

Published in *Theoretical and Applied Genetics* [Volume 122, Number 5](#), 989-1004, DOI:
<http://dx.doi.org/10.1007/s00122-010-1504-5>

This is author version postprint archived in the official open access repository of ICRISAT
www.icrisat.org

Genetic structure and relationships within and between cultivated and wild sorghum (*Sorghum bicolor* (L.) Moench) in Kenya as revealed by microsatellite markers

^{1,2,3}E. Mutegi, ^{2,4}F. Sagnard F. ⁵K. Semagn, ⁴M. Deu, ⁶M. Muraya, ⁷B. Kanyenji, ²S. de Villiers, ²D. Kiambi, ⁸L. Herselman, ⁸M. Labuschagne

¹Kenya Agricultural Research Institute (KARI), National Genebank, P.O. Box 30148, Nairobi, Kenya.

²International Crops Research Institute for the Semi-Arid Tropics (ICRISAT-Nairobi), P.O. Box 39063-00623, Nairobi, Kenya.

³ Present address: Ohio State University, Department of Evolution, Ecology, and Organismal Biology, 318 W. 12th Ave., Columbus, OH 43210-1293, USA.

⁴CIRAD, UMR DAP, F-34398 Montpellier , France

⁵International Maize and Wheat Improvement Center (CIMMYT), P.O. Box 1041, Village Market-00621, Nairobi, Kenya.

⁶Leibniz Institute of Plant Genetics and Crop Plant Research, Corrensstrasse 3, 06466 Gatersleben, Germany

⁷KARI - Embu Research Station, PO Box 27, Embu – Kenya.

⁸Department of Plant Sciences, University of the Free State, P.O. Box 339, Bloemfontein, South Africa.

Corresponding author e-mail: e_ mutegi.1@yahoo.com

Abstract

Understanding the extent and partitioning of diversity within and among crop landraces and their wild/weedy relatives constitutes the first step in conserving and unlocking their genetic potential. This study aimed to characterize the genetic structure and relationships within and between cultivated and wild sorghum at country scale in Kenya, and to elucidate some of the underlying evolutionary mechanisms. We analyzed a total of 439 individuals comprising 329 cultivated and 110 wild sorghums using 24 microsatellite markers. We observed a total of 295 alleles across all loci and individuals, with 257 different alleles being detected in the cultivated sorghum gene pool and 238 alleles in the wild sorghum gene pool. We found that the wild sorghum gene pool harboured significantly more genetic diversity than its domesticated counterpart, a reflection that domestication of sorghum was accompanied by a genetic bottleneck. Overall, our study found close genetic proximity between cultivated sorghum and its wild progenitor, with the extent of crop-wild divergence varying among cultivation regions. The observed genetic proximity may have arisen primarily due to historical and/or contemporary gene flow between the two congeners, with differences in farmers' practices explaining inter-regional gene flow differences. This suggests that deployment of transgenic sorghum in Kenya may lead to escape of transgenes into wild-weedy sorghum relatives. In both cultivated and wild sorghum, genetic diversity was found to be structured more along geographical level than agro-climatic level. This indicated that gene flow and genetic drift contributed to shaping the contemporary genetic structure in the two congeners. Spatial autocorrelation analysis revealed a strong spatial genetic structure in both cultivated and wild sorghums at the country scale, which could be explained by medium- to long-distance seed movement.

Keywords: Genetic structure, relationships, crop-wild gene flow, *Sorghum bicolor*, autocorrelation, Kenya

Introduction

Plants demonstrate a wide range of diversity in their morphology, adaptation and ecology, the product of many years of evolutionary divergence and diversification. Characterising the extent and partitioning of this diversity across populations and/ or distribution ranges for target taxa, coupled with an understanding of the mechanisms through which it arises and maintained, has been of interest to plant genetic resources conservation and crop improvement programs alike (Frankel and Hawkes 1975; Frankel et al. 1995). In particular, assessment of the level and patterns of genetic diversity is useful for several purposes, including (i) determining the level of genetic variability for facilitating identification of subsets of core or mini-core collections with possible utility for specific breeding purposes (Mohammadi and Prasanna 2003); (ii) estimating any possible loss of genetic diversity during conservation programs; (iii) assisting the selection of diverse parental combinations to create segregating progenies with maximum genetic variability for further selection (Barrett and Kidwell 1998); and (iv) estimating the relative strengths of the evolutionary forces (mutation, natural selection, migration or gene flow, genetic drift) and population properties such as population size, breeding system, population structure, and dispersal mechanisms. Moreover, information on the genetic relationships between crops and their wild/weedy relatives is useful in estimating the extent and dynamics of crop-wild gene flow. Although gene flow between crops and their relatives has been taking place since the dawn of agriculture (Ellstrand et al. 1999; Haygood et al. 2003), there are fears that transgenes will escape from genetically modified (GM) crops to sexually compatible wild and weedy relatives via gene flow. Depending on the relative fitness conferred by such transgenic traits in recipient wild/weedy relatives, potential harmful consequences include increased invasiveness, weediness, genetic erosion and in extreme cases extinction of populations (Ellstrand 1992; Snow and Moran-Palma 1997; Bhatia and Mitra 2003; Conner et al. 2003; Haygood et al. 2003; Cleveland and Soleri 2005; Thies and Devare 2007; Auer 2008; Chandler and Dunwell 2008).

Sorghum (*Sorghum bicolor* L. Moench.) is one of the world's most important cereals. Sorghum was domesticated and diversified in Africa before moving to other parts of the world (Dogget 1988) and continues to play an important food security role in Africa. In 2007 for example, over 40 million ha of land was dedicated to sorghum production globally, out of which 60 percent was found in Africa (FAO 2008). Besides its use as a cereal crop sorghum is extensively used for fodder, construction material, brooms, syrup and beer. In Kenya, sorghum is grown in all but one administrative province. It is an important food crop and dietary staple in the country's arid and semi-arid lands which account for over 80 percent of the total land area. A wide diversity of sorghum landraces is cultivated under equally diverse agro-climatic conditions and practices by subsistence farmers in different communities of Kenya (Mutegi et al. 2010). Moreover, morphologically and geographically diverse wild relatives of domesticated sorghum in the primary and tertiary gene pools are known to occur in the country (Clayton and Renvoize 1982; Mutegi et al. 2010). Wild relatives of sorghum are recognised as broad genetic base reservoirs and potential sources for resistance and adaptation traits in breeding programs (Gurney et al. 2002; Kamala et al. 2002; Reed et al. 2002; Rao Kameswara et al. 2003; Rich et al. 2004) and deserve special conservation attention. Unfortunately, we are not aware of any documented studies on the extent and partitioning of diversity in cultivated sorghum or its wild relatives at national scale in Kenya. Such information is important for designing appropriate conservation and utilization programs.

Cultivated sorghum (*S. bicolor* ssp. *bicolor*) is taxonomically conspecific with its proposed wild progenitor (*S. bicolor* ssp. *verticilliflorum*) and the stabilised weedy derivative of their hybridisation (*S. bicolor* ssp. *drummondii*) (Harlan and De Wet 1972; Dogget 1988). All subspecies of *S. bicolor* are inter-fertile under sympatric conditions, leading to a continuum of wild-weedy-domesticated complex forms that have been documented to occur in many sorghum growing parts of Africa (Dogget and Majisu 1968; Dogget 1988; Tesso et al. 2008; Mutegi et al. 2010). Moreover, cultivated and wild sorghum occupy diverse ecological landscapes and have over the years been subjected to diverse biotic and abiotic selection pressures across their geographic range. Wide genetic diversity is therefore expected in the landraces of cultivated sorghum and their wild-weedy relatives in Africa.

Levels and patterns of diversity within and between cultivated and wild sorghum gene pools have been reported before (e.g. Morden et al. 1990; Aldrich and Doebley 1992; Aldrich et al. 1992; Cui et al. 1995; Deu et al. 1995; Casa et al. 2005). These previous studies have shown that (i) there is low to moderate genetic differentiation among cultivated and wild sorghum gene pools, (ii) portions of the wild gene pool most genetically similar to cultivars originated in central-north-east Africa and (iii) genetic diversity is greater in wild-weedy than in cultivated sorghums. Little is known about the extent, patterns and direction of introgression between cultivated and wild sorghum. Moreover, most of these previous results were obtained from *ex situ* collections from genebanks and need validation using exhaustive samples obtained *in situ* at different spatial scales. This is especially needed for Africa, the center of origin and primary diversification of sorghum. Attempts have been made to use *in situ* collected samples but such studies have been limited to separate investigations of genetic diversity and structure in either cultivated sorghum (Djè et al. 1998; Djè et al. 1999; Ayana et al. 2000b; Ayana et al. 2001; Ghebru et al. 2002; Barnaud et al. 2007; Deu et al. 2008; Sagnard et al. 2008; Barro-Kondombo et al. 2010) or its closest wild relatives (Ayana et al. 2000a). Our study applied microsatellite markers to analyse cultivated sorghum and its closest wild relatives sampled from different growing regions in Kenya, in order to elucidate patterns of diversity within and among the two congeners, and to shed more light on their genetic and evolutionary relationships. We sought to address three questions: (i) What is the extent of diversity within cultivated and wild gene pools of sorghum in Kenya?; (ii) Are cultivated and wild sorghum gene pools genetically differentiated?; and (iii) How is genetic diversity in cultivated and wild sorghum gene pools structured?

Materials and methods

Plant materials

Cultivated and wild sorghum seed samples were collected in farmer fields in the crop's four main growing areas of Kenya: (i) Turkana, in the northern parts of the Rift Valley bordering Sudan and Ethiopia; (ii) western/Nyanza region covering the Kisii Highlands and lowlands around Lake Victoria; (iii) eastern/central region covering the Highlands east of Mt. Kenya and the drier lowlands of Meru, Kitui and Machakos administrative districts; and (iv) coastal areas of the country including the Taita Hills and adjacent areas as well as the farming systems in the Indian Ocean hinterlands (Fig. 1). Three collection trips were undertaken between June 2006 and July 2007 in order to capture differences in cropping seasons amongst the four growing regions. Passport data associated with each collected sample and farmer knowledge of cultivated varieties as well as of wild and weedy sorghum distribution, ecology and dynamics were recorded. Geographic coordinates and elevation data of each collection point were recorded using a handheld global positioning system (GPS) (eTrex Summit HC, Garmin). In addition, 93 georeferenced samples of cultivated sorghum were obtained from the National Genebank of Kenya (see Fig. 1) to cover the north-eastern region bordering Ethiopia (34 samples), central Rift Valley (55 samples) and parts of Kisii highlands in western/Nyanza (4 samples) that were not covered in the collection trips due to logistical constraints. In total, 439 samples comprising 110 wild and 329 cultivated sorghum varieties were assembled. The highest number of samples for cultivated sorghum originated from eastern/central (90), followed by western/Nyanza (72), Rift Valley (55), Turkana (42), Coast (36) and north-eastern (34) regions. For wild sorghum, the highest number of samples originated from eastern/central (41), followed by Coast (39), Turkana (17) and western/Nyanza (13) regions. Overall, samples were representative of Kenya's sorghum growing agro-climatic and ethno-linguistic diversity. A copy of each sample collected in this study was deposited at the National Genebank of Kenya for long-term conservation and future utilization. Detailed description of the collection, including the ecogeographical variation across the country is published elsewhere (Mutegi et al. 2010).

DNA isolation and genotyping

Seeds from each sample of cultivated and wild sorghum were grown at room temperature for two weeks in potted plastic trays in the laboratory. To break seed dormancy in wild sorghum, glumes were removed using a scalpel blade and seeds soaked overnight in water at 35°C before planting. Subsequently, only one seedling per sample was used for DNA extraction. Total genomic DNA was extracted from freshly harvested leaves (4-6 cm) using a modified version of the high throughput mini-prep 3% Cetyl Trimethyl Ammonium Bromide (CTAB) method described by Mace et al. (2003). The quality and quantity of the isolated DNA was determined by comparing the fluorescence of aliquots of DNA samples with a known concentration of λ -DNA after running them on a 0.8% agarose gel that contained 0.3 μ g/ml ethidium bromide solution.

The strategy of sampling one individual per sample was guided by the need to maximize the number of landrace and wild/weedy relatives to be genotyped on country scale. This approach has proved sufficient to

detect large-scale inter-sample evolutionary trends in crops and/or their wild relatives (e.g. Matsuoka et al. 2002, Fukunaga et al. 2005, Mariac et al. 2006; Deu et al. 2008; Barro-Kondombo et al. 2010) provided that the number of loci is sufficient.

PCR amplification and genotyping

Thirty SSR markers (Table 1) with a good genome coverage were analysed using the M13-tailed primer method (Schuelke 2000) to label amplicons for visualisation on an ABI 3730 (Applied Biosystems) capillary sequencer. Forward primers were 5'-tailed with a 19-base pair (bp) M13 universal sequence, 5'-CACGAGCTTGTAACGACXXXXXXXXXXXXX-3', where the X's denote microsatellite-specific primer sequences (See Table 1 for details).

Polymerase chain reaction (PCR) was performed in 10 µl reaction volumes, containing 2.5 ng of template DNA, 0.2 units of Amplitaq Gold *Taq* DNA polymerase (Applied Biosystems), 1X PCR buffer (10 mM Tris-HCl pH 8.3, 50 mM KCl, 1.5 mM MgSO₄), 0.16 mM dNTPs, 2 µM sequence-specific reverse primer, 0.04 µM 5'-M13 tailed sequence-specific forward primer and 0.16 µM 5'-fluorescently labelled M13 universal sequence primer in a GeneAmp PCR system 9700 thermocycler (Applied Biosystems). The M13 universal sequence primer was 5'-tagged with VIC, NED, FAM or PET fluorescent dyes in order to facilitate post-PCR multiplexing. The PCR program was as described by Folkertsma et al. (2005)

After the PCR, a few samples from each primer pair were randomly selected and checked for proper amplification and product intensity on to 2% agarose. For high throughput and low cost genotyping, PCR products were separated by pooling post-PCR products based on fluorescent dye and/or fragment size. Depending on band intensity on agarose gel, 1.5-3.0 µl of PCR products from each of the 6-FAM, VIC, NED and PET-labelled PCR products were pooled together and the final volume adjusted to 15 µl by adding the required volume from a mix of an injection solution (HiDi) and size standard (GS500 LIZ). PCR fragments were denatured and size-fractionated using ABI 3730 Capillary DNA Sequencer (PE-Applied Biosystems) as described in the user's manual. The peaks were sized and the alleles called using GeneMapper software version 3.7 and the internal size standard GS500LIZ. Positive control samples (genotype BTx623, pool-A, pool-B, pool-C) were included in all PCR for verifying the repeatability of each PCR and allele calls.

Data scoring and analyses

Although automated DNA sequencers and corresponding software's contribute substantially to increased throughput rates for large-scale genotyping projects, different factors (e.g., plus-A amplification, stuttering, incorrect allele sizing) cause ambiguity in allele binning. Thus, the software AlleloBin (Prasanth et al. 2006) was used to classify observed microsatellite allele sizes into representative discrete alleles using the least-square minimization algorithm of Idury and Cardon (1997). All statistical analyses were performed on the adjusted data. Twenty four out of the 30 SSR markers showed high reproducibility, with high consistency between the expected (based on sequence information) and observed allele sizes for the 4

positive control samples (BTx623, pool-A, pool-B, pool-C). Therefore, only the 24 SSRs were used in the analysis (Table 1).

Investigated levels of genetic diversity

The extent and partitioning of microsatellite diversity was investigated at three different levels: (i) sorghum type (cultivated or wild), (ii) geographic region of origin and (iii) agro-climatic zone of origin. Six geographic zones that corresponded to the six sorghum growing areas sampled in this study, namely Turkana, western/Nyanza, eastern/central, Coast, Rift Valley and north-eastern were recognised. Agro-climatic zones were defined according to the agro-climatic zone map of Kenya (Sombroek et al. 1982) which recognises seven agro-climatic zones based on annual rainfall and potential for evaporation: I (humid with 1100-2700 mm of annual rainfall), II (sub-humid with 1000-1600 mm of annual rainfall), III (semi-humid with 800-1400 mm of annual rainfall), IV (semi-humid to semi-arid with 600-1100 mm of annual rainfall), V (semi-arid with 450-900 mm of annual rainfall), VI (arid with 300-550 mm of annual rainfall) and VII (very arid with 150-350 mm of annual rainfall).

Standard parameters of genetic diversity, including total number of alleles (A^t), number of rare alleles (A^r , alleles with a frequency $< 5\%$ per group), private alleles (A^p , alleles unique to a group), observed heterozygosity (H_o) and unbiased expected heterozygosity or gene diversity (H_e) were computed using GENETIX version 4.05 (Belkhir et al. 2004). Since the observed number of alleles is highly dependent on sample size, the program FSTAT (Goudet 2002) was employed to compute the mean allelic richness across all loci (R_s) for each defined level of genetic structure. In addition the software HP-RARE 1.2 (Kalinowski 2005) was used to compute and compare the private allelic richness (\prod_{taxon}^S) between cultivated and wild sorghum. The two programs implement the rarefaction statistical method first used by Hulbert (1971) to estimate species diversity. The method allows for unbiased comparisons among populations of unequal sample sizes by calculating a standardised estimate of allelic richness for a fixed sample size. Overall differences in R_s , \prod_{taxon}^S and H_e between cultivated and wild sorghum were assessed for significance using Wilcoxon's signed-rank test as implemented in the software GenStat (VSN International Ltd. 2007). The Kruskal-Wallis test was used to test for differences among geographic and agro-climatic zones in their allelic richness and gene diversity using the software R (R Development Core Team 2007).

To investigate the genetic relationships between all pairs of cultivated and wild sorghum plants, a genetic dissimilarity matrix was computed using the simple matching procedure and subsequently used as an input for principal coordinate analysis (PCoA) in the software DARwin 5.0 (Perrier and Jacquemoud-Collet 2006). The pairwise deletion option was chosen to ensure that dissimilarity calculations were done only for pairs of genotypes where allelic scores were obtained for at least 70% of all loci. At this level of threshold, eighteen individuals were eliminated from the final calculation because of having too many missing data.

Genetic structure

Three complementary approaches were used to explore genetic diversity structure: Fixation index (F_{ST}), analysis of molecular variance (AMOVA), and model-based clustering. The observed diversity was for

each approach, partitioned between the sorghum types (cultivated or wild) and among geographic and agro-climatic zones for cultivated and wild sorghum separately.

The software GENETIX 4.04 was used to compute the Weir and Cockerham (1984) θ , an unbiased estimator of F_{ST} . Pairwise F_{ST} calculations were used to compare the level of genetic differentiation among geographic regions for cultivated and wild sorghum separately and between the two sorghum gene pools. The F_{ST} values were tested for significance using the permutations procedure (10000 permutations). Genetic differentiation in cultivated and wild sorghum was further analysed using the analysis of molecular variance (AMOVA; Excoffier et al. 1992) procedure, implemented in the software ARLEQUIN 3.11 (Excoffier et al. 2005). The significance of partitioning of the genetic variance components among the various groups was tested using 10000 permutations.

The Bayesian model-based clustering method implemented in the software STRUCTURE 2.2.3 (Pritchard et al. 2000) was also used to explore genetic structure first by pooling cultivated and wild sorghum individuals and then separately for each sorghum type. The basic admixture model with unlinked loci, uncorrelated allele frequencies and with no *a priori* population information was used. STRUCTURE was run by varying the number of clusters (K) from 2 to 10 using the web resources of the Computational Biology Service Unit (CBSU) from Cornell University (<http://cbsuapps.tc.cornell.edu/structure.aspx>). Each K was run 10 times with a burn-in length of 500000 and a post-burning data collection length of 1×10^6 . The most likely number of genetic clusters was estimated using the ad hoc statistic ΔK (Evanno et al. 2005), which is based on the second order rate of change of $P(X|K)$, the posterior probability of the data with respect to a given K. According to Evanno and co-workers (2005), the peak value of the distribution of ΔK is located at the most likely value of K. We illustrated the peak value by plotting ΔK values against successive K values. The proportion of individuals genome assigned to each cluster (Q) for the most likely number of clusters was summarized by way of bar plots.

Spatial structure of genetic variation

To investigate the spatial structure of genetic diversity in cultivated and wild sorghum at country scale, spatial autocorrelation analysis was performed as implemented in the software SPAGeDi (Hardy and Vekemans 2002). A geographic distance matrix was generated from the latitude/longitude coordinates associated with each sample using the software Geographic Distance Matrix Generator version 1.2.3 (http://biodiversityinformatics.amnh.org/open_source/gdmg). In each case, 20 distance classes were defined such that there were approximately equal numbers of pairwise comparisons in each class. Within each class, the relative kinship coefficient (r_{ij}) was estimated using the method of Ritland (1996). This index represents the correlation in allelic states between homologous genes and weighs allele distribution by the inverse of allele frequency, thus giving more weight to rare alleles. This way the approach results in lower sampling variance, hence, is powerful for detecting genetic structure (Hardy and Vekemans 2002). The cultivated sorghum samples obtained from the genebank for north-eastern and Rift Valley regions did not have corresponding wild sorghum samples and were therefore eliminated from the analysis to allow for comparison of patterns between cultivated and wild sorghum. The significance of the estimated values of kinship coefficient and regression slope was tested by permuting individuals among locations 1000 times.

The relationship between genetic relatedness and geographic distance was visualised in correlograms using the software R, with the 95% confidence interval envelope under the null hypothesis of no spatial structure indicated.

Further, the method of Rousset (2000) was used to indirectly infer the extent of gene dispersal between cultivated and wild sorghum individuals. This approach is based on the analytical model of isolation-by-distance, which predicts that the genetic distance between individuals (\hat{a}) (Rousset 2000) increases approximately linearly with the logarithm of spatial distance. Rousset's measure of genetic distance, \hat{a} , was computed for each pair of individuals using the program SPAGeDi (Hardy and Vekemans 2002). Ten distance classes consisting of approximately equal numbers of individual pairwise genetic distance comparisons were defined. Subsequently, the pairwise genetic distance estimates were regressed on the logarithm of spatial distance, providing a regression slope (*blog*) and an estimate of the coefficient of determination (r^2). The significance of the regression slope was tested by a randomisation procedure whereby individuals were permuted among locations 1000 times to assess the distribution of the slope values under the null hypothesis of no correlation between geographic and genetic distance. P-values were estimated as the proportion of this distribution lying higher than the observed slope value. The program R was used to plot estimates of the pairwise genetic distance between individuals of cultivated and wild sorghum against logarithmic spatial distance, with the regression line shown.

Results

Extent of genetic diversity

The number of alleles per marker varied from 3 in Xtxp136 to 25 in SbAGB02 (Table 1). The 24 SSR markers amplified a total of 295 different alleles among all the 439 samples, an average of 12.3 alleles per marker (Table 2). The number of alleles detected among the 329 cultivated sorghum samples was 257, out of which 173 (67%) were rare and 15 (5%) private alleles. In comparison, 238 alleles were observed in the 110 wild sorghum samples, with 122 (51%) being rare and 13 (5%) private alleles. The cultivated gene pool was observed to harbour lower genetic diversity than the wild gene pool, based on significantly lower mean allelic richness ($P \leq 0.05$), private allelic richness ($P \leq 0.05$) and gene diversity ($P < 0.001$) values. We found significant differences among regions in the levels of allelic richness ($3.45 \leq R_s \leq 5.59$; $P \leq 0.05$) and gene diversity ($0.3396 \leq H_e \leq 0.5595$; $P \leq 0.001$) for both cultivated and wild sorghum. Estimates of allelic richness and gene diversity for cultivated and wild sorghum in each region are presented in form of box plots in Fig. 2. For cultivated sorghum, the mean allelic richness ranged from 3.45 in Turkana to 5.59 in the coast, whereas the mean gene diversity ranged from 0.3396 in Turkana to 0.5595 in eastern/central. For wild sorghum, the mean allelic richness ranged from 3.67 in Turkana to 6.00 in the coast, while mean gene diversity ranged from 0.4836 in western/Nyanza to 0.6806 in the coast. No significant differences ($P \geq 0.05$) were evident for either allelic richness or gene diversity among the agro-climatic zones for cultivated sorghum as well as for its wild progenitor.

Genetic relationships within and between cultivated and wild sorghum

Genetic relationships among individuals of cultivated and wild sorghum are presented as a biplot of the first plane of a PCoA (Fig. 3). In total the first plane accounted for 13.7% of the total variability, 6.1% of which is accounted for by axis 1 and 5.3% by axis 2. Generally, the separation between cultivated and wild sorghum gene pools was low, although cultivated sorghum from north-eastern is clearly separated from both the cultivated and wild counterparts from the other regions. For the rest of the samples, there appeared to be high levels of overlap both among regions and among sorghum types.

Patterns of genetic differentiation

Overall, the level of differentiation between cultivated and wild sorghum based on F_{ST} was moderate but highly significant ($F_{ST}=0.062$; $P < 0.001$) (Table 3). Similarly, AMOVA and Bayesian cluster analysis showed close genetic proximity between cultivated and wild sorghum in Kenya. For AMOVA, only 6.5% of the total genetic diversity was partitioned to the variation between cultivated and wild sorghum, compared to 93.6% that was partitioned to the variation within the two sorghum types. Similarly, the Bayesian model-based cluster analysis at $K=2$ (i.e. assuming only two genetic groups) failed to identify distinct differentiation among cultivated and wild sorghum individuals (Fig. 4). In this analysis, we assumed that an individual was only exclusively assigned to a particular genetic cluster if at least 85% of its genome (i.e. $q_i \geq 0.85$) is found in it; otherwise it was assumed to be jointly assigned to two clusters probably due to admixture. Using this arbitrary threshold, at least 26.8% of cultivated sorghum individuals and 36.4% of their wild sorghum counterparts were assigned to both cluster 1 and 2.

In cultivated sorghum, genetic differentiation was high among regions ($F_{ST}=0.187$; $P < 0.001$), but moderate among agro-climatic zones ($F_{ST}=0.077$; $P < 0.001$). Similar trends were observed for wild sorghum, with moderate genetic differentiation among regions ($F_{ST}=0.097$; $P \leq 0.001$), and low genetic differentiation among agro-climatic zones ($F_{ST}=0.054$; $P \leq 0.001$). The outcome of comparing the extent of genetic differentiation within and between cultivated and wild sorghum among the regions (pairwise F_{ST}) is presented in Table 4. Overall, all F_{ST} values were significantly greater than zero ($P \leq 0.05$).

The level of divergence among regions was variable in both cultivated and wild sorghum, with the former generally exhibiting greater F_{ST} values than the latter. In cultivated sorghum the lowest level of inter-regional genetic differentiation was observed between eastern/central and coastal regions ($F_{ST}=0.03$) and the highest between Turkana and north-eastern regions ($F_{ST}=0.44$). Notably, there was high level of inter-region genetic similarity between eastern/central and coastal regions ($F_{ST}=0.03$), western/Nyanza and coastal regions ($F_{ST}=0.05$) and eastern/central and western/Nyanza ($F_{ST}=0.07$). Contrastingly, Turkana and north-eastern cultivated sorghum pools appeared to be clearly distinct, both with each other and with the rest of the cultivated sorghum pools. Similar trends were revealed in wild populations, with substantial inter-regional similarities among coastal, eastern/central and western regions ($0.06 \leq F_{ST} \leq 0.10$), that were coupled with substantial distinctiveness of Turkana populations in relation to those from other regions ($0.13 \leq F_{ST} \leq 0.17$, $P < 0.001$).

We detected lower levels of genetic divergence between cultivated and wild sorghum within ($0.03 \leq F_{ST} \leq 0.18$; shaded in grey) than among ($0.05 \leq F_{ST} \leq 0.33$) regions (Table 4). Within regions, the highest level of crop-wild genetic divergence was recorded in Turkana, and the least in western/Nyanza ($F_{ST}=0.03$). The level of crop-wild genetic divergence was comparable and moderate in the coastal ($F_{ST}=0.11$) and eastern ($F_{ST}=0.10$) regions.

The genetic structure of the 329 cultivated sorghum individuals based on the Bayesian model-based algorithm implemented in STRUCTURE is shown for $K=7$ (Fig. 5a&b). Evanno's *ad hoc* ΔK method determined $K=7$ to be the most likely number of genetic clusters for the entire cultivated sorghum pool (Supplementary Fig. S1). The mean proportion of genome assigned in each of the seven clusters (Q_i) is presented for each geographic region in Fig. 5b. The identified genetic structure corresponded closely to geographic origin for Turkana ($Q_i=0.94$ in cluster 5), North-eastern ($Q_i=0.70$ in cluster 1) and to some considerable extent for Rift valley ($Q_i=0.65$ in cluster 7). For each of these regions, the largest proportion of genome was assigned predominantly to a single and largely unique cluster. Contrastingly, cultivated sorghum individuals from the coastal, eastern/Central and western/Nyanza regions tended to be jointly assigned to more than one cluster (Fig. 5b).

For the 110 wild sorghum individuals, STRUCTURE identified two genetic clusters (Supplementary Fig S2), which clearly did not correspond to the geographic origin (Fig. 6a&b). Of the four regions the highest level of genetic uniformity was observed in western/Nyanza, where the wild sorghum individuals were assigned almost exclusively to one of the two genetic structure ($Q_i=0.95$ in cluster 1).

When cultivated and wild sorghum data was pooled together and STRUCTURE run from K=1 to K=10, Evanno's *ad hoc* ΔK method determined K=5 to be the most likely number of genetic clusters (Supplementary Fig. S3). The wild forms were separated largely into four of the five genetic groups, the largest of which (cluster 3) was least shared with the cultivated counterparts (Fig. 7). Cultivated sorghum on the other hand was generally represented in all the five clusters, with Turkana and north-eastern collections being restricted into single and largely distinct genetic groups. Notably, cultivated sorghum from north-eastern appeared to share minimal ancestry with cultivated and wild sorghum individuals from other regions. For the rest of the regions, cultivated and wild sorghum individuals seemed to overlap in most of the clusters (Fig. 7).

Spatial genetic structure

Outcome of spatial autocorrelation analyses in cultivated and wild sorghum is presented as correlograms (Fig. 8a and 8b). There was a clear decrease in relatedness among individuals with increasing geographical distance, a reflection of strong spatial genetic structure in both cultivated and wild sorghum. Cultivated sorghum had a mean regression slope (*blog*) value of -0.015 ($P < 0.001$) and a coefficient of determination value (r^2) of 0.045, while wild sorghum had a *blog* value of -0.017 ($P < 0.001$), and a r^2 value of 0.055. Furthermore, kinship coefficient values were positive and significant within a similar range of about 180 km for the two sorghum conspecifics. Negative and significant kinship coefficient values were clearly evident in wild sorghum beyond 600 km, while in cultivated sorghum significant negative values did not show a consistent pattern. The outcome of investigations on the relationship between crop-wild genetic distance and geographic distance at country level is presented in Fig. 8c. Rousset's genetic distance (\hat{a}) between pairs of cultivated and wild sorghum individuals increased linearly with logarithmic distance (slope=0.149, Permutation test: $P \leq 0.001$, $r^2=0.028$), a pattern typical of isolation by distance.

Discussion

Extent of genetic diversity in cultivated and wild sorghum

Mean gene diversity across the 24 SSR markers for cultivated sorghum in Kenya ($H_e=0.59$) is similar to values reported for microsatellites in Niger ($H_e=0.61$) by Deu et al. (2008) and in South Africa ($H_e=0.60$) by Uptmoor et al. (2003), but slightly lower than values estimated for Eritrea by Ghebru et al. (2002) and for Morocco by Djè et al. (1999). In the wild sorghum gene pool, the mean gene diversity estimated for Kenya across the 24 SSR markers ($H_e=0.69$) was higher than that estimated for a set of accessions selected to represent a wide geographic sampling in Africa ($H_e=0.59$) by Casa et al. (2005). As noted by Deu and co-workers (2008), however, comparisons of the magnitude of genetic diversity between different studies is difficult as it may be complicated by differences in amongst others underlying sampling schemes, number of SSR surveyed, size of SSR repeats and location of the SSR on the genome.

During the process of domestication, evolutionary processes of founder effect, population bottleneck and artificial selection are all expected to reduce genetic diversity of the crop in relation to its wild progenitor (Ladizinsky 1999; Gepts 2004). This view was supported in the present study by findings that cultivated sorghum harboured lower genetic diversity (in terms of overall allelic richness, private allelic richness and gene diversity) than its proposed wild progenitor. Our findings were consistent with previous comparisons between cultivated and wild sorghum using various genetic markers (Morden et al. 1990; Aldrich and

Doebley 1992; Cui et al. 1995; Casa et al. 2005). The significantly higher private allelic richness in the wild sorghum relative to its cultivated counterpart is of great importance to the conservation and utilisation of sorghum genetic resources. These findings support the widely held view that crop wild relatives are potential sources of important and unique genes for crop improvement programs and therefore deserve special conservation and utilisation attention.

Divergence between cultivated and wild sorghum

We found close genetic proximity between individuals of cultivated and wild sorghum at the national level, based on PCoA, F_{ST} and Bayesian model-based cluster analyses. Furthermore, pairwise F_{ST} showed the extent of crop-wild genetic divergence to be generally lower within than among regions. Considered together these findings may reflect important historical gene flow between cultivated sorghum and its wild relatives in situ. Cultivated and wild sorghum are inter-fertile, with natural hybrids between *ssp. bicolor* and *ssp. verticilliflorum* being well documented within and around cultivation in Africa (Dogget and Majisu 1968; Dogget and Prasada Rao 1995; Tesso et al. 2008). In a study related to the present one, Mutegi et al. (2010) recorded putative crop-wild hybrid plants within and around cultivated sorghum and other fields across the crop's growing regions of Kenya. Such crop-wild hybrids may potentially act as conduits for escape and persistence of transgenes in wild and/or weedy relative populations through introgressive hybridisation as has demonstrated between GM oilseed rape (*Brassica napus* L.) and its wild relatives (Halfhill et al. 2004).

Furthermore, we found the extent of divergence between cultivated and wild sorghum to vary substantially among regions. This finding may reflect inter-regional differences in the extent of introgression between cultivated and wild sorghum, probably due to differences in farmer practices. For example, during our sample collection surveys, two contrasting weedy sorghum management practices were observed between western/Nyanza and Turkana regions. Farmers in western/Nyanza were observed to tolerate putative crop-wild hybrids in their sorghum fields following harvest, often resulting in populations of wild sorghum in recently harvested sorghum fields. Such remnant populations were also observed in abandoned (fallow) fields within close proximity to cropped sorghum fields. Such a scenario may enhance hybridisation between cultivated and wild sorghum on-farm. In Turkana most of the sorghum is grown under irrigation with farmers holding small (50 X 100 m) plots. Farmers therefore practice more intensive weeding on their sorghum fields. Consequently we encountered very few putative crop-wild hybrids within and around sorghum fields during the surveys. Not surprisingly, relative to other growing regions, genetic divergence between cultivated and wild sorghum was least in western/Nyanza and highest in Turkana.

Patterns of genetic differentiation in cultivated and wild sorghum

Our study found the extent of genetic diversity to vary significantly among regions but not among agro-climatic zones for both cultivated and wild sorghum. Furthermore, we found greater levels of genetic differentiation among geographic regions than among agro-climatic zones for both cultivated and wild sorghum. Our results suggest that diversity of cultivated sorghum and its wild counterpart in Kenya is structured more along geographical than climatic factors. Deu et al (2008) found the same for cultivated sorghum in Niger, with moderate differentiation among regions ($F_{ST}=0.07$) and lower differentiation among

annual rainfall classes ($F_{ST}=0.03$). The main evolutionary forces responsible for producing genetic structure in plant populations are gene flow, selection associated with environmental heterogeneity and/or farmer preferences and random genetic drift (Hartl and Clark 1997; Neal 2004). Because most SSR loci are presumed to be selectively neutral, environmental factors are expected to offer minimum contribution to the observed genetic structure. In contrast, geographic isolation limits the level of gene flow among populations and should therefore contribute significantly to contemporary genetic structure. The higher level of geographic structure observed in cultivated sorghum ($F_{ST}=0.187$) compared to its wild progenitor ($F_{ST}=0.097$) may point to greater intra-regional genetic proximity in the former than the latter. Such intra-regional genetic proximity in cultivated sorghum would arise through seed exchanges among farmers. Mutegi et al. 2010 reported sorghum seed systems in Kenya to be largely traditional, with farmers playing a major role in the selection and exchange of seeds.

Bayesian model-based cluster analysis nevertheless showed poor correspondence between the observed genetic structure and geographic origin both for cultivated and wild sorghum. Similar results have been reported in Africa for cultivated sorghum (Ayana et al. 2000b; Ghebru et al. 2002; Nkongolo and Nsapatu 2003; Deu et al. 2008) and for its wild relatives (Ayana et al. 2000a). In our study, the only exception was Turkana and north-eastern cultivated sorghum pools, each of which was clustered into a distinct and unique genetic group. Pairwise F_{ST} analysis showed variable levels of genetic differentiation among regions for both cultivated and wild sorghum. Considered together our results may reflect contemporary and/or historical seed-mediated gene flow among the geographic regions, with varying amounts of seed exchanges among regions. Because Turkana and north-eastern are relatively geographically remote from other growing regions, the two regions have experienced minimal if any cross-regional seed-mediated genetic exchanges. Interestingly, Turkana and north-eastern cultivated sorghum gene pools appear to be genetically distinct even though the regions are geographically proximal. The two regions are however physically separated by Lake Turkana and Mt. Kulal, both of which might have acted as barriers to cross-regional seed-mediated gene flow. Another plausible explanation for clear genetic separation of north-eastern and Turkana sorghum both from themselves and from the rest of the country is separate evolutionary history. For example, North-eastern sorghum may be part of the Ethiopian sorghum gene pool, having originated from the Boran agro-pastoralist ethnic group whose distribution spans across Kenya-Ethiopia border. This hypothesis however needs further testing with evolutionary genetic studies that incorporate materials from among other neighbouring countries such as Ethiopia, Sudan and Uganda.

Spatial genetic structure

We observed largely patterns of strong spatial genetic structure in both cultivated and wild sorghum, with surprisingly significant positive spatial autocorrelation within approximately 180 km for the two congeners. This suggests that similar evolutionary factors may underlie the observed pattern of spatial genetic structure in the two congeners. Among factors that can explain the strong spatial structure in cultivated and wild sorghum, one could consider seed-mediated and/or pollen-mediated gene flow. In plants, spatial distribution of genetic variation is primarily determined by seed and pollen dispersal, habitat distribution, micro-environmental selection and genetic drift (Levin and Kerster 1974; Epperson 1993). Seed and pollen dispersal (with or without interbreeding) causes similarity between neighbouring populations, whereas

distant populations differ for the studied autocorrelation coefficient (Sokal and Oden 1978; Epperson 1993; 2004).

In the present study, a number of farmers originally from the western/Nyanza and eastern/central regions were noted to have migrated with their sorghum landraces into the coastal region. In a recent study, Mutegi et al. (2010) documented incidences of medium to long distance seed exchanges in sorghum among growing regions, mostly through inter-ethnic marriage relationships, but also appreciably through formal distribution of improved varieties via government and non-governmental extension systems. Finally, there are two plausible explanations for the near identical patterns of spatial genetic structure in cultivated and wild sorghum; inadvertent dispersal and establishment of wild sorghum seed via cultivated sorghum seed systems and pollen-mediated crop-wild gene flow in sites of sympatric occurrence.

Implications for biosafety regulations

Important historical gene flow between cultivated and wild sorghum is strongly suggested by two findings in this study; low level of divergence between cultivated and wild sorghum gene pools and significant isolation-by-distance between pairs of cultivated and wild sorghum individuals. The level of divergence between cultivated and wild sorghum varied among geographic regions, probably a reflection of intra-region differences in the level of crop-to-wild gene flow. Differences in farmer practices such as weedy sorghum management and/or seed selection are some of the factors that could explain this inter-region variation in the extent of crop-to-wild gene flow. Furthermore, the pattern of increased genetic similarity between geographically close pairs of cultivated and wild sorghum individuals relative to isolated ones (isolation-by-distance) as revealed in this study is of further biosafety significance. It suggests that crop-to-wild gene flow in sorghum is spatially predictable, and transgene escape into cultivated and/or wild-weedy relatives of the crop is higher within and around cultivated fields compared to natural habitats away from cultivation. Overall, this study suggests that deployment of GM sorghum in Kenya may lead to escape and persistence of transgenes into wild-weedy sorghum relatives, with the rate of crop-to-wild gene flow being variable among growing regions. However, the extent and direction of gene flow remains unknown, as does the consequence of transgenes escape and persistence in wild sorghum populations. Biosafety regulators could benefit from further studies on the extent and direction of crop-wild gene flow on-farm, and from studies on the fitness effect of transgenic traits in wild-weedy relatives of crop sorghum.

Implications for utilization and conservation of germplasm

Significantly higher levels of genetic diversity were revealed in wild sorghum relative to its domesticated congeners. These results, together with findings of 2.6 unique alleles per locus in the wild sorghum gene pool, are of interest to sorghum breeding. The high genetic diversity could potentially be exploited in broadening the genetic base of sorghum breeding germplasm, while the unique diversity imply that wild sorghum is a potential source of novel genes such as pests and disease resistance. The genetic potential of wild relatives of sorghum, particularly as sources of resistance to pests and diseases, is well documented such as for sorghum shoot fly (Kamala et al. 2009), sorghum midge (Sharma and Franzmann 2001), green bug (Duncan et al. 1991), downy mildew (Kamala et al. 2002) and ergot (Reed et al. 2002). Moreover, the substantial genetic variability and differentiation revealed in sorghum landraces of Kenya should be

incorporated in breeding programs by developing different populations with a broad genetic base. In addition to safeguarding the landrace diversity, this will help to create new genetic recombinations that can be exploited in response to new breeding challenges.

Finally, levels of genetic diversity within the two *S. bicolor* conspecifics differed significantly among geographic regions. Adequate measures need to be put in place for systematic conservation of these important genetic resources using complementary *ex situ* and *in situ* approaches. Such approaches could also benefit from further studies on the extent and partitioning of diversity in the country, including wild sorghum samples from natural habitats away from cultivated lands and from the regions not sampled in the present study (Rift Valley and north-eastern).

Acknowledgement

This study formed part of the project, “Environmental risk assessment for the introduction of genetically modified sorghum in Mali and Kenya” funded by the United States Agency for International Development (USAID) Biotechnology and Biodiversity Interface (BBI) Program. We are deeply indebted to the late Dr. Fabrice Sagnard (Principle Investigator), who offered exemplary leadership and immense scientific contribution to the entire project. We acknowledge Caroline Mwongera, Charles Marangu and Bernard Rono who participated in collections as well as farmers from various sorghum growing areas of Kenya and the National Genebank of Kenya for providing the seed samples used in this study.

References

- Aldrich PR, Doebley J (1992) Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild *Sorghum bicolor*. *Theor Appl Genet* 85:293-302
- Aldrich PR, Doebley J, Schertz KF, Stec A (1992) Patterns of allozyme variation in cultivated and wild *Sorghum bicolor*. *Theor Appl Genet* 85:451-460
- Auer C. (2008) Ecological risk assessment and regulation for genetically-modified ornamental plants. *Crit Rev Plant Sci* 27:255-271
- Ayana A, Bekele E, Bryngelsson T (2000a) Genetic variation in wild sorghum (*Sorghum bicolor* ssp *verticilliflorum* (L.) Moench) germplasm from Ethiopia assessed by random amplified polymorphic DNA (RAPD). *Hereditas* 132:249-254
- Ayana A, Bryngelsson T, Bekele E (2000b) Genetic variation of Ethiopian and Eritrean sorghum (*Sorghum bicolor* (L.) Moench) germplasm assessed by random amplified polymorphic DNA (RAPD). *Genet Resour Crop Evol* 47:471-482
- Ayana A, Byngelsson T, Bekele E (2001) Geographic and altitudinal allozyme variation in sorghum (*Sorghum bicolor* (L.) Moench) landraces from Ethiopia and Eritrea. *Hereditas* 135:1-12
- Barnaud A, Deu M, Garine E, Mckey D, Joly HI (2007) Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. *Theor Appl Genet* 114:237-248
- Barrett BA, Kidwell KK (1998) AFLP-based genetic diversity assessment among wheat cultivars from the Pacific Northwest. *Crop Sci* 38: 1261-1271
- 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 (2004) GENETIX 405, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier.
- Bhatia CR, Mitra R (2003) Consequences of gene flow from genetically engineered crops. *Curr Sci India* 84:138-141

- Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S (2005) Diversity and selection in Sorghum: simultaneous analyses using simple sequence repeats. *Theor Appl Genet* 111:23-30
- Chandler S, Dunwell JM (2008) Gene flow, risk assessment and the environmental release of transgenic plants. *Crit Rev Plant Sci* 27:25-49
- Clayton WD, Renvoize RD (1982) Poaceae. *Flora of Tropical East Africa*, Part 3 AA Balkema, Rotterdam
- Cleveland DA, Soleri D (2005) Rethinking the risk management process for genetically engineered crop varieties in small-scale, traditionally based agriculture. *Ecol Soc* 10:1-33
- Conner AJ, Glare TR, Nap JP (2003) The release of genetically modified crops into the environment. Part II. Overview of ecological risk assessment. *Plant J* 33:19-46
- Cui YX, Xu GW, Magill CW, Schertz KF, Hart GE (1995) RFLP-based assay of *Sorghum bicolor* (L) Moench. genetic diversity. *Theor Appl Genet* 90:787-796
- Deu M, Hamon P, Chanterreau J, Dufour P, D'Hont A, Lanaud C (1995) Mitochondrial DNA diversity in wild and cultivated sorghum. *Genome* 38:635-645
- Deu M, Sagnard F, Chanterreau 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* 116:903-913
- Djè Y, Ater M, Lefèbvre C, Vekemans X (1998) Patterns of morphological and allozyme variation in sorghum landraces of northwestern Morocco. *Genet Resour Crop Evol* 45:541-548
- Djè Y, Forcioli D, Ater M, Lefèbvre C, Vekemans X (1999) Assessing population genetic structure of sorghum landraces from North-western Morocco using allozyme and microsatellite markers. *Theor Appl Genet* 99:157-163
- Dogget H (1988) *Sorghum*. Longman Scientific and Technical, Essex
- Dogget H, Majisu BN (1968) Disruptive selection in crop development. *Heredity* 23:1-23
- Dogget H, Prasada Rao KE (1995) *Sorghum*. In: Smartt J, Simmonds NW (eds) *Evolution of Crop Plants*, 2nd edn. Longman Group, Essex, pp 140-159

- Duncan RR, Bramel-Cox PJ, Miller FR (1991) Contributions of introduced sorghum germplasm to hybrids development in the USA. In: Shands HL, and Wiesner LE (eds) Use of Plant Introductions in the Cultivar Development. Crop Science Society of America, Madison, USA, pp 69 - 101
- Ellstrand NC (1992) Gene flow by pollen: Implications for plant conservation genetics. *Oikos* 63:77-86
- 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
- Epperson BK (1993) Recent advances in correlation analysis of spatial patterns of genetic variation. *Evol Biol* 27:95-155
- Epperson BK (2004) Multilocus estimation of genetic structure within populations. *Theor Popul Biol* 65:227-237
- Evanno S, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611-2620
- Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479-491
- Excoffier L, Laval LG, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol Bioinform* 1:47-50
- FAO (2008) FAOSTAT. <http://faostat.fao.org>
- Folkertsma RF, Rattunde HFW, Chandra S, Raju GS, 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
- Frankel OH, Brown AHD, Burdon JJ (1995) The conservation of plant diversity. Cambridge University Press, New York
- Frankel OH, Hawkes JG (1975) Crop genetic resources for today and tomorrow. Cambridge University Press Cambridge, New York
- Fukunaga K, Hill J, Vigouroux Y, Matsuoka Y, Sanchez G, Liu K, Buckler ES, Doebley J. 2005 Genetic diversity and population structure of teosinte. *Genetics* 169:2241-2254

- Gepts P (2004) Crop domestication as a long-term selection experiment. In: Jannick J (Ed.) Plant Breeding Reviews, Volume 24, Part 2: Long-term selection: Crops, Animals, Bacteria, John Wiley & Sons, New York
- Ghebru B, Schmidt RJ, Bennetzen JL (2002) Genetic diversity of Eritrean sorghum landraces assessed with simple sequence repeat (SSR) markers. *Theor Appl Genet* 105:229-236
- Goudet J (2002) FSTAT, a program to estimate and test gene diversity and fixation indices. (version 2932).
- Gurney AL, Press MC, Scholes JD (2002) Can wild relatives of sorghum provide new sources of resistance or tolerance against *Striga* species? *Weed Sci* 42:317-324
- Halfhill MD, Zhu B, Warwick SI, Raymer PI, Millwood RJ, Weissinger AK, Stewart Jr NC (2004) Hybridization and backcrossing between transgenic oilseed rape and two related weed species under field conditions. *Environ Biosafety Res* 3: 73-81
- Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol* 2:618-620
- Harlan JR, De Wet JMJ (1972) A simplified classification of cultivated sorghum. *Crop Sci* 12:172-177.
- Hartl DL, Clark G (1997) Principles of population genetics. Sinauer Associates Inc, Sunderland
- Haygood R, Ives AR, Andow DA (2003) Consequences of recurrent gene flow from crops to wild relatives. *Proc R Soc Lond B* 270:1879-1886
- Hulbert SH (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52:577-586
- Idury RM, Cardon LR (1997) A simple method for automated allele binning in microsatellite markers. *Genome Res* 11: 1104 - 1109
- Kalinowski S (2005) HP-RARE 10: a computer program for performing rarefaction on measures of allelic richness. *Mol Ecol* 5:187-189
- Kamala V, Sharma HC, Manohar Rao D, Varaprasad KS, Bramel PJ (2009) Wild relatives of sorghum as sources of resistance to sorghum shoot fly, *Atherigona soccata*. *Plant Breeding* 128:137-142
- Kamala V, Singh SD, Bramel PJ, Rao DM (2002) Sources of resistance to downy mildew in wild and weedy sorghums. *Crop Sci* 42:1357-1360
- Ladizinsky G (1999) Plant evolution under domestication. Kluwer Academic Publishers, London

- Levin DA, Kerster HW (1974) Gene flow in seeds plants. *Evol Biol* 7:139-220
- Mace EM, Buhariwalla HK, Crouch JH (2003) A high-throughput DNA extraction protocol for tropical molecular breeding programs. *Plant Mol Biol Rep* 21:459a-459h.
- Mariac C, Luong V, Kapran I, Mamadou A, Sagnard F, Deu M, Chantereau J, Gérard B, Ndjeunga J, Bezançon G, Pham JL, Vigouroux Y (2006) 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
- 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
- Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop Sci* 43:1235-1248
- Morden CW, Doebley JF, Schertz KF (1990) Allozyme variation among the spontaneous species of *Sorghum* section *Sorghum* (Poaceae). *Theor Appl Genet* 80:296-304
- 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
- Neal D (2004) Introduction to population biology. Cambridge University Press Cambridge, New York
- Nkongolo KK, Nsapato L (2003) Genetic diversity in *Sorghum bicolor* (L.) Moench accessions from different ecogeographical regions in Malawi assessed with RAPDs. *Genet Resour Crop Evol* 50:149-156
- Perrier X, Jacquemoud-Collet JP (2006) DARwin software. <http://darwin.cirad.fr/darwin>
- Prasanth V, Chandra S, Jayashree B, Hoisington D (2006) AlleloBin - A program for allele binning of microsatellite markers based on the the algorithm of Idury and Cardon (1997). ICRISAT International Crops Research Institute for the Semi-Arid Tropics.
- 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, Austria (<http://www.r-project.org>).
- Rao Kameswara N, Reddy LJ, Bramel PJ (2003) Potential of wild species for genetic enhancement of some semi-arid food crops. *Genet Resour Crop Evol* 50:707-721
- Reed JD, Ramundo BA, Claflin LF, Tuinstra MR (2002) Analysis of resistance to ergot in sorghum and potential alternate hosts. *Crop Sci* 42:1135-1138
- Rich PJ, Grenier U, Ejeta G (2004) Striga resistance in the wild relatives of sorghum. *Crop Sci* 44:2221-2229
- Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res* 67:175-185
- Rousset F (2000) Genetic differentiation between individuals. *J Evol Biol* 13:58-62
- Sagnard F, Barnaud A, Deu M, Barro C, Luce C, Billot C, Rami JF, Bouchet S, Dembelé D, Pomies V, Calatayud C, Rivallan R, Joly H, vom Brocke K, Touré A, Chantereau J, Bezançon G, Vaksman M (2008) Multi-scale analysis of sorghum genetic diversity: Understanding the evolutionary processes for *in situ* conservation. *Cah Agric* 17:114-121
- Schuelke M (2000) An economic method for the fluorescent labelling of PCR fragments. A poor man's approach to genotyping for research and high throughput diagnostics. *Nat Biotechnol* 18:233-234
- Sharma HC, Franzmann BA (2001) Host plant preference and oviposition responses of the sorghum midge, *Stenodiplosis sorghicola* (Coquillett) (Dipt., *Cecidomyiidae*) towards wild relatives of Sorghum. *J. Appl Ent* 125: 109-114
- Snow AA, Moran-Palma P (1997) Commercialization of transgenic plants: potential ecological risks. *BioScience* 47:86-96
- Sokal RR, Oden NL (1978) Spatial autocorrelation in biology 2. Some biological implications and four applications of evolutionary and ecological interest. *Biol J Linn Soc* 10:249
- Sombroek WC, Braun HMH, van der Pour BJA (1982) Explanatory soil map and agro-climatic zone map of Kenya. Report E1:1-56
- 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

Thies JE, Devare MH (2007) An ecological assessment of transgenic crops. *J Dev Stud* 43:97-129

Uptmoor R, Wenzel W, Friedt W, Donaldson G, Ayisi K, Ordon F (2003) Comparative analysis on the genetic relatedness of *Sorghum bicolor* accessions from Southern Africa by RAPDs, AFLPs and SSRs. *Theor Appl Genet* 106:1316-1325

VSN International Ltd (2007) GenStat Discovery Edition 3 VSN International Ltd. Hernal Hempstead, UK.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358-1370

Table 1 List of the 24 microsatellite loci used in the analysis. The repeat motif, forward and reverse sequences, scored allelic range, number of scored alleles, expected heterozygosity (H_e) and observed heterozygosity (H_o) are shown for each locus.

Locus name	Core motif	Chromosome	Forward primer sequence (5'-3')	Reverse primer sequence(5'-3')	Size range	N
gpsb123	(AC)7 (GA)5	8	ATAGATGTTGACGAAGCA	GTGGTATGGGACTGGA	284-316	
mSbCIR238	(AC)26	2	AGAAGAAAAGGGGTAAGAGC	CGAGAAACAATTACATGAACC	64-112	
mSbCIR240	(TG)9	8	GTTCTTGGCCCTACTGAAT	TCACCTGTAACCCCTGTCTTC	101-113	
mSbCIR246	(CA)7.5	7	TTTTGTTGCACTTTTGAGC	GATGATAGCGACCACAAATC	87-109	
mSbCIR248	(GT)7.5	5	GTTGGTCAGTGGTGGATAAA	ACTCCCATGTGCTGAATCT	84-122	
mSbCIR262	(CATG)3.25	10	GCACCAAAATCAGCGTCT	CCATTTACCCGTGGATTAGT	210-242	
mSbCIR276	(AC)9	3	CCCCAATCTAACTATTTGGT	GAGGCTGAGATGCTCTGT	220-238	
mSbCIR300	(GT)9	7	TTGAGAGCGGCGAGGTAA	AAAAGCCCAAGTCTCAGTGCTA	98-122	
SbAGB02	(AG)35	7	CTCTGATATGTCGTTGTGCT	ATAGAGAGGATAGCTTATAGCTCA	91-143	
Xcup02	(GCA)6	9	GACGCAGCTTTGCTCCTATC	GTCCAACCAACCCACGTATC	185-203	
Xcup14	(AG)10	3	TACATCACAGCAGGGACAGG	CTGGAAAGCCGAGCAGTATG	201-241	
Xcup53	(TTTA)5	1	GCAGGAGTATAGGCAGAGGC	CGACATGACAAGCTCAAACG	178-202	
Xcup61	(GAG)7	3	TTAGCATGTCCACCACAACC	AAAGCAACTCGTCTGATCCC	186-210	
Xcup63	(GGATGC)4	2	GTAAAGGGCAAGGCAACAAG	GCCCTACAAAATCTGCAAGC	135-153	
Sb4-72	(AG)16	6	TGCCACCACTCTGGAAAAGGCTA	CTGAGGACTGCCCAAATGTAGG	176-214	

Locus name	Core motif	Chromosome	Forward primer sequence (5'-3')	Reverse primer sequence(5'-3')	Size range	N
Xisep0310	CCAAT(4)	2	TGCCTTGTGCCTTGTTATCT	GGATCGATGCCTATCTCGTC	148-218	
Xtxp010	(CT)14	9	ATACTATCAAGAGGGGAGC	AGTACTAGCCACACGTCAC	129-151	
Xtxp012	(CT)22	4	CGTCTTCTACCGCGTCCT	CATAATCCCACTCAACAATCC	165-209	
Xtxp015	(TC)16	5	CACAAACACTAGTGCCTTATC	CATAGACACCTAGGCCATC	201-235	
Xtxp057	(GT)21	6	GGAAC TTTGACGGGTAGTGC	CGATCGTGATGTCCCAATC	217-265	
Xtxp114	(AGG)8	3	CGTCTTCTACCGCGTCCT	CATAATCCCACTCAACAATCC	191-239	
Xtxp136	(GCA)5	5	GCGAATAGCATCTTACAACA	ACTGATCATTGGCAGGAC	233-239	
Xtxp320	(AAG)20	1	TAAACTAGACCATATACTGCCATGATAA	GTGCAAATAAGGGCTAGAGTGTT	248-305	
Xtxp40	(GGA)7	7	CAGCAACTTGCACTTGTC	GGGAGCAATTTGGCACTAG	103-142	

^aH_e=Expected heterozygosity or Gene diversity ^bH_o=Observed heterozygosity

Table 2 Comparative genetic diversity estimates for Kenya's sorghum gene pool

Gene pool	N	A ^t	A ^r (< 5%)	A ^p	R _s	\prod_{taxon}^S	H _e	H _o
Wild	110	238	122	13	9.68	2.60	0.6872	0.1762
Cultivated	329	257	173	15	8.63	1.22	0.5876	0.1100
Overall	439	295	204	-	12.12	-	0.6304	0.1267

N=Number of samples, A^t=Total number of alleles, A^r=Number of rare alleles, A^p= Number of private alleles, R_s=allelic richness, \prod_{taxon}^S =Private allelic richness, H_e=Expected (unbiased) heterozygosity, H_o=Observed heterozygosity.

Table 3 F_{ST}-based genetic differentiation of the sorghum gene pool at various levels

Sorghum gene pool	Differentiation level	Mean F _{ST}
Overall	Between cultivated and wild sorghum	0.062***
Cultivated	Among geographic regions	0.187***
	Among agro-climatic zones	0.077***
Wild	Among geographic regions	0.097***
	Among agro-climatic zones	0.054***

*** Highly significant P-value (P < 0.001)

Table 4 Estimates of pairwise FST among collections of cultivated and wild sorghum within and among different geographical regions. Letters C, E, N, R, T and W represent coast, eastern/central, north-eastern, Rift Valley, Turkana and western/Nyanza regions, respectively

		Cultivated						Wild			
		C	E	N	R	T	W	C	E	T	W
Cultivated	C	-	0.03	0.24	0.11	0.25	0.05				
	E		-	0.24	0.09	0.24	0.07				
	N			-	0.30	0.44	0.31				
	R				-	0.31	0.13				
	T					-	0.27				
	W						-				
Wild	C	0.11	0.13	0.23	0.17	0.31	0.14	-	0.06	0.13	0.10
	E	0.10	0.10	0.26	0.14	0.31	0.13		-	0.13	0.09
	T	0.15	0.17	0.33	0.24	0.18	0.20			-	0.17
	W	0.05	0.07	0.31	0.13	0.32	0.03				-

Figures

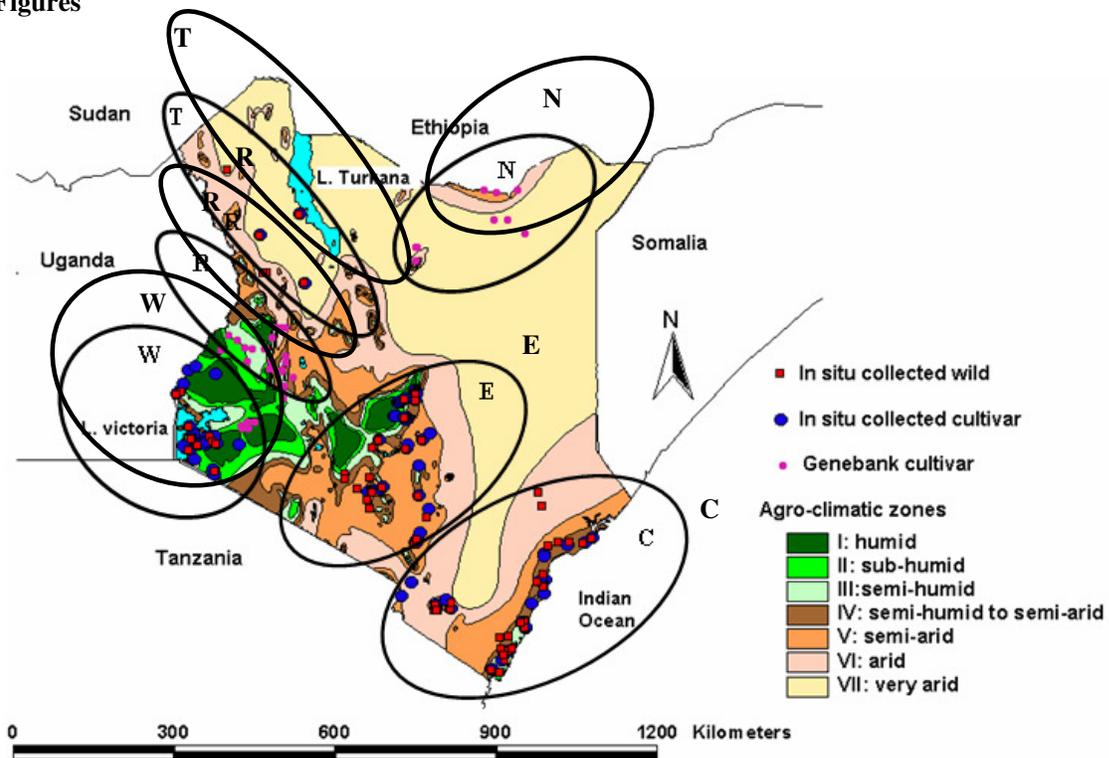


Fig. 1 Agro-climatic map of Kenya with origin and distribution of cultivated and wild sorghum used in the study shown. Six sorghum growing regions were covered: Coast (C), Eastern/central (E), North-eastern (N), Rift valley (R), Turkana (T) and Western/Nyanza (W)

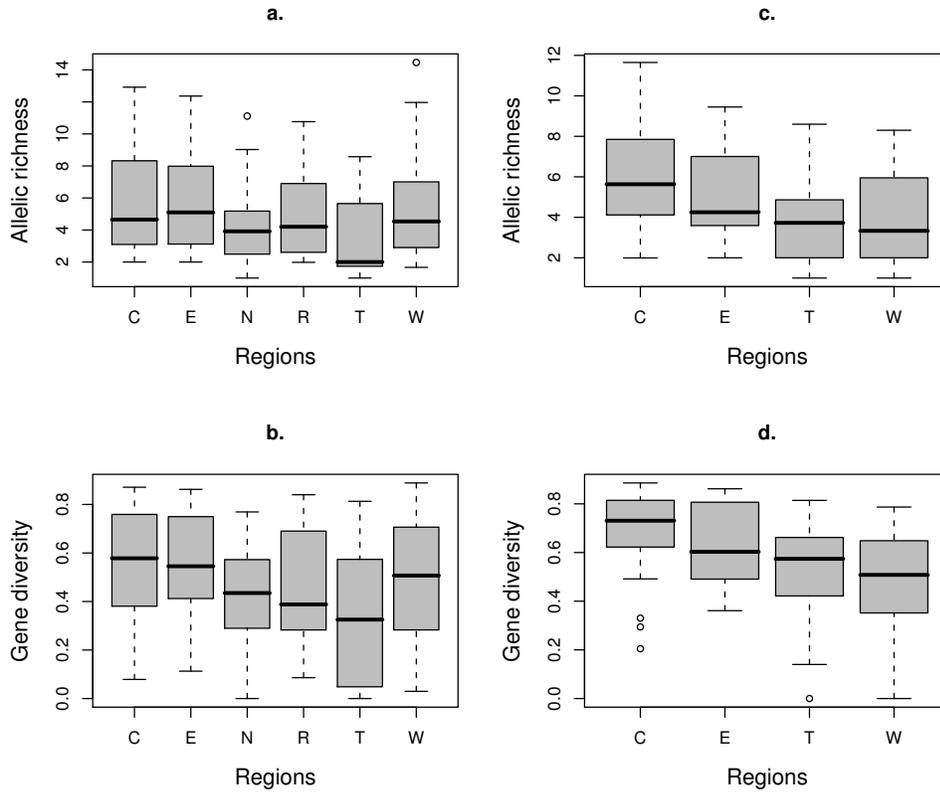


Fig. 2 Box plots showing inter-regional differences in allelic richness and gene diversity for cultivated (a & b) and wild (c & d). Letters C, E, N, R, T and W represent coast, eastern/central, north-eastern, Rift Valley, Turkana and western/Nyanza regions, respectively. The box represents the interquartile range (50% of values); while the line across the box indicates the median. The lines running vertically from the box (whiskers) extend to the highest and lowest values, excluding outliers which are denoted by circles.

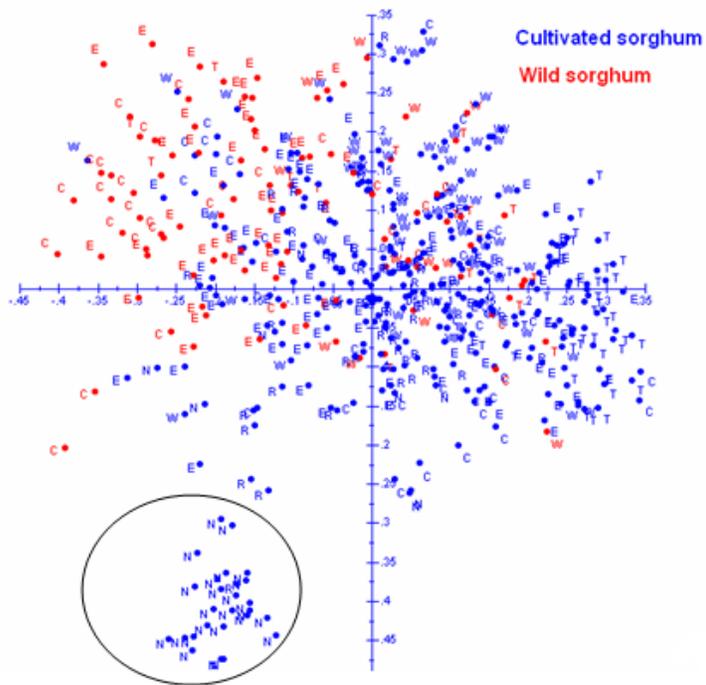


Fig. 3 Biplot of the axes 1 and 2 of the principle coordinate analysis based on the dissimilarity of 24 SSR markers for cultivated and wild sorghum. Letters C, E, N, R, T and W represent coast, eastern/central, north-eastern, Rift Valley, Turkana and western/Nyanza regions, respectively



Fig. 4 Estimated population structure at $K = 2$ for the entire sorghum gene pool ordered by type and membership fraction. Each individual is represented by a vertical line, which is partitioned into coloured segments segments that represent the individual's membership fraction in K clusters.

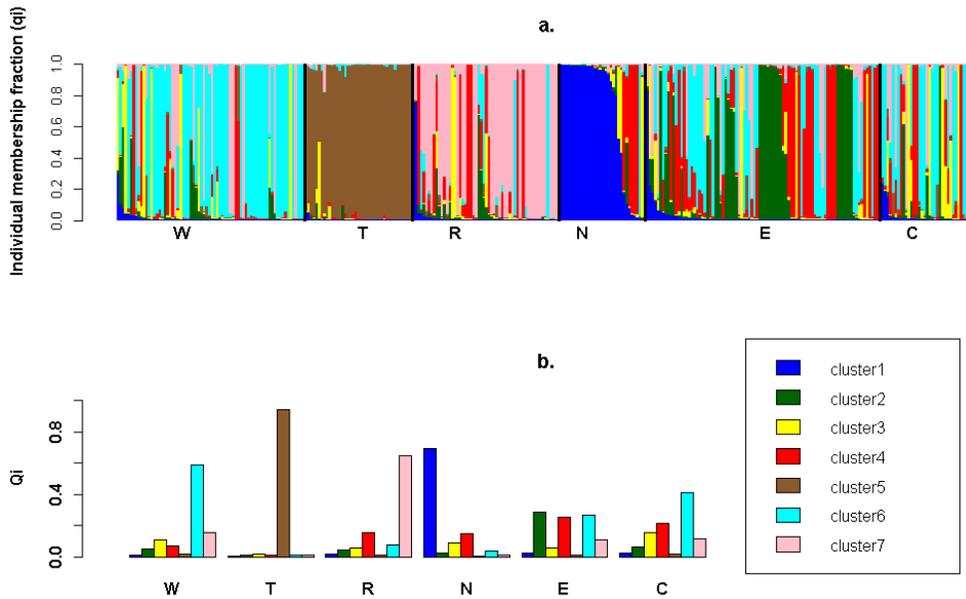


Fig. 5 Structure of the genetic diversity of the 329 cultivated sorghum plants at $K = 7$: (a) bar plot of overall partitioning and individual plant assignment (sorted by geographic origin), and (b) plots of the mean proportion of genome (Q_i) assigned in each of the $K = 7$ clusters for the group of individuals in each region. Letters C, E, N, R, T and W represent coast, eastern/central, north-eastern, Rift Valley, Turkana and western/Nyanza regions, respectively

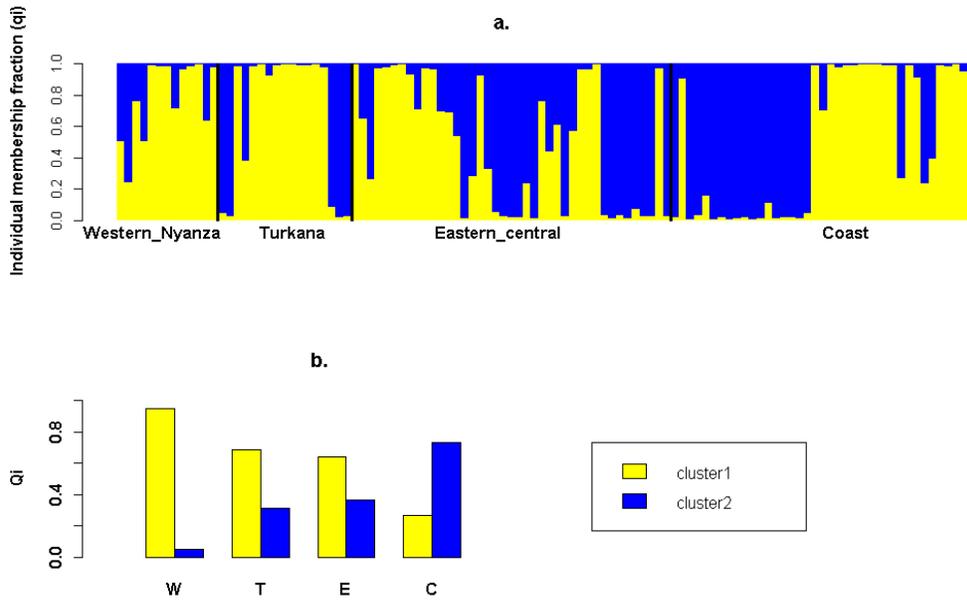


Fig. 6 Structure of the genetic diversity of the 110 wild sorghum plants at $K = 2$: (a) bar plot of overall partitioning with assignment of individual plants into the identified genetic clusters (sorted by geographic origin), and (b) plots of the overall proportion of genome (Q_i) assigned in each of the $K = 2$ clusters for the group of individuals in each region. Letters C, E, T and W represent coast, eastern/central, Turkana and western/Nyanza regions, respectively

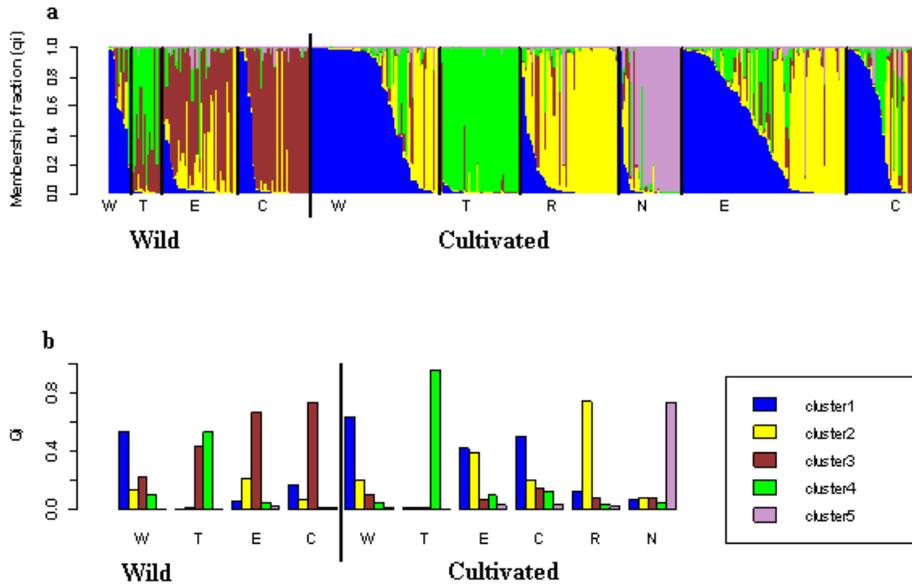


Fig. 7 Structure of the genetic diversity of the pooled cultivated and wild sorghum individuals at $K = 5$: (a) bar plot of overall partitioning with assignment of individual plants into the identified genetic clusters (sorted by geographic origin), and (b) plots of the overall proportion of genome (Q_i) assigned in each of the $K = 5$ clusters for the group of individuals in each region. Letters C, E, N, R, T and W represent coast, eastern/central, north-eastern, Rift Valley, Turkana and western/Nyanza regions, respectively

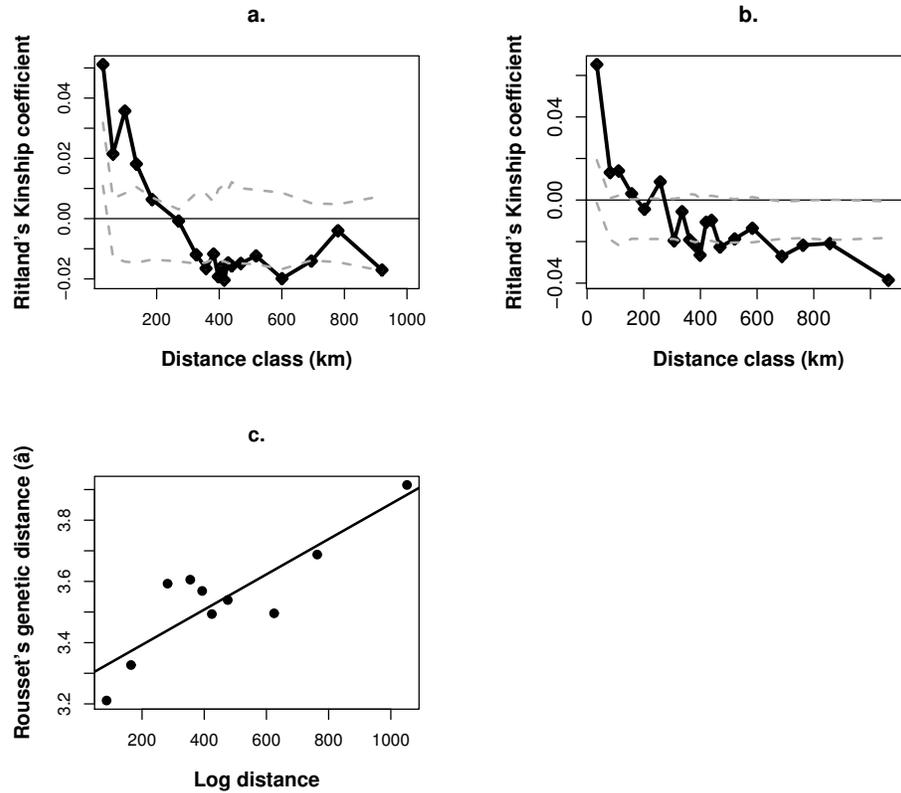


Fig. 8 Spatial patterns of genetic relatedness within and between cultivated and wild sorghum. Correlograms of pairwise relatedness (Ritland kinship coefficient) among individuals of cultivated and wild sorghum are presented in 8a and 8b, respectively. In each case, the dashed lines represent upper and lower 95% confidence limit envelopes around the null hypothesis of no spatial structure. A plot of the regression of pairwise genetic distance among cultivated and wild sorghum individuals on geographic distance is presented in 8c.

Article Title: Genetic structure and relationships within and between cultivated and wild sorghum (*Sorghum bicolor* (L.) Moench) in Kenya as revealed by microsatellite markers

Journal name: Theoretical and Applied Genetics

^{1,2,3}E. Mutegi, ^{2,4}F. Sagnard F. ⁵K. Semagn, ⁴M. Deu, ⁶M. Muraya, ⁷B. Kanyenji, ²S. de Villiers, ²D. Kiambi, ⁸L. Herselman, ⁸M. Labuschagne

¹Kenya Agricultural Research Institute (KARI), National Genebank, P.O. Box 30148, Nairobi, Kenya.

²International Crops Research Institute for the Semi-Arid Tropics (ICRISAT-Nairobi), P.O. Box 39063-00623, Nairobi, Kenya.

³ Present address: Ohio State University, Department of Evolution, Ecology, and Organismal Biology, 318 W. 12th Ave., Columbus, OH 43210-1293, USA.

⁴CIRAD, UMR DAP, F-34398 Montpellier , France

⁵International Maize and Wheat Improvement Center (CIMMYT), P.O. Box 1041, Village Market-00621, Nairobi, Kenya.

⁶Leibniz Institute of Plant Genetics and Crop Plant Research, Corrensstrasse 3, 06466 Gatersleben, Germany

⁷KARI - Embu Research Station, PO Box 27, Embu – Kenya.

⁸Department of Plant Sciences, University of the Free State, P.O. Box 339, Bloemfontein, South Africa.

Corresponding author e-mail: e_ mutegi@yahoo.com

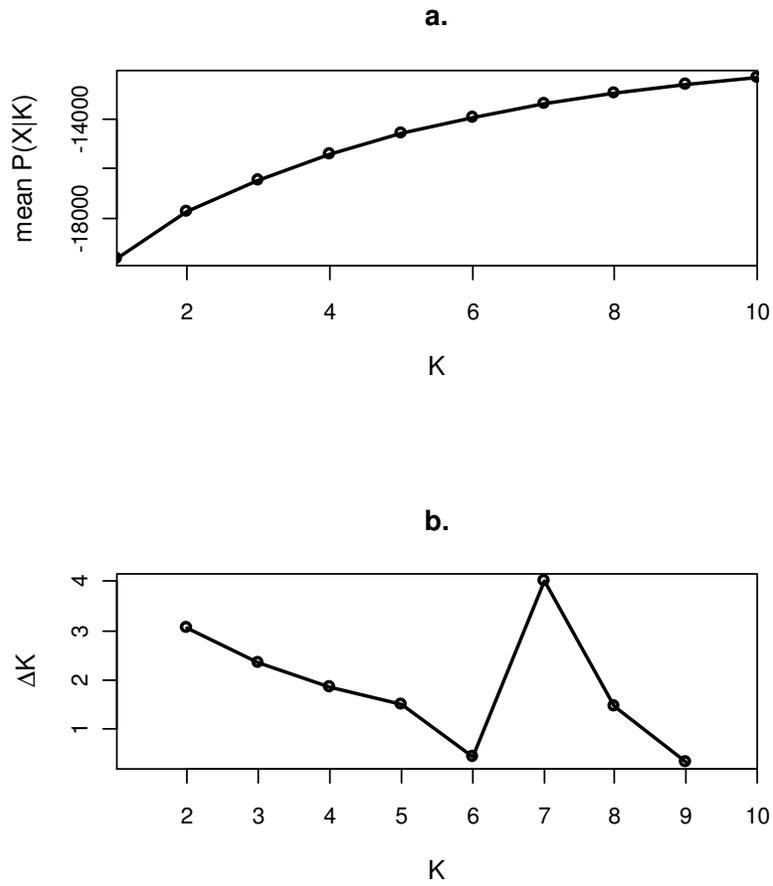


Figure S1. Results of STRUCTURE analysis for cultivated sorghum dataset. (a) Plot of mean posterior probability values with respect to K (i.e $P(X|K)$) for 10 independent runs for each value of K between $K=1$ and $K=10$. (b) Plot of ΔK , the second order rate of change of $P(X|K)$ for $K=2$ to $K=9$. The peak value of ΔK was at $K=7$, suggesting seven genetic clusters in the cultivated sorghum gene pool.

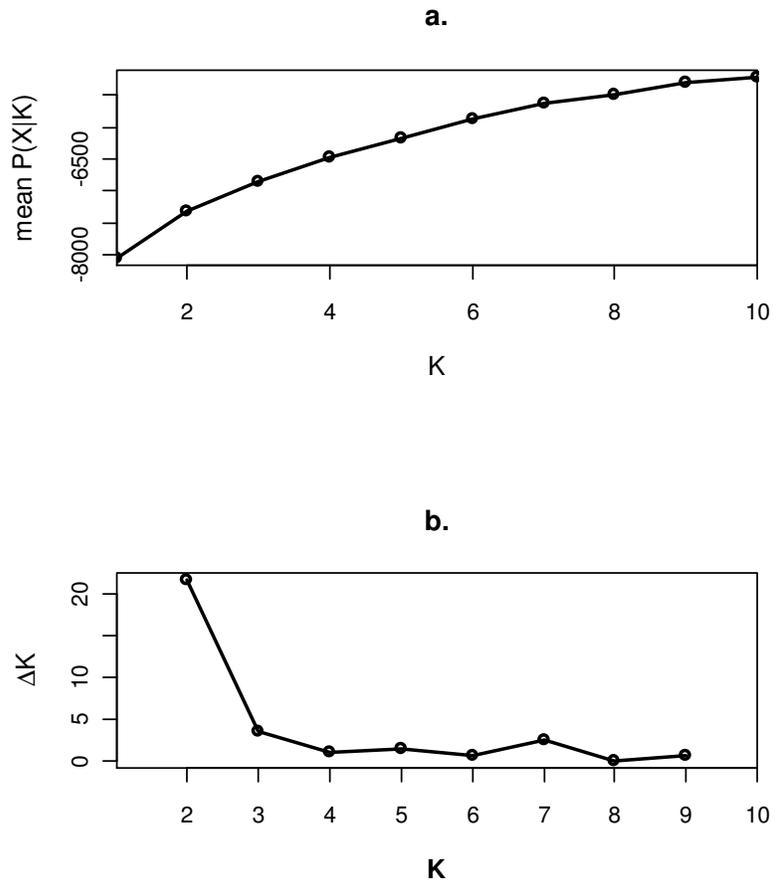


Figure S2. Results of STRUCTURE analysis for wild sorghum dataset. (a) Plot of mean posterior probability values with respect to K (i.e $P(X|K)$) for 10 independent runs for each value of K between K=1 and K=10. (b) Plot of ΔK , the second order rate of change of $P(X|K)$ for K =2 to K=9. The peak value of ΔK was at K=2, suggesting two genetic clusters in the wild sorghum gene pool.

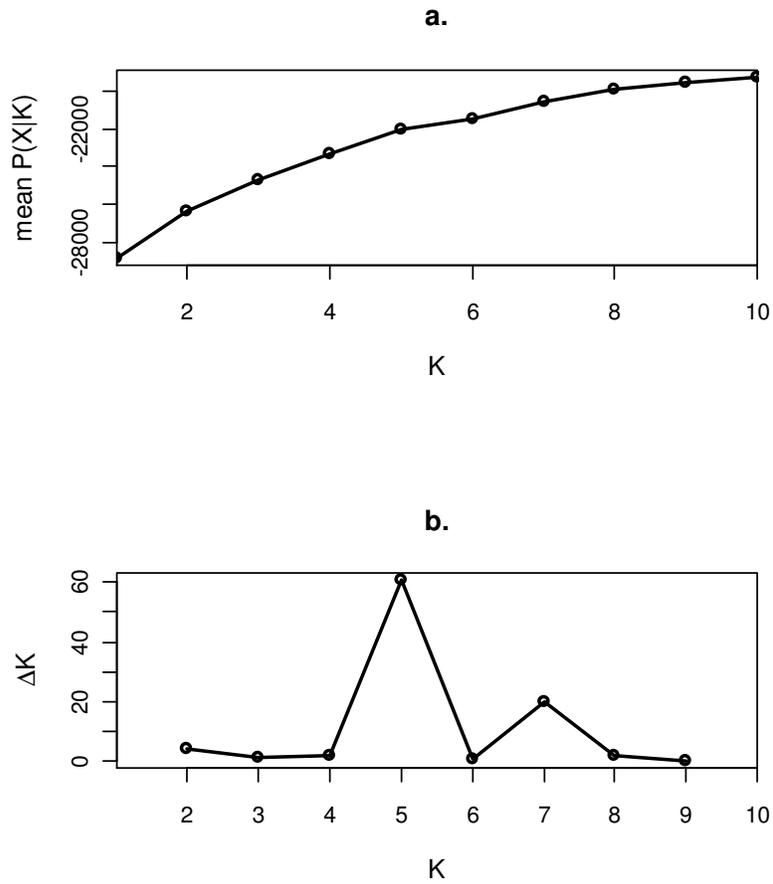


Figure S3. Results of STRUCTURE analysis for the entire dataset (cultivated and wild sorghum combined). (a) Plot of mean posterior probability values with respect to K (i.e $P(X|K)$) for 10 independent runs for each value of K between $K=1$ and $K=10$. (b) Plot of ΔK , the second order rate of change of $P(X|K)$ for $K = 2$ to $K=9$. The peak value of ΔK was at $K=5$, suggesting five genetic clusters in the combined gene pool of cultivated and wild sorghum.