Theor Appl Genet (2009) 120:13–23
DOI 10.1007/s00122-009-1155-6

ORIGINAL PAPER

Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection

Ming L. Wang · Chengsong Zhu · Noelle A. Barkley · Zhenbang Chen · John E. Erpelding · Seth C. Murray · Mitchell R. Tuinstra · Tesfaye Tesso · Gary A. Pederson · Jianming Yu

Received: 27 May 2009 / Accepted: 30 August 2009 / Published online: 16 September 2009
© Springer-Verlag 2009

Abstract Sweet sorghum has the potential to become a versatile feedstock for large-scale bioenergy production given its sugar from stem juice, cellulose/hemicellulose from stalks, and starch from grain. However, for researchers to maximize its feedstock potential a first step includes additional evaluations of the 2,180 accessions with varied origins in the US historic sweet sorghum collection. To assess genetic diversity of this collection for bioenergy breeding and population structure for association mapping, we selected 96 accessions and genotyped them with 95 simple sequence repeat markers. Subsequent genetic diversity and population structure analysis methods identified four subpopulations in this panel, which correlated well with the geographic locations where these accessions originated or were collected. Model comparisons for three quantitative traits revealed different levels of population structure effects on flowering time, plant height, and brix. Our results suggest that diverse germplasm accessions curated from different geographical regions should be considered for plant breeding programs to develop sweet sorghum cultivars or hybrids, and that this sweet sorghum panel can be further explored for association mapping.

Communicated by M. Frisch.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1155-6) contains supplementary material, which is available to authorized users.

M. L. Wang (✉) · N. A. Barkley · G. A. Pederson
USDA-ARS, PGRCU, Griffin, GA 30223, USA
e-mail: mingli.wang@ars.usda.gov

C. Zhu · T. Tesso · J. Yu (✉)
Department of Agronomy, Kansas State University,
Manhattan, KS 66506, USA
e-mail: jyu@ksu.edu

Z. Chen
Crop and Soil Sciences Department,
University of Georgia, Griffin, GA 30223, USA

J. E. Erpelding
USDA-ARS, Tropical Agriculture Research Station,
Mayaguez, PR 00680, USA

S. C. Murray
Department of Crop and Soil Sciences,
Texas A&M University, College Station, TX 77843, USA

M. R. Tuinstra
Department of Agronomy, Purdue University,
West Lafayette, IN 47907, USA

Introduction

Sorghum [*Sorghum bicolor* (L.) Moench] has three distinct features: high photosynthetic efficiency as a C_4 crop for converting solar energy to biomass; high water use efficiency for growing in high temperature and drought prone areas; and high tolerance to stressed environments and soil types for growing in poor and marginal lands (Sasaki and Antonio 2009). These characteristics make sorghum an excellent feedstock for bioenergy production (Reddy et al. 2005; Rooney et al. 2007; Vermerris et al. 2007). Cultivated sorghums can be divided into three main categories based on end product utilization: grain sorghum for starch, sweet sorghum for sugar, forage and energy sorghums for biomass. There are virtually no biological or taxonomic barriers/boundaries among these cultivated forms for hybridization, and they all belong to the same species:

S. bicolor (Ritter et al. 2007). Their morphological divergence may have arisen from natural and human selection and utilization. Out of these three types, sweet sorghum offers a unique opportunity to be exploited as a bioenergy crop.

Sweet sorghums (or sorgos) are typically characterized by low grain yields, but high biomass production. The tall, juicy stalks contain 10–25% sugars (mainly sucrose, glucose, and fructose) near the time of seed maturity. The first sweet sorghum grown in the US was Chinese Amber in 1853, and since then, additional sweet sorghum landraces were introduced from Africa, Australia, and other countries (Smith and Frederiksen 2000). Because of its high sugar content, sweet sorghum was cultivated mainly for fodder, syrup, molasses, sugar, and for small-scale ethanol production (Hunter and Anderson 1997). Less breeding effort has been devoted to sweet sorghum than to grain sorghum, whose production mainly relies on F_1 hybrids, whereas sweet sorghum production has traditionally been using pure-line cultivars. Historically, sweet sorghum cultivars were developed and released during the early to middle 1900s in the US. Subsequently, many sweet sorghum lines with high sugar content and disease resistance were developed by the USDA-ARS, Sugar Crops Field Station (Meridian, MS), Texas Agricultural Experiment Station (Weslaco, TX), Kansas Agricultural Experiment Station (Manhattan, KS), and Georgia Agricultural Experiment Station (Griffin, GA). Initially, the US sweet sorghum collection was maintained in the Sugar Crops Field Station until the program closed in 1983, whereupon the sweet sorghum collections ('MN') and sweet sorghum lines ('MER') were transferred to the USDA-ARS, Plant Genetic Resources Conservation Unit (PGRCU), Griffin, Georgia. Presently, 2,180 'MN' and 'MER' entries are located in GRIN database (Pederson and Spinks 2006), representing the US historic sweet sorghum collection (GRIN 2008).

Sweet sorghum has advantages in bioethanol production processing because it requires fewer chemical reaction steps and less energy from primary feedstock to the end product than grain and forage sorghums. Furthermore, the cost to cultivate sweet sorghum can be as little as three times lower than that of sugarcane (Reddy et al. 2005). Plant breeding and genetics are expected to play a major role in enhancing the potential of bioenergy crops such as sweet sorghum and more studies assessing the genetic diversity of sweet sorghum for plant improvement on this front are now available. For example, for determining the genetic relationship between sweet sorghum and grain sorghum, 31 sweet sorghum lines from the Australia's Commonwealth Scientific and Industrial Research Organization (CSIRO) collection and 64 grain sorghum lines were genotyped by amplified fragment length polymorphism (AFLP) markers (Ritter et al. 2007). Results showed that sweet sorghum lines could not be separated from grain sorghum lines; rather, they clustered with a similar racial origin, suggesting that sweet sorghum is of polyphyletic origin within *S. bicolor* ssp. *bicolor*. In a second study, sixty-eight sweet sorghum and four grain sorghum cultivars from the US germplasm collection were genotyped with 41 SSR markers (Ali et al. 2008). Similar to the CSIRO study, sweet sorghum cultivars did not form a separate cluster from grain sorghum cultivars, instead ten distinct clusters containing both types were observed. Recently, 125 sweet sorghum lines were genotyped with 47 SSR markers plus 322 single nucleotide polymorphism (SNP) markers (Murray et al. 2009). The assayed accessions were classified into three groups: historic and modern syrup types, amber types, and modern sugar/energy types. Given different germplasm and different sets of markers were used in these studies, it is understandable that no consistent grouping pattern was revealed. Further study with a global sample of sweet sorghum accessions and adequate markers is required.

The genetic structure and diversity of sweet sorghum accessions in the US historic collection have not been evaluated. Only a few accessions in this historic collection have been used in the previous studies (Ritter et al. 2007; Ali et al. 2008; Murray et al. 2009) and also only a limited number of SSR markers (<50) were employed. The objectives of this study were to determine the genetic diversity of a subset of accessions from the US historic sweet sorghum collection and to assess population structure to determine the potential of this panel for association mapping in subsequent studies.

Materials and methods

Plant germplasm and phenotyping

From the 2,180 accessions in the US historic sweet sorghum collection, 96 accessions (4.4%) were randomly selected. Most accessions had "MN" or "MER" designations except four accessions (PI 563295, PI 586541, PI 533998, and PI 535792). Their identifiers and collection sites/origins are listed in Table 1. Seeds of the accessions were obtained from the USDA-ARS, PGRCU and planted non-replicated evaluations in Isabela, Puerto Rico in 2007 and Manhattan, Kansas in 2007 and 2008. Flowering time, plant height, and brix data were collected in Manhattan, Kansas in 2007 and 2008. The trait means were used to assess the potential of association mapping within this sweet sorghum panel. Flowering time was recorded as the interval after planting to the time when 50% of the plants in a row had pollen shedding on 50% of the spikelets of the panicle; plant height was the distance from the ground to

Table 1 The 96 accessions from the US historic sweet sorghum collection used in the study and their subpopulation membership coefficients

No.	PI No.	Identifier[a]	Collection site/origin	G1	G2	G3	G4
1	PI 563295[b,c]	IS 9606, Rio	United States, Maryland	0.013	0.018	0.620	0.350
2	PI 145633	MN 425, Tugela ferry	South Africa	0.896	0.029	0.056	0.019
3	PI 273465	MN 4561, Jerima	Nigeria	0.033	0.156	0.784	0.027
4	PI 266917	MN 4514, Minnesota amber	United States	0.524	0.017	0.066	0.394
5	PI 52606[d]	MN 2680, African kafir	South Africa, Transvaal	0.925	0.038	0.015	0.022
6	PI 152727	MN 855, Malwal Tonj	Kenya	0.034	0.027	0.914	0.025
7	PI 643008	MN 2751, Grif 16016	United States, Mississippi	0.036	0.035	0.832	0.097
8	PI 157035	MN 2501, Nyagwang No. 56	Kenya	0.175	0.191	0.343	0.290
9	PI 156018	MN 1615, E-22	Zaire	0.169	0.091	0.631	0.109
10	PI 586541[b,c,d]	BJ-15, Tracy	Australia	0.705	0.021	0.260	0.014
11	PI 152971[d]	MN 1060, Awanlek	Sudan	0.026	0.368	0.110	0.496
12	PI 154846	MN 1502, Kabiri	Uganda	0.302	0.055	0.570	0.074
13	PI 641862	MN 715, Collier	United States, Mississippi	0.782	0.022	0.176	0.021
14	PI 273969	MN 4578, 2016	Ethiopia	0.034	0.521	0.424	0.021
15	PI 180348	MN 2917, Juar	India	0.073	0.810	0.100	0.017
16	PI 179749	MN 2906, Juar	India	0.013	0.791	0.052	0.144
17	PI 152963	MN 1052, Thok (B)	Sudan	0.022	0.048	0.803	0.127
18	PI 154988	MN 1470, S.A.2	Swaziland	0.736	0.116	0.136	0.012
19	PI 154987	MN 1469, S.A.1	Swaziland	0.325	0.053	0.060	0.562
20	PI 145626	MN 417, Manyoble	South Africa	0.876	0.032	0.077	0.015
21	PI 253986	MN 4138, K1137	Syria	0.037	0.637	0.312	0.015
22	PI 257600	MN 4487, No. 6 Gambela	Ethiopia	0.032	0.031	0.303	0.635
23	PI 181080[b,d]	MN 2931, Honey sorghum	India	0.015	0.825	0.139	0.020
24	PI 653411[b,c,d]	MER 71-1, M 81E	United States, Mississippi	0.450	0.148	0.031	0.371
25	PI 156463	MN 2342, Dobbs	Tanzania	0.015	0.031	0.161	0.793
26	PI 157030	MN 2496, Andiwo III 57	Kenya	0.102	0.018	0.669	0.212
27	PI 152733	MN 861, Merrisa (BarI)	Sudan	0.029	0.015	0.411	0.544
28	PI 641835	MN 23, Rex	United States, Mississippi	0.963	0.010	0.012	0.014
29	PI 641810	MN 46, Colman (Y)	United States, Mississippi	0.779	0.199	0.013	0.009
30	PI 181899	MN 2938, Aleppo No. 41	Syria	0.156	0.807	0.021	0.015
31	PI 641821	MN 313, Honey drip	United States, Mississippi	0.723	0.058	0.151	0.068
32	PI 152816	MN 934, Wad fur white	Sudan	0.214	0.023	0.061	0.701
33	PI 22913[c,d]	MN 235, Chinese amber	China	0.066	0.069	0.783	0.082
34	PI 218112	MN 3382, IS 2352	Pakistan	0.025	0.612	0.334	0.029
35	PI 175919	MN 2870, IS 12833	Turkey	0.858	0.014	0.116	0.011
36	PI 273955	MN 4566, 1883	Ethiopia	0.020	0.031	0.932	0.017
37	PI 255239	MN 4458, Caxa	Mexico, Sonora	0.165	0.141	0.656	0.037
38	PI 266927	MN 4512, Co. 1	India	0.017	0.503	0.470	0.010
39	PI 641817[b,c]	MN 10, Early sumac	United States, Mississippi	0.747	0.026	0.204	0.023
40	PI 154750	MN 1306, Serere	Uganda	0.071	0.028	0.261	0.639
41	PI 302252	MN 4299, A-6026	China	0.618	0.158	0.211	0.013
42	PI 152860	MN 963, Merasi	Sudan	0.025	0.008	0.376	0.592
43	PI 511355[c,d]	MER 81-2, Smith	United States, Texas	0.375	0.083	0.106	0.436
44	PI 152961	MN 1050, Malnal	Sudan	0.030	0.010	0.293	0.667
45	PI 643016	MN 2761, Grif 16024	United States, Mississippi	0.010	0.053	0.926	0.011
46	PI 157804	MN 2513, Feterita abu derega	Sudan	0.047	0.178	0.538	0.237
47	PI 152828	MN 1133, U.T. 23	Zaire	0.487	0.175	0.317	0.020
48	PI 156871	MN 2443, Tutobo	Zaire	0.143	0.013	0.818	0.027
49	PI 170787	MN 2826, 1716	Turkey	0.934	0.006	0.047	0.013
50	PI 155149	MN 1581, Dhurra No. 7	Yemen	0.024	0.322	0.633	0.022
51	PI 641834	MN 22, Planter	United States, Mississippi	0.577	0.040	0.352	0.031

Table 1 continued

No.	PI No.	Identifier[a]	Collection site/origin	G1	G2	G3	G4
52	PI 156178	MN 2014, 1159	Malawi	0.370	0.038	0.558	0.034
53	PI 155885	MN 1644, IS 12598	Tanzania	0.020	0.012	0.351	0.617
54	PI 155760	MN 1973, Namuse	Malawi	0.518	0.013	0.450	0.020
55	PI 196583	MN 3080, No. 1	China, Taiwan	0.936	0.022	0.025	0.017
56	PI 641807	MN 3, Atlas	United States, Mississippi	0.911	0.012	0.066	0.011
57	PI 251672	MN 4135, 1035	Yugoslavia	0.625	0.016	0.024	0.335
58	PI 156268	MN 2155, Chedomaba	Malawi	0.044	0.291	0.624	0.040
59	PI 154944	MN 1426, L31 Emiroit	Uganda	0.077	0.045	0.054	0.824
60	PI 176766	MN 2873, 9387	Turkey	0.010	0.101	0.872	0.016
61	PI 152751	MN 880, Nytwal	Sudan	0.039	0.019	0.103	0.839
62	PI 156393	MN 2277, 1427	Tanzania	0.019	0.021	0.839	0.122
63	PI 92270	MN 2740, 12748	China, Beijing	0.023	0.435	0.518	0.023
64	PI 181083	MN 2934, Kamandri	India	0.041	0.701	0.228	0.030
65	PI 144134	MN 395, Inyangentombi	South Africa, Natal	0.680	0.055	0.091	0.174
66	PI 196049	MN 3053, IS 2131	Ethiopia	0.031	0.063	0.888	0.018
67	PI 651497[c]	MER-67-10, Theis	United States, Mississippi	0.400	0.006	0.215	0.379
68	PI 155516	MN 1705, Maska	Zambia	0.025	0.014	0.406	0.555
69	PI 157033	MN 2499, Ifube No. 18	Kenya	0.014	0.018	0.097	0.871
70	PI 156487	MN 2363, 1519	Tanzania	0.205	0.021	0.306	0.468
71	PI 533998[b,c,d]	NSL 51352, Brawley	United States, Texas	0.820	0.010	0.145	0.024
72	PI 217691	MN 4534, Nagad el mur	Sudan	0.026	0.046	0.399	0.529
73	PI 651495[b,c,d]	MER 64-12, Dale	United States, Mississippi	0.327	0.007	0.650	0.016
74	PI 586443	MN 818, IS 27818	Hungary	0.028	0.015	0.547	0.411
75	PI 196598	MN 3095, No. 18	China, Taiwan	0.822	0.109	0.052	0.017
76	PI 156217	MN 2109, 1250	Malawi	0.291	0.013	0.335	0.361
77	PI 303658	MN 4607, Nerum boer	Sudan, Southern	0.087	0.013	0.309	0.592
78	PI 641909	MN 1170, Red losinga	Sudan	0.018	0.029	0.049	0.904
79	PI 180487	MN 2921, Juar	India	0.018	0.541	0.318	0.123
80	PI 257599	MN 4486, No. 5 Gambela	Ethiopia	0.010	0.011	0.663	0.316
81	PI 152966	MN 1055, Ayuak	Sudan	0.015	0.016	0.243	0.726
82	PI 156203	MN 2089, 1232	Malawi	0.406	0.013	0.545	0.035
83	PI 156890	MN 2462, Dura huria	Zaire	0.373	0.038	0.570	0.019
84	PI 535792	N 107	United States, Nebraska	0.641	0.019	0.318	0.022
85	PI 583832[b,c,d]	MER 76-6, Top 76-6	United States, Georgia	0.206	0.013	0.029	0.752
86	PI 641815	MN 9, Early folger	United States, Mississippi	0.894	0.040	0.023	0.044
87	PI 152771	MN 889, Rahmetalla gallaba	Sudan	0.041	0.254	0.262	0.443
88	PI 147224	MN 519, B. 35	India	0.013	0.497	0.478	0.012
89	Grif 16306[e]	MN 4017, Greenleaf Sudan 9	Sudan	0.087	0.085	0.811	0.017
90	PI 651493[b,c,d]	MER 65-2, Ramada	United States, Mississippi	0.130	0.033	0.093	0.745
91	PI 653616[b,c]	MER 69-13, Wray	United States, Mississippi	0.407	0.018	0.273	0.303
92	PI 653617[b,c]	MER 68-2, Keller	United States, Mississippi	0.405	0.030	0.297	0.269
93	PI 152923	MN 1012, Duro el jack	Sudan, Kordofan	0.274	0.028	0.364	0.335
94	PI 145632	MN 424, Tegevini	South Africa	0.899	0.010	0.076	0.014
95	PI 154844[b,c]	MN 1500, Grassl	Uganda	0.041	0.010	0.072	0.877
96	PI 653737[f]	MN 4254, A-5740	United States, Oklahoma	0.202	0.053	0.706	0.039

[a] The "MN" and "MER" designations are listed along with the plant name or other collection identifiers

[b] These 12 accessions were also used in the study of Ritter et al. 2007

[c] These 14 accessions were also used in the study of Ali et al. 2008

[d] These 11 accessions were also used in the study of Murray et al. 2009

[e] *Sorghum x drummondii*

[f] *Sorghum propinquum* (Kunth) Hitchc.

the tip of the panicle; and brix (a measure of soluble solids) was measured 3–4 weeks after flowering with a handheld refractometer at three nodes (4, 7, and 9) of a typical plant in a row, and the mean was recorded for the row. Eighteen accessions did not flower; nonetheless, days to flowering, the length of the season for these accessions, and measured brix was recorded with the last accessions to flower.

Genotyping with SSRs

Fresh leaf tissue was collected and lyophilized from a single plant which represents most of the plants within an accession morphologically. DNA was extracted by using the E.Z.N.A. Plant DNA kit from Omega Bio-Tek (Doraville, GA). DNA concentration was determined by measurement with a fluorimeter. All DNA samples were dissolved and diluted in 0.1X TE (1 mM Tris, 0.1 mM EDTA, pH 8.0) to a final concentration of 10 ng/μl for use in PCR. Simple sequence repeat (SSR) markers were mostly from published literatures (Kong et al. 2000; Schloss et al. 2002; Menz et al. 2002) with a few newly developed markers from International Crops Research Institute for the Semi-Arid Tropics (ICRISAT, T. Hash, personal comm.). Out of 120 tested SSRs, 95 (79%) showed polymorphisms across a screening panel of eight sweet sorghum accessions. The names for the 95 polymorphic DNA markers are listed in Supplementary Table 1, together with their amplicon sizes, chromosome locations, and sequence positions obtained from the published sorghum genome sequence (Paterson et al. 2009). Different PCR amplification conditions and separation methods were used based on amplicon size. For 50 SSRs with an amplicon size greater than 180 bp, non-labeled primers were used following the method described by Wang et al. (2006b) for PCR reaction mixture and amplification cycling programs. Amplicons were separated on a 5% acrylamide/bis non-denature gel with Mega-gel system (C-DASG-400-50, CBS Scientific Co., Del Mar, CA) under the condition of 300 V for 1 h. Separated amplicons were stained with ethidium bromide, visualized by UV light and sized by referencing with a 50 bp DNA ladder (Invitrogen, Carlsbad, CA) on the same gel. For 45 SSRs with an amplicon size smaller than 180 bp, fluorescently labeled forward primers and non-labeled reverse primers were used for the PCR reaction mixture and amplification cycling programs following the method described by Wang et al. (2006a) with the separation of the amplicons performed on capillary gel system (Beckman CEQ 8000, Fullerton, CA). After the calling of the amplicon size, the scored data were inspected to remove minor non-specific DNA bands from mega-gel images and extra stutter peaks from CEQ chromatograms.

Statistical analysis

DNA marker profile

Alleles produced from each accession were scored based on size comparison to a molecular weight ladder. All automated scoring from the Beckman software package, Fragments, were evaluated for accurate automatic allele calling. PowerMarker version 3.25 (Liu and Muse 2005) was used to calculate Chord distance (Cavalli-Sforza and Edwards 1967) among accessions, to compute other molecular diversity statistics, and to construct the consensus neighbor joining (NJ) tree with 100 runs of bootstrapping.

Population structure analysis

The program STRUCTURE version 2.2 (Falush et al. 2003; Falush et al. 2007; Pritchard et al. 2000) was used to detect population structure and to assign individuals to subpopulations. This program employs model-based clustering in which a Bayesian approach identifies clusters based on a fit to the Hardy–Weinberg equilibrium and linkage equilibrium. The STRUCTURE program was run ten times for each number of subpopulation (*k)* value, ranging from 1 to 10, using the admixture model with 20,000 replicates for burn-in and 5,000 replicates during analysis. The final population subgroups were determined based on (1) likelihood plot of these models; (2) stability of grouping patterns across ten runs; and (3) germplasm information about the materials under study. Based on this information, we chose $k = 4$ as the optimal grouping. Out of the ten runs for $k = 4$, the run with the highest likelihood value was selected to assign the posterior membership coefficients (Q) to each accession. A graphical bar plot was then generated with the posterior membership coefficients (Fig. 1), and plots were also generated for $k = 2$ and 3 for result interpretation (Supplementary Fig. 1).

An analysis of molecular variance (AMOVA) was performed using Arlequin 3.11 (Excoffier et al. 2005) to evaluate population differentiation among the four subpopulations. Furthermore, the genetic distances among these four subgroups were calculated as Nei's minimum distance and pairwise F_{st}. To validate the genetic structure and test for different models, principal component analysis (PCA) and nonmetric multidimensional scaling (nMDS) were conducted to construct plots of the most significant axes for grouping pattern verification and to obtain axes for further model testing and association mapping (Patterson et al. 2006; Price et al. 2007; Zhu and Yu 2009). While PCA decomposes the overall variation among accessions measured by molecular markers into orthogonal axes, nMDS transforms data into Euclidean distances and generates a configuration to represent the structure contained in the original data (Zhu and Yu 2009). Kinship (K) was

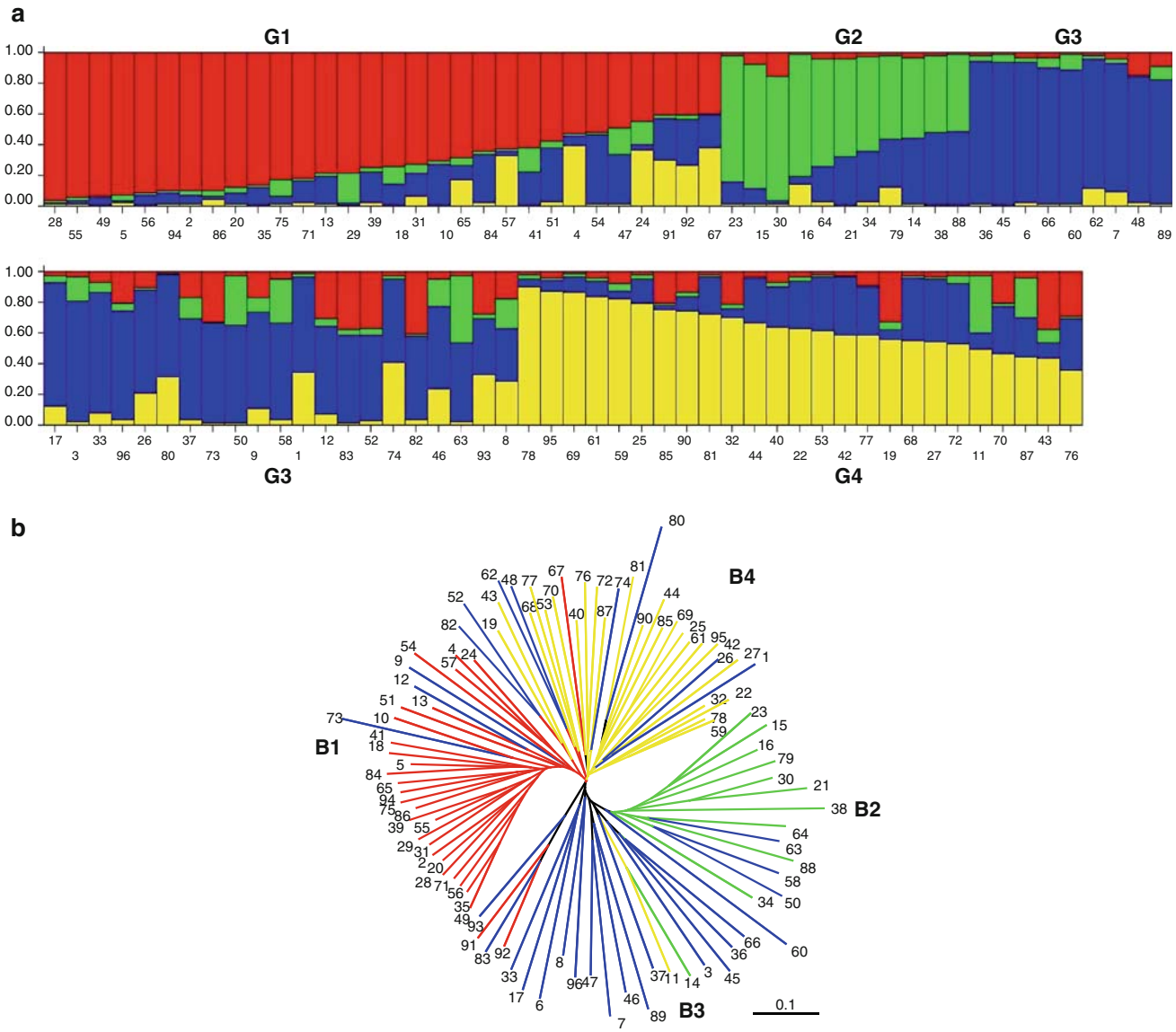

Fig. 1 Genetic diversity of sweet sorghum accessions. **a** Population structure analysis. *Numbers* on the *y*-axis show the subgroup membership, and the numbers on the *x*-axis show the accession number. **b** Neighbor-joining tree analysis. The numbers at the tip of the tree branch indicate the accession number. The *colors of the bar* and the *tree branch* indicate the four groups identified through the STRUCTURE program (G1 = *red*, G2 = *green*, G3 = *blue*, and G4 = *yellow*)

calculated with SPAGeDi (Loiselle et al. 1995; Hardy and Vekemans 2002). The combined display of the color coded subpopulation memberships from STRUCTURE with other analyses (i.e., NJ, PCA, and nMDS) and the world map followed that of a recent genetic diversity study in common bean (Kwak and Gepts 2009).

Model comparisons

To assess the effect of population structure on association mapping of various quantitative traits measured in the sweet sorghum panel, model comparisons using the 95 SSR markers were conducted. These markers were not expected to have significant effects on the quantitative traits because of their random selection and the low genome coverage. Following the previously recommended procedures (Yu et al. 2006; Zhu and Yu 2009), various mixed models with subpopulation membership percentage, PCA, and nMDS as fixed covariates and with kinship as a random effect were tested. The dimension of PCA and nMDS was determined for each trait individually (Zhu and Yu 2009). Among all possible models, namely the simple model, Q, PCA, nMDS, Q + K, PCA + K, and nMDS + K, the best fit model was determined for each trait based on the Bayesian information criterion (BIC). The selected model was then used to test individual SSR markers, and the quantile–

quantile plots of the *F* test statistics for the SSR markers were plotted to assess the control of type I errors.

Results

Profile of SSR markers

For the 95 SSR markers genotyped across the 96 sweet sorghum accessions, 705 alleles were detected with an average of 7.6 alleles per locus and an average major allele frequency of 0.55. The mean polymorphism information content (PIC) was 0.54, and the mean genetic diversity was 0.58 (Table 2). About one-third of the markers produced at least one taxon specific allele. As expected, PI 653737 produced the greatest number of taxon specific alleles since this accession is classified as *Sorghum propinquum* (Kunth) Hitchc.

Population structure and genetic diversity

STRUCTURE analysis can help to identify clusters of genetically similar accessions. Thus, the subpopulations from STRUCTURE analysis numbered four groups (G): G1, G2, G3, and G4 (Fig. 1). G1 contained mostly US historic syrup type sweet sorghum accessions; G2 lines were primarily race durra accessions and from Asia; and G4 lines were mostly landraces from Africa and a few modern US lines developed for sugar and energy production (Table 1). G3 accessions had mixed origins and formed a loose group since the other three groups clustered, which was evident when the results from different numbers of subpopulations ($k = 2$, 3, and 4) were compared (Supplementary Fig. 1). As a result, the level of genetic diversity within G3 (0.61) was higher than for G1 (0.46), G2 (0.46), and G3 (0.49) (Table 2). G3 also contained the most alleles per locus and had the highest total number of alleles of the four identified subpopulations. For the AMOVA analysis, substantially more genetic variation within subpopulations (86%) was observed than between subpopulations (14%). Additionally, the genetic distances among these four subgroups measured by Nei's minimum distance and pairwise F_{st} were consistent, the genetic distance between G1 and G2 being the largest and between G3 and G4 being the smallest (Table 3).

Neighbor-joining analysis clustered 96 accessions into four branches (B), which was generally consistent with the results from STRUCTURE analysis with a few exceptions (Fig. 1). First, B1 contained 29 accessions with 26 accessions from subpopulation G1 and three accessions from subpopulation G3. All accessions from South Africa, 13 of 22 accessions from the US and three of five accessions from China were in this branch. Second, B2 contained 14 accessions, including ten accessions from subpopulation G2 and four accessions from subpopulation G3. All seven accessions from India and the two Syrian accessions were clustered in this branch. Third, B3 contained 20 accessions that corresponded primarily with subpopulation G3. Within B3, 11 accessions (>50%) were from central East Africa (Sudan, Ethiopia, and Kenya) and West Africa (Zaire and Nigeria), and included five accessions from the US that may have originated or was developed from parental lines from Africa. Finally, B4 contained 33 accessions and was similar to subpopulation G4. The dominant component of this branch was accessions from east Africa (Zambia, Tanzania, Uganda, Malawi, Ethiopia, and Sudan). Among these four branches, B3 was more diverged than the other branches.

Genetic clustering and geographic origin

Plotted on the first two dimensions of nMDS, the G1, G2, and G4 subpopulations detected by population structure analysis were adequately separated from each other but the G3 group was more dispersed (Fig. 2). When the sample position on the two dimensional nMDS plot was linked with the geographic origin on the world map, a clearer pattern emerged. The samples originating or collected from close geographic regions were generally clustered by both population structure and nMDS methods (Fig. 2). Also, the results of PCA were similar (Supplementary Fig. 2). Overall, the general pattern in genetic relationship was

Table 2 Summary statistics for the whole group of sweet sorghum accessions and subpopulations detected by structure analysis based on 95 SSR markers

Statistics	Overall	G1	G2	G3	G4
Sample size	96	30	11	30	25
Total number of alleles	705	400	302	566	414
Number of alleles per locus	7.61	4.17	3.15	5.89	4.31
Major allele frequency	0.55	0.66	0.64	0.52	0.62
Genetic diversity	0.58	0.46	0.46	0.61	0.49
PIC[a]	0.54	0.42	0.41	0.57	0.46

[a] Polymorphism information content

Table 3 Genetic distance between sweet sorghum groups from structure analysis

Group	G1	G2	G3	G4
G1	0	0.24	0.11	0.16
G2	0.32	0	0.13	0.22
G3	0.12	0.14	0	0.08
G4	0.20	0.28	0.08	0

The top diagonal is Nei's minimum distance and the bottom diagonal is pairwise F_{st}

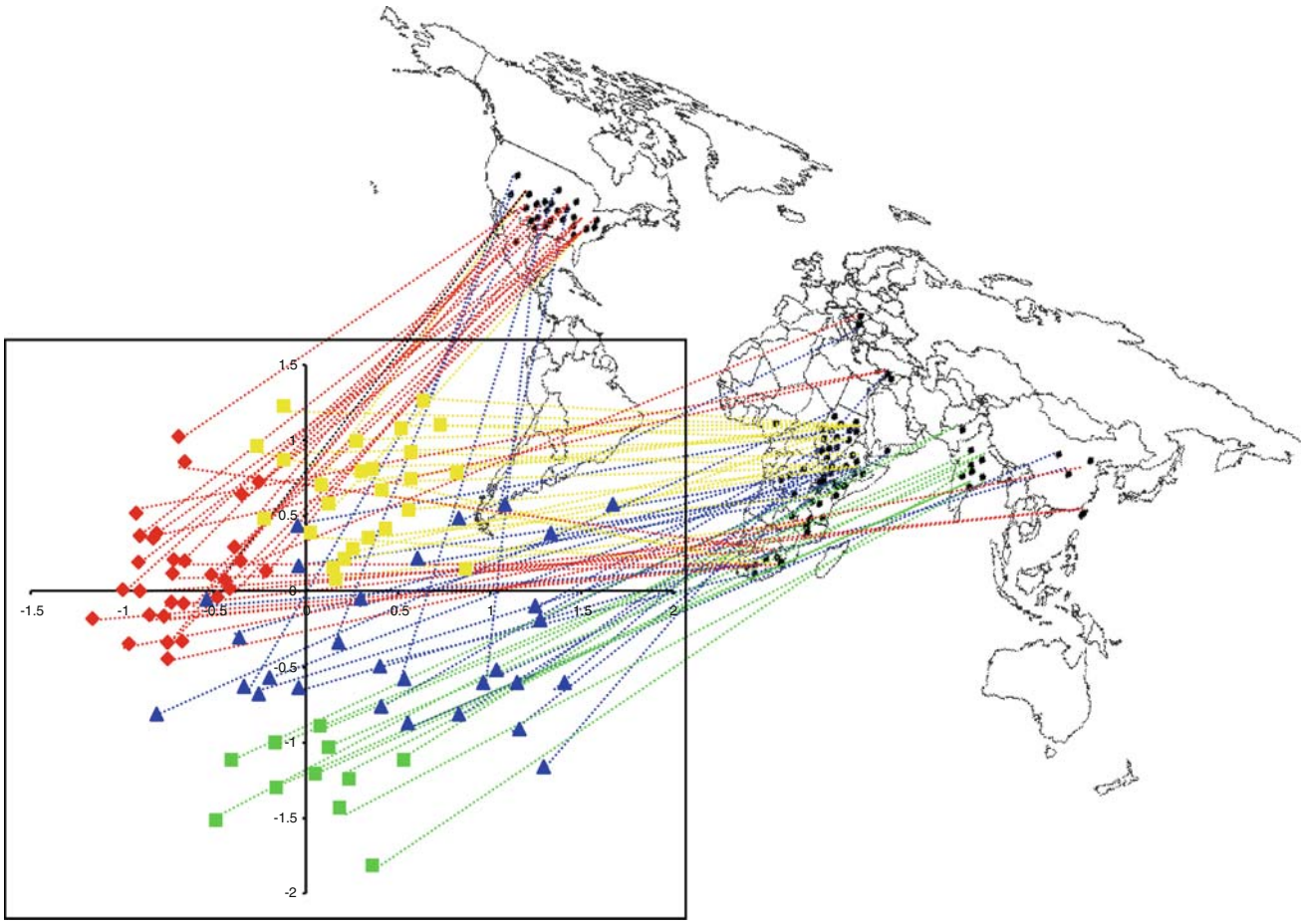

Fig. 2 Combined display of geographical distribution and genetic analysis of 96 sweet sorghum accessions with nMDS. The *lower left plot* shows the results of nMDS with the horizontal axis being the first dimension and the vertical axis being the second dimension. The *dotted lines* link the positions of accessions in the nMDS plot and their geographical origins on the world map. The *colors* indicate population membership identified by population structure analysis

consistent among various methods and further supported by genetic distances among subpopulations and the geographic distances among these accessions. However, the genetic divergence among these subpopulations was generally weak as can be seen from the NJ tree, nMDS plot, genetic distance measurements (Nei's minimum distance and F_{st}), and the change in likelihood values from STRUCTURE analysis. In retrospect, the random sampling selection process prevented the inclusion of accessions with high genetic relatedness that would lead to obvious close clusters.

Model comparison and association sample type

After completing the genetic analysis, the potential of this sweet sorghum collection as an association mapping panel was examined by testing these random SSR markers against phenotypes. Dimension determination for nMDS indicated that different dimensions should be included for testing associations for the three quantitative traits: flowering time (4), plant height (2), and brix (2). Further model comparisons determined that the mixed model with population structure and kinship had the best fit for flowering time, but the mixed model with only kinship had the best fit for plant height and brix. Accordingly, we tested selected models using individual SSR markers. Quantile–quantile plots of *F*-statistics verified the adequate control of false positives for the mixed model because the deviation of the observed *F*-statistics for these 95 SSR markers from the expected value was minimal (Fig. 3). Further, no highly significant SSR marker associated with any of the three quantitative traits was identified. Although molecular marker analyses revealed weak population differentiation among these accessions, the combined assessment of these analyses and the model testing with phenotypes demonstrated this sweet sorghum panel can be classified as a type II association sample with a low level of relatedness (Zhu and Yu 2009).

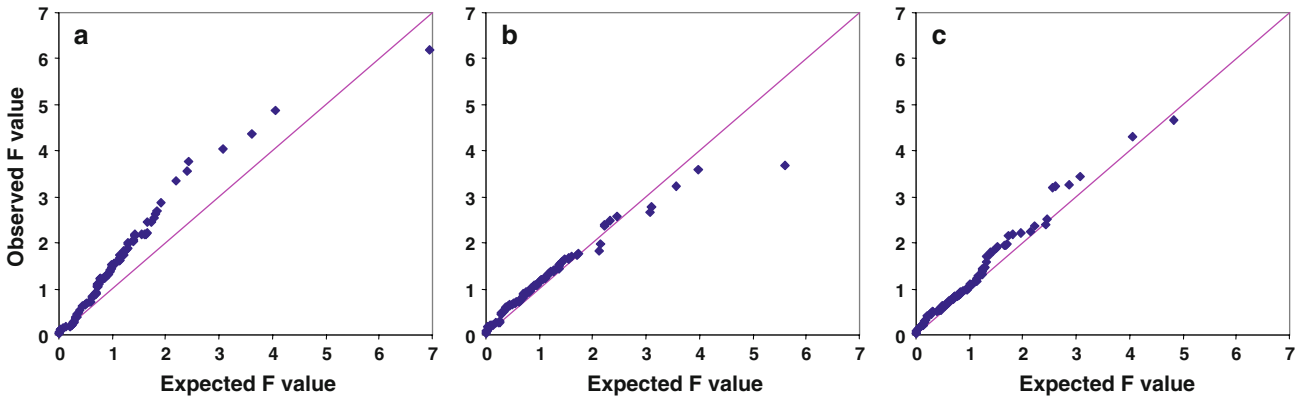

Fig. 3 Quantile-quantile plots of *F* test statistics with 95 SSR markers for three quantitative traits showing the adequate control of a type I error with the mixed model for **a** flowering time, **b** plant height, and **c** brix

Discussion

SSR genotyping

At the start of this project, the sorghum genome sequence was not available. Therefore, some SSR markers were genetically mapped, but their physical distances were not known. However, with the sorghum genome sequenced (Paterson et al. 2009; http://www.phytozome.net/sorghum, verified 25 May 2009), we found that primer sequences for three markers, *Xtxp 168*, *Gpsb 69*, and *Xtxp 250*, were located on more than one chromosome. In addition, the marker *Gpsb 69* revealed eighteen alleles. This made tracking the genotype scores challenging. We also found a high probability of multiple alleles being amplified from several loci instead of alleles from a single locus for this marker. Interestingly, the marker sequence of *Xtxp 40* was completely covered by *Xtxp 36*. It is recommended that the chromosome locations of the markers be checked prior to marker selection for molecular analysis.

The average number of alleles revealed per locus in this study was 7.6, lower than 10.4 from a previous study in grain sorghum (Wang et al. 2006a) and higher than 3.2 from another sweet sorghum study (Ali et al. 2008). This is possibly due to three factors: the polymorphism of SSR markers, the diversity of germplasm accessions, and the sensitivity of DNA fragment separation systems. Based on first comparison, it could also be speculated that the transition from grain sorghum to sweet sorghum, regardless the number of sweet origins, narrowed the genetic diversity.

Genetic diversity and population structure

SSR marker data have frequently been used as a tool to examine the dynamics of differentiation in population and enable specific analyses (Matsuoka et al. 2002; Liu et al. 2003; Barkley et al. 2006; Kwak and Gepts 2009). For example, structure analysis can estimate the number of subpopulations and the genetic relatedness among assessed accessions. Also, neighbor-joining tree analysis can cluster the assayed accessions into branches. Overall, the results obtained from these two separate analyses support each other (Liu et al. 2003; Kwak and Gepts 2009). The small discrepancy between groupings in this study is not unexpected. While neighbor joining tree analysis assigned a fixed branch position to each accession, population structure analysis resulted in a subpopulation membership percentage and the highest percentage was used to assign the group for easy interpretation. It should be pointed out that the one-dimension clustering methods in NJ would be difficult to capture adequate information if complicated genetic relationship exists among accessions. Accordingly, focusing on the overall tree pattern in result interpretation, rather than specific accessions, is recommended. Nevertheless, the general pattern of genetic diversity and population structure was further verified with the nMDS plot and the world map of origins.

Ideally, the reference to the world map of origins would be more valid if all accessions are landraces (Kwak and Gepts 2009). However, since the plant breeding work in sweet sorghum is relatively limited and many of those US accessions are not many generations, or cycles of breeding, away from their landrace counterparts, whose true origins would probably be clustered. Reasonable results were shown in the attempt to connect population structure and genetic diversity analysis with the world map of origins. Ultimately, all these analyses yielded four groups of sweet sorghum accessions. Furthermore, the branch separating pattern with the NJ tree and the spread of accessions in the nMDS plot agreed with the genetic distance measurements among subpopulations and the AMOVA analysis, indicating that this historic sweet sorghum collection is diverse and well suited for association mapping.

Sweet sorghum accessions

Unlike previous studies on sweet sorghum diversity (Ritter et al. 2007; Ali et al. 2008; Murray et al. 2009), more SSR markers (95) were used in this study and their physical locations on chromosome were also provided. This may facilitate sweet sorghum breeders to choose flanking SSR markers in breeding programs if the chromosome locations of their targeted trait are known. Furthermore, the accessions evaluated in this study (mainly focused on the US historic sweet sorghum collection) were different from the previous studies. Therefore, the conclusions drawn from this study would be specific to the US collection. However, comparing with this study, there were 12, 14, and 11 accessions in common from CSIRO, Nebraska, and Texas, respectively (see Table 1). Six accessions (PI 586541, PI 653411, PI 533998, PI 651495, PI 583832, and PI 651493) were examined in all studies. In all four studies, PI 583832 (Top 76-6) and PI 651493 (Ramada) were classified into the same group as the G4 and B4 (i.e., landraces from Africa and modern US cultivars for sugar and energy production) in the current study. Also, in all four studies, PI 586541 (Tracy) and PI 533998 (Brawley) were classified into the same group as G1/B1 (i.e., CSIRO's group 1, Nebraska's group I, and Texas' historical and modern syrup type). Next, the classification of PI651459 (Dale) by NJ in B1 was consistent with the three previous studies, but not with the classification in G3 by the population structure analysis. The classification of PI 653411 (M 81E), however, indicated a mixed origin of this accession. Ultimately, we found no consistent grouping pattern of this accession in any two studies.

Plant breeding for bioenergy

Sweet sorghum has emerged as an ideal feedstock for bioethanol production to exploit alternative bioenergy. Indeed, significant genetic potential exists in the sweet sorghum germplasm collection. The results of this study demonstrated that genotyping sweet sorghum accessions with DNA markers can quickly reveal the genetic structure and gauge the genetic diversity of each accession. Therefore, the information from the genetic structure and diversity can help breeders choose parents for developing sweet sorghum cultivars and hybrids. Clearly, germplasm panels are useful genetic resources for sweet sorghum improvement, and thus, research should continue focusing on African and southern Asian sorghum germplasm. Also, researchers should be aware that additional sweet sorghum accessions exist and are conserved in the US sorghum collection and at other centers, specifically, Institute of Crop Germplasm Research-Chinese Academy of Agricultural Sciences (ICGR-CAAS, Beijing, China) and ICRISAT (Patancheru, Andhra Pradesh, India). In the end, while the plant breeding effort in major crops has been successful with the field-based pedigree breeding method and recently enhanced by molecular technologies, our research indicates that breeding for bioenergy crops, however, could benefit from the results of molecular analysis at the start in germplasm selection and breeding scheme design.

Prospects for association mapping

So far, both linkage analysis with a biparental cross and association mapping with a sweet sorghum panel have been conducted in sweet sorghum and identified QTL have been reported (Ritter et al. 2008; Natoli et al. 2002; Murray et al. 2008a, b). As the sequencing and genotyping costs keep decreasing, association mapping will see a wider application in different crop species (Zhu et al. 2008). In the current study, the 95 SSR markers were genotyped as background markers to assess the genetic diversity and population structure of this sweet sorghum panel. With three phenotypic traits, we further tested the feasibility of mixed model association mapping methods and classified this panel as a type II sample (i.e., no strong population structure and a low level of kinship overall) (Zhu and Yu 2009). Future research may focus on candidate genes or genomic regions that have been identified for various traits related to bioenergy (e.g., brix, stem biomass, juice yield, bagasse yield, sugar content, and sugar yield) (Murray et al. 2008a, b). Transferring the knowledge gained from model species and biochemical pathway analysis will be critical to genetic studies carried out on bioenergy crops. Alternatively, large scale SNP genotyping can be conducted for genome-wide association analysis.

In summary, molecular marker analysis revealed the geographic distribution pattern of genetic diversity and population structure in the US historic sweet sorghum collection. Therefore, genomic mapping and plant breeding in bioenergy crops should be greatly facilitated by the increasingly powerful genomic technologies that yield quantitative information of the type this research demonstrates.

Acknowledgments The authors gratefully thank Dr. Tom Hash for sharing unpublished SSR marker information, Dr. Xiyin Wang for searching the chromosome location of assayed DNA markers, Ms. Lee-Ann Chalkley and Tiffany Fields for processing and distribution the sorghum accessions, and Ms. Merrelyn Spinks for helping search the GRIN database. This project is supported by USDA-ARS, Kansas Grain Sorghum Commission, Kansas State University Targeted Excellence Program in Sorghum Translational Genomics, and the National Research Initiative (NRI) Plant Genome Program of the USDA Cooperative State Research, Education and Extension Service (CSREES) Grant no. 2006-35300-17155.

References

Ali ML, Rajewski JF, Baenziger PS, Gill KS, Eskridge KM, Dweikat I (2008) Assessment of genetic diversity and relationship among a collection of US sweet sorghum germplasm. Mol Breed 21:497–509

Barkley NA, Roose ML, Krueger RR, Federici CT (2006) Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). Theor Appl Genet 112:1519–1531

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinf Online 1:47–50

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

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574–578

GRIN (2008) National Genetic Resources Program (NGRP). Germplasm Resources Information Network (GRIN). Nat Germplasm Resour Lab, Beltsville, MD. Available at http://www.ars-grin.gov

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

Hunter E, Anderson I (1997) Sweet sorghum. In: Janick L (ed) Horticultural reviews, vol 21. Wiley, New York, pp 40–73

Kong L, Dong J, Hart GE (2000) Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple-sequence repeats (SSRs). Theor Appl Genet 101:438–448

Kwak M, Gepts P (2009) Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). Theor Appl Genet 118:979–992

Liu K, Muse SV (2005) PowerMarker: integrate analysis environment for genetic marker data. Bioinformatics 21:2128–2129

Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). Am J Bot 82:1420–1425

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

Menz MA, Klein RR, Mullet JE, Obert JA, Unruh NC, Klein PE (2002) A high-density genetic map of *Sorghum bicolor* (L) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Mol Biol 48:483–499

Murray SC, Sharma A, Rooney WL, Klein PE, Mullet JE, Mitchell SE, Kresovich S (2008a) Genetic improvement of sorghum as a biofuel feedstock I: QTL for stem sugar and grain nonstructural carbohydrates. Crop Sci 48:2165–2179

Murray SC, Rooney WL, Mitchell SE, Sharma A, Klein PE, Mullet JE, Kresovich S (2008b) Genetic improvement of sorghum as a biofuel feedstock II: QTL for stem and leaf structural carbohydrates. Crop Sci 48:2180–2193

Murray SC, Rooney WL, Hamblin MT, Mitchell SE, Kresovich S (2009) Sweet sorghum genetic diversity and association mapping for brix and height. Plant Genome 2:48–62

Natoli A, Gorni C, Chegdani F, Ajmone Marsan P, Colombi C, Lorenzoni C, Marocco A (2002) Identification of QTLs associated with sweet sorghum quality. Maydica 47:311–322

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genetics 2:e190

Pederson GA, Spinks M (2006) Utilization old data to improve germplasm documentation: sweet sorghum collection. In: The abstract of 2006 international annual meetings of ASA-CSSA-SSSA, 12–16 November, Indianapolis, IN, p 1232b. Available at http://www.agronomy.org

Price AL, Patterson NJ, Plenge RM, Weiblatt ME, Shadick NA, Reich D (2007) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

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

Reddy B, Ramesh S, Reddy S, Ramaiah B, Salimath P, Kachapur R (2005) Sweet sorghum–a potential alternate raw material for bio-ethanol and bio-energy. Intl Sorghum Millets Newsl 46:79–86

Ritter KB, McIntye CL, Godwin ID, Jordan DR, Chapman SC (2007) An assessment of the genetic relationship between sweet and grain sorghums, within *Sorghum bicolor* ssp. bicolor (L.) Moench, using AFLP. Euphytica 157:161–167

Ritter KB, Jordan DR, Chapman SC, Godwin ID, Mace ES, McIntyre CL (2008) Identification of QTL for sugar-related traits in a sweet x grain sorghum (*Sorghum bicolor* L Moench) recombination inbred population. Mol Breed 22:367–384

Rooney WL, Blumenthal J, Bean B, Mullet JE (2007) Designing sorghum as a dedicated bioenergy feed stock. Biofuels Bioprod Biorefin 1:147–157

Sasaki T, Antonio BA (2009) Sorghum in sequence. Nature 457:547–548

Schloss SJ, Mitchell SE, White GM, Kukatla R, Bowers JE, Patterson AH, Kresovich S (2002) Characterization of RFLP probe sequences for gene discovery and SSR development in *Sorghum bicolor* (L.) Moench. Theor Appl Genet 105:912–920

Smith CW, Frederiksen RA (2000) Preface, p. vii-ix. In: Smith CW, Frederiksen RA (eds) Sorghum: origin, history, technology and production. Wiley, New York

Vermerris W, Saballos A, Ejeta G, Mosier NS, Ladisch MR, Carpita NC (2007) Molecular breeding to enhance ethanol production from corn and sorghum stover. Crop Sci 47:S142–S153

Wang ML, Dean R, Erpelding J, Pederson G (2006a) Molecular genetic evaluation of sorghum germplasm differing in response to fungal diseases: Rust (*Puccinia purpurea*) and anthracnose (*Collectotrichum graminicola*). Euphytica 148:319–330

Wang ML, Mosjidis JA, Morris JB, Dean RE, Jenkins TM, Pederson GA (2006b) Genetic diversity of Crotalaria germplasm assessed through phylogenetic analysis of EST-SSR markers. Genome 49:701–715

Yu J, Pressoir G, Briggs WH, Vroh BI, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen D, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhu C, Yu J (2009) Nonmetric multidimensional scaling corrects for population structure in whole genome association studies. Genetics 182:875–888

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20