Recombination and Gene Conversion in a 170-kb Genomic Region of *Arabidopsis thaliana*

Bernhard Haubold,[1,2] Jürgen Kroymann, Andreas Ratzka, Thomas Mitchell-Olds and Thomas Wiehe[3]

Max-Planck-Institut für Chemische Ökologie, Department of Genetics and Evolution, D-07745 Jena, Germany

Manuscript received September 5, 2001
Accepted for publication April 22, 2002

ABSTRACT

Arabidopsis thaliana is a highly selfing plant that nevertheless appears to undergo substantial recombination. To reconcile its selfing habit with the observations of recombination, we have sampled the genetic diversity of *A. thaliana* at 14 loci of ~500 bp each, spread across 170 kb of genomic sequence centered on a QTL for resistance to herbivory. A total of 170 of the 6321 nucleotides surveyed were polymorphic, with 169 being biallelic. The mean silent genetic diversity (π_s) varied between 0.001 and 0.03. Pairwise linkage disequilibria between the polymorphisms were negatively correlated with distance, although this effect vanished when only pairs of polymorphisms with four haplotypes were included in the analysis. The absence of a consistent negative correlation between distance and linkage disequilibrium indicated that gene conversion might have played an important role in distributing genetic diversity throughout the region. We tested this by coalescent simulations and estimate that up to 90% of recombination is due to gene conversion.

GENOME projects facilitate evolutionary studies, which in turn help to interpret the information uncovered by large-scale sequencing (Charlesworth *et al.* 2001). As a consequence, interest in the population genetics of the model plant *Arabidopsis thaliana* has grown steadily over the past decade. Three central observations have emerged from the analyses of the seven or so loci that have been subjected to comparative sequencing in this cruciferous weed (Kawabe *et al.* 1997; Purugganan and Suddith 1998, 1999; Kuittinen and Aguadé 2000; Savolainen *et al.* 2000): (i) There is an excess of rare polymorphisms, (ii) a number of genes have alleles that fall into two distinct classes (allelic dimorphism), and (iii) there is more recombination than might be expected, given that *A. thaliana* is a selfer.

The excess of rare polymorphisms, often indicated by a negative value of Tajima's *D*, is perhaps the least surprising of these findings. Most structural genes are subject to purifying selection, leading to an excess of rare frequency segregating sites. The converse, *i.e.*, an excess of genetic diversity (π), is an infrequent but often highly significant exception, as seen, for example, at the *Rpm1* resistance locus of *A. thaliana* (Stahl *et al.* 1999).

There is a slight tension between the observation of an excess of rare polymorphisms and allelic dimorphism. Extreme cases of the latter correspond to a deficiency of rare polymorphisms, as observed at the *Rpm1* locus (Stahl *et al.* 1999), and may therefore be evidence for balancing selection. However, the reason for the apparent dimorphism may simply be that in a sample of n sequences the expected time required for the last two lineages to coalesce is equal to that taken by the first $n - 2$ sequences (Kingman 1982a,b). In other words, even neutral genealogies tend to have deep splits, and since branch lengths are proportional to the number of segregating sites, apparent dimorphism might result from such a neutral process.

The most enigmatic observation concerns recombination. The outcrossing rate of *A. thaliana* has been estimated as 0.3% (Abbot and Gomes 1989), which is very low. On the other hand, Miyashita *et al.* (1999) found no significant linkage disequilibrium among 472 AFLP markers scored in 38 ecotypes. This contrasts with the situation in another well-studied selfing plant species, wild Barley (*Hordeum spontaneum*). Its outcrossing rate has been estimated as 1.6% (Brown *et al.* 1978) and in an extensive allozyme study 20 out of 28 populations investigated displayed significant genome-wide linkage disequilibrium (Brown *et al.* 1980). However, using an extension of the test for linkage disequilibrium applied to *H. spontaneum* (Haubold *et al.* 1998), Sharbel *et al.* (2000) detected highly significant linkage disequilib-

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AF471728–AF472273.

[1] *Corresponding author:* Lion Bioscience AG, Waldhofer Str. 98, D-69123 Heidelberg, Germany.
E-mail: bernhard.haubold@lionbioscience.com

[2] *Present address:* Lion Bioscience AG, Waldhofer Str. 98, D-69123 Heidelberg, Germany.

[3] *Present address:* Freie Universität Berlin und Berliner Centrum für genombasierte Bioinformatik, Arnimallee 22, 14195 Berlin, Germany.

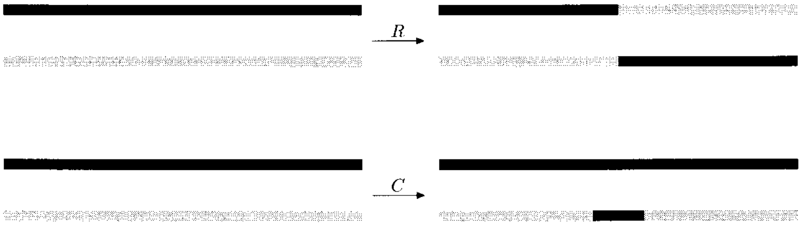

FIGURE 1.—The effects of reciprocal recombination (R) and gene conversion (C) on the distribution of genetic material.

rium among 79 amplified fragment length polymorphism (AFLP) loci scored in 142 ecotypes (henceforth referred to as *accessions*) of *A. thaliana*. Nevertheless, the extent of recombination in *A. thaliana* has remained unclear, prompting the present study.

We employed the genomic sequence of the Columbia accession of *A. thaliana* to sample the genetic diversity among 39 accessions at 14 loci of ~500 bp each in a region spanning 170 kb on chromosome 5. The region was chosen because a polymorphism for production of defensive metabolites maps within this interval (KROYMANN *et al.* 2001).

In this study we examine the mode of recombination in this region. Depending on the way the Holliday Junction is resolved, recombination may result either in reciprocal recombination or in gene conversion (Figure 1). Reciprocal recombination affects a series of homologous loci downstream of the recombination break point. Gene conversion, on the other hand, leads to the alteration of single segments only. Therefore, recombination causes the decay of linkage disequilibrium with distance, while no such effect results from gene conversion if conversion tracts are short (WIEHE *et al.* 2000). We have applied this idea to our data and discovered that a substantial input from gene conversion is likely.

MATERIALS AND METHODS

Plant material and DNA sequencing: The 39 accessions used in this study are listed in Figure 3. Primers were designed using the published sequence of the accession Columbia and the software PRIMER3 (ROZEN and SKALETSKY 1998). Total DNA was extracted from leaves of single plants, amplified, and sequenced directly on both strands. All primer pairs are shown in Table 1.

Data analysis: *Alignment:* DNA sequences were aligned using PILEUP (GCG Wisconsin Package) and all computations were carried out after gap removal.

Nucleotide diversity: Most loci in our sample contained coding as well as noncoding regions (Table 2). To compute the average number of silent substitutions between pairs of sequences (π_s) we included the third codon positions of the coding regions as well as the complete noncoding segment and applied

$$\hat{\pi}_{\text{silent}} = \frac{n}{L(n-1)} \sum_{i=1}^{n-1} \sum_{j>i} m_{ij},$$

where n is the sample size, L the number of silent positions in the alignment, and m_{ij} denotes the number of mismatches between the ith and jth haplotype.

Confidence intervals for π_s were estimated using the bootstrap procedure (EFRON 1979) across taxa: Rows of the aligned data matrix were resampled with replacement and the average number of pairwise mismatches per nucleotide was recalculated 10,000 times. The resulting mismatch values were sorted, and the 2.5 and 97.5% quantiles were looked up in the sorted array.

Pairwise linkage disequilibrium: We use the normalized linkage disequilibrium, D', to quantify pairwise linkage disequilibria (LEWONTIN 1964). Consider two biallelic loci, $\mathscr{X}$ and $\mathscr{Y}$, and denote the probability of finding allele "1" at locus $\mathscr{X}$ by p_1 and at locus $\mathscr{Y}$ by q_1. The frequencies of the possible four haplotypes are denoted P_{00}, P_{10}, P_{01}, and P_{11}. We define $d = P_{00}P_{11} - P_{01}P_{10}$ as the linkage disequilibrium. Then, the desired disequilibrium measure is

$$D' = \begin{cases} \dfrac{d}{\min(p_1(1-q_1),\,(1-p_1)q_1)}, & \text{if } d > 0 \\ \dfrac{d}{\max(-p_1q_1,\,-(1-p_1)(1-q_1))}, & \text{if } d < 0 \\ 0, & \text{if } d = 0 \text{ and } \min(p_1(1-q_1), \\ & \quad (1-p_1)q_1) \neq 0 \\ 1, & \text{otherwise}. \end{cases}$$

The correlation between D' and pairwise distance was tested using Mantel tests as described by MANLY (1994, p. 72ff).

Now consider three loci, $\mathscr{X}$, $\mathscr{Y}$, $\mathscr{Z}$, where the distance between $\mathscr{X}$ and $\mathscr{Y}$ is less than or equal to one-tenth of the distance between $\mathscr{X}$ and $\mathscr{Z}$. We then define the ratio

$$Q = \frac{\log(D'_{\mathscr{X}\mathscr{Y}})}{\log(D'_{\mathscr{X}\mathscr{Z}})}.$$

Its expectation is equal to 1 for gene conversion and <1 for reciprocal recombination (WIEHE *et al.* 2000). The distribution of $\tilde{Q} = \log_{10}(Q)$ as determined either from real or simulated data is an indicator of the relative frequency of recombination and gene conversion. We further consider the sign of $\tilde{Q}$ as a derived random variable:

$$\text{sign } \tilde{Q} = \begin{cases} +1 & \text{if } \tilde{Q} > 0 \\ \ \ 0 & \text{if } \tilde{Q} = 0 \\ -1 & \text{if } \tilde{Q} < 0, \end{cases}$$

and $\langle \text{sign } \tilde{Q} \rangle = -\text{Prob}(\tilde{Q} < 0) + \text{Prob}(\tilde{Q} > 0)$.

Simulating the distribution of $\tilde{Q}$: We used a coalescent program distributed by HUDSON (2002) to generate gene samples. This program implements both reciprocal recombination as well as the model of gene conversion developed by WIUF and HEIN (2000). The chosen input parameters corresponded to our data, that is, sample size = 39 and number of sites = 168,037, which is equal to the distance between the first and the last nucleotide in our data set. The mutation parameter θ was computed by scaling the value observed in our 14 loci, which cover only a fraction of the surveyed region, resulting in $\theta =$

TABLE 1

Primers used to amplify the 14 loci investigated in this study

ID	Name	Forward primer	Reverse primer
1	MRN17.3	TGCTTCTCGTGGTCATAAGG	CGCTCAACAGAAACACCAAC
2	MRN17.9	CGCATACATTAAAAGGGTACGG	TCCGAGAAAGATTCGAACAAC
3	MRN17.11	CAACTCTCCAGGCAACACAA	GCAGATAAGAAACCAGCGGA
4	s-H	TTGAGGGGTTAGGGAGGGAAGGA	CAAGAAAACCGAAGAAGAAAACAAAAGACT
5	MRN17.18	CGCAAATTACGTCAGGAGTG	AACCGGTTCACTGTTCTTCTC
6	MRN17.20	TGGTCGGAAAATAATAAGGAGGTT	TTGCCGGCGATAATGAAAAAG
7	MRN17.21	CGGAAAGCCTTAGGAGTTGG	CAAAACCCTTTTGGCCTGAC
8	MYJ24.4	CCCCATTGATCAAGGAAGCATAAA	GTGAAGTTCTAGTAATCCGACAGG
9	MYJ24.5	GAGGAGATGCAAAGAGAGATAACAG	GCGGCTAAAACAGATAGCTACTTC
10	MYJ24.6	TACTCGCCGAAAGAGAAACC	ACCGCAAACGTAATGACTCC
11	MYJ24.7	TGATGAAGAAGGCCGTAGAAG	GTGATTTGCCCTCCATATCC
12	MYJ24.8	GGAGATGTTACCACCAGATGTTC	CTGCTGCTTCCTCCTCAGTC
13	MYJ24.11	TCCTCCTCCATCTCCATCAC	TCCCAATTCTCTCCAGCATC
14	MYJ24.13	CGTCGCTCAGCTTCTTTACC	GCCTGTCCACTATATCCTCCTG

1320. The tract length for gene conversion was 300 bp, a number smaller than the mean conversion tract length of 352 bp found in Drosophila (Hilliker *et al.* 1994) and larger than the estimated tract length of 30 bp at the human leukocyte antigen (HLA) locus (Parham *et al.* 1995; Wiehe *et al.* 2000). From the simulated samples we excluded all mutations outside of our 14 loci and grouped the remaining polymorphisms into blocks, which (1) contained at least two polymorphisms (to exclude point mutations as a potential source of variation), (2) extended at most as far as the polymorphism pattern across the aligned haplotypes did not change, and (3) did not extend beyond locus boundaries.

An example of such a block is provided by positions 136,818–137,110 in our data set (Figure 3), which all belong to locus 10 and have an identical haplotype structure. Although the adjacent polymorphic position 138,531 also maintains the haplotype structure, it is not included in the block as it belongs to locus 11 (Figure 2). From triplets of such blocks of polymorphisms we computed $\tilde{Q}$ as outlined above. Distances were defined as base pairs between block midpoints.

Multilocus disequilibrium: Multilocus disequilibrium was investigated by treating each distinct sequence at the 14 loci as an allele and calculating the number of loci at which each pair of haplotypes differed. The observed variance of this

TABLE 2

Coding parts and annotations of the 14 loci investigated

ID	Name	Nuc.	Nuc.-gaps	Coding	Annotation
1	MRN17.3	409	406	325–409	s.t. Synechocystis sp. PCC6803 sll0362 (alanyl-tRNA synthetase; alaS)
2	MRN17.9	519	519	1–378	s.t. lysosomal pro-x carboxypeptidase precursor (EC 3.4.16.2)
3	MRN17.11	324	324	1–324	s.t. *A. thaliana* histone H2B-like protein
4	s-H	604	580	None	None
5	MRN17.18	503	464	1–26; 247–353; 473–464	s.t. *Caenorhabditis elegans* cosmid T27F7
6	MRN17.20	464	452	254–461	Unknown protein
7	MRN17.21	546	414	1–150; 394–425	s.t. serine carboxypeptidase precursor (EC 3.4.16.0)
8	MYJ24.4	602	594	1–112; 164–271; 406–533	Acetyl-CoA synthetase-like protein
9	MYJ24.5	568	462	1–129	Putative protein
10	MYJ24.6	335	335	1–335	s.t. *Dictyostelium discoideum* ThyB
11	MYJ24.7	429	428	206–302; 377–428	s.t. R07E5.1 protein
12	MYJ24.8	456	429	1–45; 130–296; 391–430	s.t. dr1 protein homolog
13	MYJ24.11	476	462	1–197; 284–464	s.t. hypothetical 37.3-kd protein in ycf29-psbe intergenic region
14	MYJ24.13	537	451	1–50; 338–451	s.t. ATP-dependent Clp proteinase (EC 3.4.21.92) chain P homolog

Annotations were taken from the Arabidopsis Information Resource (http://www.arabidopsis.org); "s.t." denotes "similar to."

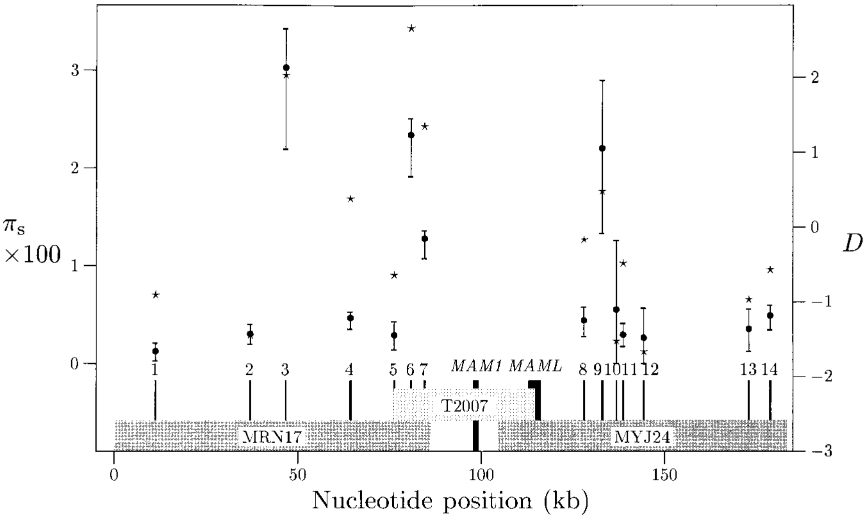

FIGURE 2.—Silent genetic diversity, π_s (●), including 95% confidence intervals, and Tajima's D (★) across the 170-kb region studied. Tajima's D was significant at locus 6 ($P = 0.01$), perhaps indicating balancing selection. MRN17, T2007, and MYJ24 designate bacterial P1 clones of genomic *A. thaliana* DNA used in the Arabidopsis genome sequencing project (ARABIDOPSIS GENOME INITIATIVE 2000); *MAM1* encodes a methylthioalkylmalate synthase involved in glucosinolate chain elongation (KROYMANN *et al.* 2001); and *MAML* encodes a duplication of *MAM1*.

"mismatch distribution," V_D, was then compared to the variance expected under linkage equilibrium, V_e (BROWN *et al.* 1980; HAUBOLD *et al.* 1998). The ratio between these variances serves as a measure of the strength of multilocus association in the sample

$$I_A^s = \left(\frac{V_D}{V_e} - 1\right)\frac{1}{l - 1},$$

where l is the number of loci and I_A^s is the standardized index of association (MAYNARD SMITH *et al.* 1993; HUDSON 1994). We used the software LIAN to calculate the I_A^s and to test its significance (HAUBOLD and HUDSON 2000).

RESULTS

Sequence data and genetic diversity: A total of 39 accessions were sequenced at 14 loci distributed over a 170-kb region (Figure 2). After gap removal this amounted to 6321 nucleotides in 39 accessions. Of this data set, 170 sites distributed among 35 haplotypes were polymorphic (Figure 3). With the exception of one hypervariable position (46,750; Figure 3), all segregating sites had only two nucleotide states. In addition, there were three heterozygous positions in accession Kondara (37,062, 37,304, and 37,351; Figure 3). The hypervariable and heterozygous sites were removed from the computation of pairwise disequilibria.

Most of the loci were located within predicted genes and contained both protein-coding as well as noncoding parts (Table 2). However, locus 4 was entirely noncoding, while loci 3 and 10 consisted of coding sequence only. With the exception of loci 6 and 9, functions had been assigned to the investigated loci in the context of the Arabidopsis genome project (ARABIDOPSIS GENOME INITIATIVE 2000). These functions were diverse, ranging from putative alanyl-tRNA synthetase (locus 1) to histone (locus 3), peptidases (loci 7 and 14), and acetyl-CoA synthetase (locus 8; Table 2).

The genetic diversity varied by a factor of 30 between $\pi_s = 0.001$ at locus 1 and $\pi_s = 0.030$ at locus 3 across the region (Figure 2). To assess whether these diversity values were compatible with neutral equilibrium expectations, we investigated the frequency spectrum of the single-nucleotide polymorphisms using Tajima's D test statistic (TAJIMA 1989). This test is based on the assumption that the data have not been subject to recombination. We explored this assumption by computing the minimum number of recombination events for each locus (R_m; HUDSON and KAPLAN 1985). Only locus 7 showed evidence of a recombination event and with this background information we proceeded to calculate Tajima's D.

The only locus with a significant value of Tajima's D was locus 6 ($D = 2.66$, $P = 0.01$; Figure 2). Unfortunately, its function is unknown. Further, the signs of the test statistics showed no consistent pattern, with 9 out of the 14 loci having $D < 0$ and the rest $D > 0$ (Figure 2).

Multilocus linkage disequilibrium: HANFSTINGL *et al.* (1994) hypothesized that recombination in *A. thaliana* was frequent enough to erode linkage disequilibrium between sites just 350 bp apart. Since all the loci investigated in our survey were >350 bp apart (Figure 2), we assessed the strength of association between these loci by calculating the standardized index of association, I_A^s, which is zero under linkage equilibrium (HUDSON 1994). For our sample $I_A^s = 0.179$, a value significantly >0 ($P < 10^{-4}$).

Phylogeny: Given that there was strong linkage disequilibrium between the surveyed loci, we used the exploratory tool of statistical geometry to investigate the phylogeny of the genomic region (EIGEN *et al.* 1988; MAYNARD SMITH 1989). Statistical geometry proceeds by first generating a phylogeny on the basis of the parsimony criterion for each quartet of sequences in the sample. These phylogenies are averaged to generate the graph shown in Figure 4. Note that there are three ways in which this graph can be reduced to a conventional unrooted tree: Collapse dimensions X and Y of the cen-

```
                                                                                              11111111111111111111111111111111111111111111111111111111111111111111111111111111111111
          1133333333333333344444444444666666666667777778888888888888888888888888888888888888888222222222222333333333333333333333333333333333333333333333333333444444444777777777777777
          1166677777777777766666666666444444444456666660000000000000000000000111144444444444447777888888882222222223333333333333333333333366666666667777888884444444442223333888888889
          2389900122222333366667777889111223444590112236666677777778888888999000022334446666778999011233338888999990000111111111222223333388888889900115666711222224588911226788880
          8361216823569025724791567561477772789880244892466603456680116678259001469011395588122469436317991478337783459124455899900444334441444467582601300774800224906786913633556 2
          2555612293045421012294027379456294082212219079117879194683064590968069879616483701099731988908782869576783019260717556712037361348378912729630145891823139589843945909033 0

AA-0      CATTCGGGCCGAGCACGCACTAGACCCGACAACTTTATTCAGTGTGGACCAAATTGCTCGTGGAACCGCAAAGTTTTTATCTTCCGTAGTTGCCCCCATGTCAGAAAGTATCGGATGAACAAAATTGACACTCCAATGTCTGGACACATCTTTGGCGACTAATCGACAGG
AG-0      ........T.................................................................................................................................................................
BL-1      ...........T.....AG.CGTGTAG........CC.............CG.CG.T..A....C...T.GCTA..A..GGGA.............TTGA...................TGG.....................T..........................
BLA-10    .T.....A...T....A....GTG.AG....T..............AGATCGCC.ATAAAC.CCCTTTT..................C...C....T.GAGA...G.TGTAAA.GAATTT..TG.CCTA..............T............A............A
CAL-0     ...........T.....AG.CGTG.AG...................AGATCGCC.ATAAAC.CCCTTTT..C.AC.A..GGGA..C.........................................................T..................A.......
CAN-0     .....A.....T.....AGTCGTG.AG..TT.......A.GTA.......CG.C..T..A....C...T.GCTA..A..GGGA.TC.TAC.C..GTTTGA...................TGG.....................T.G..........A............A
CNT-1     ...........T.....AG.CGTG.AG...................AGATCGCC.ATAAAC.CCCTTTT..................C...C....T.GAGA...G.TGTAAA.GAATTT..TG.CCTA..............T.G..................A.T...
COL-0     ..........................................................................................................................................................................
CVI-0     T..GT.....AT..G..AG.CGTG.AG..TT.T........T..G.AGATCGCC.ATAAAC.CCCTTTT..CTA..A..GGGAA..G.........TTGA...................TGG.....................T............A..CTG.......A
DI-0      ..........................................................................................................................................................................
DI-1      ......................C.......................AGATCGCC.ATAAAC.CCCTTTT..C.AC.A..GGGA............................................................T..........................
DIJON_G   ..C........T.....AGTCGTG.AG........CC.............CG.CG.T..A....C...T.GCTA..AG.GGGA.............TTGA...................TGG.....................T.............G......A.....
EI-2      ...........T.....AGTCGTG.AG.C......CC.A.GTA.......CG.C..T..A....C...T.GCTA..A..GGGA.TC.TAC.C..GTTTGA...................TGG.....................T.G................A.......
GY-0      ...............................................................................................................................................T..........................
HODJA     ......A....TAT.T.AG.CGTG.AG........C.....T...T..................C...............................TTGA...................TGG.....................T.....................T...A
ITA-0     .........A.TA....AG.CGTG.AG..TT..AA...........AGATCGCC.ATAAAC.CCCTTTT..CTA..A..GGGAATC.TA.....GTT.GA......T.G.AAAAG....T....C..................T............A............A
KA-0      ...........T.....AG.CGTGTAG..TT..AA....T..........CG.C..T..A....C...T..CTA..A..GGGA.............TTGA...................TGG.....................T..........................
KAS-1     ......A....TAT.T.AG.CGTG.AG........C.....T...T..................C....................C.........................................................T.G..........A.A....G.....A
KIL-0     .T..T......T..........C..A..C......CC.A.GTA.......CG.C..T..A....C...T.GCTA..A..GGGA.TC.TAC.C..GTTTGA...................TGG.....................T.G........................
KONDARA   ......R....TAY.Y.AG.CGTG.AG........C.....T...T..................C...............................TTGA...................TGG.....................T.....................T...A
LER-0     ..C........T.....AGTCGTG.AG........CC.............CG.CG.T..A....C...T.GCTA..AG.GGGA.............TTGA...................TGG.....................T.............G......A.....
LIP-0     ...........T.....AGTCGTG.AG..TT..AA......T........CG.C..T..A....C...T..CTA..A..GGGA.............TTGA...................TGG.....................T............A............A
LM-2      ......................C.......................AGATCGCC.ATAAAC.CCCTTTT..C.AC.A..GGGA...................GA.......................................T.G........................
MA-0      ..........................................................................................................................................................................
MRK-0     ...........T.....AG.CGTG.AG..........C.................................CTA..A.TGGGA..C.......T.................................................T..........................
MT-0      ..........................................................................................................................................................................
NO-0      ...........T.....AG.CGTGTAG..TT..AA....T..........CG.C..T..A....C...T..CTA..A..GGGA.............TTGA...................TGG.....................T............A..........T.A
PA-1      ...........T.....AG.CGTG.AGA...............A..AGATCGCC.ATAAAC.CCCTTTT..C.AC.A..GGGA..C.........................................................T.........A..A...........AA
PETERGOF  ..C........T.....AGTCGTG.AG........CC.............CG.CG.T..A....C...T.GCTA..AG.GGGA.............TTGA...................TGG.....................T.............G......A.....
PI-0      ...........T.....AG.CGTGTAG........CC.............CG.CG.T..A....C...T.GCTA..A..GGGA.............TTGA...................TGG.....................T............A..CTG..A.....
PLA-0     .T.....A...T....A....GTG.AG....T..............AGATCGCC.ATAAAC.CCCTTTT..................C...C....T.GAGA...G.TGTAAA.GAATTT..TG.CCTA.........................................
POG-0     .T.........T.....AG.CGTG.AG...................AGATCGCC.ATAAAC.CCCTTTT..................C...C....T.GAGA...G.TGTAAA.GAATTT..TG.CCTA..............T.G..................A.T...
SEI-0     ...........T.....AG.CGTGTAG..TT..AA....T..........CG.C..T..A....C...T..CTA..A..GGGA.............TTGA...................TGG.....................T..........................
SORBO     ...........T.....AGTCGTG.AG..TT..............................C..C....T.C.ACAA..GGGAATC.TA...T.GTTTGA....G..............TGG.....................T............A............A
SU-0      ...........T...A.AG.CGTGTAG..TT..AA...........AGATCGCC.ATAAAC.CCCTTTT..CTA..A.TGGGA..C.......T..T.GA......T.G.AAAAG....T....C....TAAGTGTCAGTCTCTT.GTGACAA.ATA..CTG.....T.A
TAC       ...........T.....AG.CGTG.AG...................AGATCGCC.ATAAAC.CCCTTTT..................C...C....T.GAGA...G.TGTAAA.GAATTT..TG.CCTA..............T.G..................A.T...
TSU-1     ...........T.....AG.CGTGTAG..TT..AA....T..........CG.C..T..A....C...T..CTA..A..GGGA.............TTGA...................TGG.....................T.............G......A.....
WL-0      ...........TA....AG.CGTGTAG..TT..AA....T..........CG.C..T..A....C...T..CTA..A..GGGA.............TTGA...................TGG.....................T..........................
YO-0      ...........T...A.AG.CGTGTAG..TT..AA...........AGATCGCC.ATAAAC.CCCTTTT..CTA..A.TGGGA..C.......T..T.GA......T.G.AAAAG....T....C....TAAGTGTCAGTCTCTT.GTGACAA.ATA..CTG.....T.A
```

FIGURE 3.—Prettyplot of all 170 polymorphic sites. Positions are indicated by numbers, which should be read top to bottom. At each site a dot indicates agreement with the nucleotide shown in the top row.

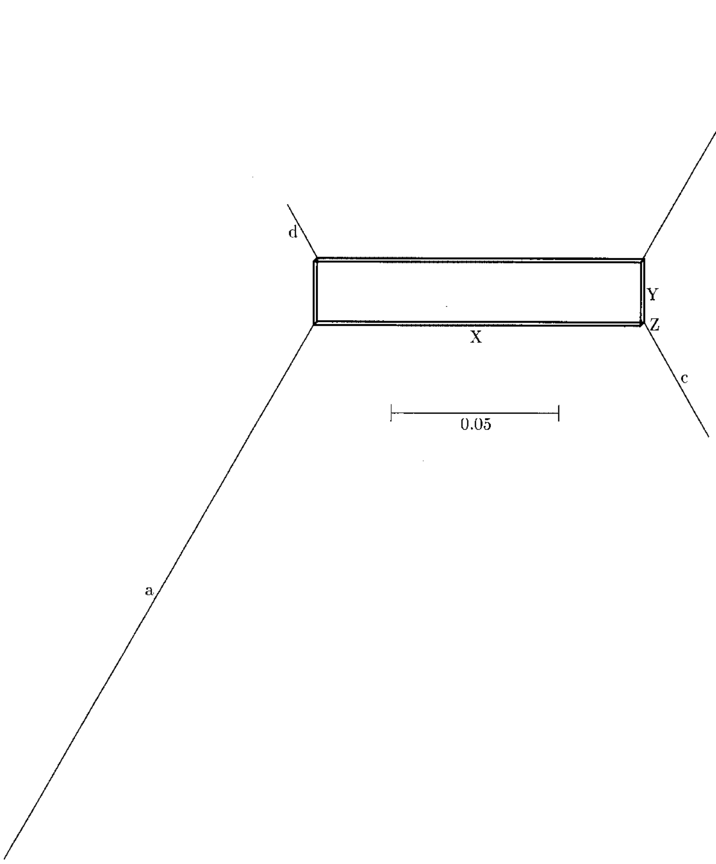

FIGURE 4.—Statistical geometry phylogeny for the combined nucleotide data. *X*, *Y*, and *Z* indicate the dimensions of the three-dimensional box from which the terminal branches (a–d) stick out. If the data were tree-like, the small *Z*-dimension as well as the larger *Y*-dimension would be zero. The bar indicates the number of substitutions per polymorphic site.

tral box, collapse dimensions *Z* and *Y*, or collapse dimensions *Y* and *Z*. In other words, a statistical geometry graph simultaneously represents the three unrooted trees that can be formed from four taxa. If no recombination has taken place, only one of these three possible trees should be supported by the data. High support for all three possible trees is indicated by a large central box.

For our data the deviation from the ideal tree topology was considerable (Figure 4), and we did not attempt to further reconstruct the phylogenetic history of the region.

Disequilibrium as a function of distance: Given that over a stretch of 170 kb the phylogeny of *A. thaliana* does not conform to a tree, reciprocal recombination or gene conversion has probably contributed considerably to the evolution of this species. Under reciprocal recombination the disequilibria between pairs of polymorphic sites are expected to fall off exponentially with distance. In contrast, gene conversion should generate no distance effect on disequilibria, if the average tract length is short.

We started our investigation of the relationship between distance and disequilibrium by grouping the single-nucleotide polymorphisms (SNPs) into 24 "blocks" as outlined in MATERIALS AND METHODS. Pairwise linkage disequilibria between these blocks were negatively correlated with distance ($r = -0.11$, $P = 6 \times 10^{-5}$; Figure 5A). However, if only haplotype pairs with four alleles were included in the analysis, *i.e.*, allele pairs where a recombination event could be detected, the negative correlation between distance and linkage disequilibrium turned positive ($r = 0.353$, $P = 10^{-5}$; Figure 5B). There is no neutral mechanism that results in a significant positive correlation between distance and disequilibrium. When we removed the one locus with a significant Tajima's *D* from the analysis (locus 6), the correlation between distance and disequilibrium vanished altogether ($r = 0.07$, $P > 0.05$; Figure 5C). This indicated that reciprocal recombination may not have been the primary mechanism for exchanging homologous DNA in the region.

To quantify the mode of recombination more directly, we applied a statistical test designed to distinguish between reciprocal recombination and gene conversion (WIEHE *et al.* 2000).

Gene conversion *vs.* reciprocal recombination: We carried out coalescent simulations with a recombination rate of one-tenth the rate of mutation, which appears to be a reasonable value given estimates in the literature (KUITTINEN and AGUADÉ 2000) and the observation of genome-wide linkage disequilibrium (SHARBEL *et al.* 2000). In our simulations we distributed this "effective" rate of recombination between reciprocal recombination and gene conversion. A graph of the mean value

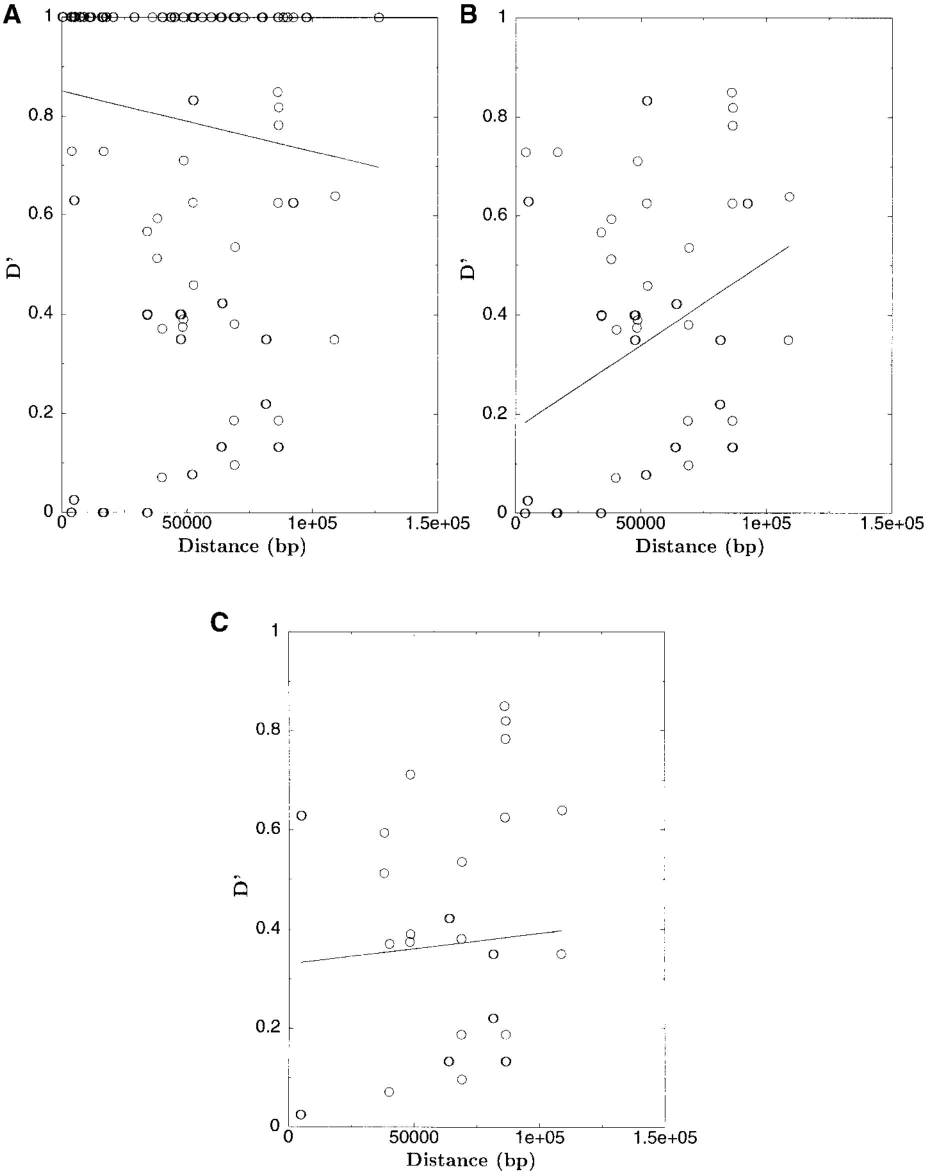

Figure 5.—Linkage disequilibrium as a function of distance. (A) All pairs of blocks of haplotypes included in the analysis. (B) Only those pairs of blocks are included where all four possible haplotypes were present, *i.e.*, where a recombination event certainly has taken place. (C) Same as B, except that the nonneutral locus 6 was removed from the analysis.

of sign $\tilde{Q}$, the sign of the random variable $\tilde{Q}$, as a function of the percentage of gene conversion returned a maximum-likelihood estimate of 90% conversion (Figure 6).

DISCUSSION

Completely asexual reproduction halves the rate of adaptation compared to panmixis and is therefore usually regarded as a rare exception, if it exists at all (Fisher 1930/1999, p. 123). This may appear surprising, given the large number of selfing plant species and other asexual organisms, including bacteria. However, in most selfing plants inbreeding is not complete and even the existence of purely clonal bacterial populations has been doubted (Feil *et al.* 2001). Different accessions of *A. thaliana* can be crossed in the laboratory, which forms the basis of the large amount of classical mapping work carried out using this organism. However, in the wild *A. thaliana* is a selfer with a very low outcrossing rate of 0.3% (Abbot and Gomes 1989). Recent studies of this plant's molecular population genetics suggested that in spite of its selfing habit, it underwent recombination rather frequently (Kuittinen and Aguadé 2000), leading to a decay of linkage disequilibrium in worldwide samples over ~250 kb (Nordborg *et al.* 2002). In this study we contribute to the clarification of the apparent contradiction between selfing and the molecular data.

Nucleotide polymorphism: The genetic diversity in the *MAM* region is highly variable (Figure 2). In 13 out of the 14 cases the polymorphisms do not contradict neutral expectations. The one exception (locus 6, Figure 2) is currently annotated as a gene of unknown function. Every new genome that is sequenced reveals a large number of predicted genes to which no function can be assigned. Given that sequencing is usually easier than elucidating a gene's function, comparative se-

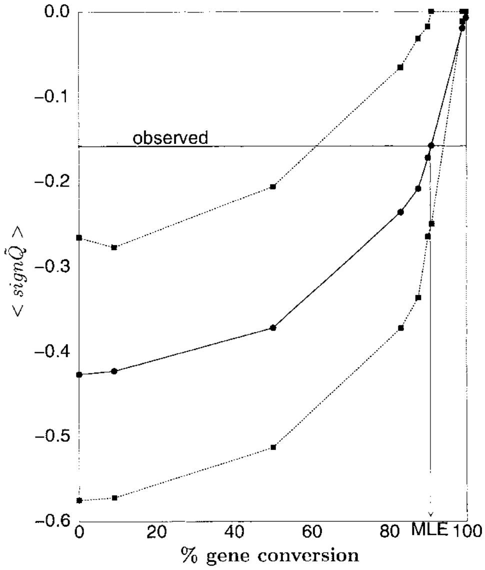

FIGURE 6.—The mean of the test statistic sign $\tilde{Q}$as a function of the extent of gene conversion. Plotted are mean values (●) and 50% confidence intervals (■). All blocks of polymorphisms were included in the analysis and the corresponding maximum-likelihood estimate (MLE) for gene conversion is ~90%. See MATERIALS AND METHODS for details on the test statistic $\tilde{Q}$.

quencing combined with tests of neutrality might point to those genes whose products are most relevant to an organism's biology.

Multilocus disequilibrium: If multiple loci have been investigated, linkage disequilibrium can be assessed either by performing pairwise tests or by calculating the overall linkage disequilibrium. Pairwise tests are difficult to interpret, as they are not independent from each other. The test based on the mismatch distribution used in this study does not suffer from this uncertainty about its interpretation (HAUBOLD *et al.* 1998). Moreover, it leads to the discovery of strong linkage disequilibrium not only in our data set, but also in a set of genome-wide AFLP markers (SHARBEL *et al.* 2000). A lack of genome-wide linkage disequilibrium as suggested by MIYASHITA *et al.* (1999) would be hard to reconcile with the selfing habit of *A. thaliana* and previous findings in other selfing plant species (BROWN *et al.* 1980).

Phylogeny: The average phylogeny differed from an ideal tree topology, which indicated that there was substantial recombination in the region (Figure 4). This difference becomes more pronounced if the genome-wide AFLP data published by SHARBEL *et al.* (2000) is subjected to statistical geometry (Figure 7). This is not surprising, since disequilibrium decreases exponentially with distance. However, even in this situation the loci display significant genome-wide linkage disequilibrium (SHARBEL *et al.* 2000). Having rejected the two extreme hypotheses of no recombination and of linkage equilibrium, we were interested in investigating the rate and mode of recombination.

Disequilibrium as a function of distance: It is clear that linkage disequilibrium should reflect genetic distance rather than physical distance. However, genetic positions are rather unreliable over short distances and hence we have used physical positions as a substitute (NORDBORG *et al.* 2002).

In the *MAM* region linkage disequilibrium apparently decreases with distance (Figure 5A). However, a positive correlation with distance was observed when we analyzed only pairs of blocks displaying all four possible haplotypes (Figure 5B). The puzzle of finding a positive correlation was resolved when we removed the one locus with significant evidence for selection from the sample. The resulting data set showed no correlation between distance and disequilibrium (Figure 5C). It is clear that three haplotypes can be generated by mutation alone, while four haplotypes between two markers must be the result of recombination, assuming no recurrent mutation. Hence, Figure 5C shows a sample that has certainly been shaped by recombination, while in Figure 5A the pairs of positions may or may not have been affected by recombination. Nevertheless, under neutrality and reciprocal recombination the two samples should yield a similar decay of linkage disequilibrium with distance. This suggests that gene conversion has shaped the distribution of polymorphisms in this region.

Mode of recombination: Gene conversion has been at the center of recent empirical and theoretical population genetic studies. LANGLEY *et al.* (2001) investigated the extent of linkage disequilibrium in the *su(s)* and *su(*w^a*)* loci on the *Drosophila melanogaster* X chromosome that are located in a region of reduced crossing over. In spite of low reciprocal recombination, the authors observed a similar genomic scale of linkage disequilibrium at the *su(s)* and *su(*w^a*)* loci as found in regions with normal rates of crossing over. This suggests that gene conversion is high in this region (LANGLEY *et al.* 2001).

WIUF and HEIN (2000) have introduced gene conversion into coalescent models. These authors noted that there was no statistic available to assess the relative extent of recombination and gene conversion. Such a statistic, $\tilde{Q}$, has been provided by WIEHE *et al.* (2000) and we carried out coalescent simulations to explore the utility of $\tilde{Q}$ when applied to a data set such as ours.

These simulations were based on the assumption of neutrality, which may not apply throughout the region, especially at locus 6 (Figure 2). Removal of this locus from the plot of linkage disequilibrium as a function of distance resulted in zero correlation between the two variables (Figure 5C), which would be expected with high gene conversion rates.

We show that with an effective recombination rate of one-tenth the rate of mutation ($c/\mu = \frac{1}{10}$) and a 90% gene conversion rate the experimental data can be explained quite adequately (Figure 6). However, this calculation should be treated with caution, as the distribu-

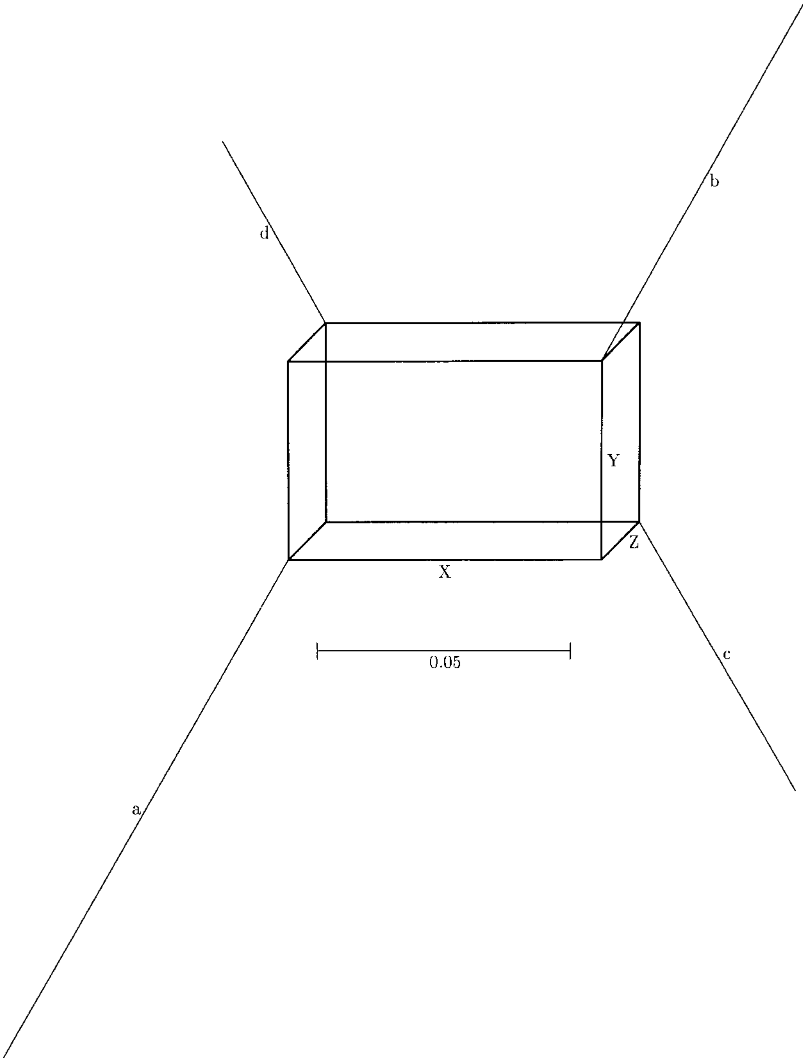

Figure 7.—Statistical geometry phylogeny for the 87 AFLP loci in 115 accessions in *A. thaliana* published by Sharbel *et al.* (2000). *X*, *Y*, and *Z* indicate the dimensions of the three-dimensional box from which the terminal branches (a–d) stick out. If the data were tree-like, the small *Z*-dimension as well as the larger *Y*-dimension would be zero. The bar indicates the number of substitutions per locus.

tion of $\tilde{Q}$ has a large variance and is far from normal. Nevertheless, this study demonstrates that there is no need to invoke a high rate of recombination to account for the experimental data; $c/\mu = \frac{1}{10}$ is sufficient. There is no contradiction between knowing that as a selfer *A. thaliana* must have a low rate of recombination and observing recombination events at the molecular level.

We thank Richard Hudson for providing his gene conversion simulation code and comments. Thanks are also due to U. Priedemuth for help with data handling. This work was supported by the Max-Planck-Gesellschaft; the Bundesministerium für Forschung, Germany, BMBF grant 0312705A to T.W.; the U.S. National Science Foundation (grant DEB-9527725); and the European Union.

LITERATURE CITED

Abbot, R. J., and M. F. Gomes, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411–418.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature **408:** 796–815.

Brown, A. H. D., D. Zohary and E. Nevo, 1978 Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch in Israel. Heredity **41:** 49–62.

Brown, A. H. D., M. W. Feldman and E. Nevo, 1980 Multilocus structure of natural populations of *Hordeum spontaneum.* Genetics **96:** 523–536.

Charlesworth, D., B. Charlesworth and G. A. T. McVean, 2001 Genome sequences and evolutionary biology, a two-way interaction. Trends Ecol. Evol. **16:** 235–242.

Efron, B., 1979 Bootstrap methods: another look at the Jackknife. Ann. Stat. **7:** 1–26.

Eigen, M., R. Winkler-Oswatitsch and A. Dress, 1988 Statistical geometry in sequence space: a method of quantitative comparative sequence analysis. Proc. Natl. Acad. Sci. USA **85:** 5913–5917.

Feil, E. J., E. C. Holmes, D. E. Bessen, M.-S. Chan, N. P. Day *et al.*, 2001 Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA **98:** 182–187.

Fisher, R. A., 1930/1999 *The Genetical Theory of Natural Selection*, variorum edition. Oxford University Press, Oxford.

Hanfstingl, U., A. Berry, E. A. Kellogg, I. T. Costa III, W. Rüdiger *et al.*, 1994 Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics **138:** 811–828.

Haubold, B., and R. R. Hudson, 2000 Lian 3.0: detecting linkage disequilibrium in multilocus data. Bioinformatics **16:** 847–848.

Haubold, B., M. Travisano, P. B. Rainey and R. R. Hudson, 1998 Detecting linkage disequilibrium in bacterial populations. Genetics **150:** 1341–1348.

Hilliker, A. J., G. Harauz, M. Gray, S. H. Clark and A. Chovnick, 1994 Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster.* Genetics **137:** 1019–1026.

Hudson, R. R., 1994 Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. J. Evol. Biol. **7:** 535–548.

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.

Kawabe, A., H. Innan, R. Terauchi and T. Miyashita, 1997 Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana.* Mol. Biol. Evol. **14:** 1303–1315.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

Kroymann, J., S. Textor, J. Tokuhisa, K. Falk, S. Bartram *et al.*, 2001 A gene controlling variation in *Arabidopsis thaliana* glucosinolate composition is part of the methionine chain elongation pathway. Plant Physiol. **127:** 1077–1088.

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

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen and J. M. Braverman, 2001 Linkage disequilibria and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster X* chromosome. Genetics **156:** 1837–1852.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Manly, B. F. J., 1994 *Multivariate Statistical Methods: A Primer*, Ed. 2. Chapman & Hall, London.

Maynard Smith, J., 1989 Trees, bundles or nets? Trends Ecol. Evol. **4:** 302–304.

Maynard Smith, J., N. H. Smith, C. G. Dowson and B. G. Spratt, 1993 How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:** 4384–4388.

Miyashita, N. T., A. Kawabe and H. Innan, 1999 DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. Genetics **152:** 1723–1731.

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.

Parham, P., E. J. Adams and K. L. Arnett, 1995 The origins of HLA-A,B,C polymorphism. Immunol. Rev. **143:** 141–180.

Purugganan, M. D., and J. I. Suddith, 1998 Molecular population genetics of the *Arabidopsis CAULIFLOWER* regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. Proc. Natl. Acad. Sci. USA **95:** 8130–8134.

Purugganan, M. D., and J. I. Suddith, 1999 Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the APETALA3 and PISTILLA genes of *Arabidopsis thaliana.* Genetics **151:** 839–848.

Rozen, S., and H. Skaletsky, 1998 Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

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

Sharbel, T. F., B. Haubold and T. Mitchell-Olds, 2000 Genetic isolation by distance in *Arabidopsis thaliana:* biogeography and postglacial colonization of Europe. Mol. Ecol. **9:** 2109–2118.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman and J. Bergelson, 1999 Dynamics of disease polymorphism at the *Rpm1* locus of *Arabidopsis.* Nature **400:** 667–671.

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

Wiehe, T., J. Mountain, P. Parham and M. Slatkin, 2000 Distinguishing recombination and intragenic gene conversion by linkage disequilibrium patterns. Genet. Res. **75:** 61–73.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Communicating editor: S. W. Schaeffer