The Plant Genome
Original Research

Evidence for a Selective Sweep on Chromosome 1 of Cultivated Sorghum

Alexandra M. Casa,* Sharon E. Mitchell, Jeffrey D. Jensen, Martha T. Hamblin, Andrew H. Paterson, Charles F. Aquadro, and Stephen Kresovich

A.M. Casa, S.E. Mitchell, M.T. Hamblin, and S. Kresovich, Institute for Genomic Diversity, Cornell Univ., Ithaca, NY 14853; J.D. Jensen and C.F. Aquadro, Dep. of Molecular Biology and Genetics, Cornell Univ., Ithaca, NY 14853; A.H. Paterson, Plant Genome Mapping Laboratory and Comparative Grass Genomics Center, Univ. of Georgia, Athens, GA 30602. DNA sequences were deposited in GenBank under accession numbers DQ459071, DQ462793–DQ463100. Received 1 February 2006. *Corresponding author (amc56@cornell.edu).

Abstract

Recently, a simple sequence repeat (SSR)-based genome-wide diversity scan of *Sorghum bicolor* (L.) Moench identified several candidate loci with patterns of variation consistent with directional selection in cultivated lines. Data were insufficient, however, to determine if selection had actually occurred at or near candidate SSR loci or if the unusual diversity patterns observed were due to the effects of demographic factors such as population bottlenecks or mating system. In the present study, we collected DNA sequences from 10 segments within a 99 kb region flanking one of the previously identified candidates, SSR locus *Xcup15*, located near the distal end of chromosome 1. We performed statistical tests both to address alternative hypotheses to selection and to aid in localizing the selection target. Analyses of genomic DNA sequences from a panel of 17 cultivated and 13 wild accessions indicated that cultivated lines had reduced diversity in this region (about one-third of the diversity present in wild sorghums) and a moderate degree of differentiation was observed between cultivated and wild groups (F_{st} = 0.15). Several features of the data support the hypothesis that recent directional selection shaped diversity patterns around *Xcup15*, including overall low levels of variation and extensive haplotype structure (a predominant haplotype occurred over the 99 kb region) in cultivated sorghum, and a derived fixed difference at the 5′ untranslated region (UTR) of a protein phosphatase 2C (PP2C) gene between cultivated and wild sorghums. Moreover, two of the four tests employed to detect deviations from the neutral, equilibrium model, the Hudson Kreitman Aguadé (HKA), and the composite likelihood ratio (CLR) tests indicated that patterns of diversity in the *Xcup15* region were consistent with a selective sweep. Although we were unable to rule out demography as a possible explanation for the diversity patterns observed along this region, this study supported previous findings based on SSR diversity and identified candidates for the target of selection; the confirmation of which will require functional and association studies.

Published in Crop Sci 46(S1) (2006). Published 8 Aug. 2006.
doi:10.2135/cropsci2006.0001tpg

677 S. Segoe Rd., Madison, WI 53711 USA

During the past decade, SSRs have been extensively used for quantifying neutral genetic diversity in plant in situ populations and ex situ germplasm collections (Garris et al., 2005). Although a great deal of valuable genetic information for managing both in situ and ex situ collections has been generated, these studies have been retrospective rather than predictive (Kresovich et al., 2006). It would be desirable, therefore, to use information gathered by such analyses to concurrently identify agronomically or horticulturally useful diversity. With this point in mind, we have recently used population genetics-based analysis of genome-wide SSR diversity in *S. bicolor* to locate candidate genomic regions that may have undergone diversifying and, in particular, directional selection (Casa et al., 2005).

Population genetics theory predicts that intense directional selection, as would be experienced during crop domestication, is expected to dramatically reduce variation at the genomic target of selection and at linked neutral loci, a phenomenon known as *genetic hitchhiking* (Maynard Smith and Haigh, 1974). Following a selective sweep, new mutations aris-

Abbreviations: SSR, simple sequence repeat; BAC, bacterial artificial chromosome; EST, expressed sequence tag; HKA, Hudson Kreitman Aguadé; CLR, composite likelihood ratio; GOF, goodness-of-fit; LD, linkage disequilibrium; MITE, miniature inverted repeat transposable element; MLE, maximum likelihood estimate; PP2C, protein phosphatase 2C; UTR, untranslated region.

ing in the selected region initially result in a skew in the site frequency spectrum (i.e., excess of rare alleles). Selection might also lead to genetic differentiation as a consequence of allele frequency shifts between selected and nonselected populations (e.g., a domestication-associated allele will quickly increase in frequency in the cultivated populations). In a genome-wide scan of diversity at neutral markers such as SSRs, loci that show unusual patterns of allelic variation relative to genome-wide averages (i.e., locus-specific reductions in diversity, excess of rare alleles, or increased population differentiation) may be linked to targets of selection. Genome-wide scans of diversity have been used in this manner to identify candidate genomic regions in organisms such as humans, *Drosophila*, *Arabidopis*, and maize (*Zea mays* L.) (Vigouroux et al., 2002; Kauer et al., 2003; Kayser et al., 2003; Aranzana et al., 2005). Once a candidate region has been identified, it may be possible to identify the target by surveying adjacent genomic regions and looking for a return to neutral patterns of variation (Schlotterer, 2003).

In a genome-wide scan of diversity at neutral markers such as SSRs, loci that show unusual patterns of allelic variation relative to genome-wide averages may be linked to targets of selection.

Sorghum bicolor, a tropical grass probably domesticated in eastern Africa 3000 to 6000 years ago (Kimber, 2000), is the fifth most important grain crop worldwide (FAO, 2004). Because of its ability to tolerate drought, soil toxicities, and temperature extremes more effectively than other cereals including maize, grain sorghum is a pillar of food security in the semi-arid zones of western and central Africa. Sorghum's global socioeconomic importance has prompted substantial interest in characterizing levels of genetic diversity using molecular markers (Dje et al., 2000; Grenier et al., 2000; Menz et al., 2004). More recently, studies of DNA sequence diversity (Hamblin et al., 2004, 2005, 2006) have indicated that sorghum has both lower nucleotide diversity and more extensive linkage disequilibrium (LD) than maize. Compared with more distantly related rice (Garris et al., 2003), however, sorghum has less extensive LD.

Recently, an SSR-based genome-wide scan of diversity in *S. bicolor* identified several loci with patterns of variation deviating from neutral expectations (Casa et al., 2005), but data were not sufficient to determine whether the apparent signal of selection resulted from a true selective event or from demographic factors such as population bottlenecks or mating system. For example, bottlenecks can produce locus-specific effects that resemble the effects of directional selection (Thornton and Andolfatto, 2006), and the degree of genetic differentiation between populations is usually higher in self-pollinating species than in outcrossers, independent of selection (Hamrick and Godt, 1996).

Here, we sequenced a bacterial artificial chromosome (BAC) clone containing the previously identified candidate SSR locus, *Xcup15*, which exhibited the highest genetic differentiation between wild and cultivated *S. bicolor* (F_{st} = 0.76) (Casa et al., 2005). We also collected and analyzed DNA sequence data from a panel of 17 cultivated and 13 wild sorghum accessions to determine if patterns of variation in this region of the *S. bicolor* genome show evidence of a domestication-related selective sweep.

Materials and Methods

Comparative Analysis

Identification of BAC Clones Containing Candidate SSR and Sequencing

As reported in an earlier study, primers that amplify candidate SSR locus *Xcup15* were developed from the DNA sequence of *S. bicolor* restriction fragment length polymorphism probe pSB1790 (Schloss et al., 2002). The BAC clones from BTx623, an elite *S. bicolor* inbred line, were obtained from the Clemson University Genomics Institute (www.genome.clemson.edu/groups/bac/, verified 5 May 2006) and clones containing *Xcup15* were identified by hybridization to an overgo probe (SOG0602) derived from pSB1790. Restriction fragment length polymorphism probe pSB1790 maps to *S. bicolor* chromosome 1 at 227.9 cM (P.J. Brown, 2006, personal communication) on the BTx623 × IS3620C map (Menz et al., 2002) and at 106.9 cM on the *S. bicolor* BTx623 × *S. propinquum* map (Bowers et al., 2003). We should note that *S. bicolor* chromosome 1 corresponds to linkage group (LG) A of Peng et al. (1999) and to LG C of Chittenden et al. (1994).

DNA was extracted from single BAC clones and used as templates in PCRs to confirm the presence of locus *Xcup15*. From these clones, a single BAC, c0156b06, was selected for complete DNA sequencing because of its central position on the *S. bicolor* physical map (www.genome.arizona.edu, verified 5 May 2006) relative to the location of *Xcup15* and to the other BACs evaluated.

Table 1. *Sorghum bicolor* accessions analyzed including cultivated (ssp. *bicolor*) types, wild (ssp. *arundinaceum*) and outgroup (*S. propinquum*).

Accession ID	ICRISAT† ID	Origin	Subspecies	Race
Cultivated				
BTx623		U.S. inbred line	*bicolor*	
NSL50875	IS7171	Chad	*bicolor*	guinea
NSL51030‡	IS3817	Mali	*bicolor*	guinea
NSL51365	IS6272	India	*bicolor*	guinea
NSL55243	IS917	Algeria	*bicolor*	durra
NSL56003	IS8822	Kenya	*bicolor*	bicolor
NSL56174	IS8539	Ethiopia	*bicolor*	kafir
NSL77034	IS10400	Uganda	*bicolor*	kafir
NSL77217	IS10747	Chad	*bicolor*	bicolor
NSL87666	IS7115	Central African Rep	*bicolor*	caudatum
NSL87902	IS14790	Cameroon	*bicolor*	durra
NSL92371	IS14318	Swaziland	*bicolor*	bicolor
PI152702	IS12568	Sudan	*bicolor*	caudatum
PI221607	IS2361	Nigeria	*bicolor*	caudatum
PI267408	IS2724	Uganda	*bicolor*	guinea
PI267539	IS2901	India	*bicolor*	kafir
PI585454	IS25061	Ghana	*bicolor*	bicolor
Wild				
L-WA13		Sudan	*arundinaceum*	verticilliflorum
L-WA15‡		Sudan	*arundinaceum*	verticilliflorum
L-WA17	IS14215	Angola	*arundinaceum*	verticilliflorum
L-WA22‡	IS14235	Angola	*arundinaceum*	verticilliflorum
L-WA29	IS14254	Angola	*arundinaceum*	verticilliflorum
L-WA38	IS14279	South Africa	*arundinaceum*	verticilliflorum
L-WA42	IS14313	South Africa	*arundinaceum*	verticilliflorum
L-WA55		Benin	*arundinaceum*	arundinaceum
L-WA59	IS14300	South Africa	*arundinaceum*	arundinaceum
L-WA63	IS14359	Malawi	*arundinaceum*	arundinaceum
L-WA67‡	IS14485	Sudan	*arundinaceum*	aethiopicum
L-WA88	IS18808	Egypt	*arundinaceum*	virgatum
PI302233		former Soviet Union	*arundinaceum*	arundinaceum
Outgroup				
S. propinquum		The Philippines	NA§	NA

† International Crops Research Institute for the Semi-Arid Tropics.
‡ Data from these accessions were not used in the CLR and GOF tests.
§ NA, not applicable.

Randomly sheared libraries with inserts ranging from 1.0 to 4.0 kb were constructed and shotgun sequencing was performed with pGEM-T (Promega Corporation, Madison, WI) vector primers by MWG-Biotech (Ebersberg, Germany). Sequence reads were generated to ≈8-fold coverage and assembled using Sequencher (Gene Codes Corporation, Ann Arbor, MI) followed by visual inspection of chromatograms. The DNA sequence of BAC clone c0156b06 was deposited in the National Center for Biotechnology Information (NCBI) nucleotide database (GenBank) under accession number DQ459071.

BAC Annotation and Sequence Comparisons

Genes were predicted using the Rice Genome Automated Annotation System (http://ricegaas.dna.affrc.go.jp/, verified 5 May 2006). This system utilizes several gene prediction programs including FGENESH (trained with monocot sequences), GENESCAN (trained with *Arabidopsis* or maize sequences), and the Rice Hidden Markov Model (RiceHHM). We annotated genes only where all prediction programs were in agreement. We also queried against the rice genome sequence (www.gramene.org, verified 5 May 2006) and the NCBI protein (nr), nucleotide (nt), and the expressed sequence tag (EST) databases. Predicted gene sequences were considered to be expressed if they were at least 99.8% similar to *S. bicolor* ESTs or EST consensus sequences. This criterion for sequence similarity was determined by pairwise comparisons of nucleotide diversity (π) observed in coding regions of cultivated sorghum (average $\pi = 0.0020$ or about one single nucleotide polymorphism, SNP, every 500 bp) (Hamblin et al., 2006). Repetitive sequences were identified by searching against both the Poaceae RepBase (www.girinst.org, verified 5 May 2006) and the TIGR Gramineae Repeat Database (http://tigrblast.tigr.org/euk-blast/index.cgi?project=osa1, verified 5 May 2006). PipMaker (Schwartz et al., 2000) was used both to align DNA sequences from rice BAC clone OSJNBa0003A09 (GenBank accession AC118132), identified by similarity searches above, and *S. bicolor* BAC c0156b06 and to generate sequence identity and dot plots.

Diversity Analysis

Plant Material

DNA sequences around *Xcup15* were collected from 30 *S. bicolor* accessions including both cultivated (subsp. *bicolor*) ($n = 17$) and wild (subsp. *arundinaceum*) ($n = 13$) lines and a weedy relative, *S. propinquum* (Table 1). These accessions, comprising all *S. bicolor* subspecies and races, were chosen to maximize geographic distribution, morphological variation, and genetic diversity as assessed by variation at 74 SSR loci (Casa et al., 2005). This sampling strategy was devised to minimize the effects of population structure on tests of selection. Seeds from cultivated material (landraces) were

obtained either from the National Center for Genetic Resources Preservation (USDA-ARS, Ft. Collins, CO) or the Plant Genetic Resources Conservation Unit (USDA-ARS, Griffin, GA), and seeds from wild accessions were provided by Mitchell R. Tuinstra (Agronomy Department, Kansas State University). *Sorghum propinquum* leaves were obtained from the Plant Genome Mapping Laboratory (University of Georgia). Information on geographic 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, verified 5 May 2006).

DNA Sequencing and Assembly

Total genomic DNA was isolated from individual seedlings following a standard CTAB extraction protocol and used as template in PCRs following previously established protocols (Casa et al., 2005) except for segments (loci) *9a* and *9b* (see Results and Discussion), where annealing temperature was 62°C. PCR products were prepared for direct sequencing by treatment with exonuclease I (New England Biolabs, Ipswich, MA) and shrimp alkaline phosphatase (Promega Corporation, Madison, WI) following the enzyme manufacturers' instructions. Single-pass sequencing was performed at the Cornell University BioResource Center. Most individuals were homozygous so double-pass sequences were obtained only when putative heterozygotes were encountered. DNA sequences were assembled using Sequencher and alignments were visually inspected and manually edited. Each set of sequence chromatograms was inspected independently by at least two people. DNA sequences were deposited in the NCBI PopSet database under accession numbers DQ462793-DQ463100.

DNA Sequence Analysis

Summary statistics including levels of diversity based on both the average number of nucleotide differences per site between two sequences (π) and number of segregating sites (θ), interspecific divergence, and F_{st}, were calculated using DnaSP v. 4.0 (Rozas et al., 2003). Insertion–deletion polymorphisms were excluded from these analyses. Three statistics were employed to evaluate deviations from the neutral, equilibrium model:

(i) The HKA test (Hudson et al., 1987) was used to compare ratios of polymorphism to divergence for sampled regions assuming a neutral model (i.e., no selection). Each locus was tested against a reference locus comprised of pooled data from 204 loci (Hamblin et al., 2006). For intraspecific polymorphism the following parameters were used: $S = 1075$, $N = 16$, and $L = 138243$, where S is the number of variable sites, N is the sample size, and L is the total number of nucleotide sites surveyed in a sample of cultivated sorghum. For interspecific divergence we used $K = 1948$ and $L = 136626$, where, K is the average number of differences between cultivated *S. bicolor* and *S. propinquum* and L is the number of nucleotide sites evaluated. A Bonferroni correction was applied to account for multiple comparisons.

(ii) Tajima's D (Tajima, 1989) was employed to test for an excess of rare alleles. Following a selective sweep, new mutations arise in the selected region resulting in a skew in the distribution of nucleotide polymorphisms (site frequency spectrum). The population bottleneck associated with sorghum's domestication is, however, expected to affect the site frequency spectrum genome-wide; in particular, the variance of D will be much larger than under a neutral equilibrium model. Critical values of D were obtained from coalescent simulations of a simple bottleneck model that produces the same average number of segregating sites and the same average D as was observed in a genome-wide survey of variation in cultivated sorghum, and in which most of the parameters were estimated based on independent data (Hamblin et al., 2006): the average ancestral population mutation parameter ($4N_e\mu$) was fixed at 3.8 based on variation in wild *S. bicolor*; the population recombination parameter ($4N_e r$) was fixed at 0.01 bp (Hamblin et al., 2005). The time of the bottleneck was $0.025(4N_e)$ generations ago, which would correspond to about 14 000 generations ago if all our assumptions were correct (although this is considerably longer ago than is suggested by archeological data, namely 3000 to 6000 years ago, more recent bottlenecks were incompatible with the observed average value of D). Assuming that the size of the current population and the ancestral population are the same, the intensity of the bottleneck (the size of the bottlenecked population relative to its duration) required to produce the observed value of S was 2.1, equivalent to a 128-fold reduction in population size. The distribution of D values generated by 10 000 simulations of this model had a 95% confidence interval of −1.96, +2.35.

(iii) The CLR test (Kim and Stephan, 2002) was employed for detecting directional selection along a recombining chromosome. This test compares the likelihood of observed patterns of DNA sequence variation under a selective sweep model compared with a neutral equilibrium model of evolution. The CLR test was also used to generate maximum likelihood estimates (MLEs) of the location of the putative selected site (X) and the strength of selection ($\alpha = 2N_e s$). The following parameters were used: NCD = 0 (number of coding regions), $R_n = 0.023$ [scaled recombination rate ($4N_e r$) per nucleotide, where $r = 4 \times 10^{-8}$ (Hamblin et al., 2005) and $4N_e = 570\,000$

(Hamblin et al., 2006)], θ – 1 (Watterson's estimate of θ from data; Watterson, 1978), Nrepl 1 (number of replicates), LBs 1 and RBs 100250 (left and right boundaries on the candidate region where beneficial mutation might be located), and int*X* 1000 (interval between initial guesses of *X*). Recombination rate was assumed constant across the region and θ was estimated from the data in order to make the CLR test conservative (Kim and Stephan, 2002). The frequency of the beneficial allele was set to 1. This method assumes the selected site was fixed very recently. Only accessions for which DNA sequences were available for all loci were included in the analysis (see Table 1). Variable sites were coded as either ancestral (0) (if the nucleotide at the variable position was shared with *S. propinquum*) or derived (1).

Distinguishing between Positive Selection and Demographic Factors

A goodness-of-fit (GOF) test (Jensen et al., 2005) was performed for discriminating whether CLR test rejections were due to selection or to nonequilibrium demographic effects. To determine significance, GOF values obtained from our polymorphism data were compared with those estimated from 1000 data sets simulated under a selection scenario using the maximum likelihood parameter estimates of the location and intensity of selection from the CLR test. In this way, given that the dataset has rejected neutrality in favor of selection, the GOF sets the CLR test selection model as the null and determines whether the sweep model explains the data well, or whether the data simply poorly fit a neutral, equilibrium model.

Results and Discussion

Gene Annotation and Comparative Analysis

To identify functional regions that might be targets of selection, and to obtain sequence information allowing additional polymorphism surveys in the vicinity of *Xcup15*, we identified and sequenced *S. bicolor* BAC clone c0156b06. Within this 112 592 bp clone, 20 complete genes were predicted (Table 2). Three of the predicted genes were associated with *S. bicolor* retrotransposons (either reverse transcriptases or polyproteins). Of the 17 remaining genes, 15 had homologs in rice ($P < e^{-16}$), 13 of which were collinear in a region on rice chromosome 3 (nucleotides 1 991 425 to 2 070 448) (Table 2). Homologous *S. bicolor* transcripts could be identified for 65%

Table 2. Predicted genes within *Sorghum bicolor* BAC clone c0156b06.

Gene	Strand	BAC Coordinates (bp)	SbEST†	Rice locus‡	Rice Ch. 3 Coordinates (bp) ‡	Protein
1	–	740–3 450	TC105178	Os03g04490	2 070 448–2 073 478	cyclin-dependent kinase inhibitor 3
2§	+	7 111–13 234				reverse transciptase, non-LTR retroelement
3	–	27 861–28 529				predicted protein
4	+	30 393–32 521		Os03g04480	2 057 097–2 059 411	corA-like Mg^{++} transporter
5	–	33 200–34 488	TC108263	Os03g04470	2 055 195–2 056 765	expressed protein
6	–	35 308–38 802	TC94332	Os03g04460	2 051 233–2 055 005	voltage-dependent anion channel
7	–	40 757–43 144	TC100511 CD213858	Os03g04450	2 046 824–2 049 512	expressed protein
8	+	48 485–51 284	TC96080	Os03g04440	2 041 085–2 045 015	expressed protein
9	–	52 635–55 405	TC105183	Os03g04430	2 037 231–2 040 749	protein phosphatase 2C
10	+	67 987–68 637				predicted protein
11	+	70 167–71 201				copia polyprotein
12	+	71 455–73 917				copia reverse transcriptase
13	+	78 304–82 596	TC105890			expressed protein, myb/SANT domain
14	+	88 113–95 337	TC96811 TC92295	Os03g04410	2 007 254–2 014 008	aconitate hydratase 1
15	+	99 277–101 549	TC110997 TC100900	Os03g04400	1 999 173–2 002 401	expressed protein, DNAJ domain
16	+	102 196–104 583		Os03g04390	1 996 085–1 998 759	pentatricopeptide repeat (PPR) domain protein
17	–	105 192–105 674		Os03g04380	1 995 394–1 996 038	NADH-ubiquinone oxidoreductase complex I protein (LYR family)
18	+	105 999–106 721	TC108604			expressed protein
19	–	108 378–108 809	TC101567	Os03g04370	1 994 485–1 995 179	expressed protein
20	+	109 434–111 035		Os03g04360	1 991 425–1 993 969	phosphate:H^+ symporter

† Accession numbers for *Sorghum bicolor* transcript consensus sequences (TCs) (TIGR Sorghum Gene Index) and ESTs (NCBI nucleotide database).

‡ Gramene database.

§ Shading indicates *Sorghum bicolor* genes that were not collinear in rice.

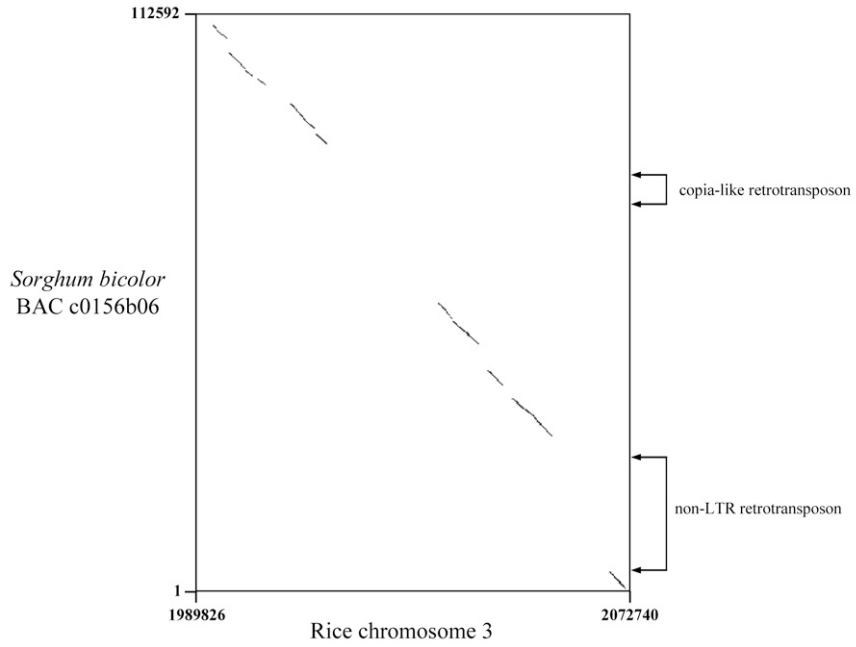

Fig. 1. Dot plot showing areas of sequence similarity between *Sorghum bicolor* BAC c0156b06 and rice chromosome 3. The *x* axis shows the rice genome coordinates (nucleotides), and sorghum coordinates are on the *y* axis. Note that the sorghum BAC sequence is in opposite orientation to the rice sequence (*i.e.*, lower right line in the diagonal corresponds to sorghum gene 1 on Table 2). Arrows on the right side of the figure span the sorghum retrotransposon insertions.

(11/17) of the nontransposable element-related genes. We should note that failure to identify an EST does not necessarily mean that the gene is not expressed. Although there are ≈200 000 *S. bicolor* ESTs in the public domain (www.ncbi.nlm.nih.gov/; verified 5 May 2006), these do not capture all expressed genes.

Comparative sequence analyses indicated that gene order and orientation were well conserved between the 112 592 bp region containing *Xcup*15 (on *S. bicolor* chromosome 1) and an 82 915 bp region on rice chromosome 3 (Fig. 1). Areas of noncollinearity were primarily located in two regions corresponding to the *S. bicolor* retroelements and their associated flanking sequences (Fig. 1).

Sequence Diversity Assessment

To assess diversity in cultivated and wild accessions within the *Xcup15* region, genomic DNA sequences were collected from 10 segments (loci) ranging in size from 240 to 1964 bp and spanning 99 kb centered on *Xcup15* (Supplemental Table 1). Because regions with higher neutral mutation rates provide greater power to detect reductions in variation (and less selective constraint), we assayed mostly intronic and intergenic sequences (79.2% of the ≈7.7 kb of DNA sequence obtained from each individual was from noncoding DNA). In only one instance (locus *7*) was coding sequence solely analyzed. The candidate SSR, *Xcup15*, resides within locus *9b* (Tables 3 and S-1).

Levels of within and between species variation (diversity and divergence, respectively) in the sampled region are shown in Table 3. Cultivated sorghums were invariant at six of the 10 loci and average nucleotide diversity (π) was 0.0008 (range was 0.0–0.0071), a considerably lower estimate than obtained in a previous study of other genomic regions in the same sorghum accessions (average π was 0.0023) (Hamblin et al., 2006). In general, levels of diversity based on the number of segregating sites (θ) were lower than those based on π (Table 3). Notably, locus *9-10b* was unusually diverse. This locus, from an intergenic region rich in miniature inverted repeat transposable elements (MITEs) between the PP2C gene and a predicted protein, accounted for most (≈90%) of the variation detected within cultivated sorghum.

In contrast to the cultivated material, wild accessions were polymorphic at all loci. Average diversity levels based on π (0.0027) and θ (0.0031) were similar and about three times higher than in cultivated sorghum. Accession L-WA15 was heterozygous at three loci and three wild samples had a MITE insertion within locus *9-10a* (data not shown). As observed in cultivated lines, locus *9-10b* exhibited the highest levels of variation (Table 3). Notably, a ≈1 kb transposon-like insertion was observed within locus *9b* (which includes SSR locus *Xcup15*) of *S. propinquum* (outgroup). This insertion was absent in all cultivated and wild accessions.

Diversity and divergence trends for cultivated and wild sorghums within the genomic region containing *Xcup15* are shown in Fig. 2. Directional selection is expected to reduce levels of diversity in cultivated relative to wild sorghum around the selection target. Previous values based on genome-wide estimates of nucleotide diversity have indicated that cultivated accessions exhibit about two-thirds the diversity observed in wild material (Hamblin et al., 2005). In the *Xcup15* region, however, cultivated lines were even less diverse, showing one-third the diversity of wild accessions. The contrast is very striking, however, when polymorphism data for locus *9-10b*, an extreme outlier with similar polymorphism levels in both cul-

Table 3. DNA sequence diversity for loci sampled within the *Xcup*15 region in cultivated and wild *Sorghum bicolor*.

Locus	*N*†	*S*‡	Length§ bp	π (×1000)¶	θ (×1000)¶	Divergence (×100)¶
Cultivated						
1	17	1	830	0.27	0.36	1.09
7	17	0	743	0.00	0.00	0.14
8	17	0	655	0.00	0.00	2.29
9a	17	0	579	0.00	0.00	1.04
9b#	16	0	1963	0.00	0.00	2.75
9–10a	17	1	411	0.29	0.72	1.64
9–10b	17	4	240	7.11	4.93	1.46
13	17	2	859	0.27	0.69	1.05
14	17	0	588	0.00	0.00	1.36
15	17	0	867	0.00	0.00	1.85
Average	16.9	0.8	773	0.79	0.67	1.46
Wild						
1	13	9	825	4.66	3.52	0.64
7	13	2	743	0.59	0.87	0.14
8	14	11	655	4.70	5.28	2.10
9a	13	4	579	1.90	2.23	0.88
9b	12	12	1931	1.72	2.06	2.75
9–10a	14	9	366	4.83	7.73	1.41
9–10b	11	4	240	7.12	5.69	1.50
13	12	1	859	0.35	0.39	1.05
14	14	4	589	1.18	2.14	1.34
15	13	3	867	0.68	1.12	1.85
Average	12.9	5.9	765	2.77	3.10	1.36

† Number of accessions for which complete DNA sequence data were obtained. In some cases, *N* is greater than the number of accessions surveyed because heterozygous accessions were counted as two chromosomes.
‡ Number of segregating sites.
§ Lengths may not match those reported in Table S-1 because insertion–deletion polymorphisms were excluded from this analysis.
¶ The actual numbers were multiplied by this to obtain the reported numbers.
# Locus contains *Xcup15*.

tivated and wild sorghums (see above), were excluded from the analysis. Here, the amount of variation in cultivated sorghum was only 5% of that observed in wild accessions. The magnitude of this reduction in diversity is comparable with that reported for domestication-related genes in maize. In contrast to genome-wide estimates that indicate that maize contains ≈57% of the variability found in its progenitor (Wright et al., 2005), the promoter regions of the maize *teosinte branched1* (*tb1*) (Doebley et al., 1995) and *teosinte glume architecture1* (*tga1*) (Dorweiler et al., 1993) alleles possess 3% (Wang et al., 1999) and 5% (Wang et al., 2005), respectively, of the variation observed in wild relatives, the teosintes. Both *tb1* and *tga1* have been shown to be targets of domestication-related selection in maize (Wang et al., 1999, 2005).

Our data indicate that average nucleotide divergence (1.4%) (Table 3) between cultivated *S. bicolor* and *S. propinquum* in the *Xcup15* region was similar to previous estimates (Hamblin et al., 2004, 2006). Divergence at one locus (*9b*), however, is twice as high (2.8%) (Fig. 2). This locus contains the *Xcup15* SSR and encompasses part of the 5′ UTR and upstream region of the PP2C gene. These patterns probably reflect differences in underlying mutation rates and/or functional constraint on these regions.

F_{st} measures the level of genetic differentiation between populations (here, cultivated and wild sorghums) based on allele frequencies. Under a scenario of directional selection in cultivated sorghum, F_{st} values are expected to be higher at the selection target and adjacent loci, but diminish with distance as recombination prevents the unusual differentiation associated with selection from occurring. Although the average value of F_{st} observed across the entire *Xcup15* region (0.15) is comparable with a previous estimate based on genome-wide SSR data (F_{st} = 0.13) (Casa et al., 2005), loci corresponding to the third intron (locus *9a*) and the 5′ UTR (locus *9b*) of the PP2C gene revealed a considerably greater degree of differentiation (0.52 and 0.46, respectively) (Fig. 2 and Table 4). Thus, the F_{st} analysis suggests that selection may have occurred in or near loci *9a* and *9b* (the PP2C gene).

A Candidate for Directional Selection in Cultivated Sorghum

Several features of the data lend support to the hypothesis that recent directional selection has shaped diversity patterns around *Xcup15*. This assertion is based in part on previous observations on levels of DNA sequence diversity from *S. bicolor* and also polymorphism data obtained from other grass species. First, diversity levels across this region in cultivated accessions were very low, about one-third of previous estimates of genome-wide diversity using the same accessions (Hamblin et al., 2006).

Second, simulation studies have shown that LD increases after a selective sweep (Przeworski, 2002; Kim and Nielsen, 2004). A particular haplotype (extending for at least 99 kb) predominated among cultivated sorghums while wild accessions showed no such haplotype structure (Fig. 3). Previous estimates in sorghum have indicated that LD decays, on average, by 15 kb (Hamblin et al., 2005). Although low levels of polymorphism in the *Xcup15* region precluded our ability to assess LD levels, the haplotype structure in cultivated sorghums was unusual

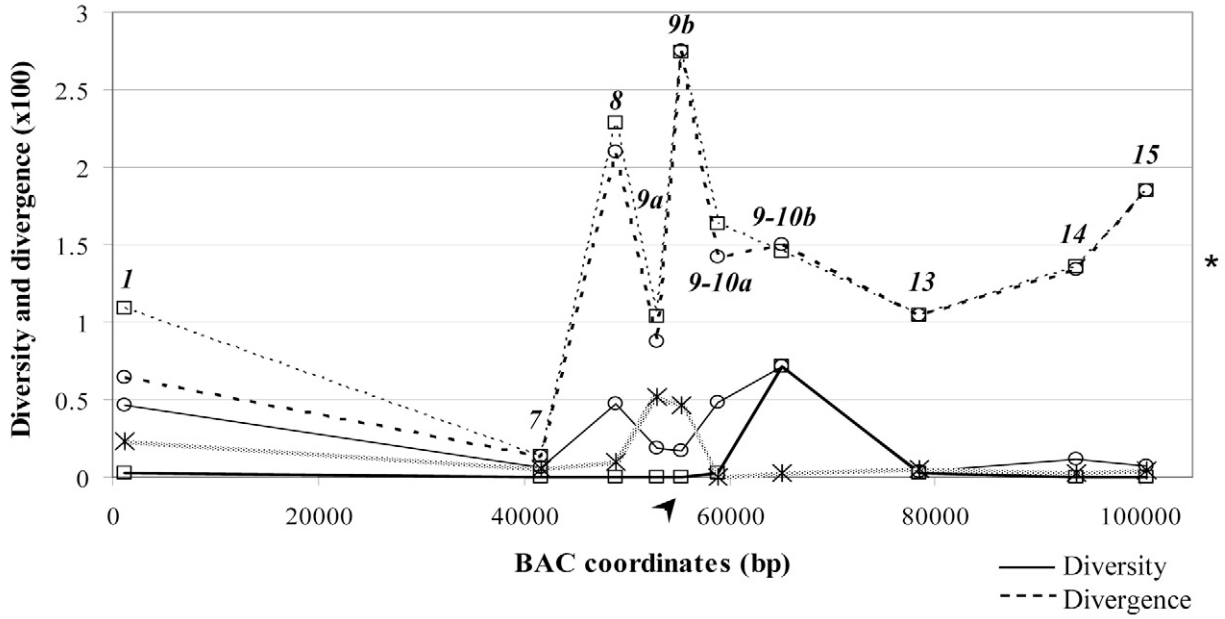

Fig. 2. Diversity, divergence, and population differentiation (F_{st}) for 10 loci in a 99 kb region flanking *Xcup15* in cultivated and wild *Sorghum bicolor* lines. The *y* axis shows levels of diversity (×100) and divergence (×100) and *S. bicolor* BAC c0156b06 coordinates (bp) are on the *x* axis (actual numbers were multiplied by 100 to obtain the reported numbers). Solid and dashed trend lines represent diversity and divergence, respectively. Cultivated and wild sorghum accessions are denoted by open squares and circles, respectively. The solid gray line with asterisks shows F_{st} values, a measure of population differentiation between cultivated and wild lines. Loci sampled are designated by numbers and letters along the lines (see Table 3). The asterisk to the right of the graph denotes average divergence levels between cultivated sorghum and *S. propinquum* based on genome-wide estimates. The arrow head along the *x* axis indicates the approximate location of SSR locus *Xcup15*.

and resembled that observed in swept regions of other species. For example, DNA sequence data from maize, a randomly mating outcrossing species (Brown and Allard, 1970), have suggested that selection produces higher LD. In a survey of six genes (1.2–10 kb in length) in a diverse set of tropical and semitropical lines of maize, Remington et al. (2001) found that LD declined rapidly (within 200–1500 bp) for five genes but that it decayed much more slowly (within ≈10 kb), for sugary 1 (*su1*). Subsequent analysis showed that *su1*, an enzyme in the starch biosynthesis pathway, had been under directional selection during either domestication or breeding (Whitt et al., 2002). Extended LD has been also been detected around the maize allele of the *Y1* gene that encodes for yellow endosperm (Palaisa et al., 2003). In rice, nucleotide diversity data surrounding the *xa5* locus, a bacterial blight resistance gene, showed significant LD between sites 100 kb apart for resistant accessions but no significant association among susceptible types (Garris et al., 2003). Rice, like sorghum, is a predominantly selfing species although outcrossing rates in rice (<1%) (Rong et al., 2004) are much lower than estimates for sorghum (5–30%) (Ollitraut, 1987; Doggett, 1970).

And finally, support for recent selection in the region we have studied in sorghum comes from our identification of a fixed G→A transition between wild (including *S. propinquum*) and cultivated accessions at position 56 122 bp of the BAC clone (corresponding to the 5′ UTR of the PP2C gene) and ≈105 bp upstream of SSR *Xcup15* (Fig. 3). Previous analysis of variation across a total of 23 174 bp (Hamblin et al., 2005) never yielded a fixed difference between DNA sequences from wild and cultivated sorghums. Moreover, DNA sequence alignment of this region to sequences from sugarcane, maize, and rice indicated that these taxa exhibit the same nucleotide (G) observed in the wild sorghums at position 56 122 bp, confirming that the A allele in the cultivated is derived. The serine–threonine phosphatase (PP2C gene) that harbors this fixed transition was most similar to *Arabidopsis thaliana* gene At3g51370 and belongs to one of the largest gene families described in plants. According to Kerk et al. (2002), *Arabidopsis* contains 69 such genes. Moreover, the PP2C *Arabidopsis* homolog is a member of the least studied groups of phosphatases, class *D* (Schweighofer et al., 2004). Serine–threonine phosphatases have been implicated in mechanisms such as abscisic acid (ABA) signal transduction, regulation of flower development (Schweighofer et al., 2004) and seed germination (Yoshida et al., 2006). Two sorghum domestication-related QTLs

co-localize with *Xcup15*, one for plant height (Lin et al., 1995) and the other for primary branch number in the inflorescence (P.J. Brown, 2006, personal communication). Although the prospects are tantalizing, we have no evidence at present that the PP2C gene does or does not influence any of these phenotypes in *S. bicolor*. The high level of LD (haplotype structure) in the cultivated lines should also lead to caution in the acceptance of the PP2C gene as being the actual target of selection without additional functional and/or association studies (see below).

Statistical Evidence for Selection

We employed statistical methods to determine if the patterns of diversity observed in cultivated sorghums in the genomic region surrounding *Xcup15* differed significantly from an equilibrium neutral model and in a manner consistent with a selection scenario in cultivated sorghum. Directional selection (i.e., fixation of a favorable mutation) will result in decreased variation at linked neutral regions and the size of the affected region is a function of both the regional rate of recombination and the strength of selection. To test if differences among loci in the amount of diversity within species relative to divergence between species were significant we employed the HKA test. Because the amount of DNA sequence variation observed within a species (diversity) is expected to be proportional to the amount of DNA sequence divergence between species at neutrally evolving loci (Kimura, 1983), significant differences in these ratios might suggest the local effects of selection. If a particular locus shows a low ratio of diversity to divergence relative to other loci, for example, directional selection may have been responsible for the reduced diversity and the locus possibly encodes or influences a domestication-related trait. Conversely, higher diversity than expected under a neutral evolution model might indicate the effects of balancing or diversifying selection (the locus could be involved in local adaptation or crop improvement). Results from HKA tests for the 10 loci surveyed are presented in Table 4. Among the comparisons performed for cultivated sorghum (each of 10 loci vs. a "reference locus" composed of genome-wide data) (see Materials and Methods), only locus *9b* (the same locus that showed the fixed nucleotide difference between cultivated and wild sorghums) exhibited a significant *P* value (0.0009) after applying the Bonferroni correction. This finding indicates a deficiency of polymorphism in cultivated lines relative to divergence and is consistent with expectations under a model of recent directional selection. None of the HKA tests performed on loci from the wild accessions were significant (Table 4). Although not ideal, comparison of wild data to the cultivated reference locus was carried out due to the lack of an appropriate reference dataset derived exclusively from the wild sorghums, and is conservative for detection of directional selection.

Table 4. F_{st}, HKA test *P* values, and Tajima's *D* for wild and cultivated sorghums.

Locus	F_{st}	HKA *P* Value		Tajima's *D*	
		Cultivated	Wild	Cultivated	Wild
1	0.23	0.146	0.141	–0.49	1.28
7	0.05	0.468	0.307	–	–0.91
8	0.10	0.022	0.542	–	–0.44
9a	0.52	0.101	0.645	–	–0.49
9b	0.46	0.0009†	0.161	–	–0.69
9–10a	0	0.247	0.093	–1.16	–1.44
9–10b	0.03	0.505	0.417	1.35	0.92
13	0.05	0.308	0.191	–1.50	–0.19
14	0.03	0.066	0.943	–	–1.48
15	0.04	0.018	0.209	–	–1.23
Average	0.15			–0.45	–0.47

† Significant locus ($P < 0.01$) after Bonferroni correction.

Another feature of the sequence data that can be used to infer the action of selection is the frequency distribution of polymorphisms. Assuming no recombination, a selective sweep of a new mutation or unique variant eliminates all linked neutral variation. With time, as the population recovers from the sweep, new mutations will accumulate, initially at low frequencies. This skew towards an excess of rare variants is measured by Tajima's *D*, which compares the difference between two measures of diversity, θ_w and θ_π. The θ_w estimate (θ in Table 3) is based on the number of segregating sites and is, therefore, affected mostly by low frequency variants, while θ_π (π in Table 3) is based on average nucleotide diversity and is mostly influenced by intermediate frequency alleles. Because the means of these two estimators are expected to be equal under neutrality (see Fay and Wu, 2005), significantly negative values of *D* are consistent with directional selection whereas significantly positive values are consistent with balancing selection. Results for Tajima's *D* (Table 4) indicated a predominance of low-frequency polymorphisms in both cultivated (average $D = -0.45$) and wild (average $D = -0.47$) sorghums. Although these results are in the direction expected under a directional selection scenario, none of the loci (*D* ranged from –1.50 to +1.35) differed significantly from expectations under either an equilibrium neutral model or a simple bottleneck model. Therefore, the Tajima's *D* results provide no evidence for a recent selective sweep of a single new or unique variant. This result is not surprising considering that the power of this test for detecting a selective sweep is restricted within a fairly narrow time interval following the sweep (Simonsen et al., 1995).

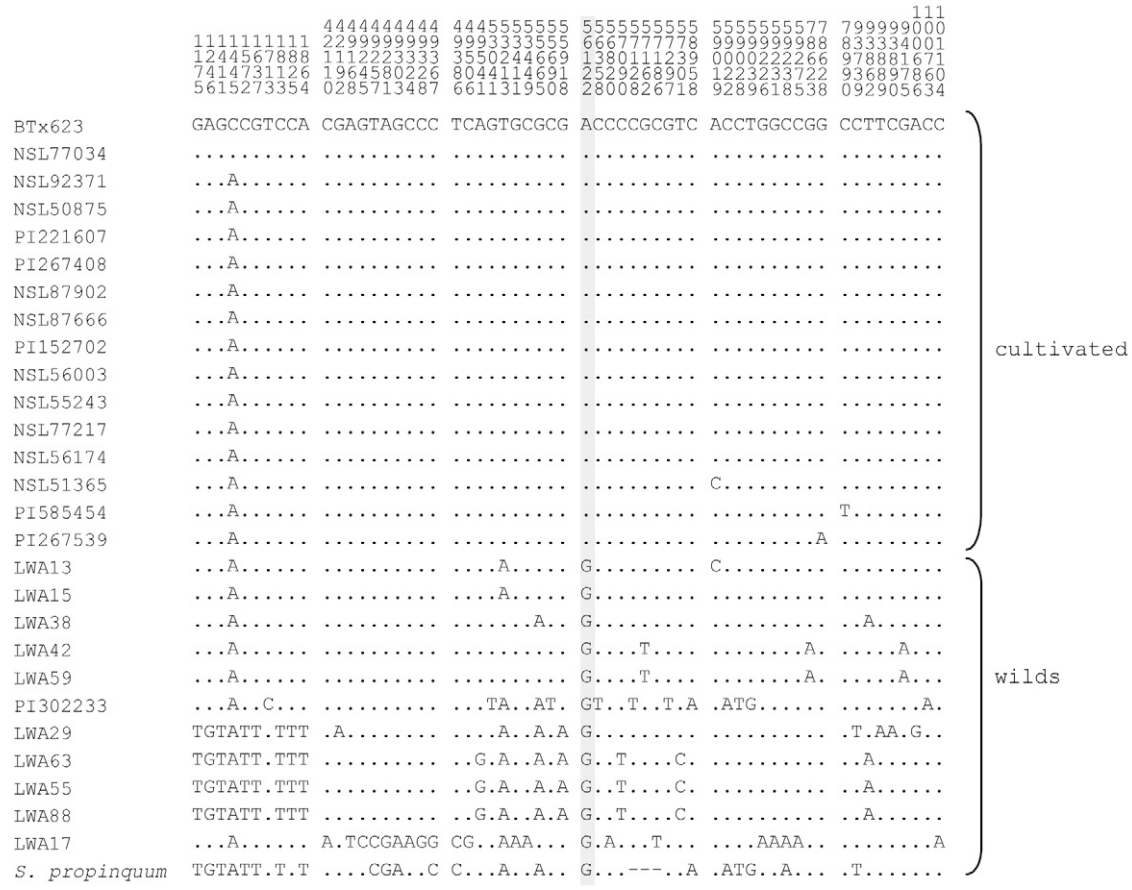

Fig. 3. Haplotypes observed in the *Xcup15* region of cultivated and wild sorghum excluding locus *9–10b*. Numbers across the top of the figure indicate the site coordinate within the BAC sequence (bp), and the shaded area denotes the position of the derived fixed nucleotide difference between cultivated and wild sorghum accessions. Dots indicate sequence identity to reference sequence. Segregating nucleotide sites are shown only for accessions with no missing data. Insertion–deletion polymorphisms were excluded from the analysis. Mutations unique to *S. propinquum* are not shown.

Unlike the previous tests, in which loci are tested individually, the likelihood-based statistical test or CLR evaluates the significance of a local reduction of variation along a physically linked but not necessarily contiguous stretch of DNA (see Materials and Methods). Departure from neutrality is, therefore, tested with sequence data from all loci simultaneously. Moreover, the CLR estimates the strength and location of directional selection from DNA sequence data. We tested polymorphism data for the cultivated and wild groups separately and also for the combined dataset to evaluate species-wide patterns. Results from this composite likelihood analysis rejected the neutral equilibrium model in favor of a strong selective sweep or hitchhiking model (MLE of the strength of selection or $\alpha = 10087$) only in the combined data set. When population size (N_e) is set to 142 500 (see Materials and Methods), the MLE of α suggests a selection coefficient (s) of 0.035. This value of s is similar to those obtained for the *tga1* (s = 0.03–0.04) (Wang et al., 2005) and *tb1* (s = 0.04–0.08) (Wang et al., 1999) genes of maize. As indicated above, both loci have been shown to be targets of domestication-related selection in maize (Wang et al., 1999, 2005). In addition, the CLR test located the target of selection at position 26 107 bp of the BAC clone sequence (between genes 2 and 3) (Table 2) and ≈30 kb upstream of the fixed transition (at 56 122 bp) observed between wild and cultivated sorghums (see above). Except for multiple transposable element-related coding sequences, the region containing the predicted target comprises the longest expanse of DNA containing no predicted genes (Table 2). It is worth noting, however, that simulation studies have recently demonstrated that the MLE of the target of selection is less reliable in partially sequenced regions, having a very large relative mean square error relative to estimates based on complete sequence (J.D. Jensen, 2006, personal communication). In order to quantify this result, 95% confidence intervals were calculated via parametric bootstrap and were seen to encompass ≈39% of the total region, between positions 6487 and 45722. To improve precision of our localization, therefore, we would need

to collect contiguous DNA sequence polymorphism data from across the entire 99 kb sample region (a very significant sequencing effort).

Distinguishing Selection from Demographic Factors

Results from the CLR test indicated that patterns of diversity in this region of the sorghum genome are a better fit to a selective sweep model than to an equilibrium neutral model. This test, however, is not robust to undetected population structure or a recent bottleneck (Jensen et al., 2005), processes that can generate large deviations from equilibrium and patterns of sequence variation that resemble those expected under a selection scenario. For example, an alternative interpretation of the diversity patterns observed for cultivated and wild sorghums (Fig. 2) could involve demographic amplification of ancestral stochastic variation via a population bottleneck associated with cultivation. Alternatively, this pattern could represent a preexisting sweep signal (i.e., selection occurred in the wild sorghums and was amplified in cultivated lines through one or more bottlenecks) (see Pool et al., 2006).

To address this issue, we took the maximum likelihood estimates from the CLR test and employed them in the GOF which has been shown to have high sensitivity for discriminating between a hitchhiking model and nonequilibrium demography (Jensen et al., 2005). Results from the GOF test suggest that the hitchhiking model fits the data poorly ($P = 0.12$; the lower the value the worse the fit) and, therefore, the signal detected by the CLR method can not be distinguished from demography. We should note, however, that other factors might account for the poor fit observed with the GOF test. First, Jensen et al. (2005) have indicated that deviations from a simple selection model (one that assumes a single, recent, and complete sweep) can generate a large Λ_{GOF} (and therefore a small P value), even if selection has taken place. Additionally, joint analysis of the wild and cultivated data artificially created population structure (see F_{st} results, Table 4), which has been shown to lead to false positives with the CLR test (Jensen et al., 2005). Furthermore, the sweep model (Kim and Stephan, 2002) assumes that the data are sampled from a random mating population at equilibrium. Sorghum, however, is a predominantly selfing species and it is not a population in equilibrium (Hamblin et al., 2005, 2006). While the GOF test appears to be robust to violations of a number of these assumptions in *Drosophila* (Jensen et al., 2005), the effects of these violations are as of yet unexplored in a species such as sorghum. Thus, the results of the CLR test should be viewed only as being consistent with, and not evidence for, recent strong selection in this region of the sorghum genome.

Implications for Identifying Targets of Directional Selection

The power to detect directional selection is directly proportional to the amount of within-species diversity. That is, higher levels of variation provide more power for detecting significant reductions in variation likely associated with selection (Wright et al., 2005; Yamasaki et al., 2005; Hamblin et al., 2006). Cultivated sorghum exhibits one-fourth of the amount of genetic variation observed in a comparable sample of geographically and genetically diverse maize landraces (Hamblin et al., 2004, 2005). Therefore, the low levels of diversity observed within sorghum, coupled with the relatively low divergence to the outgroup (*S. propinquum*), represent major factors limiting our ability to unambiguously determine the target of selection in this genomic region.

When employing genome-wide scans of diversity to identify signals of selection, there are both advantages and disadvantages associated with having extensive haplotype structure or LD. For example, species with fairly extensive LD such as rice and sorghum require lower marker density for suitable genome coverage compared with species in which LD decays much more rapidly (e.g., maize). Conversely, extensive haplotype structure also hinders exact localization of the selection target. Because one major haplotype was observed along the 99 kb *Xcup15* region of cultivated sorghum, at least 12 predicted genes (1, 3, 4, 5, 6, 7, 8, 9, 10, 13, 14, and 15; Table 2) should be considered as potential selection candidates. Given that we were unable to establish the precise boundaries of this putative sweep, genes outside this range are possible candidates as well, despite the evidence of a fixed derived mutation at the PP2C noted in the previous section.

The number of genes that one needs to consider as selection candidates will also depend on the interplay between recombination and genome organization. A major difference between the genome organization of maize and sorghum (as well as rice and *Arabidopsis*) is the interspersion patterns of genes and repetitive sequences. Sorghum and rice have very compact genomes (≈772 and 470 Mb, respectively, Arumuganathan and Earle, 1991; Goff et al., 2002) and gene density tends to be high (Goff et al., 2002; Kim et al., 2005). Gene density in maize, on the other hand, is much lower (SanMiguel et al., 1996; Tikhonov et al., 1999), with genes separated by large blocks of highly methylated repetitive elements (Bennetzen et al., 1994) that are recombinationally suppressed. For example, although LD extends for up to 90 kb upstream of *tb1*

(Clark et al., 2004), a gene that has played a major role in the morphological transition from teosinte to maize (Doebley et al., 1995) and is the best documented target of strong directional selection in plants (Wang et al., 1999), *tb1* is the only gene present within this range. The remaining 90 kb upstream region is composed almost entirely of transposable elements.

Implications for Association Studies and Future Directions

Genome-wide scans of diversity performed in highly diverse panels of maize have yielded dozens of candidates associated with domestication and/or crop improvement (Vigouroux et al., 2002; Wright et al., 2005; Yamasaki et al., 2005). The success of population genetics-based approaches in maize, therefore, prompted us to evaluate this methodology applied to a selfing species as a way of identifying targets of directional selection. As this study reveals, DNA sequence polymorphism data support our initial findings based on SSR genome-wide scans of diversity (Casa et al., 2005) that recent directional selection likely shaped diversity patterns around locus *Xcup15*. Thus, as has been shown in maize, population genetics-based approaches can also lead to the successful identification of candidate genomic regions in sorghum. However, the domestication process in sorghum may not have been as simple as it apparently has been in maize (see Matsuoka et al., 2002). While we assume a single, recent, and complete sweep, it is possible that the history of cultivated sorghum was complex and involved multiple domestication events and/or postdomestication gene flow between wild and cultivated sorghum.

This study has also revealed that unambiguous identification of the target of directional selection in sorghum might not be as straight forward as it presumably has been in maize, because of the overall low levels of variation, more extensive LD, and other departures from equilibrium in sorghum (Hamblin et al., 2006). This challenge might also be faced when such studies are conducted in species that exhibit genomic characteristics and mating systems similar to sorghum. As with the genomic signatures of directional selection, we do not really know what the signal of diversifying selection (pertaining to traits such as flowering time, plant height, and disease resistance) will look like in sorghum. From a practical point of view, however, use of directed (i.e., starting from traits of interest instead of random scans of diversity) and integrated approaches (i.e., combining population development, QTL mapping, and assessment of variation in diversity panels) should pave the way for the successful identification of functionally interesting alleles for crop improvement and line development in *S. bicolor*.

Acknowledgments

We thank Hong Sun (Cornell University) for help with collection of DNA sequence data, Baohua Wang (Cornell University) for database assistance, John Bowers (University of Georgia) for the overgo hybridizations, Mitch Tuinstra (Kansas State University) for providing seeds of wild *S. bicolor* accessions, and Gael Pressoir (Cornell University), Amanda Garris (USDA-ARS, Geneva, NY), and two anonymous reviewers for their comments on the manuscript. This work was funded by NSF grant DBI 0115903 (to AHP, CFA, and SK).

References

Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. 2005. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 1(5):e60 10.1371/journal.pgen.0010060.

Arumuganathan, K., and E. Earle. 1991. Estimation of nuclear DNA content of plants by flow cytometry. Plant Mol. Biol. Rep. 9:208–218.

Bennetzen, J.L., K. Schrick, P.S. Springer, W.E. Brown, and P. SanMiguel. 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome 37:565–576.

Bowers, J.E., C. Abbey, S. Anderson, C. Chang, X. Draye, A.H. Hoppe, R. Jessup, C. Lemke, J. Lennington, Z. Li, Y.R. Lin, S.C. Liu, L. Luo, B.S. Marler, R. Ming, S.E. Mitchell, D. Qiang, K. Reischmann, S.R. Schulze, D.N. Skinner, Y.W. Wang, S. Kresovich, K.F. Schertz, and A.H. Paterson. 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, A.D.H. and R.W. Allard. 1970. Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. Genetics 66:133–145.

Casa, A.M., S.E. Mitchell, M.T. Hamblin, H. Sun, J.E. Bowers, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2005. Diversity and selection in sorghum: Simultaneous analyses using simple sequence repeats (SSRs). Theor. Appl. Genet. 111:23–30.

Chittenden, L.M., K.F. Schertz, Y.R. Lin, R.A. Wing, and A.H. Paterson. 1994. A detailed RFLP map of *Sorghum bicolor* × *S. propinquum* suitable for high-density mapping suggests ancestral duplication of sorghum chromosomes or chromosomal segments. Theor. Appl. Genet. 87:925–933.

Clark, R.M., E. Linton, J. Messing, and J. Doebley. 2004. Pattern of diversity in the genomic region near the maize domestication gene, *tb1*. Proc. Natl. Acad. Sci. USA 101:700–707.

Dje, Y., M. Heuertz, C. Lefèbvre, and X. Vekemans. 2000. Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. Theor. Appl. Genet. 100:918–925.

Doebley, J., A. Stec, and C. Gustus. 1995. *teosinte branched1* and the origin of maize: Evidence for epistasis and the evolution of dominance. Genetics 141:333–346.

Doggett, H. 1970. *Sorghum*. Longmans Green and Co., London.

Dorweiler, J., A. Stec, J. Kermicle, and J. Doebley. 1993. *Teosinte glume architecture1*: a genetic locus controlling a key step in maize evolution. Science (Washington, DC) 262:233–235.

FAO. 2004. Production yearbook 2002. No. 56. FAO Statistic Series No. 176. FAO, Rome.

Fay, J.C., and C.-I. Wu. 2005. Detecting hitchhiking from patterns of DNA polymorphism. p. 65–77. *In* D. Nurminsky (ed.) Selective

sweep. Eurekah.com and Kluwer Academic/Plenum Publ., New York.

Garris, A.J., A.M. Casa, and S. Kresovich. 2005. Molecular technologies and their role in maintaining and utilizing genetic resources. p. 757–759. *In* R.M. Goodman (ed.) Encyclopedia of plant and crop science. Marcel Dekker, New York.

Garris, A.J., S.R. McCouch, and S. Kresovich. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). Genetics 165:759–769.

Goff, S.A., D. Ricke, T.-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. Markus Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W. Sun, L. Chen, B. Cooper, S. Park, T.C. Wood, L. Mao, P. Quail, R. Wing, R. Dean, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R.M. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science (Washington, DC) 296:92–100.

Grenier, C., M. Deu, S. Kresovich, P.J. Bramel-Cox, and P. Hamon. 2000. 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.

Hamblin, M.T., A.M. Casa, H. Sun, S.C. Murray, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2006. Challenges of detecting directional selection after a bottleneck: Lessons from *Sorghum bicolor*. Genetics 173:953–964.

Hamblin, M.T., S.E. Mitchell, G.M. White, J. Gallego, R. Kukatla, R.A. Wing, A.H. Paterson, and S. Kresovich. 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.

Hamblin, M.T., M.G. Salas Fernandez, A.M. Casa, S.E. Mitchell, A.H. Paterson, and S. Kresovich. 2005. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171:1247–1256.

Hamrick, J.L., and M.J.W. Godt. 1996. Effects of life history traits on genetic diversity in plant species. Philos. Trans. R. Soc. Lond. B Biol. Sci. 351:1291–1298.

Hudson, R.R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Jensen, J.D., Y. Kim, V.B. DuMont, C.F. Aquadro, and C.D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401–1410.

Kauer, M.O., D. Dieringer, and C. Schlotterer. 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., S. Brauer, and M. Stoneking. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol. Biol. Evol. 20:893–900.

Kerk, D., J. Bulgrien, D.W. Smith, B. Barsam, S. Veretnik, and M. Gribskov. 2002. The complement of protein phosphatase catalytic subunits encoded in the genome of *Arabidopsis*. Plant Physiol. 129:908–925.

Kim, J.-S., M.N. Islam-Faridi, P.E. Klein, D.M. Stelly, H.J. Price, R.R. Klein, and J.E. Mullet. 2005. Comprehensive molecular cytogenetic analysis of sorghum genome architecture: Distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. Genetics 171:1963–1976.

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513–1524.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777.

Kimber, C. 2000. Origins of domesticated sorghum and its early diffusion to India and China. p. 3–98. *In* C.W. Smith and R.A. Frederiksen (ed.) Sorghum. John Wiley, New York.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge, U.K.

Kresovich, S., A.M. Casa, A.J. Garris, S.E. Mitchell, and M.T. Hamblin. 2006. Improving the connection between effective crop conservation and breeding. p. 90–95 *In* K. Lamkey and M. Lee (ed.) Plant breeding: The Arnel R. Hallauer International Symposium. Blackwell Publ., Ames, IA.

Lin, Y.R., K.F. Schertz, and A.H. Paterson. 1995. Comparative analysis of QTLs affecting plant height and maturity across the *Poaceae*, in reference to an interspecific sorghum population. Genetics 141:391–411.

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

Maynard Smith, J., and J. Haigh. 1974. The hitchhiking effect of a favourable gene. Genet. Res. 23:23–35.

Menz, M.A., R.R. Klein, J.E. Mullet, J.A. Obert, N.C. Unruh, and P.E. Klein. 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.

Menz, M.A., R.R. Klein, N.C. Unruh, W.L. Rooney, P.E. Klein, and J.E. Mullet. 2004. Genetic diversity of public inbreds of sorghum determined by mapped AFLP and SSR markers. Crop Sci. 44:1236–1244.

Ollitraut, P. 1987. Evaluation génétique des sorghos cultivés (*Sorghum bicolor* L. Moench) par l'analyse conjointe des diversités enzymatique et morphophysiologique. Univ. of Paris, Orsay, France.

Palaisa, K., M. Morgante, M. Williams, and A. Rafalski. 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15:1795–1806.

Peng, Y., K.F. Schertz, S. Cartinhour, and G.E. Hart. 1999. Comparative genome mapping of *Sorghum bicolor* (L.) Moench using an RFLP map constructed in a population of recombinant inbred lines. Plant Breed. 118:225–235.

Pool, J.E., V. Bauer Dumont, J.L. Mueller, and C.F. Aquadro. 2006. A scan of molecular variation leads to the narrow localization of a selective sweep affecting both Afrotropical and cosmopolitan populations of *Drosophila melanogaster*. Genetics 172:1093–1105.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S.I. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA. 98:11 479–11 484.

Rong, J., H. Xia, Y. Zhu, Y. Wang, and B.-R. Lu. 2004. Asymmetric gene flow between traditional and hybrid rice varieties (*Oryza sativa*) indicated by nuclear simple sequence repeats and implications for germplasm conservation. New Phytol. 163:439–445.

Rozas, J., J.C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497.

SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A.M. Berhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the inter-

genic regions of the maize genome. Science (Washington, DC) 274:765–768.

Schloss, S.J., S.E. Mitchell, G.M. White, R. Kukatla, J.E. Bowers, A.H. Paterson, and S. Kresovich. 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. 2003. Hitchhiking mapping—Functional genomics from the population genetics perspective. Trends Genet. 19:32–38.

Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. Pipmaker—A web server for aligning two genomic DNA sequences. Genome Res. 10:577–586.

Schweighofer, A., H. Hirt, and I. Meskiene. 2004. Plant PP2C phosphatase: Emerging functions in stress signaling. Trends Plant Sci. 9:236–243.

Simonsen, K.L., G.A. Churchill, and C.F. Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141:413–429.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Thornton, K.R., and P. Andolfatto. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172:1607–1619.

Tikhonov, A.P., P.J. SanMiguel, Y. Nakajima, N.M. Gorenstein, J.L. Bennetzen, and Z. Avramova. 1999. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc. Natl. Acad. Sci. USA 96:7409–7414.

Vigouroux, Y., M. McMullen, C.T. Hittinger, K. Houchins, L. Schulz, S. Kresovich, Y. Matsuoka, and J. Doebley. 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.

Wang, H., T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigouroux, M. Faller, K. Bomblies, L. Lukens, and J.F. Doebley. 2005. The origin of the naked grains of maize. Nature (London) 436:714–719.

Wang, R.-L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. Nature (London) 398:236–239.

Watterson, G.A. 1978. The homozygosity test of neutrality. Genetics 88:405–417.

Whitt, S.R., L.M. Wilson, M.I. Tenaillon, B.S. Gaut, and E.S.I. Buckler. 2002. Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. USA. 99:12959–12962.

Wright, S., I. Vroh, S. Schroeder, M. Yamasaki, J. Doebley, M.D. McMullen, and B. Gaut. 2005. The genomic extent of artificial selection in maize. Science (Washington, DC) 308:1310–1314.

Yamasaki, M., M. Tenaillon, I. Vroh, S. Schroeder, H. Sanchez-Villeda, J. Doebley, B. Gaut, and M.D. McMullen. 2005. A large scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17:2859–2872.

Yoshida, T., N. Nishimura, N. Kitahata, T. Kuromori, T. Ito, T. Asami, K. Shinozaki, and T. Hirayama. 2006. *ABA-Hypersensitive Germination3* encodes a protein phosphatase 2C (AtPP2CA) that strongly regulates abscisic acid signaling during germination among Arabidopsis protein phosphatase 2Cs. Plant Physiol. 140:115–126.