Theor Appl Genet (2015) 128:1765–1775
DOI 10.1007/s00122-015-2545-6

ORIGINAL PAPER

# Genetic differentiation analysis for the identification of complementary parental pools for sorghum hybrid breeding in Ethiopia

**Taye T. Mindaye[1,2] · Emma S. Mace[3] · Ian D. Godwin[4] · David R. Jordan[1]**

Received: 18 August 2014 / Accepted: 27 April 2015 / Published online: 30 May 2015


**Abstract**
***Key message*** **The potential for exploiting heterosis for sorghum hybrid production in Ethiopia with improved local adaptation and farmers preferences has been investigated and populations suitable for initial hybrid development have been identified.**
*Abstract* Hybrids in sorghum have demonstrated increased productivity and stability of performance in the developed world. In Ethiopia, the uptake of hybrid sorghum has been limited to date, primarily due to poor adaptation and absence of farmer's preferred traits in existing hybrids. This study aimed to identify complementary parental pools to develop locally adapted hybrids, through an analysis of whole genome variability of 184 locally adapted genotypes and introduced hybrid parents (R and B). Genetic variability was assessed using genetic distance, model-based STRUCTURE analysis and pair-wise comparison of groups. We observed a high degree of genetic similarity between the Ethiopian improved inbred genotypes and a subset of landraces adapted to lowland agro-ecology with the introduced R lines. This coupled with the genetic differentiation from existing B lines, indicated that these locally adapted genotype groups are expected to have similar patterns of heterotic expression as observed between introduced R and B line pools. Additionally, the hybrids derived from these locally adapted genotypes will have the benefit of containing farmers preferred traits. The groups most divergent from introduced B lines were the Ethiopian landraces adapted to highland and intermediate agro-ecologies and a subset of lowland-adapted genotypes, indicating the potential for increased heterotic response of their hybrids. However, these groups were also differentiated from the R lines, and hence are different from the existing complementary heterotic pools. This suggests that although these groups could provide highly divergent parental pools, further research is required to investigate the extent of heterosis and their hybrid performance.

Communicated by H.-C. Jing.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-015-2545-6) contains supplementary material, which is available to authorized users.

✉ Taye T. Mindaye
tayabo@gmail.com

[1] Queensland Alliance for Agriculture and Food Innovation, Hermitage Research Facility, The University of Queensland, 604 Yangan Rd, Warwick, QLD 4370, Australia

[2] Ethiopian Institute of Agricultural Research, Melkassa Research Center, 436, Nazareth, Ethiopia

[3] Department of Agriculture, Forestry and Fisheries, Hermitage Research Facility, 604 Yangan Rd, Warwick, QLD 4370, Australia

[4] School of Agriculture and Food Sciences, The University of Queensland, Brisbane, QLD 4072, Australia

## Introduction

Sorghum (*Sorghum bicolor* (L.) Moench) is a drought- and heat-tolerant $C_4$ tropical crop with wide agro-ecological adaptations. It is a major staple food for over 500 million people in semi-arid tropical Africa and Asia (Rao et al. 2014). In Ethiopia, sorghum has been grown in diverse agro-ecologies for centuries using traditional farming systems, where it is produced both for its grain, used for leavened bread, porridge and locally produced beverages, and biomass for animal feed, fire wood and construction of fences. However, the increased demand for grain due to human population growth, coupled with more frequently

occurring droughts, has led to a requirement for technologies that increase productivity while maintaining adaptability to the changing environments.

The first commercial sales of hybrid maize seed were made in the USA in 1924 (Crow 1998) and subsequently cereal breeders have exploited the phenomena of heterosis in $F_1$ hybrids to increase the productivity of a range of crops including grain sorghum, sunflowers and canola (Duvick 1999; Reddy et al. 2006; Miller 1999). Sorghum hybrids have demonstrated superior performance compared to inbred cultivars in a range of environments, particularly under drought stress (Haussmann et al. 1998). Evaluation of hybrid sorghum in Ethiopia began in the 1970s using hybrids produced predominantly from introduced parental lines (Gebrekidan 1980). Despite demonstrable increases in grain yield, hybrids developed using the introduced parental lines were not accepted in Ethiopia. A number of factors contributed to the lack of acceptance. In particular the introduced hybrid parents were developed to perform in high-input mechanized farming systems in temperate climates and hence were neither suitable to the farming production systems practised in Ethiopia nor for the principal end uses of food consumption and plant biomass (Cavatassi et al. 2011; Mekbib 2006).

Commercial $F_1$ sorghum hybrids have been produced almost exclusively through the use of the $A_1$ cytoplasmic male sterility (CMS) system first described by Stephens and Holland (1954). This is a three-line system. A new male sterile female parent (commonly called an A line) is created by backcrossing a maintainer line (B line) into a female parent with the maternally inherited $A_1$ cytoplasm. After sufficient backcrossing, the new A line and its maintainer B line have a near-identical nuclear genome with different cytoplasm. The fertile $F_1$ hybrid is then produced by crossing the sterile A line with a male fertile restorer parent or R line that carries dominant nuclear genes which restore fertility in the hybrid with the $A_1$ cytoplasm. Fertility restoration is a multigenic trait with most sorghum lines being full or partial restorers (Jordan et al. 2010, 2011; Klein et al. 2005). The CMS system imposes constraints on the diversity of the heterotic pools, particularly B lines (Jordan et al. 2010, 2011). As a result the genetic base of B lines has remained narrow due to the difficulty in recovering good maintainer lines from B x R crosses which discourages breeders from making such crosses (Menz et al. 2004; Jordan et al. 2011). This is further exacerbated because the B lines are derived predominantly from the kafir race, which has more limited genetic variability compared to the other races of cultivated *S. bicolor* (e.g., Deu et al. 2006; Menz et al. 2004). For this reason, a practical initial strategy for developing locally adapted hybrids for Ethiopia would be to utilize existing B lines and cross these with new R lines which are locally adapted genotypes with farmer-preferred traits and increased levels of drought tolerance.

Previous studies (Betrân et al. 2003; Ganapathy et al. 2012; Van Inghelandt et al. 2010) have demonstrated that information on genetic variability of the genotypes used in hybrid breeding can be used to identify distinct heterotic parental pools, and in particular have indicated that hybrids developed using distantly related inbred lines tend to have improved heterotic responses (Reif et al. 2005). Studies in maize (Parentoni et al. 2001; Qi et al. 2010), sorghum (Gabriel 2005; Jordan et al. 2003) and sunflower (Darvishzadeh 2012) have all demonstrated a positive correlation between the genetic distance between parental lines and grain yield heterosis. In sorghum to date, a limited number of studies have used molecular markers to estimate genetic variability among male sterile and male fertile genotypes (Ahnert et al. 1996; Menz et al. 2004; Perumal et al. 2007). However, the advent of next-generation sequencing technologies and whole genome-based profiling provides new opportunities to fast-track the in-depth analysis of the genetic variability of the parental pools and to further study the relationship between heterosis and complex traits, including grain yield. Despite the wider use of the concept of heterosis for hybrid development, there is no clear understanding yet on the genetic base of heterosis with dominance, overdominance and epistasis mechanisms being suggested as possible causes for its expression (reviewed by Thiemann et al. 2009). A recent study in sorghum (Ben-Israel et al. 2012) identified overdominant heterosis mechanisms using heterotic trait loci (HTL) mapping, predominantly in pericentric-heterochromatic regions. In a recent study on maize grain yield heterosis, Thiemann et al. (2014) also identified additive heterotic quantitative trait loci (QTLs) in the pericentromeric regions of the genome. The heterochromatic region has lower rates of genetic recombination than euchromatin (Morris et al. 2013; Paterson et al. 2009), and previous studies in sorghum have demonstrated that heterochromatin is rich in QTLs and genes linked to important agronomic traits (e.g., Mace and Jordan 2011). It has previously been speculated (e.g., Larièpe et al. 2012; Mace and Jordan 2011) that genetic divergence in the heterochromatic regions between parental lines could result in the accumulation of favorable alleles in the $F_1$ hybrids that are linked in repulsion phase.

Previous studies analyzing genetic diversity in the sorghum gene pool have shown a high degree of correspondence between sorghum racial classification and marker-based grouping (Brown et al. 2011; Ramu et al. 2013). To date, a number of studies have also been conducted on sorghum germplasm from Ethiopia using limited numbers of molecular markers, e.g., with RAPD markers (Ayana et al. 2000), SSR and AFLP markers (Geleta et al. 2006) and SSR and ISSR markers (Desmae 2007), and have shown

**Table 1** Details of the seven groups of sorghum inbred lines based on their origin, agro-ecological adaptation and fertility restoration (B and R)

| Origin | Agro-ecology | Genotype number |
|---|---|---|
| Introduced R lines | | 68 |
| Introduced B lines | | 17 |
| Ethiopian landraces | Highland | 21 |
| | Intermediate | 14 |
| | Lowland (G1) | 16 |
| | Lowland (G2) | 13 |
| Ethiopian improved inbred | Lowland | 25 |
| | Intermediate | 10 |
| Total | | 184 |

the extent of genetic diversity within the landrace collections of the different agro-ecological adaptation zones. However, a comprehensive genome-wide variability study, focused particularly on understanding genetic variability of the inbred lines with the view of establishing complementary heterotic parental groups for hybrid breeding in Ethiopia, has not yet been undertaken.

In this study, we use high-throughput genotyping-by-sequencing approaches to generate whole genome profiles for Ethiopian landraces and improved inbred genotypes. We compare these profiles with those of introduced inbred lines to: (1) compare the genetic variability of Ethiopian sorghum landraces and improved inbred genotypes with introduced hybrid parents; and (2) identify complementary parental pools for future hybrid breeding in Ethiopia aimed at developing farmer-preferred hybrids adapted to different agro-ecological zones.

## Materials and methods

### Genetic materials

A total of 184 sorghum genotypes and inbred lines were selected based on their origin, maturity type, plant height and resistance to biotic and abiotic stresses (Table S1). These consisted of representative sets of 64 Ethiopian landrace genotypes of the three major agro-ecologies, highland (>1900 m), intermediate (1600–1900 m) and lowland (<1600 m) of the north eastern part of the country. Additionally, 35 genotypes classified as improved lines from the Ethiopian national sorghum pedigree breeding program bred using landraces and introduced genotypes as parents for adaptation to the intermediate and lowland agro-ecologies of the country were included. The remaining 85 genotypes were introduced inbred lines from international programs for hybrid breeding in Ethiopia classified as R (restorer) and A/B (maintainer) lines (Table 1).

### Sorghum genotyping

Total genomic DNA was extracted from 2-week-old seedlings as described by DArT P/L (DArT, www.diversityarrays.com). The samples were genotyped following an integrated DArT and genotyping-by-sequencing (GBS) methodology involving complexity reduction of the genomic DNA to remove repetitive sequences using methylation-sensitive restrictive enzymes prior to sequencing on next-generation sequencing platforms (DArT, www.diversityarrays.com). The sequence data generated were then aligned to the sorghum reference genome sequence (Paterson et al. 2009) to identify SNP (Single Nucleotide Polymorphism) markers.

### Genetic differentiation analysis of sorghum

Analysis was conducted using a combination of SNP data and genotype origin and adaptation to allocate the sorghum genotypes into groups and assess the relative genetic distance of these groups with the existing hybrid parental pools (B and R lines). The Sokal and Michener dissimilarity index was used to generate dissimilarity matrices (Sokal and Michener 1958), based initially on the complete set of SNP markers and then also based on subsets of SNPs located in the euchromatic and heterochromatic regions in the genome separately. Correlation between the genetic distances was computed using the three subsets of SNPs: euchromatic, heterochromatic and whole genome. Principal Coordinate Analysis (PCoA) and Unweighted neighbor-joining cluster analyses, using DARwin 5.0 statistical software (Perrier et al. 2003), were then used to identify pattern of genetic differentiation within and between the groups of sorghum genotypes. Comparisons were made between the neighbor-joining trees constructed using SNP marker subsets individually, with a consensus tree using the complete data set generated using the software DARwin 5.0 (Perrier et al. 2003). Polymorphic information content (PIC) and expected heterozygosity were computed for each SNP marker locus. The PIC value was calculated using the formula $PIC = 1 - \Sigma p_i^2$; where $p_i^2$ referred to the sum of the allelic frequency of each SNP for the tested genotypes and inbred groups (Anderson et al. 1993). Specific heterotic groups were then identified through a combination of genotype origin and agro-ecological adaptation zones, in addition to the outputs from the PCoA and cluster analyses. Analysis of molecular variance (AMOVA) and pair-wise population diversification analysis were conducted using Arlequin ver 3.0 statistical software (Excoffier and Schneider 2005).

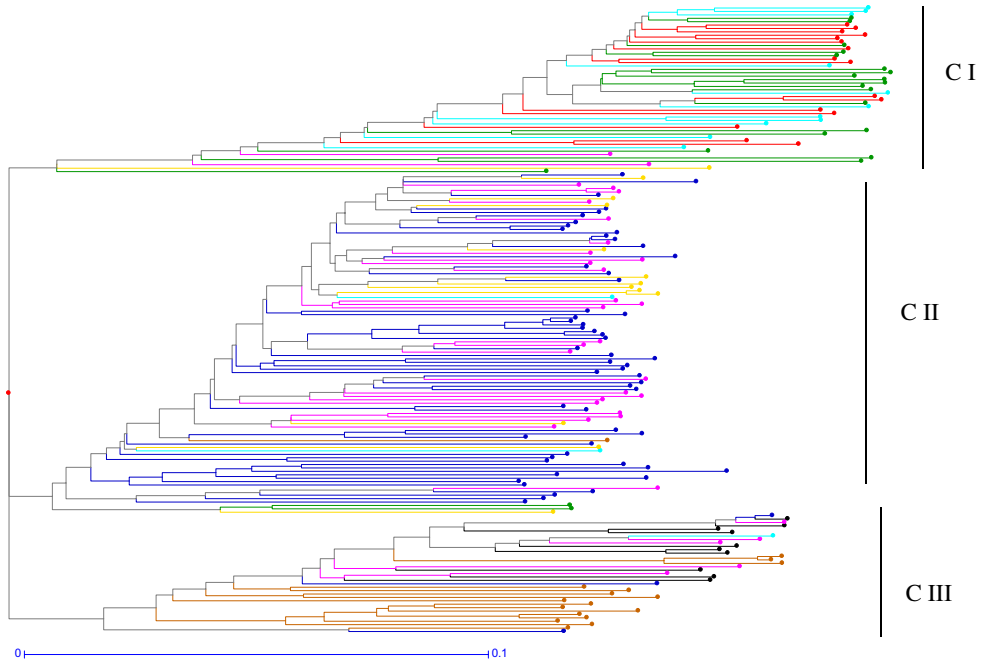


**Fig. 1** Neighbor-joining tree using the simple matching similarity coefficient based on 11,788 SNP markers for 184 sorghum genotype groups color coded as follows, introduced R lines (*blue*), introduced B lines (*orange*), Ethiopian genotypes: improved lowland (*pink*), improved intermediate (*black*), landrace lowland G1 (*red*), landrace lowland G2 (*yellow*), landrace intermediate (*light blue*), landrace highland (*green*)

Population structure analysis was also conducted for the identified groups of sorghum genotypes using the Bayesian model-based clustering algorithm in the software STRUCTURE ver 2.2 (Pritchard et al. 2000) using the SNP data excluding markers with >10 % missing data points. The admixture model with correlated allelic frequencies was used assuming regions of the genome in common across groups for each genotype (Falush et al. 2003). The model was run for the burn-in period of $1 \times 10^4$ with Markov Chain Monte Carlo (MCMC) replicates of $1 \times 10^4$ for five iterations for each population size ($k = 1$–10), and the probability values were averaged across runs for each cluster. The size of the population ($k$) was determined by the estimated logarithm of likelihood Ln P(D) for each subpopulation, where the lower variance between runs was considered as the appropriate population size (Casa et al. 2008), based on the second-order rate of change of the likelihood ($\Delta K$) (Evanno et al. 2005).

## Results

### Pattern of genetic grouping of sorghum genotypes

Sequence-based genotyping of 184 sorghum genotypes identified 11,788 polymorphic genome-wide SNP markers. The majority of the SNPs were located in the euchromatic region of the genome (84 %) with the remaining SNPs located in the heterochromatic regions. The average genetic dissimilarity between pairs of inbred lines based on the total genome-wide SNP set was 0.27, ranging from a minimum of 0.01 to a maximum of 0.39. Sub-setting the SNPs based on location in the heterochromatic or euchromatic regions produced similar overall genetic dissimilarity values across genotypes, with a high correlation coefficient between the SNP subsets ($r = 0.89$) and the whole genome set and the euchromatic SNPs ($r = 0.99$) and between the whole genome set and the heterochromatic SNPs ($r = 0.92$). Similarly, comparison between the neighbor-joining trees constructed using the euchromatic and heterochromatic SNP subsets with the consensus tree constructed with the whole genome SNP set showed a high degree of concordance (0.642 and 0.601).

The neighbor-joining tree generated using the genome-wide SNPs grouped the genotypes into three major clusters (Fig. 1) consisting of (1) the Ethiopian landrace genotypes adapted to the highland and intermediate agro-ecologies, (2) the Ethiopian improved inbred lines adapted to lowland and intermediate agro-ecologies and (3) introduced R and B lines. The lowland-adapted Ethiopian landrace genotypes separated into two different clusters. Cluster I consisted almost exclusively of the Ethiopian landraces, containing 46 (72 %) of the landrace genotypes, consisting of the majority of the highland and intermediate

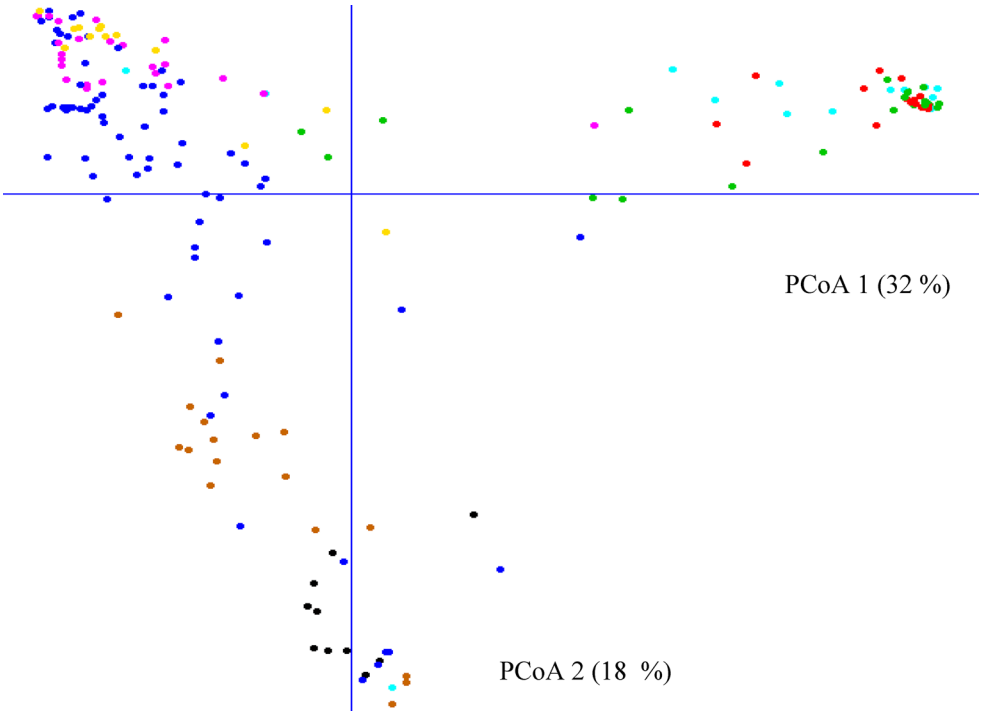


**Fig. 2** PCoA analysis for 184 sorghum genotype groups based on 11,788 SNPs color coded as follows: introduced R lines (*blue*), introduced B lines (*orange*), Ethiopian genotypes: improved lowland (*pink*), improved intermediate (*black*), landrace lowland G1 (*red*), landrace lowland G2 (*yellow*), landrace intermediate (*light blue*), landrace highland (*green*)

agro-ecological zone-adapted landrace genotypes (90 and 79 %, respectively) and approximately half (55 %) of the lowland agro-ecological zone-adapted genotypes (termed the G1 subset). Cluster II contained the remaining subset of the Ethiopian landrace genotypes adapted to the lowland agro-ecology (termed the G2 subset), together with the Ethiopian improved inbred lines also adapted to the lowland agro-ecology. The majority (95 %) of the introduced R lines were also grouped into cluster II. Finally, cluster III contained the Ethiopian improved inbred lines adapted to the intermediate agro-ecology, in addition to the introduced B lines.

A principal coordinate analysis explained 50 % of the total variation across the first two axes, i.e., 32 and 18 % for axis one and axis two, respectively (Fig. 2). The introduced B lines were located predominantly in quadrant III of the PCoA with a small proportion (23.5 %) also located in quadrant IV. The introduced R lines were spread more widely across quadrants II and III with a small proportion of the introduced R lines grouping with the introduced B lines in quadrant III. The Ethiopian improved inbreds adapted to the intermediate agro-ecological zones grouped with the introduced B lines in quadrants III and IV. The most distinct grouping of genotypes from the introduced B and R lines was in quadrant I, which consisted predominantly of Ethiopian landraces adapted to both the highland and intermediate agro-ecological zones, together with the G1 subset of the Ethiopian landraces adapted to the lowland agro-ecology. The G2 subset of the Ethiopian landrace genotypes adapted to the lowland agro-ecology grouped with Ethiopian improved lines adapted to the lowland agro-ecology along with the introduced R lines in quadrant II. The groupings based on the PCoA were also in accordance with the racial classification. The two subsets (G1 and G2) of the Ethiopian landraces adapted to the lowland agro-ecology also corresponded to the racial classification. The G1 subset was more divergent from existing R and B lines, grouping with the durra racial types of Ethiopian landraces adapted to both the highland and intermediate agro-ecological zones. The G2 subset, grouping with the Ethiopian improved inbreds adapted to the lowland agro-ecology and to the introduced R lines, consisted predominantly of caudatum and mixed racial types. These two subsets of lowland-adapted Ethiopian landraces were identified as a separate group in subsequent analysis due to the clear genetic and racial differentiation observed between them.

### Model-based genetic structure of sorghum genotypes

Genetic structures of each genotype and differentiation between the identified groups based on genome composition were carried out using STRUCTURE. The logarithm likelihood value plateaued when $k = 4$, with no significant change for sub-populations between 4 and 8 (data not shown). Based on the second-order rate of change of the logarithm likelihood for each sub-population (Evanno et al. 2005), the optimum $\Delta K$ value was obtained for population $K = 3$ (Figure S2). The model-based grouping based on three sub-populations confirmed the clear distinction of all the three groups of Ethiopian landrace genotypes, highland,

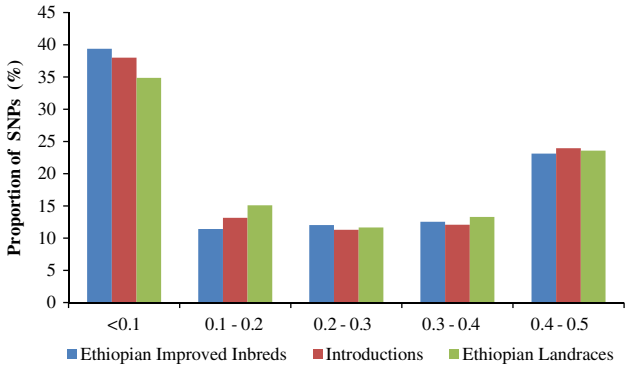


**Fig. 3** The proportion of SNPs (*y*-axis) based on PIC values (*x*-axis) for the Ethiopian improved inbreds and landrace genotype groups and introduced *inbred lines*

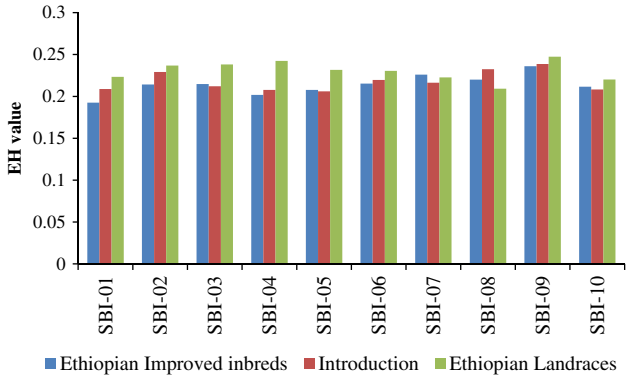


**Fig. 4** Expected Heterozygosity (EH) of sorghum genotypes based on their origin (Ethiopian improved inbreds, introduced genotypes and Ethiopian landrace) for each chromosome

intermediate and the G1 subset of lowland-adapted genotypes from all other groups of inbreds (Fig. 5). The genome composition of the Ethiopian improved inbred lines adapted to the lowland agro-ecologies was similar (60 %) to the introduced R lines. The G2 subset of lowland-adapted Ethiopian landrace genotypes also had genome composition similarity with the lowland-adapted Ethiopian improved inbreds and introduced R lines. The model-based grouping also showed the high degree of similarity in genome composition (70 %) between the introduced B lines and the Ethiopian improved inbred lines adapted to the intermediate agro-ecologies.

## Genome-wide and within-group genetic variability of sorghum genotypes

The individual SNP PIC values ranged between 0.01 and 0.5 with an average value of 0.25 across the eight groups identified. Just over a quarter (27 %) of the SNP markers had low PIC values (<0.1) with the improved inbred lines from Ethiopia and introduced genotypes having a greater proportion of the SNPs with low PIC values as compared to the landrace groups from Ethiopia (Fig. 3). The expected heterozygosity for each chromosome showed similarity between the Ethiopian improved inbreds and introduced genotype groups in contrast to the Ethiopian landrace groups, which had consistently higher heterozygosity values across 8 of the 10 chromosomes (Fig. 4). Across all chromosomes, the average expected heterozygosity was comparable between the euchromatic and heterochromatic regions for all genotypes groups (0.25 in the euchromatin, 0.24 in the heterochromatin). There were similar patterns of variability across genotype groups, with 3 chromosomes (SBI-01, SBI-03 and SBI-08) having higher heterozygosity values in the euchromatin in comparison to heterochromatin (Figure S1).

The polymorphic information content (PIC) and expected heterozygosity were calculated genome-wide for each of the eight identified genotype groups (Table 2). The Ethiopian landrace genotype groups had the highest PIC and expected heterozygosity values of all groups with the exception of the G1 subset of lowland-adapted landrace genotypes. The lowest PIC and heterozygosity values were obtained within the Ethiopian improved inbreds adapted to the intermediate agro-ecology and the G1 subset of lowland-adapted landrace genotypes, which had PIC values of 0.12 and 0.13, respectively. Although the G1 subset of lowland-adapted landraces had a high proportion of group-specific SNPs, this group also had the highest proportion of monomorphic SNPs (56 %) after the Ethiopian improved inbred lines adapted to the intermediate agro-ecologies (63 %). The introduced R lines had the highest PIC (0.19) and lowest proportion of monomorphic SNPs (24 %) as compared to the introduced B lines which had a PIC value of (0.16) and a high proportion of monomorphic SNPs (49 %). The Ethiopian landrace group adapted to the intermediate agro-ecological zone had the highest number of group-specific SNPs (428), followed by introduced R lines (262) (Table 2). The introduced R lines had higher levels of polymorphism compared to both Ethiopian improved inbred groups and the two subsets of lowland-adapted landrace genotype groups. An analysis of molecular variance revealed that variation between inbred groups accounted for 31 % of the total variation (data not shown).

## Pair-wise genetic differentiation of genotype groups

Pair-wise genetic distances, as calculated by FST values between the eight groups, were found to be significant ($p < 0.001$) for all pair-wise groupings except between the highland- and intermediate-adapted Ethiopian landrace genotype groups (Table 3). The introduced B lines had high levels of pair-wise genetic distances with all

**Table 2** Polymorphic information content (PIC), total heterozygosity (H), number of unique SNPs (Unique) and percentage of monomorphic markers (Mono) for each group of inbred lines based on agro-ecological adaptation and fertility restoration (B and R)

| Diversity indices | Introduced | | Ethiopian improved | | Ethiopian landraces | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|
| | R lines | B lines | Lowland | Intermediate | Lowland (G1) | Lowland (G2) | Intermediate | Highland | |
| PIC | 0.19 | 0.16 | 0.15 | 0.12 | 0.13 | 0.15 | 0.2 | 0.2 | 0.25 |
| H | 0.19 ± 0.2 | 0.17 ± 0.21 | 0.16 ± 0.18 | 0.14 ± 0.2 | 0.14 ± 0.2 | 0.17 ± 0.2 | 0.22 ± 0.2 | 0.21 ± 0.2 | 0.25 ± 0.2 |
| Unique | 262 | 50 | 20 | 32 | 73 | 147 | 428 | 185 | |
| Mono | 24 | 49 | 46 | 63 | 56 | 48 | 35 | 31 | |

**Table 3** Pair-wise genetic distance based on FST value for sorghum genotypes grouped based on agro-ecological adaptation and fertility restoration (B and R)

| Group of genotypes | Introduced | Ethiopian improved | | | Ethiopian landrace | | |
|---|---|---|---|---|---|---|---|
| | B | Lowland | Intermediate | Highland | Intermediate | Lowland G1 | Lowland G2 |
| Ethiopian improved | | | | | | | |
| Lowland | 0.373** | | | | | | |
| Intermediate | 0.236** | 0.478** | | | | | |
| Ethiopian Landrace | | | | | | | |
| Highland | 0.389** | 0.380** | 0.433** | | | | |
| Intermediate | 0.349** | 0.347** | 0.395** | -0.009$^{ns}$ | | | |
| Lowland (G1) | 0.496** | 0.491** | 0.557** | 0.040* | 0.053** | | |
| Lowland (G2) | 0.357** | 0.043** | 0.469** | 0.341** | 0.314** | 0.481** | |
| Introduced R | 0.252** | 0.038** | 0.359** | 0.334** | 0.300** | 0.424** | 0.061** |

*ns* not significant at ($p < 0.05$); * significant ($p < 0.01$); ** significant at ($p < 0.000$)

other groups. The pair-wise genetic distance between the introduced B lines and R lines (0.252) was lower than the overall mean FST values across all groups (0.314). Indeed 68 % of the remaining pair-wise comparisons (19/28) had higher pair-wise genetic distance values than the introduced B lines versus R lines comparison. The highest pair-wise genetic distance between the Ethiopian genotype groups and introduced B lines was between the G1 subset of lowland-adapted landrace genotypes (FST = 0.496) followed by landrace genotypes adapted to the highland agro-ecology (FST = 0.389). All the agro-ecological zone adaptation groups within the Ethiopian landraces were more genetically distinct from the introduced B lines in comparison to the B lines versus R lines pair-wise genetic distance. The only Ethiopian genotype group with a lower pair-wise genetic distance to the introduced B lines, in comparison to the B lines versus R lines FST value (0.252), was the group of Ethiopian improved inbred lines adapted to the intermediate agro-ecological zone (FST = 0.236). In addition, the most similar group to the introduced R lines was the Ethiopian improved inbred lines adapted to the lowlands (FST = 0.038) and the G2 subset of landrace genotypes adapted to the lowland agro-ecology (FST = 0.061). These data also highlighted the overall higher genetic similarity between the Ethiopian landrace genotype groups compared to Ethiopian improved inbred groups. Furthermore, there was dissimilarity between genotypes adapted to different agro-ecological zones with the exception of the G2 subset of lowland-adapted landraces from the other 3 groups of landrace genotypes.

## Discussion

Genetic variability and complementary divergent parental pools are pre-requisites for the successful exploitation of heterosis through $F_1$ hybrids. In sorghum the development of new parental pools, in particular female parents (B lines), is complicated by constraints imposed by the cytoplasmic male sterility system. We present here the first large-scale genome-wide genetic variability analysis focused on Ethiopian genotypes and show clear differentiation between Ethiopian genotype groups and existing hybrid parental pools used for commercial hybrid development in the developed world. The subset of Ethiopian lowland-adapted landraces and improved inbreds (G2), which were predominantly caudatum racial types, were found to be genetically

similar to the introduced R lines, indicating the potential for these two groups of Ethiopian genotypes to have similar patterns of heterotic responses to existing commercial hybrids when crossed with introduced B lines. These provide encouraging initial targets of complementary locally adapted genotypes to exploit heterosis in Ethiopia. In addition, a second potential and more divergent germplasm pool was identified, consisting of the Ethiopian landraces adapted to the highland and intermediate agro-ecological zones and the second subset of lowland-adapted landraces (G1), which were predominantly durra racial types. These genotypes were the most divergent from introduced B lines, indicating the potential for increased heterotic response based on parental divergence. However, as they represent very different germplasm to the existing commercial hybrid system, they present a more risky option for developing hybrids as they are less likely to have similar complementary alleles to the introduced R lines that have been selected to combine well with introduced B lines.

### Genome-wide differentiation detected within and among sorghum genotypes groups

In general, the Ethiopian landrace genotype groups across all agro-ecological adaptation zones contained more genetic diversity compared to both the Ethiopian improved inbreds and introduced genotype groups as measured by the highest number of uniquely polymorphic SNPs (17 %) and highest heterozygosity (0.22). These results are in line with previous studies in sorghum, which have also reported high genetic variation in landrace genotypes compared to improved genotypes (Mace et al. 2013; Menz et al. 2004).

The cluster- and model-based STRUCTURE analysis revealed distinct grouping and differentiation of the Ethiopian landrace groups from Ethiopian improved inbred lines adapted to two of the ago-ecological zones and from introduced R and B genotype groups. A similar result was also reported by Ayana et al. (2000) who observed genetic differentiation between Ethiopian landraces from introduced genotypes using RAPD markers. The genotype grouping observed in the current study followed racial-based classification, in line with previous reports (e.g., Deu et al. 2006; Ramu et al. 2013). The majority of the landrace genotypes predominantly adapted to highland (90 %), intermediate (79 %) and lowland (55 %) agro-ecological zones of durra racial type grouped in cluster I along with a small number of bicolor racial types. Of the remaining landrace genotypes, the G2 subset of the lowland-adapted landrace genotypes were grouped in cluster II with the caudatum racial types and mixed races of the introduced R lines and the Ethiopian improved inbreds adapted to the lowland agro-ecology. The predominance of the durra and caudatum racial types within the Ethiopian landrace genotypes included in the current study could reflect the geographic sampling. The majority of the landrace genotypes included in the study adapted to the highland and intermediate agro-ecological zones, of durra racial type originated from the North Eastern part of the country where sorghum adaptation is influenced by drought stress. The durra and caudatum racial types are highly suited to the local end use requirements, which focus on grain quality for home consumption and marketability of the seed.

Both the Ethiopian improved inbred groups and introduced genotype groups shared 67 % of their polymorphic SNPs and were located in cluster II and III of the neighbor-joining tree, indicating the commonality of the genetic background currently used for hybrid breeding. The study also revealed a clear genetic differentiation in agro-ecological adaptation within the improved inbred lines from Ethiopia, with the lowland-adapted genotypes located predominantly in cluster II, as distinct from the intermediate genotypes in cluster III, which grouped with the kaffir racial types. There was also clear genetic differentiation between genotype groups that were adapted to the same agro-ecological zone. The improved inbred lines adapted to the intermediate agro-ecological zone were genetically distinct from the landraces also adapted to the intermediate agro-ecological zone, which is likely due to the selection pressure imposed on the improved inbred lines for increased adaptation to high rainfall and humid environments.

### The Ethiopian lowland-adapted genotypes as a potential complementary parental pool for hybrid breeding for local adaptation

Development of complementary cytoplasmic male sterile (CMS) seed parent (B) and fertility restorer (R) male parental pools is the primary step in hybrid breeding, where the latter can be more readily identified from the breeding and landrace populations due to the prevalence of dominant restorer genes. However, owing to the complex nature of CMS and predominance of fertility restorer genes in the nuclear genome, B line development is both a challenging and costly process (Jordan et al. 2010). The B line pools developed to date are adapted to mechanized farming system in temperate environments. However, given the dominance of plant height (Quinby and Karper 1954) and the dominance or additivity of other important traits, hybrids developed between the existing B lines and locally adapted genotypes should produce $F_1$ hybrids that are more acceptable to Ethiopian farmers. Thus, a logical initial strategy for exploiting the utility of hybrid technology in Ethiopia would be to use the existing B lines with locally adapted and high-yielding restorer parental lines to produce hybrid

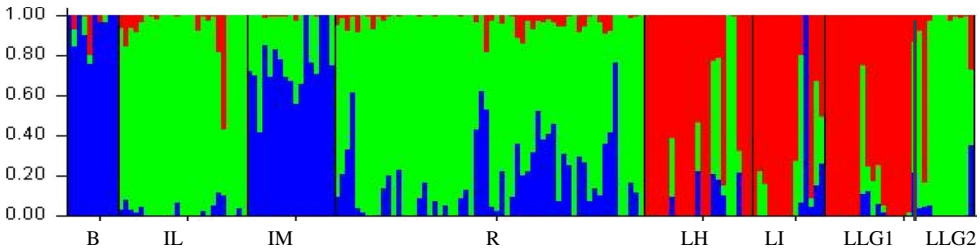


**Fig. 5** Model-based estimation of population structure for $(K = 3)$ for inbred lines identified as follows: introduced B lines (B), introduced R lines (R), Ethiopian improved lowland (IL) and improved intermediate (IM), Ethiopian landrace highland (LH), intermediate (LI), and lowland (LLG1 and LLG2). Each group is separated by a *black* vertical line

cultivars that have better adaptation to Ethiopian environments and are better suited to local end uses and farmer preferences.

With the exception of the Ethiopian improved inbred lines adapted to the intermediate agro-ecological zone, all of the Ethiopian genotype groups were genetically distinct from the introduced B lines (Figs. 2, 5). Based on STRUCTURE analysis, the Ethiopian improved inbreds adapted to the intermediate agro-ecological zone were unique among the local genotypes in sharing at least 70 % of the genome composition with the introduced B lines. Additionally, these two groups (Ethiopian inbreds adapted to the intermediate zone and the B lines) shared the lowest pair-wise genetic diversity (FST value) of all pair-wise comparison of the B line groups with the other inbred groups and had the highest proportion of monomorphic SNPs. These results indicate the potential for lower heterotic performance of any hybrids derived between these two groups.

Although the remaining local genotype groups were all genetically differentiated from the introduced B lines, only two local groups showed similarity with the existing introduced R lines; Ethiopian improved inbred lines and the G2 subset of landrace genotypes adapted to the lowland agro-ecology. The high degree of genetic similarity and genome composition as revealed through STRUCTURE analysis has provided evidence that a similar pattern of heterotic response could be expected if these Ethiopian landraces and improved inbreds adapted to the lowland agro-ecology were used in hybrid combination with introduced B lines in place of introduced R lines. The high precision of predicted breeding values for genotypes showing similar genetic grouping and linkage disequilibrium patterns has been previously reported in maize (Windhausen et al. 2012). The Ethiopian lowland-adapted landraces (G2) and improved inbred lines are expected to have similar patterns of heterotic expression as observed between the existing B and R line pools, with additional benefits of developing locally adapted hybrids containing farmers preferred traits. This represents a clear initial target for local breeding programs in Ethiopia to further explore the benefits of hybrid technology with the existing B line pools.

In contrast, the Ethiopian landraces adapted to the highland and intermediate agro-ecologies and the G1 subset of lowland-adapted genotypes were not only genetically differentiated from the introduced B lines but were also genetically distinct from the existing introduced R lines. Previous studies have reported that hybrids derived from divergent parental pools are expected to have higher heterosis due to increased level of general combining ability (Reif et al. 2005), although some studies have indicated that the extent of genetic distance for expression of heterosis is not critical (Lee et al. 2007). Additionally, the complementary trait complexes that have been selected for over the last few decades in the existing hybrid parental populations may reduce the likelihood that new inter-population combinations would be superior, even when these populations are highly divergent. However, the value of adaptation traits for the different agro-ecologies in Ethiopia including drought and other biotic stress, coupled with high degree of genetic divergence with the existing introduced B lines, indicates the value in further investigating the benefit of hybrid technology using these locally adapted landrace genotype groups with the existing B lines.

## Conclusion

Genetic analysis of the Ethiopian landraces and improved inbreds adapted to the three major agro-ecologies revealed the presence of unique alleles and distinct groupings, identifying six Ethiopian groups based on origin, agro-ecological adaptation and whole genome sequence-based SNP genetic differentiation. The Ethiopian improved inbreds and subset of landraces adapted to the lowland agro-ecology had similar patterns of differentiation with the introduced R lines and were identified as a primary target to develop locally adapted, farmer-preferred hybrids. The highly divergent Ethiopian landrace group adapted to the highland and intermediate agro-ecologies in addition to the

subset of lowland-adapted landraces were also identified as a potential second heterotic group to use with existing B lines, which would need further investigation to verify their suitability to develop locally adapted hybrids.

**Author contribution statement** TTM collected DNA samples for genotyping, conducted marker data analysis and wrote the manuscript; ESM contributed to data analysis and in writing the manuscript; IDG and DRJ contributed to data analysis and interpretation.

**Acknowledgments** The authors are thankful to AusAID (Australian Agency for International Development) for the financial support to undertake this research and sponsoring PhD scholarship to TTM and QAAFI (Queensland Alliance for Agriculture and Food Innovation) for research support.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

Ahnert D, Lee M, Austin DF, Livini C, Woodman WL, Openshaw SJ, Smith JSC, Porter K, Dalton G (1996) Genetic diversity among elite sorghum inbred lines assessed with DNA. Crop Sci 36:1385–1392

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrels ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186

Ayana A, Bryngelsson T, Bekele E (2000) Genetic variation of Ethiopian and Eritrean sorghum (*Sorghum bicolor* L. Moench) germplasm assessed by random amplified polymorphic DNA (RAPD). Genet Resour Crop Ev 47:471–482

Ben-Israel I, Kilian B, Nida H, Fridman E (2012) Heterotic Trait Locus (HTL) mapping identifies intra-locus interactions that underlie reproductive hybrid vigour in *Sorghum bicolor*. PLoS One 7:6

Betrân FJ, Ribaut JM, Beck D, Gonzalez de León D (2003) Genetic diversity, specific combining ability and heterosis in tropical maize under stress and non-stress environments. Crop Sci 43:797–806

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

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Frank CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30–40

Cavatassi R, Lipper L, Narloch U (2011) Modern variety adoption and risk management in drought prone areas: insights from the sorghum farmers of eastern Ethiopia. Agr Econ 42:279–292

Crow JF (1998) Anecdotal, historical and critical commentaries on genetics. Genetics 148:923–928

Darvishzadeh R (2012) Phenotypic and molecular marker distance as a tool for prediction of and heterosis and $F_1$ performance in sunflower (*Helianthus annus* L.) under well-watered and water stressed conditions. Aust J Crop Sci 6:732–738

Desmae H (2007) Genetic diversity and variability in grain quality of sorghum (*Sorghum bicolor* L. Moench) landraces from North–Eastern Ethiopia. PhD thesis. University of Queensland

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

Duvick DN (1999) Heterosis: Feeding people and protecting natural resources. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in Crops. Madison, Wisconsin, USA, pp 19–29

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

Excoffier LGS, Schneider S (2005) Arlequin ver. 3.0 An integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50

Falush D, Stephens M, Pritchard JK (2003) Inference in population structure using multi-locus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Gabriel K (2005) A study of heterotic relationships in sorghum. PhD Thesis. Texas A & M University

Ganapathy KN, Gomash SS, Rakshit S, Ambekar SS, Ghorade RB, Birdar BD, Saxena U, Patil JV (2012) Genetic diversity reveals utility of SSR markers in classifying parental lines and elite genotypes of sorghum (*Sorghum bicolor* L. Moench). Aust J Crop Sci 6:1486–1493

Gebrekidan B (1980) Breeding and yield evaluation of hybrid sorghum and its production prospects in Ethiopia. Ethiopian J Agri Sci 2:101–114

Geleta N, Labuschagne MT, Viljoen C (2006) Genetic diversity analysis in sorghum germplasm as estimated by AFLP, SSR and morpho agronomical markers. Biodivers Conserv 15:3251–3265

Haussmann BIG, Obilana AB, Blum A, Ayiecho PO, Schipprack W, Geiger HH (1998) Hybrid performance of sorghum and its relationship to morphological and physiological traits under variable drought stress in Kenya. Plant Breed 117:223–229

Jordan DR, Tao Y, Godwin ID, Henzell RG (2003) Prediction of hybrid performance in grain sorghum using RFLP markers. Theor Appl Genet 106:559–567

Jordan DR, Mace ES, Henzell PE, Klein RR (2010) Molecular mapping and candidate gene identification of the *Rf2* gene for pollen fertility restoration in sorghum (*Sorghum bicolor* L. Moench). Theor Appl Genet 120:1279–1287

Jordan DR, Klein RR, Sakrewski KG, Henzell RG, Klein PE, Mace ES (2011) Mapping and characterization of $RF_5$ a new gene conditioning pollen fertility restoration in $A_1$ and $A_2$ cytoplasm in sorghum (*Sorghum bicolor* L. Moench). Theor Appl Genet 123:383–396

Klein RR, Klein PE, Mullet JE, Minx P, Rooney WL, Schertz KF (2005) Fertility restorer locus *Rf1* of sorghum (*Sorghum bicolor* L. Moench) encodes a pentatricopeptide repeat protein not present in the collinear region of rice chromosome 12. Theor Appl Genet 111:994–1012

Larièpe A, Mangin B, Jasson S, Combes V, Dumas F, Jamin P, Lariagon C, Jolivot D, Madur D, Fievet J, Gallais A, Gubreuil P, Charcosset A, Moreau L (2012) The genetic basis of heterosis: multi parental quantitative trait loci mapping reveals contrasted level of apparent over dominance among traits of agronomic interest in maize (*Zea mays* L.). Genetics 190:795–811

Lee EA, Ash MJ, Good B (2007) Re-examining the relationship between degree of relatedness, genetic effects and heterosis in Maize. Crop Sci 47:629–635

Mace ES, Jordan DR (2011) Integrating sorghum whole genome sequence information with a compendium of sorghum QTL studies reveals uneven distribution of QTL and gene-rich regions with significant implications for crop improvement. Theor Appl Genet 123:169–191

Mace ES, Tai S, Gilding EK, Li Y, Prentis P, Bian L, Campbell B, Hu W, Innes DJ, Han X et al (2013) Whole-genome sequencing reveals untapped genetic potential in African's indigenous cereal crop sorghum. Nat Commun 4:2320

Mekbib F (2006) Farmer and formal breeding of sorghum (*Sorghum bicolor* (L.) Moench) and the implications for integrated plant breeding. Euphytica 152:163–176

Menz MA, Klein RR, Unruh NC, Rooney WL, Klein PE, Mullet JE (2004) Genetic diversity of public inbred of sorghum determined by mapped AFLP and SSR markers. Crop Sci 44:1236–1244

Miller JF (1999) Oilseeds and heterosis. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in Crops. Madison, Wisconsin, pp 399–404

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Kresovich S (2013) Population genomic and genome-wide association studies of agro climatic traits in sorghum. Proc Natl Acad Sci 110:453–458

Parentoni SN, Magalahâes JV, Pacheco CAP, Santos MX, Abadie T, Gama EEG, Guimarâes PEO, Merelles WF, Lopes MA, Vsconcelos MJV, Paiva E (2001) Heterotic groups based on yield-specific combining ability data and phylogenetic relationship determined by RAPD markers for 28 tropical maize open pollinated varieties. Euphytica 121:197–208

Paterson AH et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Perrier X, Floria A, Bonnet F (2003) Data analysis methods. In: Hamon P, Sequin, Perrier M, Glaszmann JC (eds) Genetic diversity of cultivated tropical plants. Enfield, Science Publishers. Montpellier, pp 43–76

Perumal R, Krishnaramanujam R, Menz MA, Katilé S, Dahlberg J, Magill CW, Rooney WL (2007) Genetic diversity among sorghum races an working groups based on AFLPs and SSRs. Crop Sci 47:1375–1383

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

Qi X, Kimati J, Li Z, Jiang L, Cui Y, Liu B (2010) Heterosis analysis using AFLP markers revealed moderate correlations between specific combing ability and genetic distance in maize inbred lines. Afr J Biot 9:1568–1572

Quinby JR, Karper RE (1954) Inheritance of height in sorghum. Agron J 46:211–216

Ramu P, Billot C, Rami JF, Senthilvel S, Updahyaya HD, Reddy LA, Hash CT (2013) Assessment of genetic diversity in the sorghum reference using EST-SSR markers. Theor Appl Genet 126:2051–2064

Rao SP, Reddy BVS, Nagaraj N, Upadhyaya HD (2014) Sorghum production for diversified uses. In: Wang YH, Upadhyaya HD, Kole C (eds) Genetics, genomics and breeding of *Sorghum*. CRC Press, FL, pp 1–27

Reddy VG, Upadhyaya HD, Gowda CLL (2006) Current status of sorghum genetic resources at ICRISAT: their sharing and impacts. Int Sorghum Millets Newsl 47:9–13

Reif JC, Hallauer AR, Melchinger AE (2005) Heterosis and heterotic pattern in maize. Maydica 50:215–223

Sokal RR, Michener CD (1958) A statistical method for evaluating system relationship. University of Kansas. Science Bulletin. pp 1409–1438

Stephens JC, Holland RF (1954) Cytoplasmic male sterility for hybrid seed sorghum production. Agron J 46:20–23

Thiemann A, Meyer S, Scholten S (2009) Heterosis in plants: manifestation in early seed development and prediction approaches to assist hybrid breeding. Chinese Sci Bull 54:2363–2375

Thiemann A, Fu J, Seifert F, Grant-Downton RT, Schrag TA, Pospisil H, Scholten S (2014) Genome-wide meta-analysis of maize heterosis reveals the potential role of additive gene expression at pericentromeric loci. BMC Plant Biol 14:88

Van Inghelandt D, Melchinger AE, Lebreton Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120:1289–1299

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink JL, Sorrells ME, Melchinger AE et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3• Genes/Genome/Genetics 2:1427–1436