

Short title: Pearl millet genomic diversity

Genetic Diversity, Population Structure, and Linkage Disequilibrium of Pearl Millet

Desalegn D. Serba*, Kebede Muleta, Paul St. Anand, Amy Bernando, Guihua Bai, Ramasamy Perumal, and Elfadil Bashir

D. Serba, R. Perumal, E. Bashir, Kansas State University, Agricultural Research Center-Hays, Hays, 1232 240th Avenue, Hays, KS 67601, USA ; K. Muleta, G. Morris, Kansas State University, Department of Agronomy, Manhattan, Kansas; P. St. Anand, A. Bernando, G. Bai, Hard Winter Wheat Genetics Research Unit, USDA-ARS, Manhattan, Kansas.

*Corresponding author (Email: ddserba@ksu.edu)

Abbreviations

GBS-genotyping-by-sequencing

GRIN- Germplasm Resource Information Network

PCA- Principal Component Analysis

PGRC- Plant Gene Resources of Canada

PMiGAP- Genetic diversity analysis in a pearl millet inbred germplasm association panel

SMIL-Sorghum and Millet Innovation Lab

SNP-single-nucleotide polymorphisms

TASSEL- Trait Analysis by aSSociation, Evolution and Linkage

Abstract

Pearl millet [*Pennisetum glaucum* (L.) R. Br.] is one of the most extensively cultivated cereals in the world, after rice, wheat, maize, barley and sorghum. It is the main component of traditional farming systems and a staple food in the arid and semi-arid regions of Africa and South Asia. However, its genetic improvement is lagging behind other major cereals and the yield is still low. Genotyping-by-sequencing (GBS)-based single nucleotide polymorphism (SNP) markers were screened on a total of 400 inbred lines and germplasm accessions from different geographic regions to assess genetic diversity, population structure and linkage disequilibrium (LD). By mapping the GBS reads to the reference genome sequence, we discovered 82,112 genome-wide SNPs. The telomeric regions of all seven chromosomes have the higher SNP density than in pericentromeric regions. Model-based clustering analysis of the population revealed a hierarchical genetic structure of six subgroups that mostly overlap with the geographic origins or sources of the genotypes but with differing levels of admixtures. A neighbor-joining phylogeny analysis of the population revealed that germplasm from West Africa rooted the dendrogram with much diversity within each subgroup. Greater LD decay was observed in the West African sub-population than in the other sub-populations, indicating a long history of recombination among landraces from West Africa. Also, selection signature analysis detected significantly different selection histories among subpopulations. This results have potential application in the development genomic-assisted breeding in pearl millet and heterotic grouping of the lines for improved hybrid performance.

Key words: pearl millet, genetic diversity, genotyping-by-sequencing, high throughput markers

Introduction

Pearl millet [*Pennisetum glaucum* (L.) R. Br.] *syn* *Cenchrus americanus*] is an important cereal crop extensively cultivated in arid and semiarid regions. It ranks sixth in area of production in the world after rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), maize (*Zea mays* L.), barley (*Hordeum vulgare* L.) and sorghum (*Sorghum bicolor* (L.) Moench) (FAO, 2014). It is cultivated on more than 30 million hectares; with a majority of the area in Africa and the Indian subcontinent (Gupta et al., 2015). It is the main component of traditional farming systems in West Africa and the Indian subcontinent. More than 500 million people depend on it as their staple food (National Research Council, 1996). Its high photosynthetic efficiency and dry matter production capacity (Yadav and Rai, 2013) make pearl millet a highly desirable crop for farmers in adverse agro-climatic regions where other cereals are likely to fail to produce economic yields. It is also grown as temporary summer pasture or cover crop in the Americas and other continents.

Pearl millet is a naturally cross-pollinating species with protogynous flowering and traditional cultivars are random-mating populations with considerable heterozygosity and heterogeneity. Hybrid breeding has become a major approach for pearl millet improvement and it has brought a progressive yield improvement, especially in India (Yadav and Rai, 2013; Kumara et al., 2014). The development of a cytoplasmic male-sterility system (CMS) (Burton, 1958) has facilitated hybrid seed production. Greater productivity is possible through genetic diversification of hybrid parents, if hybrids are developed based on heterosis prediction using parental genomic information (Gupta et al., 2018). There are several semi-dwarf inbred parental lines that were developed for hybrid breeding in the US. Assessment of genetic variability

permits the identification of genetically diverse parental materials which can enhance hybrid vigor and yield stability in variable climates (Hausmann et al., 2012; Bashir et al., 2015). Analysis of molecular diversity, population structure, and linkage disequilibrium in different sets of materials enables the identification of heterotic parental lines for enhanced hybrid vigor. Genetic diversity analysis in a pearl millet inbred germplasm association panel (PMiGAP), which represents cultivated germplasm in different areas and possessing a high gene diversity, was structured into six subpopulations (Sehgal et al., 2015). Those subpopulations supported pedigree differences and/or different characteristics of specific lines rather than their geographic origin. Also, new germplasms introduced from various sources, mainly the Germplasm Resource Information Network (GRIN) and the Plant Gene Resources of Canada (PGRC), have been collected from different geographic areas in Africa and elsewhere by multiple scientists for the purpose of preservation and utilization. However there is limited information as to the genetic variability and heterotic potential of these resources. To fill this void, inbred lines developed as seed and pollen parents and germplasm lines need to be assessed for molecular diversity using next-generation markers.

Genetic divergence between crossing parents is very important either to generate variation for selection or maximize hybrid vigor. Hence, formation of heterotic groups among the breeding populations is an essential breeding task to enhance hybrid vigor. However, there is limited research in evaluation of germplasm and breeding materials for heterotic groupings in pearl millet. Morphological traits and pedigree information have been used to characterize germplasm used for development of parents and open-pollinated varieties (Gupta et al., 2011). However, morphological traits are influenced by environment and do not measure diversity accurately.

Assessment of genetic diversity, population structure, and linkage disequilibrium is necessary to facilitate identification of heterotic groups, breeding via genomics-assisted breeding, and resource conservation. Knowledge of population structure and genetic diversity of breeding populations, germplasm, and parental lines used in the breeding program is also strikingly essential for association mapping studies, genomic selection, and genomics-assisted breeding.

The genetic improvement of pearl millet lags behind the major cereals mainly because of lack of investment in research and low yields. Genome research on pearl millet started almost at the same time as other cereals (Liu et al., 1994), but then lagged behind as the major emphasis of the genomics era was skewed to model species and major crops. Nevertheless, some efforts were made in the last few years to invigorate the genomics research in pearl millet. Genetic linkage mapping using different populations (Qi et al., 2004; Senthilvel et al., 2008; Pedraza-Garcia et al., 2010; Supriya et al., 2011), high throughput markers development and QTL mapping for important agronomic traits and stress tolerance (Yadav et al., 2004; Sehgal et al., 2012; Moumouni et al., 2015; Punnuri et al., 2016), and study of population genomics (Hu et al., 2015; Sehgal et al., 2015) have been conducted. A draft genome sequence of pearl millet ($2n=2x=14$) that can serve as a reference for further development of genomics-assisted breeding has been released (Varshney et al., 2017).

Studying the whole plant genome and its relationship with important traits facilitate cultivar development for improved yield, stress tolerance, and enhanced quality traits. A recent pearl millet whole-genome sequence (Varshney et al., 2017) is a remarkable milestone in generating genomic resources for molecular breeding. Assembly of the whole-genome sequence

and annotation of 24,000 genes indicates that enrichment of wax biosynthesis genes (Varshney et al., 2017) providing the probable genetic reason for heat and drought tolerance of the crop.

NGS technologies have considerably accelerated the investigation into the composition of genomes and their functions. Using NGS for high-throughput marker discovery and application has been too limiting to serve as a starting point for preliminary heterotic group formation in pearl millet. This genomic diversity study of inbred lines and new germplasm accessions was conducted to classify the resources for future breeding efforts. NGS-based single-nucleotide polymorphisms (SNPs) have become the marker of choice in plant breeding (Nadeem et al., 2018). Genotyping-by-sequencing (GBS), a rapid, cost-effective and reduced representation sequencing method, is a common approach for profiling genome-wide nucleotide variation in many species (Elshire et al., 2011). It has become ideal for simultaneous discovery and genotyping of thousands of SNPs across a wide range of species (Poland et al., 2012). Herein we used genome-wide GBS-SNPs to assess genetic diversity, population structure, and linkage disequilibrium (LD) of parental inbred lines developed for hybrid breeding and new germplasm lines collected from different geographic locations for trait discovery and integration.

Materials and Methods

Plant Materials

A total of 400 accessions comprising 203 inbred lines that were developed as parents for hybrid breeding and 197 germplasm lines from different sources were included in this study (Supplementary Table 1). Among them, 155 were parental inbred lines developed by Kansas State University, 27 by the University of Georgia, and seven by the University of Nebraska-

Lincoln; 200 germplasm accessions include 50 from the GRIN-Plant Genetic Resources Conservation Unit, Griffin, GA and 149 from PGRC. The germplasm accessions were diverse in geographic origin mainly from Africa, the Middle East, and India (Figure 1). Two inbred lines (16-861 and 16-911) with poor quality sequences were removed from the pool and the analysis was conducted on 398 accessions.

DNA Extraction

The seeds were germinated in 96-cell trays and grown in a greenhouse at Kansas State University. Approximately 70 to 100 mg fresh leaf tissue was collected from 2-4 plants per line 15 days after emergence. Freshly collected tissue in 96-well plates was freeze-dried for 48 hours to rapidly remove water. A 4.5 mm steel ball was added to each sample and capped plates were oscillated on a matrix mill (Retsch, Haan, Germany) at 30 cycles per second for 4 minutes to grind the tissue.

Genomic DNA was extracted from leaf tissue using a standard high-throughput 2% CTAB and chloroform:isoamyl (24:1) alcohol method in which 4 mM TCEP (tris (2-carboxyethyl) phosphine) was used in place of 2-mercaptoethanol and supplemented with 2% polyvinylpolypyrrolidone and 40 ug RNase. Sample DNA concentrations were assayed using a Quant-iT PicoGreen dsDNA HS assay kit (ThermoFisher, Waltham, MA, USA) on a FLUOstar Omega fluorescence plate reader (BMG LABTECH, Cary, NC, USA) and normalized to 20 ng μl^{-1} with 10 mM TRIS.

GBS Library Construction, Sequencing and SNP Calling

About 200 ng of genomic DNA was digested with *Pst*I (5'-CTGCA/G-3') and *Msp*I (5'-C/CGG-3') restriction enzymes (New England Biolabs, Ipswich, MA, USA). The DNA fragments from each sample were ligated to unique barcoded-adapters for identification and to allow pooling of samples for DNA sequencing and analysis.

GBS libraries were constructed as described by Mascher et al. (2013) with some modifications (Supplementary info). All adaptors and primers used for library construction and sequencing were described for Ion Torrent sequencing in Mascher et al., (2013). The concentration of adenosine 5'-triphosphate (Millipore Sigma, St. Louis, MO) used in the ligation reaction was increased to 1.25 mM, purified ligated DNA pools were quantified using the Qubit dsDNA HS assay kit (Thermo Fisher Scientific, Waltham, MA) and 7.5 ng DNA was used per 25 μ l PCR reaction. After amplification, the libraries were purified using the QIAquick PCR purification kit (Qiagen, Valencia, CA) and resuspended in a 30 μ l elution buffer, then quantified using the Qubit dsDNA HS assay kit (Thermo Fisher Scientific, Waltham, MA, USA).

Libraries were size-selected using a E-Gel system (Thermofisher.com) and 200 to 300 bp long fragments were recovered, quantified using the Qubit fluorometric quantitation system (Thermofisher.com), and normalized to a working concentration of 60 pM. Libraries were prepared for sequencing and loaded onto chips (PI v3) using the CHEF system (Thermofisher.com) and sequenced on an Ion Torrent Proton sequencer (Thermofisher.com) following manufacturer's instructions and using default analysis parameters. Each library was sequenced three times. Sequence reads from the Ion Torrent system were of variable length.

Prior to analysis, all sequencing reads had 80 poly-A bases appended to their 3' end so that TASSEL 5.0 would attempt to use reads shorter than 64 bases rather than discarding short reads.

The draft pearl millet genome sequence (Varshney et al., 2017) was used as a reference to map GBS reads and identify SNPs using the TASSEL 5.0 GBSv2 discovery pipeline (Bradbury et al. 2007, www.maizegenetics.net). The minimum locus coverage for SNP calls was 0.19 and the minimum minor allele frequency (MAF) was 0.002. All other TASSEL 5.0 settings were the defaults.

Population Structure Analysis

The millet accessions were first categorized based on their origin or source to assess the diversity within and among geographic areas and breeding programs. The Bayesian model-based quantitative assessment of population sub-clustering among the 398 pearl millet accessions was assessed using ADMIXTURE (Alexander et al., 2009). The analysis was performed based on a subset of genotypic data obtained by pruning adjacent SNP markers that are in strong LD according to the criterion of a 50 SNP window size and $r = 0.5$ using PLINK 1.9 program (Purcell et al., 2007). The percent membership of each of the accession to a sub-population was assessed assuming hypothetical subpopulations (K) ranging from 1 to 10. The most probable value of K corresponding to the number of subpopulations in the accessions was determined based on the cross-validation error parameters in the ADMIXTURE program. A cross-validation folds at 10% and a block bootstrap with 2,000 iterations were used in the analysis.

Population structure was further examined with principal components analysis (PCA) using the R package SNPRelate (Zheng et al., 2012). The genetic relationship between accessions was also determined based on the neighbor joining tree algorithm according to shared-allele distance between each pair of accession using the phylogenetic tree analysis in TASSEL software v5.2.35 (Bradbury et al., 2007). The neighbor joining tree cladogram generated by TASSEL was visualized in FigTree v1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>). Genome-wide SNP variations, including minor allele frequency, observed, and expected heterozygosity for SNP markers, were examined using VCFtools (Danecek et al., 2011). To identify genomic regions shaped by natural selection in the pearl millet population, possible reductions of nucleotide diversity between population subgroups was investigated by analyzing different ratios of nucleotide diversity (π) across the entire genome. In addition, pairwise genome-wide π , and Tajima's D test statistics (Tajima, 1989) were calculated across the genome using VCFtools (Danecek et al., 2011).

Linkage Disequilibrium Analysis

Genome-wide LD was estimated for the panel of 398 genotypes and for each subgroup (as determined by the population structure, which mostly overlapped with geographic origin). LD between pairs of SNP markers was investigated as squared allele frequency correlation (r^2) between pairs of intra-chromosomal SNPs with known genomic positions. LD among SNP markers across the genome was estimated using TASSEL v5.2.35 (Bradbury, et al., 2007). The average pattern of genome-wide LD decay over genetic distance was constructed as a scatterplot of r^2 values against the corresponding genetic distance between markers. The LD decay curve

was fitted using a non-linear regression developed by Hill and Weir (Hill and Weir, 1988), as modified by Remington (Remington et al., 2001).

Genome-wide Genetic Differentiation and Nucleotide Diversity

Pairwise estimates of genetic differentiation (F_{ST}) between different subgroups defined by population structure and geographic origin were calculated using the Weir and Cockerham's method (Weir and Cockerham, 1984). Using the VCFtools program (Danecek et al., 2011), specific outlying variants were filtered out from genetic variation data, and genome-wide F_{ST} estimates were compared between one sub-population and all remaining populations. Genome-wide distribution of selection signature was visualized by plotting Weir and Cockerham's F_{ST} against chromosomes positions. The top 0.1% F_{ST} was used to set the threshold to highlight regions for signature of selection. Nucleotide diversity within each sub-population was calculated based on a non-overlapping sliding window of 1 Mbp using VCFtools.

Results

Genome-Wide SNP Discovery

Ion Proton sequencing GBS libraries of 400 samples generated more than 540 million unique reads and 103,186,800 SNP data points. All the raw sequencing reads for all the accessions have been submitted to the NCBI Sequence Read Archive (SRA) and deposited under the accession ID, "BioProject ID": PRJNA532596. After filtering the SNPs for >20% missing, 1% MAF, and InDels, we obtained 82,112 SNPs markers (Table S3) that were distributed over all seven chromosomes. The largest number of SNPs was discovered on chromosome 1 (38,710) followed

by chromosome 2 (36,854) (Table 1). An additional 35,714 SNPs were mapped to the scaffolds not yet assigned to specific chromosomes.

Marker density ranged from 0 to 360 per Mb across the genome. Average marker density was approximately 48.3 SNPs per Mb of the genome. Markers were plotted to visualize the density and distribution of SNPs across all chromosomes (Figure 2). Genome-wide marker density showed that SNPs are more abundant in the telomeric regions of the chromosome arms than the pericentromeric regions (exact centromere location unknown). In some cases, such as on chromosome 5, more SNPs were discovered on one arm of the chromosome than on the other.

Minor allele frequency and allelic combination of SNPs were analyzed. The number of loci with a MAF less than 10 % is exceedingly larger than more frequent loci (Suppl. Figure 1A). Allele combinations indicate that the translocation (A-G, C-T) rate was more than double that of the transversion rate (A-C, G-T) (Suppl. Figure 1B).

Population Structure

The model-based clustering analysis of the population revealed a hierarchical genetic structure in the population with six subgroups (Figure 3A). The six likely subgroups were determined by assessing the ΔK statistics against 4-6 K values (Figure 3B). The six subgroups mostly overlap with the geographic origin or the source of the accessions. Accessions from West Africa were largely categorized to subgroup 1, with a few admixtures. Breeding populations from the USA and India were largely admixtures from almost all subgroups. Accessions from southern Africa

and the Middle East were clustered into distinct subgroups. Accessions obtained from eastern Africa formed several subgroups and with admixture to other subgroups.

PCA was conducted to assess the population subdivisions. PC_1 explained 5.0% of the total genetic variation found in the population and separated mostly inbred lines from landrace accessions, while PC_2 explained only 2.1% of the variation and separated the Middle East and East Africa from West Africa, USA, southern Africa, and India accessions (Fig. 3C). PCA revealed that geographic origin or source of the accessions is the most important cause of population structure. The accessions from the Middle East and East Africa were distinct from the rest of the populations. The inbred lines developed in the USA and India breeding programs are closely related to each other. The PCA also showed that the accessions from West Africa and southern Africa are closely related.

A neighbor-joining phylogeny analysis based on shared allele distance showed that germplasm from West Africa formed the root of the dendrogram (Figure 4). The analysis revealed a clustering pattern similar to the population structure found by PCA. The dendrogram, identified 12 sub-clusters with variable sizes. The first cluster had four genotypes that were collected from Niger and Nigeria. Clusters two through five were of West African origin. The sixth cluster included East African germplasm. The Southern African accessions were clustered together in cluster seven. The eighth cluster contained West African accessions and Tift 01-02. Information from USDA-GRIN and PGRC shows that ten of the genotypes in this eighth cluster were previously collected from Niger, two from Gambia, two from Mali, and the remaining one from Burkina Faso. Cluster 10 was the most complex of all with accessions from India, USA,

and East Africa. Cluster 11 included germplasm accessions from the Middle East, but mostly from Yemen. Then, the 12th cluster was formed with inbred lines developed in the US and some from India.

Linkage Disequilibrium

The degree to which alleles at two loci are associated was assessed to elucidate the patterns of genome-wide LD decay within each sub-population and across the whole population. Genome-wide LD decay in the West African sub-population was shorter than in all other sub-populations. The initial (maximum) value of average genome-wide LD (r^2) in the West African accessions declined to 0.1 at 60 kb (Figure 5). Conversely, accessions from India showed the longest LD decay (r^2 remained above 0.2 even at 200 kb). The extent of LD decay in the rest of the sub-population ranged between the West African and Indian sub-populations. The respective genetic distances at which initial LD decreased to $r^2 = 0.1$ were 500 kb, 350 kb and 82 kb in the accessions from southern Africa, the Middle East, and breeding population from the US, respectively. The initial value of average genome-wide LD (r^2) in the total population was reduced to 0.1 at about 18 kb.

Selection Signature

Genetic differentiation between population sub-groups and genome-wide patterns of nucleotide variation within each sub-population were assessed. Average genome-wide F_{ST} estimates between accessions originating from India with all other accessions ($F_{ST} = 0.006$) indicated that the Indian sub-population is less differentiated from all other sub-populations (Table 2, Figure 6). Yet, individual SNP loci that showed F_{ST} values above the threshold were detected between

these sub-populations. Average F_{ST} estimates between the East Africa sub-population and with all other accessions were also low ($F_{ST} = 0.035$). In contrast, the sub-population from the Middle East was relatively highly differentiated from all others (average F_{ST} were 0.072), followed by the inbred lines from US breeding programs ($F_{ST} = 0.060$). Considering the top 0.1% F_{ST} threshold, signatures of selection were detected at several genomic regions for the different sub-populations. The largest number of SNP loci with F_{ST} values above the threshold value ($N=55$) was detected for the sub-population of US breeding lines, followed by the West African ($N = 50$) sub-population. The lowest number of SNP loci with F_{ST} values above the threshold ($N=26$) was detected for the Indian sub-population.

Average genome-wide nucleotide diversity for the whole population was estimated to be $5.0E-04$. The highest average nucleotide diversity in a 1 kbp window was from East Africa and the Middle East (average $\pi = 5.12E-04$), followed by the West Africa (average $\pi = 5.04E-04$) sub-population. The lowest average nucleotide diversity ($\pi = 4.23E-04$) was found in the inbred lines from the US.

Discussion

Pearl millet is one of several understudied species, referred to as orphan crops (National Research Council, 1996). Limited genetic diversity studies have been made in pearl millet, especially using NGS-based high throughput SNP genotyping methods. Polymorphisms at the single nucleotide level are responsible for most of the diversity among individuals, and they often influence the expression of genes and genome evolution of a species (Shastry, 2009). SNPs are also an ideal high-throughput marker for identifying genes associated with important traits.

SNPs are the simplest and the most abundant of all genetic polymorphisms and can be found in coding, non-coding, and intronic regions of genes possessing diverse biological functions (Zheng et al., 2011). SNPs may affect transcription factor binding, gene splicing, protein folding, and many other factors at gene and transcript levels (Deng et al., 2017). The application of genomics enables the study of the genotypes and their relationship with complex phenotypic traits in plant breeding (Pérez-de-Castro et al., 2012). Integrating molecular genetics with traditional breeding significantly shortens the breeding cycle and improves selection accuracy (Lande and Thompson, 1990). Genome-wide analyses of SNPs also enable a better understanding of the selective forces that operate on a population and form a strong link between genotype and phenotype.

In plant breeding, availability of genetic diversity is a key factor for the improvement of a crop. Unless there is enough genetic variation in a breeding population, no significant improvement can be made to the crop for productivity. Therefore, understanding the structure of diversity and identification of distinct germplasm with complementary traits for hybridization provides the foundation for effective and sustained pearl millet population breeding, and for synthetic and hybrid development, based on the concept of heterotic groups (Bashir et al., 2014). A major challenge facing plant breeders is how to utilize genomic resources towards effective improvement. Study of germplasm accessions using high throughput markers offers a better understanding of the genomic diversity and population structure of the germplasm. It also facilitates gene discovery, molecular breeding, and can speed identification of heterotic groups for hybrid development.

In this study, we identified 82,112 high quality SNPs that are distributed throughout the genome. We observed an average marker density of approximately 48.3 SNPs per Mb across the genome and over all accessions. In addition to the variation between chromosomes, SNP density also varied across chromosome arm positions. There was non-uniform distribution along the chromosome arms with SNPs more abundant in the telomeric region than the pericentromeric regions either because of differences in gene coverage, actual SNP density due to recombination rates across the chromosomes, bias from the restriction enzymes, more repetitive DNA in the centromere region that leads to fewer uniquely aligned reads, or the combination of two or more of the afore mentioned reasons. It is common for the relative gene density and recombination rate to increase with the relative distance from the centromere (Akhunov et al., 2003). A trend of higher recombination rates towards the telomeres and more densely distribution SNPs was reported in common wheat (*Triticum aestivum*) (IWGSC, 2014). Likewise, the density of genes was up to 14-fold higher in the distal compared to pericentromeric region of wild emmer (*T. turgidum* ssp. *dicoccoides*) chromosome arms (Avni et al., 2017). We found the same pattern in pearl millet and likely identified many allelic variants of genes underlying important traits. These evidences show that more repetitive DNA in the pericentromeric region is the cause of less SNP density in the region.

The model-based population structure analysis clustered the population into six subgroups that mostly overlapped with the geographic origin of the accessions or the source of the genotypes. Accessions from West Africa had the least admixture while the breeding populations developed in the USA and India had the most admixtures. These results also indicated that these two regions utilized similar genetic accessions available mainly from West Africa. The slight divergence from the West African accessions is probably because of selection for certain traits in

reference to adaptation and agronomic performance in local environments. The detection of similar pearl millet inbred lines from India and the USA agrees with the germplasm and breeding materials exchange history between the USA and India through the collaboration with the International Sorghum and Millet Collaborative Research Support Program (INTSORMIL) and the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) (Andrews et al., 1993; Serba et al., 2017). This finding agrees with the genetic structure analysis in PMiGAP (Sehgal et al., 2015) which indicates that for the USA and India, pedigrees are a larger factor in genetic diversity than are the country of origin or agro-ecological zone.

We conducted PCA to assess the diversity among and within population subgroups. The first two components of the PCA explained less than 10 % of the variation, indicating that there is great diversity within these subgroups. PCA captures the correlation between genotypes and ranks the detected correlation to infer continuous axes of genetic variation (Price et al., 2006). Our results revealed that the germplasm from the Middle East and East Africa were diverged from the rest of the subgroups. Previous studies of genetic diversity observed a higher level of diversity among germplasm from the center of origin, mainly Senegalese landraces, than in accessions from southern Africa and Asia (Hu et al., 2015). The divergence of the East African and the Middle East accessions from the rest of the population suggests a secondary center of diversity for pearl millet in that geographic region. The diffusion of pearl millet to the east, following the domestication in Sahelian West Africa (Burgarella et al., 2018), might have brought adaptive genome diversification to the new environment. The utilization of those lines in Indian and USA breeding programs is low, possibly because of limited germplasm collection, selection preferences, or lack of information about genetic diversity.

The nonrandom association of alleles at different loci in the genome (LD) is a sensitive gauge of the genetic forces that structure a population (Slatkin, 2008). LD based on Pearson correlations (r^2) is a squared value of the correlation between pairs of markers across the genome, and LD decay is an estimate of the change in r^2 values over genetic distance (Weir and Cockerham, 1984). LD is a clue to understanding past evolutionary events, can aid in mapping genes that are associated with complex quantitative traits, and can explain the joint evolution of linked sets of genes. LD is also a key feature of the organization of genetic variation in natural populations (Kim et al., 2007). Therefore, systematic characterization of the patterns of LD across the genome is decisive for the design of association studies (Kim et al., 2007; Mather et al., 2007) and determines the transfer of alleles in marker-assisted breeding programs. Our estimation of LD found short decay window in the range of 200 kb genome in pearl millet. The LD was shorter in West African germplasm as compared to all other sources. The shorter LD block in West African germplasm is in agreement with a previous report of substantially shorter LD blocks in Senegalese landraces compared to other global accessions (Hu et al., 2015). Historical recombination, drift, and selection all influence the extent of LD in a population (Kauppi et al., 2003). Since West Africa is the main center of diversity (Oumar et al., 2008; Manning et al., 2011; Burgarella et al., 2018), those landraces have a long history of recombination that has shaped the nonrandom association of SNPs. As compared to related crops like sorghum, there is rapid LD decay in pearl millet. The LD decay estimated in sorghum using different marker systems was consistently decayed within 1-30 kb (Bouchet et al., 2012; Wang et al., 2013) which is also similar to rice (Mather et al., 2007).

Pearl millet has a large complex genome with abundant sequence repeats (Senthilvel et al., 2008; Varshney et al., 2017), genome duplications, inversions, and translocations as compared to well-studied related species genomes (Devos and Gale, 2000). Signatures of selection refers to favorable alleles in some region of the genome that have been preferentially increased in frequency and fixed in a population because of their functional importance in conferring enhanced fitness or productivity (Pérez O'Brien et al., 2014). Pearl millet is grown in different agro-ecologies and both natural and artificial selection have played important roles in the adaptation process. Lower genetic variability and greater LD from the causative location (Stephan et al., 2006) are the typical characteristics of selection signature. Therefore, useful signatures of selection can be obtained where many of the groups have only minor genetic differences and may not be clearly separated in a principal component analysis.

Identification of these targets harboring positive selection signatures from the SNP polymorphism data would have an important role in gene discovery, allele transfer, genomic selection, and marker assisted-selection. We assessed genome-wide signatures of selection by estimating pairwise genetic differentiation (F_{ST}) between different subgroups defined by population structure and geographic origin. F_{ST} is the ratio of the variance in the frequency of an allele between subpopulations and the variance of the allelic state in the total population (Weir and Cockerham, 1984). Differing genome-wide patterns of nucleotide variation within each subpopulation, as measured by the frequency of alleles in different populations, signifies a signature in that particular population, indicating which alleles were under positive selection. From the F_{ST} estimates for each subpopulation, we observed that accessions from the Middle East were the most highly differentiated from the remaining populations. The East African sub-population was

the second most highly differentiated from other populations. The sub-populations formed by inbred lines developed in the US had higher differentiation than the germplasm lines from India.

In line with adaptation and probable agronomic differences, signatures of selection were detected at several genomic regions for these different sub-populations. The largest number of SNP loci with F_{ST} values above the threshold value ($N=55$) was detected for the sub-population of US breeding lines, followed by the West African sub-population ($N = 50$). The least number of SNP loci with F_{ST} values above the threshold value ($N=26$) was detected for the Indian sub-population. The highest genome-wide average nucleotide diversity (1 kb window), was in the East African sub-population, followed by the Middle East and West African sub-populations. The lowest nucleotide diversity was found in the USA sub-population and may be due to targeted selection for semi-dwarf, photoperiod neutral, and early maturity types, or to limited germplasm utilization. A signature of selection study using animal genomes suggested that many different populations will be necessary for a full description of selective signatures across the genome (Barendse et al., 2009).

Conclusions

GBS_SNPs captured much of the genome variation within the populations studied. Analysis of population structure revealed that diversity was high within the pearl millet subgroups studied. LD decay was shortest in populations from the main center of diversity, likely because of historical recombination among landraces. Signature of selection analysis revealed a number of outlying SNPs which may be associated with important traits, local to each subgroup. The high level of admixture in the US and Indian subgroups indicated that a large number of similar

germplasm lines has been shared between those subpopulations for parental inbred line development. Incorporating more of the East African and Middle East germplasm lines will benefit the development of diverse parental lines for increased heterosis in the US and Indian breeding programs.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

Funding for this research was provided by the United States Agency for International Development under Cooperative Agreement No. AID-OAA-A-13-00047 with the Kansas State University Sorghum and Millet Innovation Lab (SMIL). The contents are solely the responsibility of the authors and do not necessarily reflect the views of USAID or others. We also acknowledge the Plant Genetic Resource Program of USDA-GRIN and Plant Gene Bank of Canada (PGBC) for providing germplasm accessions from their collections. We are also grateful to Dr. Wayne Hanna for providing inbred lines developed at Tifton Georgia. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by any part herein. USDA is an equal opportunity provider and employer. This is contribution number xx-xxx-x from the Kansas Agricultural Experiment Station.

References

- Akhunov, E.D., A.W. Goodyear, S. Geng, L.-L. Qi, B. Echaliier, B.S. Gill, Miftahudin, J.P. Gustafson, G. Lazo, S. Chao, O.D. Anderson, A.M. Linkiewicz, J. Dubcovsky, M. La Rota, M.E. Sorrells, D. Zhang, H.T. Nguyen, V. Kalavacharla, K. Hossain, S.F. Kianian, J. Peng, N.L. V Lapitan, J.L. Gonzalez-Hernandez, J.A. Anderson, D.-W. Choi, T.J. Close, M. Dilbirligi, K.S. Gill, M.K. Walker-Simmons, C. Steber, P.E. McGuire, C.O. Qualset, and J. Dvorak. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res.* 13(5): 753–763.
- Alexander, D.H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19(9): 1655–1664.
- Andrews, D.J., J.F. Rajewski, and K.A. Kumar. 1993. Pearl millet: New feed grain crop. p. 198–208. *In* Janick, J., Simon, J.E. (eds.), *New Crops*. Wiley, New York.
- Avni, R., M. Nave, O. Barad, K. Baruch, S.O. Twardziok, H. Gundlach, I. Hale, M. Mascher, M. Spannagl, K. Wiebe, K.W. Jordan, G. Golan, J. Deek, B. Ben-Zvi, G. Ben-Zvi, A. Himmelbach, R.P. MacLachlan, A.G. Sharpe, A. Fritz, R. Ben-David, H. Budak, T. Fahima, A. Korol, J.D. Faris, A. Hernandez, M.A. Mikel, A.A. Levy, B. Steffenson, M. Maccaferri, R. Tuberosa, L. Cattivelli, P. Faccioli, A. Ceriotti, K. Kashkush, M. Pourkheirandish, T. Komatsuda, T. Eilam, H. Sela, A. Sharon, N. Ohad, D.A. Chamovitz, K.F.X. Mayer, N. Stein, G. Ronen, Z. Peleg, C.J. Pozniak, E.D. Akhunov, and A. Distelfeld. 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357(6346): 93 LP-97.
- Barendse, W., B.E. Harrison, R.J. Bunch, M.B. Thomas, and L.B. Turner. 2009. Genome wide signatures of positive selection: The comparison of independent samples and the

- identification of regions associated to traits. *BMC Genomics* 10(1): 178.
- Bashir, E.M.A., A.M. Ali, A.M. Ali, A.E. Melchinger, H.K. Parzies, and B.I.G. Haussmann. 2014. Characterization of Sudanese pearl millet germplasm for agro-morphological traits and grain nutritional values. *Plant Genet. Resour.* 12(1): 35–47.
- Bashir, E.M.A., A.M. Ali, A.M. Ali, E.T.I. Mohamed, A.E. Melchinger, H.K. Parzies, and B.I.G. Haussmann. 2015. Genetic diversity of Sudanese pearl millet (*Pennisetum glaucum* (L.) R. Br.) landraces as revealed by SSR markers, and relationship between genetic and agro-morphological diversity. *Genet. Resour. Crop Evol.* 62(4): 579–591.
- Bouchet, S., D. Pot, M. Deu, J.-F. Rami, C. Billot, X. Perrier, R. Rivallan, L. Gardes, L. Xia, P. Wenzl, A. Kilian, and J.-C. Glaszmann. 2012. Genetic Structure, Linkage Disequilibrium and Signature of Selection in Sorghum: Lessons from Physically Anchored DArT Markers. *PLoS One* 7(3): e33470.
- Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19): 2633–2635.
- Burgarella, C., P. Cubry, N.A. Kane, R.K. Varshney, C. Mariac, X. Liu, C. Shi, M. Thudi, M. Couderc, X. Xu, A. Chitkineni, N. Scarcelli, A. Barnaud, B. Rhoné, C. Dupuy, O. François, C. Berthouly-Salazar, and Y. Vigouroux. 2018. A western Sahara centre of domestication inferred from pearl millet genomes. *Nat. Ecol. Evol.* 2: 1377–1380.
- Burton, G.W. 1958. Cytoplasmic male-sterility in pearl millet (*Pennisetum glaucum*) (L.). *Agron. J.* 40: 230.
- Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project

- Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* 27(15): 2156–2158.
- Deng, N., H. Zhou, H. Fan, and Y. Yuan. 2017. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget* 8(66): 110635–110649.
- Devos, K.M., and M.D. Gale. 2000. Genome relationships: the grass model in current research. *Plant Cell* 12(5): 637–646.
- Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6(5): e19379.
- Gupta, S.K., R. Bhattacharjee, K.N. Rai, and M.S. Kumar. 2011. Characterization of ICRISAT-bred restorer parents of pearl millet. 9(December): 1–5.
- Gupta, S.K., T. Nepolean, C.G. Shaikh, K. Rai, C.T. Hash, R.R. Das, and A. Rathore. 2018. Phenotypic and molecular diversity-based prediction of heterosis in pearl millet (*Pennisetum glaucum* L. (R.) Br.). *Crop J.* 6(3): 271–281.
- Gupta, S.K., K.N. Rai, P. Singh, V.L. Ameta, S.K. Gupta, A.K. Jayalekha, R.S. Mahala, S. Pareek, M.L. Swami, and Y.S. Verma. 2015. Seed set variability under high temperatures during flowering period in pearl millet (*Pennisetum glaucum* L. (R.) Br.). *Field Crops Res.* 171: 41–53.
- Hausmann, B.I.G., H. Fred Rattunde, E. Weltzien-Rattunde, P.S.C. Traoré, K. vom Brocke, and H.K. Parzies. 2012. Breeding strategies for adaptation of pearl millet and sorghum to climate variability and change in West Africa. *J. Agron. Crop Sci.* 198(5): 327–339.
- Hill, W.G., and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* 33(1): 54–78.

- Hu, Z., B. Mbacké, R. Perumal, M.C. Guèye, O. Sy, S. Bouchet, P.V.V. Prasad, and G.P. Morris. 2015. Population genomics of pearl millet (*Pennisetum glaucum* (L.) R. Br.): Comparative analysis of global accessions and Senegalese landraces. *BMC Genomics* 16(1): 1048.
- IWGSC. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345(6194).
- Kauppi, L., A. Sajantila, and A.J. Jeffreys. 2003. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.* 12(1): 33–40.
- Kim, S., V. Plagnol, T.T. Hu, C. Toomajian, R.M. Clark, S. Ossowski, J.R. Ecker, D. Weigel, and M. Nordborg. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* 39: 1151.
- Kumara, C.D., C. Bantilan, A. Rajalaxmi, K.N. Rai, O.P. Yadav, S.K. Gupta, N.P. Singh, and S.D. Moses. 2014. Development and diffusion of pearl millet improved cultivars in India: Impact on growth and yield stability. *Work. Pap. Ser. No. 52. ICRISAT, Patancheru, India.*: 76.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124(3): 743 LP-756.
- Liu, C.J., J.R. Witcombe, T.S. Pittaway, M. Nash, C.T. Hash, C.S. Busso, and M.D. Gale. 1994. An RFLP-based genetic map of pearl millet (*Pennisetum glaucum*). *Theor. Appl. Genet.* 89(4): 481–487.
- Manning, K., R. Pelling, T. Higham, J.L. Schwenniger, and D.Q. Fuller. 2011. 4500-Year old domesticated pearl millet (*Pennisetum glaucum*) from the Tilemsi Valley, Mali: New insights into an alternative cereal domestication pathway. *J. Arch. Sci.* 38(2): 312–322.

- Mather, K.A., A.L. Caicedo, N.R. Polato, K.M. Olsen, S. McCouch, and M.D. Purugganan. 2007. The Extent of Linkage Disequilibrium in Rice (*Oryza sativa* L.). *Genetics* 177(4): 2223 LP-2232.
- Moumouni, K.H., B.A. Kountche, M. Jean, C.T. Hash, Y. Vigouroux, B.I.G. Hausmann, and F. Belzile. 2015. Construction of a genetic map for pearl millet, *Pennisetum glaucum* (L.) R. Br., using a genotyping-by-sequencing (GBS) approach. *Mol. Breed.* 35(1): 1–10.
- Nadeem, M.A., M.A. Nawaz, M.Q. Shahid, Y. Doğan, G. Comertpay, M. Yıldız, R. Hatipoğlu, F. Ahmad, A. Alsaleh, N. Labhane, H. Özkan, G. Chung, and F.S. Baloch. 2018. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Equip.* 32(2): 261–285.
- National Research Council. 1996. *Lost Crops of Africa. Volume I: Grains.* National Academy Press, Washington, D.C.
- Oumar, I., C. Mariac, J.-L.L. Pham, and Y. Vigouroux. 2008. Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theor. Appl. Genet.* 117(4): 489–497.
- Pedraza-Garcia, F., J.E. Specht, and I. Dweikat. 2010. A new PCR-based linkage map in pearl millet. *Crop Sci.* 50: 1754–1760.
- Pérez-de-Castro, A.M., S. Vilanova, J. Cañizares, L. Pascual, J.M. Blanca, M.J. Díez, J. Prohens, and B. Picó. 2012. Application of Genomic Tools in Plant Breeding. *Curr. Genomics* 13(3): 179–195.
- Pérez O'Brien, A.M., Y.T. Utsunomiya, G. Mészáros, D.M. Bickhart, G.E. Liu, C.P. Van Tassell, T.S. Sonstegard, M.V.B. Da Silva, J.F. Garcia, and J. Sölkner. 2014. Assessing signatures of selection through variation in linkage disequilibrium between taurine and

- indicine cattle. *Genet. Sel. Evol.* 46(1): 19.
- Poland, J.A., P.J. Brown, M.E. Sorrells, and J.L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7(2).
- Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38: 904.
- Punnuri, S.M., J.G. Wallace, J.E. Knoll, K.E. Hyma, S.E. Mitchell, E.S. Buckler, R.K. Varshney, and B.P. Singh. 2016. Development of a high-density linkage map and tagging leaf spot resistance in pearl millet using genotyping-by-sequencing markers. *The Plant Genome* 9(2): 1–13.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81(3): 559–575.
- Qi, X., T.S. Pittaway, S. Lindup, H. Liu, E. Waterman, F.K. Padi, C.T. Hash, J. Zhu, M.D. Gale, and K.M. Devos. 2004. An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, *Pennisetum glaucum*. *Theor. Appl. Genet.* 109(7): 1485–1493.
- Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* 98(20): 11479–11484.
- Sehgal, D., V. Rajaram, I.P. Armstead, V. Vadez, Y.P. Yadav, C.T. Hash, and R.S. Yadav. 2012.

- Integration of gene-based markers in a pearl millet genetic map for identification of candidate genes underlying drought tolerance quantitative trait loci. *BMC Plant Biol.* 12(1): 9.
- Sehgal, D., L. Skot, R. Singh, R.K. Srivastava, S.P. Das, J. Taunk, P.C. Sharma, R. Pal, B. Raj, C.T. Hash, and R.S. Yadav. 2015. Exploring potential of pearl millet germplasm association panel for association mapping of drought tolerance traits. *PLoS One* 10(5): e0122165.
- Senthilvel, S., B. Jayashree, V. Mahalakshmi, P.S. Kumar, S. Nakka, T. Nepolean, and C. Hash. 2008. Development and mapping of simple sequence repeat markers for pearl millet from data mining of expressed sequence tags. *BMC Plant Biol.* 8: 119.
- Serba, D.D., R. Perumal, T.T. Tesso, and D. Min. 2017. Status of global pearl millet breeding programs and the way forward. *Crop Sci.* 57: 2891–29050.
- Shastri, B.S. 2009. SNPs: Impact on Gene Function and Phenotype. p. 3–22. *In* Komar, A.A. (ed.), *Single Nucleotide Polymorphisms: Methods and Protocols*. Humana Press, Totowa, NJ.
- Slatkin, M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9(6): 477–485.
- Stephan, W., Y.S. Song, and C.H. Langley. 2006. The Hitchhiking Effect on Linkage Disequilibrium Between Linked Neutral Loci. *Genetics* 172(4): 2647–2663.
- Supriya, A., S. Senthilvel, T. Nepolean, K. Eshwar, V. Rajaram, R. Shaw, C.T. Hash, A. Kilian, R.C. Yadav, and M.L. Narasu. 2011. Development of a molecular linkage map of pearl millet integrating DArT and SSR markers. *Theor. Appl. Genet.* 123(2): 239–250.
- Tajima, F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* 123(3): 585–595.

- Varshney, R.K., C. Shi, M. Thudi, C. Mariac, J. Wallace, P. Qi, H. Zhang, Y. Zhao, X. Wang, A. Rathore, R.K. Srivastava, A. Chitikineni, G. Fan, P. Bajaj, S. Punnuri, S.K. Gupta, H. Wang, Y. Jiang, M. Couderc, M.A.V.S.K. Katta, D.R. Paudel, K.D. Mungra, W. Chen, K.R. Harris-Shultz, V. Garg, N. Desai, D. Doddamani, N.A. Kane, J.A. Conner, A. Ghatak, P. Chaturvedi, S. Subramaniam, O.P. Yadav, C. Berthouly-Salazar, F. Hamidou, J. Wang, X. Liang, J. Clotault, H.D. Upadhyaya, P. Cubry, B. Rhoné, M.C. Gueye, R. Sunkar, C. Dupuy, F. Sparvoli, S. Cheng, R.S. Mahala, B. Singh, R.S. Yadav, E. Lyons, S.K. Datta, C. Tom Hash, K.M. Devos, E. Buckler, J.L. Bennetzen, A.H. Paterson, P. Ozias-Akins, S. Grando, J. Wang, T. Mohapatra, W. Weckwerth, J.C. Reif, X. Liu, Y. Vigouroux, and X. Xu. 2017. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* 35(10): 969–976.
- Wang, Y.-H., H.D. Upadhyaya, A.M. Burrell, S.M.E. Sahraeian, R.R. Klein, and P.E. Klein. 2013. Genetic Structure and Linkage Disequilibrium in a Diverse, Representative Collection of the C4 Model Plant, *Sorghum bicolor*. *G3, Genes|Genomes|Genetics* 3(5): 783 LP-793.
- Weir, B.S., and C.C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38(6): 1358.
- Yadav, R.S., C.T. Hash, F.R. Bidinger, K.M. Devos, and C.J. Howarth. 2004. Genomic regions associated with grain yield and aspects of post-flowering drought tolerance in pearl millet across stress environments and tester background. *Euphytica* 136(3): 265–277.
- Yadav, O.P., and K.N. Rai. 2013. Genetic improvement of pearl millet in India. *Agri. Res.* 2(4): 275–292.
- Zheng, L.-Y., X.-S. Guo, B. He, L.-J. Sun, Y. Peng, S.-S. Dong, T.-F. Liu, S. Jiang, S. Ramachandran, C.-M. Liu, and H.-C. Jing. 2011. Genome-wide patterns of genetic variation

in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol.* 12(11): R114.

Zheng, X., D. Levine, J. Shen, S.M. Gogarten, C. Laurie, and B.S. Weir. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28(24): 3326–3328.

Figure Legends:

Figure 1. Origin or sources of 400 pearl millet genotypes used in this study.

Figure 2. Genome-wide single-nucleotide polymorphism (SNP) distributions of 82,112 high quality SNPs detected by GBS of 398 inbred lines and accessions on the seven chromosomes of pearl millet.

Figure 3. Population structure of the 398 pearl millet genotypes: (A) Bayesian posterior probability of membership determined by the model-based clustering method for hypothetical subpopulations, K values of 4-6. The color of the vertical bar on the x-axis represents the proportion of membership of each accession in each subgroup or clusters. The geographic regions of the accessions origin are also indication on the x-axis, which are designated as EA/India (East Africa/India), ME (Middle East), SA (Southern Africa), USA and WA (West Africa). (B) Plot of ΔK statistics vs K values, which determines the likely number of subpopulations. (C) Population structure based on principal component analysis (PCA).

Figure 4. Neighbor-joining dendrogram of 398 pearl millet inbred lines and germplasm accessions based on 82,112 SNP markers. Genotypes have been assigned colors based on geographic origin or source.

Figure 5. Genome-wide LD (r^2) decay in the whole population of 398 genotypes and subgroups. Average r^2 (squared allele frequency correlation between pairs of intra-chromosomal SNPs) were plotted against the corresponding genetic distance between markers.

Figure 6. Genome scans for differentiation among population sub-groups calculated using Weir and Cockerham's F_{ST} method. The x-axis represents the chromosomes and y-axis the F_{ST} values. Each point represents a SNP marker. The red horizontal line represents the threshold based on the top 0.1% F_{ST} values. A) East Africa; B) India; C) Middle East; D) Southern Africa; E) United States; F) West Africa.

Table Legends:

Table 1: Number of SNPs detected, and marker gaps from GBS of 398 genotypes.

Table 2: Pairwise F_{ST} statistics and genome-wide nucleotide diversity per 1 kbp.

Supplementary Files:

Supplementary file. Table S1. List of inbred lines and germplasm accessions used in this study.

Supplementary file: Table S2. SNP loci in each subpopulation that showed F_{ST} value above 0.1% threshold.

Supplementary file: Table S3. SNP location, SNP alleles and 400 bp flanking sequences for 82,112 markers developed. The marker name is made up of "Chromosome_Position", the SNP bases in a paired format like "C/T", and 400 bp flanking sequences (marker is at 200th bp) are arranged in columns. The genotype calls for each individual accession are in rows.

Supplementary file: Figure S1. Frequency distribution of the minor alleles for each of the SNPs scored in a population sample of 398 (A) and allelic combinations of SNPs in the population against the reference genome (B).

Supplementary file: Figure S2. Genome-wide LD (r^2) decay in the whole population of 398 genotypes. Average r^2 (squared allele frequency correlation between pairs of intra-chromosomal SNPs) were plotted against the corresponding genetic distance between markers.

Figure 1

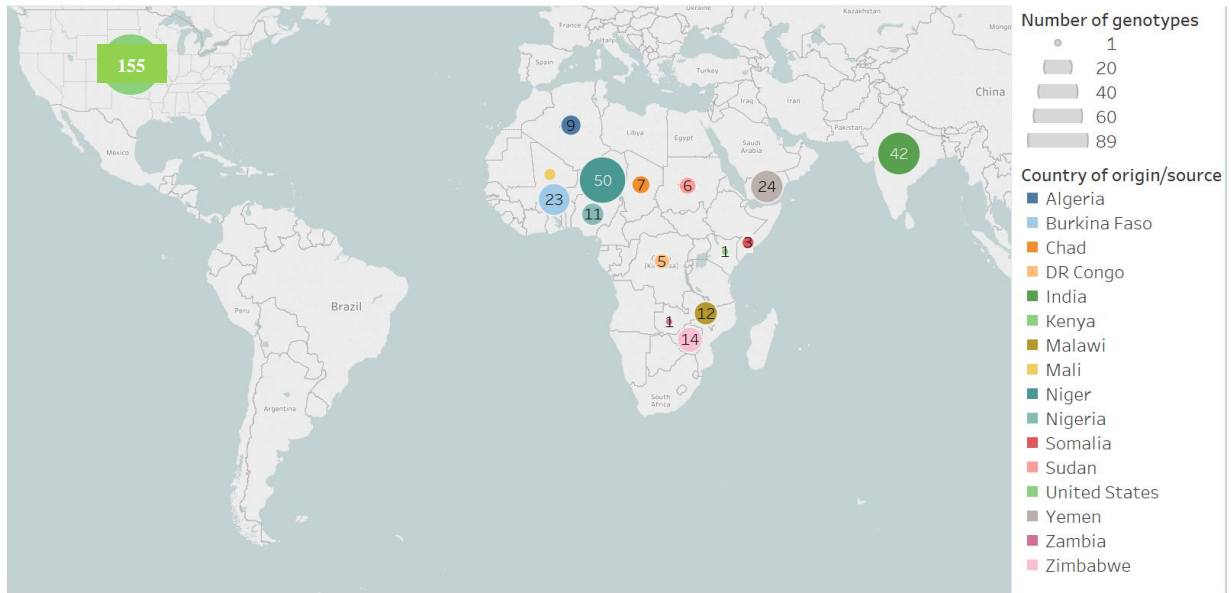


Figure 2

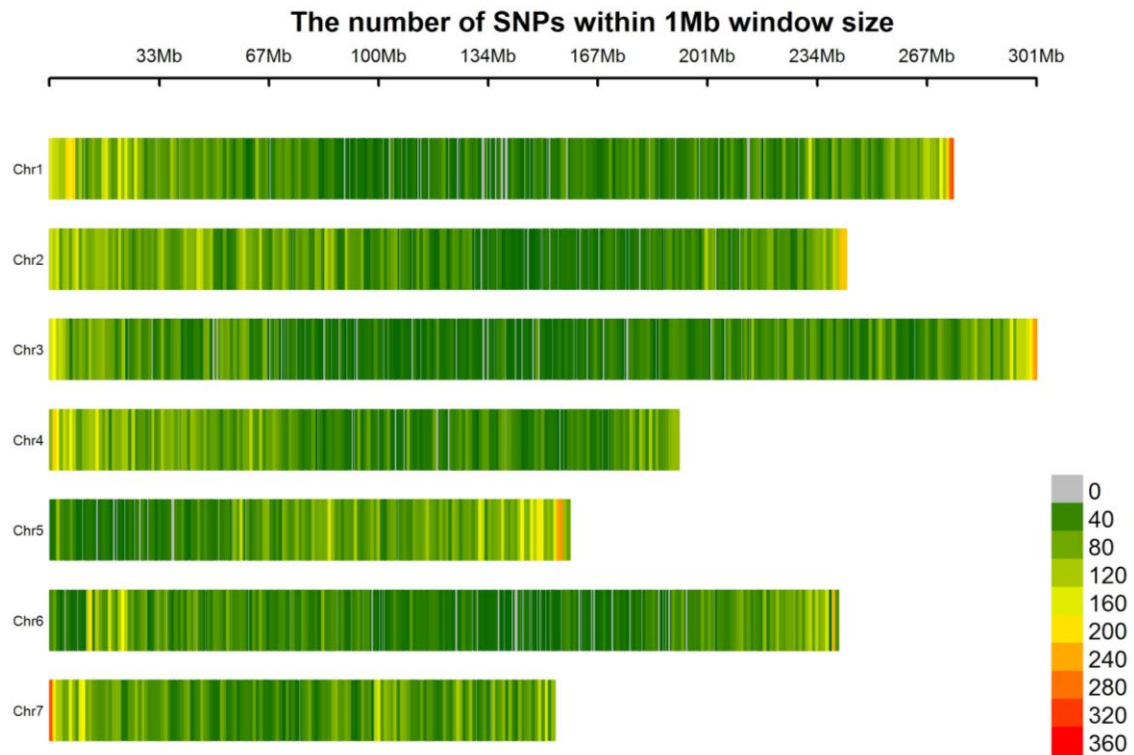


Figure 3

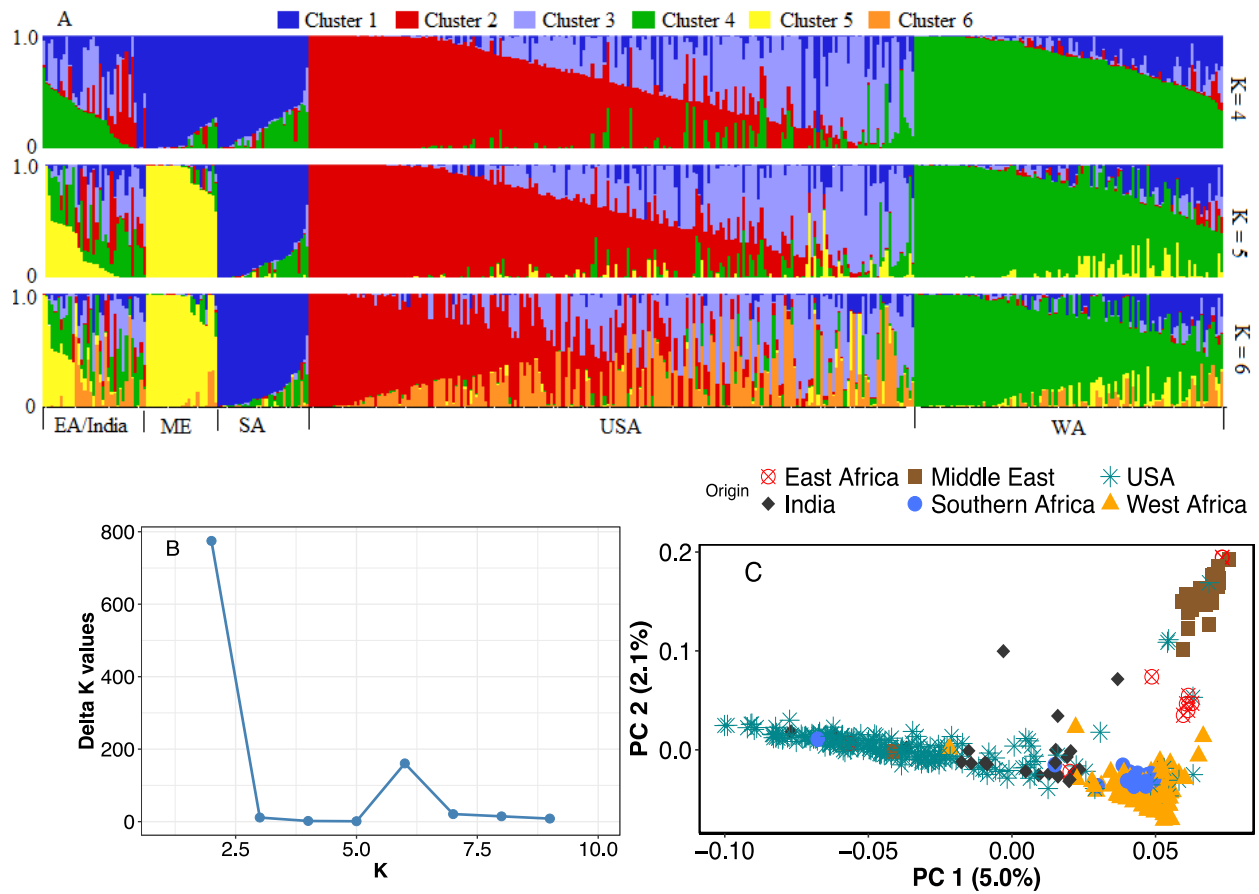


Figure 4

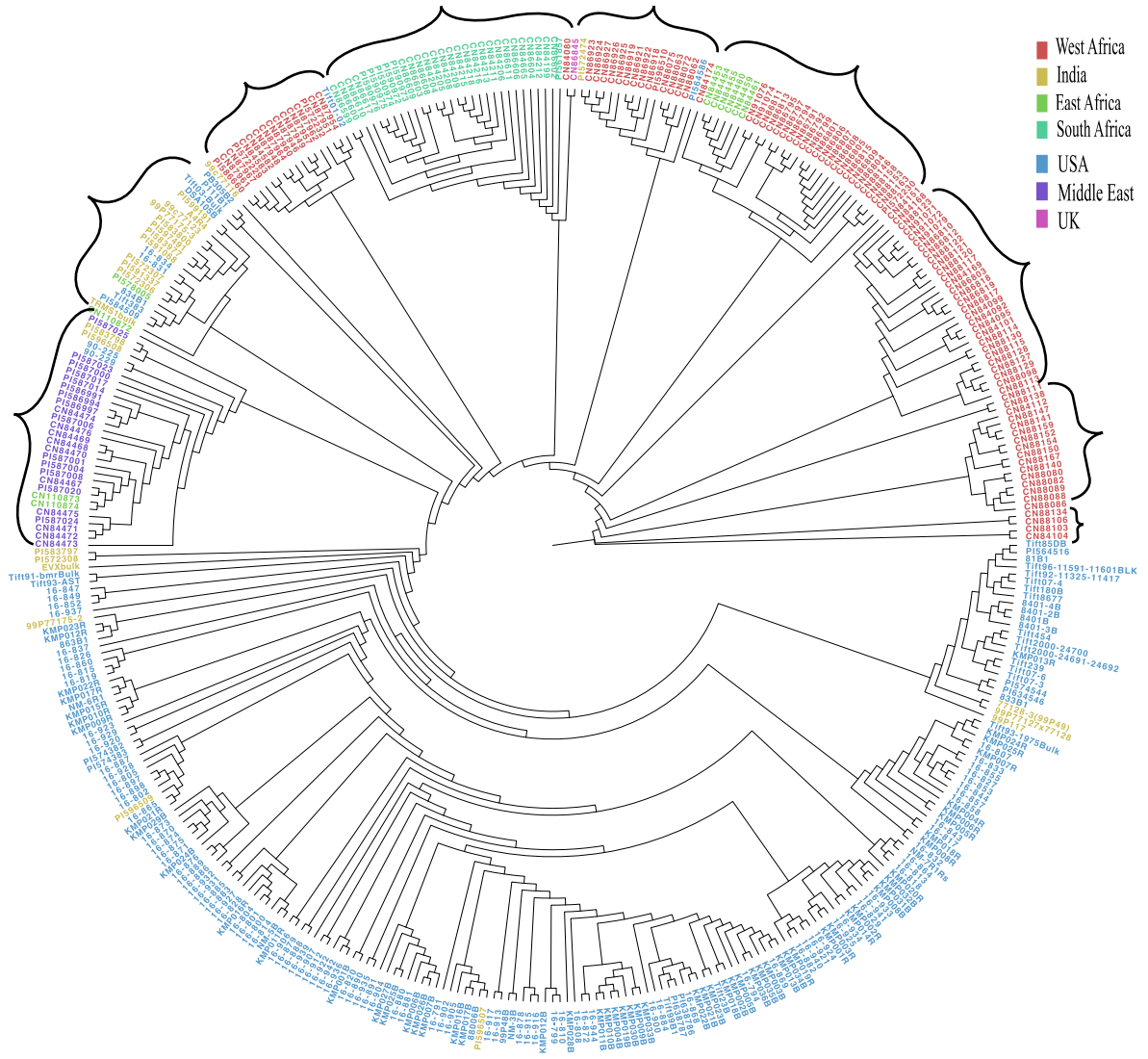


Figure 5

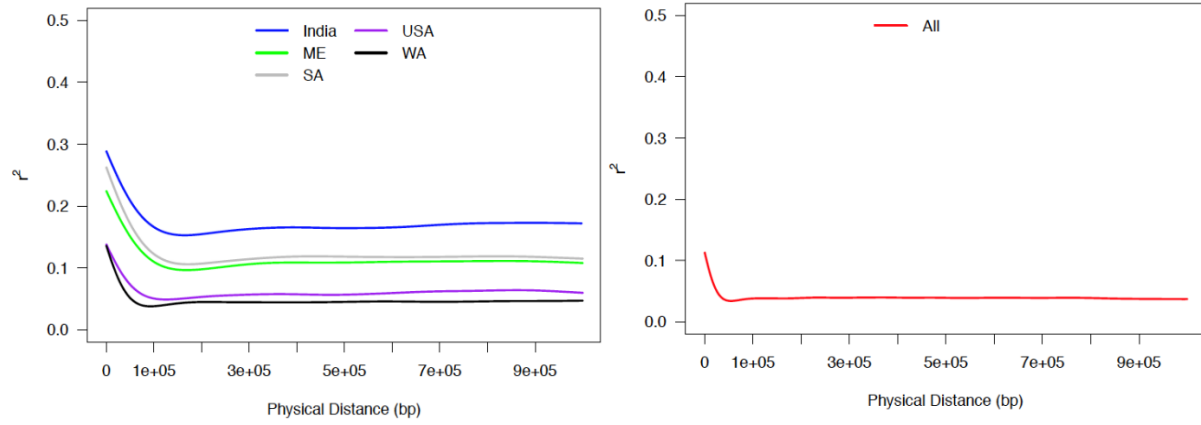


Figure 6

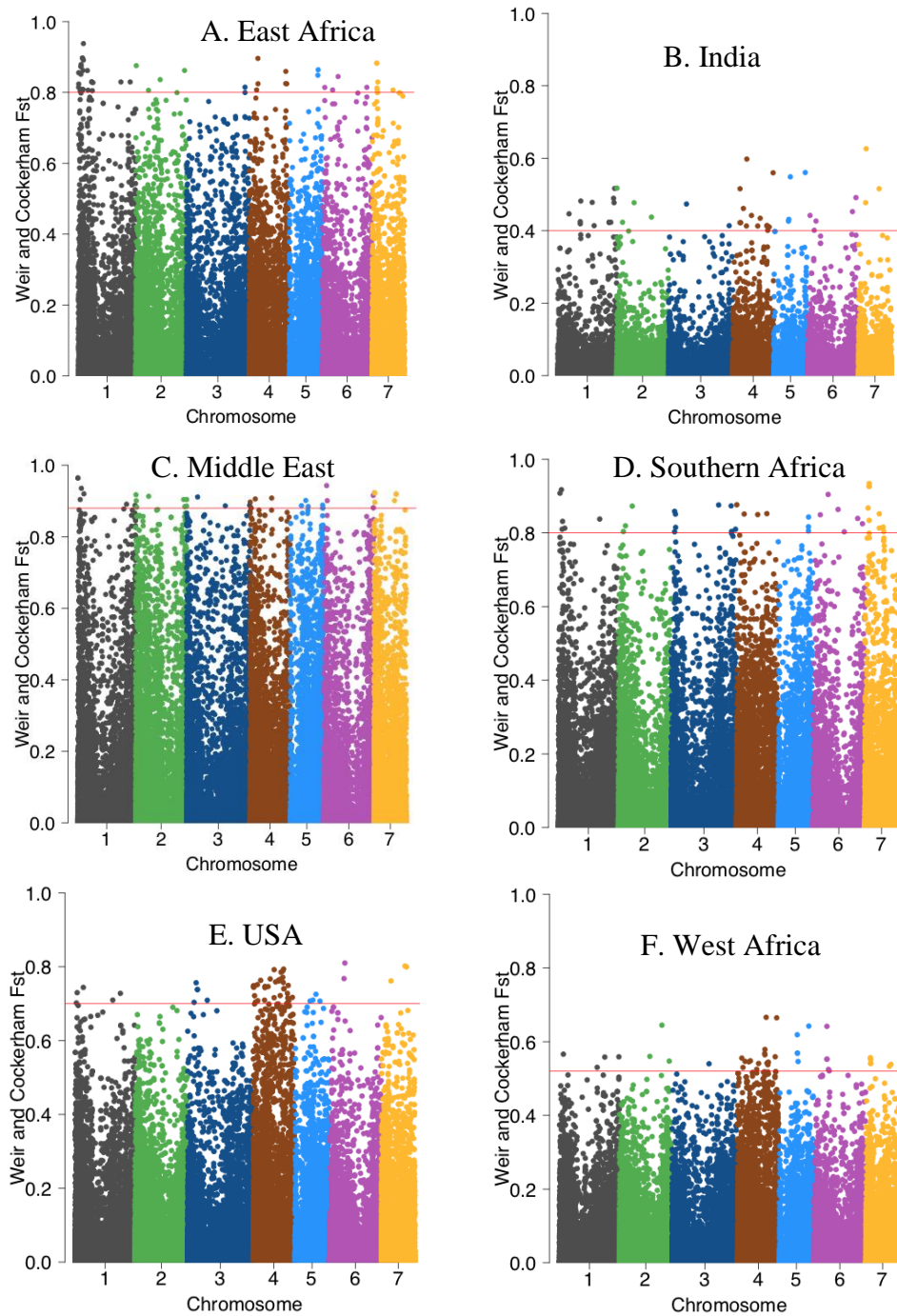


Table 1: Number of SNPs detected, and marker gaps from GBS of 398 pearl millet genotypes.

Chrom.	Chrom. Length (bp)	No. of markers on chrom	Average marker gap on chrom. bp)	Max marker gap on chrom. (bp)
1	275,468,192	38,710	7,116	1,251,556
2	242,893,347	36,854	6,591	626,660
3	300,905,882	35,477	8,482	867,928
4	191,808,916	28,053	6,837	669,411
5	158,669,458	27,191	5,835	865,327
6	240,561,232	29,395	8,184	992,253
7	154,007,176	26,573	5,796	582,530
U	515,119,487	35,714	14,423	1,849,934
Total/mean	2,079,433,690	257,967	63,264*	963,200*

Data prior to filtering for indels, minor allele frequency and missing; * refers to mean

Table 2: Pairwise F_{ST} statistics and genome-wide nucleotide diversity per 1 kbp.

Sub-population	F_{ST}	Nucleotide diversity/1kbp
East Africa	0.034595	5.12E-04
India	0.0063818	4.95E-04
Middle East	0.071837	5.12E-04
South Africa	0.047805	4.95E-04
United States	0.059203	4.23E-04
West Africa	0.043301	5.04E-04