DOI: 10.1534/genetics.105.054312

Challenges of Detecting Directional Selection After a Bottleneck: Lessons From *Sorghum bicolor*

Martha T. Hamblin,*,[1,2] Alexandra M. Casa,*,[1] Hong Sun,* Seth C. Murray,* Andrew H. Paterson,[†] Charles F. Aquadro[‡] and Stephen Kresovich*

**Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, [†]Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602 and [‡]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853*

Manuscript received December 5, 2005
Accepted for publication March 13, 2006

ABSTRACT

Multilocus surveys of sequence variation can be used to identify targets of directional selection, which are expected to have reduced levels of variation. Following a population bottleneck, the signal of directional selection may be hard to detect because many loci may have low variation by chance and the frequency spectrum of variation may be perturbed in ways that resemble the effects of selection. Cultivated *Sorghum bicolor* contains a subset of the genetic diversity found in its wild ancestor(s) due to the combined effects of a domestication bottleneck and human selection on traits associated with agriculture. As a framework for distinguishing between the effects of demography and selection, we sequenced 204 loci in a diverse panel of 17 cultivated *S. bicolor* accessions. Genomewide patterns of diversity depart strongly from equilibrium expectations with regard to the variance of the number of segregating sites, the site frequency spectrum, and haplotype configuration. Furthermore, gene genealogies of most loci with an excess of low frequency variants and/or an excess of segregating sites do not show the characteristic signatures of directional and diversifying selection, respectively. A simple bottleneck model provides an improved but inadequate fit to the data, suggesting the action of other population-level factors, such as population structure and migration. Despite a known history of recent selection, we find little evidence for directional selection, likely due to low statistical power and lack of an appropriate null model.

MULTILOCUS surveys of sequence variation can, in principle, be used to identify targets of selection, since neutral loci are all consistent with a common set of population parameters, while recently selected loci are not (reviewed in Schlotterer 2003 and Storz 2005). A frequent goal of such studies is the identification of targets of adaptive evolution in derived populations that have recently experienced a change of environment and are typically not at equilibrium. Unfortunately, selection in the context of a genomewide departure from equilibrium may be hard to detect because many loci may have low variation by chance and the frequency spectrum of variation may be perturbed in ways that resemble the effects of selection. In these cases, it may be possible to define an alternative nonequilibrium model that describes patterns of variation at neutral loci; outliers in this model are inferred to have experienced selection (Ometto *et al.* 2005; Stajich and Hahn 2005; Thornton and Andolfatto 2006). Other approaches to this problem are to compare variation in the derived population with that of the putatively ancestral one (Schlotterer 2002; Glinka *et al.* 2003) or to consider outliers in the empirical distribution as candidate targets of selection (Schmid *et al.* 2005).

Domesticated species are a special case of derived populations: their population genetic characteristics are a complicated product of the characteristics of the ancestral population modified by demographic events, such as bottlenecks, migration, and nonrandom mating, and by varying degrees of selection on genes underlying traits important to farmers and breeders. Selection during domestication may target alleles that are neutral in the wild ancestor and are segregating at moderate frequencies; in such cases, predictions of simple models of selection on new mutations will not be met (Orr and Betancourt 2001; Innan and Kim 2004; Przeworski *et al.* 2005). Domesticated plants may also experience introgression to/from wild relatives due to the lack of reproductive isolation between wild ancestors and their very recently derived (and abundant) cultivated descendants, resulting in cultivated individuals that carry wild alleles and outgroup taxa that carry cultivated alleles. Successful use of genomewide scans to identify domestication genes has been largely limited to maize, where patterns of variation at both SSRs and SNPs have revealed loci that appear to have lost more variation than can be accounted for by the domestication

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. DQ427111–DQ430705

[1]These authors contributed equally to this work.

[2]*Corresponding author:* IGD, 156 Biotechnology Bldg., Cornell University, Ithaca, NY 14853. E-mail: mth3@cornell.edu

bottleneck (Vigouroux *et al.* 2002; Tenaillon *et al.* 2004; Wright *et al.* 2005; Yamasaki *et al.* 2005).

We are studying the population genetics of the cultivated tropical grass *Sorghum bicolor*, which was most probably domesticated in eastern Africa 3000–6000 years ago, subsequently spread to the entire African continent, and reached Asia during the first millennium (Kimber 2000). Much more recently (mid-nineteenth century), sorghum was brought to the United States. There is a great deal of morphological diversity in cultivated sorghum, in characters such as seed size and color, panicle architecture, and height. Furthermore, because sorghum is grown in environments that differ dramatically from one another (*e.g.*, in rainfall, temperature, soil type, and elevation), there must be physiological diversity as well (Doggett 1988). We are interested in characterizing the patterns of genetic variation in worldwide samples of grain sorghum, with the goal of identifying loci that are important in domestication, local adaptation, and agronomic performance. In a previous study of genomewide patterns of sequence variation (Hamblin *et al.* 2004), 95 loci of average length 275 bp were surveyed. In this study, we resequenced 204 additional loci of average length 671 bp distributed throughout the genome in a sample of 16 cultivated grain sorghum landraces (*i.e.*, not elite, modern cultivars) as well as the reference elite cultivar BTx623 and an outgroup, *S. propinquum*. We characterize the overall patterns of sequence variation, compare those patterns to the predictions of some simple bottleneck models, and test for evidence of both directional and diversifying selection.

MATERIALS AND METHODS

Plant material: Genomewide diversity was assessed in 17 *S. bicolor* accessions (16 landraces and BTx623) and one individual of a wild relative, *S. propinquum* (Table 1). These accessions composed all racial types of cultivated grain sorghum and represented a wide geographic sampling from the species' center of diversity (Africa). Seeds of the cultivated material (landraces) were obtained from either the National Center for Genetic Resources Preservation [United States Department of Agriculture (USDA)/Agricultural Research Service (ARS), Fort Collins, CO] or the Plant Genetic Resources Conservation Unit (USDA/ARS, Griffin, GA) and information on geographical origin and racial classification was gathered primarily from the System-wide Information Network for Genetic Resources database (http://singer.cgiar.org/Search/SINGER/search.htm). BTx623 seeds were kindly provided by William Rooney (Texas A&M University).

Loci and primer design: We tested a total of 293 genetically mapped loci. These comprised sequenced genomic RFLP probes and cDNA clones from several species including sorghum ($n = 160$), johnsongrass (68), sugarcane (14), maize (21), buffelgrass (8), barley (2), oat (3), wheat (1), rice (13), and Arabidopsis (3) (locus name, map positions, GenBank accession numbers, and locus origin can be found at http://igd.tc.cornell.edu). Loci derived from species other than *S. bicolor* were analyzed only if sorghum homologs could be identified through database searches.

When possible, we extended the length of the original DNA sequence of each locus through iterative searches against the GenBank GSS and EST databases. Primers were designed from these contigs so that in some cases primers amplified regions flanking the original locus. Loci that could be successfully amplified from BTx623 and *S. propinquum* DNA were then amplified from the panel of 16 sorghum accessions.

Sequencing and analysis: Total genomic DNA was isolated from individual seedlings following a standard CTAB extraction protocol (Doyle and Doyle 1987) and used as template in PCRs following previously established protocols (Casa *et al.* 2005). PCR products were prepared for direct sequencing by treatment with exonuclease I (New England Biolabs) and shrimp alkaline phosphatase (Promega) following the manufacturers' instructions. Single-pass sequencing was performed at the BioResource Center (Cornell University) using a single PCR primer, as most individuals were homozygous. Double-pass sequences were obtained only when putative heterozygotes were observed. Chromatograms were assembled into contigs using Sequencher (Gene Codes, Ann Arbor, MI) software. Alignments were then visually inspected and manually edited. Each set of sequence chromatograms was inspected independently by two or three people.

Summary statistics of diversity and divergence were obtained with DnaSP version 4.0 (Rozas *et al.* 2003). Blocks of three or more contiguous SNPs were disregarded. Insertion/deletion polymorphisms were not considered in either the diversity or divergence analyses. Loci were tested for departure from neutrality using the method of Hudson *et al.* (1987) implemented by the multilocus HKA software (Jody Hey, available at http://lifesci.rutgers.edu/~heylab/index.html). One locus, pSB1812, was not tested with all 203 other loci because it caused some sort of (unknown) incompatibility in the simulations. When tested with a subset of the data, this locus showed no unusual pattern of polymorphism and divergence. Significance of individual loci was assessed by removing the most significant locus and testing the remaining $n - 1$ loci iteratively until no significant locus was detected. The HKA program was also used to test for significance of average D (Tajima 1989) under the standard neutral model and to estimate θ ($4N_e\mu$) on the basis of both polymorphism and divergence.

Assignment of coding regions: To determine whether sequenced regions corresponded to protein-coding loci, consensus sequences obtained from aligned loci were used in database searches (blastn and blastx) using GenBank default parameters. For classifying a region as an open reading frame, we used the criteria described by Hamblin *et al.* (2004).

Simulations: Models of population history were simulated using the program ms (Hudson 2002) with the following assumptions: (1) ancestral θ ($4N_e\mu$) is the same as θ in the wild population today and is estimated to be 0.0057/bp on the basis of variation at 24 loci in 4–26 accessions of *S. bicolor* ssp. *verticilliflorum* (A. Zamora and our unpublished data); (2) θ at individual loci as estimated by the program HKA (see above) was scaled relative to the ancestral θ; (3) ancestral N_e is calculated as follows: $4N_e\mu = 0.0057$; therefore $N_e = 0.0057/4/1 \times 10^{-8} = 1.43 \times 10^5$, where a neutral mutation rate of 1×10^{-8}/bp/generation is based on sequence divergence at 11 loci between maize and sorghum (Swigonova *et al.* 2004); (4) the time of domestication is assumed to be in the range of 3000–6000 years ago, on the basis of archeological evidence, roughly equivalent to 0.005–0.01 when scaled in $4N_e$ generations, at 1 generation/year; and (5) $4N_er$ is estimated as $4 \times (1.43 \times 10^5) \times (4 \times 10^{-8}$ crossovers/bp/generation) $\times$ locus length $\times$ 0.46, where $0.46 = (1 - F)$ to account for a self-pollination rate of 0.7 (Hamblin *et al.* 2005). A recent analysis of genetic *vs.* physical distance on a chromosomal scale came to an identical estimate of average r (0.25 Mb/cM) for euchromatic regions

(Kim *et al.* 2005). We simulated a small set of models where all parameters were fixed except for the size of the bottlenecked population, which was fit to the observed value of average *S*, and tested whether other summary statistics produced by the model were consistent with the data.

Tests of haplotype number: The probability of observing *K* haplotypes given an observed number of segregating sites, *S*, was obtained using the program haploconfig (Innan *et al.* 2005). This program generates gene genealogies on the basis of an input value (or range) of θ and accepts only those that have the observed number of segregating sites. Using the overall average value of θ for this data set (1.5), large numbers of segregating sites are very unlikely to be observed and the acceptance rate becomes unreasonably low. We therefore performed simulations where θ was chosen to maximize the acceptance rate. For $S < 18$, the acceptance rate was ≤5%, but for large values of θ, the probability of observing any particular value of *S* becomes quite small even when θ is optimized. These simulations were performed (1) without recombination, which is very conservative, and (2) with $4N_e r = 5$, which is slightly conservative on the basis of empirical estimates of crossing over (see *Simulations*).

RESULTS

The goal of this study was to characterize genomewide patterns of sequence diversity for cultivated sorghum as a framework for distinguishing between demographic and selective factors as causes of patterns observed at individual loci. PCR primers were designed for 293 genetically mapped loci, distributed throughout the genome (Bowers *et al.* 2003), and tested in BTx623 and *S. propinquum*. Of the 293 primer pairs tested, 89 were discarded due to either failed amplification or amplification of multiple products. Our final data set, therefore, consisted of 204 loci, for an average spacing of 5.2 cM (range 0–30 cM) between loci. These loci, of average size 671 bp, were sequenced in a panel of 17 cultivated accessions (Table 1), selected to represent a morphologically, geographically, and genetically diverse subset of 73 lines previously assessed with simple sequence repeats (SSRs) (Casa *et al.* 2005). In addition, all loci were sequenced in one accession of *S. propinquum*. While most individuals were homozygous at most loci, as expected for a species with a high rate of self-pollination (Doggett 1988), a total of 56 heterozygous genotypes were identified at 46 loci. A majority of these instances of heterozygosity were observed in one individual, accession NSL87902 (*S. bicolor* race durra from Cameroon), which exhibited two alleles at 39 loci; in three instances (on chromosomes 5, 6, and 7) blocks of 3 or more consecutive loci were heterozygous, spanning 5.4–13.8 cM, suggesting a fairly recent history of outcrossing in this accession.

TABLE 1

Sorghum accessions evaluated in this study

Accession ID[a]	ICRISAT ID[b]	Species	Racial type	Country of origin
BTx623	—	*S. bicolor*	—	United States
NSL56003	IS8822	*S. bicolor*	Bicolor	Kenya
NSL77217	IS10747	*S. bicolor*	Bicolor	Chad
NSL92371	IS14318	*S. bicolor*	Bicolor	Swaziland
PI585454	IS25061	*S. bicolor*	Bicolor	Ghana
PI152702	IS12568	*S. bicolor*	Caudatum	Sudan
PI221607	IS2361	*S. bicolor*	Caudatum	Nigeria
NSL87666	IS7115	*S. bicolor*	Caudatum	Central African Republic
NSL55243	IS917	*S. bicolor*	Durra	Algeria
NSL87902	IS14790	*S. bicolor*	Durra	Cameroon
NSL50875	IS7171	*S. bicolor*	Guinea	Chad
PI267408	IS2724	*S. bicolor*	Guinea	Uganda
NSL51365	IS6272	*S. bicolor*	Guinea	India
NSL51030	IS3817	*S. bicolor*	Guinea	Mali
PI267539	IS2901	*S. bicolor*	Kafir	India
NSL56174	IS8539	*S. bicolor*	Kafir	Ethiopia
NSL77034	IS10400	*S. bicolor*	Kafir	Uganda
KFS1		*S. propinquum*		India

[a] Accession number. NSL, National Seed Storage Laboratory; PI, Plant Introduction.

[b] International Crops Research Institute for the Semi-Arid Tropics.

Patterns of sequence diversity across the genome: Approximately 138 kb of DNA sequence (alignment gaps excluded) were surveyed per individual. Forty-three loci (~21%) were invariant within cultivated *S. bicolor*. One of these loci was also invariant across species. Levels of polymorphism and divergence at individual loci are presented as supplemental data (Table S1 at http://www.genetics.org/supplemental/). Table 2 presents summary statistics from this study as well as those from our previous study (Hamblin *et al.* 2004). Levels of nucleotide diversity and divergence are consistent with the previous estimates based on a different set of loci in a different species-wide sample that was also chosen to capture racial and geographic diversity, but without prior knowledge of genetic relationships.

A total of 324 indels were identified in 96 of the 204 loci evaluated within *S. bicolor* and ranged from 1 to

TABLE 2

Average summary statistics compared with previous study

Study	No. of loci	*N*	Length (bp)	*S*, θ_w (×1000)	Var(*S*)	π (×1000)	*D*	Var(*D*)	D_a (%)
Hamblin *et al.* (2004)	95	22	308	2.1, 1.8	8.3	2.1	0.29	1.33	1.1
This study	204	16	671	5.0, 2.2	42.0	2.2	−0.08	1.40	1.3

D_a, net divergence (Nei 1987); *S*, number of segregating sites; θ_w, Watterson's (1975) estimator of $4N_e\mu$/bp; π, nucleotide diversity.

TABLE 3

Average summary statistics for transcribed regions

Functional category	No. of loci	No. bases sampled	π (×1000)	θ_w (×1000)
Synonymous	155	11,140	3.8	3.5
Nonsynonymous	155	35,380	0.6	0.4
Intron	113	36,194	2.4	2.3

π, nucleotide diversity; θ_w, Watterson's (1975) estimator of $4N_e\mu$/bp.

15 per locus. In most cases (306/324, ~95%), length polymorphisms were short (<20 bases). This proportion is similar to that observed in maize, where 92% of nonmicrosatellite indels were <20 bp in length (Tenaillon *et al.* 2002). DNA secondary structure prediction programs and BLAST searches revealed that 4 of the 13 indels >20 bp contained insertions similar to miniature inverted repeat transposable elements (MITEs). When queried against the EST database, these putative transposable elements had matches to sorghum EST sequences derived mostly from stress-induced libraries (*e.g.*, wound, salt, and heat shock).

Consensus DNA sequence from each locus was used in BLAST searches (see materials and methods) to identify exons and introns for analysis by functional category. Results from this partitioning are presented in Table 3. Across loci, the numbers of synonymous and nonsynonymous changes within *vs.* between species showed no departure from the neutral equilibrium expectation (McDonald and Kreitman 1991) (Table 4). This result contrasts with that of our earlier, smaller study, which found a significant excess of replacement polymorphism (Hamblin *et al.* 2004). Diversity levels in introns were lower ($\pi = 0.24\%$) than at synonymous sites ($\pi = 0.38\%$), consistent with observations in other species.

Genomewide departures from equilibrium: The frequency spectrum of variation was measured by two statistics: D (Tajima 1989), which summarizes the folded frequency spectrum (*i.e.*, sites at frequency 1 are equivalent to sites at frequency $n - 1$), and H (Fay and Wu 2000), which becomes more negative when derived alleles are in high frequency relative to other alleles. Observed averages and variances of these statistics were compared to their expectations under the standard neutral model (first row of Table 5). While the average value of D is close to its neutral expectation of zero, the average value of H is quite negative, and the variances of both D and H are very large. Likewise, the variance of S is very large (42): in 100 coalescent simulations of our 204-locus data set under the standard neutral model (SNM) with average $S = 5$, the variance of S ranged from 13.3 to 26.4, with a median value of 18.7.

Haplotype number should be an increasing function of S, but this relationship is weak in our data set: the maximum number of haplotypes observed was eight, even though loci with 20 or more segregating sites were observed. The expected number of haplotypes, given S, depends on θ as well as the rate of recombination, so it is most easily estimated by simulations (Depaulis and Veuille 1998). For all loci with $S \geq 5$, we tested whether the number of haplotypes was unusual, using simulations based on θ and conditioned on S (see materials and methods). These simulations showed that a large fraction of loci had significantly fewer haplotypes than expected (Table 6). Even under a very conservative (and unrealistic) assumption of no recombination, 30% of tested loci had a significantly low haplotype number. These results are consistent with a previous observation that linkage disequilibrium (LD) in sorghum is more extensive than expected under assumptions of equilibrium (Hamblin *et al.* 2005).

These genomewide departures from equilibrium have important consequences for the interpretation of test statistics at individual loci. Of the 160 individual values of D that we observed, 28 (17%) would be considered significant under an equilibrium model with recombination; given the large variance in D, however, this number clearly includes many false positives. Similarly, most of the 32 significantly negative values of H (20%) can likely be attributed to the sixfold greater variance of H relative to that of the SNM. In the case of D, it is not only the large variance that raises concerns in interpretation, but also the haplotype structure. For the 10 loci with the most strongly negative D (all < -1.9), a feature that is associated with selective sweeps, the

TABLE 4

Polymorphism and divergence of synonymous and nonsynonymous variation

Comparison	Synonymous	Nonsynonymous	*P*-value
This study			
Within cultivated *S. bicolor*	153	90	0.42
Between *S. bicolor* and *S. propinquum*	221	113	
Hamblin *et al.* (2004)			
Within cultivated *S. bicolor*	32	34	0.004
Between *S. bicolor* and *S. propinquum*	66	27	

TABLE 5

Comparison of models with observed summary statistics

Model	θ[b]	*f*	Time of BN[a] T_0	Time of BN[a] Years	Var(*S*)	Mean *D*	Var(*D*)	Mean *H*	Var(*H*)
Observed					42.0	−0.08	1.40	−1.19	8.21
SNM	1.5	—	—	—	18.7	−0.02	0.76	0.01	1.34
					<0.01	*0.23*	*<0.01*	*<0.01*	*<0.01*
BN1	3.8	2.6	0.005	2,850	31.2	0.46	1.37	−0.94	4.90
					0.06	*<0.01*	*0.38*	*0.09*	*0.04*
BN2	3.8	2.5	0.010	5,700	31.2	0.33	1.38	−0.92	4.73
					0.09	*<0.01*	*0.45*	*0.08*	*0.04*
BN3	3.8	2.3	0.015	8,550	31.1	0.19	1.39	−0.91	5.22
					0.07	*<0.01*	*0.46*	*0.03*	*0.05*
BN4	3.8	2.2	0.020	11,400	29.8	0.09	1.38	−0.84	4.65
					0.04	*<0.01*	*0.43*	*0.02*	*0.07*
BN5	3.8	2.1	0.025	14,250	30.8	−0.01	1.36	−0.79	4.65
					0.03	*0.18*	*0.35*	*0.02*	*0.06*

All simulations had an average $S = 5.0$, obtained by fitting *f*, the size of the bottlenecked population relative to its duration. The summary statistics are the median of 100 mean values in simulations of 204 loci each. The numbers in italics are the proportion of times that the observed value, or one more extreme, resulted from a simulation. *D*, Tajima's *D* (1989); *H*, Fay and Wu's *H* (2000); *S*, number of segregating sites.

[a] Time of the beginning of the bottleneck, expressed in units of $4N_e$ generations (T_0) or in years, based on assumptions explained in text.

[b] Average $4N_e\mu$ over 204 loci.

distribution of variation among haplotypes is not that expected following simple directional selection, namely a star-shaped genealogy. Rather, these negative *D* values are all a consequence of a large number of singletons falling on one to three lineages (Table 7). The genealogy that we observe at these loci could result from recombination during a selective sweep, which may produce an excess of high-frequency derived variants detectable by the *H* statistic (see Figure 2 of Fay and Wu 2000). While the power of the *H* statistic drops off quickly after the fixation of a favorable allele (Przeworski 2002), selection during the domestication of sorghum is recent enough (<0.08 N_e generations ago) that power should be ≥30%, so some of the very low *H* values in our data set may be evidence of selection. However, as Przeworski (2002) has shown, a significant *H* statistic is "not a unique signature of positive selection." Population structure can give rise to an excess of significant *H* values, as could introgression from a divergent population. Therefore, in the absence of independent evidence, these results must be interpreted with caution. Neutral processes likely explain most of the extreme values of *H* observed in this data set.

TABLE 6

The fraction of loci with too few haplotypes

S	No. of loci	Loci with $p(K) < 0.05$ (%) $4N_er = 0$	Loci with $p(K) < 0.05$ (%) $4N_er = 5$
$5 \le S \le 10$	48	9 (19)	17 (35)
$S > 10$	23	12 (52)	17 (74)

K, number of different haplotypes observed; *S*, number of segregating sites.

Tests of selection based on levels of polymorphism and divergence: We used the HKA method of Hudson *et al.* (1987) to test whether there is a consistent relationship between polymorphism and divergence at unlinked loci, as would be expected if all loci have the same effective population size and time of divergence to the outgroup. A multilocus HKA test of our data showed that the data set as a whole is highly unlikely under a neutral model, with an HKA statistic of 458.8 (202 d.f., $P < 0.0005$). It is not clear how robust the HKA test is to departures from equilibrium; the large variance in *S* that we observe is among, not within, loci, and the test assumes that θ varies among loci. However, it is likely that a bottleneck will inflate the variance in ratios of polymorphism to divergence, making the test anticonservative (*e.g.*, Hammer *et al.* 2004; Haddrill *et al.* 2005). These results, therefore, should only be used to identify outliers, rather than as a test of significance. Of the 5% of loci that contributed the most to the HKA statistic, only one (PRC0378) showed a deficiency of polymorphism relative to divergence (Table 8), the expected signature of directional selection.

One possible reason for the lack of evidence of directional selection is that the low variation in sorghum, as well as the relatively low divergence to the outgroup, leads to poor statistical power. When 21% of loci have no

TABLE 7

Configuration of singletons at loci with $D < -1.9$

Locus	Chromosome, cM[a]	*D*	*H*	*S*	No. of singletons	*n*	No. of lineages with singletons
PRC0170	4, 79.3	−2.34	−1.98	15	14	17	2
S0078	9, 98.5	−2.28	−9.82	13	13	17	1
HHUK27	8, 74.7	−2.17	−8.38	9	9	17	2
pSB0771	1, 91.6	−2.13	−4.74	8	8	17	1
BCD0349	8, 4.6	−2.12	−9.83	12	11	16	3
HHUK22	4, 98.5	−2.12	0.04	16	15	17	1
pSB0095	6, 41.6	−2.08	−2.98	7	7	17	1
pSB0745	6, 50.0	−2.11	2.20	18	16	16	1
PRC0407	3, 70.8	−2.06	−6.65	7	7	16	3
pSB0521	6, 60.8	−1.96	−3.15	15	11	17	2

[a] Chromosome and genetic map position.

variation, it is difficult to conclude that any particular instance of low variation is unusual unless divergence is extremely high. There were seven loci with a deficiency of variation in the 10% that contributed most to the HKA statistic. We collected additional data for four of these loci; if these regions actually have experienced directional selection, increasing the number of sites surveyed should increase the significance of the departure. In three of these cases, polymorphism and divergence were less unusual when the additional data were included, suggesting that the apparent deficiency of polymorphism was due to chance. In the fourth case, however (PRC0378, Table 8), the additional data resulted in a greater departure from the neutral expectation, suggesting that this region may be a good candidate for having experienced directional selection.

The sequence for PRC0378 was obtained from an EST library of rhizome cDNAs from *S. halepense* and maps to a QTL for rhizome traits in a cross between *S. bicolor* and *S. propinquum* (Hu *et al.* 2003). The QTL extends over a large region of the sorghum map, however, so this could easily be coincidence. Since neither wild nor cultivated *S. bicolor* has rhizomes, underground stems involved in clonal propagation and associated with perenniality, a selective sweep associated with a rhizome trait would likely have occurred in the ancestral population rather than during domestication.

TABLE 8

Ten most unusual loci as assessed by HKA test

Locus	Chromosome, cM[a]	*S* observed	*S* expected	*P*(HKA)
pSB0745	6, 50.0	18	6.93	0.0004
HHUK22	4, 98.5	16	6.17	0.0012
pHER1E07	8, 61.6	13	4.92	0.0026
pSB1804	4, 63.1	29	12.74	0.0042
pRC0170	4, 79.3	15	6.18	0.0163
pSB0643	6, 61.2	17	7.30	0.0285
pSB0521	6, 60.8	15	6.65	0.0961
PRC0378	1, 15.4	0	11.14	>0.10
pSB0739	3, 30.8	20	10.05	>0.10
pSHR0114	2, 5.4	10	4.5	>0.10

[a] Chromosome and genetic map position.

As for the six loci that appear to have an excess of variation, three are among the loci that are highlighted in Table 7 because of their excess of singletons on a few lineages, and the fourth, pHER-1E07, shares this pattern. It is the large number of singletons, rather than a signature characteristic of diversifying selection, that has produced the significant test statistic at those loci. The two remaining loci, pSB0643 and pSB1804, have positive *D* values and gene genealogies that are more consistent with diversifying selection. Locus pSB0643 is closely linked to *Dw2*, a locus associated with variation in plant height that is expected to show local adaptation. Variation at pSB0643 is distributed in two clades of 7 and 10 individuals with 12 fixed differences between them (Figure 1a); there is no obvious geographic or phenotypic structure to the groups (see Table 1). Variation at locus pSB1804, which has homology to a transport-like protein in rice (XP_466724), is also distributed in two clades (Figure 1b). There are 16 fixed differences, including 9 amino acid differences, between them; 2 unique haplotypes do not fall into either clade. Again, there is no obvious geographic or phenotypic structure to the variation.

Nonequilibrium models: Given that multiple features of the data—the variance of the frequency spectrum, variance of *S*, and haplotype structure—depart strongly from equilibrium expectations, and given that sorghum has a history of domestication, it is reasonable to ask to what extent a recent bottleneck model of population history can explain these patterns. We examined a small number of simple bottleneck models in which most of the parameters were estimated on the basis of independent data (explained in detail in MATERIALS AND METHODS). The average ancestral population mutation

A

Accession	
BTx623	GGATAGGGCA TCGCCGG
NSL87902a	.........G ....
NSL87902b	....
NSL50875	....
NSL51030	....
NSL51365	....
NSL56174	....
PI221607	....
PI267408	....
PI585454	....
PI267539	TAGCG.ACT. CAATAAA
NSL77217	TAGCG.A.T. .A.TAAA
NSL55243	TAGCG.A.T. .A.TAAA
NSL56003	TAGCG.A.T. .A.TAAA
PI152702	TAGCG.A.T. .A.TAAA
NSL92371	TAGCGTA.T. .A.TAAA
NSL77034	TAGCGTA.T. .A.TAAA
KFS1	....

B

Accession	rsrrsrrrss rrsrrsrrsr rrrsrrssnn nnn
BTx623	GCCGGGTTAC GATGCTGTAG AGGGGTTAGG GCT
NSL50875	....
NSL87902	A..... T..... .....C.... A..
NSL77217	....C.... A..
PI152702	....A..... .....C.... A..
PI221607	.... T....C.... A..
PI585454	.... T....C.... A..
NSL55243	..TAA..GG. .TC.AC.CT- .AA..CACCA .TC
NSL51365	.TT..AA.G. .....A ...A..... ...
NSL87666	.TT..AA.G. .TC..CTC.. ..A.A.ACCA ..C
NSL51030	.TT..AA.G. .TC..CTC.. ..A.A.ACCA ..C
NSL56174	.TT..AA.G. .TC..CTC.. ..A.A.ACCA ..C
PI267539	.TT..AA.G. .TC..CTC.. ..A.A.ACCA ..C
NSL92371	.TT..AA.G. .TC..CTC.. ..A.A.ACCA ..C
PI267408	.TT..AA.GA .TC..CTC.. ..A.A.ACCA ..C
NSL77034	.TT..AA.GA .TC..CTC.. ..A.A.ACCA ..C
NSL56003	.TT..AA.GA .TC..CTC.. ..A.A.ACCA ..C
KFS1	.TT..AA.GA .TC..CTC.. ..A.A.ACCA ..C

Figure 1.—Haplotypes of variable sites at two loci with excess polymorphism. Variation at (A) locus pSB0643 and (B) locus psB1804. Information on the origin and racial type of accessions can be found in Table 1. Locus pSB0643 does not contain coding sequence. (B) r, replacement; s, synonymous; n, noncoding.

parameter ($4N_e\mu$) was fixed at 3.8 (with $4N_e\mu$ at individual loci scaled accordingly), the population recombination parameter ($4N_er$) was fixed at 0.01/bp, and it was assumed that the size of the current population and the ancestral population are the same (*i.e.*, $N_0 = 1$). Using our estimate of ancestral $4N_e\mu$, we inferred N_e, which then allowed us to calculate a plausible time of the bottleneck (T_0) on the basis of archeological data: a domestication time 3000–6000 years ago would correspond to 0.005–0.010 $\times$ ($4N_e$) generations. The intensity of the bottleneck (*i.e.*, length and size reduction) was adjusted to produce the observed value of *S*, and other resulting summary statistics were recorded.

We used a T_0 of 0.005 as a starting point and increased it incrementally to assess the fit of the resulting summary statistics (Table 5). For all models tested, the variances of the summary statistics were larger and *H* was more negative. However, models that were consistent with the estimated time of the bottleneck, *i.e.*, $T_0 = 0.005$–0.010, had average *D* values much larger than we observed (BN1 and BN2 in Table 5). Only when the time since the bottleneck was increased to 0.025, arguably equivalent to 14,000 years ago, did *D* approach the observed value (BN5). Conversely, BN1 and BN2 produced average values of *H* that were closer to the observed data, but the fit to *H* was worse in BN5.

The median values produced by these models (Table 5) show that none of these models is a good fit to the data, but they do not tell us whether the models exclude the data. Because the variances are large, any given iteration of a particular model might, by chance, produce a result that is much closer to the observed data. Indeed, Figure 2 shows that all five bottleneck models could produce the observed variance in *S*, variance in *D*, average *H*, and variance of *H*, although these values are in some cases in the tails of the distributions. However, for the recent bottlenecks, the range of average *D* (in 100 simulations) does not include the observed value.

DISCUSSION

Genomewide sequence variation in a species-wide sample of cultivated *S. bicolor* is strongly perturbed from equilibrium expectations with regard to the variance of the number of segregating sites, the variance of the site frequency spectrum, and haplotype configuration. This presents a serious challenge for identification of loci that may have experienced selection, as many tests for selection are based on these same properties of the data. This problem has been recognized and extensively explored in the model organisms of population genetics—Drosophila, humans, and Arabidopsis—all of which have at least some populations that clearly are not at equilibrium (*e.g.*, Andolfatto and Przeworski 2000; Pluzhnikov *et al.* 2002; Glinka *et al.* 2003; Marth *et al.* 2004; Nordborg *et al.* 2005; Schmid *et al.* 2005).

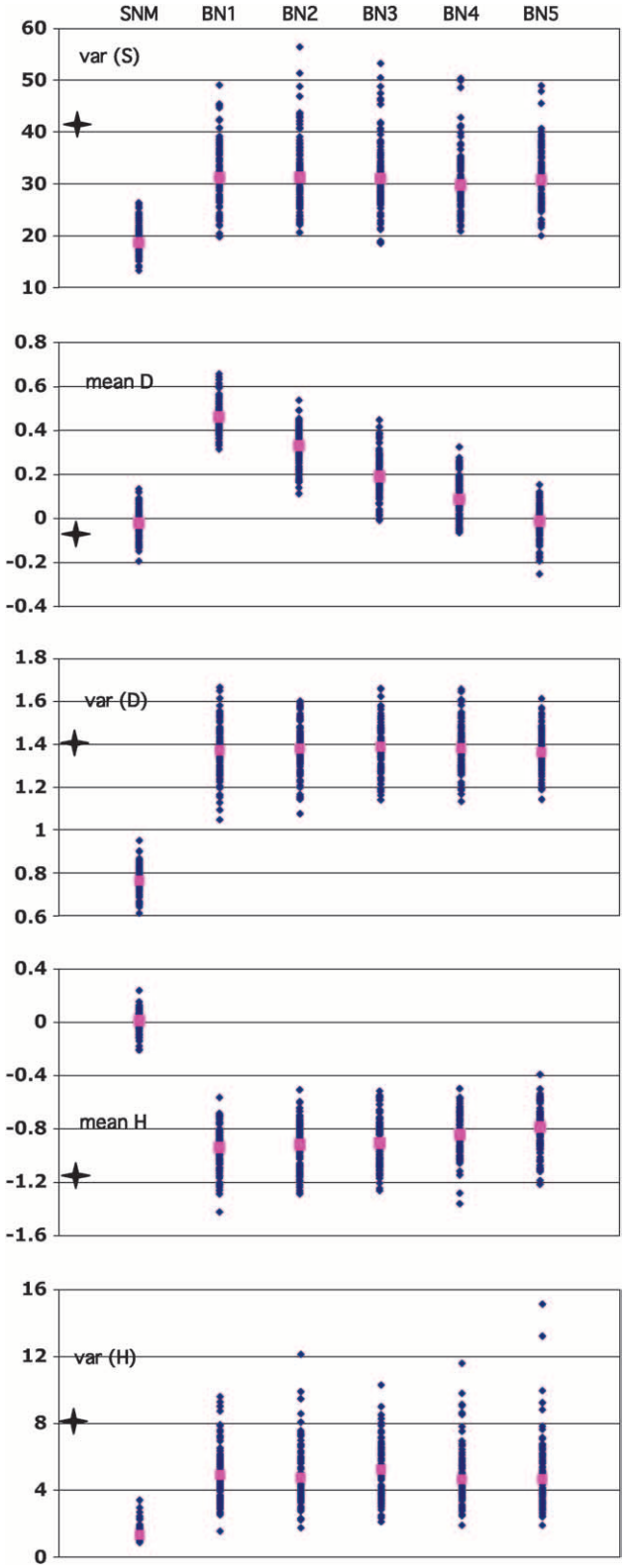

FIGURE 2.—Simulated distributions of summary statistics under various models. Each point is the average value from one simulated set of 204 loci under the given model. The lighter squares are the median values of 100 simulations for each model. For parameters of the models, see Table 5. The crosses indicate the observed value of the statistic.

Recently, there has been considerable effort to explore the effects of demographic changes and population structure on summary statistics and to find alternative models that reproduce some or many features of the data. While it has been possible for both maize and Drosophila to find fairly simple alternative models that capture many features of the data (*e.g.*, HADDRILL *et al.* 2005; WRIGHT *et al.* 2005), this has not been possible in Arabidopsis (NORDBORG *et al.* 2005; SCHMID *et al.* 2005). Our results suggest that the evolutionary history of sorghum, like that of Arabidopsis, has been too complex to allow use of a simple alternative model.

The time of a domestication bottleneck: The statistic that varied the most among models was Tajima's *D* (Figure 2). A bottleneck 0.005 or 0.010 $4N_e$ generations ago, our best estimate based on independent data (see MATERIALS AND METHODS), produces average *D* values that are strongly positive due to the loss of rare alleles. Models that place the bottleneck further in the past produced a better fit to the data (implying that our estimate of ancestral N_e is too large), in which *D* is very slightly negative. This is because additional time since the bottleneck allows for the accumulation of new, rare variants at loci where variation had been eliminated (a star-shaped genealogy). Thus there are more strongly negative and more strongly positive *D*'s than expected under neutrality, producing a large variance in *D*, but an average near zero, as we observe. The strongly negative *D*'s in our data set, however, are not produced by star-shaped genealogies (see RESULTS and Table 7), suggesting that this scenario is not appropriate and that fitting a model to summary statistics can be misleading. Furthermore, the fit to *H*, the variance of *H*, and the variance of *S* are all marginal (see Figure 2).

To reconcile an older bottleneck (*i.e.*, $T_0 = 0.025$) with a domestication time of 6000 years ago would require either that our estimate of μ be twofold larger or that our estimate of ancestral $4N_e\mu$/bp (θ_A) be twofold smaller. Thus some combination of higher μ, smaller θ_A, and older time of domestication may be able to explain the data. Our estimate of θ_A may, in fact, be somewhat inflated, if our sample of wild accessions includes individuals from subpopulations that did not contribute to the cultivated population. However, a twofold smaller θ_A would be very similar to $4N_e\mu$ in cultivated sorghum, inconsistent with a bottleneck.

Another piece of evidence to be considered is that a different species-wide sorghum sample produced a sequence data set with a positive average value of *D* (0.29; see Table 2). This positive value is consistent with a more recent bottleneck and with the other independent data. For this reason, we suggest that a more recent bottleneck is more likely correct for cultivated sorghum, but that the true evolutionary model includes other factors (*e.g.*, population structure and migration) that generate a sufficiently large variance that both positive and near-zero average *D*'s can be observed when different loci, or

different individuals, are sampled. Some models that include population structure and migration do have sufficiently large variances, although they are a poor fit to other aspects of the data (not shown).

Sampling issues: The discrepancy in average Tajima's *D*, as well as a difference in the ratio of nonsynonymous to synonymous polymorphism, in different samples raises some issues about sampling (see Tables 2 and 4). Both of these samples were intended to be representative of species-wide diversity in *S. bicolor*, but were chosen using slightly different criteria. The Hamblin *et al.* (2004) sample ($n = 22$ accessions) was chosen to maximize geographic distribution and morphological variation, as no genetic data were available at that time. Criteria for the sample in the current study ($n = 17$ accessions) also included maximization of genetic diversity, as assessed by variation at 74 SSR loci (Casa *et al.* 2005); it contains 8 accessions in common with the 2004 study.

Independent samples chosen randomly from a true single population should have similar properties, *i.e.*, the lineages should be exchangeable; the fact that they are not suggests that all members of the population do not share the same history. In this case, the sample was not drawn from a geographically restricted locality but instead is scattered, a sampling technique that is appropriate when species-wide variation is of interest and when natural populations do not exist. When each individual comes from a different deme, so that only the collecting phase of the genealogy (coalescence and migration among demes) is captured by the sample, the properties of the sample should be similar to those of a panmictic population (Wakeley 2004). In our data, it appears that the chance sampling of more divergent haplotypes has contributed enough low frequency variants to bring the average *D* close to zero even though another sample showed the expected effect of a recent bottleneck, namely a positive average *D*. These divergent haplotypes could be due to population structure in the ancestral and/or current population, in a scenario that violates the assumptions of Wakeley's analysis. We may have inadvertently increased the probability of sampling divergent lineages in this study, since we maximized genetic distance in selecting the sample. Given the star-shaped sorghum genealogy based on variation at SSRs (see Figure 1 in Casa *et al.* 2005), we did not expect this effect; however, the Casa *et al.* study may have been too small to detect structure. Early studies based on small numbers of individuals and markers concluded that Arabidopsis has no population structure, but later studies have come to quite different conclusions (Nordborg *et al.* 2005; Schmid *et al.* 2006). Unlike Nordborg *et al.* (2005), we are unable to attribute the divergent haplotypes to one or two lineages; many accessions contributed divergent haplotypes at different loci.

With regard to the discrepancy in the ratio of nonsynonymous to synonymous polymorphism, the difference is due to the number of replacement variants detected, as the level of synonymous polymorphism is very similar in the two studies. This could be a consequence of sampling but could also be due to sample size if amino acid polymorphism is slightly deleterious, since some fraction of low frequency amino acid variants that would be observed in a sample of 22 would be missed in a sample of 16.

Lack of evidence of directional selection: Because of sorghum's history of domestication, which was very recent in evolutionary terms, we expected to see evidence of directional selection at some loci but found none when we used the multilocus HKA test (Table 8). Strong evidence of directional selection was also lacking in two previous studies of genomewide variation in *S. bicolor* (Casa *et al.* 2005; Hamblin *et al.* 2004). Considering the three studies together, a total of 445 loci (371 sequenced loci totalling167 kb, plus 74 SSR loci) have been surveyed. While the loci surveyed had been genetically mapped, they were mapped in a cross between *S. bicolor* and *S. propinquum* and thus were not biased to be variable within *S. bicolor*.

If many loci in the data set have experienced selection, the multilocus HKA test may be overly conservative because the overall distribution is non-neutral. To assess the effect that this might have had on our results, we also performed HKA tests using pooled data from a putatively neutral reference subset of 22 loci chosen by the following criteria: the loci had (1) a small (<0.5) deviation in the HKA test of all loci and (2) a $|D|$ statistic of <0.5. In tests of each locus *vs.* the pooled data, 42 were significant at the 0.05 level: 23 loci with a deficiency of variation and 19 loci with an excess of variation. This far exceeds the 10 loci expected by chance, suggesting that some of these loci are true outliers. Conversely, none of the loci met the Bonferroni-corrected significance criterion of $P < 0.05/204 = 0.0002$, so this procedure does not change our conclusions.

Comparisons with maize: In cultivated maize, Wright *et al.* (2005) estimated that a minimum of 2–4% of the genome has experienced directional selection. If the proportion in sorghum were similar, and the probability that any random locus has experienced directional selection were 0.02, then the chance that 445 randomly chosen loci include zero selected loci would be quite small (0.00012). Furthermore, because of more extensive LD in sorghum (Hamblin *et al.* 2005), we expect the footprint of selection to be more extensive, increasing the percentage of the genome affected. There are a number of possible factors that might explain the difference between our results and those obtained for maize. Some of those factors have to do with the biology and history of the organism, while others have to do with the methodological approaches used to study them.

The simplest and most obvious explanation for our negative results is that low variation in sorghum provides

poor power for detecting directional selection and that much larger regions need to be sequenced to obtain sufficient power. While the average size of regions sequenced in this study is twice the average size surveyed in Wright *et al.* (2005), the average number of segregating sites per locus is roughly one-fourth. This implies that genomewide scans in organisms with low variation may require at least several kilobases of sequence at each locus. Sorghum and maize also have a major difference in mating system: sorghum's high frequency of self-pollination reduces the effective rate of recombination, which increases linkage disequilibrium and likely contributes to the very large variances associated with summary statistics. Finally, the domestication process in sorghum and maize may have been quite different. While maize has undergone a very dramatic change in plant growth habit and ear morphology compared to the progenitor teosinte (Doebley and Stec 1993), the morphological changes in domesticated sorghum are more subtle and fewer major genes may be involved. Selection in sorghum may have acted more frequently on standing variation than on new or rare mutations; in such cases, the signal of directional selection may be weak (Orr and Betancourt 2001; Innan and Kim 2004; Przeworski *et al.* 2005). Clear signals of selection may also have been obscured if domestication spread through a structured population, producing patterns of neutral variation very different from those expected following selection in a panmictic population (Slatkin and Wiehe 1998; Santiago and Caballero 2005).

Tests for directional selection in maize have largely been based on simulations that model the bottleneck from teosinte to maize (Vigouroux *et al.* 2002; Tenaillon *et al.* 2004; Wright *et al.* 2005), while our approach in this study was to look for a deficiency of variation relative to divergence to the only available sequence of an outgroup, *S. propinquum*, which is not the direct ancestor of cultivated sorghum. In cases where maize genes have been tested by both methods, it has been suggested that the comparison between maize and teosinte produces more significant results. Tenaillon *et al.* (2004), for example, concluded that "the addition of parviglumis [teosinte] data has improved significantly the ability to infer selection," when they found strong evidence of selection on *ts2* and *d8*, two genes for which the HKA test had produced equivocal results. In a recent article (Yamasaki *et al.* 2005), both methods were used to test for evidence of selection during domestication, and the difference in results was small; using the simulations, six loci showed significant evidence for selection, while five were significant by the HKA test. The difference was greater for loci showing evidence of improvement (*i.e.*, selection in inbreds only). Whether this difference represents increased power or a higher false-positive rate is not known.

Comparisons with Drosophila: Detecting evidence of adaptation in the context of a population bottleneck is a problem also faced in population genetic studies of Drosophila, as non-African (*i.e.*, derived) populations show genomewide departures from equilibrium and reduced variation relative to African populations. In some cases, a bottleneck model can explain most of the reduction in variation in non-African populations (Kauer *et al.* 2003; Haddrill *et al.* 2005; Thornton and Andolfatto 2006). While multilocus patterns are often suggestive of adaptive evolution in derived populations (Orengo and Aguade 2004; Schofl and Schlotterer 2004), it has been hard to demonstrate that selection has acted on a particular locus (but see Bauer Dumont and Aquadro 2005; Catania and Schlotterer 2005 for exceptions). Furthermore, apparent signals of selection in non-African populations often appear, upon further study, to be amplified signals of selection in the ancestral African populations (Li and Stephan 2005; Beisswanger *et al.* 2006; Pool *et al.* 2006), rather than evidence of adaptation to temperate environments. Interestingly, Ometto *et al.* (2005) found that a large proportion (>60%) of unusual loci in a non-African sample of Drosophila had an excess, rather than a deficiency, of polymorphism. As in our case, this may reflect increased power for detecting diversifying selection in a population with overall low levels of variation. Thus the difficulty of detecting directional selection in a bottlenecked population is not unique to domesticated species and presents important challenges for the field of population genetics.

Conclusions: Detecting a locus-specific reduction in variation, diagnostic of an episode of directional selection, is difficult when levels of variation at neutral loci are already low. When low levels of variation are caused by a bottleneck, they are accompanied by perturbations of the frequency spectrum and increased LD that further decrease the signal-to-noise ratio. In the case of *S. bicolor* it appears that, in addition to a bottleneck, other population-level phenomena have contributed to the observed departures from equilibrium. We base this conclusion on the observation that simple bottleneck models are inconsistent with the data and that, while several loci had extreme negative values of Tajima's *D* statistic, in not one case was this due to a star-like genealogy. Other phenomena that might underlie this observation could be, for example, ancestral population structure, multiple domestications, or introgression from wild conspecifics or congeners. More sophisticated models (incorporating data from wild populations) may, in principle, allow disentanglement of these factors. Phenotypic analyses in experimental populations, however, complementing population genetic analyses, may prove to be a more fruitful strategy for elucidating the genetic basis of the cultivated phenotype (Wright and Gaut 2005).

We thank Sharon Mitchell for designing some of the PCR primers; Joey Bedell for providing access to sequence data prior to publication; Alejandro Zamora for providing estimates of diversity in wild sorghum;

Jeff Jensen, Kevin Thornton, and John Pool for advice about modeling; Kangyu Zhang for providing source code for haploconfig; Baohua Wang for help with data formatting and submission to GenBank; Tessa Dumont, Amanda Garris, Gael Pressoir, and two anonymous reviewers for comments on the manuscript. This project was supported by grant DBI0115903 from the National Science Foundation to A.H.P., C.F.A., and S.K. A.M.C., C.F.A., S.K., and A.H.P. designed the study. A.M.C., H.S., and S.C.M. collected the data. M.T.H., A.M.C., and S.C.M. analyzed the data. M.T.H. wrote the paper.

LITERATURE CITED

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

Bauer Dumont, V., and C. F. Aquadro, 2005 Multiple signatures of positive selection downstream of Notch on the X chromosome in *Drosophila melanogaster*. Genetics **171:** 639–653.

Beisswanger, S., W. Stephan and D. De Lorenzo, 2006 Evidence for a selective sweep in the *wapl* region of *Drosophila melanogaster*. Genetics **172:** 265–274.

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.

Casa, A. M., S. E. Mitchell, M. T. Hamblin, H. Sun, J. E. Bowers *et al.*, 2005 Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theor. Appl. Genet. **111:** 23–30.

Catania, F., and C. Schlotterer, 2005 Non-African origin of a local beneficial mutation in D. melanogaster. Mol. Biol. Evol. **22:** 265–272.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788–1790.

Doebley, J., and A. Stec, 1993 Inheritance of the morphological differences between maize and teosinte: comparison of results for two F2 populations. Genetics **134:** 559–570.

Doggett, H., 1988 *Sorghum*. Longman Scientific & Technical, Essex, England.

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

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Glinka, S., L. Ometto, S. Mousset, W. Stephan and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165:** 1269–1278.

Haddrill, P. R., K. R. Thornton, B. Charlesworth and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res. **15:** 790–799.

Hamblin, M. T., S. E. Mitchell, G. M. White, J. Gallego, R. Kukatla *et al.*, 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 *et al.*, 2005 Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass sorghum bicolor. Genetics **171:** 1247–1256.

Hammer, M. F., D. Garrigan, E. Wood, J. A. Wilder, Z. Mobasher *et al.*, 2004 Heterogeneous patterns of variation among multiple human X–linked loci: the possible role of diversity-reducing selection in non-Africans. Genetics **167:** 1841–1853.

Hu, F. Y., D. Y. Tao, E. Sacks, B. Y. Fu, P. Xu *et al.*, 2003 Convergent evolution of perenniality in rice and sorghum. Proc. Natl. Acad. Sci. USA **100:** 4050–4054.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

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

Innan, H., and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. USA **101:** 10667–10672.

Innan, H., K. Zhang, P. Marjoram, S. Tavare and N. A. Rosenberg, 2005 Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics **169:** 1763–1777.

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.

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

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, New York.

Li, H., and W. Stephan, 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. Genetics **171:** 377–384.

Marth, G. T., E. Czabarka, J. Murvai and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166:** 351–372.

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

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

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol **3:** e196.

Ometto, L., S. Glinka, D. De Lorenzo and W. Stephan, 2005 Inferring the effects of demography and selection on Drosophila melanogaster populations from a chromosome-wide scan of DNA variation. Mol. Biol. Evol. **22:** 2119–2130.

Orengo, D. J., and M. Aguade, 2004 Detecting the footprint of positive selection in a european population of *Drosophila melanogaster*: multilocus pattern of variation and distance to coding regions. Genetics **167:** 1759–1766.

Orr, H. A., and A. J. Betancourt, 2001 Haldane's sieve and adaptation from the standing genetic variation. Genetics **157:** 875–884.

Pluzhnikov, A., A. Di Rienzo and R. R. Hudson, 2002 Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161:** 1209–1218.

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.

Przeworski, M., G. Coop and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution Int. J. Org. Evolution **59:** 2312–2323.

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.

Santiago, E., and A. Caballero, 2005 Variation after a selective sweep in a subdivided population. Genetics **169:** 475–483.

Schlotterer, C., 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics **160:** 753–763.

Schlotterer, C., 2003 Hitchhiking mapping—functional genomics from the population genetics perspective. Trends Genet. **19:** 32–38.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar and T. Mitchell-Olds, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601–1615.

Schmid, K. J., O. Torjek, R. Meyer, H. Schmuths, M. H. Hoffmann *et al.*, 2006 Evidence for a large-scale population structure of

Arabidopsis thaliana from genome-wide single nucleotide polymorphism markers. Theor. Appl. Genet. **112:** 1104–1114.

Schofl, G., and C. Schlotterer, 2004 Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-African D. simulans. Mol. Biol. Evol. **21:** 1384–1390.

Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. **71:** 155–160.

Stajich, J. E., and M. W. Hahn, 2005 Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. **22:** 63–73.

Storz, J. F., 2005 Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol. Ecol. **14:** 671–688.

Swigonova, Z., J. Lai, J. Ma, W. Ramakrishna, V. Llaca *et al.*, 2004 Close split of sorghum and maize genome progenitors. Genome Res. **14:** 1916–1923.

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

Tenaillon, M. I., M. C. Sawkins, L. K. Anderson, S. M. Stack, J. Doebley *et al.*, 2002 Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics **162:** 1401–1413.

Tenaillon, M. I., J. U'Ren, O. Tenaillon and B. S. Gaut, 2004 Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. **21:** 1214–1225.

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.

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., 2004 Metapopulation models for historical inference. Mol. Ecol. **13:** 865–875.

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

Wright, S. I., and B. S. Gaut, 2005 Molecular population genetics and the search for adaptive evolution in plants. Mol. Biol. Evol. **22:** 506–519.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310–1314.

Yamasaki, M., M. I. Tenaillon, I. Vroh Bi, S. G. Schroeder, H. Sanchez-Villeda *et al.*, 2005 A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell **17:** 2859–2872.

Communicating editor: M. Aguade