources of the Computational Biology Service Unit from Cornell University.

References

Arnold MJ (2004) Natural hybridization and the evolution of domesticated, pest and disease organisms. Mol Ecol 13:997–1007

Barnaud A, Deu M, Garine E, McKey D, Joly H (2007) Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. Theor Appl Genet 114:237–248

Barnaud A, Deu M, Garine E, Chantereau J, Bolteu J, Koïda EO, Mc Key D, Joly HI (2009) A weed–crop complex in sorghum: the dynamics of genetic diversity in a traditional farming system. Amer J Bot 96(10):1869–1879

Barro-Kondombo C, Sagnard F, Chantereau J, Deu M, vom Brocke K, Durand P, Gozé E, Zongo JD (2010) Genetic structure among sorghum landraces as revealed by morphological variation and microsatellite markers in three agroclimatic regions of Burkina Faso. Theor Appl Genet 120:1511–1523

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2002) Genetix 4.04, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France. (Available from http://www.univ-montp2.fr/~genetix/genetix/genetix.htm)

Bezançon G, Pham JL, Deu M, Vigouroux Y, Sagnard F, Mariac C, Kapran I, Mamadou A, Gérard B, Ndjeunga J, Chantereau J (2009) Changes in the diversity and geographic distribution of cultivated millet (*Pennisetum glaucum* [L.] R. Br.) and sorghum (*Sorghum bicolor* (L.) Moench) varieties in Niger between 1976 and 2003. Genet Resour Crop Evol 56:223–236

Brown PJ, Myles S, Kresovich S (2011) Genetic support for phenotype-based racial classification in Sorghum. Crop Sci 51:224–230

de Wet JMJ (1978) Systematics and evolution of *Sorghum* sect. *Sorghum* (Gramineae). Am J Bot 65(4):477–484

de Wet JMJ, Harlan JR (1971) The origin and domestication of *Sorghum bicolor*. Econ Bot 25:129–134

de Wet JMJ, Harlan JR, Price EG (1970) Origin of variability in the spontanea complex of *Sorghum bicolor*. Am J Bot 57(6):704–707

de Wet JMJ, Harlan JR, Price EG (1976) Variability in *Sorghum bicolor*. In: Harlan JR, de Wet JMJ, Stemler ABL (eds) Origins of African plant domestication. Mouton, The Hague, pp 453–463

Deu M, Hamon P, Chantereau J, Dufour P, D'Hont A, Lanaud C (1995) Mitochondrial DNA diversity in wild and cultivated sorghum. Genome 38:635–645

Deu M, Rattunde F, Chantereau J (2006) A global view of genetic diversity in cultivated sorghums using a core collection. Genome 49:168–180

Deu M, Sagnard F, Chantereau J, Calatayud C, Hérault D, Mariac C, Pham JL, Vigouroux Y, Kapran I, Traoré PS, Mamadou A, Gérard B, Ndjeunga J, Bezançon G (2008) Niger-wide assessment of in situ sorghum genetic diversity with microsatellite markers. Theor Appl Genet 116:903–916

Deu M, Sagnard F, Chantereau J, Calatayud C, Vigouroux Y, Pham JL, Mariac C, Kapran I, Mamadou A, Gérard B, Ndjeunga J, Bezançon G (2010) Spatio-temporal dynamics of genetic diversity in *Sorghum bicolor* in Niger. Theor Appl Genet 120:1301–1313

Doggett H (1988) Sorghum, 2nd edn. Longman Scientific and Technical, London

Ejeta G, Grenier C (2005) Sorghum and its weedy hybrids. In: Gressel J (ed) Crop ferality and volunteerism. Taylor & Francis, Boca Raton, pp 123–135

Ellstrand NC (2003) Current knowledge on gene flow in plants: implications for transgene flow. Phil Trans R Soc Lond B 358:1163–1170

Ellstrand NC, Prentice HC, Hancock JF (1999) Gene flow and introgression from domesticated plants into their wild relatives. Annu Rev Ecol Syst 30:539–563

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Folkertsma RT, Rattunde HFW, Chandra S, Soma Raju W, Hash CT (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, Papa R (2003) Possible effects of (trans)gene flow from crops on the genetic diversity from landraces and wild relatives. Environ Biosafety Res 2:89–103

Goudet J (2002) FSTAT, a program to estimate and test gene diversity and fixation indices (version 2.9.3.2. Available from http://www.unil.ch/izea/softwares/fstat.html)

Gressel J (2005) Introduction—the challenges of ferality. In: Gressel J (ed) Crop ferality and volunteerism. Taylor & Francis, Boca Raton, pp 1–7

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Hardy OJ, Charbonnel N, Fréville H, Heuertz M (2003) Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. Genetics 163:1467–1482

Harlan JR, de Wet JMJ (1971) Toward a rational classification of cultivated plants. Taxon 20(4):509–517

Harlan JR, de Wet JMJ (1972) A simplified classification of cultivated sorghum. Crop Sci 12:172–176

Jarvis DI, Hodgkin T (1999) Wild relatives and crop cultivars: detecting natural introgression and famer selection of new genetic combinations in agroecosystems. Mol Ecol 8:159–173

Jarvis A, Lane A, Hijmans RJ (2008) The effect of climate change on crop wild relatives. Agric Ecosyst Environ 126:13–23

Kalinowski ST (2005) HP-RARE1.0: a computer program for performing rarefaction on measures of allelic richness. Mol Ecol Notes 5:187–189

Kameswara Rao N, Reddy LJ, Bramel PJ (2003) Potential of wild species for genetic enhancement of some semi-arid food crops. Genet Resour Crop Evol 150:707–721

Kouressy M, Traoré S, Vaksmann M, Grum M, Maikano I, Soumaré M, Traoré PS, Bazile D, Dingkuhn M, Sidibé A (2008) Adaptation des sorghos du Mali à la variabilité climatique. Cah Agric 17(2):95–100

Lewis PO, Zaykin D (2001) Genetic data analysis: computer program for the analysis of allelic data. Available from http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php

Mariac C, Robert T, Allinne C, Remigereau MS, Luxereau A, Tidjani M, Seyni O, Bezançon G, Pham JL, Sarr A (2006a) Genetic diversity and gene flow among pearl millet crop/weed complex: a case study. Theor Appl Genet 113:1003–1004

Mariac C, Luong V, Kapran I, Mamadou A, Sagnard F, Deu M, Chantereau J, Gérard B, Ndjeunga J, Bezancon G, Pham JL, Vigouroux Y (2006b) Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers. Theor Appl Genet 114:49–58

Martel C, Réjasse A, Rousset F, Bethenod M-T, Bourguet D (2003) Host-plant-associated genetic differentiation in Northern French populations of the European corn borer. Heredity 90:141–149

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci USA 99:6080–6084

Maxted N, Ford-Lloyd BV, Jury S, Kell S, Scholten M (2006) Towards a definition of a crop wild relative. Biodivers Conserv 15:2673–2685

Michalakis Y, Excoffier L (1996) A genetic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061–1064

Muraya MM, Sagnard F, Parzies HK (2010) Investigation of recent populations bottlenecks in Kenyan wild sorghum populations (*Sorghum bicolor* (L.) Moench ssp. *verticilliflorum* (Steud.) De Wet) based on microsatellite diversity and genetic disequilibria. Genet Resour Crop Evol 57:995–1005

Mutegi E, Sagnard F, Muraya M, Kanyenji B, Rono B, Mwongera C, Marangu C, Kamau J, Parzies H, de Villiers S, Semagn K, Traoré PS, Labuschagne M (2010) Ecogeographical distribution of wild, weedy and cultivated *Sorghum bicolor* (L.) Moench in Kenya: implications for conservation and crop-to-wild gene flow. Genet Resour Crop Evol 57:243–253

Mutegi E, Sagnard F, Semagn K, Deu M, Muraya M, Kanyenji S, de Villiers S, Kiambi D, Herselman L, Labuschagne M (2011) Genetic structure and relationships within and between cultivated and wild sorghum (*Sorghum bicolor* (L.) Moench) in Kenya as revealed by microsatellite markers. Theor Appl Genet 122:989–1004

Papa R, Gepts P (2003) Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. Theor Appl Genet 106:239–250

Papa R, Acosta J, Delgado-Salinas A, Gepts P (2005) A genome-wide analysis of differentiation between wild and domesticated *Phaseolus vulgaris* from Mesoamerica. Theor Appl Genet 111:1147–1158

Perrier X, Jacquemoud-Collet JP (2006) DARwin software. http://darwin.cirad.fr/darwin

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Risterucci AM, Grivet L, N'Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet 101:948–955

Rousset F (1999) Genetic differentiation within and between two habitats. Genetics 151:397–407

Rousset F (2000) Genetic differentiation between individuals. J Evol Biol 13:58–62

Rousset F (2008) GENEPOP' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour 8:103–106

Sagnard F, Barnaud A, Deu M, Barro C, Luce C, Billot C, Rami JF, Bouchet S, Dembélé D, Pomiès V, Calatayud C, Rivallan R, Joly H, vom Brocke K, Touré A, Chantereau J, Bezançon G, Vaksmann M (2008) Analyse multiéchelle de la diversité génétique des sorghos: compréhension des processus évolutifs pour la conservation in situ. Cah Agric 17(2):114–121

Slatkin M (1995) A measure of population subdivision based on microsatellite allelic frequencies. Genetics 139:457–462

Snowden JD (1936) The cultivated races of sorghum. Adlard, London, pp 1–274

Teshome A, Fahrig L, Torrance JK, Lambert JD, Arnason TJ, Baum BR (1999) Maintenance of sorghum (*Sorghum bicolor*, Poaceae) landrace diversity by farmers' selection in Ethiopia. Econ Bot 53:79–88

Tesso T, Kapran I, Grenier C, Snow A, Sweeney P, Pedersen J, Marx D, Bothma G, Ejeta G (2008) The potential for crop-to-wild gene flow in sorghum in Ethiopia and Niger: a geographic survey. Crop Sci 48:1425–1431

Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, Smith JSC, Jaqueth J, Smith OS, Doebley J (2005) An analysis of genetic diversity across the maize genome using microsatellites. Genetics 169:1617–1630

Zizumbo-Villarreal D, Colunga-Garcia Marin P, Payro de la Cruz E, Delgado-Valerio P, Gepts P (2005) Population structure and evolutionary dynamics of wild–weedy–domesticated complexes of common bean in a Mesoamerican region. Crop Sci 45:1073–1083