DOI: 10.1534/genetics.105.041566

Equilibrium Processes Cannot Explain High Levels of Short- and Medium-Range Linkage Disequilibrium in the Domesticated Grass *Sorghum bicolor*

Martha T. Hamblin,* Maria G. Salas Fernandez,* Alexandra M. Casa,* Sharon E. Mitchell,* Andrew H. Paterson† and Stephen Kresovich*,[1]

**Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853 and* †*Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602*

Manuscript received February 10, 2005
Accepted for publication June 2, 2005

ABSTRACT

Patterns of linkage disequilibrium (LD) are of interest because they provide evidence of both equilibrium (*e.g.*, mating system or long-term population structure) and nonequilibrium (*e.g.*, demographic or selective) processes, as well as because of their importance in strategies for identifying the genetic basis of complex phenotypes. We report patterns of short and medium range (up to100 kb) LD in six unlinked genomic regions in the partially selfing domesticated grass, *Sorghum bicolor.* The extent of allelic associations in *S. bicolor,* as assessed by pairwise measures of LD, is higher than in maize but lower than in Arabidopsis, in qualitative agreement with expectations based on mating system. Quantitative analyses of the population recombination parameter, ρ, however, based on empirical estimates of rates of recombination, mutation, and self-pollination, show that LD is more extensive than expected under a neutral equilibrium model. The disparity between ρ and the population mutation parameter, θ, is similar to that observed in other species whose population history appears to be complex. From a practical standpoint, these results suggest that *S. bicolor* is well suited for association studies using reasonable numbers of markers, since LD typically extends at least several kilobases but has largely decayed by 15 kb.

THE extent of allelic associations, commonly called linkage disequilibrium (LD), is of great interest in many species because of its implications for the design and feasibility of association studies and genome-wide scans to identify the genetic basis of complex traits. Genome-wide patterns of LD are fundamentally the product of two processes: (1) a new mutation occurs and is necessarily associated with the variants on the chromosome on which it arises, and (2) recombination places that mutation on a different genetic background, breaking the association. Thus the rate of recombination (r) is a key parameter in the process of LD decay. The relationship between r and LD is also affected by demographic factors. More specifically, the extent of LD is a reflection of the population recombination parameter, $4N_e r$, or ρ, where N_e (effective population size) is a function of the long-term historical size of the population, population structure, and mating system. Similarly, the mutational process that generates associations is summarized in the population mutation parameter $4N_e\mu$, or θ, where μ is the mutation rate. At equilibrium in a randomly mating population without selection, the extent of allelic associations is simply a function of the relative rates of mutation and recombination, since $4N_e$ cancels out in the ratio ρ/θ. Nevertheless, even in this simplest of scenarios, LD decay varies widely in unlinked regions due to the substantial stochasticity of the evolutionary process (Nordborg and Tavare 2002).

When a population departs from the panmictic equilibrium model, we can no longer accurately estimate ρ and θ from empirical data, and observed levels of LD may be inconsistent with empirically determined rates of recombination under simple demographic models. In the case of a partially self-pollinating species, the resulting reduction in heterozygous genotypes means that the effective rate of recombination is lower for a given rate of crossing over, so LD is, on average, expected to be more extensive. This is an equilibrium effect and can be accounted for by appropriate scaling (Nordborg and Donnelly 1997; Nordborg 2000). In other cases, however, mating is random but nonequilibrium population history leads to inconsistencies that are less easily explained. For example, LD in Drosophila is more extensive than expected on the basis of genetic distances and estimates of N_e from θ (Andolfatto and Przeworski 2000). In humans, LD can be either more or less extensive than expected, depending on the range

Sequence data from this article have been deposited with the GenBank library under accession nos. AY748249–AY748278 and AY761251–AY762036.

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

of interlocus comparisons (Pritchard and Przeworski 2001). Selection can also generate LD, although the locus-specific effects of selection can be hard to distinguish from the noise generated by neutral processes (Huttley *et al.* 1999; Kim and Nielsen 2004).

Our knowledge of the extent of LD in plants is limited (Flint-Garcia *et al.* 2003). Only in Arabidopsis, with a fully sequenced genome and high-density genetic map, has it been possible to conduct analyses comparable to those in Drosophila and humans, namely to evaluate LD over large, defined physical distances in the context of local rates of recombination. Very extensive LD on the order of 250 kb has been observed in this highly self-pollinating species (Hagenblad and Nordborg 2002; Nordborg *et al.* 2002), although a recent genome-wide study has shown that LD at most loci decays within 25–50 kb (Nordborg *et al.* 2005). In contrast, LD in maize, an outcrosser, decays within a few hundred base pairs in diverse samples (Tenaillon *et al.* 2001), although the extent of LD increases when narrower samples of germplasm (Remington *et al.* 2001; Ching *et al.* 2002; Jung *et al.* 2004) or targets of selection (Clark *et al.* 2004; Palaisa *et al.* 2004) are analyzed. Studies of gene-sized regions in both Populus (Ingvarsson 2004) and loblolly pine (Brown *et al.* 2004) indicate low levels of LD in these highly outcrossing trees. While these contrasts are frequently explained in terms of mating system, LD in wild barley, with a selfing rate similar to Arabidopsis, has patterns of LD more similar to maize (Lin *et al.* 2002). Very extensive LD is observed in rice (Garris *et al.* 2003; Semon *et al.* 2004) where the effects of mating system are likely confounded with population structure.

The extent of LD is a key issue in the design and feasibility of association mapping methods (Flint-Garcia *et al.* 2003). Association studies (also called LD mapping) have been successful in maize (*e.g.*,Thornsberry *et al.* 2001), where they may be limited to candidate genes because of the small extent of LD. *Sorghum bicolor,* a largely (~70%) self-pollinating domesticated grass (Rooney and Smith 2000) that is important for human nutrition in semiarid regions of sub-Saharan Africa (FAO 1996), is closely related to maize but has a smaller and less complex genome (Draye *et al.* 2001). In a previous study (Hamblin *et al.* 2004), we reported that LD over very short distances in sorghum was more extensive than in maize, suggesting that sorghum may be suitable for LD mapping of genes underlying complex, agronomically important traits common to both species. However, our limited data did not reveal anything about the scale over which LD dissipates, and we had no information about local relationships between genetic and physical distance for the regions analyzed. A physical map of the sorghum genome is being assembled and integrated with the genetic map (Draye *et al.* 2001; Mullet *et al.* 2002; Bowers *et al.* 2003), which ultimately will allow for estimation of rates of recombination, as well as LD, over fairly large distances, on the scale of centimorgans and megabases.

Meanwhile, genomic regions represented by BAC clones containing genetic markers can be used to sample patterns of LD and the relationship between physical and genetic distance on a scale of tens of kilobases. Six fully sequenced BAC clones, representing five different chromosomes, were used for this purpose. The goals of this study were to examine the pattern of pairwise associations among a large number of single-nucleotide polymorphisms (SNPs) in six large (40–100 kb) unlinked regions and to estimate ρ (the population recombination parameter, $4N_e r$), θ (the population mutation parameter, $4N_e\mu$), and r (the rate of recombination per base pair per generation) for those same regions. Analysis of this data set allowed us to assess the contribution of mating system and recombination rate to patterns of LD in sorghum and provides a general picture of those patterns that may prove useful in LD-based methods of genetic analysis.

MATERIALS AND METHODS

Sorghum accessions: Our panel of 32 *S. bicolor* accessions was a subset of 104 diverse accessions that had been characterized at 76 SSR loci (Casa *et al.* 2005) and were chosen to represent all of the population clusters identified by phylogenetic analysis. These include two U.S. inbred lines: BTx623 and RTx430; 22 land races: PI510985 and PI510906 from Botswana, NSL83707 from Cameroon, NSL50875 from Chad, PI22913 from China, PI267525 and PI267523 from Egypt, PI257595 from Ethiopia, PI221607 from Ghana, NSL51365, NSL87088 and NSL51836 from India, PI213900 from Kenya, NSL51032 from Mali, PI221655, PI221540, and NSL50744 from Nigeria, NSL51397 from South Africa, PI152702 from Sudan, NSL55751 and NSL77034 from Uganda, PI287624 from Zimbabwe; two subspecies of *drummondii* from Sudan: L-WA12 and L-WA71; six subspecies of *verticilliflorum*: L-WA13 from Sudan, L-WA22 and L-WA28 from Angola, L-WA42 from South Africa, L-WA55 from Benin, and L-WA88 from Egypt. The *verticilliflorum* subspecies is believed to be the ancestor of cultivated sorghum; all the subspecies are fully interfertile. One *S. propinquum* accession was used as an outgroup for estimates of ρ (see below). DNA was prepared from young leaves of individual plants according to the method of Doyle and Doyle (1987).

Choice of regions for resequencing: Annotation of predicted genes was used to identify putative intronic regions between 500 and 1700 bp in length that could be amplified from PCR primers on the basis of flanking exon sequence. Note that this is a technical consideration only and that the functional status of the sequence has no bearing on the analysis.

PCR products ranged from 700 to 1700 bp in size (Table 1) and were prepared for sequencing by treatment with shrimp alkaline phosphatase and exonuclease I digestion. PCR primers (and internal primers as necessary) were used for cycle sequencing with ABI Big Dye V. 3.1, and sequencing reactions were analyzed on a 3730 or 3700 sequencer at the Bioresources Center at Cornell University.

Analysis: Chromatograms were trimmed and edited manually using Sequencher 4.2 (Gene Codes, Ann Arbor, MI). When necessary, text files were exported and aligned using Multalin (http://prodes.toulouse.inra.fr/multalin/multalin.html) or Se-Al (http://evolve.zoo.ox.ac.uk). Although sorghum is usually homozygous at most loci, some heterozygous individuals

were observed (Table S7 at http://www.genetics.org/supplemental/). In these cases, the heterozygous individual was considered to have two chromosomes at that region only. Except for TASSEL (see below), none of our analyses required that phase be known. Summary statistics of DNA sequence polymorphism were estimated by DnaSP (Rozas and Sanchez-DelBarrio 2003). Multilocus tests of polymorphism and divergence (Hudson *et al.* 1987) and the variance of Tajima's *D* (Tajima 1989) were performed with Jody Hey's program HKA (http://lifesci.rutgers.edu/heylab). Coalescent simulations of a population bottleneck were performed with Hudson's program ms (Hudson 2002).

Files of variable sites generated by MEGA (Kumar *et al.* 2001) were formatted for LD estimation after removal of sites for which the minor allele had a frequency of <10% (Remington *et al.* 2001). In a few regions we observed an individual that appeared to result from introgression from a divergent wild relative, as it differed from all other alleles in the sample at many sites. For example, in region 2, of a total of 91 segregating sites, accession LWA22 contributed 51. Such individuals were removed from the sample for those regions only. Note that, because these individuals contribute only singletons, they have no effect on estimates of LD.

In a small number of cases, some accessions produced sequence of insufficient quality for calling all bases, but most of the polymorphic SNPs could nonetheless be reliably scored and were added to the LD analysis of the full sequence data. These sequences were not included in estimates of sequence diversity. These exceptions to the strategy of full resequencing are noted in Table 1. The programs dipdat (kindly provided by Dick Hudson) and maxdip (http://genapps.uchicago.edu/maxdip/index.html) were used to estimate r^2 (Hill 1974) and ρ (*i.e.*, $4N_e r$), respectively, from genotypic (*i.e.*, unphased diploid) data (Hudson 2001). LD triangle plots were constructed using TASSEL (http://www.maizegenetics.net/bioinformatics/tasselindex.htm) after removal of genotypes containing more than one heterozygous site, since TASSEL requires that phase be known.

We also used the program PHASE (Li and Stephens 2003) to obtain estimates of ρ and of the variation in recombination rates across the surveyed regions. Default parameters were used, except that the −X10 option was used to increase the number of iterations in the final run, as suggested in the documentation. Point estimates of ρ and λ (the factor by which the recombination rate in an interval between two loci exceeds the background rate) were obtained by taking the median value from 1000 iterations, as suggested in the documentation. An interval was considered a potential hotspot if the fifth percentile of λ was >1.0.

Estimation of rates of recombination: We used a recombinant inbred line (RIL) population, DNAs of which were kindly provided by Tom Hash of the International Crop Research Institute for the Semiarid Tropics. There were 244 lines in the population, which had been self-pollinated and advanced to the F_6 generation by single-seed descent. Within each BAC sequence, we identified at least two simple sequence repeats that showed length variation between the parents of the RIL population (BTx623 and IS18551). We designed primers for fragment analysis and scored the markers in the 244 DNAs. Recombination per generation (r) was estimated using the formula: observed recombination fraction $= 2r/(1 + 2r)$ (Burr *et al.* 1988) and divided by the number of base pairs between markers.

RESULTS

We characterized the decay of linkage disequilibrium in six unlinked regions of the *S. bicolor* genome represented by six fully sequenced BAC clones (Table 1). LD was estimated on the basis of SNP variation in several amplicons of ~700–1700 bp spaced irregularly across each region (Figure 1, Table 1), spanning a total distance of between 38 and 103 kb/BAC. Our panel of 32 *S. bicolor* accessions (see materials and methods), which includes both cultivated and wild representatives, was chosen to maximize diversity and to capture as much of the evolutionary history of *S. bicolor* as possible. All 32 lines were fully sequenced for most regions (for exceptions, see materials and methods). The numbers of SNPs observed in each subregion (*i.e.*, amplicon) vary considerably (Table 1). Only SNPs for which the minor allele was present three or more times in the sample were included in the analyses of LD; all subsequent references to SNPs refer to this subset of 249 of 427 total SNPs observed.

Pairwise estimates of LD: We calculated r^2 for all pairwise comparisons among SNPs within the same region and plotted those values as a function of physical distance. Figure 2, which shows these plots for each region separately, reveals a great deal of heterogeneity in the decay of LD among these regions. While the pooled data indicate that, on average, r^2 falls below 0.1 by 15–20 kb, in region 3 many values of r^2 remain >0.2 even at distances >35 kb, and associations in region 5 are essentially absent at distances of 5–10 kb. To assess "background LD," we calculated r^2 between pairs of sites in regions 1 and 2 (different chromosomes) and between regions 2 and 3 (both chromosome 1); the mean values of r^2 are 0.035 and 0.024, respectively.

Figure 2 also shows that the relationship between physical distance and r^2 is not strong, particularly over shorter distances. When logarithmic trend lines are fit to the data, the coefficients of determination vary from 0.03 for region 2 to 0.61 for region 4. When only distances <4 kb are included, the relationship between r^2 and distance almost entirely disappears: only region 4 has a coefficient of determination >0.03.

There is considerable interest in whether LD is "block-like," as strong haplotype structure, if real, may simplify LD-mapping studies (Zhang *et al.* 2002). The plots in Figure 2 obscure patterns of LD among blocks of sites, so we have also made triangle plots of the data (Figure S3 at http://www.genetics.org/supplemental/). Again, the patterns are very different among the regions. In region 4, for example, there are only a few isolated associations of *any* significance (< 0.01) between sites >1 kb apart. Region 5 has associations of $P < 0.0001$ that extend >15 kb, but those associations are patchy, not block like. In contrast, region 1 has an almost solid block of associations of $P < 0.001$ extending >13 kb. These patterns are not easy to compare because the spacing of subregions and the numbers of SNPs observed in each are not uniform; nonetheless it is apparent that some regions of the sorghum genome have extended haplotypes while others do not.

TABLE 1

Regions surveyed and informative SNPs observed

Subregion	Region 1 (AF010283, chromosome 3) First[a]	Last[a]	No. of SNPS[b]	Region 2 (AF5034333, chromosome 1) First	Last	No. of SNPS	Region 3 (AF366906, chromosome 1) First	Last	No. of SNPS
A	2658	3289	9	45198	45416	3	72052	73369	4
B	5797	6183	10	69772	69988	3	108953	109669	12
C	11421	12028	9	73731	73952	2	109776	110691	14
D	16897	17388	5	79496	79584	3	122557	123605	5
E	29923	30971	41	87209	87209	1	134222	135252	3
F	40631	41068	4						
	No. of pairwise comparisons:		3003	No. of pairwise comparisons:		66	No. of pairwise comparisons:		703
	Total distance:		38,410	Total distance:		42,011	Total distance:		63,200

Subregion	Region 4 (AY144442, chromosome 8) First	Last	No. of SNPS	Region 5 (AF466200, chromosome 4) First	Last	No. of SNPS	Region 6 (AF527809, chromosome 5) First	Last	No. of SNPS
A	70877	71517	4	25293	25553	3	6898	7165	2[c]
B	138484	139314	8[c]	57940	58463	23	15766	16050	2
C	163703	164258	8	60864	61476	28	39653	40350	2
D	166190	166190	1[c]	67489	67818	10	46711	47371	26
E	173198	173577	8	72761	73230	6			
	No. of pairwise comparisons:		406	No. of pairwise comparisons:		2415	No. of pairwise comparisons:		496
	Total distance:		102,700	Total distance:		47,937	Total distance:		40,473

[a] Nucleotide positions (in GenBank sequence) of the first and last SNPs scored in each sequenced segment.
[b] The number of SNPs whose minor allele was present three or more times.
[c] Not all bases were called in this region (see MATERIALS AND METHODS).

Multilocus estimates of LD: As is evident in Figure 2, pairwise estimates of LD are noisy and often difficult to interpret (NORDBORG and TAVARE 2002). A statistic that summarizes LD over an entire region is ρ, the population recombination parameter (see Introduction). A number of methods have been proposed for estimating ρ, but many have been found to perform poorly, particularly when sample sizes are small (WALL 2000). Due to the nature of our data set (namely, fairly large, discontinuous regions), we chose to use the composite likelihood (CL) method of HUDSON (2001) as well as the *p*roduct of *a*pproximate *c*onditionals (PAC) method of LI and STEPHENS (2003). Consistent with simulation studies that showed that these estimators perform well, the two estimates for each region do not differ by more than about twofold (Table 2). ρ_{PAC} was lower than ρ_{CL} in five of six cases, but the relative order of the values, by region, was similar for the two methods.

The CL method allows for gene conversion as well as crossing over and estimates the ratio of the two processes, given a certain mean gene conversion tract length (l) (see MATERIALS AND METHODS). Allowing for gene conversion can result in lower estimates of ρ; however, the impact on this data set is modest. Assuming

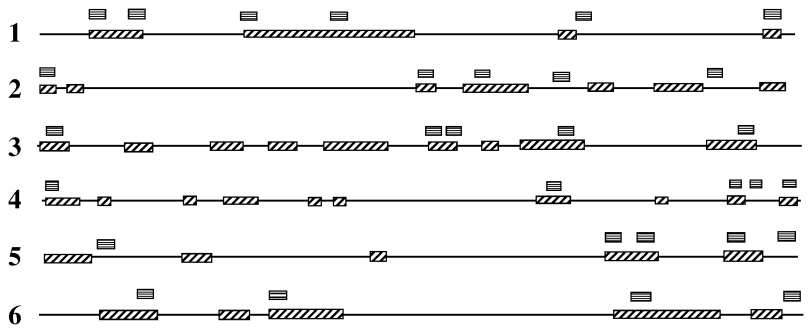

FIGURE 1.—Gene content (diagonally striped boxes) and sequenced subregions (horizontally striped boxes) for the six regions. See Table 1 for more information.

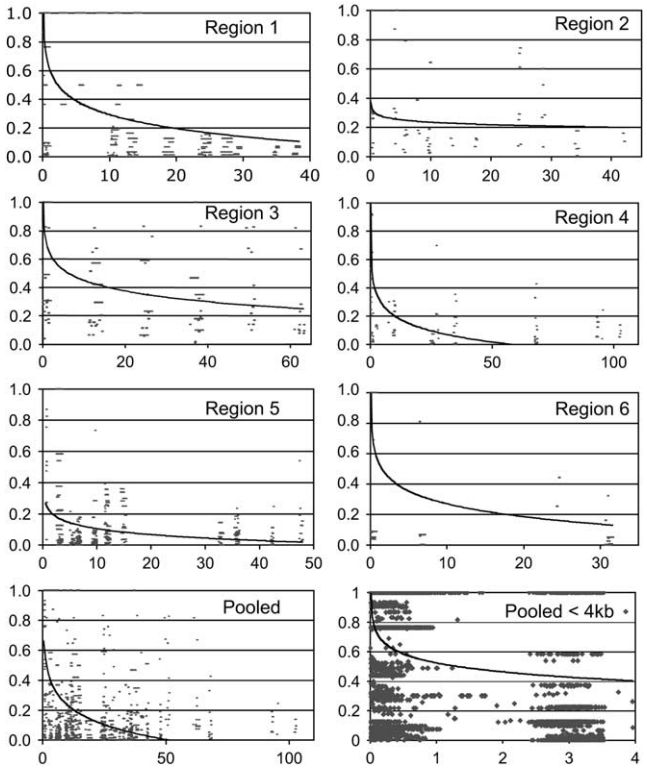

FIGURE 2.—Plots of r^2 (*y*-axis) *vs.* distance (in kilobases; *x*-axis) for individual regions and pooled data. The curves are logarithmic trend lines fit to the data.

$l = 300$ or 500, a model without gene conversion (*i.e.*, $f = 0$) fit the data best for all regions except regions 2 and 5 (Table 2). For region 2, the likelihood curve was extremely shallow, so that a model without gene conversion fit the data almost as well as several other models

TABLE 2

Estimates of the population recombination parameter ρ

Region	ρ_{PAC} (90% C.I.)[a]	ρ_{CI}[b]	f[c]	Hotspots[d]
1	1.76 (0.69, 4.51)	3.04	0	1E/1F; λ = 5.7
2	1.11 (0.22, 4.16)	1.56	5[e]	None
3	0.63 (0.19, 1.88)	1.31	0	None
4	7.53 (3.26,18.15)	16.08	0	Within 4B; λ = 6.4
5	3.15 (1.04, 9.29)	6.14	1	None
6	4.03 (0.95,14.87)	2.50	0	None

[a] The median value (of 1000 iterations) of ρ as estimated by the method of LI and STEPHENS (2003) $\times 10^{-4}$; the 90% C.I. reports the 5th and 95th percentiles of the sampled values of ρ.

[b] ρ as estimated by the method of HUDSON (2001) $\times 10^{-4}$.

[c] The ratio of gene conversion to recombination as estimated by the method of HUDSON (2001), assuming a tract length of 300 bp.

[d] Intervals (within or between subregions) for which the rate of recombination is higher than the background rate for this region; λ is the ratio of ρ for this interval to ρ for the region as a whole (see text).

[e] See text for a discussion of these results for region 2.

TABLE 3

Recombination frequency scored in RIL population

Region	Marker locations	Distance (kb)	*n*	No. of recombinants	r[a] (95% C.I.)
2	39938 129460	89.5	244	4	4.5 (1.8, 11.5)
3	4300 139974	135.6	238	3	2.3 (0.8, 6.7)
4	70675 139390	68.7	229	5	8.0 (3.4, 18.8)
5	7738 147967	140.2	243	4	2.9 (1.1, 7.4)
6	4990 95543	90.5	242	7	8.1 (3.9, 16.5)

[a] The recombination rate per base pair per generation ($\times 10^{-8}$).

that included high rates of gene conversion. Support for a model including gene conversion was strong only for region 5 but only a modest amount of gene conversion ($f = 1$) was inferred; ρ_{CL} for this region was 7.55 under a model with $f = 0$.

For both methods, the estimates of ρ (Table 2) vary ~12-fold across the six genomic regions, and the 90% C.I.'s for regions 3 and 4 do not overlap. Variation in recombination rate among regions of the genome is a possible factor underlying these differences. To test this hypothesis, we experimentally measured rates of recombination for five of our regions. (Recombination was not measured for region 1 because the entire BAC clone is only 40 kb in size and recombination events were not likely to be observed.) These studies, shown in Table 3, indicate that rates of recombination vary ~4-fold from region to region and may explain a portion of the difference in patterns of LD that we observe, in particular the difference between regions 3 and 4. None of the estimates of r, however, are significantly different. Variation in r explains only ~25% of the variation in ρ_{CL}, but explains almost 60% of the variation in ρ_{PAC}.

While rates of recombination based on crossover frequencies can be estimated over fairly large regions, there is interest in whether variation in recombination rates may occur on a much finer scale; except in the case of major hotspots (*e.g.*, YAUK *et al.* 2003), such variation can be detected only by inference from population genetic data (FEARNHEAD and DONNELLY 2001). The method of LI and STEPHENS (2003), which we used to estimate ρ, allows rates of recombination to vary along the region and reports a value of λ for each interval between adjacent SNPs (see MATERIALS AND METHODS). The vast majority (98%) of λ-values were between 0.5

TABLE 4

Comparison of ρ and θ for total and cultivated samples

		Total			Cultivated only		
Region	r/bp[a]	ρ[b]	θ[b,c]	ρ/θ	ρ[b]	θ[b,c]	ρ/θ
1	—	3.04	67.11	0.045	6.36	49.59	0.128
2	4.5	1.56	14.64	0.107	0.73	11.76	0.062
3	2.3	1.31	32.98	0.040	0.38	27.82	0.014
4	8.0	16.08	42.93	0.375	5.14	34.29	0.150
5	2.9	6.14	74.05	0.083	3.83	71.84	0.053
6	8.1	4.03	30.30	0.133	2.68	10.76	0.249

[a] $\times 10^{-8}$.
[b] $\times 10^{-4}$.
[c] Average nucleotide diversity (Nei 1987) including all observed sequence polymorphisms.

and 2.0, indicating a fairly uniform rate of recombination. Two exceptions, noted in Table 2, were found. In both cases, point estimates of λ were ~6, and 95% of the iterations produced a λ-value >1.0. While this does not constitute a test of significance, and does not account for multiple tests (252 intervals were tested), these two intervals are clearly outliers in the data set.

Cultivated sorghum has experienced a domestication bottleneck (Aldrich and Doebley 1992), which is expected to affect patterns of LD (Pritchard and Przeworski 2001). Furthermore, admixture of accessions from the wild and cultivated populations could result in elevated LD (Pritchard and Przeworski 2001). However, estimates of ρ for the cultivated accessions only are almost all (five of six) lower than that for the total sample (Table S5 at http://www.genetics.org/supplemental/), suggesting that the effect of the bottleneck is more important than that of admixture. This is not surprising, given that the wild accessions are only moderately differentiated from the cultivated ones and that there is little structure among the cultivated accessions (Casa *et al* 2005). The exception is region 1, for which ρ is considerably higher in the cultivated sample. This is because the cultivated sample had 47 SNPs as compared to 78 in the total sample, and a large number of alleles in strong LD were present only in the wild accessions (most of the SNPs in subregions 1A–1D). When wild accessions were eliminated, the number of SNPs in region 5 dropped from 70 to 62, and for the four other regions there was no difference. In general, most of the polymorphisms that were observed only in the wild accessions were in low frequency and had not been included in the analysis. The greater LD in the cultivated sample is therefore presumably due to the loss of haplotypes that provide evidence of recombination among the same pairs of alleles.

Because the ability to detect evidence of recombination depends on the presence of informative polymorphisms, it is useful to look at the ratio of ρ to θ in comparing ρ across regions. Table 4 shows the ratio of ρ and θ for the total sample and the cultivated accessions only. (In this analysis, we use the higher estimate of ρ, usually ρ_{CL}, which is conservative for testing the hypothesis that LD is higher than expected.) This ratio, which ranges from 0.040 to 0.375 in the total sample and from 0.014 to 0.249 in the cultivated sample, indicates that recombination is relatively infrequent relative to mutation.

DISCUSSION

In this study of linkage disequilibrium in *S. bicolor*, we present estimates of the population recombination parameter, ρ, based on six large, unlinked, fully resequenced regions for which we have also estimated the local rate of recombination per base pair. We find that the extent of allelic associations in sorghum, as assessed by pairwise measures of LD, is higher than in maize but lower than in rice and Arabidopsis, in qualitative agreement with expectations based on differences in mating system. Multilocus estimates of the population recombination parameter ρ, however, are among the lowest observed in any species, including Arabidopsis (Kuittinen and Aguade 2000; Hagenblad and Nordborg 2002; Nordborg *et al.* 2005). In attempting to account for these observations, several factors should be considered.

Estimation of ρ: A number of different methods have been proposed for estimating ρ, many of which do not perform well when tested against simulated data with known values of ρ (Wall 2000; Hudson 2001): point estimates are often far from the known value, and/or confidence intervals are very large. Likewise, different estimators of ρ may produce very different results for the same empirical data (*e.g.*, Tenaillon *et al.* 2002). Therefore it is reasonable to ask how much confidence we have in the estimates that we report. True confidence intervals for estimates of ρ are not trivial to obtain, particularly when the data collection scheme is not simple (as in our case). For ρ_{PAC}, however, it is easy to obtain an approximation of the uncertainty from the distribution of sampled ρ-values; each of these ~90% credible intervals (Table 2) contains the corresponding ρ_{CL}. In fact, the two estimates are in all cases within about twofold of each other, suggesting that they are not very far from the true value.

Testing an equilibrium model: Both mating system and rates of recombination affect LD in predictable ways that can be accounted for in an equilibrium model. If accounting for these factors does not explain the data, then we must invoke nonequilibrium phenomena such as population structure and history and selection.

The effects of mating system, recombination rate, and mutation rate on ρ in a partially selfing organism can be described by the equation $\rho/\theta = (r/\mu)(1 - F)$, where F is the inbreeding coefficient (Hagenblad and Nordborg

2002). The rate of self-pollination in sorghum is ~70% (Rooney and Smith 2000), so $F = 0.7/(2 - 0.7) = 0.54$. On the basis of synonymous substitutions between maize and sorghum at multiple loci (Swigonova *et al.* 2004), we use an estimate of $\mu = 1 \times 10^{-8}$/bp/generation. We therefore expect that r/μ should be on the order of $(r/1 \times 10^{-8}) \times (0.46)$, a value that is >1 for all the regions in our study. Actual values of ρ/θ are ~5–33 times lower (Table 4), using the *higher* of our two estimates of ρ. While mutation rates may vary among regions, as reflected by the differences in θ, it appears that the effects of mating system and recombination rate cannot account for the low values of ρ observed in this sample. Much higher rates of self-pollination (>90%) would be required to fit these data with an equilibrium model. While some cultivated sorghum accessions do have rates in this range, sorghum spent the vast majority of its evolutionary history as a wild species with an outcrossing rate believed to be ≥30% (Doggett 1988). Rates of recombination much lower (~10^{-9}/bp/generation), or mutation much higher (~2×10^{-7}/bp/generation), could also resolve the discrepancy, but such values are not plausible.

Population structure and history: Population structure can contribute to elevated LD (Pritchard and Przeworski 2001), so our wide sampling could bias estimates of LD upward if population structure were strong. However, we know from studies of SSR diversity that there is little structure in sorghum populations (Casa *et al.* 2005). Furthermore, results of Wakeley and Lessard (2003) suggest that our sampling strategy may minimize the impact of any population structure on estimates of LD. In samples drawn from just two demes, they show that correlations in histories of alleles are increased. The properties of "scattered" samples like ours, however, where each individual comes from a different deme, approach those of samples from panmictic populations. In any case, estimates of ρ in the cultivated sample are *lower* than those in the total (*i.e.*, admixed) sample, including wild accessions. These results are consistent with the findings for maize, where samples that capture greater genetic diversity (*i.e.*, more ancestral recombination events) show increasingly less LD (Flint-Garcia *et al.* 2003). Thus our sampling strategy could be considered conservative for testing the hypothesis that LD in sorghum is more extensive than expected.

Nonequilibrium population history can result in discrepancies between ρ and θ, as has been observed in humans and Drosophila (Frisse *et al.* 2001; Wall *et al.* 2002), and this factor is likely to be important in sorghum, which has experienced a domestication bottleneck. In our data, the cultivated subsample has 58% of the segregating sites of the total sample (Table S6 at http://www.genetics.org/supplemental/). Consistent with the effects of a bottleneck, the frequency spectrum of variation in the cultivated data set shows a strong departure from the neutral expectation: coalescent simulations show that both the mean (0.45) and the variance of Tajima's D (1.42) are significantly too large. We have explored a small range of recent simple bottleneck models to attempt to find one that is consistent with the domestication history of sorghum and with our empirical data. Using $\theta = 0.0056$ based on variation in wild *S. bicolor* (unpublished data from our lab), and $\mu = 1 \times 10^{-8}$/bp/generation, we estimate ancestral N_e to be 1.4×10^5. A domestication event 5000–6000 years ago (Kimber 2000) would thus correspond to ~$0.01(4N_e)$ generations. Using coalescent simulations, we were unable to find bottleneck parameters for which both the average D and the variance of D were close to what we observe. We are currently performing more exhaustive analyses with a larger data set that will be published elsewhere. Nonetheless, while the details remain to be determined, it seems reasonable to conclude that a nonequilibrium history has perturbed the frequency spectrum and also elevated linkage disequilibrium in this sample.

The weak relationship between association and distance for alleles that are <4 kb apart (Figure 2) is an interesting observation that may also be due to population history. If a bottleneck generated excess LD, that excess LD will have decayed over time more quickly for alleles that are farther apart (*i.e.*, for which r is larger), while closely linked alleles may still retain the signature of that demographic event. A similar pattern is observed at some loci in Arabidopsis; *e.g.*, see Figure 4 in Shepard and Purugganan (2003).

Selection: While the data may be consistent with a strictly neutral, nonequilibrium model, this of course does not preclude that selection may also have influenced the observed patterns of LD. In particular, we might expect that directional selection associated with domestication has played a role in the cultivated subsample. A multilocus HKA test of polymorphism and divergence (Hudson *et al.* 1987) in that subsample showed that the data were highly unlikely under a neutral model ($P = 0.00001$), but there was no convincing evidence of selection at any particular loci. It is possible that the departure is due to demography rather than to selection. Interestingly, subregions A–D of region 1 in the cultivated sample have only 10–15% of the diversity of the total sample, possibly due to selection at the *shrunken2* locus near subregion 1A, but this has not resulted in higher estimates of LD for this region in the cultivated sample (see results). Region 3, with the highest LD, contains the *phytochrome A* locus, for which a previous study showed no evidence of selection in sorghum (White *et al.* 2004). Thus, while we cannot rule out that selection may have played a role in shaping observed patterns of LD, there is no clear relationship in the data between the extent of LD and any evidence of a history of selection.

Features of recombination in sorghum: In maize, where LD is much less extensive, it appears that most

recombination occurs within genes rather than in intergenic regions, perhaps because the substantial variation in transposable element complements from chromosome to chromosome disrupts the homology necessary for recombination to occur (Fu *et al.* 2002). Our sequencing strategy was not designed to address this question, and most evidence for recombination occurs between the sequenced subregions in genomic areas that consist of both coding and noncoding sequence. There is evidence of only 12 recombination events [using the four-gamete test of Hudson and Kaplan (1985)] *within* the 27 fully resequenced subregions, almost all of which correspond to introns. However, the small number of bases surveyed represents only ~7% of the total length of the genomic regions analyzed (24 of 330 kb). Interestingly, the most gene-rich region, region 3, has the highest LD, while regions 4 and 5, relatively sparse in genes, have the least. Furthermore, rates of recombination, as assessed by the PAC method of Li and Stephens (2003), appeared to be quite uniform across each region. These observations, although anecdotal, do not suggest that recombination in sorghum is concentrated primarily within genes.

Haubold *et al.* (2002) and Nordborg *et al.* (2005) concluded that, in Arabidopsis, most "recombination" is in fact caused by gene conversion. High rates of gene conversion have also been implicated in the discrepancy between short- and long-range LD in humans (Frisse *et al.* 2001; Przeworski and Wall 2001; Andolfatto and Wall 2003) and a deficit of LD in short regions of low recombination in Drosophila (Andolfatto and Wall 2003). In sorghum, if there is any discrepancy between short- and long-range LD, it goes in the other direction: short-range LD is more extensive than expected, relative to long-range LD. Consistent with this observation, we saw little evidence for gene conversion, and there is little evidence of recombination within subregions, the scale on which short gene conversion tracts would be observed.

To understand and make use of information about LD, we made empirical estimates of r per base pair. Rates of recombination within and between species vary tremendously, such that a genetic interval defined by two markers 1 cM apart may correspond to 50 kb or 5 Mb of DNA, depending on the organism and the local rate of recombination. In wheat, for example, rates as low as 0.04 cM/Mb and as high as 8.5 cM/Mb have been measured (Gill *et al.* 1996a,b). In maize, measurement of recombination nodules per cytological distance vary >12-fold across chromosome 1 (Tenaillon *et al.* 2002). In sorghum, our estimates of r varied only ~4-fold in five unlinked regions, in the range of 2–8 cM/Mb, and the differences were not significant. All measured rates were higher than the estimate of 1.5 cM/Mb on the basis of total genome size and genetic map distance. This suggests that other genomic regions, not sampled in this study, may be recombinationally relatively inert.

Conclusions: The extent of LD in sorghum is greater than that in maize, where it is generally low (Tenaillon *et al.* 2001), and less than that in Arabidopsis (Nordborg *et al.* 2005). This qualitative observation is consistent with the mixed mating system of sorghum producing intermediate levels of effective recombination. Quantitative analyses, however, based on estimated rates of recombination, mutation, and selfing, show that both ρ and the ratio of ρ to θ are lower than expected under an equilibrium model. These analyses, as well as the frequency spectrum of polymorphism, suggest that a genome-wide departure from equilibrium underlies this phenomenon.

The greater extent of LD in sorghum makes it amenable for association studies using a limited number of markers. Genotyping of a few SNPs per gene in many cases can capture most of the haplotypic variation.

We thank Tom Hash for providing the DNAs for the RILs; David Witonsky for help with Maxdip; Don Viands for advice in estimating rates of recombination; Matthew Stephens for help with interpreting the results of PHASE; Chip Aquadro, Ed Buckler, Magnus Nordborg, and anonymous reviewers for comments on the manuscript; and Joy Bergelson for editorial assistance. Support for this project came from grant DBI0115903 from the National Science Foundation to A.H.P. and S.K.

Note added in proof: J.-S. Kim, M. N. Islam-Faridi, P. E. Klein, D. M. Stelly, H. J. Price, R. R. Klein and J. E. Mullet (2005, Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. Genetics **171** (in press)) have recently estimated the average genome-wide rate of recombination for euchromatic regions in *S. bicolor*. Their estimate is very similar to ours: 0.254 Mbp/cM or 4×10^{-8}/bp.

LITERATURE CITED

Aldrich, P. R., and J. Doebley, 1992 Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild Sorghum bicolor. Theor. Appl. Genet. **85:** 293–302.

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

Andolfatto, P., and J. D. Wall, 2003 Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. Genetics **165:** 1289–1305.

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.

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley and D. B. Neale, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. USA **101:** 15255–15260.

Burr, B., F. A. Burr, K. H. Thompson, M. C. Albertson and C. W. Stuber, 1988 Gene mapping with recombinant inbreds in maize. Genetics **118:** 519–526.

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.

Ching, A. S., K. S. Caldwell, M. T. Jung, M. Dolan, O. H. Smith *et al.*, 2002 SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. **3:** 19.

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

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

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

Draye, X., Y. R. Lin, X. Y. Qian, J. E. Bowers, G. B. Burow *et al.*, 2001 Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. Plant Physiol. **125:** 1325–1341.

FAO, 1996 *The World Sorghum and Millet Economies: Facts, Trends and Outlook.* Food and Agricultural Organization of the United Nations, Rome.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Flint-Garcia, S. A., J. M. Thornsberry and E. S. Buckler, IV, 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. **54:** 357–374.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Fu, H., Z. Zheng and H. K. Dooner, 2002 Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc. Natl. Acad. Sci. USA **99:** 1082–1087.

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

Gill, K. S., B. S. Gill, T. R. Endo and E. V. Boyko, 1996a Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. Genetics **143:** 1001–1012.

Gill, K. S., B. S. Gill, T. R. Endo and T. Taylor, 1996b Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics **144:** 1883–1891.

Hagenblad, J., and M. Nordborg, 2002 Sequence variation and haplotype structure surrounding the flowering time locus FRI in *Arabidopsis thaliana.* Genetics **161:** 289–298.

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.

Haubold, B., J. Kroymann, A. Ratzka, T. Mitchell-Olds and T. Wiehe, 2002 Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana.* Genetics **161:** 1269–1278.

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

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

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

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

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

Huttley, G. A., M. W. Smith, M. Carrington and S. J. O'Brien, 1999 A scan for linkage disequilibrium across the human genome. Genetics **152:** 1711–1722.

Ingvarsson, P. K., 2005 Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). Genetics **169:** 945–953.

Jung, M., A. Ching, D. Bhattramakki, M. Dolan, S. Tingey *et al.*, 2004 Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. Theor. Appl. Genet. **109:** 681–689.

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

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.

Kuittinen, H., and M. Aguade, 2000 Nucleotide variation at the CHALCONE ISOMERASE locus in *Arabidopsis thaliana.* Genetics **155:** 863–872.

Kumar, S., K. Tamura, I. B. Jakobsen and M. Nei, 2001 MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics **17:** 1244–1245.

Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165:** 2213–2233.

Lin, J. Z., P. L. Morrell and M. T. Clegg, 2002 The influence of linkage and inbreeding on patterns of nucleotide sequence diversity at duplicate alcohol dehydrogenase loci in wild barley (*Hordeum vulgare* ssp. *spontaneum*). Genetics **162:** 2007–2015.

Mullet, J. E., R. R. Klein and P. E. Klein, 2002 Sorghum bicolor—an important species for comparative grass genomics and a source of beneficial genes for agriculture. Curr. Opin. Plant Biol. **5:** 118–121.

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., and S. Tavare, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

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.

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.

Palaisa, K., M. Morgante, S. Tingey and A. Rafalski, 2004 Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. Proc. Natl. Acad. Sci. USA **101:** 9885–9890.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Przeworski, M., and J. D. Wall, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet Res. **77:** 143–151.

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

Semon, M., R. Nielsen, M. Jones and S. McCouch, 2004 The population structure of African cultivated rice *Oryza glaberrima* (Steud.): evidence for elevated levels of LD caused by admixture with *O. sativa* and ecological adaptation. Genetics **169:** 1639–1647.

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.

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, 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.

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.

Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen *et al.*, 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. **28**(3): 286–289.

Wakeley, J., and S. Lessard, 2003 Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. Genetics **164:** 1043–1053.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203–216.

White, G. M., M. T. Hamblin and S. Kresovich, 2004 Molecular evolution of the phytochrome gene family in sorghum: changing rates of synonymous and replacement evolution. Mol. Biol. Evol. **21:** 716–723.

Yauk, C. L., P. R. Bois and A. J. Jeffreys, 2003 High-resolution sperm typing of meiotic recombination in the mouse MHC Ebeta gene. EMBO J. **22:** 1389–1397.

Zhang, K., P. Calabrese, M. Nordborg and F. Sun, 2002 Haplotype block structure and its applications to association studies: power and study designs. Am. J. Hum. Genet. **71:** 1386–1394.

Communicating editor: J. Bergelson