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') 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' 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' 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' early flowering phenotype. This study demonstrates the utility of LD mapping for elucidating the genetic basis of natural, ecologically relevant variation in Arabidopsis.
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 photoperiodic 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 (Ler-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 naturally occurring alleles in the CRY2 genomic region that, like the CRY2edi allele, are associated with early flowering under short-day photoperiods. Unlike CRY2edi, 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 CRY2edi. 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 F1P919; 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 AT1G04430, 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 capillary automated sequencer (Applied Biosystems). Sequence management was carried out using BioEdit version 2.09.1 (Tom Hall, North Carolina State University). A. thaliana is a predominantly selfing species, and no heterozygosities were 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 (Roza and Roza 1999). Levels of nucleotide diversity per silent site were estimated as \( \pi \) (Tajima 1983) and \( \theta_w \) (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 \( F \) (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
10 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 flowering time (see supplemental data, Table S1). The sample sizes for association tests depended on the number of ecotypes for which both phenotypic and genotypic data were available (10 hr of light at 20°C; 14 hr of dark at 18°C). Rosette leaves were counted on the first day that the inflorescence stalk was visible, and the four adjacent flanking sequences alone 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°C; 10 hr of dark at 18°C) and short-day conditions (10 hr of light at 20°C; 14 hr of dark at 18°C). Rosette leaves were counted on the first day that the inflorescence stalk was visible. 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 about 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 (ANOVA) were 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 86 bp downstream of the stop codon. Ninety SNPs and 13 indels were observed across the entire gene. The silentsite 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).

Table 1: Features of sequence variation at the Arabidopsis thaliana CRY2 genomic region

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene ID</th>
<th>Position</th>
<th>Length (bp)</th>
<th>π</th>
<th>Tajima’s D</th>
<th>Fu &amp; Li’s D*</th>
</tr>
</thead>
<tbody>
<tr>
<td>aldo-keto reductase-like</td>
<td>AT1G04420</td>
<td>40.51–41.51</td>
<td>1053</td>
<td>0.0243</td>
<td>0.3286</td>
<td>1.5687*</td>
</tr>
<tr>
<td>putative malate dehydrogenase</td>
<td>AT1G04410</td>
<td>41.95–42.84</td>
<td>906</td>
<td>0.0117</td>
<td>0.1106</td>
<td>1.2456</td>
</tr>
<tr>
<td>CRY2</td>
<td>AT1G04400</td>
<td>44.34–47.54</td>
<td>3218</td>
<td>0.0125</td>
<td>0.3479</td>
<td>1.5890*</td>
</tr>
<tr>
<td>hypothetical</td>
<td>AT1G04390</td>
<td>49.68–50.60</td>
<td>926</td>
<td>0.0150</td>
<td>0.1553</td>
<td>1.3794</td>
</tr>
<tr>
<td>2-oxoglutarate-dept dioxygenase</td>
<td>AT1G04380</td>
<td>54.95–55.87</td>
<td>920</td>
<td>0.0063</td>
<td>0.1410</td>
<td>1.0760</td>
</tr>
</tbody>
</table>

* Based on TAIR genome annotation.

| Position along BAC clone F19P19 sequence, in megabases).
| Length of aligned, sequenced region.

* Estimates based only on silent sites; *P < 0.02.
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 Arabidopsis 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 and HAP A; 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 Arabidopsis 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<sup>367</sup> allele causing early flowering under short-day conditions (El-Assal et al. 2001). A screen of >100 Arabidopsis accessions indicates that the CRY2<sup>367</sup> 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 Arabidopsis. 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<sup>Q</sup>) and those with S (HAP A<sup>S</sup>) at codon 127 (Figure 1A). The HAP A<sup>Q</sup> and HAP A<sup>S</sup> 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 Ler ecotype, which was not observed with the sequencing of Ler in this study. This difference may reflect genetic variation between the specific lines examined (Ler-2 and Ler-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
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 \( \text{HAP A}^\circ, \text{HAP A}^\circ, \) and \( \text{HAP B}; \) see supplemental data, Table S3 at http://www.genetics.org/supplemental/) were typed in accessions from the unstratified ecotype sample, and associations between \( \text{CRY2} \) haplogroups and RLN were tested using one-way analyses of variance. For the long-day growth chamber treatment, we found no association between major \( \text{CRY2} \) 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 \( \text{HAP A}^\circ \) and \( \text{HAP B} \) alleles bolt significantly earlier than those with \( \text{HAP A}^\circ \) alleles (16.34 ± 0.84 and 15.10 ± 1.16 rosette leaves at bolting, for \( \text{HAP A}^\circ \) and \( \text{HAP B}, \) respectively; 19.26 ± 0.49 rosette leaves for \( \text{HAP A}^\circ \)). Similarly, for field conditions, \( \text{HAP A}^\circ \) and \( \text{HAP B} \) alleles are associated with earlier-flowering ecotypes (26.40 ± 0.97 and 20.83 ± 1.90 for \( \text{HAP A}^\circ \) and \( \text{HAP B}, \) respectively; 28.13 ± 0.58 for \( \text{HAP A}^\circ \)), although in this case pairwise Fisher’s protected least significant difference (PLSD) tests (Calderola et al. 1998) indicate that \( \text{HAP A}^\circ \) is not significantly different from \( \text{HAP A}^\circ \) but rather shows significantly later flowering than \( \text{HAP B} \) (Figure 2C). Taken together, these data suggest that under short-day, but not long-day photoperiod conditions, the \( \text{HAP A}^\circ \) and \( \text{HAP B} \) haplogroups are associated with significantly earlier flowering than is the more common \( \text{HAP A}^\circ \) haplogroup.

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

The haplogroup dimorphism and the very strong linkage disequilibrium observed in \( \text{CRY2} \) extends across this entire 16-kb region. Elevated levels of nucleotide diversity (\( \pi \)) 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 \( \text{CRY2} \)-region haplotype tree confirms this pattern of haplogroup dimorphism and extensive haplotype structure (Figure 3). This tree has the same basic topology as the \( \text{CRY2} \) tree (Figure 1A), differing primarily in the degree of differentiation among the previously identified haplogroups. As with \( \text{CRY2} \) alone, most polymorphisms in each of the flanking regions occur along the internal branch separating the \( \text{HAP A} \) and \( \text{HAP B} \) groups (72.7–90.4% of substitution polymorphisms among the four loci). Moreover, like the \( \text{CRY2} \) tree, the extended haplotype tree is nearly free of homoplasy (\( H1 = 0.004 \), 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 mutation. An x indicates the location of four consecutive nucleotide polymorphisms associated with the *CRY2* exon 127 Q/S polymorphism. Shaded ellipses designate haplotype groups. 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 A³* 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 A³* and *HAP A⁴* 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 A³* and the *HAP A⁴* 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 A³*. 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 A³*.

**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 A³* to *A⁸*, *HAP A⁴*, 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 A³*-haplotypes. This potentially heterogeneous group is designated *HAP A⁰* (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
haplotypes (N\textsubscript{P}0.85, N\textsubscript{P}0.81). Ecotypes possessing all three of the major derlying 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\textsuperscript{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\textsuperscript{EDI} has a major effect on flowering time variation. Unlike the natural early flowering alleles of the former two genes, however, CRY2\textsuperscript{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\textsuperscript{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\textsuperscript{A} 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\textsuperscript{EDI} allele (haplotype A4

DISCUSSION

Haplotypedefinition 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\textsuperscript{A} and HAP B. In the case of HAP A\textsuperscript{A}, 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 \~\sim2.5 kb upstream and \~\sim25 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 \~\sim65-kb genomic interval centered on CRY2.

The evolution of flowering time in Arabidopsis: In recent years, geneticists have isolated three genes—CRY1 (Johanson et al. 2000; Le Corre et al. 2002), FLC (Gazzani et al. 2003; Michael 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\textsuperscript{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\textsuperscript{EDI} has a major effect on flowering time variation. Unlike the natural early flowering alleles of the former two genes, however, CRY2\textsuperscript{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.
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 A8 and HAP B have additive effects of −1.46 and −2.08 rosette leaves upon bolting in short days, respectively, when compared to HAP A6. Under field conditions, HAP B has an additive effect of −3.65 rosette leaves upon bolting compared to HAP A6. 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. 2005). 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 A8 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 CRY2B1 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 A8 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 CRY2B1 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 haplgroup, the prevalence of the earlier-flowering A8 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 A8 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 A8 is a radical change that is otherwise conserved across vascular plants (Figure 1C). Third, the polymorphisms that distinguish HAP A8 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 A8, 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


Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of
Linkage Disequilibrium Mapping in Arabidopsis


