

A. M. Casa · S. E. Mitchell · M. T. Hamblin
H. Sun · J. E. Bowers · A. H. Paterson
C. F. Aquadro · S. Kresovich

Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats

Received: 1 October 2004 / Accepted: 3 February 2005 / Published online: 28 April 2005
© Springer-Verlag 2005

Abstract Although molecular markers and DNA sequence data are now available for many crop species, our ability to identify genetic variation associated with functional or adaptive diversity is still limited. In this study, our aim was to quantify and characterize diversity in a panel of cultivated and wild sorghums (*Sorghum bicolor*), establish genetic relationships, and, simultaneously, identify selection signals that might be associated with sorghum domestication. We assayed 98 simple sequence repeat (SSR) loci distributed throughout the genome in a panel of 104 accessions comprising 73 landraces (i.e., cultivated lines) and 31 wild sorghums. Evaluation of SSR polymorphisms indicated that landraces retained 86% of the diversity observed in the wild sorghums. The landraces and wilds were moderately differentiated ($F_{st}=0.13$), but there was little evidence of population differentiation among racial groups of cultivated sorghums ($F_{st}=0.06$). Neighbor-joining analysis showed that wild sorghums generally formed a distinct group, and about half the landraces tended to cluster by race. Overall, bootstrap support was low, indicating a history of gene flow among the various

cultivated types or recent common ancestry. Statistical methods (Ewens-Watterson test for allele excess, $\ln RH$, and F_{st}) for identifying genomic regions with patterns of variation consistent with selection gave significant results for 11 loci (approx. 15% of the SSRs used in the final analysis). Interestingly, seven of these loci mapped in or near genomic regions associated with domestication-related QTLs (i.e., shattering, seed weight, and rhizomatousness). We anticipate that such population genetics-based statistical approaches will be useful for re-evaluating extant SSR data for mining interesting genomic regions from germplasm collections.

Electronic Supplementary Material Supplementary material is available for this article at <http://dx.doi.org/10.1007/s00122-005-1952-5>

Communicated by H.H. Geiger

A. M. Casa · S. E. Mitchell · M. T. Hamblin
H. Sun · S. Kresovich (✉)
Institute for Genomic Diversity,
Cornell University, Ithaca, NY 14853, USA
E-mail: sk20@cornell.edu
Tel.: +1-607-2551492
Fax: +1-607-2556249

J. E. Bowers · A. H. Paterson
Applied Genetics Technology Center
and Comparative Grass Genomics Center,
University of Georgia, Athens, GA 30602, USA

C. F. Aquadro
Department of Molecular Biology and Genetics,
Cornell University, Ithaca, NY 14853, USA

Introduction

Sorghum [*Sorghum bicolor* (L.) Moench] is the fifth most important cereal grown worldwide in terms of both production and area planted (FAO 2004). Like millet and fonio, sorghum is genetically suited to hot and dry agroecologies where it is difficult to grow most food grains. Sorghum, therefore, is a pillar of food security in the semi-arid zones of Western and Central Africa. Three *S. bicolor* subspecies are recognized, including cultivated types (i.e., landraces) (ssp. *bicolor*), a wild complex that is widely distributed and ecologically diverse (ssp. *verticilliflorum*), and weedy types from hybridizations between domesticated and wild sorghums (ssp. *drummondii*) (De Wet 1978). Furthermore, five races (i.e., *bicolor*, *caudatum*, *durra*, *guinea* and *kafir*) and ten hybrid races that combine characteristics of two or more races have been described within ssp. *bicolor* (Harlan and De Wet 1972).

Because of its global socio-economic importance, there has been substantial interest in characterizing the levels of genetic diversity present within sorghum using both phenotypic and molecular markers (Dean et al. 1999; Dje et al. 2000; Ghebru et al. 2002; Grenier et al. 2000a, b; Menz et al. 2004; Uptmoor et al. 2003), and

these analyses have provided a foundation of genetic data for making informed decisions regarding the management and utilization of genetic resources. More importantly, molecular diversity data can potentially bridge conservation and use when employed as a tool for mining germplasm collections for genomic regions associated with adaptive or agronomically-important traits (i.e., genes that have been important in adaptation to local environments or are associated with phenotypes selected by farmers or breeders).

Population genetics theory predicts that intense directional selection pressure should dramatically reduce allelic diversity, both at the target of selection and at linked neutral loci, due to genetic “hitchhiking” (Maynard Smith and Haigh 1974). Methods based on this prediction are now used routinely by human and *Drosophila* geneticists for identifying genomic regions that may have experienced selection or a “selective sweep” (Harr et al. 2002; Kayser et al. 2003; Payseur et al. 2002). Candidate regions are identified by surveying neutral markers throughout the genome for unusual patterns of allelic variation, such as skews in allele frequency distributions within populations or groups of individuals. Because crop species experienced strong selective pressures during domestication, they offer a unique opportunity to test the use of neutrally evolving markers for identifying genes controlling domestication-related traits (Vigouroux et al. 2002). In the work presented here, we used SSR loci to determine diversity and evaluate genetic relationships and population structure within and between cultivated and wild sorghum accessions. Concurrently, we applied statistical methods based on population genetics theory to identify regions of the *S. bicolor* genome that may have undergone human-driven selection during domestication.

Materials and methods

Plant material

Genome-wide diversity was assessed in 104 sorghum [*Sorghum bicolor* (L.) Moench] accessions, including both landraces ($n=73$) and wild sorghums (wilds) ($n=31$) [listed as electronic supplementary information (ESM) S1]. These accessions comprised all *S. bicolor* subspecies and racial types and represented a wide geographic sampling from the species’ center of diversity (Africa). Seeds of the cultivated material (landraces) were obtained either from the National Center for Genetic Resources Preservation (USDA/ARS, Ft. Collins, Colo.) or the Plant Genetic Resources Conservation Unit (USDA/ARS, Griffin, Ga.), and information on geographical origin and racial classification was gathered primarily from the System-wide Information Network for Genetic Resources database (<http://singer.cgiar.org/Search/SINGER/search.htm>). Wild accessions were provided by Mitchell R. Tuinstra (Agronomy Department, Kansas State University), and relevant

information on these was obtained from either Dr. Tuinstra or the National Plant Germplasm System (<http://www.ars-grin.gov/npgs/searchgrin.html>).

Simple sequence repeat loci

A total of 98 simple sequence repeats (SSRs) were assayed, including 47 di-, 19 tri-, and 23 tetranucleotide or longer motifs, and nine imperfect/compound repeats. These loci were developed from restriction fragment length polymorphism (RFLP) probe sequences (Schloss et al. 2002), small-insert genomic libraries (Brown et al. 1996; Kong et al. 2000), and bacterial artificial chromosome (BAC) end sequences (Bhatramakki et al. 2000). Locus designation followed the naming convention of Bhatramakki et al. (2000). All SSR loci were mapped in an F_2 population derived from an interspecific cross between *S. bicolor* inbred line BTx623 and *S. propinquum* (Chittenden et al. 1994). Loci that do not appear on the published genetic map (Bowers et al. 2003) were mapped in our laboratory (to be submitted elsewhere). A list of the SSRs assayed, including primer sequences, information on repeat motif and length, and genomic location are provided as ESM (S2).

DNA extraction, PCR, and data collection

DNA was isolated from individual seedlings following a standard CTAB extraction protocol (Doyle and Doyle 1987). PCRs were performed as previously described (Matsuoka et al. 2002) using primers with fluorescent labels [6-FAM, HEX, TET, NED, TAMRA, VIC or PET; Applied Biosystems Inc. (ABI), Foster City, Calif.]. Samples containing 0.5 μ l PCR product, 0.05 μ l size standard (GeneScan-500 LIZ; ABI), 9.0 μ l formamide, and 0.45 μ l water were denatured (95°C/5 min) and assayed on a capillary DNA sequencer (Model 3730XL; ABI). Fragment sizes were determined based on migration relative to the internal lane size standard using GENEMAPPER VER. 3.0 software (ABI) and the “local Southern” sizing algorithm.

Analysis of genetic diversity

Number of alleles (A), allele frequencies (p_i), and expected heterozygosity (H_e) (also referred to as unbiased gene diversity, D) for each locus was calculated with POWERMARKER VER. 3.0 (Liu and Muse 2002). Here, we use expected heterozygosity and gene diversity interchangeably. Allelic richness (R_s), a measure of the number of alleles corrected for sample size differences, was estimated using FSTAT (Goudet 2001). Relative loss of diversity in cultivated versus wild was calculated according to Vigouroux et al. (2002). Mann-Whitney (MW) tests were used to determine whether estimates were significantly different between groups. Genetic dis-

tance (Nei et al. 1983) matrices and cluster analyses based on neighbor-joining (Saitou and Nei 1987) were carried out with POPULATIONS VER.1.2.28 (<http://www.pge.cnrs-gif.fr/bioinfo/populations/index.php>). Branch support was determined by bootstrap (1,000 replicates) (Efron 1979). A matrix containing SSR allele information for the germplasm tested is available from <http://igd.tc.cornell.edu>.

Statistics for detecting selection

To identify candidate loci linked to genomic regions that might have experienced selection, we used three statistics: the Ewens-Watterson test for allele excess (EW), lnRH, and *F_{st}*. Invariant loci were excluded from these analyses. EW was calculated using BOTTLENECK VER.1.2 (Cornuet and Luikart 1996). Accessions were assigned to either cultivated or wild groups and evaluated separately under the infinite allele mutation model (IAM) of microsatellite evolution (see Results and discussion section). Probability values were obtained based on 1,000 replications. lnRH, a comparison of the relative amount of gene diversity between two populations or groups of individuals (see Kauer et al. 2003), was estimated as follows:

$$\ln[(RH)] = \ln \left[\frac{\left(\left(\frac{1}{1-He_{Pop1}} \right)^2 - 1 \right)}{\left(\left(\frac{1}{1-He_{Pop2}} \right)^2 - 1 \right)} \right]$$

where He_{Pop1} is the gene diversity in population 1 (in our study landraces), and He_{Pop2} is the gene diversity in population 2 (wilds). A Kolmogorov-Smirnov test was performed to ascertain whether the observed lnRH values followed a normal distribution. For detecting genomic regions under selection, we standardized lnRH by the mean and standard deviation across all loci (Kauer et al. 2003). The standardized distribution, therefore, had a mean of 0 and a standard deviation of 1, and 95% of loci were expected to have values between 1.96 and -1.96. Any locus with a value falling outside this interval was significant ($P < 0.05$) (i.e., the locus might have experienced selection). Unbiased estimates of *F_{st}* (Weir and Cockeram 1984) between cultivated and wild sorghum accessions and among cultivated races were calculated from allele frequency data using POWERMARKER VER. 3.0. Several methods, largely based on demography, have been developed to derive expected distributions of *F_{st}* values (Beaumont and Nichols 1996; Akey et al. 2002; Balloux and Goudet 2002). However, for most organisms, including sorghum, knowledge of the intrinsic factors that affect population demography is lacking. We, therefore, adopted a strictly empirical approach for determining *F_{st}* significance. That is, we assumed that outliers (loci with the highest *F_{st}* values) were most likely linked to genomic regions that had experienced selection (Kayser et al. 2003).

Results and discussion

Genome-wide diversity scan

The 98 SSRs evaluated in this study provided a fair coverage of the genome, with between 8 and 16 loci represented in each of the ten sorghum linkage groups. We excluded 22 loci with $\geq 20\%$ null alleles (primarily due to poor amplification in the wild sorghums). Our final dataset, therefore, was based on results from 76 SSRs.

As expected, the dinucleotide repeat-containing SSRs were, generally, more variable than those with longer motifs. Moreover, *Xgap* and *Xtxp* SSRs usually had more alleles than *Xcup* loci. This finding was probably due to differing SSR origins. The *Xgap* and *Xtxp* markers were isolated from either small-insert genomic libraries (Brown et al. 1996; Kong et al. 2000) or bacterial artificial chromosome (BAC)-end sequences (Bhatramakki et al. 2000). These loci, therefore, were more likely to include noncoding regions than the *Xcup* SSRs that were developed from low-copy RFLP probe sequences located primarily near or in genes (Schloss et al. 2002).

It is noteworthy that only 16 of 72 polymorphic loci (23%) followed either a stepwise (Kimura and Ohta 1978) or two-phase (Di Rienzo et al. 1994) mutation model of microsatellite evolution. Although the stepwise mutation model (SMM) has been widely accepted as the main mode through which SSRs evolve, our results indicated that most (approx. 77%) of the sorghum loci did not fit this model (loci that followed the SMM are indicated in ESM S3). This proportion is only slightly lower than what has been reported for maize, where 87% of SSR loci exhibited alleles that did not reflect the differences expected based on the variation in repeat size (Matsuoka et al. 2002). Because sorghum SSRs did not conform to the SMM, we assumed the infinite alleles model (IAM) (Wright 1949; Kimura and Crow 1964) in relevant statistical analyses. The IAM is not ideal for describing the mutational processes underlying SSR evolution in sorghum, and a greater understanding of this issue is needed for developing improved computational tools for population and evolutionary analyses in grasses.

Genetic diversity in wild and cultivated sorghums

Allelic richness (R_s) and gene diversity (He) values for sorghum landraces, cultivated races, and wild subspecies are shown in Table 1. In addition, a more detailed breakdown of these estimates by locus is provided in ESM S3. Four loci, *Xcup18*, -21, -65 and *sugar transporter* were monomorphic, and a fifth locus, *Xcup58*, was invariant in landraces. As seen in previous studies of sorghum germplasm with diverse phenotypes and geographic origins (Grenier et al. 2000b), the majority of

Table 1 Allelic richness (R_s) and gene diversity (He) in *Sorghum bicolor* landraces, cultivated races, and wild subspecies

Sample group	Allelic richness (R_s)	Gene diversity (He)
Landraces total	4.941	0.510
Race bicolor	2.986	0.513
Caudatum	2.696	0.453
Durra	2.898	0.479
Guinea	2.755	0.461
Kafir	1.953	0.293
Wilds total	6.187	0.590

alleles detected in this study (64%) were “rare” [i.e., allele frequency (p_i) < 0.05]. Because allele number is dependent on sample size and we assayed twice as many landraces as wild accessions, we estimated R_s , a measure that is independent of sample size. The results indicated that R_s was significantly higher in wild sorghums than in landraces [6.2 and 4.9, respectively; MW tests ($P < 0.05$)] (Table 1). Although landraces exhibited 86% of the gene diversity observed in the wilds, they were significantly less diverse than the wilds (average He in landraces was 0.51 compared to 0.59 in wilds; MW tests $P < 0.05$) (Table 1). As observed for other crops (Vigouroux et al. 2005), reduced variation in cultivated types typically results from population bottlenecks during domestication. We did observe four loci (*Xcup11*, -16, -19 and -53), however, that exhibited up to twofold more variation in landraces than in wilds (ESM S3). This result might reflect our heavy sampling of landraces, the occurrence of balancing selection in cultivated sorghums, and/or founder effects associated with domestication.

For sorghum races bicolor, caudatum, durra, and guinea, there were no significant differences in diversity (Table 1). The kafir accessions, however, were significantly less diverse than the other races as measured both by allelic richness and gene diversity (MW tests $P < 0.001$). Other studies based on SSRs (Dje et al. 2000) and, more recently, on DNA sequences (Hamblin et al. 2004) have shown similar results. This lower diversity could reflect the relative genetic isolation of race kafir in southern Africa. Because it is adapted to a more temperate environment, race kafir tends to be more photoperiod insensitive than the other sorghum races (Grenier et al. 2001). Both geographic isolation and temperate adaptation, therefore, may have contributed to limited opportunities for matings between kafir and other racial types.

Phenetic analysis

Neighbor-joining analysis indicated that wild sorghums belonging to *ssp. verticilliflorum* generally formed a coherent group that failed to cluster with the landraces (Fig. 1, group I). Moreover, wild accessions from the same geographic region tended to be genetically more similar to each other than to those from more distant

locations, indicating the presence of population structure. *S. bicolor ssp. drummondii*, an annual weed thought to be a natural hybrid between *ssp. bicolor* and *verticilliflorum* (De Wet 1978), grouped mainly with a subset of the cultivated *ssp. bicolor* (namely, race bicolor) (Fig. 1, group II). Race bicolor resembles spontaneous weedy sorghums but lacks the ability to disperse seeds naturally (i.e., seeds do not shatter) (de Wet 1978). Because its long, clasping glumes, elongated seed, and open panicles are considered to be primitive characters, bicolor is thought to be the race most closely related to wild sorghums (Harlan and de Wet 1972). The remaining cultivated types were loosely partitioned into three groups: one consisting primarily of bicolor and durra accessions (Fig. 1, group II), another containing the kafir accessions and a mixture of several other racial types (group III), and a third group comprising mainly the guinea and caudatum accessions (group IV). We should emphasize that branch support for groupings of cultivated accessions was low (most bootstrap values ≤ 20), indicating a history of gene flow among the various races and/or recent common ancestry.

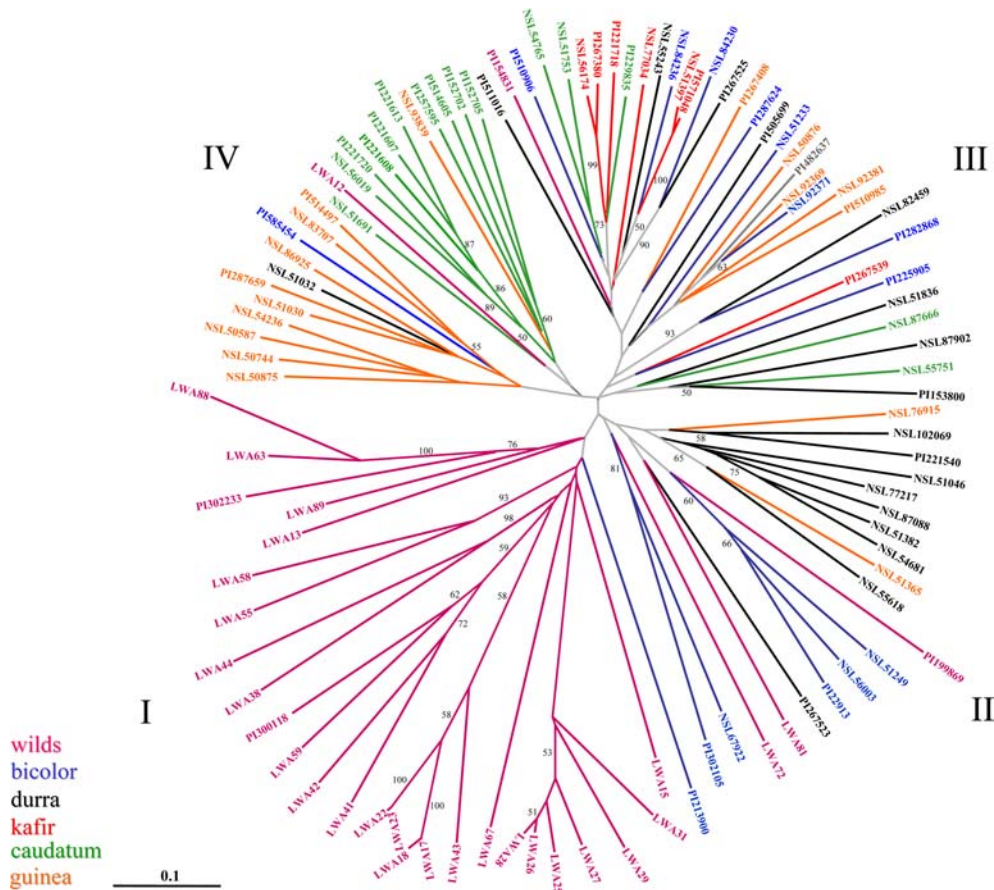
Partitioning variation

Levels of variation can be attributed to within- and between-population components. F_{st} (Wright 1951) measures genetic differentiation and allows one to determine how much variation is due to differences between populations. There was a moderate degree of differentiation between the wilds and landraces (average $F_{st} = 0.13$) and little differentiation among the cultivated races (average $F_{st} = 0.06$). This relatively low F_{st} might reflect the diverse composition of the germplasm evaluated (i.e., accessions were deliberately chosen to include the racial and geographic diversity represented in sorghum), the occurrence of pollen or seed flow, and/or recent divergence.

Using SSR diversity to detect the signals of selection

Results from statistical analyses for detecting patterns of variation that might be indicative of selection are shown in Fig. 2. Of the 72 polymorphic SSRs analyzed, 11 loci showed significant deviation from the population average or from neutrality in at least one statistical test (*Xcup14*, -40, -01, -33, -22, -60, -15, -07, -09, *Xgap256* and -10) (Fig. 2). Most loci (8/11) were identified by the EW test for allele excess (Fig. 2). This statistic, equivalent to the Watterson’s homozygosity test (Watterson 1978), compares gene diversity (He) to that expected, given the number of alleles observed at a locus. The premise underlying EW is that under neutrality, allele number and frequency distribution result from equilibrium between mutation and genetic drift. Following a selective sweep, however, a surplus of low-frequency alleles is generated as new mutations occur in the

Fig. 1 Neighbor-joining phenogram depicting genetic relationships among *Sorghum bicolor* accessions. Wild accessions and the various cultivated races of sorghum are color-coded. Numbers along branches denote bootstrap support (shown only for values greater than 50)



selected region. Therefore, the power of EW increases over time until equilibrium is re-established in the population.

Seven of the eight loci identified by the EW test exhibited a deficiency of heterozygosity (i.e., EW values were negative) (Fig. 2), which is consistent with the expectation under a directional or positive selection scenario (i.e., trait selection by farmers or breeders). One locus (*Xcup40*), however, showed an excess of heterozygosity (EW value was positive), and the allele frequency distribution at *Xcup40* (data not shown) was consistent with diversifying selection in cultivated sorghum (i.e., selection stemming from adaptation to environment, pests, or disease). Since the distribution of allele number versus gene diversity over all loci was skewed towards a reduction in allele number compared to the neutral steady-state distribution (data not shown), our significance levels for the EW test were probably conservative.

While EW compares observed and expected frequency distribution at each locus within groups (e.g., landraces), lnRH compares variability between groups (landraces and wilds). Because, lnRH is a ratio of variances in gene diversity, this test has an identical expectation for all loci independent of the SSR mutation rate and the effective population size. A recent population-specific selective sweep however, would be expected to change that ratio at loci linked to the selected site but

not at genomic regions evolving neutrally (for a detailed discussion, see Schlotterer 2002; Kauer et al. 2003). The power of lnRH is highest immediately following the selective event (when the number of alleles in the population is dramatically reduced) but declines over time as new alleles are generated by mutation and population diversity increases (when the power of EW is highest). For most SSRs (53/72), non-standardized lnRH values were negative, indicating a lower variability in landraces than in wilds. Following standardization, lnRH statistics for two loci, *Xgap256* and *Xcup09*, were marginally significant ($P < 0.05$) (Fig. 2).

Besides providing insights into how genetic variation is partitioned (see above), *F_{st}* can also be used for identifying loci under selection. Unlike lnRH, the power of *F_{st}* does not decline over time as new variants arise as long as different alleles predominate in the groups being compared. *F_{st}* values for three loci (*Xcup15*, *Xcup60*, and *Xgap256*) were at least fourfold higher (> 0.50) than the genome-wide average (0.13) (Fig. 2) and were considered to be most likely linked to genomic regions under selection.

Approximately 15% of the SSRs evaluated had patterns of variation consistent with selection according to at least one of the tests. Only 1 of the 11 loci identified, *Xgap256*, was significant by all methods. We should emphasize that deviation from neutral expectation is only the first step towards the identification of target

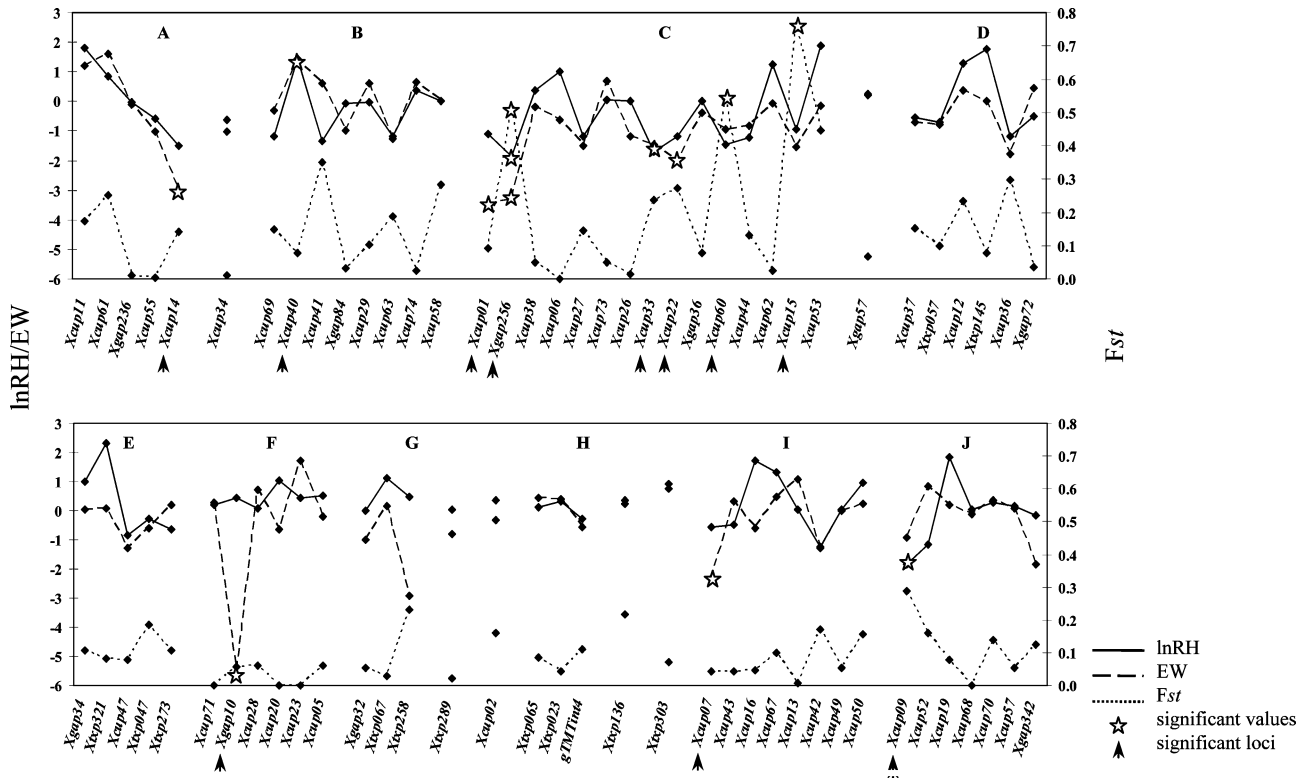


Fig. 2 Plot of F_{st} , $\ln RH$, and EW (deficiency of heterozygosity) values for 72 SSR loci by linkage group. *Right-hand side* y -axis scales, F_{st} ; *left-hand side* $\ln RH$ and EW . Loci designations are on the x -axis. F_{st} values are connected by the dotted lines, $\ln RH$ by solid lines, and EW by dashed lines. Stars indicate significant values

($P < 0.05$), and corresponding loci are denoted by arrows. Map positions of loci (centiMorgans) are presented as ESM S2. Loci not mapped in the *S. bicolor* \times *S. propinquum* population appear at the right end of each linkage group (A–J) and are not connected to mapped loci by lines

genes. The incorporation of both DNA sequence diversity and map location of candidate loci will be invaluable for associating candidates with quantitative trait loci (QTLs) for traits that were/are under selection. Seven loci identified in this study map in or near regions associated with domestication-related QTLs. For example, *Xcup33* maps to the shattering region on linkage group (LG) C (Paterson et al. 1995a, b), *Xcup14* is located near a QTL for leaf width and seed weight on LG A (Paterson et al. 1995a), and *Xgap10*, 22, 60, *Xgap10*, and 256 map to major QTLs for rhizomatousness on LGs C and F (Hu et al. 2003; Paterson et al. 1995b). Results from independent diversity studies can also lend support to the selection hypothesis. For example, the low variation we observed for cultivated sorghum at locus *Xgap256* was consistent with findings by Grenier et al. (2000b) where this locus had both the lowest number of alleles (7 compared to an average 19.3) and gene diversity (0.27 compared to average 0.80) of all loci evaluated in approximately 700 sorghum landraces from the world collection. We should also note that although significance at some loci was marginal, we have now obtained DNA sequence data from loci linked to two of the candidate SSRs, and these results were also consistent with a selective sweep (i.e., variation deviated significantly from neutral expectation).

Possibilities and caveats

SSRs have been extensively used in sorghum for germplasm characterization (Grenier et al. 2000b), evaluation of breeding material (Smith et al. 2000; Menz et al. 2004) and genetic diversity analyses (Dje et al. 2000; Ghebru et al. 2002; Uptmoor et al. 2003). Here, we applied population genetics principles to SSR diversity data in an attempt to identify sorghum genomic regions that might have experienced recent selection. One of the advantages of using this strategy is that no a priori information is required, either of traits or function, for identifying regions of the genome underlying functional or adaptive diversity. Indeed, about one-third of the candidate loci identified in this study had no association with previously characterized QTLs. The effectiveness of genome-wide scans, however, relies heavily on both a knowledge of population structure and appropriate sampling. In our sorghum panel, we found only weak evidence of population structure between cultivated and wilds as well as among the cultivated sorghum races. Therefore, loci showing significant deviation from mutation-drift equilibrium (as measured by the EW test) or high levels of genetic differentiation (F_{st}) are more likely to be associated with selection events and not false signals produced by population structure.

SSRs are usually not the targets of selection, and the probability that variation will be reduced at a linked site during a selective sweep depends not only on the strength of selection, but also on SSR mutation mode and rate, and the rate of recombination between the selected site and the marker assayed. Because higher mutation rates translate into higher diversity, depending on when selection took place, SSR loci may quickly recover lost diversity and, therefore, give non-significant deviation from neutrality (Vigouroux et al. 2005). In a species like maize where linkage disequilibrium (LD) decays rapidly (r^2 falls to 0.24 by 200 bp) (Remington et al. 2001; Tenaillon et al. 2001), a signal detected by an SSR would indicate that the target of selection is nearby (possibly within a few hundred nucleotides). However, rapid LD decay also means that to detect selection a candidate region would need to be saturated with markers. Because LD decays much more slowly than in maize, *S. bicolor* might be more amenable to population genetics-based assays for selection and association analyses. Recent studies have shown that in sorghum, r^2 declines to 0.2, on average, by 15 kb (Hamblin et al. 2004; M.T. Hamblin, unpublished data). Compared to maize, therefore, a smaller number of markers would be required both to scan the sorghum genome for selection signals and to detect associations between genotype and phenotype.

Conclusions

While the area of functional genomics is still in its infancy, we anticipate that the use of population diversity-based approaches will allow the mining of germplasm collections and extant SSR diversity data for identifying interesting genomic regions. Certainly molecular data from species with little population structure and intermediate levels of LD would be well suited for re-analysis. We must stress that while these approaches may be advantageous for identifying genomic regions that differ from the average observed in the genome, some of these departures may also result from non-equilibrium population history. Functional studies (e.g., mutant screening, genetic complementation, expression analysis, biochemical localization and characterization, etc) are still required to establish causation.

Acknowledgements We thank Patrick Brown and Seth Murray (Cornell University) for assistance with data collection and Mitchell Tuinstra (Kansas State University) for his comments and suggestions. This work was funded by NSF grant DBI 0115903 (to AHP, CFA and SK).

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12:1805–1814

- Balloux F, Goudet J (2002) Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Mol Ecol* 11:771–783
- Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc London Ser B* 263:1619–1626
- Bhattarakkhi D, Dong J, Chhabra AK, Hart G (2000) An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. *Genome* 43:988–1002
- Bowers JE, Abbey C, Anderson S, Chang C, Draye X, Hoppe AH, Jessup R, Lemke C, Lenington J, Li Z, Lin YR, Liu SC, Luo L, Marler BS, Ming R, Mitchell SE, Qiang D, Reischmann K, Schulze SR, Skinner DN, Wang YW, Kresovich S, Schertz KF, Paterson AH (2003) A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics* 165:367–386
- Brown SM, Hopkins MS, Mitchell SE, Wang TY, Kresovich S, Duncan RR, Senior ML, Gonzales-Candelas F (1996) Multiple methods for the identification of polymorphic simple sequence repeats in sorghum [*Sorghum bicolor* (L.) Moench]. *Theor Appl Genet* 93:190–198
- Chittenden LM, Schertz KF, Lin YR, Wing RA, Paterson AH (1994) A detailed RFLP map of *Sorghum bicolor* × *S. pro-pinquum*, suitable for high-density mapping, suggests ancestral duplication of sorghum chromosomes or chromosomal segments. *Theor Appl Genet* 87:925–933
- Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144:2001–2014
- De Wet MJJ (1978) Systematics and evolution of *Sorghum* Sect. *Sorghum* (Gramineae). *Am J Bot* 65:477–484
- Dean RE, Dahlberg JA, Hopkins MS, Mitchell SE, Kresovich S (1999) Genetic redundancy and diversity among 'orange' accessions in the US National Sorghum Collection as assessed with simple sequence repeat (SSR) markers. *Crop Sci* 39:1215–1221
- Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci USA* 91:3166–3170
- Dje Y, Heuertz M, Lefebvre C, Vekemans X (2000) Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. *Theor Appl Genet* 100:918–925
- Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19:11–15
- Efron B (1979) Bootstrap methods: another look at the jackknife. *Ann Stat* 7:1–26
- FAO (2004) Production yearbook 2002 no. 56. FAO Statistic Series no. 176, Rome
- Ghebru B, Schmidt RJ, Bennetzen JL (2002) Genetic diversity of Eritrean sorghum landraces assessed with simple sequence repeat (SSR) markers. *Theor Appl Genet* 105:229–236
- Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from <http://www.unil.ch/izea/software/fstat.html>
- Grenier C, Bramel-Cox PJ, Noirot M, Prasada Rao KA, Hamon P (2000a) Assessment of genetic diversity in three subsets constituted from the ICRISAT sorghum collection using random vs. non-random sampling procedures. A. Using morpho-agronomical and passport data. *Theor Appl Genet* 101:190–196
- Grenier C, Deu M, Kresovich S, Bramel-Cox PJ, Hamon P (2000b) Assessment of genetic diversity in three subsets constituted from the ICRISAT sorghum collection using random vs. non-random sampling procedures. B. Using molecular markers. *Theor Appl Genet* 101:197–202
- Grenier C, Bramel-Cox PJ, Harmon P (2001) Core collection of sorghum: I. stratification based on eco-geographical data. *Crop Sci* 41:234–240

- Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA, Paterson AH, Kresovich S (2004) Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics* 167:471–483
- Harlan JR, De Wet MJM (1972) A simplified classification of cultivated sorghum. *Crop Sci* 12:172–176
- Harr B, Kauer M, Schlotterer C (2002) Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 99:12949–12954
- Hu FY, Tao DY, Sacks E, Xu P, Li L, Yang Y, McNally K, Khush GS, Paterson AH, Li ZK (2003) Convergent evolution of perenniality in grasses. *Proc Natl Acad Sci USA* 100: 4050–4054
- Kauer MO, Dieringer D, Schlotterer C (2003) A microsatellite variability screen for positive selection associated with the “out of Africa” habitat expansion of *Drosophila melanogaster*. *Genetics* 165:1137–1148
- Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol Biol Evol* 20:893–900
- Kimura M, Crow J (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738
- Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc Natl Acad Sci USA* 75:2868–2872
- Kong L, Dong J, Hart G (2000) Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple sequence repeats (SSRs). *Theor Appl Genet* 101:438–448
- Liu K, Muse S (2002) POWERMARKER: new genetic data analysis software (version. 3.0). Available from <http://www.powermarker.net/>
- Matsuoka Y, Mitchell SE, Kresovich S, Goodman M, Doebley J (2002) Microsatellites in *Zea*-variability, patterns of mutations, and use for evolutionary studies. *Theor Appl Genet* 104:436–450
- Maynard Smith J, Haigh J (1974) The hitchhiking effect of a favourable gene. *Genet Res* 23:23–35
- Menz MA, Klein RR, Unruh NC, Rooney WL, Klein PE, Mullet JE (2004) Genetic diversity of public inbreds of sorghum determined by mapped AFLP and SSR markers. *Crop Sci* 44:1236–1244
- Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J Mol Evol* 91:153–170
- Paterson AH, Lin YR, Li Z, Schertz KF, Doebley JF, Pinson SRM, Liu SC, Stansel JW, Irvine JE (1995a) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* 269: 1714–1718
- Paterson AH, Schertz KF, Lin YR, Liu SC, Chang YL (1995b) The weediness of wild plants: molecular analysis of genes responsible for dispersal and persistence of johnsongrass (*Sorghum halepense* L. Pers.). *Proc Natl Acad Sci USA* 92: 6127–6131
- Payseur BA, Cutter AD, Nachman MW (2002) Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol Biol Evol* 19:1143–1153
- Remington DL, Thornsberry JM, Matsuoka Y, Wilson, LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 98:11479–11484
- Saitou N, Nei M (1987) The neighbor-joining method—a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Schloss SJ, Mitchell SE, White GM, Kukatla R, Bowers JE, Paterson 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
- Schlotterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160:753–763
- Smith, JSC, Kresovich S, Hopkins MS, Mitchell SE, Dean RE, Woodman WL, Lee M, Porter K (2000) Genetic diversity among elite sorghum inbred lines assessed with simple sequence repeats. *Crop Sci* 40:226–232
- Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays*). *Proc Natl Acad Sci USA* 98:9161–9166
- Uptmoor R, Wenzel W, Friedt W, Donaldson G, Ayisi K, Ordon F (2003) Comparative analysis on the genetic relatedness of *Sorghum bicolor* accessions from Southern Africa by RAPDs, AFLPs and SSRs. *Theor Appl Genet* 106:1316–1325
- Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc Natl Acad Sci USA* 99:9650–9655
- Vigouroux Y, Mitchell SE, Matsuoka Y, Hamblin MT, Kresovich S, Smith JSC, Jaqueth J, Smith OS, Doebley J (2005) An analysis of genetic diversity across the maize genome using microsatellites. *Genetics* (in press)
- Watterson GA (1978) The homozygosity test of neutrality. *Genetics* 88:405–417
- Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution* 38: 1358–1370
- Wright S (1949) Adaptation and selection. In: Jepsen GL, Simpson GG, Mayr E (eds) *Genetics, paleontology, and evolution*. Princeton University Press, New Jersey, pp 365–389
- Wright S (1951) The genetical structure of populations. *Ann Eugenics* 15:323–354