Comparative Population Genetics of the Panicoid Grasses: Sequence Polymorphism, Linkage Disequilibrium and Selection in a Diverse Sample of *Sorghum bicolor*

Martha T. Hamblin,* Sharon E. Mitchell,* Gemma M. White,*,[1] Javier Gallego,*,[2] Rakesh Kukatla,* Rod A. Wing,[†,3] Andrew H. Paterson[‡] and Stephen Kresovich*,[4]

**Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, [†]Department of Agronomy, Clemson University, Clemson, South Carolina 29634-0359 and [‡]Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602*

Manuscript received September 25, 2003
Accepted for publication January 28, 2004

ABSTRACT

Levels of genetic variation and linkage disequilibrium (LD) are critical factors in association mapping methods as well as in identification of loci that have been targets of selection. Maize, an outcrosser, has a high level of sequence variation and a limited extent of LD. Sorghum, a closely related but largely self-pollinating panicoid grass, is expected to have higher levels of LD. As a first step in estimation of population genetic parameters in sorghum, we surveyed 27 diverse *S. bicolor* accessions for sequence variation at a total of 29,186 bp in 95 short regions derived from genetically mapped RFLPs located throughout the genome. Consistent with its higher level of inbreeding, the extent of LD is at least severalfold greater in sorghum than in maize. Total sequence variation in sorghum is about fourfold lower than that in maize, while synonymous variation is fivefold lower, suggesting a smaller effective population size in sorghum. Because we surveyed a species-wide sample, the mating system, which primarily affects population-level diversity, may not be primarily responsible for this difference. Comparisons of polymorphism and divergence suggest that both directional and diversifying selection have played important roles in shaping variation in the sorghum genome.

IDENTIFICATION of the genetic variation underlying traits important in domestication and improvement of crops is an area of great interest to both evolutionary and applied biologists. Classical genetic approaches to this problem, such as quantitative trait loci (QTL) mapping, test for an association between a trait and a gene in experimental populations in which the numbers of segregating alleles and meioses are both small. In recent years, methods have been developed that test for such an association in population samples (*i.e.*, groups of unrelated individuals) in which the numbers of alleles and meioses are much larger. Together, these methods provide a strategy for moving from low- to high-resolution mapping of traits, with the ultimate identification of quantitative trait nucleotides (QTNs; Long and Langley 1999).

Characterization of basic population genetic parameters is an essential prerequisite to any approach that analyzes variation in population samples: the power and resolution of haplotype mapping and association studies depend critically on levels of genetic variation, linkage disequilibrium (LD), and population structure. Thus, knowledge of population genetic parameters is a prerequisite to moving beyond mapping in experimental populations. Population genetic analysis can also provide a complementary approach to mapping studies by the identification of loci that have been targets of selection during the process of domestication or crop improvement. These methods can be applied to candidate genes identified through mapping or "reverse" genetics (Wang *et al.* 1999) or used to scan the genome for targets of selection without a prior hypothesis (Vigouroux *et al.* 2002). Tests for evidence of selection can be made only in reference to average genome-wide patterns of neutral variation.

Mating system is an important variable in population genetics: it influences effective population size and effective rate of recombination, which in turn influence levels of genetic variation and linkage disequilibrium (Nordborg and Donnelly 1997). Study organisms that vary in mating system are therefore likely to vary in their suitability for various types of population-based genetic

Sequence data from this article have been deposited in the GenBank Popset library under accession nos. AY234336–AY234362, AY502964–AY504423, AY514060–AY514119, and AY517934–AY518080 and in the GSS library (*S. propinquum* data) under nos. CG993079–CG993165 and CL147585–CL147591.

[1]*Present address:* PIE Department, IACR-Rothamsted, Harpenden, Hertfordshire AL5 2JQ, United Kingdom.

[2]*Present address:* Facultad de Ciencias Experimentales y de la Salud, Sección de Biología Celular y Genética, Universidad San Pablo-CEU, Urb. Montepríncipe, 28668 Madrid, Spain.

[3]*Present address:* Department of Plant Sciences, University of Arizona, Tucson, AZ 85721-0036.

[4]*Corresponding author:* Institute for Genomic Diversity, 157 Biotechnology Bldg., Cornell University, Ithaca, NY 14853.
E-mail: sk20@cornell.edu

analysis. For example, in a species with moderate levels of linkage disequilibrium, haplotype mapping can be accomplished with a reasonable density of markers, but identification of QTNs may not be possible (Nordborg *et al.* 2002; Rafalski 2002). Thus, it may be desirable to exploit closely related species that differ in mating system as a way to move systematically from lower- to higher-resolution analyses.

Maize (*Zea mays* L. ssp. *mays*) and sorghum (*Sorghum bicolor* [L.] Moench) are closely related species that differ dramatically in mating system. Together with pearl millet (*Pennisetum glaucum*), they show considerable synteny in their genomes (Gale and Devos 1998), are expected to share a genetic basis for many agronomically important traits, and can be considered one experimental system of panicoid grass crops. Basic population genetic analyses have shown that maize, an outcrosser, has a very high level of sequence variation and a very limited extent of LD (Remington *et al.* 2001; Tenaillon *et al.* 2001). Because sorghum is largely self-pollinating, it is expected to have higher levels of LD and homozygosity, both of which greatly facilitate LD mapping (Nordborg *et al.* 2002). Furthermore, sorghum may prove to be more tractable than maize for genetic analysis of some phenotypes as its genome is only about one-fourth the size of that of maize, and it has single copies of genes that are duplicated in maize.

As a first step in exploring the merits of sorghum for LD mapping and population genetic analyses, we have assessed sequence variation and LD in 95 short regions (123–444 bp) located throughout the genome, including coding and noncoding sequences. These regions, which correspond to mapped restriction fragment length polymorphism (RFLP) loci, were sequenced in a panel of 27 *S. bicolor* accessions representing elite inbred lines, the five races of *S. bicolor* ssp. *bicolor* (caudatum, durra, bicolor, guinea, and kafir), and three races of *S. bicolor* ssp. *verticilliflorum* (*arundinaceum*, *aethiopicum*, and *verticilliflorum*). Members of this panel display a wide range of geographic and phenotypic diversity. In addition, one accession of *S. propinquum* was sequenced at all loci to serve as an outgroup. Divergence data from *S. propinquum* allow inferences about differences in neutral mutation rates across the genome, and the relationship between polymorphism and divergence allows inferences about the possible role of selection in the evolution of particular loci. Identification of targets of selection may prove valuable in the search for candidate genes underlying important phenotypes.

MATERIALS AND METHODS

Plant material: Accessions and their attributes are listed in Table 1. The three subspecies *verticilliflorum* accessions are wild sorghum; all other *S. bicolor* are cultivated. Five of the *S. bicolor bicolor* accessions were exotic lines that had been converted to day-length insensitivity and short stature by crossing to United States inbred line BTx406 followed by repeated backcrossing to the exotic parent. Leaves from one individual from each accession were harvested for extraction of DNA according to the method of Doyle and Doyle (1987).

RFLP probe sequences and primer development: Sequence information was available for clones of *Pst*I-digested BTx623 genomic DNA ("pSB" clones) that had been developed as RFLP probes (Schloss *et al.* 2002). Our goal was to survey sequence variation at 10 loci for each of the 10 linkage groups for a total of ~100 loci. In anticipation of some failures, 129 mapped RFLP loci were chosen to cover as much of the genome as possible. PCR primers were developed for these loci and tested on a panel of DNAs from four accessions: BTx3197, BTx623, RTx430, and *S. propinquum*. Loci that did not amplify from all four accessions were dropped from the set. Of the 102 successful loci, 96 were chosen for amplification in the larger set of 28 accessions. One locus was found to be duplicated and was discarded.

Sequencing and analysis: PCR products were prepared for sequence analysis by treatment with exonuclease I and shrimp alkaline phosphatase. Cycle sequencing with ABI (Columbia, MD) Big Dye, followed by analysis on an ABI 3700, was performed in the Bioresource Center at Cornell University and at Clemson University. PCR primers were used as sequencing primers. Most PCR products were sequenced with both forward and reverse primers, but in the event that one reaction failed, a single-pass sequence was used.

Chromatograms were assembled into contigs for each locus using both Seqscape (ABI) and Sequencher (Gene Codes, Ann Arbor, MI) software. Our method relied on initial semiautomated identification of variation by Seqscape software (Applied Biosystems) followed by visual inspection and confirmation using Sequencher. Every single-nucleotide polymorphism (SNP) was confirmed by inspection of the chromatograms by at least two different experienced individuals. For purposes of estimating levels of polymorphism on the basis of nucleotide substitution, we removed blocks of three or more contiguous SNPs that were completely associated with each other, since these are likely to arise through insertion/deletion events rather than through nucleotide substitution.

Although sorghum is a predominantly self-pollinating species and therefore usually homozygous at most loci, some heterozygous individuals were observed at eight loci. In these cases, the heterozygous individual was considered to have two chromosomes at that region only. With the exception of LD analysis (see below), the phase of SNPs was unimportant in our analyses. DnaSP version 3 (Rozas and Rozas 1999) was used to calculate diversity and divergence statistics. Insertion/deletion variation was not considered in these analyses.

Each locus was tested for departure from neutrality by the method of Hudson *et al.* (1987) as implemented in Jody Hey's multilocus Hudson-Kreitman-Aguadé (HKA) program (http://lifesci.rutgers.edu/heylab/DistributedProgramsandData.htm#HKA). The simulations were run 10,000 times.

Linkage disequilibrium: The program dipdat (kindly provided by R. R. Hudson) was used to estimate D' and r^2, measures of linkage disequilibrium, as functions of distance. This program uses the maximum-likelihood method of Hill (1974) to estimate these measures from diploid genotype data. Positions at which the rare allele was present in less than three copies were not included in the analysis. For comparisons involving sites within the same locus, distance was measured in base pairs. For comparisons involving sites at different loci, distance was measured in centimorgans as reported by Bowers *et al.* (2003).

Fisher's exact tests of the interlocus comparisons were implemented in DnaSP. Individuals that were heterozygous at more than one site within a linkage group were eliminated from this analysis, as phase in those cases could not be inferred.

Assignment of coding regions: Most of the loci sequenced were anonymous genomic regions. To classify as many sites as possible by functional category, we performed database searches (blastn and blastx) to identify those regions for which there was good evidence of a transcribed open reading frame. The sequence of the surveyed region was submitted to a blastx search against the nonredundant protein database of GenBank using default parameters. Criteria were as follows:

1. If the region showed a 98–100% sequence match to a *S. bicolor* expressed sequence tag (EST) from the CGGC database or a >95% sequence match to a *Z. mays* EST from the Institute for Genomic Research database or GenBank, a score of >50 in a blastx query of the protein database was sufficient to consider it a coding region. Scores only slightly >50 usually represented short stretches of high similarity.
2. In the absence of a strong match in either the sorghum or the maize EST databases, it is still possible that a region encodes a rare transcript. In such cases, a region with a blastx score of >80 was required for the region to be considered coding. In most of these cases, the region also had a strong match with genomic or EST sequence from rice. An exception to this requirement was locus 640, at which polymorphisms were observed more frequently at synonymous sites than if they were occurring at random. In this case, the pattern of polymorphisms provided convincing evidence that the region codes for protein, even though the blastx score was only 75 and there was no good EST match in either maize or sorghum.

RESULTS

Our goal was to characterize levels and patterns of sequence variation across the sorghum genome in a diverse panel of germplasm (Table 1) and to identify regions that appear to depart from average patterns. The final data set represents loci that could be amplified and successfully sequenced in our panel of 27 *S. bicolor* and one *S. propinquum* (see MATERIALS AND METHODS). Not all individuals were successfully amplified or sequenced for all loci, so the sample size varies from locus to locus, averaging 24.7 chromosomes/locus (range is 14–30). The sample size is greater than the number of accessions in a few cases because of the presence of some heterozygous individuals (see MATERIALS AND METHODS). At most loci (87), all individuals were homozygous at all sites. At 8 loci, a few individuals were heterozygous at one or more sites. Accessions BTx406, BTx3197, 152702, 267380, SC0033, and SC0155 were heterozygous at two loci, and accessions 195684, 56174, and LWA4 were heterozygous at 3 loci.

Total sequence diversity in *S. bicolor*: Standard summary statistics of sequence variation for each locus are presented in Table 2, arranged by linkage group; LG designations follow CHITTENDEN *et al.* (1994). It should be noted that our panel of accessions, which includes one individual from each of several populations of two different subspecies, does not represent a sample of individuals randomly chosen from one panmictic population. An important consequence of our sampling is that the variances of statistics of interest are likely to be larger than the standard variances assumed in tests based on these statistics (WAKELEY 1996).

Only base-substitution polymorphisms are included in the statistics reported in Table 2. Although 46 loci had at least one indel, only 26 loci had indel variation polymorphic in *S. bicolor*. Most length variation was found between *S. bicolor* and *S. propinquum* where it sometimes appeared to be complex and difficult to align. A total of 238 SNPs were observed in 29,186 bases surveyed, yielding an average of one SNP every 123 nucleotides in this sample. This is about one-fourth the frequency observed in a comparable sample in maize (TENAILLON *et al.* 2001). The average level of nucleotide diversity, as well as sequence variation based on the number of segregating sites (WATTERSON 1975), is 0.23%, compared to 0.96% in maize (TENAILLON *et al.* 2001). In comparisons to other selfing plants, total sequence variation as well as synonymous site variation in both Arabidopsis (AGUADÉ 2001; SHEPARD and PURUGGANAN 2003) and wild barley (MORRELL *et al.* 2003) worldwide samples is about threefold higher than that of sorghum. In both cases, the higher diversity results from the presence of highly diverged haplotypes at some loci.

If the three wild sorghum accessions are removed from the sample, the number of bases surveyed increases to 29,306 while the number of segregating sites decreases to 198. Nucleotide diversity is reduced only slightly, to 0.21%, because the SNPs unique to the wild accessions are usually singletons. Removal of the wild accessions increases the average D to 0.299, indicating that alleles in cultivated *S. bicolor* tend to be skewed toward intermediate frequency.

Evidence for directional and diversifying selection: Estimates of sequence diversity (π) at individual loci ranged from 0 to 1.5%. Variation in levels of diversity is expected as a consequence of evolutionary variance, sampling variance due to the small number of nucleotides surveyed per locus, and differences in neutral mutation rate among loci. The neutral mutation rate can be estimated by the amount of divergence between species, in this case *S. propinquum*, which varies from 0 to 9.8% and averages ~1.2% (Table 2). Polymorphism and divergence are expected to increase and decrease together across the genome when a changing neutral mutation rate underlies both phenomena, while a dramatic change in the relationship between polymorphism and divergence suggests the local effects of selection. We plotted π and divergence as a function of genetic map position across each linkage group (see Figure 1). These plots illustrate how dramatically the relationship between polymorphism and divergence can change, even at fairly closely linked loci.

To test whether differences in mutation rate alone could account for the observed differences in polymorphism, we employed the method of HUDSON *et al.*

TABLE 1

Accessions and their geographic and racial associations

Accession no.	U.S. source	Origin	Species	Subspecies	Race
NSL 82459	Fort Collins, CO	Cameroon	*bicolor*	*bicolor*	Bicolor
NSL 56003	Fort Collins, CO	Kenya	*bicolor*	*bicolor*	Bicolor
PI 22913	Griffin, GA	China	*bicolor*	*bicolor*	Bicolor
NSL 50875	Fort Collins, CO	Chad	*bicolor*	*bicolor*	Guinea
NSL 92381	Fort Collins, CO	Malawi	*bicolor*	*bicolor*	Guinea
NSL 51030	Fort Collins, CO	Mali	*bicolor*	*bicolor*	Guinea
NSL 102069	Fort Collins, CO	Botswana	*bicolor*	*bicolor*	Durra
PI 246712	Griffin, GA	India	*bicolor*	*bicolor*	Durra
PI 195684	Griffin, GA	Ethiopia	*bicolor*	*bicolor*	Durra
PI 152702	Griffin, GA	Sudan	*bicolor*	*bicolor*	Caudatum
PI 257595	Griffin, GA	Ethiopia	*bicolor*	*bicolor*	Caudatum
PI 514605	Griffin, GA	Senegal	*bicolor*	*bicolor*	Caudatum
NSL 56174	Fort Collins, CO	Ethiopia	*bicolor*	*bicolor*	Kafir
PI 267380	Griffin, GA	Zimbabwe	*bicolor*	*bicolor*	Kafir
NSL 77034	Fort Collins, CO	Uganda	*bicolor*	*bicolor*	Kafir
PI 225905	Griffin, GA	Zambia	*bicolor*	*verticilliflorum*	Arundinaceum
IS14569	Manhattan, KS	Kenya	*bicolor*	*verticilliflorum*	Verticilliflorum
IS14567	Manhattan, KS	Sudan	*bicolor*	*verticilliflorum*	Aethiopicum
PI534163 (C)[a]	Griffin, GA	Sudan	*bicolor*	*bicolor*	Caudatum
SC0326 (C)	Griffin, GA	Ethiopia	*bicolor*	*bicolor*	Caudatum
SC0706 (C)		Sudan	*bicolor*	*bicolor*	Caudatum
PI534132 (C)	Griffin, GA	Ethiopia	*bicolor*	*bicolor*	Durra
PI534155 (C)	Griffin, GA	Ethiopia	*bicolor*	*bicolor*	Durra-bicolor
BTx406 (E)[b]	College Station, TX	USA	*bicolor*	*bicolor*	—
BTx623 (E)	College Station, TX	USA	*bicolor*	*bicolor*	—
BTx3197 (E)	College Station, TX	USA	*bicolor*	*bicolor*	—
RTx430 (E)	College Station, TX	USA	*bicolor*	*bicolor*	—
S. propinquum	Athens, GA	India	*propinquum*	—	—

[a] A converted line (see text).
[b] An elite line.

(1987), known as the HKA test. This test compares polymorphism and divergence at multiple unlinked loci: under neutrality, all loci should be consistent with one estimate of effective population size and divergence time, given a constant neutral mutation rate at each locus. The overall χ^2 statistic for the data set was 145.11, which has a *P*-value of 0.00061, and none of the 10,000 simulations had a χ^2 statistic that high, indicating that selection has altered patterns of polymorphism and divergence in these data. On the other hand, none of the individual cell values had a *P*-value <0.10, so there was not strong evidence that any particular locus had been under selection. In Table 3, we show the 10 loci that had the greatest deviation from expected values (indicated by asterisks in Figure 1). Of these 10 loci, 4 show a deficiency and 6 show an excess of polymorphism relative to divergence, suggesting that both directional and diversifying selection have played a role in sorghum evolution. When the three wild accessions are removed from the analysis, the results change very little (data not shown). We have no information about regional rates of recombination in sorghum, so the contribution of background selection (Charlesworth *et al.* 1993) to reductions in variation cannot be taken into account.

Short-range and long-range linkage disequilibrium: Sorghum is a predominantly self-pollinating species (estimates of outcrossing range from 2 to 35% depending on panicle type; Dje *et al.* 2000; Rooney and Smith 2000) and is therefore expected to show higher levels of LD than outcrossing species like maize (Nordborg 2000), which has a selfing rate of ~10% (Kahler *et al.* 1984). Smaller effective population size, indicated by sorghum's lower level of sequence diversity, will also lead to higher levels of LD. In Figure 2, we show r^2 as a function of distance for comparisons within loci, pooled over the entire data set. A logarithmic trend line fit to the data indicates that average r^2 drops to ~0.5 by 400 bp. For this same set of comparisons, only 29 of 329 $|D'|$ values were <1.0. Since none of the comparisons involve SNPs >400 bp apart, we are unable to estimate the decay of LD over longer intragenic distances. However, even in this limited data set, there is a clear contrast with maize, for which Tenaillon *et al.* (2001) found that r^2 dropped to 0.24 by 200 bp and to 0.15 by 500

TABLE 2

Polymorphism and divergence of 95 loci arranged by linkage group

Locus	Length	*n*	No. of segregating sites	π	*D*	Div
			LG-A			
1718	315	23	0	0.00	[a]	2.84
1432	270	26	1	0.28	−1.156	1.85
0581	341	27	2	0.99	−0.745	0.00
0688	246	27	2	0.60	−1.512	1.22
1433	300	27	0	0.00	[a]	0.33
1421	283	27	0	0.00	[a]	1.42
1233	303	26	1	0.70	−0.311	0.59
1938	279	23	0	0.00	[a]	0.36
0289	400	25	1	1.20	1.347	1.03
1138	422	27	0	0.00	[a]	2.61
			LG-B			
1028	320	27	3	0.69	−1.734	1.83
1675	282	20	0	0.00		2.06
1224	165	24	1	2.61	1.027	0.30
1476	200	30	3	5.22	0.890	1.14
0075	159	25	4	11.22	1.821	0.91
1382	214	23	1	1.88	0.834	2.26
1253	186	28	2	3.80	0.783	0.64
1801	178	23	1	0.49	−1.161	0.56
0774	298	24	0	0.00	[a]	0.34
0669	269	20	0	0.00	[a]	0.00
1944	229	22	1	0.76	−0.641	1.31
			LG-C			
0878	369	15	1	0.36	−1.159	2.17
0897	450	25	1	0.34	−0.698	0.64
1059	337	27	1	0.22	−1.154	1.19
0446	396	25	1	0.20	−1.158	1.26
0851	386	26	3	2.88	1.013	0.36
1760	298	26	0	0.00	[a]	1.34
0874	352	24	0	0.00	[a]	0.49
0033	288	27	7	8.90	0.733	1.15
0062	407	26	6	6.60	2.097	0.40
1777	307	25	0	0.00	[a]	0.00
0088	342	27	1	1.42	1.399	2.67
			LG-D			
1460	237	26	0	0.00	[a]	0.43
0725	320	23	0	0.00	[a]	0.63
0487	376	24	8	3.68	−1.136	1.35
1895	123	27	0	0.00	[a]	0.00
1773	297	27	1	1.06	0.336	1.23
1047	409	27	0	0.00	[a]	0.75
1104	252	25	2	4.13	2.090	1.00
1436	302	26	4	3.39	−0.060	0.91
1310	388	27	4	3.05	0.375	0.68
0747	304	25	1	0.50	−0.698	3.63
0161	279	27	1	0.27	−1.150	3.60
1379	324	26	3	4.79	2.387	0.19
			LG-E			
1654	342	26	1	0.43	−0.714	0.29
0544	200	18	0	0.00	[a]	0.00
1823	309	19	0	0.00	[a]	2.27
1101	313	22	4	4.98	1.180	0.07
1647	292	26	2	2.88	1.291	3.66
1549	221	14	1	2.24	1.121	0.51

(*continued*)

TABLE 2

(Continued)

Locus	Length	*n*	No. of segregating sites	π	*D*	Div
0318	331	19	5	3.96	−0.262	0.36
1396	307	26	1	0.88	0.054	2.40
			LG-F			
0907	410	24	0	0.00	[a]	1.95
0871	247	26	1	0.31	−1.156	0.41
1489	282	26	1	0.27	−1.156	0.36
0455	308	25	2	0.76	−1.214	6.27
1354	187	21	1	1.38	−0.133	0.01
0176	386	27	1	0.19	−1.154	0.78
0193	441	27	4	1.69	−0.742	1.27
1601	217	26	4	9.03	2.324	0.42
1446	320	19	1	0.33	−1.165	2.19
1056	436	20	10	8.09	0.878	0.29
1445	291	21	6	8.08	1.290	2.17
			LG-G			
1057	288	24	2	1.48	−0.444	0.36
0122	303	22	3	3.09	0.352	0.57
1140	332	26	2	2.45	1.178	1.55
1229	370	26	1	1.27	1.303	0.30
1866	265	26	4	1.16	−1.887	1.88
1905	334	25	14	10.90	−0.063	0.96
0347	278	25	0	0.00	[a]	1.44
1505	423	26	6	4.26	0.434	0.53
			LG-H			
0860	367	26	1	0.21	−1.156	3.68
0616	261	26	1	0.29	−1.556	0.38
1464	273	20	2	3.51	1.639	1.29
1248	210	26	1	1.01	−0.311	0.00
0640	375	28	15	15.49	1.677	9.80
1249	347	27	5	1.26	−1.856	0.28
0543	376	26	6	5.81	1.153	1.05
0914	313	23	1	0.28	−1.161	3.83
1218	330	26	13	12.17	0.609	0.88
			LG-I			
0355	341	25	0	0.00	[a]	0.88
0142	422	27	4	4.43	2.117	0.35
0742	385	24	0	0.00	[a]	0.00
1236	321	19	1	0.33	−1.165	1.87
1684	159	29	4	1.73	−1.889	0.53
0703	347	27	1	0.76	0.017	0.29
0027	253	24	0	0.00	[a]	0.40
			LG-J			
0811	213	26	0	0.00	[a]	0.48
0716	359	27	7	5.29	0.136	0.95
0501	173	27	2	3.03	0.022	1.99
0402	345	25	12	8.50	−0.262	0.75
0067	308	27	2	0.48	−1.512	0.63
0540	409	21	3	2.79	0.973	0.53
0639	372	27	12	4.47	−1.555	1.32
1108	297	27	0	0.00	[a]	0.67
Total	29,186		238	2.25	−0.001	1.18

n, sample size (chromosomes); π, nucleotide diversity (Nei 1987) × 1000; *D*, Tajima's *D* (Tajima 1989); Div, net nucleotide divergence (Nei 1987) between *S. bicolor* and *S. propinquum* × 100.

[a] *D* could not be calculated because there was no variation.

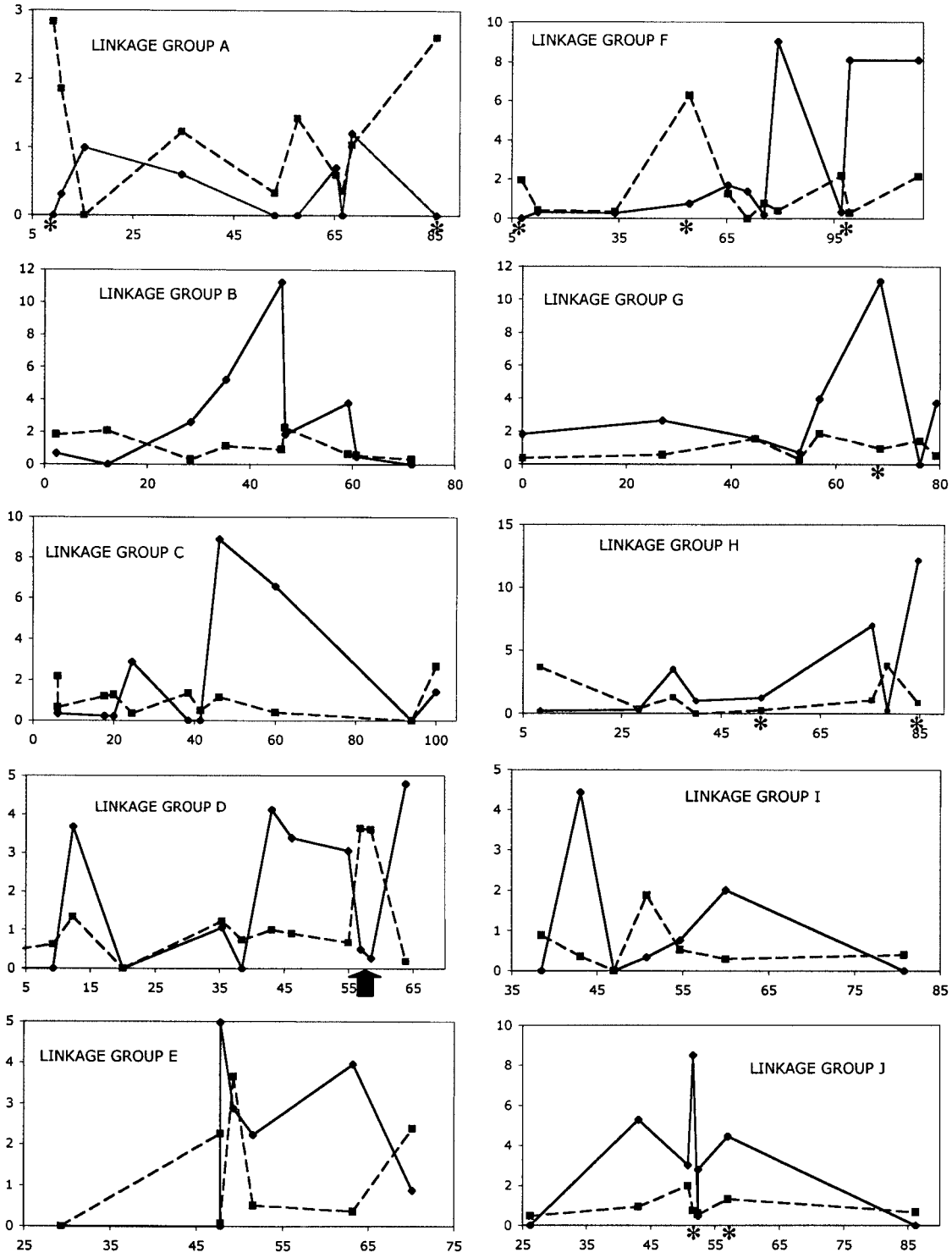

FIGURE 1.—Polymorphism and divergence across each linkage group. The *x*-axis is genetic position in centimorgans. Solid lines with diamonds represent nucleotide diversity within *S. bicolor* multiplied by 1000; the average value is ~2.2 on this scale. Dashed lines with squares represent net divergence between *S. bicolor* and *S. propinquum* multiplied by 100; the average value is ~1.2 on this scale. Locus 640 was removed from the representation of LG H because of its extremely high divergence. Asterisks indicate the positions of loci listed in Table 3. The solid arrow indicates the position of the loci associated with domestication QTL mentioned at the end of the DISCUSSION.

bp. Even in a narrower sample of maize germplasm, where LD is expected to be higher, REMINGTON *et al.* (2001) found that r^2 at five of six genes dropped to between 0.2 and 0.4 by 400 bp. We also looked at the associations between variants at different loci, where distances are measured in centimorgans rather than in base pairs (Table 4). Fisher's exact tests showed that 8.7% of interlocus comparisons were significant at the

TABLE 3

Loci showing an unusual level of variation as assessed by the multilocus HKA test

Locus	Segregating sites observed/expected	Divergence observed/expected	Location[a]	Cell value[b]
1718	0/3.3	8.0/4.7	A (9.2)	4.34
1138	0/4.2	11.0/6.8	A (85.4)	5.17
0907	0/3.0	8.0/5.0	F (7.7)	3.87
0455	2/8.3	19.1/12.9	F (54.7)	4.71
1056	10/4.7	2.8/8.1	F (99.3)	6.71
1905	14/7.4	4.9/11.6	G (68.5)	6.17
1249	5/2.4	1.2/3.8	H (53.1)	3.81
1218	13/6.9	4.9/11.1	H (84.7)	5.73
0402	12/6.1	4.0/9.9	J (51.6)	6.12
0639	12/6.9	5.7/10.9	J (57.0)	4.12

[a] Linkage group assignment (position in centimorgans).
[b] Summed contribution to χ^2 from both polymorphism and divergence; the mean cell value for all loci is 1.53.

0.05 level, in contrast to the 1.5% significant interlocus comparisons found by Tenaillon *et al.* (2001). Thus, in agreement with theoretical predictions, sorghum's selfing behavior and smaller effective population size seem to produce stronger long-distance allelic associations than those of maize.

Variation in protein-coding regions: All loci were analyzed to determine whether there was good evidence that the sequence encodes protein and, if so, to establish the reading frame for codon-based analyses (see materials and methods). Of the 29,186 nucleotides surveyed, 11,025 (38%) from 52 loci were classified as coding sequence. Since the remaining sequence could not be assumed to be noncoding, no analysis was done of noncoding sequence as a functional class. Average nucleotide diversity (π) at synonymous sites and nonsynonymous sites is 0.39 and 0.09%, respectively (estimates for each locus are provided at http://www.genetics.org/supplemental). We also estimated the average levels of θ_W (Watterson 1975) for purposes of comparisons with maize: average θ_W at synonymous sites is 0.34%, compared to 1.73% in maize, while the average level at nonsynonymous sites is 0.09%, compared to 0.39% in maize. The ratio of synonymous to nonsynonymous variation, 3.8, is between that of maize (4.43) and humans (2.65), both of which are smaller than that of Drosophila (8.67; Tenaillon *et al.* 2001).

Both positive and negative selection can alter the ratio of nonsynonymous to synonymous changes. When most variation is neutral, the ratio of synonymous to nonsynonymous mutations is the same within and between species. A departure from this expectation can be detected with a 2 × 2 test of independence (McDonald and Kreitman 1991), although the effects of selection are very hard to detect at individual loci, particularly when the number of nucleotides surveyed is small. How-

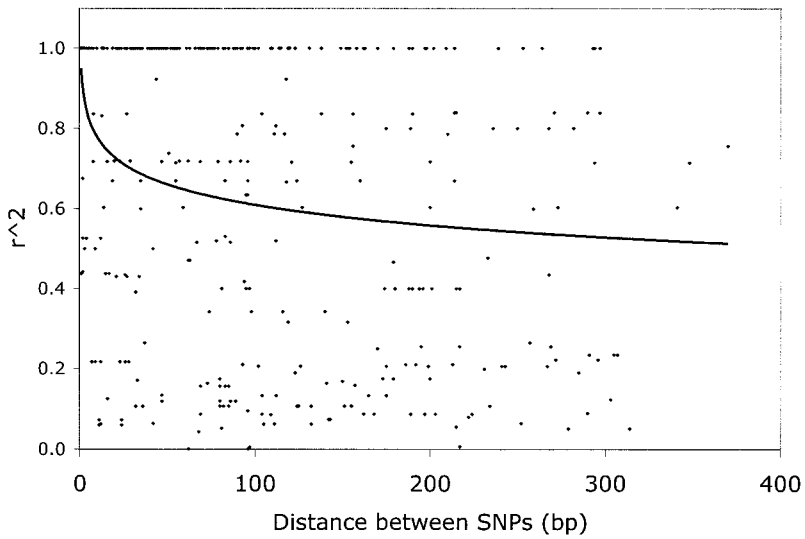

Figure 2.—Linkage disequilibrium (r^2) *vs.* distance within loci. A total of 359 pairwise estimates of r^2 were calculated from 28 loci across the genome (see materials and methods). The line is a logarithmic trend line fit to the data by Microsoft Excel.

TABLE 4

Average LD between loci in the same linkage group as a function of genetic distance

n	Centimorgans apart	D'	r^2
154	0–5	0.601	0.056
216	6–10	0.576	0.042
254	11–20	0.457	0.091
77	21–30	0.493	0.070
192	31–40	0.402	0.070
176	41–50	0.485	0.072
72	>50	0.558	0.063

n, the number of pairwise comparisons. Centimorgans are rounded to the closest integer.

TABLE 5

Polymorphism and divergence of synonymous and nonsynonymous variation

	Pooled (52 loci)		Locus 640		Pooled loci − 640	
	S	*N*	*S*	*N*	*S*	*N*
Within	37	37	5	3	32	34
Between	78	47	12	20	66	27
P-value	0.087		0.250		0.004	

Within, within *S. bicolor*; Between, between *S. bicolor* and *S. propinquum*; *S*, the number of mutations at synonymous sites; *N*, the number of mutations at nonsynonymous sites.

ever, we can test whether genome-wide patterns of variation depart from the neutral expectation. In particular, we were interested in testing for an excess of replacement polymorphisms, as has been observed in several recent studies of variation in humans (Sunyaev *et al.* 2000; Fay *et al.* 2001) and Arabidopsis (Bustamante *et al.* 2002). The data are shown in Table 5: when data from all loci are pooled, there is a trend toward an excess of replacement polymorphisms, and it is close to statistically significant.

One locus, 640, is a clear outlier in this study. This locus putatively encodes a homolog of *Mla1*, a mildew-resistance gene characterized in barley. Disease resistance genes are known to have very rapid rates of evolution and to accumulate amino acid differences at a much higher rate than the average (Bishop *et al.* 2000; Bergelson *et al.* 2001). Indeed, locus 640 accounts for almost half (20/47) of the total amino acid differences observed between *S. bicolor* and *S. propinquum*. When locus 640 is tested alone, the trend is toward an excess of amino acid *fixations*, opposite to that observed in the pooled data. Removal of locus 640 from the pooled data results in a highly significant test statistic indicating that, genome-wide, there are more nonsynonymous polymorphisms in *S. bicolor* than expected.

Relationships among races: *S. bicolor*, having originated in eastern Africa, has been classified into five racial groups on the basis of morphology, and previous studies based on allozyme, RFLP, and simple sequence repeat variation have concluded that both geography and racial structure contribute to the genetic relationships among accessions (Aldrich *et al.* 1992; Deu *et al.* 1994; Dje *et al.* 2000). The extent of genetic divergence among the races, as measured by F_{st}, varies considerably (Table 6). Kafir and durra, which have 29 fixed differences between them and share only 11 of 97 polymorphisms, are the most divergent pair. Part of this divergence can be attributed to the relatively lower variation within these two races compared to the others, as low variation causes an increase in F_{st} (Charlesworth 1998). Of the five *S. bicolor* ssp. *bicolor* races, bicolor is the most variable, consistent with it being the most primitive of the cultivated sorghums (Kimber 2000). The next most variable is caudatum, followed in descending order by guinea, durra, and kafir.

DISCUSSION

The panicoid grass crops provide an opportunity for efficient identification of genetic variation underlying common phenotypes of agronomic interest. Correspondence of QTL locations (Paterson *et al.* 1995) suggests that many such traits may have been subjected to convergent selection in different grasses, so the identification of the underlying gene in one taxon may often account for variation in other related taxa. The suitability of each species for various higher-resolution strategies such as LD mapping, association studies, and screens for targets of selection will depend on its particular level of genetic variation and extent of LD, both of which are affected by mating system. Among the panicoid grasses, these population genetic parameters have previously been estimated only in maize (Remington *et al.* 2001; Tenaillon *et al.* 2001). To provide a similar framework for studies in sorghum, we have surveyed genome-wide sequence variation in a diverse panel of germplasm.

Sequence diversity: This study shows that sorghum has about one-fourth the total variation of maize, from which sorghum is thought to have diverged ~16.5 million years ago (Gaut and Doebley 1997). On the basis of synonymous sites alone, the fraction drops almost to one-fifth. The small discrepancy between total and synonymous variation may result from the different proportions of coding and noncoding sequences included in the sequences used to estimate "total" variation. In addition, levels of total variation may be affected by different patterns of nearly neutral evolution in the two species (see below).

The lower level of variation in sorghum may be due to a number of factors. First, there was a bias away from sequences with higher mutation rates, since 27

TABLE 6

Genetic differentiation between races of *S. bicolor*

	No. of polymorphisms within each race:					
	Caudatum: 105	Kafir: 31	Guinea: 83	Durra[a]: 64	Bicolor: 88	Verticilliflorum: 107
Caudatum	—	0.450	0.173	0.336	0.271	0.261
Kafir	5/14/105	—	0.341	0.665	0.540	0.452
Guinea	1/57/111	1/15/92	—	0.420	0.251	0.251
Durra	6/37/115	29/11/97	15/23/120	—	0.190	0.340
Bicolor	0/41/113	8/17/93	0/41/106	0/23/97	—	0.213
Verticil	1/40/135	7/16/109	0/39/137	6/24/126	2/37/126	—

The top half is F_{ST} calculated according to Hudson *et al.* (1992). The bottom half is the number of fixed differences/number of shared polymorphisms/number of total polymorphisms.

[a] Accession PI 195684 was eliminated from this analysis due to missing data at a large number of loci.

loci (21% of the 129 loci tested) that could not be amplified in *S. propinquum* were dropped from the study. Another possibility is that genome-wide mutation rates in sorghum are lower than those in maize. Considering replication errors alone, the fairly recent common ancestry of maize and sorghum makes this hypothesis implausible. However, the presence of duplicated genes in maize may allow for relaxed constraint and divergent evolution in paralogues, which may increase the neutral mutation rate (Ohta 1993; Clegg *et al.* 1997; Kondrashov *et al.* 2002) although it has no effect on the underlying mutational process.

Since variation is a function of both neutral mutation rate and effective population size, it is likely that sorghum has an effective population size (N_e) considerably smaller than that of maize. To what extent this simply reflects differences in census population size is difficult to say. However, because there is seldom a very good correspondence between census size and effective size, other factors must be considered. A domestication "bottleneck" may have been more severe in sorghum than in maize, which has retained ~70% of the variation present in ancestral teosinte (Eyre-Walker *et al.* 1998; White and Doebley 1999). Our sampling is not adequate to address this question, but other studies that measured allozyme or RFLP variation in larger numbers of accessions estimated that cultivated sorghum retains 60–70% of the variation in its wild relatives (Aldrich *et al.* 1992; Cui *et al.* 1995), similar to the estimate for maize. On the other hand, the average Tajima's *D* in cultivated sorghum (0.299) is considerably higher than that in maize, where it is close to zero (Tenaillon *et al.* 2001), possibly indicating a greater effect of a bottleneck. Population structure may also contribute to the higher *D* statistic.

The effects of self-pollination on population genetics: Another factor affecting the difference in sequence variation between sorghum and maize may be their respective mating systems, specifically, that maize is primarily an outcrosser and sorghum is primarily self-pollinated. There is considerable theoretical work on the effects of self-pollination on population genetics. In a completely self-pollinating species, effective population size, and hence polymorphism, is reduced by half (Pollak 1987). Furthermore, the effective rate of recombination is reduced because most individuals are homozygous at most loci. Background selection caused by the elimination of deleterious alleles therefore has a very important effect (Charlesworth *et al.* 1993), reducing variation as much as 10-fold, depending on the deleterious mutation rate. Hitchhiking effects of directional selection will also be stronger in a self-pollinating species. (All these effects would be intermediate in a partially selfing organism.) Indirectly, mating system may also affect population structure, since selfing species are more likely to be colonizers and may have a more fragmented distribution. The effects of a fragmented population structure are complicated, but can result in smaller effective population size in some situations (Whitlock and Barton 1997; Wakeley and Aliacar 2001).

Several empirical studies have compared patterns of sequence variation in selfing species to those in closely related outcrossing species. In the genus Lycopersicon, Baudry *et al.* (2001) found that two self-compatible tomato species were 4- to 40-fold less variable than the least variable of three self-incompatible species, a far greater difference than could be accounted for by mating system alone. In Leavenworthia (Liu *et al.* 1999), sequence variation at *PgiC* was also greatly reduced in the self-pollinating species. However, at the *Adh* locus in *Arabidopsis lyrata* and *A. thaliana* (Savolainen *et al.* 2000), the results were less clear and depended critically on whether the species were compared on the basis of *within*-population or *across*-population variation. In all three of these studies, each sample was composed of individuals from a single location. Theoretical models that predict reduced N_e in self-pollinators are also based on single population samples. Savolainen *et al.* (2000), who surveyed more than one such population, showed that variation within individual selfing *A. thaliana* popu-

lations is low compared to that in outcrossing *A. lyrata* populations, but that across-population variation is similar in the two species. A larger study by Wright *et al.* (2003) found that within-population variation in *A. lyrata* was 10-fold higher than that in *A. thaliana*, but that species-wide variation in *A. thaliana* was intermediate between that of *A. lyrata petraea* and *A. lyrata lyrata*. Their results suggested that (a) factors other than mating system contribute to the observed differences in variation and that (b) the effects of population subdivision and demographic history make it difficult to infer population genetic parameters from levels and patterns of sequence variation.

The analyses of Wright *et al.* (2003) were based on analyses of variation both within and between populations, so our data are not directly comparable. However, their results suggest that, because our sample includes individuals from many disparate populations, mating system may not be the primary explanation for the lower level of variation in sorghum relative to maize. And while it is reasonable to conclude that a fivefold reduction in synonymous site variation reflects a smaller effective population size, it is not possible to make a quantitative statement about that difference.

While one might expect comparisons to other self-pollinating species to provide some insight on the effect of mating system on levels of sequence variation, the comparisons to wild barley and Arabidopsis, which show severalfold higher levels of species-wide variation, are likely to be confounded by other factors. There are deeply diverged lineages at many loci in both these species, as well as strong geographic structure in barley, suggesting that the population histories of these species are quite different from that of cultivated sorghum. The comparison to maize is more easily interpreted, in that the two species are closely related and both have been domesticated and dispersed by humans within the last 10,000 years.

Linkage disequilibrium: The extent to which linked sites will have a correlated evolutionary history is a function of both effective population size and recombination rate, both of which are affected by mating system (Nordborg 2000), although sorghum's lower N_e may largely be due to other factors. Consistent with the predicted effects of self-pollination and reduced effective population size, sorghum has a greater extent of LD than does maize. Our sequencing strategy did not allow us to plot the decay of LD with physical distance, but short-range intralocus associations are much stronger than those in maize, and significant interlocus associations are severalfold more common. On the other hand, the vast majority of interlocus associations are not significant, and the relationship between polymorphism and divergence changes dramatically at fairly short genetic distances (*e.g.*, Figure 1), suggesting that recombination has decoupled the evolutionary histories of most loci that are not tightly linked. We are currently surveying variation that spans tens of kilobases, and preliminary results suggest that LD dissipates within 10 kb or less (M. T. Hamblin, unpublished data). Thus it appears that the high, but partial, rate of self-pollination in sorghum produces a pattern of LD that is intermediate between that of maize and Arabidopsis (Nordborg *et al.* 2002). It is worth noting, however, that comparison with wild barley also reveals that mating system is not a simple predictor of levels of LD or sequence variation: barley is highly self-pollinating but is more similar in diversity and LD to maize than to sorghum (Morrell *et al.* 2003).

Excess amino acid polymorphism: Effective population size not only determines levels of neutral variation, but also affects patterns of nearly neutral variation, although this process is still not well understood (Ohta 2002). We have found evidence for an excess of amino acid polymorphism in sorghum, a pattern that has also been observed in Arabidopsis (Bustamante *et al.* 2002) and humans (Sunyaev *et al.* 2000; Fay *et al.* 2001). This pattern is thought to be due to the presence of variants that are subject to selection coefficients on the order of the reciprocal of N_e and may explain the difference in ratios of synonymous to nonsynonymous variation in species of different effective size. This ratio is smaller in sorghum than in maize, consistent with this theory. Also consistent is the fact that amino acid polymorphisms in sorghum have a lower average frequency than synonymous polymorphisms: while π for synonymous sites is $>\theta_W$, there is essentially no difference between π and θ_W for nonsynonymous sites.

An alternative explanation is that, in this diverse group of accessions, human selection and/or local adaptation have favored different protein alleles in different environments (see below). Africa, where sorghum diversification occurred, has a particularly wide range of habitats ranging from humid tropics to desert (Kimber 2000), a situation that could produce strong diversifying selection. Association studies with nonsynonymous SNPs could address this interesting possibility. (Note that the racial groups analyzed in Table 6 do not correspond to geographical subpopulations; the durra sample, for example, consists of accessions from India, Ethiopia, and Botswana.)

The effects of selection on sequence variation: Candidate genes for association studies are typically identified through integration of QTL mapping, molecular genetics, and bioinformatics approaches. Population genetic analyses can complement this strategy by identifying regions that have been subject to selection (Vigouroux *et al.* 2002). This approach is likely to be particularly fruitful in crop species, where recent human selection is known to be responsible for much of the useful phenotypic variation.

Selection by humans to improve the agronomic properties of crops is expected to produce characteristic signatures of selection at loci underlying those traits (see, *e.g.*, Wang *et al.* 1999). Genes underlying "domesti-

cation traits," such as the retention of seeds, should show a signature of directional selection, namely a deficiency of variation relative to divergence. We observed several loci in our study that have this signature (Table 3), suggesting that genes in these regions may have been targets of selection. The genomic region affected by a selective event may be relatively larger in sorghum than in maize or other largely outcrossing taxa, due to the reduced effective rate of recombination.

In contrast to targets of directional selection, loci that have responded to selection from local conditions may show an elevated level of diversity in a species-wide sample such as ours, although they might show reduced variation within a local population. Six of the most unusual loci in our HKA tests (Table 3) departed in the direction of excess polymorphism. Of these six, loci 1056, 1218, and 1249 have five, seven, and one nonsynonymous polymorphism(s), respectively, while coding sites were not identified in the other three loci. Interestingly, theoretical work (Nordborg *et al.* 1996) has shown that high rates of selfing increase the signal-to-noise ratio for diversifying selection, making it easier to detect than in outcrossing species.

Our power to detect strong evidence of selection at particular loci in this study is impaired because detection of selection was not the major motivation of the study and the amount of data at any one locus is quite small. None of the departures that we identify in Table 3 is significant; they simply identify candidate regions for further investigation. Conversely, there are regions not highlighted in Table 3 for which independent evidence suggests that they may be associated with phenotypes under selection. On LG D, for example, π at loci 747 (57 cM) and 161 (59 cM) is eightfold less than average, while divergence is more than three times the average (see arrow in Figure 1). These loci are within the likelihood intervals for QTL affecting tillering, regrowth (Paterson *et al.* 1995), and leaf morphology (R. Ming and A. H. Paterson, unpublished results), traits likely to have been under strong directional selection during sorghum domestication.

Conclusions: On the basis of a survey of almost 30,000 sites throughout the genome of *S. bicolor*, we find a frequency of SNPs about one-fourth of that observed in a comparable sample of maize accessions. There is no evidence of a skew to rare alleles; thus many of these SNPs are found in the frequency range useful for LD mapping and association studies. While the high level of intralocus LD in sorghum may prevent phenotypic differences from being attributed to individual sequence variants, interlocus LD does not appear to be so high as to reduce the utility of genome scans. Comparisons of polymorphism and divergence suggest that both directional and diversifying selection have played important roles in the evolutionary history of sorghum and that identification of the targets of that selection may provide important insights into the genetic basis of agronomically important phenotypes in the grasses and grains.

M. Tuinstra, W. Rooney, and G. Peterson provided seed; C. T. Hash provided information about accessions; Maria José Aranzana provided technical assistance; J. Hey provided a program to perform the multilocus HKA test; E. Buckler, P. Morrell, M. Aguadé, and two anonymous reviewers provided comments on the manuscript. Support for this project came from grants DBI-9872649 and 01-15903 from the National Science Foundation to A.H.P. and S.K.

LITERATURE CITED

Aguadé, M., 2001 Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in Arabidopsis thaliana. Mol. Biol. Evol. **18:** 1–9.

Aldrich, P. R., J. Doebley, K. F. Schertz and A. Stec, 1992 Patterns of allozyme variation in cultivated and wild Sorghum bicolor. Theor. Appl. Genet. **85:** 451–460.

Baudry, E., C. Kerdelhue, H. Innan and W. Stephan, 2001 Species and recombination effects on DNA variability in the tomato genus. Genetics **158:** 1725–1735.

Bergelson, J., M. Kreitman, E. A. Stahl and D. Tian, 2001 Evolutionary dynamics of plant R-genes. Science **292:** 2281–2285.

Bishop, J. G., A. M. Dean and T. Mitchell-Olds, 2000 Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc. Natl. Acad. Sci. USA **97:** 5322–5327.

Bowers, J. E., C. Abbey, S. Anderson, C. Chang, X. Draye *et al.*, 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.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature **416:** 531–534.

Charlesworth, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. **15:** 538–543.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

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.

Clegg, M. T., M. P. Cummings and M. L. Durbin, 1997 The evolution of plant nuclear genes. Proc. Natl. Acad. Sci. USA **94:** 7791–7798.

Cui, Y. X., G. W. Xu, C. W. Magill, K. F. Schertz and G. E. Hart, 1995 RFLP-based assay of Sorghum bicolor (L.) Moench genetic diversity. Theor. Appl. Genet. **90:** 787–796.

Deu, M., D. L. D. Gonzalez, J. C. Glaszmann, I. Degremont, J. Chantereau *et al.*, 1994 RFLP diversity in cultivated sorghum in relation to racial differentiation. Theor. Appl. Genet. **88:** 838–844.

Dje, Y., M. Heuertz, C. Lefebvre 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.

Doyle, J. J., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small amounts of leaf tissue. Phytochem. Bull. **19:** 11–15.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. S. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. USA **95:** 4441–4446.

Fay, J. C., G. J. Wyckoff and C.-I Wu, 2001 Positive and negative selection on the human genome. Genetics **158:** 1227–1234.

Gale, M. D., and K. M. Devos, 1998 Plant comparative genetics after 10 years. Science **282:** 656–659.

Gaut, B. S., and J. F. Doebley, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA **94:** 6809–6814.

Hill, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity **33:** 229–239.

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

Hudson, R. R., M. Slatkin and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Kahler, A. L., C. O. Gardner and R. W. Allard, 1984 Nonrandom mating in experimental populations of maize Zea-Mays. Crop Sci. **24:** 350–354.

Kimber, C., 2000 Origins of domesticated sorghum and its early diffusion to India and China, pp. 3–98 in *Sorghum*, edited by C. W. Smith and R. A. Frederiksen. John Wiley & Sons, New York.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf and E. V. Koonin, 2002 Selection in the evolution of gene duplications. Genome Biol. **3:** RESEARCH0008.

Liu, F., D. Charlesworth and M. Kreitman, 1999 The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus Leavenworthia. Genetics **151:** 343–357.

Long, A. D., and C. H. Langley, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. **9:** 720–731.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

Morrell, P. L., K. E. Lundy and M. T. Clegg, 2003 Distinct geographic patterns of genetic diversity are maintained in wild barley (Hordeum vulgare ssp. spontaneum) despite migration. Proc. Natl. Acad. Sci. USA **100:** 10812–10817.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nordborg, M., 2000 Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics **154:** 923–929.

Nordborg, M., and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146:** 1185–1195.

Nordborg, M., B. B. Charlesworth and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond. Ser. B Biol. Sci. **263:** 1033–1039.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory *et al.*, 2002 The extent of linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. **30:** 190–193.

Ohta, T., 1993 Pattern of nucleotide substitutions in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics **134:** 1271–1276.

Ohta, T., 2002 Near-neutrality in evolution of genes and gene regulation. Proc. Natl. Acad. Sci. USA **99:** 16134–16137.

Paterson, A. H., Y. R. Lin, Z. Li, K. F. Schertz, J. F. Doebley *et al.*, 1995 Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science **269:** 1714–1718.

Pollak, E., 1987 On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117:** 353–360.

Rafalski, A., 2002 Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. **5:** 94–100.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA **98:** 11479–11484.

Rooney, W. L., and C. W. Smith, 2000 Techniques for developing new cultivars, pp. 329–347 in *Sorghum*, edited by C. W. Smith and R. A. Frederiksen. John Wiley & Sons, New York.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Savolainen, O., C. H. Langley, B. P. Lazzaro and H. Freville, 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing Arabidopsis lyrata and the selfing Arabidopsis thaliana. Mol. Biol. Evol. **17:** 645–655.

Schloss, S. J., S. E. Mitchell, G. M. White, R. Kukatla, J. E. Bowers *et al.*, 2002 Characterization of RFLP probe sequences for gene discovery and SSR development in Sorghum bicolor (L.) Moench. Theor. Appl. Genet. **105:** 912–920.

Shepard, K. A., and M. D. Purugganan, 2003 Molecular population genetics of the Arabidopsis CLAVATA2 region: the genomic scale of variation and selection in a selfing species. Genetics **163:** 1083–1095.

Sunyaev, S. R., W. C. Lathe, III, V. E. Ramensky and P. Bork, 2000 SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. Trends Genet. **16:** 335–337.

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

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proc. Natl. Acad. Sci. USA **98:** 9161–9166.

Vigouroux, Y., M. McMullen, C. T. Hittinger, K. Houchins, L. Schulz *et al.*, 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.

Wakeley, J., 1996 The variance of pairwise nucleotide differences in two populations with migration. Theor. Popul. Biol. **49:** 39–57.

Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905.

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

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

White, S. E., and J. F. Doebley, 1999 The molecular evolution of terminal *ear1*, a regulatory gene in the genus Zea. Genetics **153:** 1455–1462.

Whitlock, M. C., and N. H. Barton, 1997 The effective size of a subdivided population. Genetics **146:** 427–441.

Wright, S. I., B. Lauga and D. Charlesworth, 2003 Subdivision and haplotype structure in natural populations of Arabidopsis lyrata. Mol. Ecol. **12:** 1247–1263.

Communicating editor: M. Aguadé