Sequence Variation and Haplotype Structure Surrounding the Flowering Time Locus *FRI* in *Arabidopsis thaliana*

Jenny Hagenblad and Magnus Nordborg

Department of Genetics, Lund University, Sölvegatan 29, S-223 62 Lund, Sweden

Manuscript received August 17, 2001
Accepted for publication January 31, 2002

**ABSTRACT**

Linkage disequilibrium in highly selfing organisms is expected to extend well beyond the scale of individual genes. The pattern of polymorphism in such species must thus be studied over a larger scale. We sequenced 14 short (0.5–1 kb) fragments from a 400-kb region surrounding the flowering time locus *FRI* in a sample of 20 accessions of *Arabidopsis thaliana*. The distribution of allele frequencies, as quantified by Tajima’s *D*, varies considerably over the region and is incompatible with a standard neutral model. The region is characterized by extensive haplotype structure, with linkage disequilibrium decaying over 250 kb. In particular, recombination is evident within 35 kb of *FRI* in a haplotype associated with a functionally important allele. This suggests that *A. thaliana* may be highly suitable for linkage disequilibrium mapping.

---

**VARIATION** at a particular site in the genome is the result of mutations occurring along the branches of the genealogical tree relating the homologous copies of that site. Recombination allows different sites to have different genealogical trees. This makes it possible to gain insight into the stochastic process that gave rise to the trees. Without recombination, data would reflect a single realization of this process, making statistical inference a questionable project. In general, linked sites will have correlated trees, the strength of the correlation depending on the genetic distance between the loci. This correlation in genealogy between sites may be reflected in the pattern of variation at the sites, giving rise to linkage disequilibrium (e.g., Nordborg and Tavaré 2002).

Recombination is a powerful force in population genetics. The probability of a neutral mutation is typically estimated to be on the order of $10^{-8}$–$10^{-9}$/bp/meiosis in eukaryotes (e.g., Li 1997). Recombination probabilities are of the same order of magnitude: In humans, for example, 1 cM corresponds to ~1 Mb on average; i.e., the probability of recombination per base pair per meiosis must be on the order of $10^{-8}$ on average. It follows from basic population genetic theory that there will on average be as many recombination events in a sample of sequences as there are segregating sites (Nordborg 2000). Theory also predicts that most of these recombination events will be undetectable (Hudson and Kaplan 1985). In particular, it is important to realize that almost all methods that exist for detecting recombination in sequence data are designed to detect recombination when it is rare (as in horizontal transfer between bacterial lineages, for example) and are almost powerless when faced with “normal” recombination (Maynard Smith 1999). For example, whereas low levels of recombination (compared to the level of polymorphism) lead to distinctive “runs” of polymorphic sites that are incompatible with a single genealogical tree, high levels of recombination destroy most such patterns. As a consequence, it becomes very difficult to determine whether incompatible polymorphisms are caused by recombination or mutational hot spots (Templeton et al. 2000).

In general, the more polymorphic sites per recombination event, the more information we have about the underlying genealogical trees and recombination events (Hudson and Kaplan 1985). This is very important in the context of linkage disequilibrium mapping (Long and Langley 1999). A simple way of increasing the ratio of polymorphism to recombination is to look at partially selfing species (Nordborg 2000). The effective rate of recombination in such species is much reduced, essentially because recombination does not matter unless it occurs in outcrossed (heterozygous) individuals. As a result, extensive linkage disequilibrium and (some) clearly detectable recombination events are expected.

We decided to study the pattern of linkage disequilibrium and recombination in the highly selfing *Arabidopsis thaliana*. Previous polymorphism surveys in this species have found the expected high degree of linkage disequilibrium as well as evidence for some recombination (e.g., Hanfstringl et al. 1994; Innan et al. 1996). However, because these studies focused on single genes, little has

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AYO92417–AYO92756.

1Corresponding author: Molecular and Computational Biology, University of Southern California, 835 W. 37th St. SHS 172, Los Angeles, CA 90089-1340. E-mail: magnus@usc.edu
been known about the pattern in larger chromosomal regions. To fill this gap, we resequenced 14 short (0.5–1 kb) fragments from a 400-kb region on chromosome 4, in a sample of 20 accessions.

In addition to studying a larger chromosomal region, we were interested in the pattern of variation surrounding a polymorphic locus likely to be adaptively important. The pattern of variation in linked regions can reveal the history of selection on the alleles and can also be used for linkage disequilibrium mapping. *A. thaliana* is known to vary tremendously for flowering time (Laibach 1951; Napp-Zinn 1985; Karlsson et al. 1993; Nordborg and Bergelson 1999). Several studies indicated that much of this variation maps to two loci, FRI and FLC (e.g., Sanda et al. 1997). We chose to investigate the pattern of variation surrounding FRI and therefore selected fragments spanning the chromosomal region known to contain this locus. The 20 accessions were chosen so as to have a wide range of phenotypes. Ends of fragments were trimmed so as to remove sequences/deletions and substitutions, and these were treated as complex alleles or haplotypes. In the analyses below, loci with more than two alleles were left out found in *A. thaliana* genome sequence. Some of the fragments were chosen on the basis of preliminary results from other fragments. PCR conditions were optimized for each locus separately. PCR conditions and primer sequences are available upon request.

**DNA sequencing:** PCR products were purified with the QIAGEN (Valencia, CA) QIAquick PCR purification kit and used as templates for cycle sequencing with the fluorescent Bigdye Terminator ready reaction kit. Sequencing was done on ABI automated sequencers.

**Analysis:** All *A. thaliana* fragments were sequenced in both directions, and the results were aligned using “Sequencher” (http://www.genecodes.com) for base calling. Ends of fragments were trimmed so as to remove low-quality sequence. The resulting fragments are listed in Table 1, which also gives the putative gene content of each fragment on the basis of the TIGR annotation (http://www.tigr.org).

Sequences from different accessions were also aligned using “Sequencher,” with additional adjustments by hand in case of longer insertion/deletion polymorphisms. Most indels were treated as single events in the analysis; however, a few fragments contained complex polymorphisms that evidently involved repeated insertions/deletions and substitutions, and these were treated as complex alleles or haplotypes. In the analyses below, loci with more than two alleles were left out (the effect of doing this is slight). Finally, singleton

<table>
<thead>
<tr>
<th>Position</th>
<th>Length</th>
<th>Exons</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>151847–152427</td>
<td>571</td>
<td>(−355)–270, 480–(831)</td>
<td>F5I10.20</td>
<td>Unknown</td>
</tr>
<tr>
<td>207611–208160</td>
<td>550</td>
<td>(−1570)–(1039)</td>
<td>F5I10.24</td>
<td>Unknown</td>
</tr>
<tr>
<td>233725–234540</td>
<td>816</td>
<td>655–(959)</td>
<td>F6N23.19</td>
<td>Putative myb-like DNA-binding protein</td>
</tr>
<tr>
<td>255107–255556</td>
<td>450</td>
<td>39–136, 246–(504)</td>
<td>F6N23.28</td>
<td>Putative tetrahydrofolate synthase</td>
</tr>
<tr>
<td>264109–264632</td>
<td>524</td>
<td>16–135, 242–301, 400–(532)</td>
<td>F6N23.15</td>
<td>Putative potassium/H+ antiporter</td>
</tr>
<tr>
<td>264638–265102</td>
<td>465</td>
<td>(−129)–3, 185–321</td>
<td>F6N23.15</td>
<td>Putative potassium/H+ antiporter</td>
</tr>
<tr>
<td>268125–268682</td>
<td>558</td>
<td>(−98)–(840)</td>
<td>F6N23.25</td>
<td>FRI</td>
</tr>
<tr>
<td>269509–270579</td>
<td>1115</td>
<td>112–759</td>
<td>F6N23.25</td>
<td>FRI</td>
</tr>
<tr>
<td>273320–274338</td>
<td>1019</td>
<td>319–626, 790–850</td>
<td>F6N23.6</td>
<td>Putative RNA helicase</td>
</tr>
<tr>
<td>282466–282971</td>
<td>506</td>
<td>None</td>
<td>Intergenic</td>
<td></td>
</tr>
<tr>
<td>300982–301526</td>
<td>545</td>
<td>223–429, 516–(617)</td>
<td>F6N23.10</td>
<td>Unknown</td>
</tr>
<tr>
<td>301539–302035</td>
<td>497</td>
<td>(−41)–60, 143–(697)</td>
<td>F6N23.10</td>
<td>Unknown</td>
</tr>
<tr>
<td>376411–376940</td>
<td>None</td>
<td>Intergenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>233725–234540</td>
<td>816</td>
<td>655–(959)</td>
<td>F6N23.19</td>
<td>Putative myb-like DNA-binding protein</td>
</tr>
<tr>
<td>255107–255556</td>
<td>450</td>
<td>39–136, 246–(504)</td>
<td>F6N23.28</td>
<td>Putative tetrahydrofolate synthase</td>
</tr>
<tr>
<td>264109–264632</td>
<td>524</td>
<td>16–135, 242–301, 400–(532)</td>
<td>F6N23.15</td>
<td>Putative potassium/H+ antiporter</td>
</tr>
<tr>
<td>264638–265102</td>
<td>465</td>
<td>(−129)–3, 185–321</td>
<td>F6N23.15</td>
<td>Putative potassium/H+ antiporter</td>
</tr>
<tr>
<td>268125–268682</td>
<td>558</td>
<td>(−98)–(840)</td>
<td>F6N23.25</td>
<td>FRI</td>
</tr>
<tr>
<td>269509–270579</td>
<td>1115</td>
<td>112–759</td>
<td>F6N23.25</td>
<td>FRI</td>
</tr>
<tr>
<td>273320–274338</td>
<td>1019</td>
<td>319–626, 790–850</td>
<td>F6N23.6</td>
<td>Putative RNA helicase</td>
</tr>
<tr>
<td>282466–282971</td>
<td>506</td>
<td>None</td>
<td>Intergenic</td>
<td></td>
</tr>
<tr>
<td>300982–301526</td>
<td>545</td>
<td>223–429, 516–(617)</td>
<td>F6N23.10</td>
<td>Unknown</td>
</tr>
<tr>
<td>301539–302035</td>
<td>497</td>
<td>(−41)–60, 143–(697)</td>
<td>F6N23.10</td>
<td>Unknown</td>
</tr>
<tr>
<td>376411–376940</td>
<td>None</td>
<td>Intergenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>596659–597130</td>
<td>447</td>
<td>(49)–60, 143–(697)</td>
<td>F6N23.10</td>
<td>Unknown</td>
</tr>
<tr>
<td>300982–301526</td>
<td>545</td>
<td>223–429, 516–(617)</td>
<td>F6N23.10</td>
<td>Unknown</td>
</tr>
<tr>
<td>301539–302035</td>
<td>497</td>
<td>(−41)–60, 143–(697)</td>
<td>F6N23.10</td>
<td>Unknown</td>
</tr>
<tr>
<td>376411–376940</td>
<td>None</td>
<td>Intergenic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All coordinates are based on the available *Col* sequence. The positions of predicted exons are given with respect to the relevant fragment. Gene designations are based on bacterial artificial chromosomes (BACs): F5I10.20 refers to gene 20 on BAC F5I20, for example. See http://www.tigr.org for details.

**METHODS**

**Sample:** The following accessions were included in the sample: Alnus crispa, *Col*, *Dem-4*, *Got-32*, *Kent*, *Köl*, *Kondara*, *Ks-9*, *Ler*, *Lisse*, *Lund*, *MT-0*, *NC-6*, *Pu-2-3*, *Pu-2-8*, *Rsch-4*, *Shakhbazar*, *Tamm-46*, *Tsu-0*, and *Vimmerby*. Further information about these accessions can be found in Nordborg and Bergelson (1999). In addition, single individuals of *A. (Cardaminopsis) arenosa*, *Cardamine hirsuta*, *Erophila verna*, and *Capsella bursa-pastoris* were used as outgroups. Genome DNA was extracted using standard protocols.

**PCR amplification of fragments:** PCR primers were constructed by applying primer-selection software to suitably spaced intervals of the published *A. thaliana* genome sequence. Some of the fragments were chosen on the basis of preliminary results from other fragments. PCR conditions were optimized for each locus separately. PCR conditions and primer sequences are available upon request.

**DNA sequencing:** PCR products were purified with the QIAGEN (Valencia, CA) QIAquick PCR purification kit and used as templates for cycle sequencing with the fluorescent Bigdye Terminator ready reaction kit. Sequencing was done on ABI automated sequencers.

**Analysis:** All *A. thaliana* fragments were sequenced in both directions, and the results were aligned using “Sequencher” (http://www.genecodes.com) for base calling. Ends of fragments were trimmed so as to remove low-quality sequence. The resulting fragments are listed in Table 1, which also gives the putative gene content of each fragment on the basis of the TIGR annotation (http://www.tigr.org).

Sequences from different accessions were also aligned using “Sequencher,” with additional adjustments by hand in case of longer insertion/deletion polymorphisms. Most indels were treated as single events in the analysis; however, a few fragments contained complex polymorphisms that evidently involved repeated insertions/deletions and substitutions, and these were treated as complex alleles or haplotypes. In the analyses below, loci with more than two alleles were left out (the effect of doing this is slight). Finally, singleton
TABLE 2
The number of segregating sites in each fragment

<table>
<thead>
<tr>
<th>Exon</th>
<th>Noncoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragment</td>
<td>No. of sites</td>
</tr>
<tr>
<td>151847–152427</td>
<td>362</td>
</tr>
<tr>
<td>207611–208160</td>
<td>550</td>
</tr>
<tr>
<td>233725–234540</td>
<td>134</td>
</tr>
<tr>
<td>255107–255556</td>
<td>303</td>
</tr>
<tr>
<td>264109–264632</td>
<td>305</td>
</tr>
<tr>
<td>264638–265102</td>
<td>140</td>
</tr>
<tr>
<td>268125–268682</td>
<td>558</td>
</tr>
<tr>
<td>269509–270579</td>
<td>648</td>
</tr>
<tr>
<td>273320–274338</td>
<td>369</td>
</tr>
<tr>
<td>282466–282971</td>
<td>0</td>
</tr>
<tr>
<td>300982–301526</td>
<td>237</td>
</tr>
<tr>
<td>301539–302035</td>
<td>415</td>
</tr>
<tr>
<td>376411–376940</td>
<td>0</td>
</tr>
<tr>
<td>596659–597130</td>
<td>275</td>
</tr>
</tbody>
</table>

Numbers are given in the form a/b/c, where b is the number of sites with singleton mutations, c is the number of sites with more than two alleles, and a is all other polymorphic sites. NA, not applicable.

polymorphisms (i.e., those with frequency 1/20) were left out of all analyses of linkage disequilibrium. Note that some fragments contained only singletons.

To improve the power to detect recombination, the ancestral state of each polymorphic site was determined by comparison with A. arenosa, except in the case of one fragment that could not be amplified, where the other species were used instead. No attempt was made to obtain complete sequences from the outgroup species, as this would have required cloning of all fragments (A. arenosa is a tetraploid outcrosser, for example).

Simulations: The data were compared to results of simulations using the ancestral recombination graph (as described in Nordborg 2001). Two sets of simulations were run. First, to study the properties of samples of short fragments, 10,000 samples of size n = 20 were generated for each of the parameter combinations shown in Table 3. Second, to exemplify the behavior of long-range linkage disequilibrium, five such samples were generated for each of the parameter combinations in Figure 6.

RESULTS AND DISCUSSION

The 17 amplicons (14 contiguous fragments) were successfully amplified and sequenced in all accessions. Comparison of our Col sequence with that obtained by the Arabidopsis Genome Initiative (AGI) revealed one discrepancy in 8627 bp. This discrepancy was judged to be due to an error in our base calling. Similarly, three discrepancies were observed in 445 bp of overlap between contiguous amplicons (for each of the 20 accessions), suggesting an error rate of $\sim 10^{-4}$.

As shown in Table 1, 12 of the 14 fragments turned out to be located in putative genes. Two of these covered single exons, while the remaining 10 covered a mixture of coding and noncoding sequence. The pattern of polymorphism in our data is consistent with the genome annotation: Even though the number of coding and noncoding bases examined were roughly equal, insertion/deletion (indel) polymorphisms were largely limited to noncoding regions (see Table 2). Out of a minimum of 22 such polymorphisms, only 3 were in coding regions. Of these, 2 were single-base-pair indels present in single individuals and could well represent genuine deleterious mutations. The remaining indel was an in-frame Asn repeat. Indels in noncoding regions, on the other hand, were frequently both long and highly polymorphic.

Levels of polymorphism: The level of polymorphism in a sample of sequences can be quantified by $\hat{\theta}_w$, Watterson’s estimate of the neutral mutation parameter $\theta$ (Watterson 1975). To allow comparison between different regions, we estimate $\theta$ per site. Furthermore, we consider noncoding regions separately from coding regions, and, in the latter, distinguish between synonymous and nonsynonymous sites on the basis of whether a change in the DNA sequence would lead to a change in the amino acid sequence (see, e.g., Li 1997). In the present data, $\hat{\theta}_w = 6 \times 10^{-3}$ per site for both synonymous and noncoding sites, and $\hat{\theta}_w = 2 \times 10^{-3}$ for nonsynonymous sites. These values are consistent with previous
estimates (see Aguadé 2001). However, as shown in Figure 1, the estimates vary considerably between the fragments, both in terms of the relationship between the levels of polymorphism for different types of sites and in terms of the absolute levels. It is worth noting here that these estimates should be positively correlated. Under neutrality, \( \hat{\theta}_w \) depends strongly on the total branch length of the underlying genealogical tree, and the total branch lengths of linked genealogies should be more similar than the total branch lengths of unlinked genealogies. Thus the variability is more notable. This is especially significant when comparing different types of sites in the same fragment, which should have very similar genealogies. Interestingly, \( \hat{\theta}_w \) for synonymous and noncoding sites seems to be much more correlated than either is with \( \hat{\theta}_w \) for nonsynonymous sites (Figure 1). This may reflect different levels of selective constraint in the latter.

**Allele-frequency distribution:** Another important characteristic of the pattern of polymorphism is the distribution of allele frequencies. Variation in this distribution can be summarized by comparing Watterson’s estimator of \( \theta \), \( \hat{\theta}_w \), with Tajima’s estimator, \( \hat{\theta} \) (Tajima 1983). The latter is based on the average number of pairwise differences between sequences and is therefore very sensitive to the topology of the underlying genealogy, which is reflected in the distribution of allele frequencies. Figure 2 shows the behavior of these two estimators across the region. The difference between them can be quantified using Tajima’s \( D \) statistic, which has expectation zero under the standard neutral model, but will tend to be negative if there are too many rare alleles and positive if there are too many common ones (Tajima 1989). Since such deviations can be caused by selection, Tajima’s \( D \) is often used in tests for selection.

Figure 3 shows how \( D \) varies across the studied region. Note that a few values are “significantly” different from zero by the criteria typically used in molecular evolution studies. More importantly, \( D \) fluctuates wildly from highly negative to highly positive values. The variance of \( D \) observed here is 1.96: Simulations (see methods) show that the probability of observing such a high variance among 14 independent realizations of \( D \) under the standard neutral model is on the order of \( 10^{-3} \). Thus, the variance would seem to be too large to be compatible with the standard neutral model. It should be noted that the effect of linkage on \( D \) is not known; however, it seems highly likely that \( D \) for linked fragments should again be positively correlated, thus increasing the significance of the deviation from the standard model.

What is the cause of this deviation? Since, as is dis-
Figure 4.—Each point represents a comparison between a pair of polymorphic sites. The point is black if the two sites are compatible with a tree topology by the four-gametes test (see text) and white if not. The axes represent the physical position of the sites (so that points on the diagonal correspond to comparisons of each site with itself). Because of the uneven spacing of the sites, the distance between the points is not proportional to true distance. The actual position of each site is shown at the bottom. Blue and red points correspond to black and white points, respectively, but refer to two functionally important mutations in *FRI* (see text): The one giving rise to *friCol* is located at ~269 kb; the one giving rise to *friLer* is located at ~208 kb.

cussed below, the sample includes alleles of *FRI* with extremely strong phenotypic effects, the region is *a priori* unlikely to have evolved neutrally. Nonetheless, we do not think it is warranted to conclude that selection on *FRI* is responsible for the observed pattern. First of all, the observed pattern of variation in *D* is not immediately suggestive of any particular selective scenario. Second, and more importantly, we do not know whether the observed pattern is in fact typical for the genome. It has become standard practice in population genetics to “test for selection” by calculating a summary statistic such as Tajima’s *D* and finding that it deviates significantly from its neutral expectation (reviewed in Kreitman 2000). This approach assumes that most of the rest of the genome is in fact evolving according to a standard neutral model. Leaving aside the (important) issue of whether selection perhaps affects every site in the genome (Gillespie 2000), a very serious problem with the standard neutral model is that is assumes that there is no population structure—an assumption that is almost certainly false for most species. Since it is well known that population structure can affect variation in ways that mimic selection (*e.g.*, Nordborg 2001), it would clearly be prudent to reject neutrality for a particular locus by comparing it to the actual pattern of variation in the rest of the genome, rather than to an idealized distribution. For obvious reasons, this has hitherto not been possible; equally obviously, it soon will be.

**Recombination and linkage disequilibrium:** A simple way to look for recombination is the “four-gamete” test (Hudson and Kaplan 1985): If, for a pair of diallelic sites, all four possible haplotypes are present, then this is incompatible with unique mutations on a single genealogical tree: Either there must have been a repeat mutation at one of the sites or there must have been recombination between the sites. Because the probability of recombination increases with the physical distance between sites, recombination may give rise to a distinctive pattern where closely linked sites are compatible and more distantly linked sites are mostly incompatible. Multiple mutations should not normally give rise to such a pattern, since the probability of a mutation affecting one member of a pair loci should not depend on the distance between the loci (Jakobsen and Easteal 1996).

As pointed out above, recombination is expected to generate a distinctive pattern only when it is rare relative
Next we consider another reflection of recombination, namely the decay of linkage disequilibrium. The top graph of Figure 5 shows linkage disequilibrium between all pairs of polymorphic sites as a function of the distance between the sites. The bottom graph shows a histogram of the distribution of \( P \) values (under Fisher’s exact test of association) for each distance interval. Linkage disequilibrium is extensive, but decays sharply with distance within the surveyed region (i.e., over 250 kb).

These data agree qualitatively with model predictions for selfing species (NORDBORG 2000). First, the amount of linkage disequilibrium is considerably more extensive than what is typically observed in outcrossing species. For example, in Drosophila, linkage disequilibrium typically decays within a few kilobases, not 250 kb (LANGLEY et al. 2000). Second, linkage disequilibrium is far from genome wide. \( A. \) thaliana has sometimes been treated as if it were an effectively clonal species, with accessions (or “ecotypes”) evolving in a tree-like fashion. This is plainly not the case. Although linkage disequilibrium is extensive, it does decay. Heterozygous individuals are routinely observed in natural populations (E. A. STAHL, R. HUFFT, M. KREITMAN and J. BERGELSON, unpublished results), and there is no “phylogeny” of \( A. \) thaliana accessions (SHARBEL et al. 2000). The decay of linkage disequilibrium observed here appears to be consistent with the genome-wide pattern (NORDBORG et al. 2002).

What about quantitative agreement? The frequency of recombination in standard population genetics models is determined by the recombination parameter \( r \), which plays a role analogous to that of \( \theta \) for mutation. Several methods for estimating \( r \) from polymorphism data exist; however, they generally have poor statistical properties (WALL 2000), and none are well suited for the present type of data (short sequences with little recombination separated by large regions with no information). We therefore use simulations (see METHODS) and summary statistics. Like all other existing estimation methods, our simulations assume no population structure or selection, assumptions that we know to be false. The results should therefore be interpreted with caution.

First we consider the decay of linkage disequilibrium. Five genealogical histories were generated for each of five values of \( r \), and the decay of linkage disequilibrium was plotted as in the bottom graph of Figure 5. The results are shown in Figure 6. Note, first, that there is tremendous variation between realizations, at least for the lower values of \( r \). The reason for this is simple: The lower the rate of recombination, the more correlated the genealogies, and the bigger that variance between realizations. Nonetheless, a comparison of these results with Figure 5 suggests that \( r \) for the region studied is most likely \( \sim 40 \) and probably not \( > 80 \). This suggests that \( r \approx 0.2/\text{kb} \) or 0.1–0.2/fragment. As discussed earlier, \( \theta \) per fragment is estimated to be somewhere in the range of 1–3, depending on the length of the fragment and the proportion of coding to noncoding se-
Variation Surrounding FRI in A. thaliana

Figure 6.—The decay of linkage disequilibrium in data sets simulated under different assumptions about the population recombination parameter \( \rho \). These histograms should be compared with the histogram in Figure 5. The axes are the same. Each histogram was calculated from a single realization of the ancestral recombination graph with mutation parameter \( u = 50 \) and with the recombination parameter \( \rho \) specified at the top of each column.

sequence. Are these parameters compatible with the pattern of linkage disequilibrium observed within fragments? In particular, are they compatible with the somewhat surprising finding that recombination seems to have occurred in 2 of 14 fragments? To investigate this, we simulated a large number of fragment data sets using a range of values for \( \rho \) and \( \theta \). The results are shown in Table 3. Note first that whereas the probability of recombination having taken place in a sample depends only on \( \rho \), the probability of detecting this depends on \( \theta \) as well. In general, the higher the value of \( \theta \), the greater the probability of detecting a recombination event; however, a substantial fraction can never be detected (Hudson and Kaplan 1985).

It is clear from Table 3 that, even for \( \theta = 7 \), the \textit{a priori} probability of observing recombination in 2 of 14 segments is small unless \( \rho = 0.5 \) or so. Thus \( \rho = 0.1 \) would seem to be too low. If \( \rho = 0.5 \) per fragment, then \( \rho = 200 \) for the region in Figure 5, which contradicts the conclusions drawn from Figure 6. Pritchard and Przeworski (2001) reported a similar discrepancy—too little linkage disequilibrium on a fine scale and too much on a larger scale—in humans. This phenomenon deserves further study as more data become available.

Another question is whether the amount of recombination relative to mutation is compatible with the high degree of selfing in A. thaliana. It makes sense to consider the ratio of \( \rho/\theta \) because

\[
\frac{\rho}{\theta} = \frac{r}{u}
\]

where \( u \) is the neutral mutation probability per meiosis and \( r \) is the recombination fraction. This avoids dealing directly with the coalescent scaling constant (the “effective population size,” \( N_e \)). Both \( u \) and \( r \) can be estimated: \( u \) from sequence divergence between species and \( r \) using standard genetic methods (with the important caveat that the values of \( r \) that are of interest in population genetics are typically much smaller than can be directly estimated; see Andolfatto and Nordborg 1998). In a partially selfing organism,

\[
\frac{\rho}{\theta} = \frac{r}{u} (1 - F),
\]

where \( F \) is the inbreeding coefficient (Nordborg 2000). Thus, as long as \( u \) and \( r \) are the same, \( \rho/\theta \) is a selfer should be \((1 - F)\) of that in an outcrosser. If \( A. \)
thaliana is 99% selfing, as has been suggested (Abbot and Gomes 1989; Bergelson et al. 1998), then $F = 0.98$ and $(1 - F) = 2\%$. In the present data, $p/\theta$ seems to be roughly $1/10$, in agreement with previous studies (Kawabe and Miyashita 1999) and not incompatible with what is known about this ratio in outcrossing organisms (Nordborg et al. 2002). However, if $F \gg 0.98$, for instance, because of inbreeding resulting from population structure, then the ratio is too high, and the data may suggest that A. thaliana is, or has recently been, more outcrossing than is commonly believed.

**FRI haplotypes:** We selected a sample that included both early- and late-flowering accessions, hoping that (a) FRI would turn out to be responsible for a large fraction of the phenotypic variation and (b) the sample would include a sufficiently large number of alleles of each type to address questions about the history of selection on the alleles, as well as about the feasibility of linkage disequilibrium mapping. Since the initiation of this study, FRI was identified, and it became possible to identify the FRI alleles present in our sample, which turned out to have the following composition (Johnson et al. 2000):

1. The early flowering Col, Köl, and Mt-0 share a FRI loss-of-function allele, fri_col.
2. The early flowering Ler, Rsch-4, and Tsu-0 share another FRI loss-of-function allele, fri_col.
3. The remaining 14 accessions appear to have functional FRI alleles. Two of these, Shakhdara and Ks-9, are also quite early (Kondara and Pu-2-8 are intermediate). Other loci are likely to be involved here; the early flowering of Shakhdara seems to be caused by a recessive allele at the unlinked locus FLC (S. Ganéstam, J. Hagenblad, V. S. Rao, T. Kraft and M. Nordborg, unpublished results; S. Michaels, personal communication).

In other words, FRI did turn out to be responsible for a large fraction of the phenotypic variation; however, no early flowering allele had a frequency higher than three, which severely reduces our power to draw meaningful conclusions about the history of the alleles.

The pattern of recombination associated with the two functionally important FRI alleles is highlighted in Figure 4. Take fri_col first. On the telomeric side (left side in the coordinate system used), there is no evidence for recombination in 117 kb. On the centromeric side, a single incompatible site is found 32 kb from the mutation; however, this site is very likely a mutational hot spot (note that it is incompatible with all other sites). More plausible evidence for recombination is found at 108 kb. Thus fri_col may be associated with a haplotype that is well over 200 kb long in our sample. We note that this is one of the longest haplotypes in the data. In fact, if we consider all 1404 possible choices of 3 out of 20 accessions, no other combination shares such a long haplotype. This suggests that the 3 sampled copies of fri_col have an unusually recent common ancestor, perhaps as a result of positive directional selection. The relative recency of the fri_col allelic class is also suggested by the fact that there is only a single segregating site within this class (in addition to the incompatible sites in Figure 4, which may or may not be part of the allelic class, depending on where recombination occurred).

Next consider fri_col. Even if we conservatively ignore all isolated incompatible sites, there is clear evidence for recombination within 34 kb on the telomeric side and reasonable evidence for recombination within 109 kb on the other side. The haplotype associated with fri_col would thus appear to be considerably shorter than the one associated with fri_col. There is a single segregating site within this class in addition to the three isolated incompatible sites (this excludes the portions that have clearly undergone recombination).

A number of methods for estimating the ages of alleles exist; however, given the small amount of data in this study, such estimation is of questionable value. The picture is brighter if we turn from the uncertainties of historical inference to the more practical issue of linkage disequilibrium mapping. There is currently tremendous interest in using population association to map genes responsible for naturally occurring phenotypic information. The extent of linkage disequilibrium is very important in this context.
because it determines how dense a map is needed, on the one hand, and how finely loci may be mapped, on the other (Altschuler et al. 2000). In outcrossing organisms such as Drosophila (Long et al. 1998) or maize (Thornsberry et al. 2001), linkage disequilibrium appears to be restricted to a few kilobases, making it a realistic prospect that trait variation can be associated with particular sites, while at the same time making it unlikely that loci can ever be identified in genomic scans. In humans, the extent of linkage disequilibrium is subject to heated debate (e.g., Kruglyak 1999; Reich et al. 2001). The results presented here suggest that selfing organisms, like A. thaliana, rice, or barley, may be highly suitable for mapping by genomic scans for association. The presence of extensive, highly diverged haplotypes would seem to be ideal for modern inference methods (Morris et al. 2000; Toivonen et al. 2000), and the naturally inbred populations make it easy to obtain haplotype data as the complications of heterozygotes are avoided.

Finally, our data illustrate some of the limitations of linkage disequilibrium mapping rather well. First, whether a particular allele can be mapped or not depends on the history of that allele. History cannot be repeated. Thus, it is quite possible that the *fril* allele will still be surrounded by a haplotype that is several hundred kilobases long even in a much larger sample. If so, linkage disequilibrium mapping would not be particularly useful for this allele, as it is relatively painless to map genes to this scale in A. thaliana using traditional methods (such as recombinant inbred lines). However, sometimes it will work, as illustrated by the *fril* allele, where we found clear evidence for recombination within 34 kb in a sample of only three alleles. It is clear that a very large mapping population would be needed to achieve this kind of resolution using traditional methods.

Second, genetic heterogeneity is a very important issue for anyone interested in association mapping (Weiss and Terwilliger 2000). Early flowering in our sample was due to alleles at multiple loci as well due to multiple alleles at a single locus. It will clearly not be possible to map an allele unless it is responsible for a reasonably large fraction of the variation in the sample. Sampling is likely to play an important role in this context: We note that in our sample, none of the early flowering accessions from Central Asia appear to carry early flowering alleles of *FRI*, whereas almost all other early flowering accessions do.

We thank C. Dean, U. Johanson, and J. West for much help and for providing access to prepublication data without which this project would not have been possible. P. Artander and H. Ellegren graciously let us use their sequencing machines. We thank H. Innan for comments on the manuscript. This work was supported by grants from the Swedish Natural Sciences Research Council and the Erik Philip-Sörensens Foundation to M.N.

LITERATURE CITED


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


Nordborg, M., and J. Bergelson, 1999 The effect of seed and rosette cold treatment on germination and flowering time in *Arabidopsis thaliana*.


Communicating editor: O. Savolainen