DOI: 10.1534/genetics.103.024950

Linkage Disequilibrium Mapping of Arabidopsis *CRY2* Flowering Time Alleles

Kenneth M. Olsen,* Solveig S. Halldorsdottir,* John R. Stinchcombe,† Cynthia Weinig,‡ Johanna Schmitt† and Michael D. Purugganan*,[1]

*Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695, †Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912 and ‡Department of Plant Biology, University of Minnesota, Saint Paul, Minnesota 55108

Manuscript received November 25, 2003
Accepted for publication March 26, 2004

ABSTRACT

The selfing plant *Arabidopsis thaliana* has been proposed to be well suited for linkage disequilibrium (LD) mapping as a means of identifying genes underlying natural trait variation. Here we apply LD mapping to examine haplotype variation in the genomic region of the photoperiod receptor *CRYPTOCHROME2* and associated flowering time variation. *CRY2* DNA sequences reveal strong LD and the existence of two highly differentiated haplogroups (*A* and *B*) across the gene; in addition, a haplotype possessing a radical glutamine-to-serine replacement (A^S) occurs within the more common haplogroup. Growth chamber and field experiments using an unstratified population of 95 ecotypes indicate that under short-day photoperiod, the A^S and *B* haplogroups are both highly significantly associated with early flowering. Data from six genes flanking *CRY2* indicate that these haplogroups are limited to an ~65-kb genomic region around *CRY2*. Whereas the *B* haplogroup cannot be delimited to <16 kb around *CRY2*, the A^S haplogroup is characterized almost exclusively by the nucleotide polymorphisms directly associated with the serine replacement in *CRY2*; this finding strongly suggests that the serine substitution is directly responsible for the A^S early flowering phenotype. This study demonstrates the utility of LD mapping for elucidating the genetic basis of natural, ecologically relevant variation in Arabidopsis.

A major goal of modern evolutionary biology has been to understand the genetic basis of naturally occurring variation in complex traits. Linkage disequilibrium (LD) mapping of candidate gene associations is an emerging approach for identifying the genes underlying such phenotypic variation (Ardlie *et al.* 2002). Like the more established approach of quantitative trait locus (QTL) mapping, this technique infers associations between genotypes and phenotypic variation by examining genetic polymorphisms that have been shuffled into different genetic backgrounds through recombination. However, whereas QTL mapping considers only variation between two crossed individuals and relies solely on recombination events observed in their progeny, LD mapping exploits the phenotypic and genetic variation present across a natural population and draws inferences on the basis of past recombination events that have shaped the haplotype structure of that species (Nordborg and Tavaré 2002; Borevitz and Nordborg 2003). Specifically, the LD mapping approach tests for associations between phenotypic variation and haplotypes in a genomic region. This method has been successfully applied in studies of Drosophila (*e.g.*, Long *et al.* 1998), maize (Thornsberry *et al.* 2001), and humans (*e.g.*, Fullerton *et al.* 2000) to identify genes and specific polymorphisms within genes that underlie natural phenotypic variation.

The LD mapping strategy has been proposed to hold particular promise for identifying polymorphisms underlying trait variation in the wild model plant species *Arabidopsis thaliana* (Borevitz and Nordborg 2003). Because this species is predominantly selfing and shows very low effective recombination rates, its genome contains extensive blocks of LD (Nordborg *et al.* 2002; Borevitz and Nordborg 2003) and thus a well-defined haplotype structure for LD mapping. On the other hand, the physical length of haplotypes associated with such low effective recombination—up to 250 kb (Nordborg *et al.* 2002)—could potentially be a barrier to localizing causal polymorphisms, since all polymorphisms across a haplotype are linked and their phenotypic effects are indistinguishable.

Here we examine the utility of the LD mapping approach in *A. thaliana* for identifying and elucidating natural allelic variation that is associated with flowering time, a major life history trait in this species. The change from vegetative to reproductive development is a key event in the life history of plants (Simpson and Dean 2002). In *A. thaliana*, this transition to inflorescence production halts the production of rosette leaves formed

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY576055–AY576271.

[1]*Corresponding author:* Department of Genetics, Box 7614, 3513 Gardner Hall, North Carolina State University, Raleigh, NC 27695. E-mail: michael_purugganan@ncsu.edu

during the vegetative phase and determines the number of axillary meristems potentially available for branches within the inflorescence. Later-flowering ecotypes display greater potential lifetime fecundity as well as greater environmental plasticity in a suite of postreproductive traits (Dorn *et al.* 2000), perhaps in part because a greater number of axillary meristems provide greater developmental flexibility.

There is considerable intraspecific diversity in the timing of flowering in *A. thaliana*, and the genetic architecture of this trait has been under intense investigation. Over 60 genes that control flowering time have been identified (Mouradov *et al.* 2002; Simpson and Dean 2002), and several QTL that contribute to natural variation in the number of rosette leaves upon bolting have been mapped (Clarke *et al.* 1995; Alonso-Blanco *et al.* 1998; Mouradov *et al.* 2002; Simpson and Dean 2002; Ungerer *et al.* 2002, 2003). Photoperiod (day length) is one important environmental cue that affects the timing of flowering. This photoperiod response is mediated in part by the blue light photoreceptor gene *CRYPTOCHROME2* (*CRY2*). *CRY2* acts to promote flowering (Guo *et al.* 1998); null mutants of this gene lead to late flowering under long-day conditions (Koornneef *et al.* 1991; Guo *et al.* 1998). Under short-day conditions, downregulation of *CRY2* protein on a diurnal cycle leads to delayed flowering relative to long-day photoperiod (El-Assal *et al.* 2001). Recently, a natural allele of *CRY2* was shown to be responsible for the major effect *EARLY DAYLENGTH INSENSITIVE* (*EDI*) QTL identified in a cross between the Cape Verde Island (Cvi-0) and Landsberg *erecta* (L*er*-2) ecotypes (El-Assal *et al.* 2001). This allele confers early flowering under short-day photoperiod; it has been observed only in the Cvi-0 ecotype (El-Assal *et al.* 2001).

In this study, we have used a haplotype-based LD mapping approach to identify two other naturally occurring alleles in the *CRY2* genomic region that, like the $CRY2^{EDI}$ allele, are associated with early flowering under short-day photoperiods. Unlike $CRY2^{EDI}$, however, these alleles have a widespread distribution within the species and may thus play an important role in modulating natural variation in *A. thaliana* flowering time across the species range. Moreover, these alleles appear to underlie QTL of much smaller effect on flowering time than $CRY2^{EDI}$. This is the first example of the use of LD mapping in identifying and fine-mapping QTL in *A. thaliana*, and it illustrates the potential of this approach in dissecting the genetic architecture and molecular basis of adaptive variation in this selfing plant species.

MATERIALS AND METHODS

Molecular population genetic analyses: *A. thaliana* ecotypes, representing the geographical distribution of the species in Eurasia, were obtained from single-seed propagated material provided by the *Arabidopsis* Biological Resource Center (see supplemental data, Table S1 at http://www.genetics.org/supplemental/). Genomic DNA was isolated from young leaves of a single individual per ecotype using Plant DNeasy mini kits (QIAGEN, Valencia, CA).

Thirty-one ecotypes were sequenced at *CRY2* and ~1-kb portions of six flanking loci. All primers were designed from the Col-0 genomic sequence (BAC F19P19; GenBank accession no. AC000104), using Primer3 (Rozen and Skaletsky 2000; see supplemental data, Table S2 at http://www.genetics.org/supplemental/). For *CRY2*, PCR primers were designed to amplify two partially overlapping portions of the gene, together spanning ~1 kb of 5′ promoter sequence plus the entire transcriptional unit. Sequenced flanking regions included the four loci adjacent to *CRY2* (AT1G04380, AT1G04390, AT1G04410, and AT1G04420), plus two loci ~23.5 kb upstream and ~25 kb downstream of this *CRY2* genomic region (AT1G04480 and AT1G04300, respectively). PCR was performed using *Taq* DNA polymerase (Roche, Indianapolis), with amplification conditions following the polymerase manufacturer's protocols and annealing temperatures adjusted for each primer pair.

PCR products were purified using QIAquick gel extraction kits (QIAGEN) and sequenced directly using cycle sequencing with BigDye terminators (Applied Biosystems, Foster City, CA). DNA sequencing was performed with a Prism 3700 96-capillary automated sequencer (Applied Biosystems). Sequence management was carried out using BioLign version 2.09.1 (Tom Hall, North Carolina State University). *A. thaliana* is a predominantly selfing species, and no heterozygosity was observed in the genes sequenced. In several instances, rare polymorphisms were confirmed with reamplification and resequencing. GenBank accession numbers for sequenced regions are AY576055–AY576271.

DNA sequences were visually aligned, and most molecular population genetic analyses were conducted using DnaSP 3.51 (Rozas and Rozas 1999). Levels of nucleotide diversity per silent site were estimated as π (Tajima 1983) and θ_W (Watterson 1975), and the Tajima (1989) and Fu and Li (1993) tests for neutral evolution were employed. Linkage disequilibrium between parsimony-informative sites within and between genes was estimated as r^2 (Hill and Robertson 1968), with statistical significance determined by two-tailed Fisher's exact tests (Sokal and Rohlf 1981). Haplotype trees were constructed using a maximum parsimony analysis (branch and bound search, stepwise addition) in PAUP* (Swofford 2000). Insertion/deletion polymorphisms (indels) were included in the parsimony analyses, with each indel block treated as a single character.

Controlling for population structure: When attempting to identify the genetic basis of phenotypic variation, it is important to control for cryptic genetic structure (stratification) within the sample population, which can result in spurious associations between genetic and phenotypic variation (reviewed by Cardon and Palmer 2003). We performed a Bayesian analysis on a multilocus, genome-wide genotype data set [79 amplified fragment length polymorphism (AFLP) markers; Sharbel *et al.* 2000] for 104 *A. thaliana* ecotypes to identify the largest subset that could be considered a single, unstratified population. The *structure* 1.0 program (Pritchard *et al.* 2000; see also Thornsberry *et al.* 2001) was used to identify the number of genetically distinct subpopulations that maximize the likelihood of AFLP allele distributions among ecotypes and to assign ecotypes to subpopulations. Using 50,000 iterations following a burn-in of 50,000 iterations, an optimal subpopulation number was determined to be $K = 4$. Specifying K values of 5 or more led to minimal variation in likelihood values, indicating the absence of additional substructuring within these four subpopulations (see discussion in *structure*

TABLE 1

Features of sequence variation at the *Arabidopsis thaliana CRY2* genomic region

Gene	Gene ID[a]	Position[b]	Length (bp)[c]	π[d]	Tajima's *D*	Fu & Li's *D**
aldo-keto reductase-like	AT1G04420	40.51–41.51	1053	0.0243	0.3286	1.5687*
putative malate dehydrogenase	AT1G04410	41.93–42.84	906	0.0117	0.1106	1.2456
CRY2	AT1G04400	44.34–47.54	3218	0.0125	0.3479	1.5890*
hypothetical	AT1G04390	49.68–50.60	926	0.0150	0.1553	1.3794
2-oxoglutarate-dept dioxygenase	AT1G04380	54.95–55.87	920	0.0063	0.1410	1.0760

[a] Based on TAIR genome annotation.
[b] Position along BAC clone F19P19 sequence, in megabases.
[c] Length of aligned, sequenced region.
[d] Estimates based only on silent sites; $*P < 0.02$.

1.0 documentation). Ninety-five of the 104 ecotypes were assigned to a single subpopulation, with the remaining 9 ecotypes distributed among the three other subpopulations. The large, unstratified subpopulation comprising 95 ecotypes was used in all tests of association between *CRY2* variation and variation in flowering time (see supplemental data, Table S1 at http://www.genetics.org/supplemental/); this ecotype set does not include the Cvi-0 line, which carries the early flowering $CRY2^{EDI}$ allele. Because *A. thaliana* is a selfing species, the *structure* analysis may overestimate the number of subpopulations in the sample of ecotypes that were examined (Falush *et al.* 2003). Thus, our exclusion from association tests of all ecotypes that do not fall within the single largest subpopulation is a conservative approach for dealing with cryptic population structure.

Phenotypic measurements: From the set of 95 ecotypes that showed no evidence of population stratification, 10 plants per line were grown in randomized flats in the Phytotron facility of North Carolina State University. Ecotypes were grown under two photoperiod treatments: long-day conditions (14 hr of light at 20°; 10 hr of dark at 18°) and short-day conditions (10 hr of light at 20°; 14 hr of dark at 18°). Rosette leaves were counted on the first day that the inflorescence stalk was clearly identifiable. Rosette leaf number at bolting (RLN) was calculated as the mean of replicates for each ecotype.

In October 2001, 3–5 seeds from each ecotype in this subset were deposited into each of 12 randomized and blocked peat pots that had been sunk into the soil in raised beds outside of the Brown University (Providence, RI) greenhouses. Seeds were allowed to germinate naturally under the protection of metal window-screening. All plants except the seedling closest to the center of the pot were thinned; thinning began after ~14 days. RLN was calculated as for growth chamber conditions. For both of these experiments, broad-sense heritabilities and additive effects were estimated according to methods described in Falconer and Mackay (1996).

Haplotype tagging and association tests: On the basis of DNA sequences from *CRY2* and the four adjacent flanking genes, candidate single nucleotide polymorphisms (SNPs) were identified for distinguishing major haplotype groups in *CRY2* and in the extended *CRY2* genomic region. SNP genotyping was conducted by the dCAPS method (Neff *et al.* 2002), and insertion/deletion (indel) genotyping by size-fractionation of PCR-amplified fragments (see supplemental data, Table S3 at http://www.genetics.org/supplemental/). One-way analyses of variance (ANOVAs), performed in StatView, version 5.0.1 (Calderola *et al.* 1998), were used for association tests within the unstratified ecotype sample, under long-day (growth chamber) and short-day (both growth chamber and field) photoperiod conditions. Association tests were performed first using haplotypes based on *CRY2* sequences alone and then using haplotypes based on the *CRY2* genomic region. Sample sizes for association tests depended on the number of ecotypes for which both phenotypic and genotypic data were successfully obtained and thus varied slightly among association tests.

RESULTS

Nucleotide polymorphisms and haplotype structure in the *CRY2* gene: We examined nucleotide variation in the *CRY2* gene in a sample of 31 *A. thaliana* ecotypes. The sequenced region is ~3.2 kb and includes the entire transcriptional unit as well as 961 bp of sequence upstream of the translation start and 66 bp downstream of the stop codon. Ninety SNPs and 13 indels were observed across the entire gene. The silent-site nucleotide diversity, π, is 0.0125 (Table 1), which is higher than the mean level of 0.007 observed for previously studied *A. thaliana* nuclear genes (Yoshida *et al.* 2003).

SNPs and indels define a total of seven haplotypes in *CRY2*, and maximum parsimony analysis yielded a single *CRY2* haplotype tree (Figure 1A). The seven *CRY2* haplotypes are structured into two highly differentiated haplogroups (*HAP A* and *HAP B*; Figure 1A), which are separated by a long internal branch comprising most of the variation observed in the gene (77 SNPs and 13 indels). The *HAP A* and *HAP B* groups occur at frequencies of 89.2 and 10.8%, respectively (Figure 1A). Tajima's (1989) *D* and Fu and Li's (1993) *D** statistics are both positive for *CRY2* (Table 1); the significantly positive value of the latter statistic is consistent with the internal branch being longer than expected under the neutral equilibrium model. There is a perfect correlation among polymorphisms on the haplotype tree (homoplasy index, HI = 0), indicating the absence of detectable recombination across *CRY2*; this finding is confirmed by an explicit test for recombination (Rm = 0; Hudson and Kaplan 1985). Together these findings indicate both haplotype dimorphism and very strong linkage disequilibrium across *CRY2*. Fewer polymorphisms were observed in the 961-bp upstream portion

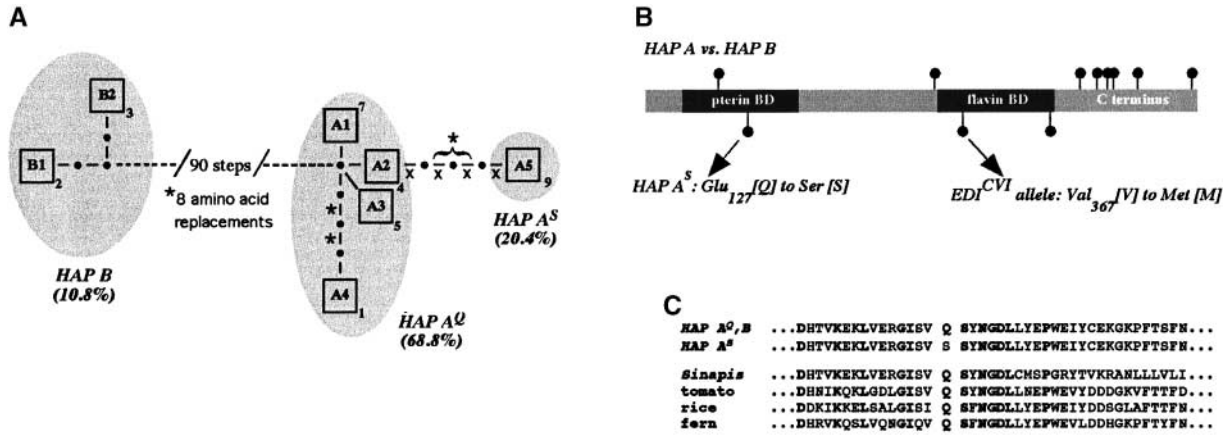

FIGURE 1.—(A) The *CRY2* haplotype tree. Haplotypes are designated as letter-number combinations inside squares and correspond to labels in GenBank accessions AY576055–AY576271. Numbers outside of squares indicate occurrences of each haplotype in 31 sequenced *A. thaliana* ecotypes. Each short line is a mutational step corresponding to a single SNP or indel polymorphism. Asterisks indicate amino acid replacement polymorphisms, and solid circles between lines represent inferred intermediate haplotypes. Shaded ellipses designate haplogroups *HAP B*, *HAP A*^{Q}, and *HAP A*^{S}; haplogroup frequencies are based on haplotype-tag genotyping of 93 of 95 ecotypes constituting a single, unstratified population sample. An x indicates the location of four consecutive nucleotide polymorphisms in exon 2, of which the middle two sites encode a glutamine (Q)-to-serine (S) substitution at codon 127. (B) Schematic of the CRY2 protein, indicating positions of observed amino acid replacement polymorphisms. Fixed differences between the *HAP A* and *HAP B* groups are indicated along the top of the protein; replacement polymorphisms within the *HAP A* group are indicated along the bottom. (C) Alignment of *A. thaliana* CRY2 amino acid sequences flanking codon 127 with sequences from taxa spanning the vascular plants. Amino acids conserved across all taxa are indicated in boldface type.

of *CRY2* than in the transcriptional unit (π = 0.0054 and 0.0248 for upstream and transcribed portions, respectively). As in the transcribed portion, most polymorphic sites in the upstream region differentiate the *HAP A* and *HAP B* haplogroups (14 of 20 SNPs plus all 4 indels).

Nonsynonymous substitutions result in a total of 11 amino acid replacement polymorphisms in *CRY2*. Eight of these occur on the long internal branch of the haplotype tree, indicating that the proteins encoded by the *HAP A* and *HAP B* groups are fixed for 8 amino acid differences (Figure 1, A and B). The 3 remaining replacement polymorphisms occur within the *HAP A* group. Two of the 3 are unique to the Cvi-0 ecotype (haplotype *A4* in Figure 1A) and occur within the flavin-binding domain of the protein. One of these 2—a Val-to-Met replacement at position 367 (see Figure 1, A and B)—has previously been shown to result in the *CRY2*^{EDI} allele causing early flowering under short-day conditions (EL-ASSAL *et al.* 2001). A screen of >100 Arabidopsis accessions indicates that the *CRY2*^{EDI} allele is found only in Cvi-0 (EL-ASSAL *et al.* 2001).

The third amino acid replacement polymorphism within *HAP A* is a radical Glu (Q)-to-Ser (S) replacement at codon 127, located in the pterin-binding domain of the CRY2 protein (Figure 1B; see also GUO *et al.* 1998; EL-ASSAL *et al.* 2001). This substitution is associated with four consecutive nucleotide substitutions in exon 2, the middle two of which encode the amino acid change (Figure 1A). The Q at this position is conserved in *CRY2* proteins across 395 million years of plant evolution, from angiosperms to ferns (Figure 1C), yet is polymorphic for a radical substitution within *A. thaliana*. This pattern suggests that the observed amino acid replacement might well be expected to affect CRY2 protein function. On the basis of this radical amino acid polymorphism, we define two major subgroups within *HAP A:* those with Q (*HAP A*^{Q}) and those with S (*HAP A*^{S}) at codon 127 (Figure 1A). The *HAP A*^{Q} and *HAP A*^{S} haplogroups occur at frequencies of 68.8 and 20.4%, respectively, in the unstratified ecotype sample and differ solely by the four consecutive nucleotide substitutions associated with this amino acid replacement (Figure 1A).

EL-ASSAL *et al.* (2001) report the occurrence of an amino acid substitution at position 188 in the L*er* ecotype, which was not observed with the sequencing of L*er* in this study. This difference may reflect genetic variation between the specific lines examined (L*er*-2 and L*er*-0 in the previous and this study, respectively).

***CRY2* haplotype tagging and associations with flowering time:** The degree of genetic differentiation that we observed in *CRY2* led us to examine whether these polymorphisms might be associated with variation in flowering time under long- and/or short-day conditions. When testing for associations between genetic and phenotypic variation, it is critical to control for cryptic population structure (stratification), which can lead to spurious positive associations (reviewed by CARDON and PALMER 2003). An unstratified set of 95 ecotypes was therefore used in all tests of phenotypic association (see MATERIALS AND METHODS). Using RLN as an indicator

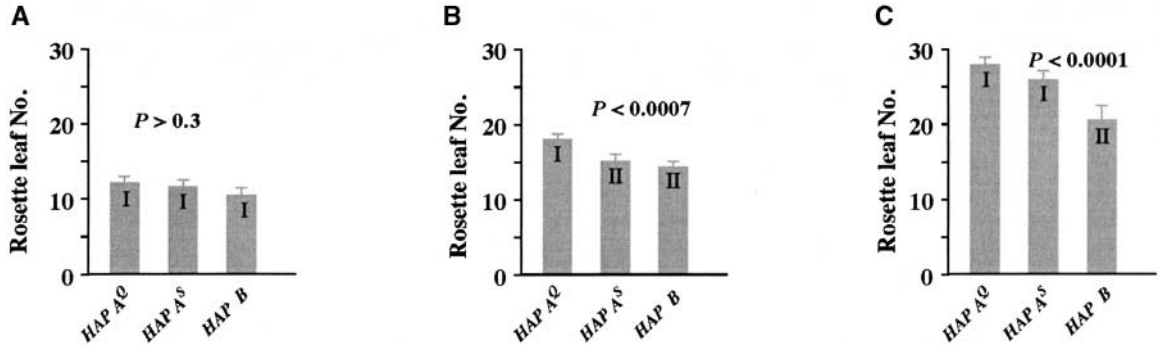

FIGURE 2.—Associations between *CRY2* haplogroups and rosette leaf number at bolting. (A) Long-day photoperiod, growth chamber. (B) Short-day photoperiod, growth chamber. (C) Short-day (overwintering) photoperiod, field. *P* values indicate significance in one-way analyses of variance. Bars that share roman numerals are not significantly different from each other at $P = 0.05$ by Fisher's PLSD.

of flowering time, we first confirmed that there is significant variation in flowering time among ecotypes (ANOVA, $P < 0.0001$ for both growth chamber conditions and field conditions). The broad-sense heritability (H^2) of this trait was found to be 0.639 and 0.481 for long- and short-day conditions, respectively, in the growth chamber and 0.526 for plants overwintered in outdoor beds (field, short-day photoperiod) in Rhode Island.

We then employed a haplotype tagging strategy (JOHNSON *et al.* 2001) for LD mapping of the flowering time alleles at this locus. Three polymorphisms that differentiate the major haplotype groups (*HAP* A^Q, *HAP* A^S, and *HAP B*; see supplemental data, Table S3 at http://www.genetics.org/supplemental/) were typed in accessions from the unstratified ecotype sample, and associations between *CRY2* haplogroups and RLN were tested using one-way analyses of variance. For the long-day growth chamber treatment, we found no association between major *CRY2* haplogroups and RLN (ANOVA, $P > 0.3$, $N = 88$; Figure 2A). In contrast, under short-day conditions—in both the growth chamber and the field experiments—there is a significant association (ANOVA, $P < 0.0007$, $N = 88$ and $P < 0.0001$, $N = 78$ for growth chamber and field, respectively; Figure 2, B and C). In the growth chamber, those accessions possessing *HAP* A^S and *HAP B* alleles bolt significantly earlier than those with *HAP* A^Q alleles (16.34 ± 0.84 and 15.10 ± 1.16 rosette leaves at bolting, for *HAP* A^S and *HAP B*, respectively; 19.26 ± 0.49 rosette leaves for *HAP* A^Q). Similarly, for field conditions, *HAP* A^S and *HAP B* alleles are associated with earlier-flowering ecotypes (26.40 ± 0.97 and 20.83 ± 1.90 for *HAP* A^S and *HAP B*, respectively; 28.13 ± 0.58 for *HAP* A^Q), although in this case pairwise Fisher's protected least significant difference (PLSD) tests (CALDEROLA *et al.* 1998) indicate that *HAP* A^S is not significantly different from *HAP* A^Q but rather shows significantly later flowering than *HAP B* (Figure 2C). Taken together, these data suggest that under short-day, but not long-day photoperiod conditions, the *HAP* A^S and *HAP B* haplogroups are associated with significantly earlier flowering than is the more common *HAP* A^Q haplogroup.

Defining the physical boundaries of *CRY2* flowering time haplotypes: Association of rosette leaf number with *CRY2* haplogroups may arise from polymorphisms within this gene or from variants at linked loci that are in disequilibrium with *CRY2* polymorphisms. To assess the physical boundaries of the *CRY2* haplogroups, we sequenced ~1-kb segments in the four genes most closely flanking *CRY2* (Table 1), using the same set of 31 accessions sequenced at *CRY2*. These flanking gene sequences, together with the polymorphism data from *CRY2*, provide a haplotype map of an ~16-kb genomic region around *CRY2*.

The haplogroup dimorphism and the very strong linkage disequilibrium observed in *CRY2* extends across this entire 16-kb region. Elevated levels of nucleotide diversity (π) are observed throughout the genomic region (Table 1), and both TAJIMA's (1989) *D* and FU and LI's (1993) *D** statistics are positive for all flanking loci (although not with uniform statistical significance; see Table 1). These patterns suggest an excess of intermediate-frequency polymorphisms that would be consistent with haplotype dimorphism throughout the region.

The *CRY2*-region haplotype tree confirms this pattern of haplogroup dimorphism and extensive haplotype structure (Figure 3). This tree has the same basic topology as the *CRY2* tree (Figure 1A), differing primarily in the degree of differentiation among the previously identified haplogroups. As with *CRY2* alone, most polymorphisms in each of the flanking regions occur along the internal branch separating the *HAP A* and *HAP B* groups (72.7–90.4% of substitution polymorphisms among the four loci). Moreover, like the *CRY2* tree, the extended haplotype tree is nearly free of homoplasy (HI = 0.004, reflecting a single homoplasious polymorphism; Figure 3). Thus, there is no evidence for recombination between sequenced haplotypes across this 16-kb region. The strong correlation of polymorphisms is

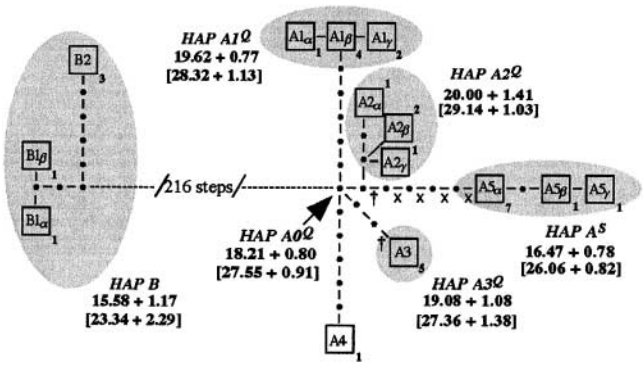

FIGURE 3.—Haplotype tree for the *CRY2* genomic region, based on DNA sequences of 31 *A. thaliana* ecotypes for the entire *CRY2* gene plus ~1-kb portions of the four genes flanking *CRY2*. Letter-number combinations inside squares indicate haplotype designations; numbers outside of squares indicate occurrence of each haplotype among 31 sequenced ecotypes. Short lines and solid circles indicate mutational steps and inferred intermediate haplotypes, respectively, as in Figure 1A. Daggers indicate the placement of a single homoplasious polymorphism. An x indicates the location of four consecutive nucleotide polymorphisms associated with the *CRY2* codon 127 Q/S polymorphism. Shaded ellipses designate haplogroups. Numbers under haplogroup names indicate rosette leaf number at bolting ±1 SE for short-day growth chamber and field (in brackets) conditions. This tree represents one of two equally parsimonious arrangements (for alternative topology see supplemental data, Figure S2 at http://www.genetics.org/supplemental/).

also evident in an LD diagram spanning this region (see supplemental data, Figure S1 at http://www.genetics.org/supplemental/).

To delimit the physical boundaries of the *A* and *B* haplogroups beyond the 16-kb *CRY2* genomic region, we sequenced ~1-kb portions of two additional genes, located ~23.5 kb upstream of this region (AT1G04480, *rpl23A*) and ~25 kb downstream (AT1G04300, encoding a MATH-domain protein), using the same 31 ecotypes as for other sequenced loci. This larger genomic region spans a total of ~65 kb around *CRY2*. A breakdown in the *CRY2*-region haplotype structure is apparent in this larger genomic region. Sequences in the upstream gene, *rpl23A*, show a pattern of haplotype structure that is incongruent with that of the *CRY2* region. Haplotype clades in this gene do not correspond to the *HAP A*, *HAP B*, and *HAP A*^{S} haplogroups found in the *CRY2* region, and a test for recombination (HUDSON and KAPLAN 1985) reveals a minimum of two recombinations between *rpl23A* and the *CRY2* region.

The downstream gene, AT1G04300, was found to contain almost no variation at all (three silent polymorphisms; $\pi = 0.0004$). This level of nucleotide diversity is far below that of neutrally evolving *A. thaliana* genes (0.007; YOSHIDA *et al.* 2003), which suggests that this portion of the genome has been affected by evolutionary forces (*e.g.*, a selective sweep) that are not observed in the *CRY2* genomic region. In addition, DNA sequences from a gene located ~7 kb farther downstream (AT1G04280) show a pattern of haplotype structure that is incongruent with that observed in the *CRY2* genomic region, with a minimum of two recombinational events inferred between the *CRY2* genomic region and this gene (HUDSON and KAPLAN 1985; $n = 13$ accessions; R. MOORE, North Carolina State University, personal communication). Taken together with data from the upstream gene *rpl23A*, these findings indicate that the dimorphism characterizing the *CRY2 HAP A* and *HAP B* haplogroups does not extend to these more distant loci and that recombination has occurred between the 16-kb *CRY2* genomic region and these genes. Thus, the haplotype structure observed at *CRY2* is limited to a 65-kb region around this gene.

Whereas the polymorphisms defining the *HAP A* and *HAP B* groups extend at least 16 kb around *CRY2*, those differentiating the *HAP A*^{Q} and *HAP A*^{S} groups are characterized by only five nucleotide sites (Figure 3). Four of these are the consecutive nucleotides in *CRY2* exon 2 associated with the Q-to-S substitution at amino acid position 127. The fifth site is a homoplasious substitution that is found in both the *HAP A*^{S} and the *HAP A*^{Q} groups. Thus, across the 16-kb region, there are no polymorphisms other than those associated with the *CRY2* Q-to-S amino acid replacement that are unique to *HAP A*^{S}. The fifth, homoplasious site can be assigned with equal parsimony to one of two nucleotide positions: a silent-site mutation within the first intron of a gene encoding an aldo-keto reductase-like protein, ~4 kb upstream of *CRY2* (Table 1; Figure 3), or a synonymous third codon position substitution at *CRY2* amino acid position 20 (see supplemental data, Figure S2 at http://www.genetics.org/supplemental/). In either case, the homoplasious mutation is not unique to *HAP A*^{S}.

Extended *CRY2* haplotypes and flowering time associations: The well-defined haplotype structure across the 16-kb *CRY2* genomic region allows us to perform association tests for this genomic region using the same haplotype tagging strategy used for *CRY2* alone. For the unstratified set of 95 ecotypes, we genotyped 12 SNP markers (see supplemental data, Table S3 at http://www.genetics.org/supplemental/) that define the five major haplotype groups in this genomic region (*HAP A1*^{Q} to *A3*^{Q}, *HAP A*^{S}, and *HAP B*; Figure 3); these haplogroups correspond to major haplotypes in the *CRY2* gene (Figure 1A). Genotyping with these markers yielded a sixth major haplogroup, which falls within the *HAP A* group but which lacks SNPs characterizing any of the named *HAP A*^{Q} haplotypes. This potentially heterogeneous group is designated *HAP A0*^{Q} (see Figure 3).

The extended haplotype association analysis confirms the tests based solely on *CRY2* haplogroups. For short-day growth chamber conditions, there is a significant association of flowering time with the *CRY2* genomic region haplogroups ($P < 0.0127$, $N = 90$; Figure 4). Earlier flowering is observed in accessions containing

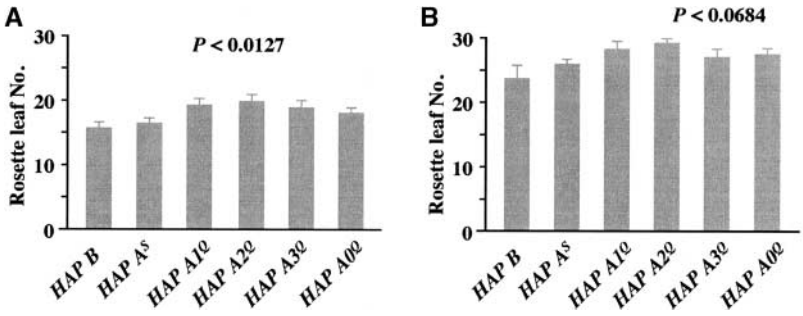

FIGURE 4.—Associations between *CRY2* genomic region haplogroups and flowering time, as measured by rosette leaf number at bolting. *P* values on charts indicate significance in one-way analyses of variance. (A) Short-day photoperiod, growth chamber. Statistically significant differences in Fisher's PLSD tests are as follows: *HAP* A^S and *HAP* $A1^Q$, $P = 0.0070$; *HAP* A^S and *HAP* $A2^Q$, $P = 0.0111$; *HAP B* and *HAP* $A1^Q$, $P = 0.0072$; *HAP B* and *HAP* $A2^Q$, $P = 0.0086$; and *HAP B* and *HAP* $A3^Q$, $P = 0.0362$. (B) Short-day photoperiod, field. Statistically significant differences in Fisher's PLSD tests are as follows: *HAP B* and *HAP* $A0^Q$, $P = 0.0415$; *HAP B* and *HAP* $A1^Q$, $P = 0.0087$; and *HAP B* and *HAP* $A2^Q$, $P = 0.0056$.

HAP A^S and *HAP B* (Figure 3), and pairwise Fisher's PLSD tests (CALDEROLA *et al.* 1998) indicate that accessions containing *HAP* A^S and *HAP B* flower significantly earlier than do the *HAP* $A1^Q$, *HAP* $A2^Q$, and *HAP* $A3^Q$ haplogroups (Figure 4). Under field conditions, haplogroup association with flowering time is not significant at $P = 0.05$ ($P < 0.0684$, $N = 81$; Figure 4; see also Figure 3); however, Fisher's PLSD tests again support the associations that show significantly earlier flowering for accessions containing *HAP B*. For the 23.5-kb upstream gene, *rpl23A*, no significant association was found between an SNP distinguishing major haplogroups at this locus and flowering time under short-day photoperiod, for either growth chamber (ANOVA, $P > 0.85$, $N = 91$) or field conditions (ANOVA, $P > 0.75$, $N = 81$). Ecotypes possessing all three of the major *CRY2* haplotypes (*HAP* A^Q, *HAP B*, and *HAP* A^S) are represented in both of these *rpl23A* haplogroups, confirming that the *CRY2* haplotype structure does not extend as far as this flanking gene.

DISCUSSION

Haplotype structure and the limits of flowering time QTL: The success of LD mapping in localizing candidate polymorphisms depends in large part on genomic patterns of linkage disequilibrium in a species and the resulting haplotype structure. In this study, we have observed two strikingly different patterns of haplotype structure for the *CRY2* haplogroups associated with early flowering under short-day conditions, *HAP* A^S and *HAP B*. In the case of *HAP* A^S, the SNPs characterizing this haplogroup comprise only four consecutive nucleotides, which include the polymorphisms responsible for the radical Q-to-S amino acid replacement at position 127 (Figure 3B). In contrast, the nucleotide polymorphisms distinguishing the *HAP B* haplogroup from *HAP A* extend across a 16-kb region around *CRY2* and include >200 SNPs in the sequenced fragments alone, of which 32 encode amino acid replacements. It is thus equally likely that any polymorphism separating the *HAP A* and *HAP B* haplogroups—including those not observed here—could underlie the associated phenotypic variation. Nonetheless, DNA sequence polymorphisms at loci ~23.5 kb upstream and ~25 kb downstream of the *CRY2* genomic region indicate that the *CRY2* haplogroup dimorphism does not extend as far as these flanking genes. Thus, although the polymorphisms underlying the *HAP B* flowering time association cannot be precisely identified, linkage disequilibrium mapping suggests that the locus responsible is localized to an ~65-kb genomic interval centered on *CRY2*.

The evolution of flowering time in Arabidopsis: In recent years, geneticists have isolated three genes—*FRI* (JOHANSON *et al.* 2000; LE CORRE *et al.* 2002), *FLC* (GAZZANI *et al.* 2003; MICHAELS *et al.* 2003), and *CRY2* (EL-ASSAL *et al.* 2001)—that harbor polymorphisms underlying quantitative variation in flowering time in wild *A. thaliana* populations. The *EDI* locus was first identified as a QTL of major effect that conferred early flowering under short-day conditions in the Cvi-0 ecotype of *A. thaliana*; positional cloning of this QTL determined that *EDI* is an allele of *CRY2* (EL-ASSAL *et al.* 2001). $CRY2^{EDI}$ has a single amino acid Val-to-Met replacement in the flavin-binding domain of the encoded receptor protein that weakens the light-induced down-regulation of CRY2 protein levels under short-day photoperiod conditions (EL-ASSAL *et al.* 2001). Like previously identified QTL alleles at *FRI* and *FLC*, $CRY2^{EDI}$ has a major effect on flowering time variation. Unlike the natural early flowering alleles of the former two genes, however, $CRY2^{EDI}$ is observed at very low frequency (in the Cvi-0 ecotype only; EL-ASSAL *et al.* 2001), which suggests that its adaptive significance may, at best, be highly geographically localized.

Our molecular population genetic and LD mapping analyses indicate the presence of at least two additional haplogroups in the *CRY2* region that, like $CRY2^{EDI}$, appear to confer early flowering under short-day conditions. Alleles of small to moderate effect on flowering time have not been previously isolated in *A. thaliana*, and the *HAP* A^S and *HAP B* alleles at *CRY2* appear to be the first alleles of this class described at the molecular level in this species. The $CRY2^{EDI}$ allele (haplotype *A4*

in Figures 1A and 3) has an additive effect of approximately −9 rosette leaves on bolting under short-day conditions (El-Assal *et al.* 2001). In comparison, *CRY2 HAP A*^{S} and *HAP B* have additive effects of −1.46 and −2.08 rosette leaves upon bolting in short days, respectively, when compared to *HAP A*^{Q}. Under field conditions, *HAP B* has an additive effect of −3.65 rosette leaves upon bolting compared to *HAP A*^{Q}. Unlike *FRI* and *FLC*, *CRY2* is a member of the photoperiod-dependent flowering time pathway, suggesting that it may be involved in determining seasonal day length cues (Mouradov *et al.* 2002; Simpson and Dean 2002). The small additive effect on flowering time associated with these *CRY2* haplogroups suggests that they may serve to modulate the floral transition rather than act as primary determinants of life history characteristics, as proposed for *FRI* or *FLC* (Johanson *et al.* 2000; Le Corre *et al.* 2002; Gazzani *et al.* 2003; Michaels *et al.* 2003). Further field analyses may determine the ecological significance of these modulating effects on early flowering.

Two lines of evidence suggest that genetic variation within the *CRY2* region may be reflecting adaptive evolution. First, the *HAP A*^{S} and *HAP B* haplogroups form two distinct clusters on the *CRY2*-region haplotype tree (Figure 3), which indicates that these early flowering haplotypes have two independent evolutionary origins; together with the *CRY*^{EDI} allele, there are thus three documented origins of early flowering in the *CRY2* genomic region. Independent evolution of early flowering alleles has also been observed in the *FRI* (Johanson *et al.* 2000; Le Corre *et al.* 2002) and *FLC* (Gazzani *et al.* 2003; Michaels *et al.* 2003) genes, and the multiple origins of this phenotype may indicate strong selection and possibly local adaptation for early flowering within this weedy species. Second, like early flowering alleles found at *FRI*, the *HAP A*^{S} and *HAP B* haplogroups occur at moderate frequencies across the species range (Figure 1A; see supplemental data, Table S1 at http://www.genetics.org/supplemental/). Thus, unlike the extremely rare *CRY*^{EDI} allele, these *CRY2* haplogroups may have widespread ecological importance. This hypothesis is supported by an analysis of the geographical distribution of these alleles, which indicates that, within the *HAP A* haplogroup, the prevalence of the earlier-flowering *A*^{S} haplotype is significantly correlated with colder mean January temperatures (T. Korves, personal communication). In addition to adaptive evolution, the genetic variation in the *CRY2* genomic region may also be reflecting the population history of *A. thaliana*. In particular, the degree of genetic divergence between the *HAP A* and *HAP B* haplogroups could reflect in part the action of genetic drift on two ancient population lineages that were isolated during periods of Pleistocene glaciation in Europe.

The identification of these *CRY2* haplotype groups as the basis for early flowering in *A. thaliana* represents an association and not a causal connection. Formal proof of a causal relationship will require transgenic complementation analysis, and the development of necessary transgenic populations is currently in progress. There are several reasons to believe, however, that the association between early flowering and at least the *HAP A*^{S} alleles may be causal. First, *CRY2* is a known candidate flowering time gene (Guo *et al.* 1998; El-Assal *et al.* 2001). Second, the amino acid replacement associated with *HAP A*^{S} is a radical change that is otherwise conserved across vascular plants (Figure 1C). Third, the polymorphisms that distinguish *HAP A*^{S} are found exclusively in *CRY2* and are not in strong LD with any of the assayed markers across the 65-kb genomic region examined (Figure 3; see also supplemental data, Figure S1 at http://www.genetics.org/supplemental/). In contrast to *HAP A*^{S}, the broad physical expanse of the *HAP B* haplogroup across this genomic region prevents the precise identification of the specific gene underlying the association with early flowering. More extensive analysis of this region will be necessary to delimit the precise boundaries of this haplotype block.

These analyses illustrate the utility of LD mapping approaches in identifying and localizing QTL within the selfing species *A. thaliana*. The application of LD mapping techniques to *A. thaliana* has been widely discussed (*e.g.*, Borevitz and Nordborg 2003), and there have been extensive analyses of the extent of LD in this species as a prelude to mapping efforts (Nordborg *et al.* 2002). The levels of LD that we observed in the *CRY2* genomic region suggest that a haplotype-based approach to mapping may be more appropriate in *A. thaliana* than the individual SNP association approach that appears to be successful in Drosophila (Long *et al.* 1998) and in maize (Thornsberry *et al.* 2001). The use of haplotypes rather than individual SNP markers in *A. thaliana* association studies exploits the more extensive haplotype structure in this species for localizing QTL and ensures that significant associations can be detected even when multiple alleles of the same additive effect are present. It thus appears that *A. thaliana* LD mapping shares more in common with human mapping approaches that exploit extended haplotypes (Johnson *et al.* 2001) and that this selfing species may provide a genetic model organism for testing LD mapping methodologies for application to human genetic studies.

We thank T. Korves, T. F. C. Mackay, and members of the Purugganan laboratory for providing unpublished data and for helpful discussions. This work was funded in part by a National Science Foundation Integrated Research Challenge in Environmental Biology grant to M.D.P, J.S., and T. F. C. Mackay.

LITERATURE CITED

Alonso-Blanco, C., S. E.-D. El-Assal, G. Coupland and M. Koornneef, 1998 Analysis of natural allelic variation at flowering time loci in the Landsberg *erecta* and Cape Verde islands ecotypes of *Arabidopsis thaliana*. Genetics **149:** 749–764.

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of

linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

Borevitz, J. O., and M. Nordborg, 2003 The impact of genomics on the study of natural variation in *Arabidopsis.* Plant Physiol. **132:** 718–725.

Calderola, J., A. Dilmaghani, J. Gagnon, K. Haycock, J. Roth *et al.*, 1998 *Statview*, Version 5.0. SAS Institute, Cary, NC.

Cardon, L. R., and L. J. Palmer, 2003 Population stratification and spurious allelic association. Lancet **361:** 598–604.

Clarke, J. H., R. Mithen, J. K. M. Brown and C. Dean, 1995 QTL analysis of flowering time in *Arabidopsis thaliana.* Mol. Gen. Genet. **248:** 278–286.

Dorn, L. A., E. H. Pyle and J. Schmitt, 2000 Plasticity to light cues and resources in *Arabidopsis thaliana*: testing for adaptive value and costs. Evolution **54:** 1982–1994.

El-Assal, S. E.-D., C. Alonso-Blanco, A. J. M. Peeters, V. Raz and M. Koornneef, 2001 A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2.* Nat. Genet. **29:** 435–440.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman Group, Essex, UK.

Falush, D., M. Stephens and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164:** 1567–1587.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor *et al.*, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67:** 881–900.

Gazzani, S., A. R. Gendall, C. Lister and C. Dean, 2003 Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. Plant Physiol. **132:** 1107–1114.

Guo, H., H. Yang, T. C. Mockler and C. Lin, 1998 Regulation of flowering time by *Arabidopsis* photoreceptors. Science **279:** 1360–1363.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

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.

Johanson, U., J. West, C. Lister, S. Michaels, R. Amasino *et al.*, 2000 Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science **290:** 344–347.

Johnson, G. C. L., L. Esposito, B. J. Barratt, A. N. Smith, J. Heward *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nat. Genet. **29:** 233–237.

Koornneef, M., C. J. Hanhart and J. H. van der Veen, 1991 A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana.* Mol. Gen. Genet. **229:** 57–66.

Le Corre, V., F. Roux and X. Reboud, 2002 DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time. Mol. Biol. Evol. **19:** 1261–1271.

Long, A. D., R. F. Lyman, C. H. Langley and T. F. C. Mackay, 1998 Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster.* Genetics **149:** 999–1017.

Michaels, S. D., Y. H. He, K. C. Scortecci and R. M. Amasino, 2003 Attenuation of *FLOWERING LOCUS C* activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis.* Proc. Natl. Acad. Sci. USA **100:** 10102–10107.

Mouradov, A., F. Cremer and G. Coupland, 2002 Control of flowering time: interacting pathways as a basis for diversity. Plant Cell **14:** S111–S130.

Neff, M. M., E. Turk and M. Kalishman, 2002 Web-based primer design for single nucleotide polymorphism analysis. Trends Genet. **18:** 613–615.

Nordborg, M., and S. Tavaré, 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.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

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

Rozen, S., and H. J. Skaletsky, 2000 Primer3 on the WWW for general users and for biologist programmers, pp. 365–386 in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by S. Krawetz and S. Misener. Humana Press, Totowa, NJ.

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.

Simpson, G. G., and C. Dean, 2002 *Arabidopsis*, the rosetta stone of flowering time? Science **296:** 285–289.

Sokal, R. R., and F. J. Rohlf, 1981 *Biometry*, Ed. 2. W. H. Freeman, New York.

Swofford, D. L., 2000 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4.0b8a. Sinauer, Sunderland, MA.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

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

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:** 286–289.

Ungerer, M. C., S. S. Halldorsdottir, J. L. Modliszewski, T. F. C. Mackay and M. D. Purugganan, 2002 Quantitative trait loci for inflorescence development in *Arabidopsis thaliana.* Genetics **160:** 1133–1151.

Ungerer, M. C., S. S. Halldorsdottir, M. D. Purugganan and T. F. C. Mackay, 2003 Genotype-environment interactions at quantitative trait loci affecting inflorescence development in *Arabidopsis thaliana.* Genetics **165:** 353–365.

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

Yoshida, K., T. Kamiya, A. Kawabe and N. T. Miyashita, 2003 DNA polymorphism at the *ACAULIS5* locus of the wild plant *Arabidopsis thaliana.* Genes Genet. Syst. **78:** 11–21.

Communicating editor: O. Savolainen