ABSTRACT

EHRENREICH, IAN MICHAEL. The Genetics of Phenotypic Variation in *Arabidopsis thaliana*. (Under the direction of Michael Purugganan.)

All organisms exhibit substantial quantitative trait variation within populations. Such variation is important because it can affect fitness and serve as the substrate for adaptive evolution. Identifying the quantitative trait genes (QTGs) responsible for phenotypic variation is necessary to understand the mechanisms that generate trait variation and to determine the historical action of natural selection on quantitative traits and QTGs. However, in most complex organisms, the genetic mapping of QTGs is difficult and presently not feasible to do systematically at a gene-level resolution. Model organisms that are both tractable in the laboratory and complex developmentally can serve as trial systems for developing broadly applicable methods for QTG mapping. Using the plant genetic model *Arabidopsis thaliana*, I have attempted to map QTGs for ecologically-significant quantitative traits – shoot branching and flowering time – through a combination of forward and reverse genetic methods. Three main research projects are reported here: i) candidate gene association mapping and linkage mapping of shoot branching; ii) regulatory network-wide candidate gene association mapping of flowering time; and iii) a survey of intra- and interspecific genetic variation at nearly half of the microRNAs (miRNAs) and their binding sites in the genome. These studies have identified strong candidate QTGs for traits that are determinants of *A. thaliana* fitness in the wild. I synthesize my results with those of other researchers in this area to highlight the achievements, future promise, and looming challenges for statistical genetics in terms of elucidating the genetic basis of trait variation.

The Genetics of Phenotypic Variation in *Arabidopsis thaliana*

by
Ian Michael Ehrenreich

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Genetics

Raleigh, North Carolina

2008

APPROVED BY:

Michael D. Purugganan Chair of Advisory Committee	Stephanie E. Curtis Co-Chair of Advisory Committee
Gregory C. Gibson	Robert G. Franks

BIOGRAPHY

Ian Ehrenreich was born and raised in Sacramento, CA. Ian attended Stanford University (Class of 2002) in Stanford, CA prior to becoming a student at NCState. It was at Stanford that Ian became interested in evolutionary genetics. Ian took Bill Durham's course entitled 'Darwin, Evolution, and the Galapagos,' which culminated in a trip to the Galapagos. Retracing the voyage of the HMS Beagle inspired Ian to study evolution and led him to join Brendan Bohannan's lab to do an honors thesis on the experimental evolution of drug-resistant bacteria. Upon graduating, Ian switched to studying biochemical adaptation in butterflies in Ward Watt's lab at Stanford and the Rocky Mountain Biological Laboratory. After leaving the butterflies, but prior to starting at NCState, Ian worked on the ecophysiology of marine cyanobacteria with John Waterbury and Eric Webb at the Woods Hole Oceanographic Institution in Woods Hole, MA. At NCState, Ian did his research on the molecular population genetics of complex trait variation in plants with Michael Purugganan, who actually moved to New York University in New York City. Ian followed Michael to NYU to finish his PhD research. Ian next plans to study the genetic basis of gene expression and complex trait variation in a variety of organisms with Leonid Kruglyak at Princeton University.

ACKNOWLEDGMENTS

Numerous people have contributed to my graduate education. I would especially like to thank my research advisor Michael Purugganan for providing me with the autonomy and support necessary to develop not only as a student, but also as a scientist and more generally as a person. Michael's emphasis on striking a balance between career and personal life, with a prioritization on the personal life, has left an indelible mark on me and has had a strong, positive impact on my happiness.

I would also like to thank my teachers and mentors throughout graduate school, in particular my past and present committee members Philip Awadalla, Stephanie Curtis, Robert Franks, and Greg Gibson. Altogether, these professors were very influential in shaping my thoughts on science and helped me to find my interests in genetics.

I also owe a debt of gratitude to numerous friends both at NCState and NYU, and to my labmates in the Purugganan lab, all of who helped make the process of graduate school more enjoyable. I would especially like to thank Ana Caicedo, my postdoc mentor in the lab during my first couple years, and my two undergraduate apprentices – Lucy Chou and Phillip Stafford.

Lastly, I must thank my family – my wife Hannah, mother Veronica, and brothers Kevan and Ryan – for their love and support throughout graduate school.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER ONE:

The molecular genetic basis of plant adaptation

The molecular genetic basis of plant adaptation

Ian M. Ehrenreich[1,2] and Michael D. Purugganan[2]

[1]Department of Genetics, Box 7614, North Carolina State University, Raleigh, North Carolina 27695 USA
[2]Department of Biology and Center for Genomics and Systems Biology, New York University, 100 Washington Square East, New York, New York 10003 USA

Corresponding author: Michael D. Purugganan, Telephone: (212) 992-9628, Email: mp132@nyu.edu.

Contributions: IME and MDP cowrote this manuscript.

This chapter is a modified version of previously published work.

Reference: Ehrenreich, I.M. and M.D. Purugganan. 2006. The molecular genetic basis of plant adaptation. *American J. Botany* 93: 953-962.

Abstract

How natural selection on adaptive traits is filtered to the genetic level remains largely unknown. Theory and quantitative trait locus (QTL) mapping have provided insights into the number and effects of genes underlying adaptations, but these results have been hampered by questions of applicability to real biological systems and poor resolution, respectively. Advances in molecular technologies have expedited the cloning of adaptive genes through both forward and reverse genetic approaches. Forward approaches start with adaptive traits and attempt to characterize their underlying genetic architectures through linkage disequilibrium mapping, QTL mapping, and other methods. Reverse screens search large sequence data sets for genes that possess the signature of selection. Though both approaches have been successful in identifying adaptive genes in plants, very few, if any, of these adaptations' molecular bases have been fully resolved. The continued isolation of plant adaptive genes will lead to a more comprehensive understanding of natural selection's effect on genes and genomes.

The nature of adaptations

Since Charles Darwin first postulated adaptation's central evolutionary role, much has been learned about how adaptations occur (1859). A major gap in our understanding remains, however, in connecting adaptive traits to their underlying molecular bases. Bridging this gap by isolating adaptive genes is important not only for discerning the evolutionary histories of individual traits, but also for clarifying how selection on phenotypes influences genetic and genomic changes.

The precise nature of the genetic architecture of adaptation — the number and effects of the genetic changes underlying adaptive traits — has proven both theoretically and empirically challenging to estimate. Ronald Fisher was the first to address this topic theoretically through his "geometric model" of phenotypic change and adaptation (1930).

He concluded that the probability a mutation will be adaptive is nearly 50% for mutations of infinitesimally small effects and approximately zero for mutations of very large effects. Fisher and others used these results to argue that adaptation occurs through the accumulation of many beneficial mutations of small effect (ORR 2005a). Motoo Kimura (1983) later challenged Fisher's findings, noting that the substitution rate of advantageous mutations under positive selection is not just dependent on the probability that a mutation is advantageous, but also on the probability of the mutation's fixation. He proffered that, relative to major effect mutations, minor effect mutations are more likely to be beneficial, but less likely to fix in a population. Kimura's results supported intermediate effect mutations as the most likely mutational class to underlie adaptations.

More recently, researchers have attempted to model the genetic basis of adaptation by focusing on DNA or protein sequence evolution (see (ORR 2005a; ORR 2005b) for a description of this research area). These studies have provided evidence for a small number of sequence changes occurring during the adaptive evolution of a gene (GILLESPIE 1991), as well as support for large relative fitness increases after the substitution of a beneficial mutation (ORR 1998). Together, these results imply that major effect mutations may be important to adaptive evolution.

Empirical results, primarily from quantitative trait locus (QTL) mapping experiments, have been largely consistent with theoretical predictions (this has been discussed elsewhere, e.g. (REMINGTON and PURUGGANAN 2003)). QTL mapping experiments have shown that the number and effect of loci controlling adaptive plant traits are variable with anywhere from one to many QTLs detectable. However, these QTL results cannot be taken entirely at face value because the degree to which QTLs represent single or multiple loci in plants is unresolved.

Though the genetic architecture of adaptation has received much discussion, the genetic dissection of more adaptive traits is necessary to expose any generalities about

the molecular basis of adaptive evolution. Whether adaptation typically proceeds through changes in regulatory or structural genes, whether certain types of mutations are most commonly utilized by natural selection, and whether the number and effect of loci that underlie adaptations are typically large or small are questions that cannot be satisfactorily answered with existing data. Pinpointing the genes that underlie adaptations will elucidate how present adaptations have evolved and will facilitate a broader understanding of how adaptations arise at the molecular level.

Identifying plant adaptive genes

A variety of methods exist for mapping genes involved in plant adaptations. Typically the methods used to map these genes attempt to detect natural selection at the molecular level or to find statistical associations or linkages between polymorphisms and adaptive traits. These techniques span many levels of genomic scale and can be used to connect adaptive traits to specific genes and polymorphisms. We discuss these methods and their merits for use in plants.

Detecting selection from molecular data. Adaptations are shaped by natural selection, which can leave a distinctive imprint on the levels and patterns of nucleotide variation in an organism's genome. Numerous statistical tests exist that use molecular variation data to identify genes that bear the signature of selection exist (NIELSEN 2001). The null hypothesis for these tests is often that the observed genetic variation is consistent with selective neutrality at the locus of interest (KIMURA 1983) and that significant departures from this neutral expectation may be indicative of the action of selection.

One class of tests for selection examines the frequencies of single nucleotide polymorphisms (SNPs) in a sample of sequenced alleles. Sequences that have evolved according to expectations of the standard neutral model are expected to display a different distribution of SNP frequencies than those that have experienced selection

(TAJIMA 1989). For example, positively selected alleles that have recently swept to fixation should possess an excess of low-frequency SNPs (KAPLAN *et al.* 1989; MAYNARD-SMITH and HAIGH 1974). In contrast, some modes of balancing selection create an excess of intermediate frequency SNPs (HUDSON and KAPLAN 1988). A number of related tests exist, such as those proposed by Tajima (TAJIMA 1989), Fu and Li (FU and LI 1993), and Fay and Wu (FAY and WU 2000), which examine whether the pattern of SNPs at a given gene is consistent with neutrality.

Linkage disequilibrium (LD), which is the nonrandom association of polymorphisms (PRITCHARD and PRZEWORSKI 2001), can also be used to identify putative targets of selection. Because neutral polymorphisms can hitchhike to high frequency if they are linked to a positively selected polymorphism, recent selective sweeps often leave long genomic tracts of linked polymorphisms that are in strong LD. Such tracts are referred to as 'long range haplotypes' and can be identified by using large polymorphism datasets (SABETI *et al.* 2002; TOOMAJIAN *et al.* 2006; VOIGHT *et al.* 2006).

Another class of tests compares levels of polymorphism within a gene to levels of divergence at that gene between the species of interest and a closely related outgroup species. These tests are founded in the expectation that polymorphism and divergence levels at a locus should be proportional under neutrality. The Hudson–Kreitman–Aguade test (the HKA test; (HUDSON *et al.* 1987)), which compares polymorphism and divergence at a gene of interest to one or more neutral reference loci, is one commonly used form of this test. Another test in this class is the McDonald–Kreitman (the MK test; (MCDONALD and KREITMAN 1991)) test, which requires data only from a single locus to test whether genes exhibit an excess or deficiency of nonsynonymous polymorphisms or substitutions at a gene.

Another indicator of selection examines d_n/d_s (or K_a/K_s) ratios in protein-coding genes, where d_n (or K_a) is the nonsynonymous substitution rate and d_s (or K_s) is the

synonymous substitution rate for a particular gene (NEI and GOJOBORI 1986). Under neutrality, the d_n/d_s ratio of a gene is expected to equal 1, and departures from this expectation can be indicative of selection (see (NIELSEN 2001) for a more detailed discussion). The d_n/d_s ratios can be constructed for orthologous sequences obtained from multiple species or individuals, or for duplicate loci. More sophisticated tests using codon-based models to examine d_n/d_s have also been successful in identifying specific amino acid positions in a protein that show a history of positive selection (YANG 1997).

A final class of tests relies on divergence in allele frequencies at particular genes between populations as an indicator of selection. This is the basis of the Lewontin–Krakauer test, which tests whether the variance of F_{ST} estimates from different loci sampled from multiple populations is larger than what might be expected by chance (LEWONTIN and KRAKAUER 1973). Significant results from this test may be indicative of selection-driven population divergence.

The tests described in this section must be used with caution as patterns suggestive of selection can also arise from demographic effects (HEIN *et al.* 2004). It may be possible to control for these demographic effects through the use of empirical distributions of test statistics obtained from genomewide sequencing projects, rather than distributions from theoretical models, for estimating statistical significance (LUIKART *et al.* 2003). Loci that exist in the tails of these empirical distributions may be regarded as candidate adaptive genes subject to further examination. It should be noted that the value of a test statistic for a gene may fall in the tails of these empirical distributions by chance, so additional experimentation is necessary to prove that a gene is indeed adaptive.

An alternative to the aforementioned tests is to use simulation-based approaches to assess selection. In particular, coalescent theory (HUDSON 1991; KINGMAN 1982) has provided a powerful opportunity to detect selection at the molecular level under a variety of evolutionary scenarios (NORDBORG 2001). Coalescent simulation can be used to

generate distributions of genealogies against which samples of alleles can be statistically compared (HEIN *et al.* 2004; ROSENBERG and NORDBORG 2002). Additionally, other non-coalescent methods, such as Poisson random field (PRF)-based methods, have been constructed to estimate the selection coefficients of sampled genes (BUSTAMANTE *et al.* 2002). These methods have proven to be of great utility in the search for genes under selection.

Genetic mapping of plant adaptive genes. Multiple techniques, including QTL mapping and LD mapping methods, exist for mapping genes underlying adaptive traits based on marker-trait associations (as reviewed in (MACKAY 2001; PHILLIPS 2005; REMINGTON *et al.* 2001b; WHITT and BUCKLER 2003)). These methods have proven successful for mapping genes for trait variation in several plant species, such as *A. thaliana*, tomato, and maize (see (REMINGTON *et al.* 2001b) for discussion).

In QTL mapping, loci controlling trait variation between two individuals are mapped to specific genomic regions. Typically with this approach, individuals that differ in traits of interest are crossed to make F_2s or recombinant inbred lines (RILs). These lines are recombinant genotypes that possess genomic regions descended from each parent. When grown in a controlled setting or a common garden, the phenotypic differences in these lines can be mapped back to the genome based on trait associations with parental markers.

QTL mapping has played a prominent role in mapping genomic regions that control phenotypic variation in many species of plants. Most of these studies have resulted in the characterization of large-sized QTLs (> 500 kb) that span hundreds of genes (e.g. Ungerer et al., 2002). Subsequent fine-mapping using nearly isogenic lines (NILs) is generally necessary to localize the gene(s) of effect within identified QTLs (TANKSLEY 1993). The resolution of QTL mapping experiments can be improved by increasing marker density, the number of RILs, and the number of generations of

intercrossing during line construction, but rarely have QTLs been localized to regions of fewer than 10 genes.

An additional caveat with QTL mapping is that the typical biparental origin of most mapping populations may lead to the identification of QTLs that are found only in one of the parental lines and are not prevalent in the general population or species. This almost certainly has been the case for some QTLs that have been identified in *A. thaliana*, which due to its species history possesses a high number of rare polymorphisms throughout the genome that are often present in only a single accession. For other plant species that have higher outcrossing rates and less population structure, QTLs are more likely to be representative of common genetic variation segregating within the population or species.

An alternative approach to mapping adaptive genes is to perform LD mapping. Because polymorphisms that are in LD with a functionally important polymorphism will also be associated with any phenotypic differences caused by that polymorphism, LD can be exploited to map the genomic regions that underlie adaptations. In practice, LD mapping requires a sample of genotyped and phenotyped individuals taken from a natural population. Associations between observed genetic variants and trait variation in this sample can then be measured, leading to the identification of specific polymorphisms or haplotypes that explain adaptive trait variation.

In LD mapping, the mapping resolution is primarily influenced by the rate of LD decay. In *A. thaliana*, LD usually decays within 5 to 10 kb (KIM *et al.* 2007; NORDBORG *et al.* 2005), which suggests that LD mapping may have a resolution orders of magnitude higher than QTL mapping (ARANZANA *et al.* 2005). In maize, LD decays faster than in *A. thaliana*, oftentimes within a couple kilobases (REMINGTON *et al.* 2001a), suggesting LD may also be a promising approach for this species. Because LD patterns vary

substantially across plant species, the utility of this method for non-model plants has yet to be determined.

LD mapping is not without its caveats. In particular, this method is prone to spurious results based on population structure. Techniques exist for assessing the extent of cryptic population structure and accounting for it in association tests (PRITCHARD and ROSENBERG 1999; PRITCHARD *et al.* 2000a; PRITCHARD *et al.* 2000b). Recent surveys in *A. thaliana* found that including estimates of population structure as covariates in association tests dramatically reduced the number of false positives throughout the genome (ARANZANA *et al.* 2005; ZHAO *et al.* 2007). Another concern for LD mapping is its lack of power for identifying associations at loci with very low minor allele frequencies. Increased sampling can reduce this problem, but not entirely solve it in some plant species. Overall, LD mapping has proven successful in plants, such as in maize (THORNSBERRY *et al.* 2001; WILSON *et al.* 2004) and Arabidopsis (e.g. (ARANZANA *et al.* 2005; CAICEDO *et al.* 2004; OLSEN *et al.* 2004)).

Contemporary studies of adaptation in plants

Although few studies of adaptation have spanned all relevant levels of biological organization (WRIGHT and GAUT 2005), several plant adaptations have been extensively examined from ecological, evolutionary, and molecular perspectives. In the following section, we will summarize some of these notable examples in which specific genes responsible for putatively or established adaptive phenotypic variation have been identified. These studies represent the application of the techniques described in the previous sections.

Adaptive trait locus mapping using selection signatures in plant genomes. For *Arabidopsis thaliana*, there have been several moderate- to large-scale screens for adaptive genes based on tests for departure from neutrality. A sequence-based screen of

334 randomly distributed genomic regions among 12 ecotypes, for example, led to the identification of 28 loci that were in the tails of the empirical distribution of various test statistics (SCHMID *et al.* 2005). In a similar screen of rapidly evolving genes between *A. thaliana* and *A. lyrata*, 14 genes among 304 compared orthologues were shown to have K_a values exceeding K_s values (BARRIER *et al.* 2003). Six of these genes were examined further by comparing within- to between-species patterns of nucleotide change in the coding regions of these loci, and these rapidly evolving genes were demonstrated to have a higher average selection intensity than previously studied genes in *A. thaliana*.

One can also use genomic screens to identify loci that have been under balancing selection or that have been involved in local adaptation. Genes subject to these types of selection pressures are expected to have high levels of intraspecific variation (HUDSON and KAPLAN 1988). A recent screen for high diversity genes in *A. thaliana* found three genomic regions with higher variation than the rest of the genome (CORK and PURUGGANAN 2005). One of these genomic regions harbored a member of a class of disease resistance genes commonly associated with balanced polymorphisms, while the putative reasons for selection for high diversity on the other loci remain unclear.

The genomic screen approach has also been used in maize to identify loci with reduced levels of diversity due to selection associated with crop domestication (WRIGHT *et al.* 2005). In this study, 774 loci were sampled from 14 maize and 16 teosinte inbred lines to estimate the severity of the bottleneck associated with the domestication of maize from teosinte. Using a bottleneck scenario to model the demographic effects of domestication, they identified two classes of genes – one class of genes whose sequence variation is consistent with the bottleneck and another class of genes that are putative domestication loci. Two to four percent of the studied loci were estimated to belong to the latter group, and extension of these results to the entire maize genome suggests that up to 1,200 genes could have been responsible for maize domestication.

Floral adaptations in Ipomoea. Floral color variation in the American morning glory (*Ipomoea*), has been implicated in adaptive evolution (CLEGG and DURBIN 2003). These adaptive changes are believed to be driven by complex interactions between these plants and their pollinators and are based on pollinator preferences for particular floral colors. Because the molecular pathways underlying floral pigmentation have been well characterized, researchers have been successful in identifying the molecular bases for much of the floral color variation in these species.

Ipomoea purpurea, which is indigenous to Mexico and was likely introduced into the southeastern U.S. concomitant with maize culture, possesses three main floral colors – blue, red, and white – that are often found in different color blends and patterns. In Mexico, most populations are fixed for blue flowers, but in the U.S. white and red flowers are not uncommon. Selection has been implicated in the maintenance of the white polymorphism, but no evidence has been found for selection on the other floral variants (CLEGG and DURBIN 2000). Ecological studies have suggested that white flowers are discriminated against by pollinators when rare but that they are maintained in populations through self-fertilization (CLEGG and DURBIN 2000).

Floral color in *Ipomoea* is determined primarily by the relative concentrations of two anthocyanin derivatives – cyanidin, which produces blue flowers, and pelargonidin, which produces red flowers. Two anthocyanin pathway genes have been identified that control floral color and patterning variation in *I. purpurea* (CLEGG and DURBIN 2003). *Flavonoid 3′-hydroxylase* (*F3′H*) has been shown to confer the dominant blue phenotype and the recessive red phenotype (ZUFALL and RAUSHER 2003). In addition, c*halcone synthase-D* (*CHS-D*), which possesses many natural alleles, was found to control floral color patterning through epistasis with other loci (HABU *et al.* 1998). Though the cloning of these genes represents a major achievement, no selective basis has been demonstrated for the maintenance of these polymorphisms in nature.

Multiple subgenera of *Ipomoea*, including the clade containing *I. quamoclit*, another well-studied species in this genus, have undergone changes from blue to red flowers to facilitate adaptive pollinator shifts from bees to birds (ZUFALL and RAUSHER 2004). Molecular analysis of anthocyanin pathway genes in *I. quamoclit* revealed that *F3′H* mRNA levels are dramatically reduced in this species relative to the predominantly blue *I. purpurea*, though biochemical analysis showed *I. quamoclit's F3′H* is still functional. In addition, *I. quamoclit*'s *dihydroflavonol reductase-B* (*DFR-B*), one of three paralogues of this gene in *Ipomoea*, was found to have numerous insertions and amino acid substitutions, as well as a 59-bp upstream shift of the stop codon. Heterologous complementation tests in the model genetic system *A. thaliana* found that the *I. purpurea DFR-B* gene could complement *A. thaliana DFR* null mutants, but that *I. quamoclit*'s *DFR-B* gene could not. These results provide strong evidence that one or both of these molecular changes – the *F3′H* regulatory changes and the *DFR-B* mutations – are responsible for the adaptive floral color transition from blue to red in *I. quamoclit*.

Flowering time variation in A. thaliana. Flowering time is a major developmental transition in plants, and this trait is likely a strong determinant of fecundity (SIMPSON and DEAN 2002). In the model genetic species *A. thaliana*, the timing of flowering varies significantly between different accessions (NORDBORG and BERGELSON 1999), although the adaptive significance of this variation is still under active exploration (see (ENGELMANN and PURUGGANAN 2006) for a discussion of this ongoing research). Flowering time in *A. thaliana* has been shown to exhibit a latitudinal cline, suggesting the possible adaptation of this trait to a geographical/climactic component (STINCHCOMBE *et al.* 2004).

Over 60 genes have been shown to regulate flowering time in *A. thaliana,* illustrating the complex molecular circuitry underlying this trait. Much less is known about the genetic controls of natural variation in flowering time, because only a handful of genes have been cloned that contribute to flowering time differences across accessions

of this species. *CRYPTOCHROME 2* (*CRY2*), which is involved in blue light photoreception, was one of the first genes to be shown to contribute to flowering time variation in *A. thaliana* (EL-ASSAL *et al.* 2001). Two amino acid polymorphisms were identified that result in altered *CRY2* protein levels during the circadian cycle, causing the early flowering of plants under short day conditions. These polymorphisms, however, were found only in an accession from the Cape Verde Islands (the Cvi ecotype), and it is unclear whether this allele is of any significance to local adaptation. A more recent association study suggested that more common haplotypes of *CRY2* could also contribute to flowering time variation in this species (OLSEN *et al.* 2004).

A similarly rare polymorphism has been observed in the *FLOWERING LOCUS M/MADS AFFECTING FLOWERING 1* (*FLM*/*MAF1*) gene of the Nd-1 accession of *A. thaliana*, which was collected from Niederzenz, Germany (WERNER *et al.* 2005). Initially identified as a QTL controlling over 60% of the flowering time variation in an RIL population derived from the Nd-1 and Columbia (Col-3 and Col-5) ecotypes, sequencing of the *FLM* genomic region in Nd-1 found that it was entirely absent in this accession. Genotyping of a larger group of accessions showed that the Nd-1 *FLM* deletion was unique to plants sampled from Niederzenz.

The most significant contributor to flowering time variation that has been characterized in *A. thaliana* to date, both in terms of effect and frequency, is the *FRIGIDA* (*FRI*) gene (JOHANSON *et al.* 2000). Molecular analysis revealed that multiple loss-of-function *FRI* alleles possessing large deletions segregate in natural populations of *A. thaliana*, at least two of which are found at moderate frequency throughout the species range (HAGENBLAD and NORDBORG 2002; HAGENBLAD *et al.* 2004; JOHANSON *et al.* 2000; LE CORRE *et al.* 2002; STINCHCOMBE *et al.* 2004). Among *A. thaliana* accessions carrying a functional *FRI* allele, there exists a *FRI* genotype-dependent latitudinal cline in flowering time under field conditions (STINCHCOMBE *et al.* 2004). Recent work has shown that the *FRI*'s association with flowering time is detectable with markers ≥100 kb

away from *FRI* (ARANZANA *et al.* 2005). This atypically extensive LD around *FRI* may be due to *FRI's* involvement in local adaptation and selective sweeps**.**

An epistatic effect of *FRI* on *FLOWERING LOCUS C* (*FLC*), which encodes a MADS box floral repressor that is known to be upregulated by *FRI*, may be responsible for this latitudinal cline in flowering time (CAICEDO *et al.* 2004). Two major *FLC* haplotype groups have been detected in *A. thaliana*, and there is significant flowering time variation associated with *FRI-FLC* two-locus genotypes. *FLC* haplotypes also show a significant latitudinal distribution, but only in functional *FRI* backgrounds. Finally, *FRI* and *FLC* show significant intergenic linkage disequilibrium, even though the two genes are found on different *A. thaliana* chromosomes. However, there is some question as whether this clinal pattern reflects some unrecognized cryptic population structure. These findings, however, suggest that epistatic selection may underlie flowering time variation in this species, although these associations need confirmation by molecular analysis of the different alleles at these loci (CAICEDO *et al.* 2004; STINCHCOMBE *et al.* 2004).

Although these four genes have been shown to contain polymorphisms that underlie natural variation in flowering time in *A. thaliana*, QTL mapping experiments suggest that there are many other loci that contribute to this trait variation (UNGERER *et al.* 2002; UNGERER *et al.* 2003; WEINIG *et al.* 2002). Interestingly, the genes controlling flowering time variation appear to differ between laboratory and field conditions (WEINIG *et al.* 2002). This emphasizes the importance of studying the genetics of adaptive traits in ecological settings. Work remains to identify these ecologically relevant genetic polymorphisms and their potential contributions to life history adaptation in this species.

Disease resistance in A. thaliana. The interaction of pathogens and their hosts has been a powerful model for the study of coevolution (BENT 1996; BERGELSON *et al.* 2001). Numerous disease resistance loci have been identified in the plant genetic model *A.*

thaliana (GRANT *et al.* 1995; KUNKEL *et al.* 1993), and multiple studies of the molecular evolutionary dynamics of pathogen resistance loci in *A. thaliana* have been conducted (CAICEDO *et al.* 1999; MAURICIO *et al.* 2003; STAHL *et al.* 1999; TIAN *et al.* 2002). These studies suggest that resistance (*R*) genes are often maintained as highly divergent alleles or presence/absence polymorphisms due to balancing selection driven by interactions between host plants and their pathogens (BERGELSON *et al.* 2001).

One extensively studied *R* gene is *RPM1*, which encodes an NBS-LRR protein that confers resistance to *Pseudomonas syringae* strains carrying either the *AvrRPM1* or *AvrB* avirulence genes (STAHL *et al.* 1999). This gene exists as a presence/absence polymorphism across the *A. thaliana* species range, and the genomic region surrounding it possesses the molecular signature of balancing selection. The creation of transgenic lines differing only in the presence or absence of *RPM1* was used to demonstrate that a fitness tradeoff at *RPM1* could be responsible for the maintenance of this balanced polymorphism (TIAN *et al.* 2003).

A larger-scale study of the 20-kb genomic region containing *RPS5*, another NBS-LRR protein-encoding *R* gene also found evidence for balancing selection (TIAN *et al.* 2002). This gene confers resistance to *Pseudomonas syringae* strains that express the *AvrPph3* avirulence gene. Like *RPM1*, this locus also possesses a widespread presence/absence polymorphism, but an additional susceptible allele that contains a frameshift mutation was discovered. The pattern of molecular variation and linkage disequilibrium at this locus suggests that balancing selection is acting upon *RPS5*. These *RPS2* and *RPS5* studies provide excellent examples of how ecologically-based selection can be connected to molecular signatures at the genomic level.

The genomics of herbivore resistance and growth rate in A. thaliana. How plants defend themselves against their predators has also played a central role in the study of the molecular evolution of ecological interactions. Glucosinolates, a class of secondary

metabolites produced by many plants, are thought to provide an important defense against herbivory in the Brassicaceae. Much is known about the biochemical pathways that produce glucosinolates in *A. thaliana* and a gene controlling variation in the types of glucosinolates synthesized, *methylthioalkylmalate synthase1* (*MAM1*), has been identified in this species (KROYMANN *et al.* 2001).

Sequencing of the *MAM* genomic region in 25 ecotypes found that a closely linked paralogue of *MAM1*, designated *MAM2*, was often present (KROYMANN *et al.* 2003). The presence of *MAM1*, *MAM2*, or both genes in the *MAM* genomic region was highly variable. A significant association was detected between the types of glucosinolates produced by an ecotype and its *MAM1/MAM2* genotype. In addition, the pattern of genetic variation at *MAM2* was indicative of balancing selection acting on this gene, suggesting that an ecological trade-off might influence this locus.

Unintentionally, further analysis of the upstream and downstream genomic regions surrounding the *MAM* cluster found two QTLs for growth rate within 100 kb of each other (KROYMANN and MITCHELL-OLDS 2005). One of these QTLs was fine-mapped to a single gene, at which balancing selection was detected from a molecular population genetic sample. The other QTL was refined to a 32-kb interval, but could not be mapped to a specific gene. Epistasis was shown to exert strong influence upon the effects of these loci on growth rate. The identification of these growth rate loci represents a major achievement, but whether they contribute to fitness in the natural habitat of *Arabidopsis* needs to be determined.

The signature of selection in domestication genes. Domestication has long been viewed as a model for adaptation (DARWIN 1859; DARWIN 1897). Investigating the evolution of crop "domestication traits" favored by early farming cultures (e.g. the loss of seed dispersal mechanisms or an increased yield under agricultural field conditions), in addition to the subsequent diversification of other crop traits that have arisen from

selective breeding to satisfy diverse human cultural preferences (e.g. different grain colors and tastes), can provide tremendous insight into the genomic signature of selection on specific traits (DOEBLEY 2004). Furthermore, though domestication may not precisely replicate adaptation, the characterization of the genes responsible for domestication traits may help to clarify the number and effects of genes that underlie the changes in selected traits.

The study of the molecular genetics of domestication is most clearly illustrated by studies of maize (*Zea mays* subsp. *mays*). Maize possesses a strikingly different morphology from its teosinte progenitor (*Z. mays* subsp. *parviglumis*), and QTL mapping studies have identified several genomic regions that harbor genes responsible for this species differentiation (DOEBLEY 2004). Two major phenotypic changes associated with maize domestication were a substantial decrease in axillary branching (DOEBLEY *et al.* 1997; GALLAVOTTI *et al.* 2004) and a reduction in glume size that exposed kernels on the corn ears (DORWEILER *et al.* 1993; WANG *et al.* 2005). Substantial effort has resulted in the characterization of the genes responsible for both of these maize domestication traits.

The increase in apical dominance in maize relative to teosinte has been mapped to the *teosinte branched1* (*tb1*) gene, which has higher expression in maize than teosinte (DOEBLEY *et al.* 1997). This gene has a significant depression in genetic variability in the 5′ non-transcribed region of maize *tb1* compared to teosinte *tb1*, suggesting that this region could have been the target of positive selection that resulted in a change in gene regulation (WANG *et al.* 1999). More extensive analysis of the genomic region surrounding this region has shown that the selective sweep associated with the fixation of *tb1* in maize extends for more than 60 kb of the upstream intergenic region (CLARK *et al.* 2004). The *barren stalk1* (*ba1*) gene, an additional domestication locus that is epistatic to *tb1*, has been characterized, and has been proposed to be a target of selection during postdomestication improvement of maize (GALLAVOTTI *et al.* 2004).

A major effect gene has also been characterized for glume reduction in maize. This gene, *teosinte glume architecture 1* (*tga1*), which was also initially identified as a QTL for domestication traits between maize and teosinte, explains up to 50% of the glume reduction associated with domestication (DORWEILER *et al.* 1993). The pattern of variation at *tga1* suggests that this locus has been the target of a selective sweep (WANG *et al.* 2005), again showing that the signature of artificial selection is often apparent at the molecular level in domesticated species. In addition to *tb1*, *ba1*, and *tga1*, hundreds of other domestication genes may exist in maize (as described in an earlier section (WRIGHT *et al.* 2005)).

Efforts to identify genes underlying domestication traits in other species have also been successful. Heterotopic expression of the *MPF2* gene, which encodes a MADS box protein, is suggested to have led to the evolution of the fruit husks in the genus *Physalis* (HE and SAEDLER 2005). Positive selection on the *CAULIFLOWER* gene has been shown to have been important in the origin of cauliflower (*Brassica oleracea* subsp. *botrytis*) and possibly broccoli (*B. oleracea* subsp. *italica*) as well (PURUGGANAN *et al.* 2000). In rice, selection on a mutation impairing proper pre-mRNA splicing of the *Waxy* gene transcript has been critical to the evolution of glutinous ("sticky") rice (OLSEN and PURUGGANAN 2002) and possibly the major variety group referred to as temperate japonica (*Oryza sativa* subsp. *japonica*; Olsen et al., 2006). Lastly, multiple genes that may have been important to differences in fruit morphology between wild and domestic tomato have been identified (FRARY *et al.* 2000; LIU *et al.* 2002). The study of the domestication of crop plants has provided many insights into the genetic targets of selection on specific traits.

Discussion

Numerous potentially adaptive genes have been characterized in plants. The cloning of adaptive and domestication genes thus far has provided substantial support for

the importance of major effect mutations to plant adaptation (e.g., the *FRI* locus in *A. thaliana* (JOHANSON *et al.* 2000); *tb1* in maize (DOEBLEY *et al.* 1997)). At present, no minor effect loci have been cloned in plants, but their existence is supported by QTL studies (a fact reported in (TANKSLEY 1993)and elsewhere). However, for no polygenic adaptation has the entire genetic architecture been determined, making it difficult to infer how many genes are typically responsible for adaptations. In addition, whether certain types of genes or mutations are more frequently responsible for adaptation is an unanswered question.

To better understand the genetic basis of plant adaptation in general, a transition from *A. thaliana* and crop species into other non-model plants is necessary. Genomic resources are being developed for genera like *Mimulus* that have longstanding traditions in ecology and evolution, but poor genomic resources. In some cases, it may be possible to use present genetic models as platforms for closely related non-models. Much research is being done in the Brassicaceae, the family containing *A. thaliana*, based on this paradigm. More examples of adaptive genes are needed in non-model plant species.

The continued characterization of adaptive genes in plants will make it possible to answer some longstanding questions about adaptation. For instance, that epistasis in genetic networks and pathways can shape how adaptations evolve is generally accepted, but the effect of natural selection on such systems is not well understood (CORK and PURUGGANAN 2004). In addition, why parallel and repeated adaptations, such as the transition to self-compatibility, have occurred numerous times in plants across both intraspecific and interspecific scales can only be fully understood once the genetic architectures for these traits have been identified (WOOD *et al.* 2005). These are just two areas in evolutionary biology that will benefit from the mapping of adaptive genes.

The research described in this paper parallels efforts of the animal research community to identify adaptive genes and to understand the effects of natural selection at

the genomic level. For instance, selection signature screens of genes throughout the human genome have identified loci that have likely been important to human adaptive evolution (BUSTAMANTE *et al.* 2005; CLARK *et al.* 2003). Efforts to classify the proportion of adaptive and deleterious mutations throughout animal genomes have provided another avenue for exploring how genomes evolve under selection (EYRE-WALKER and KEIGHTLEY 1999; SMITH and EYRE-WALKER 2002). The extension of the methods used in these examples to plants will help to determine how natural selection has shaped plant genomes.

Conclusion

Plant evolutionary genetics has experienced many successes, but much work remains in determining plant adaptive traits and their underlying genetic controls. Herein, we attempt to contribute to this area by using molecular population genetics and various mapping approaches to identify genes involved in variation in traits of ecological significance in *A. thaliana*. Chapters 2 and 3 focus on candidate gene association mapping of shoot branching and flowering time, respectively. Chapter 4 is a report of a resequencing-based examination of the possible contributions of microRNAs to trait variation in *A. thaliana*. Lastly, Chapter 5 is a synthesis of the preceding chapters, with a discussion not only of our findings, but also of where this field is heading given the advances in knowledge and technology that have occurred since this dissertation was initiated in 2004. The underlying paradigm of this dissertation is that to better understand how adaptive evolution occurs, the actual genes and, more specifically, the polymorphisms influencing adaptive trait changes must be identified. Progress in this area will help to explain the extant condition of plant biodiversity and should provide important insights into how adaptation occurs at the genetic level.

Acknowledgments

The authors would like to thank members of the Purugganan laboratory and two anonymous reviewers for a critical reading of the manuscript. This work was funded in part by a Graduate Assistance in Areas of National Need Fellowship from the U.S. Department of Education to I. M. E. and grants from the U.S. National Science Foundation's Frontiers in Integrated Biological Research and Plant Genome Research Programs to M. D. P.

References

ARANZANA, M. J., S. KIM, K. ZHAO, E. BAKKER, M. HORTON *et al.*, 2005 Genomewide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes. PLoS Genet 1: e60.

BARRIER, M., C. D. BUSTAMANTE, J. YU and M. D. PURUGGANAN, 2003 Selection on rapidly evolving proteins in the Arabidopsis genome. Genetics 163: 723-733.

BENT, A. F., 1996 Plant disease resistance genes: function meets structure. Plant Cell 8: 1757-1771.

BERGELSON, J., M. KREITMAN, E. A. STAHL and D. TIAN, 2001 Evolutionary dynamics of plant R-genes. Science 292: 2281-2285.

BUSTAMANTE, C. D., A. FLEDEL-ALON, S. WILLIAMSON, R. NIELSEN, M. T. HUBISZ *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature 416: 531-534.

CAICEDO, A. L., B. A. SCHAAL and B. N. KUNKEL, 1999 Diversity and molecular evolution of the RPS2 resistance gene in Arabidopsis thaliana. Proc Natl Acad Sci U S A 96: 302-306.

CAICEDO, A. L., J. R. STINCHCOMBE, K. M. OLSEN, J. SCHMITT and M. D. PURUGGANAN, 2004 Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci U S A 101: 15670-15675.

CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. THOMAS, A. KEJARIWAL *et al.*, 2003 Positive selection in the human genome inferred from human-chimp-mouse orthologous gene alignments. Cold Spring Harb Symp Quant Biol 68: 471-477.

CLARK, R. M., E. LINTON, J. MESSING and J. F. DOEBLEY, 2004 Pattern of diversity in the genomic region near the maize domestication gene tb1. Proc Natl Acad Sci U S A 101: 700-707.

CLEGG, M. T., and M. L. DURBIN, 2000 Flower color variation: A model for the experimental study of evolution. Proc Natl Acad Sci U S A 97: 7016-7023.

CLEGG, M. T., and M. L. DURBIN, 2003 Tracing floral adaptations from ecology to molecules. Nat Rev Genet 4: 206-215.

CORK, J. M., and M. D. PURUGGANAN, 2004 The evolution of molecular genetic pathways and networks. Bioessays 26: 479-484.

CORK, J. M., and M. D. PURUGGANAN, 2005 High-diversity genes in the Arabidopsis genome. Genetics 170: 1897-1911.

DARWIN, C., 1859 *On the Origin of Species*. Harvard University Press, Cambridge, Massachusetts.

DARWIN, C., 1897 *The Variation of Animals and Plants Under Domestication*. D. Appleton & Co., New York.

DOEBLEY, J., 2004 The genetics of maize evolution. Annu Rev Genet 38: 37-59.

DOEBLEY, J., A. STEC and L. HUBBARD, 1997 The evolution of apical dominance in maize. Nature 386: 485-488.

DORWEILER, J., A. STEC, J. KERMICLE and J. DOEBLEY, 1993 *Teosinte glume architecture 1*: a genetic locus controlling a key step in maize evolution. Science 262: 233-235.

EL-ASSAL, S. E.-D., C. ALONSO-BLANCO, A. J. 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.

ENGELMANN, K. E., and M. D. PURUGGANAN, 2006 The molecular evolutionary ecology of plant development: flowering time in *Arabidopsis thaliana*. Advances in Botanical Research 44: 505.

EYRE-WALKER, A., and P. D. KEIGHTLEY, 1999 High genomic deleterious mutation rates in hominids. Nature 397: 344-347.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

FISHER, R., 1930 *The genetical theory of natural selection*. Oxford University Press, Oxford.

FRARY, A., T. C. NESBITT, S. GRANDILLO, E. KNAAP, B. CONG *et al.*, 2000 fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85-88.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693-709.

GALLAVOTTI, A., Q. ZHAO, J. KYOZUKA, R. B. MEELEY, M. K. RITTER *et al.*, 2004 The role of barren stalk1 in the architecture of maize. Nature 432: 630-635.

GILLESPIE, J., 1991 *The causes of molecular evolution*. Oxford University Press, Oxford.

GRANT, M. R., L. GODIARD, E. STRAUBE, T. ASHFIELD, J. LEWALD *et al.*, 1995 Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. Science 269: 843-846.

HABU, Y., Y. HISATOMI and S. IIDA, 1998 Molecular characterization of the mutable flaked allele for flower variegation in the common morning glory. Plant J 16: 371-376.

HAGENBLAD, J., and M. NORDBORG, 2002 Sequence variation and haplotype structure surrounding the flowering time locus FRI in Arabidopsis thaliana. Genetics 161: 289-298.

HAGENBLAD, J., C. TANG, J. MOLITOR, J. WERNER, K. ZHAO *et al.*, 2004 Haplotype structure and phenotypic associations in the chromosomal regions surrounding two Arabidopsis thaliana flowering time loci. Genetics 168: 1627-1638.

HE, C., and H. SAEDLER, 2005 Heterotopic expression of MPF2 is the key to the evolution of the Chinese lantern of Physalis, a morphological novelty in Solanaceae. Proc Natl Acad Sci U S A 102: 5779-5784.

HEIN, J., M. SCHIERUP and C. WIUF, 2004 *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, Oxford, UK.

HUDSON, R. R., 1991 Gene genealogies and the coalescent process, pp. 1-44 in *Oxford Surveys in Evolutionary Biology*, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford, UK.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics 120: 831-840.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

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.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics 123: 887-899.

KIM, S., V. PLAGNOL, T. T. HU, C. TOOMAJIAN, R. M. CLARK *et al.*, 2007 Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat Genet 39: 1151-1155.

KIMURA, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

KINGMAN, J., 1982 The coalescent. Stochastic Processes and Their Applications 13: 235-248.

KROYMANN, J., S. DONNERHACKE, D. SCHNABELRAUCH and T. MITCHELL-OLDS, 2003 Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc Natl Acad Sci U S A 100 Suppl 2: 14587-14592.

KROYMANN, J., and T. MITCHELL-OLDS, 2005 Epistasis and balanced polymorphism influencing complex trait variation. Nature 435: 95-98.

KROYMANN, J., S. TEXTOR, J. G. TOKUHISA, K. L. FALK, S. BARTRAM *et al.*, 2001 A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. Plant Physiol 127: 1077-1088.

KUNKEL, B. N., A. F. BENT, D. DAHLBECK, R. W. INNES and B. J. STASKAWICZ, 1993 RPS2, an Arabidopsis disease resistance locus specifying recognition of Pseudomonas syringae strains expressing the avirulence gene avrRpt2. Plant Cell 5: 865-875.

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. Molecular Biology and Evolution 19: 1261-1271.

LEWONTIN, R. C., and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.

LIU, J., J. VAN ECK, B. CONG and S. D. TANKSLEY, 2002 A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc Natl Acad Sci U S A 99: 13302-13306.

LUIKART, G., P. R. ENGLAND, D. TALLMON, S. JORDAN and P. TABERLET, 2003 The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4: 981-994.

MACKAY, T. F., 2001 The genetic architecture of quantitative traits. Annu Rev Genet 35: 303-339.

MAURICIO, R., E. A. STAHL, T. KORVES, D. TIAN, M. KREITMAN *et al.*, 2003 Natural selection for polymorphism in the disease resistance gene Rps2 of Arabidopsis thaliana. Genetics 163: 735-746.

MAYNARD-SMITH, J., and J. HAIGH, 1974 The hitchhiking effect of a favorable gene. Genet Res 23: 23-35.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

NEI, M., and T. GOJOBORI, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418-426.

NIELSEN, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity 86: 641-647.

NORDBORG, M., 2001 Coalescent theory, pp. 179-212 in *Handbook of Statistical Genetics*, edited by D. BALDING, M. BISHOP and C. CANNINGS. Wiley, Chichester, UK.

NORDBORG, M., and J. BERGELSON, 1999 The effect of seed and rosette cold treatment on germination and flowering time in some Arabidopsis thaliana (Brassicaceae) ecotypes. Am J Bot 86: 470.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3: e196.

OLSEN, K. M., S. S. HALLDORSDOTTIR, J. R. STINCHCOMBE, C. WEINIG, J. SCHMITT *et al.*, 2004 Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics 167: 1361-1369.

OLSEN, K. M., and M. D. PURUGGANAN, 2002 Molecular evidence on the origin and evolution of glutinous rice. Genetics 162: 941-950.

ORR, H., 1998 The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52: 935-949.

ORR, H. A., 2005a The genetic theory of adaptation: a brief history. Nat Rev Genet 6: 119-127.

ORR, H. A., 2005b Theories of adaptation: what they do and don't say. Genetica 123: 3-13.

PHILLIPS, P. C., 2005 Testing hypotheses regarding the genetics of adaptation. Genetica 123: 15-24.

PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1-14.

PRITCHARD, J. K., and N. A. ROSENBERG, 1999 Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65: 220-228.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000a Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000b Association mapping in structured populations. Am J Hum Genet 67: 170-181.

PURUGGANAN, M. D., A. L. BOYLES and J. I. SUDDITH, 2000 Variation and selection at the CAULIFLOWER floral homeotic gene accompanying the evolution of domesticated Brassica oleracea. Genetics 155: 855-862.

REMINGTON, D. L., and M. PURUGGANAN, 2003 Candidate genes, quantitative trait loci, and functional trait evolution in plants. Int. J. Plant Sci. 164: S7-S20.

REMINGTON, D. L., J. M. THORNSBERRY, Y. MATSUOKA, L. M. WILSON, S. R. WHITT *et al.*, 2001a Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci U S A 98: 11479-11484.

REMINGTON, D. L., M. C. UNGERER and M. D. PURUGGANAN, 2001b Map-based cloning of quantitative trait loci: progress and prospects. Genet Res 78: 213-218.

ROSENBERG, N. A., and M. NORDBORG, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3: 380-390.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in Arabidopsis thaliana reveals a genomewide departure from a neutral model of DNA sequence polymorphism. Genetics 169: 1601-1615.

SIMPSON, G. G., and C. DEAN, 2002 Arabidopsis, the Rosetta stone of flowering time? Science 296: 285-289.

SMITH, N. G., and A. EYRE-WALKER, 2002 Adaptive protein evolution in Drosophila. Nature 415: 1022-1024.

STAHL, E. A., G. DWYER, R. MAURICIO, M. KREITMAN and J. BERGELSON, 1999 Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature 400: 667-671.

STINCHCOMBE, J. R., C. WEINIG, M. UNGERER, K. M. OLSEN, C. MAYS *et al.*, 2004 A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci U S A 101: 4712-4717.

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

TANKSLEY, S. D., 1993 Mapping polygenes. Annu Rev Genet 27: 205-233.

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.

TIAN, D., H. ARAKI, E. STAHL, J. BERGELSON and M. KREITMAN, 2002 Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A 99: 11525-11530.

TIAN, D., M. B. TRAW, J. Q. CHEN, M. KREITMAN and J. BERGELSON, 2003 Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. Nature 423: 74-77.

TOOMAJIAN, C., T. T. HU, M. J. ARANZANA, C. LISTER, C. TANG *et al.*, 2006 A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. PLoS Biol 4: e137.

UNGERER, M. C., S. S. HALLDORSDOTTIR, J. L. MODLISZEWSKI, T. F. 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. MACKAY, 2003 Genotype-environment interactions at quantitative trait loci affecting inflorescence development in Arabidopsis thaliana. Genetics 165: 353-365.

VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. PLoS Biol 4: e72.

WANG, H., T. NUSSBAUM-WAGLER, B. LI, Q. ZHAO, Y. VIGOUROUX *et al.*, 2005 The origin of the naked grains of maize. Nature 436: 714-719.

WANG, R. L., A. STEC, J. HEY, L. LUKENS and J. DOEBLEY, 1999 The limits of selection during maize domestication. Nature 398: 236-239.

WEINIG, C., M. C. UNGERER, L. A. DORN, N. C. KANE, Y. TOYONAGA *et al.*, 2002 Novel loci control variation in reproductive timing in Arabidopsis thaliana in natural environments. Genetics 162: 1875-1884.

WERNER, J. D., J. O. BOREVITZ, N. WARTHMANN, G. T. TRAINER, J. R. ECKER *et al.*, 2005 Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci U S A 102: 2460-2465.

WHITT, S. R., and E. S. BUCKLER, IV., 2003 Using natural allelic diversity to evaluate gene function in *Plant Functional Genomics: Methods and Protocols*, edited by E. GROTEWALD. Humana Press.

WILSON, L. M., S. R. WHITT, A. M. IBANEZ, T. R. ROCHEFORD, M. M. GOODMAN *et al.*, 2004 Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16: 2719-2733.

WOOD, T. E., J. M. BURKE and L. H. RIESEBERG, 2005 Parallel genotypic adaptation: when evolution repeats itself. Genetica 123: 157-170.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection on the maize genome. Science 308: 1310-1314.

WRIGHT, S. I., and B. S. GAUT, 2005 Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22: 506-519.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

ZHAO, K., M. J. ARANZANA, S. KIM, C. LISTER, C. SHINDO *et al.*, 2007 An Arabidopsis example of association mapping in structured samples. PLoS Genet 3: e4.

ZUFALL, R. A., and M. D. RAUSHER, 2003 The genetic basis of a flower-color polymorphism in the common morning glory, *Ipomoea purpurea*. Journal of Heredity 94: 442-448.

ZUFALL, R. A., and M. D. RAUSHER, 2004 Genetic changes associated with floral adaptation restrict future evolutionary potential. Nature 428: 847-850.

CHAPTER TWO:

The genetic architecture of shoot branching in *Arabidopsis thaliana*: A comparative assessment of candidate gene associations *vs.* quantitative trait locus mapping

The genetic architecture of shoot branching in *Arabidopsis thaliana*: A comparative assessment of candidate gene associations *vs.* quantitative trait locus mapping

Ian M. Ehrenreich[1,2], Phillip A. Stafford[1], and Michael D. Purugganan[2]

[1]Department of Genetics, Box 7614, North Carolina State University, Raleigh, North Carolina 27695 USA
[2]Department of Biology and Center for Genomics and Systems Biology, New York University, 100 Washington Square East, New York, New York 10003 USA

Corresponding author: Michael D. Purugganan, Telephone: (212) 992-9628, Email: mp132@nyu.edu.

Contributions: IME and MDP designed this study. IME and PAS conducted all experiments and analyses. IME and MDP cowrote this manuscript.

This chapter is a modified version of previously published work.

Reference: Ehrenreich, I.M., Stafford, P.A., and M.D. Purugganan. 2007. The genetic architecture of shoot branching in *Arabidopsis thaliana*: A comparative assessment of candidate gene associations *vs.* quantitative trait locus mapping. *Genetics* 176: 1223-1236.

Abstract

Association mapping focused on 36 genes involved in branch development was used to identify candidate genes for variation in shoot branching in *Arabidopsis thaliana*. The associations between four branching traits and moderate-frequency haplogroups at the studied genes were tested in a panel of 96 accessions from a restricted geographic range in Central Europe. Using a mixed model association mapping method, we identified three loci – *MORE AXILLARY GROWTH 2* (*MAX2*), *MORE AXILLARY GROWTH 3* (*MAX3*), and *SUPERSHOOT 1* (*SPS1*) – that were significantly associated with branching variation. On the basis of a more extensive examination of the *MAX2* and *MAX3* genomic regions, we find that linkage disequilibrium in these regions decays within ~10 kb and that trait associations localize to the candidate genes in these regions. When the significant associations are compared to relevant quantitative trait loci (QTLs) from previous L*er* x Col and Cvi x L*er* recombinant inbred line (RIL) mapping studies, no additive QTLs overlapping these candidate genes are observed, although epistatic QTLs for branching, including one that spans *SPS1*, are found. These results suggest that epistasis is prevalent in determining branching variation in *A. thaliana* and may need to be considered in linkage disequilibrium mapping studies of genetically diverse accessions.

Introduction

Evolutionary change in shoot architecture has played a central role in the morphological diversification of plant species, though relatively little is known about its molecular basis (SUSSEX and KERK 2001). Only a small number of genes have been implicated in shoot architectural evolution (BRADLEY *et al.* 1997; GALLAVOTTI *et al.* 2004; PURUGGANAN *et al.* 2000; PURUGGANAN and SUDDITH 1998; VOLLBRECHT *et al.* 2005; YOON and BAUM 2004), with the most comprehensively understood example coming from the maize *teosinte branched 1* (*tb1*) gene (DOEBLEY *et al.* 1997). In studies

of morphological evolution under domestication, it has been demonstrated that *tb1*, a TCP class transcription factor, was a target of selection for reduced tillering during the evolutionary origin of maize from its wild ancestor teosinte (CLARK *et al.* 2004; CLARK *et al.* 2006; WANG *et al.* 1999). A fuller understanding of the evolutionary basis of plant shoot variation can only be achieved through the continued identification of the molecular mechanisms responsible for the vast diversity in plant architectures.

One broad component of shoot architecture is branching pattern. Developmental regulation of branching occurs at several levels, including (i) node patterning, (ii) meristem determination, and (iii) axillary meristem elongation (MCSTEEN and LEYSER 2005). Several genes have been shown to affect node patterning in the model plant species *Arabidopsis. thaliana*, including *LATERAL SUPPRESSOR* (*LAS*) (GREB *et al.* 2003), *SHOOT MERISTEMLESS* (*STM*) (LONG *et al.* 1996), *REVOLUTA* (*REV*) (TALBERT *et al.* 1995), and the *REGULATORS OF AXILLARY MERISTEMS* (*RAX*) genes (KELLER *et al.* 2006; MULLER *et al.* 2006). The determination of inflorescence meristem identity is largely controlled by the floral identity genes *TERMINAL FLOWER1* (*TFL1*) (BRADLEY *et al.* 1997) and *LEAFY* (*LFY*) (WEIGEL *et al.* 1992). Branch elongation is regulated by numerous phytohormones, such as auxin, cytokinin, and abscicic acid (WARD and LEYSER 2004). Analyses of certain auxin signaling genes, such as *AUXIN RESISTANT1* (*AXR1*) (LEYSER *et al.* 1993; LINCOLN *et al.* 1990; STIRNBERG *et al.* 1999), as well of genes that appear to regulate auxin transport, such as the *MORE AXILLARY GROWTH* (*MAX*) genes (BENNETT *et al.* 2006; BOOKER *et al.* 2005; SOREFAN *et al.* 2003; STIRNBERG *et al.* 2002), have shown that this hormone plays a crucial role in coordinating branch outgrowth with the plant developmental program.

In *A. thaliana*, it has been shown that significant variation in branch number and other quantitative aspects of shoot architecture exists (UNGERER *et al.* 2002). Quantitative trait locus (QTL) mapping has historically been used as the main approach to mapping genes responsible for variation in ecologically and evolutionary significant

traits (LYNCH and WALSH 1998) and QTL mapping experiments have identified numerous loci that may be responsible for branching variation (e.g., (UNGERER *et al.* 2002; UNGERER *et al.* 2003)). Recently, however, association or linkage disequilibrium (LD) mapping has emerged as a serious alternative to identifying genes underlying quantitative phenotypes, and has been used with greater frequency in mapping traits in *A. thaliana* (MITCHELL-OLDS and SCHMITT 2006), maize (YU and BUCKLER 2006), Drosophila (MACKAY 2004), and humans (CARDON and ABECASIS 2003). Association studies in *A. thaliana* have covered a broad spectrum of genomic scales, ranging from candidate gene analysis (CAICEDO *et al.* 2004; HAGENBLAD and NORDBORG 2002; HAGENBLAD *et al.* 2004; OLSEN *et al.* 2004) to genomewide scans (ARANZANA *et al.* 2005; ZHAO *et al.* 2007).

Since rarely, if ever, is it possible pinpoint *a priori* which genes in a genetic pathway or network are likely to possess functional polymorphism(s), large-scale analyses of all genes in a trait's genetic network are a necessary next step in the candidate gene approach. The candidate gene approach is likely to be particularly useful in *A. thaliana,* since the relatively rapid decay of linkage disequilibrium (LD) in its genome within less than 25 kb suggests that trait associations will often span a few loci and possibly delimit functionally-significant polymorphism(s) to appropriate candidate loci (MITCHELL-OLDS and SCHMITT 2006). This mapping resolution will facilitate a more expedient characterization of the genetic basis of natural phenotypic variation in *A. thaliana*, although the promise of association mapping is dependent on our ability to differentiate significant results that are biologically informative from spurious associations. This continues to be a substantial challenge, as high false positive rates for association mapping have been found in *A. thaliana* (ARANZANA *et al.* 2005; ZHAO *et al.* 2007). Spurious associations typically occur when the demographic structure of a population is correlated with trait variation (PRITCHARD *et al.* 2000b), but this problem is more pronounced in Arabidopsis due to the selfing nature of this species and the geographic isolation of subpopulations with distance (ZHAO *et al.* 2007). Methods to

account for demographic structure in association mapping panels have been developed by incorporating estimates of population ancestry (PRICE *et al.* 2006; PRITCHARD *et al.* 2000b) and kinship (YU and BUCKLER 2006) into genotype-phenotype association tests. In general, these methods dramatically improve the power of association mapping to detect functional genetic variation, but do not totally resolve confounding demographic structure. Thus, results from association studies in this species must be cautiously evaluated and, ideally, replicated through other approaches, such as QTL mapping or transformation experiments.

We explore the genetics of branching variation in *A. thaliana* through association mapping focused on candidate genes involved in branch development. We initially sequenced ~600 bp fragments of 36 genes involved in shoot architectural development from a set of 24 geographically diverse *A. thaliana* accessions, and further genotyped common haplogroups (>10%) in an additional 96 geographically restricted accessions for association mapping. Several genes show nominally significant associations with branching, and we localize two of these associations to the gene level through a more extensive analysis of linked genomic regions. None of the observed associations are detected as additive QTLs within the L*er* × Col or Cvi × L*er* recombinant inbred lines (RILs), although reanalysis of the QTL mapping data shows numerous epistatic interactions underlying branching variation. These results provide an opportunity to compare candidate gene association studies with QTL mapping analyses, and suggest the possibility that epistatic interactions should be considered in association mapping investigations.

Materials and methods

Amplification and sequencing of the candidate genes and their genomic regions. Candidate genes included in this study are listed in Table 1. Amplification primers were designed in Primer3 (http://fokker.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) using

the Col-0 genome sequence. PCRs were conducted using standard reaction conditions and Perkin-Elmer thermalcyclers. Either Taq or Ex-Taq (TaKaRa, Otsu, Japan) polymerases were used for PCR amplification. PCR products were purified using either Qiagen PCR Purification Kits or Gel Extraction Kits (Qiagen, Hilden, Germany), or ExoSAP (Invitrogen, Carlsbad, CA). Sequencing reactions were performed with BigDye terminators (Applied Biosystems (ABI), Foster City, CA). Sequencing was conducted at the North Carolina State University Genome Research Lab or the New York University Genetic Analysis Center on ABI 3100, 3700, or 3730 capillary sequencers according to standard protocols.

Sequence manipulation and analysis. Sequence data was compiled into contigs using Phred/Phrap (CodonCode, Dedham, MA) and initially aligned in Biolign v2.0.9 (Tom Hall, Ibis Therapeutics, Carlsbad, CA). All polymorphisms were confirmed using the sequence trace files. Final sequence alignment was performed in Bioedit v7.0.5.2 (Tom Hall, Ibis Therapeutics, Carlsbad, CA).

Levels of polymorphism were determined based on the average number of pairwise differences between alleles (π (TAJIMA 1983)) or on the number of segregating sites (S) in the sequence sample (θ_w (WATTERSON 1975)) in DnaSP v4.0 (ROZAS *et al.* 2003). Nucleotide diversity values for nonsynonymous and synonymous sites (π_n and π_s, respectively) in coding regions, as well as Tajima's D (TAJIMA 1989), were also computed in DnaSP v4. Number of haplotypes (h) are also reported for each gene. Maximum parsimony gene genealogies were created in MEGA v3.1 (KUMAR *et al.* 2004).

LD in the *MAX2* and *MAX3* genomic regions was calculated as r^2 (HILL 1974). Polymorphisms with minor allele frequencies $\leq$ 10% were excluded from analysis. The median LD decay plot was created by grouping r^2 values into bins of 5 kb based on the distances between markers. The median r^2 value was taken from each bin and plotted

against bin midpoint. LD was plotted in TASSEL v1.9.4 (Ed Buckler lab, Cornell University, Ithaca, NY). Significance for LD was determined through 10,000 permutations.

Genotyping. Polymorphisms for genotyping were determined based on haplogroups present in the gene genealogies. For both the candidate genes and the genomic region sequencing, one polymorphism was genotyped from any branch that separated ≥ 10% of the alleles from all others using cleaved amplified polymorphic sequences (CAPS) or degenerate CAPS (dCAPS) markers. CAPS markers were chosen by comparison of the sequences of different alleles in NEBCutter v2.0 (http://tools.neb.com/NEBcutter2/index.php). dCAPS Finder v2.0 (http://helix.wustl.edu/dcaps/dcaps.html) was used to choose dCAPS markers. All primers for marker amplification were designed in Primer3. All restriction enzymes for digestion were purchased from New England Biolabs (Ipswich, MA). For *BP* and *SEU*, the alleles for these genes were sequenced from all 96 individuals, rather than genotyped with CAPS or dCAPS markers. Genotyped accessions and haplogroup assignments are provided in the Supplement.

Phenotype data for LD mapping. Data is from a controlled growth chamber experiment conducted at North Carolina State University (OLSEN *et al.* 2004). A lateral branch was defined as any elongated branch along the primary inflorescence. A basal branch was characterized as any branch emanating from the rosette. In large part, these basal branches extend from apical rosette nodes. Total branches were the sum of lateral branches and basal branches. Lateral branch nodes were considered any point along the primary inflorescence where a lateral branch could form and were counted as the number of cauline leaves. Least square (LS) means for trait values were used for the analyses in this paper. Trait values were standardized prior to running the mixed model to improve convergence.

Association tests. To decrease the possibility of spurious associations due to hidden population structure in our sample, we used a recently reported mixed model method that incorporates population ancestry estimates from the program STRUCTURE v2 (PRITCHARD *et al.* 2000a) and pairwise kinship estimates from the program SPAGeDi v1.2 (HARDY and VEKEMANS 2002). STRUCTURE runs with a prior of four ancestral populations ($K = 4$), resulted in the highest likelihood value of all K values. The ancestry estimates are based on previously described SNP data (SCHMID *et al.* 2006). Mixed model association tests were conducted in SAS v9.1.2 (SAS, Cary, NC) with a previously described program (YU and BUCKLER 2006).

Mapping epistatic QTLs. We analyzed previously published data for branching from the L*er* × Col and Cvi × L*er* RILs (UNGERER *et al.* 2002; UNGERER *et al.* 2003) in the program EPISTACY (http://www4.ncsu.edu/~jholland/Epistacy/epistacy.htm). A significance threshold of $P = 0.001$ was used. Marker combinations exhibiting epistatic interactions were evaluated for linkage to determine the span of the detected interactions. Epistatic QTL intervals were determined based on the physical positions of all linked markers that appear to represent the same epistatic interaction. In some cases, only a single marker pair was found to represent the epistatic interaction, in which case the physical spans of the epistatic QTLs could not be determined.

Results and discussion

Levels of variation and haplogroup structure of candidate branching genes. The 36 candidate genes used in this study were chosen to represent loci that control various developmental and physiological processes involved in shoot branching. One ~600 bp fragment comprised primarily of exons was sequenced from each candidate gene in a panel of 24 accessions from across the geographic range of the species. Nucleotide diversity levels at the 36 genes were lower than previously reported genomewide values for functional genes with $\pi = 0.002$ and $\theta_w = 0.003$ (NORDBORG *et al.* 2005; SCHMID *et al.*

2005). *PINOID* (*PID*), which has no polymorphisms, has the lowest nucleotide diversity of all genes, while *PINHEAD* (*PNH*) possesses the highest diversity (π and θ_w = 0.014) (Table 2). In coding regions, nucleotide diversity values for nonsynonymous polymorphisms (π = 0.001 and θ_w = 0.002) were nearly 3.5-fold lower than for synonymous polymorphisms (π = 0.004 and θ_w = 0.005), and numerous genes were not polymorphic at their nonsynonymous sites.

The average Tajima's D value observed in this sequence dataset is -1.023, the negative value arising from the prevalence of low frequency mutations in this dataset. The mean number of polymorphic sites observed per gene is 7.6, while the mean number of haplotypes per gene is 6.5, indicating that a large proportion of observed mutations are found in unique haplotypes. Most sequenced genes exhibit haplotype structure characterized by numerous low-frequency haplogroups that are differentiated from each other by a small number of mutations (Figure 1). Three genes – *AUXIN RESISTANT 2* (*AXR2*), *MORE AXILLARY GROWTH 2* (*MAX2*), and *PINHEAD* (*PNH*) – possess a relatively high number of moderate frequency mutations in comparison to other loci. Only 27 of the 36 genes (75%) possess haplogroups that are present at a frequency $\geq$ 10%.

We attempted to determine if genes involved in different developmental processes underlying shoot branching possess different levels of sequence variation. We categorized the 36 genes into three general groups based on their roles in branch development: (i) node patterning, (ii) meristem identity determining, and (iii) phytohormone signaling genes. Signaling genes have the highest levels of nucleotide diversity (π = 0.004), with the other two classes possessing similar levels of polymorphism (π = 0.001 for both classes). There are also lower nonsynonymous diversity levels for the meristem identity and patterning genes ($\pi_n \approx 0$ for both classes) than for the signaling genes (π_n = 0.002). Meristem identity genes also have a higher

proportion of loci with no nonsynonymous site variation (75%) than either node patterning (21%) or hormone signaling loci (28%).

Characteristics of the association mapping panel. Population structure in *A. thaliana* is extensive, with substantial differences in allele frequencies occurring across its geographic range (NORDBORG *et al.* 2005; SCHMID *et al.* 2005). This poses a serious challenge for association mapping since many traits in this species exhibit clines that are correlated with population structure (ARANZANA *et al.* 2005). To minimize the confounding effect of population structure on association mapping, we use 96 accessions from a geographically restricted range in Central Europe between latitudes 45°N and 55°N and longitudes 4°E and 18°E. Other studies have shown that population structure within this region is less severe than the global population structure of the species (KORVES *et al.* 2007; KORVES *et al.* in press; NORDBORG *et al.* 2005; SCHMID *et al.* 2005).

Previous association studies in our laboratory used a published AFLP dataset (SHARBEL *et al.* 2000) to correct for population structure (CAICEDO *et al.* 2004; OLSEN *et al.* 2004), but this dataset indicated minimal population stratification. More recent analyses based on SNP data have shown that, in fact, substantial population stratification is present in *A. thaliana* (NORDBORG *et al.* 2005; SCHMID *et al.* 2005), requiring the re-evaluation of our previous association results (KORVES *et al.* 2007). In this study, we utilized 115 genomewide SNPs from Schmid et al. (2005) using the program STRUCTURE (PRITCHARD *et al.* 2000a), which confirms that there is population stratification in the accessions used in this study, with the most likely number of populations (*K*) being four. The percentage membership to populations one through three is spatially correlated, with population one membership correlated to both latitude and longitude of origin of the accessions ($F_{2, 93} = 10.97$, $P < 0.001$), whereas population two and three correlated to longitude ($F_{1, 94} = 7.48$, $P = 0.008$) and latitude ($F_{1, 94} = 13.06$, $P <$

0.001), respectively. Thus, despite sampling at a smaller geographic scale, there is still detectable genealogical and geographic structure in this accession panel.

We determined the distribution of shoot branching across these 96 accessions for four branching traits under long and short day controlled growth chamber conditions: (i) lateral branch number, (ii) basal branch number, (iii) total branch number, and (iv) lateral branch node number. Substantial quantitative variation was found in these shoot branching traits (Figure 2), and broad sense heritabilities (H^2) range from .09 for basal branches in short day to .41 for lateral branches in long day, with higher H^2 values generally observed in long day than in short day (Table S1).

Branching traits exhibit geographic clines in our *A. thaliana* sample, with the majority of traits showing significant correlations with either the latitude or longitude of origin of the accessions, or both. Long day lateral branches (ANOVA, $F_{1, 94} = 7.38$, $P = 0.008$), short day lateral branch nodes (ANOVA, $F_{1, 94} = 6.94$, $P = 0.010$), and short day basal branches (ANOVA, $F_{1, 94} = 6.95$, $P = 0.010$) are all correlated with longitude, while long day lateral branch nodes are also associated with latitude (ANOVA, $F_{1, 94} = 5.56$, $P = 0.020$).

Associations between candidate gene polymorphisms and branch architecture. Of the 36 genes we initially analyzed, only 27 genes contained moderate frequency (≥10%) haplogroups that were subsequently genotyped in the association mapping panel of 96 accessions. We assessed the level of population structure in our mapping sample using these candidate gene haplogroups and found that unlike in the genomewide SNP set, the most likely number of putative ancestral populations (K) was found to be two. This suggests that our candidate genes exhibit lower levels of population stratification than genomewide markers in the set of 96 accessions.

Mixed model association tests were conducted based on haplogroup genotyping results, except for *AINTEGUMENTA* (*ANT*) for which no variation was found in our association mapping sample. The observed nominally significant associations ($P < 0.05$) per environment-trait combination ranges from one to four (Table 3), with seven genes (27%) exhibiting a nominally significant association in at least one environment-trait combination, while five genes (19%) are nominally significant in two or more environment-trait combinations. It has been shown that even with the use of the mixed model approach there is still an excess of false positive associations in Arabidopsis (ARANZANA *et al.* 2005; ZHAO *et al.* 2007). In an effort to be conservative in our evaluation of results, we recomputed association probabilities by ranking the nominal *P*-values for the 26 genes. We accepted an association as significant only if (i) the mixed model result was nominally significant ($P < 0.05$) AND (ii) the association *P*-value was at the lower 5% tail of the observed *P*-value distribution of all the candidate genes. In essence, this latter criterion allows us to use all the candidate gene associations as an empirical distribution for *P*-values obtained from the mixed model method (i.e., if ~ 20 genes are used, then one may be considered actually significant).

Based on both our criteria, one single significant association was found for each environment-trait combination, except for long day lateral branches, which had no nominally significant associations. Three genes – *MORE AXILLARY GROWTH 2* (*MAX2*), *MORE AXILLARY GROWTH 3* (*MAX3*), and *SUPERSHOOT1* (*SPS1*) – exhibited significant haplogroup-phenotype associations after the *P*-values were reassigned. The significant associations were: (i) *MAX2* and lateral and total branches in short day, (ii) *MAX3* and lateral branch nodes and total branches in long day, and (iii) *SPS1* and basal branches in long day and short day and lateral branch nodes in short day. The *SPS1* association with basal branching was the only gene-trait significant association across environmental conditions.

Patterns of LD and genotype-phenotype associations across two linked genomic regions. The patterns of polymorphism and linkage disequilibrium in the genome are the primary determinants of the resolution of association mapping (GAUT and LONG 2003). Though the extent of LD in *A. thaliana* has been characterized above 25-kb scales, we were interested in documenting it below this level of resolution across our significant genes. We sequenced ~600 bp fragments from three genes both upstream and downstream of *MAX2* and *MAX3*; these two genes were chosen because they exhibited multiple trait associations and are linked on chromosome 2, permitting the analysis of trait association patterns at both the fine (~ 35 kb) and coarse (~800 kb) genomic scales. These two regions display a 5-fold difference in polymorphism levels, with the *MAX2* region having a mean $\pi = 0.005$, while the mean π of the *MAX3* region is 0.001. The range of nucleotide diversity is also much higher in the *MAX2* ($\pi_{min} \approx 0$, $\pi_{max} = .011$) than in the *MAX3* region ($\pi_{min} \approx 0$, $\pi_{max} = .002$).

The patterns of LD are similar for both the *MAX2* and *MAX3* genomic regions, with SNP correlations primarily observed between adjacent genes. In the *MAX2* region, three haplotype blocks are distinguishable (Figure 3), with only slight LD ($r^2 \approx .5$) detectable between a pair of SNPs found in two blocks. As for the *MAX3* region, all LD with $r^2 > .4$ is found intragenically, except for LD that is detected between *MAX3* and its downstream neighbor At2g45000 (Figure 3). There is no detectable disequilibrium between the *MAX2* and *MAX3* regions. These results suggest that phenotypic associations detected at *MAX2* and *MAX3* should not span more than 10 kb, encompassing at most two to three genes (Figure 4)

To test whether our haplogroup associations at *MAX2* and *MAX3* extend to linked genes, we employed the mixed model association method on moderate-frequency haplogroups (≥10%) found across these two genomic regions. These results confirm that the trait associations span short genomic distances, localizing both the *MAX2* and *MAX3* associations to these genes and not to the linked loci (Figure 5).

Comparison of LD to QTL mapping. To replicate our candidate gene association results, we attempted to determine if our gene-trait associations correspond to previously observed additive QTLs in the L*er* × Col and Cvi × L*er* RIL populations. For lateral branches and lateral branch nodes combined, 5 QTLs were observed in both short day and long day in the L*er* × Col RILs, whereas 11 and 14 were observed in the Cvi × L*er* RILs in short day and long day, respectively (UNGERER *et al.* 2002; UNGERER *et al.* 2003). The loci exhibiting trait associations with lateral branches or lateral branch nodes were sequenced in the parental lines of each set of RILs. The L*er* × Col RILs have *MAX2* haplotypes that are members of significantly different haplogroups in the association mapping study (contrasts of the haplogroups containing the two alleles are significant for both lateral branches and total branches in short day, with $P = 0.030$ and $P = 0.047$, respectively), while the Cvi × L*er* RILs possesses relevant haplotypes for *SPS1* (contrast of the two relevant haplogroups for lateral branch nodes results in $P = 0.042$). However, for neither *MAX2* nor *SPS1* are overlapping additive QTLs present in the appropriate RIL mapping populations.

The absence of additive QTLs in the Ler x Col and Cvi x Ler mapping populations (UNGERER *et al.* 2002; UNGERER *et al.* 2003) encompassing either *SPS1* and *MAX2* was unexpected, given that we used a mixed model approach which previous reports suggested had a relatively low false positive rate (YU and BUCKLER 2006; ZHAO *et al.* 2007), a geographically-restricted mapping panel to further minimize population stratification, and a conservative re-computation of *P*-values by comparison of nominal probabilities across all tested candidate genes. There are two possibilities to account for this discrepancy. First, despite our efforts to minimize false positives, we are unable to completely remove all residual population structure that can still give rise to spurious associations. That association mapping is fundamentally confounded by the demographic structure of a population has been well documented (ZHAO *et al.* 2007), and although we

have been methodologically conservative so as to mitigate population structure effects, cryptic population structure may continue to be present that leads to spurious associations

The second possibility is that these genes indeed underlie natural variation in shoot branching, but in primarily epistatic interactions rather than direct additive effects. To assess whether epistatic QTLs involving the *MAX2* and *SPS1* genomic regions are present, we determined the number and locations of epistatic interactions involved in branching in the L*er* × Col and Cvi × L*er* RILs. Seven and nine branching epistatic interactions were found in the L*er* × Col RILs in short day and long day, respectively, while six and seven were found in the Cvi × L*er* RILs in the same growth conditions (Tables 4 and 5). Epistatic interactions were indeed found to overlap both genes in the appropriate RILs (Figure 6)**.** Epistatic QTLs that overlap *MAX2* are observed on the L*er* × Col mapping population, but are for a different branching trait-environment combination than was observed by the candidate gene association mapping study. For *SPS1*, however, an overlapping epistatic QTL was detected in the Cvi x L*er* lines that replicated the association mapping results.

It is thus possible that what we are observing is the effect of population stratification in our association studies, but in this context the population structure works to maintain epistatic interactions in a large, genetically-diverse sample. Indeed, it is well-known that epistatic effects can manifest themselves as additive genetic variation, particularly in species that are either highly inbred and/or highly structured at the population level (as reviewed in (NEIMAN and LINKSVAYER 2006)). If association mapping populations contain either inbred genotypes and/or individuals from different subpopulations, alleles involved in epistatic interactions may prove detectable through conventional association mapping techniques, but undetectable as additive QTLs through QTL mapping. The results presented here support those from recent studies that point to the prevalence of epistasis on the genetic architecture of various traits in *A. thaliana* (MALMBERG *et al.* 2005).

Candidate gene association mapping of branch variation in Arabidopsis. Shoot architecture, in particular the organization of axillary branches, is one of the most visible features that differentiate plant species (SUSSEX and KERK 2001). Characterizing the molecular basis of microevolutionary variation in branching in tractable plant models, such as *A. thaliana*, may provide clues to the molecular mechanisms underlying shoot macroevolution. This study is the first report of a large-scale candidate gene association mapping in *A. thaliana*, which evaluates many of the genes found in the genetic network that underlies shoot branching in this plant species. By screening 36 genes, we have identified at least one strong candidate gene for branching variation in this species in *SPS1* and two weaker candidates in *MAX2* and *MAX3*. Additionally, we have localized the *MAX2* and *MAX3* associations to the genes themselves.

The associations described in this study provide evidence that variation in phytohormone signaling pathways for auxin and cytokinin may play important roles in generating branching diversity in *A. thaliana*. Both auxin and cytokinin have long been known to play central roles in apical dominance and shoot branch development, so it is plausible that these signals could contribute to branching variation. This study also emphasizes the strong influence the environment exerts over quantitative variation in the shoot and its genetic architecture. The nominally significant candidate gene associations detected in each environment were typically very different, and these differences are comparable to those observed in QTL mapping experiments that have shown a large number of environment-specific QTLs for shoot architectural traits (e.g. (UNGERER *et al.* 2003)). This environmental sensitivity provides support for the importance of genotype-environment interactions in shaping shoot morphology, and it is likely that such environment-specific control of branching exists and may be of adaptive value as plants modulate their architecture to various ecological signals such as photoperiod length, available nutrients, and herbivory (BONSER and AARSSEN 1996).

Our results demonstrate the utility and difficulty of association mapping, which is increasingly being applied to a large number of trait mapping studies. This study shows that in *A. thaliana* one can conceivably obtain high resolution and localize haplogroup-phenotype associations to single genes with this approach. Detailed studies are necessary in order to validate the associations reported in this study at the causal level, and these studies are currently underway. Our results suggest, however, that candidate gene association studies can provide strong candidate quantitative trait genes, though certainly this approach is not as comprehensive as genomewide analyses that are not limited to characterized genes. Nevertheless, the challenge is to determine the biological significance of associations detected by association mapping, and the possibility that epistatic interactions may underlie the effects of many of these genes poses challenges on how association studies are replicated and validated in the future.

Acknowledgments

We are grateful to past and present members of the Purugganan lab for assistance performing experiments and analyses in this paper, as well as for thoughtful discussion regarding this manuscript. Also, we thank Tonia Korves, Ottoline Leyser, and Magnus Nordborg for comments on a previous version of this manuscript. This work was supported by a Department of Education Graduate Assistance in Areas of National Need Fellowship and a National Science Foundation (NSF) Graduate Research Fellowship to I. M. E., and by grants from the NSF's Frontiers in Integrated Biological Research and Plant Genome Research Programs to M. D. P.

References

ARANZANA, M. J., S. KIM, K. ZHAO, E. BAKKER, M. HORTON *et al.*, 2005 Genomewide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1: e60.

BENNETT, T., T. SIEBERER, B. WILLETT, J. BOOKER, C. LUSCHNIG *et al.*, 2006 The Arabidopsis MAX pathway controls shoot branching by regulating auxin transport. Curr Biol 16: 553-563.

BONSER, S., and L. AARSSEN, 1996 Meristem allocation: A new classification theory for adaptive strategies in herbaceous plants. Oikos 77: 347-352.

BOOKER, J., T. SIEBERER, W. WRIGHT, L. WILLIAMSON, B. WILLETT *et al.*, 2005 MAX1 encodes a cytochrome P450 family member that acts downstream of MAX3/4 to produce a carotenoid-derived branch-inhibiting hormone. Dev Cell 8: 443-449.

BRADLEY, D., O. RATCLIFFE, C. VINCENT, R. CARPENTER and E. COEN, 1997 Inflorescence commitment and architecture in Arabidopsis. Science 275: 80-83.

CAICEDO, A. L., J. R. STINCHCOMBE, K. M. OLSEN, J. SCHMITT and M. D. PURUGGANAN, 2004 Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci U S A 101: 15670-15675.

CARDON, L. R., and G. R. ABECASIS, 2003 Using haplotype blocks to map human complex trait loci. Trends Genet 19: 135-140.

CLARK, R. M., E. LINTON, J. MESSING and J. F. DOEBLEY, 2004 Pattern of diversity in the genomic region near the maize domestication gene tb1. Proc Natl Acad Sci U S A 101: 700-707.

CLARK, R. M., T. N. WAGLER, P. QUIJADA and J. DOEBLEY, 2006 A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet 38: 594-597.

DOEBLEY, J., A. STEC and L. HUBBARD, 1997 The evolution of apical dominance in maize. Nature 386: 485-488.

GALLAVOTTI, A., Q. ZHAO, J. KYOZUKA, R. B. MEELEY, M. K. RITTER *et al.*, 2004 The role of barren stalk1 in the architecture of maize. Nature 432: 630-635.

GAUT, B. S., and A. D. LONG, 2003 The lowdown on linkage disequilibrium. Plant Cell 15: 1502-1506.

GREB, T., O. CLARENZ, E. SCHAFER, D. MULLER, R. HERRERO *et al.*, 2003 Molecular analysis of the LATERAL SUPPRESSOR gene in Arabidopsis reveals a conserved control mechanism for axillary meristem formation. Genes Dev 17: 1175-1187.

HAGENBLAD, J., and M. NORDBORG, 2002 Sequence variation and haplotype structure surrounding the flowering time locus FRI in Arabidopsis thaliana. Genetics 161: 289-298.

HAGENBLAD, J., C. TANG, J. MOLITOR, J. WERNER, K. ZHAO *et al.*, 2004 Haplotype structure and phenotypic associations in the chromosomal regions surrounding two Arabidopsis thaliana flowering time loci. Genetics 168: 1627-1638.

HARDY, O., and X. VEKEMANS, 2002 SPAGEDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2: 618-620.

HILL, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Heredity 33: 229-239.

KELLER, T., J. ABBOTT, T. MORITZ and P. DOERNER, 2006 Arabidopsis REGULATOR OF AXILLARY MERISTEMS1 controls a leaf axil stem cell niche and modulates vegetative development. Plant Cell 18: 598-611.

KORVES, T., K. SCHMID, A. CAICEDO, C. MAYS, J. STINCHCOMBE *et al.*, 2007 Fitness effects associated with the major flowering time gene *FRIGIDA* in Arabidopsis thaliana in the field. Am Nat 169: epub.

KORVES, T., K. SCHMID, A. CAICEDO, C. MAYS, J. STINCHCOMBE *et al.*, in press Fitness effects associated with the major flowering time gene *FRIGIDA* in Arabidopsis thaliana in the field. Am Nat.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5: 150-163.

LEYSER, H. M., C. A. LINCOLN, C. TIMPTE, D. LAMMER, J. TURNER *et al.*, 1993 Arabidopsis auxin-resistance gene AXR1 encodes a protein related to ubiquitin-activating enzyme E1. Nature 364: 161-164.

LINCOLN, C., J. H. BRITTON and M. ESTELLE, 1990 Growth and development of the axr1 mutants of Arabidopsis. Plant Cell 2: 1071-1080.

LONG, J. A., E. I. MOAN, J. I. MEDFORD and M. K. BARTON, 1996 A member of the KNOTTED class of homeodomain proteins encoded by the STM gene of Arabidopsis. Nature 379: 66-69.

MACKAY, T. F., 2004 The genetic architecture of quantitative traits: lessons from Drosophila. Curr Opin Genet Dev 14: 253-257.

MALMBERG, R. L., S. HELD, A. WAITS and R. MAURICIO, 2005 Epistasis for fitness-related quantitative traits in Arabidopsis thaliana grown in the field and in the greenhouse. Genetics 171: 2013-2027.

MCSTEEN, P., and O. LEYSER, 2005 Shoot branching. Annu Rev Plant Biol 56: 353-374.

MITCHELL-OLDS, T., and J. SCHMITT, 2006 Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. Nature 441: 947-952.

MULLER, D., G. SCHMITZ and K. THERES, 2006 Blind homologous R2R3 Myb genes control the pattern of lateral meristem initiation in Arabidopsis. Plant Cell 18: 586-597.

NEIMAN, M., and T. A. LINKSVAYER, 2006 The conversion of variance and the evolutionary potential of restricted recombination. Heredity 96: 111-121.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3: e196.

OLSEN, K. M., S. S. HALLDORSDOTTIR, J. R. STINCHCOMBE, C. WEINIG, J. SCHMITT *et al.*, 2004 Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics 167: 1361-1369.

PRICE, A. L., N. J. PATTERSON, R. M. PLENGE, M. E. WEINBLATT, N. A. SHADICK *et al.*, 2006 Principal components analysis corrects for stratification in genomewide association studies. Nat Genet 38: 904-909.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000a Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000b Association mapping in structured populations. Am J Hum Genet 67: 170-181.

PURUGGANAN, M. D., A. L. BOYLES and J. I. SUDDITH, 2000 Variation and selection at the CAULIFLOWER floral homeotic gene accompanying the evolution of domesticated Brassica oleracea. Genetics 155: 855-862.

PURUGGANAN, M. D., and J. I. SUDDITH, 1998 Molecular population genetics of the Arabidopsis CAULIFLOWER regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. Proc Natl Acad Sci U S A 95: 8130-8134.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in Arabidopsis thaliana reveals a genomewide departure from a neutral model of DNA sequence polymorphism. Genetics 169: 1601-1615.

SCHMID, K. J., O. TORJEK, R. MEYER, H. SCHMUTHS, M. H. HOFFMANN *et al.*, 2006 Evidence for a large-scale population structure of Arabidopsis thaliana from genomewide single nucleotide polymorphism markers. Theor Appl Genet 112: 1104-1114.

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.

SOREFAN, K., J. BOOKER, K. HAUROGNE, M. GOUSSOT, K. BAINBRIDGE *et al.*, 2003 MAX4 and RMS1 are orthologous dioxygenase-like genes that regulate shoot branching in Arabidopsis and pea. Genes Dev 17: 1469-1474.

STIRNBERG, P., S. P. CHATFIELD and H. M. LEYSER, 1999 AXR1 acts after lateral bud formation to inhibit lateral bud growth in Arabidopsis. Plant Physiol 121: 839-847.

STIRNBERG, P., K. VAN DE SANDE and H. M. LEYSER, 2002 MAX1 and MAX2 control shoot lateral branching in Arabidopsis. Development 129: 1131-1141.

SUSSEX, I. M., and N. M. KERK, 2001 The evolution of plant architecture. Curr Opin Plant Biol 4: 33-37.

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.

TALBERT, P. B., H. T. ADLER, D. W. PARKS and L. COMAI, 1995 The REVOLUTA gene is necessary for apical meristem development and for limiting cell divisions in the leaves and stems of Arabidopsis thaliana. Development 121: 2723-2735.

UNGERER, M. C., S. S. HALLDORSDOTTIR, J. L. MODLISZEWSKI, T. F. 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. MACKAY, 2003 Genotype-environment interactions at quantitative trait loci affecting inflorescence development in Arabidopsis thaliana. Genetics 165: 353-365.

VOLLBRECHT, E., P. S. SPRINGER, L. GOH, E. S. T. BUCKLER and R. MARTIENSSEN, 2005 Architecture of floral branch systems in maize and related grasses. Nature 436: 1119-1126.

WANG, R. L., A. STEC, J. HEY, L. LUKENS and J. DOEBLEY, 1999 The limits of selection during maize domestication. Nature 398: 236-239.

WARD, S. P., and O. LEYSER, 2004 Shoot branching. Curr Opin Plant Biol 7: 73-78.

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

WEIGEL, D., J. ALVAREZ, D. R. SMYTH, M. F. YANOFSKY and E. M. MEYEROWITZ, 1992 LEAFY controls floral meristem identity in Arabidopsis. Cell 69: 843-859.

YOON, H. S., and D. A. BAUM, 2004 Transgenic study of parallelism in plant morphological evolution. Proc Natl Acad Sci U S A 101: 6524-6529.

YU, J., and E. S. BUCKLER, 2006 Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17: 155-160.

ZHAO, K., M. J. ARANZANA, S. KIM, C. LISTER, C. SHINDO *et al.*, 2007 An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet 3: e4.

Table 1. Genes included in this study

Gene	Abbreviation	Gene ID	Annotation
ABA INSENSITIVE 3	*ABI3*	At3g24650	Transciption factor
ALTERED MERISTEM PROGRAM 1	*AMP1*	At3g54720	Glutamate carboxypeptidase
AINTEGUMENTA	*ANT*	At4g37750	Transcription factor
APETALA 1	*AP1*	At1g69120	MADS transcription factor
AUXIN-REGULATED GENE INVOLVED IN ORGAN SIZE	*ARGOS*	At3g59900	Auxin-inducible gene that controls lateral organ size
AUXIN RESISTANT 1	*AXR1*	At1g05180	Ubiquitin activating enzyme E1-related protein
AUXIN RESISTANT 2	*AXR2*	At3g23050	Auxin-responsive protein / indoleacetic acid-induced protein 7 (IAA7)
AUXIN RESISTANT 3	*AXR3*	At1g04250	IAA17
AUXIN RESISTANT 6	*AXR6*	At4g02570	Cullin family protein
BREVIPEDICELLUS	*BP*	At4g08150	Homeobox protein knotted-1 like 1
BUSHY AND DWARF 1	*BUD1*	At1g18350	MAP kinase kinase7
CAULIFLOWER	*CAL*	At1g26310	MADS transcription factor
CYTOKININ-INDEPENDENT 1	*CKI1*	At2g47430	Cytokinin-responsive histidine kinase
CYTOKININ RESPONSE 1	*CRE1*	At2g01830	Histidine kinase AHK4
EMBRYONIC FLOWER 1	*EMF1*	At5g11530	Regulates reproductive development
ERECTA	*ER*	At2g26330	Receptor protein kinase
ENHANCED RESPONSE TO ABA 1	*ERA1*	At5g40280	Beta subunit of farnesyl-trans-transferase,
LATERAL SUPPRESSOR	*LAS*	At1g55580	GRAS transcription factor
LEAFY	*LFY*	At5g61850	Transcription factor
MORE AXILLARY GROWTH 1	*MAX1*	At2g26170	Cytochrome P450 (CYP711A)
MORE AXILLARY GROWTH 2	*MAX2*	At2g42620	F-box protein
MORE AXILLARY GROWTH 3	*MAX3*	At2g44990	Carotenoid cleavage dioxygenase 7
MORE AXILLARY GROWTH 4	*MAX4*	At4g32810	Carotenoid cleavage dioxygenase 8
MONOPTEROS	*MP*	At1g19850	IAA24, auxin response factor 5 (ARF5)
PINOID	*PID*	At2g34650	Serine/threonine kinase
PINFORMED 1	*PIN1*	At1g73590	Auxin efflux carrier protein
PINHEAD	*PNH*	At5g43810	Translation initiation factor
REGULATOR OF AXILLARY MERISTEMS 1	*RAX1*	At5g23000	Myb transcription factor 37
REGULATOR OF AXILLARY MERISTEMS 2	*RAX2*	At2g36890	Myb transcription factor 38
REGULATOR OF AXILLARY MERISTEMS 3	*RAX3*	At3g49690	Myb transcription factor 84
REVOLUTA	*REV*	At5g60690	Homeodomain-leucine zipper protein
SEUSS	*SEU*	At1g43850	Transcriptional co-regulator of AGAMOUS
SUPERSHOOT 1	*SPS1*	At1g16410	Cytochrome P450 (CYP79F1)
SHOOT MERISTEMLESS	*STM*	At1g62360	Knotted-like homeodomain protein
TERMINAL FLOWER 1	*TFL1*	At5g03840	Controls inflorescence meristem identity
TRANSPORT INHIBITOR RESPONSE 1	*TIR1*	At3g62980	F-box protein

Table 2. Genetic variation at the sequenced genes

Gene	Class[a]	n	Length	*S*	*h*	π	θ_w	π_n	π_s	Tajima's D
ABI3	S	23	766	7	7	0.0010	0.0025	0.0007	0.0015	-1.883
AMP1	P	24	765	7	9	0.0020	0.0025	0.0014	0.0043	-.606
ANT	P	21	805	4	4	0.0007	0.0014	0.0004	0.0019	-1.458
AP1	D	25	518	12	8	0.0029	0.0062	0	0	-1.83
ARGOS	P	17	551	4	5	0.0022	0.0022	0.0004	0	.135
AXR1	S	25	842	7	8	0.0013	0.0022	0.0021	0	-1.335
AXR2	S	23	834	24	11	0.0082	0.0080	0.0034	0.0186	.134
AXR3	S	23	717	4	5	0.0010	0.0015	0.0003	0.0049	-.931
AXR6	S	24	822	5	6	0.0010	0.0016	0.0005	0	-1.188
BP	P	22	413	10	6	0.0052	0.0068	0.0010	0.0115	-.806
BUD1	S	23	487	4	7	0.0031	0.0022	0.0027	0.0046	-1.13
CAL	D	25	520	7	8	0.0024	0.0036	0.0005	0.0020	-1.035
CKI1	S	24	794	2	3	0.0002	0.0007	0	0.0009	-1.515
CRE1	S	25	788	1	2	0.0001	0.0003	0.0001	0	-1.158
EMF1	P	17	487	2	3	0.0005	0.0012	0	0.0022	-1.504
ER	S	24	520	7	6	0.0019	0.0036	0.0003	0.0085	-1.464
ERA1	S	25	798	10	9	0.0016	0.0033	0.0013	0.0042	-1.682
LAS	P	25	806	2	3	0.0002	0.0007	0	0.0008	-1.514
LFY	D	25	544	3	4	0.0006	0.0015	0	0.0028	-1.504
MAX1	S	25	687	6	6	0.0019	0.0023	0.0008	0.0086	-.519
MAX2	S	25	760	16	11	0.0062	0.0006	0.0055	0.0084	.38
MAX3	S	25	660	4	5	0.0015	0.0016	0.0009	0.0035	-.151
MAX4	S	25	690	6	5	0.0019	0.0024	0.0036	0	-.632
MP	P	25	808	4	5	0.0008	0.0013	0.0003	0.0029	-1.019
PID	S	25	796	0	1	0	0	0	0	N/A
PIN1	S	21	812	5	5	0.0007	0.0017	0	0	-1.795
PNH	P	22	831	29	12	0.0093	0.0101	0.0022	0.0303	-.247
RAX1	P	25	568	9	8	0.0022	0.0042	0.0014	0.0050	-1.584
RAX2	P	25	437	5	4	0.0028	0.0030	0.0027	0.0032	-.269
RAX3	P	25	556	7	8	0.0018	0.0033	0.0008	0.0057	-1.469
REV	P	21	720	9	9	0.0025	0.0035	0.0021	0.0073	-1.003
SEU	P	24	535	9	9	0.0040	0.0050	0.0043	0.0030	-.622
SPS1	S	24	700	11	14	0.0028	0.0042	0.0032	0.0005	-1.125
STM	P	25	644	11	6	0.0024	0.0047	0	0.0030	-1.665
TFL1	D	24	566	2	3	0.0004	0.0010	0	0	-1.202
TIR1	S	25	814	2	3	0.0004	0.0007	0	0	-.941

[a] Functional groupings used for population genetic analyses: D = determination of meristem identity, P = shoot patterning, and S = phytohormone signaling

Table 3. Nominal *P*-values for mixed model association tests

	Long Day				Short Day			
Gene	Lateral Branches	Basal Branches	Total Branches	Lateral Branch Nodes	Lateral Branches	Basal Branches	Total Branches	Lateral Branch Nodes
AMP	0.664	0.138	0.398	0.789	0.915	0.972	0.896	0.481
AP1	0.793	0.227	0.255	0.226	0.200	0.370	0.146	0.705
ARGOS	0.225	0.0.8297	0.687	0.797	0.478	0.592	0.440	0.364
AXR1	0.898	0.580	0.942	0.531	0.276	0.696	0.159	0.465
AXR2	0.211	0.445	0.644	0.668	0.073	0.031	0.236	0.060
AXR3	0.151	0.493	0.378	0.193	0.733	0.359	0.901	0.124
AXR6	0.791	0.786	0.922	0.953	0.416	0.646	0.527	0.069
BP	0.500	0.776	0.586	0.739	0.736	0.129	0.889	0.717
BUD1	0.646	0.057	0.101	0.156	0.454	0.595	0.384	0.995
CAL	0.318	0.566	0.616	0.213	0.862	0.861	0.899	0.775
ER	0.935	0.582	0.839	0.455	0.459	0.051	0.194	0.092
ERA1	0.359	0.524	0.134	0.063	0.024	0.427	0.013	0.035
MAX1	0.503	0.251	0.737	0.863	0.513	0.124	0.460	0.138
MAX2	0.234	0.659	0.335	0.384	<0.001*	0.736	<0.001*	0.113
MAX3	0.058	0.302	0.008*	0.037*	0.322	0.053	0.286	0.135
MAX4	0.510	0.942	0.471	0.706	0.713	0.325	0.775	0.385
MP	0.490	0.879	0.375	0.212	0.267	0.327	0.358	0.820
PNH	0.964	0.299	0.947	0.966	0.767	0.518	0.632	0.788
RAX1	0.133	0.186	0.898	0.667	0.299	0.235	0.192	0.487
RAX2	0.989	0.514	0.504	0.648	0.473	0.101	0.241	0.966
RAX3	0.089	0.453	0.357	0.219	0.050	0.117	0.183	0.104
REV	0.346	0.591	0.108	0.248	0.605	0.307	0.401	0.352
SEU	0.735	0.612	0.588	0.437	0.018	0.616	0.026	0.195
SPS1	0.373	0.0158*	0.109	0.645	0.207	0.003*	0.244	0.014*
STM	0.303	0.846	0.497	0.705	0.562	0.542	0.729	0.072
TIR1	0.287	0.359	0.783	0.154	0.612	0.976	0.568	0.082

* $P < .05$ based on ranking of nominal *P*-values within environment-trait combination

Table 4. Epistatic QTLs detected in the L*er* × Col RILs

Environment	Trait	ID[a]	Marker 1	Position	Range	Marker 2	Position	Range
Long Day	Lateral Branches	E1	CATTS039	I - 51.22	45.20 - 60.78 (MI62 - BH.160L)	MI421	II - 12.27	2.24 - 25.45 (MI320 - THY_1)
		E2	MI390	IV - 9.75	6.00 - 36.73 (G3843 - PCITD23)	EMB514	V - 110.74	104.35 - 110.74 (MI70 - EMB514)
		E3	M336	II - 65.98	57.80 - 71.36 (VE017 - MI79A)	M194	V - 81.26	N/A
		E4	G17311	I - 131.83	N/A	RRS2	II - 79.29	N/A
		E5	AGP64	I - 128	126.76 – 128 (VE011 - AGP64)	MI138	V - 33.14	N/A
		E6	VE012	II - 0	N/A	G4014	III - 68.86	N/A
		E7	PCITF3	IV - 31.83	N/A	MI322	V - 27.73	N/A
		E8	M315	II - 3.73	N/A	MI473	II: 69.76	69.76 - 71.36 (MI473 - MI79A)
	Lateral Branch Nodes	E1	ATHFUS6	III - 90.25	90.25 - 93.62 (ATHFUS6 – NGA6)	MI174	V: 23.96	0 - 27.73 (PATT80 - MI322)
Short Day	Lateral Branches	E1	NGA8	IV - 25.31	13.06 - 33.64 (APP - M518)	H2A1	V - 102.00	68.69 - 110.74 (M435 - EMB514)
		E2	CDS7	I - 51.22	14.65 - 51.22 (ATTS0477 – CATTS039)	MI306	IV - 22.11	22.11 - 50.88 (MI306 - MI112)
		E3	MI103	I - 115.41	N/A	AG	IV - 55.18	N/A
		E4	MI208	I - 79.83	N/A	CA1	III - 0	N/A
	Lateral Branch Nodes	E1	M448A	IV - 21.59	2.75 - 36.73 (MI204 - PCITD23)	H2A1	V - 102.00	75.36 - 129.05 (G4028 - CATHHAN)
		E2	MI139	II - 28.11	28.11 - 29.38 (MI139 - MI148)	M448A	IV -21.59	N/A
		E3	AGP64	I -128.00	N/A	G4028	V - 75.36	N/A

[a] IDs correspond to epistatic QTLs in Figure 6

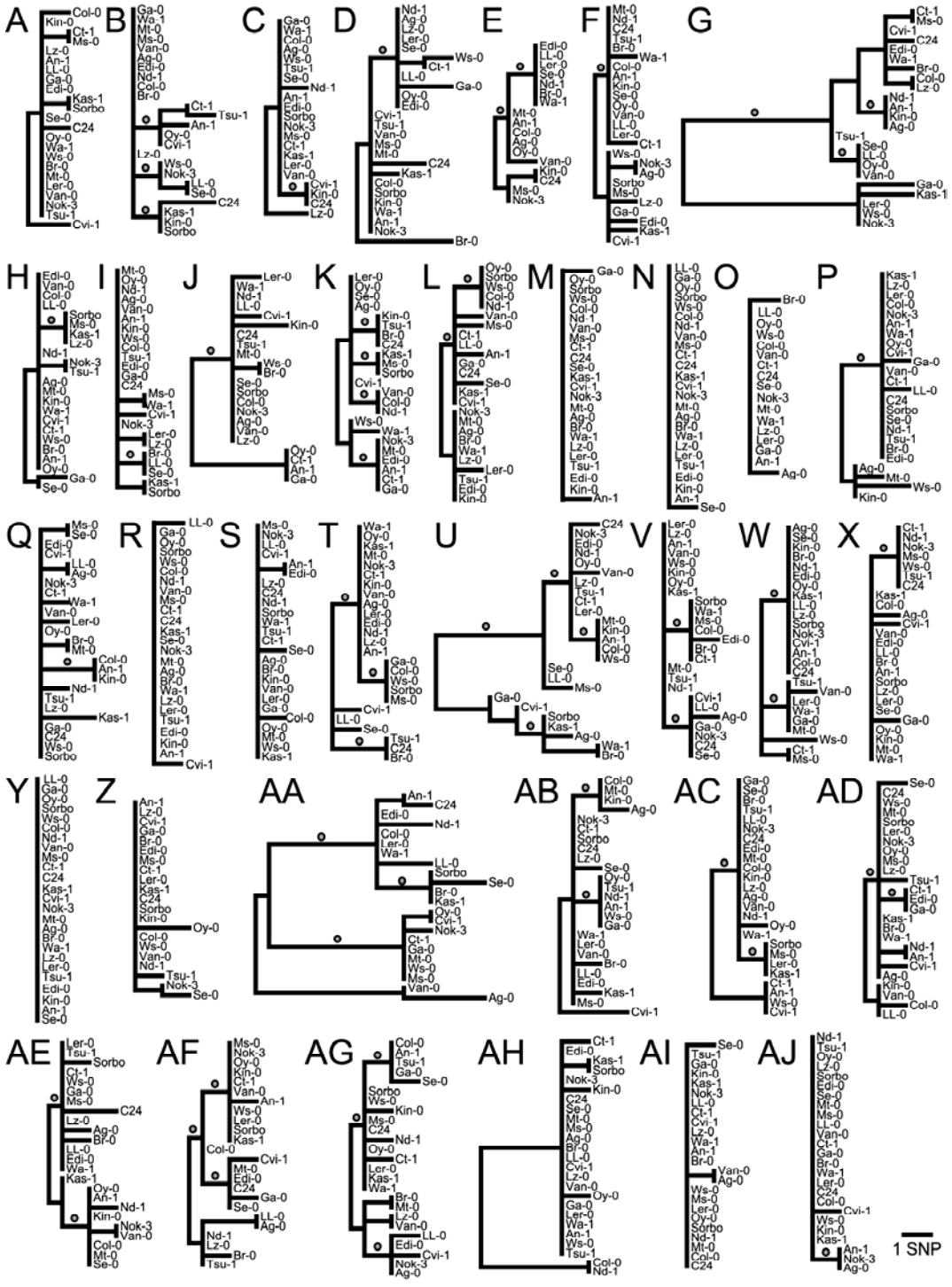
Table 5. Epistatic QTLs detected in the Cvi × L*er* RILs

Environment	Trait	ID[a]	Marker 1	Position	Range	Marker 2	Position	Range
Long Day	Lateral Branches	E1	AD.112L	III - 77.07	75.45 – 77.07 (AD.495L - AD.112L)	DF.231C	V - 26.53	20.19 - 31.80 (BH.107L - DF.184L)
		E2	EC.480C	I - 15.49	N/A	EG.75L	III - 7.25	N/A
	Lateral Branch Nodes	E1	FD.167L	IV - 51.74	37.14 - 64.34 (CD.84C - GB.490C)	DF.231C	V - 26.53	15.39 - 49.32 (BH.325L - HH.480C)
		E2	CH.200C	I - 76.01	76.01 - 79.20 (CH.200C - DF.260L)	HH.158L	III - 26.55	22.45 - 26.89 (GH.390L - EC.83C)
		E3	AXR1	I - 7.70	N/A	HH.90L	III - 80.81	N/A
		E4	GB.500C	I - 65.22	N/A	BH.92L	IV - 29.36	N/A
Short Day	Lateral Branches	E1	EG.129C	I - 57.56	49.42 - 60.96 (GB.112L - BH.162C)	AD.92L	III - 32.31	29.46 - 32.31 (GD.318C - AD.92L)
		E2	PVV4	I - 0	0 - 7.70 (PVV4 - AXR1)	HH.480C	V - 49.32	42.14 - 49.32 (GH.121L - HH.480C)
		E3	PVV4	I - 0	N/A	CC.318C	I - 108.57	N/A
	Lateral Branch Nodes	E1	AD.121C	I - 39.46	39.46 - 40.44 (AD.121C - AD.106L)	FD.345C	V - 93.14	90.49 - 109.90 (CC.262C - GD.222C)
		E2	HH.159C	IV - 60.64	56.94 - 64.34 (CH.70L - GB.490C)	DF.231C	V - 26.53	24.55 - 42.14 (AD.114C - 42.14)
		E3	CH.200C	I - 76.01	N/A	BH.96L	V - 55.96	N/A
		E4	DF.73L	I - 29.04	N/A	EC.83C	III - 26.89	N/A

[a] IDs correspond to epistatic QTLs in Figure 6

Figure 1. Maximum parsimony genealogies of the 36 candidate genes

The networks correspond to A) *ABI3*, B) *AMP1*, C) *ANT*, D) *AP1*, E) *ARGOS*, F) *AXR1*, G) *AXR2*, H) *AXR3*, I) *AXR6*, J) *BP*, K) *BUD1*, L) *CAL*, M) *CKI1*, N) *CRE1*, O) *EMF1*, P) *ER*, Q) *ERA1*, R) *LAS*, S) *LFY*, T) *MAX1*, U) *MAX2*, V) *MAX3*, W) *MAX4*, X) *MP*, Y) *PID*, Z) *PIN1*, AA) *PNH*, AB) *RAX1*, AC) *RAX2*, AD) *RAX3*, AE) *REV*, AF) *SEU*, AG) *SPS1*, AH) *STM*, AI) *TFL1*, and AJ) *TIR1*. The grey circles represent branches along which a SNP was genotyped. All genealogies are presented at the same scale.


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
AA
AB
AC
AD
AE
AF
AG
AH
AI
AJ
1 SNP

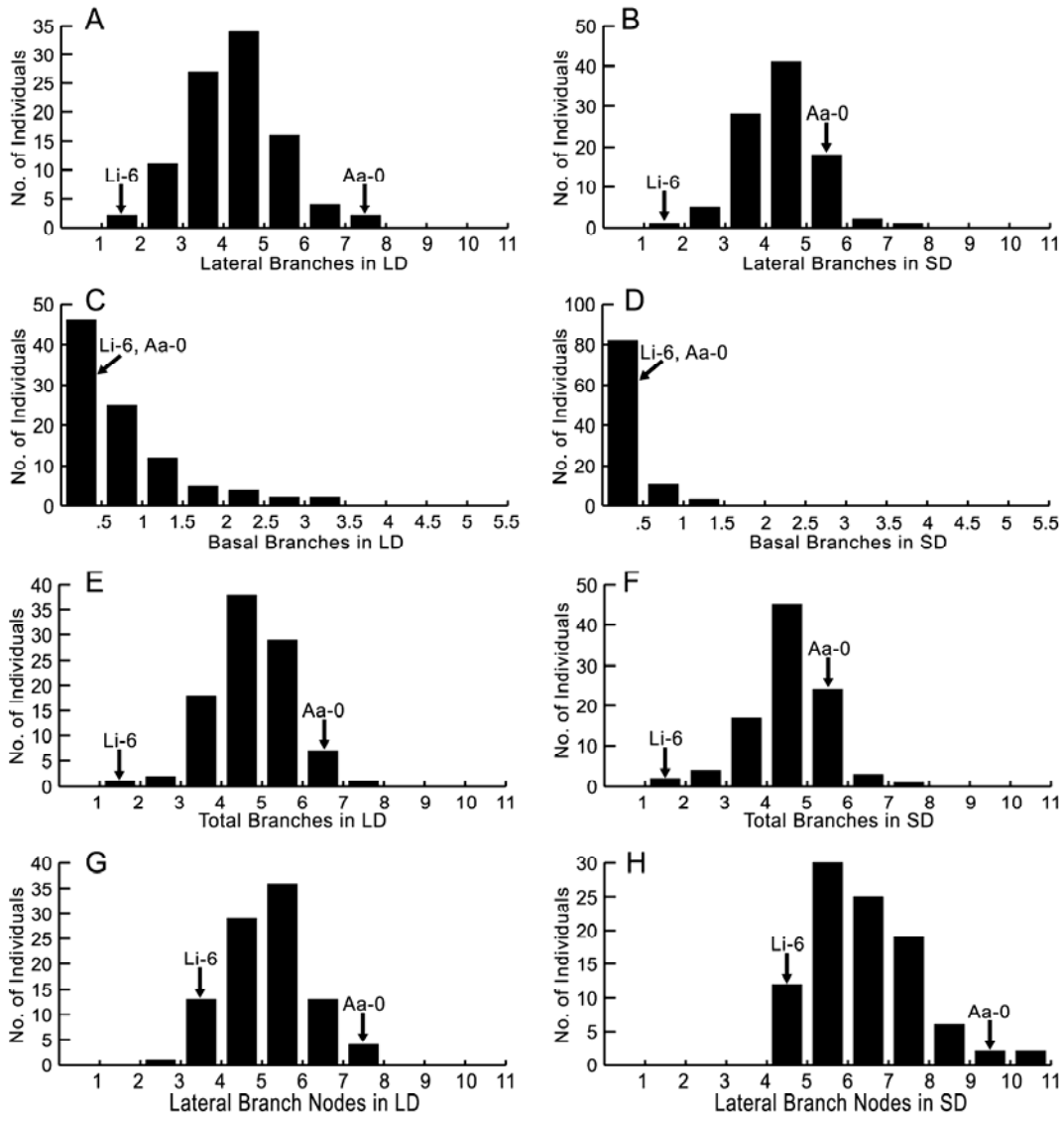


Figure 2. Trait distributions for the 96 accessions used for association mapping

Aa-0 and Li-6, which have the highest and lowest number of lateral branches in long day, respectively, are shown as reference accessions across all environment-trait combinations. Panels are A) lateral branches in long day, B) lateral branches in short day, C) basal branches in long day, D) basal branches in short day, E) total branches in long days, F) total branches in short day, G) lateral branch nodes in long day, and H) lateral branch nodes in short day. Long day and short day are abbreviated LD and SD, respectively.

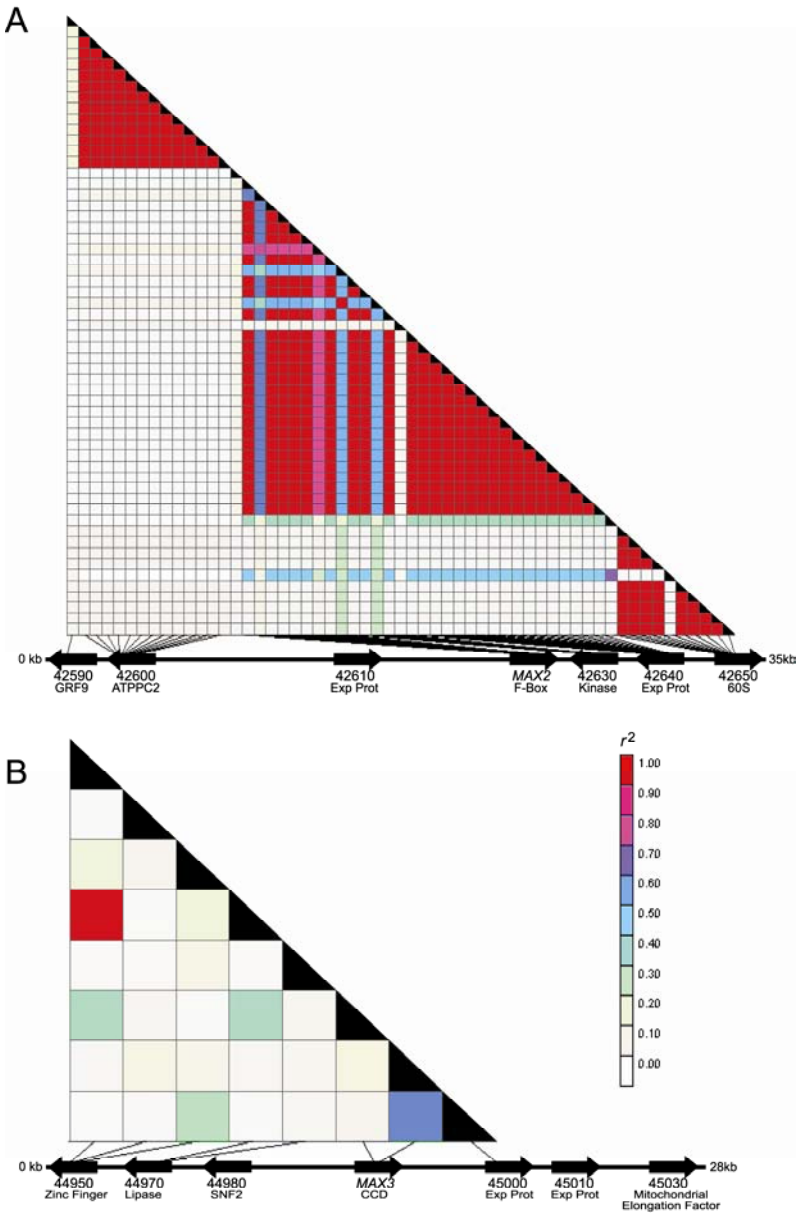


Figure 3. Linkage disequilibrium (LD) across the *MAX2* and *MAX3* regions

The *MAX2* (A) and *MAX3* (B) regions are shown according to scale. Fragments that were not polymorphic or were without common polymorphisms do not have lines connecting the LD plots to the physical maps. The At2g prefix for all numbered genes are omitted in this and subsequent figures. Gene annotations are presented below the gene numbers. 'Exp Prot' is an abbreviation for the annotation expressed protein.

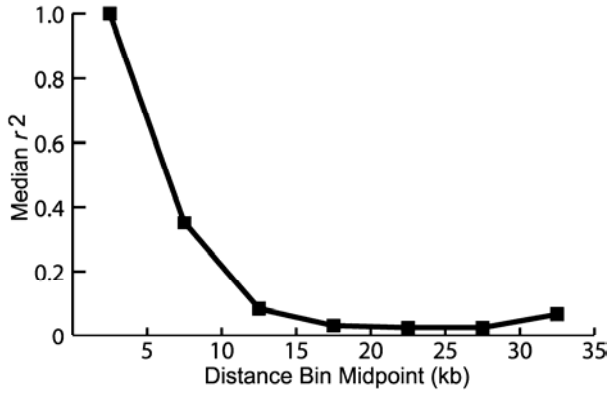


Figure 4. Decay of LD in the *MAX2* and *MAX3* regions

Median r^2 is plotted by the midpoint of each pairwise marker distance bin.

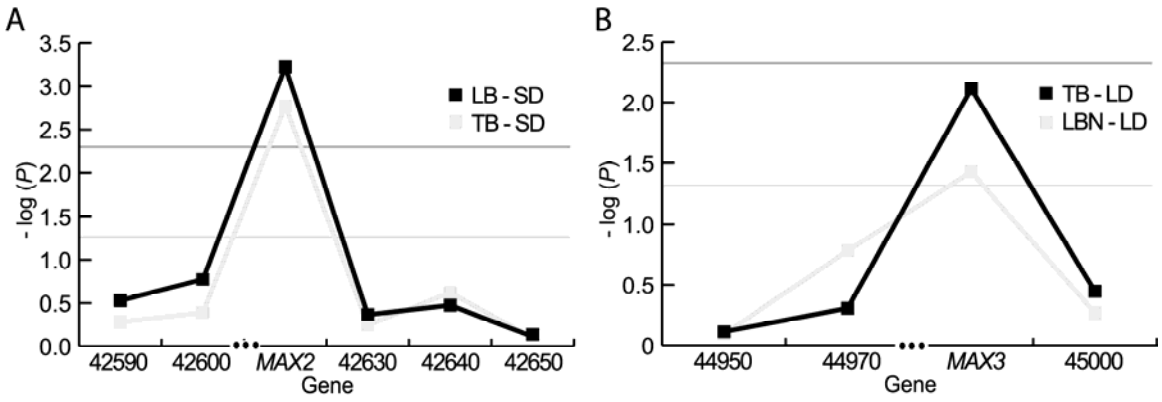


Figure 5. Trait associations across the *MAX2* (A) and *MAX3* (B) regions

P-values are plotted as – log (*P*). Genes with no polymorphisms at ≥ 10% frequency are not included. The light grey horizontal line denotes $P = 0.05$, while the dark grey line represents $P = 0.005$. LB, TB, and LBN are abbreviations for lateral branches, total branches, and lateral branch nodes, respectively, while LD and SD denote long day and short day, respectively.

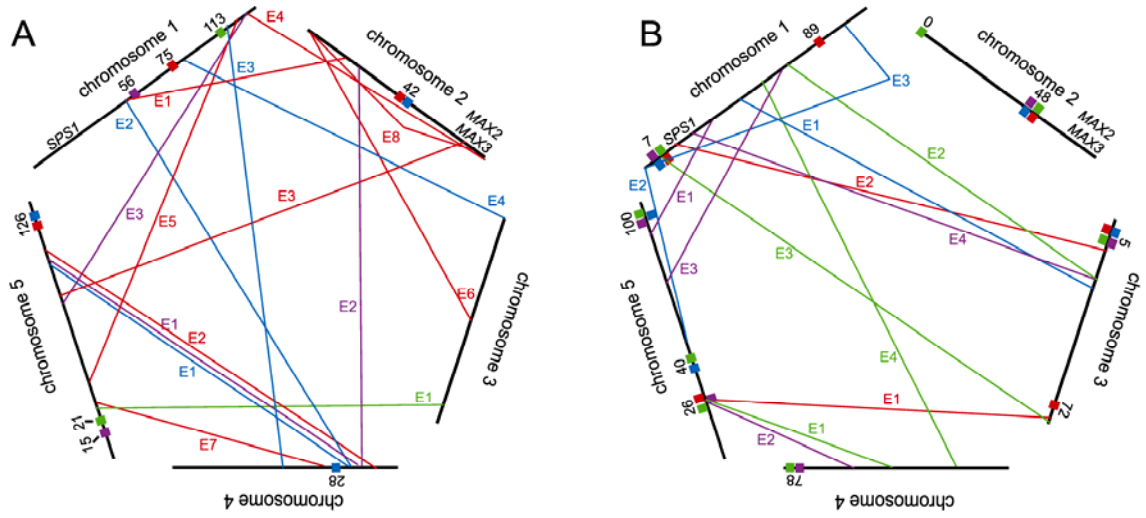


Figure 6. Genomic map of candidate gene associations and QTLs in the L*er* × Col (A) and Cvi × L*er* (B) RILs

Environment-trait combinations are colored as red, green, blue, and purple for lateral branches in long day, lateral branch nodes in long day, lateral branches in short day, and lateral branch nodes in short day, respectively. Each epistatic QTL is referenced to the table of epistatic QTLs by number. Additive QTLs are included on the map as colored rectangles at the marker location reported in Ungerer et al. (2002, 2003) using the same color scheme as for the epistatic QTLs.

Table S1. Broad-sense heritabilities (H^2) for branching traits in 96 accession association mapping panel

Trait	Mean (S.E.)	Min. 2.5%	Max. 2.5%	V_G[a]	V_E[b]	H^2[c]
		----Long Day----				
Lateral Branches	4.22 (.11)	2.06	6.91	1.05	1.487	0.41
Basal Branches	0.80 (.08)	0	2.96	0.45	1.517	0.29
Total Branches	4.72 (.10)	2.70	7.10	0.95	3.167	0.29
Lateral Branch Nodes	5.10 (.10)	3.15	7.26	0.98	2.862	0.34
		----Short Day----				
Lateral Branches	4.28 (.10)	2.09	6.37	0.60	3.248	0.18
Basal Branches	0.20 (.03)	0	1.10	0.05	0.451	0.09
Total Branches	4.45 (.10)	2.13	6.49	0.78	2.898	0.21
Lateral Branch Nodes	6.43 (.13)	4.41	9.73	0.60	2.652	0.18

Note: 10 replicates were grown per accession, though in some cases not all replicates survived to maturity.

[a] Among-ecotype variance component from ANOVA.

[b] Residual variance component from ANOVA.

[c] Calculated as $V_G/(V_G + V_E)$.

CHAPTER THREE:

Network-wide candidate gene association mapping in *Arabidopsis thaliana*

Network-wide candidate gene association mapping in *Arabidopsis thaliana*

Ian M. Ehrenreich[1,2], Yoshie Hanzawa[2], Lucy Chao[2], Paula X. Kover[3], and Michael D. Purugganan[2]

[1]Department of Genetics, Box 7614, North Carolina State University, Raleigh, North Carolina 27695 USA
[2]Department of Biology and Center for Genomics and Systems Biology, New York University, 100 Washington Square East, New York, New York 10003 USA
[3]Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PT, United Kingdom

Corresponding author: Michael D. Purugganan, Telephone: (212) 992-9628, Email: mp132@nyu.edu.

Contributions: IME and MDP designed this study. IME, YH, and LC conducted all experiments and analyses. PXK contributed the HSRILs in this study. IME and MDP cowrote this manuscript.

This chapter is a draft of a manuscript being prepared for submission to PLoS Genetics.

Abstract

In *Arabidopsis thaliana*, substantial variation exists in flowering time, yet only a fraction of the causal alleles underlying this variation have been identified. The genetic pathways responsible for flowering time in *A. thaliana* comprise one of the best characterized molecular networks in plants. We attempt to harness this extensive molecular genetic knowledge to identify potential flowering time quantitative trait genes (QTGs) through candidate gene association mapping. Based on recently reported resequencing data, haplotype tagging SNPs (htSNPs) at 51 flowering time genes, as well as at 182 randomly selected genomewide resequencing fragments, are identified. In total, greater than 1,000 htSNPS from 475 stock center accessions are genotyped and used to identify genes that may be responsible for flowering time variation. The rate of nominal significance for candidate genes and random markers are found to be qualitatively different but statistically indistinguishable. By comparing association mapping results from a panel of 275 natural accessions to those from a set of 360 inbred lines derived from the intercrossing of 19 accessions, several strong candidate flowering time genes that have not been previously reported to affect flowering time variation are identified. However, most of the associations detected in the natural accessions are not replicated in the population of outbred lines, suggesting that a high false positive rate could remain even after the use of conservative statistical approaches that control for population stratification.

Introduction

Flowering time in *A. thaliana* has become the preeminent model for understanding complex trait genetics in plants, in part because of how extensively it has been characterized via forward genetic approaches in type accessions (as described in (SIMPSON and DEAN 2002) and elsewhere). Additionally, to date, a number of quantitative trait genes (QTGs) for flowering time have been identified via both quantitative trait locus (QTL) and association mapping approaches. *CRYPTOCHROME2* (*CRY2*) (EL-ASSAL *et al.* 2001), *FRIGIDA* (*FRI*)

(JOHANSON *et al.* 2000), *FLOWERING LOCUS C* (*FLC*) (WERNER *et al.* 2005a), *FLM* (WERNER *et al.* 2005b), *PHYTOCHROME A* (*PHYA*) (MALOOF *et al.* 2001), *PHYB* (FILIAULT *et al.* 2008), *PHYC* (BALASUBRAMANIAN *et al.* 2006), and *PHYD* (AUKERMAN *et al.* 1997) have all been conclusively shown to harbor natural polymorphisms that alter flowering time. Nearly half of the functional variants at these genes are rare and even accession-specific (i.e. Minor Allele Frequency (MAF) ≤ 0.1), while the other half are common and distributed throughout the species range (i.e. MAF > 0.1). Despite this wealth of knowledge about the population and quantitative genetic bases of natural variation in flowering time in *A. thaliana*, a substantial amount of variation in flowering time remains unexplained (WERNER *et al.* 2005b).

Given that common alleles appear to be important contributors to variation in flowering time in this species and that linkage disequilibrium (LD) decays rapidly in *A. thaliana*, within several genes in most genomic regions (KIM *et al.* 2007), association mapping should be a useful approach to further identify the genetic basis of flowering time variation. Association mapping projects largely fall into two classes – genomewide and candidate gene association studies. Whereas genomewide studies focus on as many polymorphisms as is feasible regardless of the functions of the loci at which these polymorphisms reside, candidate gene studies specifically target genes with known functions in the trait of interest, with the expectation that doing so may enrich for the number of meaningful trait associations. In model organisms, such as *Arabidopsis thaliana*, candidate gene studies may be powerful because many of the genetic pathways underlying ecologically significant traits, such as flowering time, have been dissected through forward genetic approaches, providing strong targets for genes and pathways that might underlie natural variation (EHRENREICH *et al.* 2007).

However, the challenges of association mapping in *A. thaliana* are well-documented (ARANZANA *et al.* 2005; WEIGEL and NORDBORG 2005; ZHAO *et al.* 2007) and involve two main problems. First, variation in most traits is correlated with the population structure that

exists in this species, likely causing a large number of false positive genotype-phenotype correlations throughout the genome. However, it is possible to control for this stratification when conducting association tests (i.e. by using statistical methodologies that take into account different estimates of stratification, as recently described in (YU *et al.* 2006) and (ZHAO *et al.* 2007) and reviewed in (YU and BUCKLER 2006)), and using such controls can decrease the false positive rate for association mapping. Second, for complex traits that are likely to be influenced by numerous QTGs, obtaining confirmatory evidence for a number of associations simultaneously can be difficult. The use of multiple recombinant inbred line (RIL) or F_2 populations has become a common mode of cross-validation (e.g. (EHRENREICH *et al.* 2007; ZHAO *et al.* 2007)) for moderate- to large-scale studies in this species. However, the comparison of association mapping results to linkage mapping results cannot be considered definitive proof or disproof of an association because of resolution differences between the methods, the possibility that the expression of a functional polymorphism may be sensitive to its genetic background, and more simply that each allele of interest may not segregate in an examined mapping population (EHRENREICH *et al.* 2007). How to systematically replicate multiple associations detected in panels of accessions is a major issue that has yet to receive a satisfactory solution.

We attempt here to harness the molecular genetic knowledge of flowering and the rapid decay of LD in *A. thaliana* to identify potential QTGs for flowering time. Using a combination of candidate gene haplotype tagging SNPs (htSNPs) and htSNPs selected at random loci throughout the genome, we conduct network-wide association mapping of flowering time and compare our candidate gene results to those from randomly selected background loci. We then re-test a subset of our associations in an independent panel of inbred lines derived from the intercrossing of 19 accessions, which segregate for all the polymorphisms we examine. Based on the combined results from association mapping in the natural accessions and association mapping in the inbred lines, we identify several promising novel candidates for flowering time variation. Overall, though, our results suggest that the candidate gene approach does not provide a detectable advantage over the use of random loci

and that many of the associations that are detected are likely to be false positives, even after the implementation of statistical controls for population stratification. We discuss the implications of our results for genetic mapping efforts.

Results and discussion

Identification and genotyping of htSNPs at flowering time genes and background loci. *A. thaliana* typically exhibits strong LD on the scale of ~5 to 10 kb (KIM *et al.* 2007). To take advantage of this short-range LD for association mapping, we identified htSNPs representative of common haplotype structure in a panel of 25 accessions at 51 candidate genes that we recently resequenced (Figure S1; Table S1). We also identified htSNPs in 187 fragments that were also previously resequenced (hereafter referred to as '2010') (NORDBORG *et al.* 2005); these 2010 fragments comprise two groups: i) 182 fragments that were randomly selected from a subset of 733 2010 fragments that met quality criteria matching what was found in the candidate genes (hereafter referred to as '2010 Background'), and ii) five fragments that were previously identified as associated with flowering time variation in this species (hereafter referred to as '2010 Most Promising') (ZHAO *et al.* 2007).

Our method for htSNP selection was similar to a recently proposed approach (CARLSON *et al.* 2004). The htSNPs were selected by grouping all common SNPs identified at a locus based on the strength of LD between them (here, we used a threshold of $r^2 = 1$). From each group of SNPs, one was randomly selected to be the htSNP (see Methods). The median and mean numbers of htSNPs identified per candidate gene were 8 and 9.8, respectively; for the 2010 fragments, the median and mean numbers of htSNPs were 2 and 2.7, respectively. We successfully genotyped htSNPs from 475 *A. thaliana* accessions at the candidate genes and the 2010 loci. An additional 131 common SNPs were randomly selected from throughout the genome and successfully genotyped, and used primarily for population structure analyses.

Population Structure in the genotyped accessions. *A. thaliana* possesses extensive population structure that can confound genetic association studies (ARANZANA *et al.* 2005; EHRENREICH *et al.* 2007; NORDBORG *et al.* 2005; ZHAO *et al.* 2007), and we attempted to identify population structure specific to our sample using both the program STRUCTURE (FALUSH *et al.* 2003; PRITCHARD *et al.* 2000a) and the related program InStruct (GAO *et al.* 2007), which explicitly accounts for inbreeding while estimating population structure.

Runs of STRUCTURE and InStruct produced very different most likely *K* estimates, with $K = 10$ and $K = 2$ being maximal for STRUCTURE and InStruct, respectively, suggesting, as has been reported elsewhere (GAO *et al.* 2007), that the inclusion of selfing in population structure estimation can have a dramatic effect on the determination of a most likely *K* value. Ancestry assignments of accessions to particular subpopulations, however, were very similar between the two methods (see Figure 1). Results from $K = 2$ corroborate previous findings of large-scale genetic differentiation between European and Asian *A. thaliana* accessions, with a region of admixture existing in Eastern Europe presumably arising from historical isolation of these populations until ~10,000 years ago (SHARBEL *et al.* 2000). Subpopulations identified at $K > 2$ appear to differentiate subgroups of European ancestry (e.g. Portuguese-Spanish accessions, Scandinavian accessions), which constitute the bulk of our sample.

Analysis of the extent of haplotype sharing between all pairs of accessions shows that despite clear population structure detectable via model-based approaches, most individuals share a large proportion of their alleles (see Figure 2). Haplotype sharing is typically on the order of 30% to 60%, suggesting that despite pervasive population structure due to both geographic isolation and selfing, *A. thaliana* exhibits a high amount of allelic recombination and genotypic diversity.

Distribution of associations across the 2010 Background Loci. We first conducted an association analysis on the 182 2010 Background loci by testing for genotype-phenotype association at the level of haplotypes at each locus (see Methods). The phenotype data used was days to flowering (DF) and rosette leaf number (RLN) in both long day (LD) and short day (SD) growth chamber conditions, measured for 275 accessions with unique multi-locus genotypes (see Methods). Rosette leaf number has been shown to have a high genetic correlation with flowering time, and is often used as a developmental surrogate for this life history trait. Association analyses were conducted using four classes of models; (i) naïve association (ANOVA with no population structure included), (ii) structured association including STRUCTURE ancestry estimates (ANOVA with **Q** as a covariate) (PRITCHARD *et al.* 2000b; THORNSBERRY *et al.* 2001), (iii) mixed model analysis including the **K** haplotype sharing matrix (ANOVA with **K** as a random effect) (YU *et al.* 2006; ZHAO *et al.* 2007), and (iv) mixed model analysis including **K** and **Q** (ANOVA with **K** as a random effect and **Q** as a covariate) (YU *et al.* 2006; ZHAO *et al.* 2007). Models including **Q** were run separately for each per *K* value from $K = 2$ through $K = 10$ (i.e. $\mathbf{Q_2}$ through $\mathbf{Q_{10}}$).

Qualitative differences were observed in the distributions of associations for each trait (see Figures 3, S2, S3; Table 1). For example, SD-RLN exhibited a more severe bias in *P*-values than other traits when population structure controls were not included in genotype-phenotype association tests. Comparison of these models shows that the inclusion of ancestry from runs of STRUCTURE at higher *K* values dramatically reduces the proportion of loci that are nominally significant at the $P \leq 0.05$ level, with the distribution of *P*-values best approximating a uniform distribution at $K = 9$ or 10 for most traits (i.e. inclusion of $\mathbf{Q_9}$ or $\mathbf{Q_{10}}$; see Figures 3, S2, S3). We also find that the inclusion of the **K** matrix in conjunction with the $\mathbf{Q_{10}}$ matrix provides, for most traits, the greatest reduction in the nominal significance rate across the 2010 Background loci (Table 1; Figure 3).

We focused our attention on the association results for $K = 2$ and $K = 10$, since these were the most likely *K* values determined by InStruct and STRUCTURE, respectively, and

since population structure control appeared most effective around $K = 10$ (see Figures S2 and S3; Table 1). We compared the association test results for the 2010 Background loci from the naïve association, the mixed model with **K** only, the structured association with $\mathbf{Q_2}$ or $\mathbf{Q_{10}}$, and the mixed model with both **K** and $\mathbf{Q_2}$ or $\mathbf{Q_{10}}$ (Figure 3; Table 1). These results show that even with the most conservative model ($\mathbf{K} + \mathbf{Q_{10}}$), there remains an excess of nominally significant *P*-values (see Table 1). For instance, 20 2010 Background loci are nominally significant for LD-FT, more than double the nine loci that are expected if 5% of the loci were to be nominally significant by chance.

Significant associations among the 2010 loci. A number of the 2010 Background loci exhibited nominal trait associations with three or four of the examined traits. Of these loci, only two were in or closely linked to known flowering time loci. These were Chromosome 4 position 299,434, which is closely linked to *FRIGIDA* (*FRI*), and Chromosome 5 position 3,177,286, which is in the *FLOWERING LOCUS C* (*FLC*) gene. The Chromosome 4 fragment is nearly 30 kb from *FRI*; however, unusually high LD has been documented surrounding *FRI* for > 100 kb, most likely because of recent positive selection on loss-of-function *FRI* alleles (TOOMAJIAN *et al.* 2006). Interestingly, we previously reported an association at *FLC* (CAICEDO *et al.* 2004), but Zhao et al. (2007) were unable to replicate this association in their dataset.

In their paper, Zhao et al. (2007) also identified six most promising loci other than *FRI* from their analysis of ~900 loci across 96 accessions that were phenotyped for numerous flowering time-related traits. We included these 2010 Most Promising loci in our genotyping, and five – Chromosome 1 position 26,573,187, Chromosome 1 position 28,960,524, Chromosome 2 position 9,964,295, Chromosome 4 position 1,276,056, and Chromosome 5 position 7,442,039 – were successfully genotyped. Only the Chromosome 5 7,442,039 locus of these five fragments is significant in our dataset using the $\mathbf{K} + \mathbf{Q_{10}}$ model, being nominally significant in LD-FT, SD-FT, and SD-RLN and in the 5% tail of all loci in LD-FT and SD-RLN. This fragment is located in At5g22450, an uncharacterized gene that

shares identity with an unknown protein in rice. This highlights the variability in results that is becoming increasingly common in association studies in *A. thaliana* that are performed by different labs.

Association mapping using flowering time candidate genes. We next examined the flowering time candidate genes for associations with flowering time variation using the **K** + $\mathbf{Q_{10}}$ model, which appears to best correct for population stratification. To assess the significance of our candidate gene associations, we compared our candidate gene *P*-values to those observed at the 2010 Background loci. To enable comparison between the two sets of loci, we focused our analysis on one randomly chosen 'core' within each candidate gene that was comparable in size to the 2010 Background loci (see Methods).

The flowering time gene core associations were combined with those from all 2010 loci. From this combined distribution of 238 loci, we determined which candidate gene cores were in the lower 5% tail of the *P*-value distribution for each trait. Several candidate genes were empirically significant based on this approach (see Table 2 and Figure 4). We also applied a trait-wise Bonferroni correction to account for multiple testing, but only one candidate gene core association – *CONSTANS* (*CO*) – surpassed this threshold. Since Bonferroni corrections can be overly conservative, we focused on analyzing the core associations that were empirically significant. Only *CONSTANS* (*CO*) exhibits trait associations across all four traits, being empirically significant for LD-FT, SD-FT, and LD-RLN and nominally significant for SD-RLN. *FRIGIDA* (*FRI*), *FRIGIDA ESSENTIAL 1* (*FES1*), and *FLOWERING LOCUS C* (*FLC*) also were empirically significant for two traits and nominally significant for one additional trait. It is surprising that *FRI*, which is thought to explain a large fraction of the flowering time variation in *A. thaliana*, exhibits empirical significance for only LD-RLN and SD-RLN. Seven other genes exhibit nominal significance for at least one trait (see Table 2). Although these results suggest that a number of genes known to be involved in *A. thaliana* flowering are associated with natural variation in flowering time, these results cannot be regarded as causal evidence of trait association.

Several of these genes, however, including *CO*, *FES1*, and *FLC*, appear particularly promising since they have associations across multiple flowering time traits.

Comparison of candidate gene and background locus associations. Does the use of candidate genes offer an advantage over the use of random markers in association mapping? Answering this question would be particularly useful for species like *A. thaliana* in which LD decays within short physical distances, making it possible that an associated marker may actually be functionally involved in observed trait variation. Specifically, it is helpful to know whether the candidate gene cores possess a higher proportion of nominally or empirically significant associations (nominal or empirical $P \leq 0.05$), as this would be the expectation if these genes were to harbor a disproportionate number of functional polymorphisms. Qualitatively, the candidate genes do appear different from the background loci, exhibiting trait associations (nominal or empirical $P < 0.05$) based on the $\mathbf{K} + \mathbf{Q_{10}}$ mixed model nearly twice as often as the background loci (Figure S4; Tables 3 and 4). Fisher's Exact Tests were performed to determine if these apparent differences were significant, and only SD-FT exhibited a significant result ($P = 0.005$), displaying an excess of nominally significant candidate gene core associations relative to the 2010 background loci (9 out of 51 nominally significant candidate gene cores, 9 out of 182 nominally significant 2010 Background loci). All other tests resulted in *P*-values > 0.1. These results suggest that candidate genes, despite qualitative differences in their rates of trait association, do not exhibit statistically significantly elevated patterns of genotype-phenotype association relative to random loci.

A re-evaluation of the data at the SNP level. An obvious limitation of using candidate gene cores is that relevant polymorphisms may exist outside of the cores. To determine if any promising candidate genes were missed through our core approach, we ran association tests with the $\mathbf{K} + \mathbf{Q_{10}}$ model for all SNPs (Figure 5). Only two candidate gene htSNPs were significant after a trait-wise Bonferroni correction; these were in *CO* and *GAI*. 50 candidate gene htSNPs were in the 5% tail of all SNPs in at least one environment; however, these

SNPs represented only 27 of the flowering time genes. *FLC*, *GA1*, and *HOS1* each had four empirically significant htSNPs, whereas *ELF5*, *FD*, *FES1*, *TFL2*, and *VIN3L* each had three significant htSNPs. Based on our genotyping scheme, such patterns of multiple htSNPs exhibiting association can arise due simply to strong but imperfect LD between the different htSNPs. To determine if this was indeed the case, we compared the patterns of multi-trait association across htSNPs at multiply represented genes, expecting htSNPs at the same gene to typically exhibit similar patterns of association. However, in only 21% of all comparisons across htSNPs at the same locus did two SNPs exhibit the same associations across all traits. Overall, the strongest evidence for association was found at the genes *CO*, *ELF5*, and *FES1*, which each had at least one htSNP that is associated with every trait. Three genes – *GA1*, *GAI*, and *PHYD* – had an htSNP that exhibited associations with three traits.

Association mapping in Recombinant Inbred Lines from Heterogeneous Stock (HSRILs). We initially tried to determine if we could replicate our results using the comparison of our candidate gene associations to QTLs from Recombinant Inbred Lines (RILs), as has become common in association studies in this species (e.g. (EHRENREICH *et al.* 2007; ZHAO *et al.* 2007)). However, we found that no available RIL population segregates for more than a small fraction of our most promising SNPs. This was problematic and led us to try to replicate our associations in a set of HSRILs recently generated from the intercrossing of 19 different progenitors and that segregates for all common SNPs that were genotyped in the natural accessions (Paula Kover, unpublished). As part of the genotyping of these HSRILs, 270 htSNPs from the flowering time genes were genotyped, including 26 htSNPs that appeared promising from the SNP-based association tests in the accessions. The HSRILs were phenotyped for the same traits as those analyzed in the natural accessions (i.e. LD-DF, LD-RLN, SD-DF, SD-RLN) and association results were compared between the two populations. No htSNP that was genotyped in both sets of lines exhibited an identical pattern of association between the two populations; however, eight htSNPs had at least a single common association ($P \leq 0.05$) between the populations (Table 5). These SNPs occurred in *CO*, *FLC* (two SNPs), *GA1* (two SNPs), *PHYD*, *TFL2*, and *VIN3*. Several marginally

nonsignificant associations ($0.05 < P \leq 0.1$) were also found in the HSRILs, providing additional support for *CO*, *PHYD*, and *VIN3* as potential QTGs, and weaker support for *FES1*. These results give corroborative evidence for a subset of the discovered associations, suggesting that they may be biologically meaningful.

Lack of population structure at significant candidate gene htSNPs. We were particularly interested in determining if the htSNPs that were significant in our analyses exhibit unusual levels of population stratification that may cause them to be associated with flowering time despite the use of appropriate controls in association testing. We calculated global F_{ST} using 201 accessions that exhibit near-complete membership (≥ 0.9) to one of the subpopulations estimated by STRUCTURE at $K = 10$, and used two sets of population assignments based on STRUCTURE results at $K = 2$ and $K = 10$. The distributions of F_{ST} were distinct for these two different sets of population assignments, with the mean global F_{ST} assignments at 0.11 for $K = 2$ and 0.34 for $K = 10$. The htSNPs mentioned in the previous section, with the exception of one *FLC* htSNP, which was excluded since it had a minor allele frequency < 0.05 in the accessions used for this analysis, all possessed F_{ST} values close to or below the average F_{ST} at both K values, suggesting that the associations at these promising htSNPs are not due to extreme population stratification.

Polymorphisms generating the replicated flowering time gene associations. We examined the original resequencing data to determine whether the replicated htSNPs with trait associations might themselves have functional effects or whether other polymorphisms that could be functional were linked and in LD with the associated htSNPs. This analysis is summarized in Table 6. Interestingly, only two of the replicated htSNPs were in exonic regions and both of these SNPs cause amino acid changes relative to the Columbia reference sequence. All other replicated htSNPs were in non-exonic sites (i.e. intergenic, intronic, UTR). Three replicate htSNPs had other polymorphisms in perfect LD with them; these were *CO* position 347, *GA1* position 7762, and *PHYD* position 3094. Whereas the *CO* and *GA1* htSNPs had a small number of non-exonic SNPs in perfect LD with them, the *PHYD*

htSNP has 43 SNPs and indels in perfect LD with it within the resequencing data. These *PHYD* polymorphisms occur on 3 of the 25 sequenced chromosomes (12% frequency). *PHYD* has previously been reported to be involved in trait variation in *A. thaliana* (AUKERMAN *et al.* 1997), but the frameshift-causing indel that was identified as the causal polymorphism in this earlier study does not occur in the haplotype that is associated with flowering time in our data.

***Association mapping in A. thaliana*.** Association mapping has become a prominent method for the genetic mapping of trait variation in many species. *A. thaliana* possesses abundant trait variation that remains to be explained at the molecular level (ALONSO-BLANCO *et al.* 2005). The mapping of the QTGs underlying this diversity will facilitate an improved understanding of the molecular processes controlling these traits (KOORNNEEF *et al.* 2004) and will make possible the study of the evolutionary forces shaping the diversity of these traits and their causal QTGs (MITCHELL-OLDS and SCHMITT 2006). Our study begs the question of what is the best way to proceed in mapping the determinants of this trait variation.

Two key methodological issues arise from our study. First, in systems that have been extensively characterized by traditional forward genetic screens, does the knowledge of the genes that control trait development provide useful information about which genes will control variation in the same trait? Our study does not conclusively answer this question, but provides suggestive evidence that there is an improvement in the rate of associated loci gained by having candidate genes but that it is small at best.

Second, even after the implementation of conservative controls for population structure, we still find that most of our detected associations are not replicated in an independent population in which they segregate. This suggests that many of the loci that come up as significant may be highly stratified and consequently spurious even after controls for stratification; however, in terms of genetic differentiation across subpopulations, these

associated alleles appear average or below average. One possible reason for the discrepancy between what we observe in the accessions and the HSRILs could be that our htSNPs are not the causal SNPs underlying many of the detected associations, but instead are merely in disequilibrium with the causal SNPs. In such a case, if recombination were to occur between the causal SNP and an htSNP during the breeding of the HSRILs, then the association would not be possible to replicate in the HSRILs using the same htSNP. In addition, the environments used for growing the accessions and the HSRILs were slightly different because these experiments were conducted at different universities, and it is possible that this difference could have had an effect on the genotype-phenotype associations present in each experiment. Lastly, as we have stated previously (EHRENREICH *et al.* 2007), the possibility that epistatic relationships that appear as additive effects in accessions due to historical population structure and selfing are disrupted during the construction of inbred mapping populations cannot be discounted. How common background-sensitive functional polymorphisms are in this species is unclear, but it is plausible that this variance conversion effect could be at play in our study. The cause of the discrepancies between these experiments cannot be resolved with the present data. Overall, however, the cumulative evidence that we present suggests that we have identified worthwhile follow-up targets for validation as potential QTGs.

For genetic model systems like *A. thaliana*, in which it is possible to construct crosses of any desired pair or larger set of accessions, the future of the genetic mapping of trait variation may not be in association mapping, but instead in the construction of new mapping populations. The HSRILs used in this study are representative of a general movement in quantitative genetics to use genetic mapping approaches that combine linkage and association mapping to get true positives that control trait variation at a high resolution (CHURCHILL *et al.* 2004; MACDONALD and LONG 2007; YU *et al.* 2008). Which mapping approach is preferable may also depend on the genetic architecture of the trait being examined, as different considerations may take precedence when mapping truly complex,

highly polygenic traits like flowering time as opposed to monogenic or lowly polygenic traits.

We have extensively surveyed an entire genetic network to identify putative QTGs. Overall, we have strongest evidence for the genes *CO*, *FLC*, *GA1*, and *TFL2* as being potential flowering time QTGs. These results do not suggest that a particular clique of the flowering time network is responsible for flowering time variation, but instead that flowering time variation can arise from several nodes in the network. We have laid the foundation for additional research that can validate the effects of these genes to determine which of them are actual QTGs and what the molecular mechanisms of these true positives are.

Materials and methods

Resequencing Data. The resequencing data used in this paper are from several sources. Data for 48 of the flowering time candidate genes will soon be published by our lab for 24 accessions. These same accessions are among the 96 used by Nordborg et al. in generating their 2010 data. For the 2010 data, we used only the alleles from the 24 accessions overlapping those used by Hanzawa et al. (unpublished) in addition to the Columbia reference allele. Previously published resequencing data were used for the genes *CRY2, FLC,* and *FRI*, and the specific accessions and the total number of accessions used in these studies are variable and different from the Hanzawa et al. data (CAICEDO *et al.* 2004; OLSEN *et al.* 2004; STINCHCOMBE *et al.* 2004).

HtSNP Selection. HtSNPs were chosen using an algorithm that grouped all common SNPs ($p \geq 0.1$) in a multiple sequence alignment for a locus into bins based on their patterns of LD, with the threshold for binning being $r^2 = 1$. Sites with gaps or missing data were ignored by the binning procedure. From each bin, one SNP was randomly selected to be the htSNP representing that bin. Because the resequencing data used to identify candidate gene htSNPs were oftentimes more than an order of magnitude longer than the 2010 data, we randomly subsampled a polymorphic 600 bp region from each candidate gene that we refer to as the

'core.' These cores were used to promote the comparison of our candidate gene data to what we observe throughout the genome at the 2010 loci. Candidate gene cores were identified during the selection of htSNPs, so that they could be comprehensively genotyped, with remaining haplotype structure outside the core covered by genotyped non-core htSNPs. Although this approach facilitated direct comparison of our data and the 2010 data, it is limited when causal polymorphisms or informative SNPs linked to them exist outside the candidate gene cores.

The identified htSNPs were genotyped in a panel of 475 accessions. The DNA used for genotyping was isolated from the leaves of plants grown under 24 hrs light for three weeks at New York University. Qiagen 96-well DNAeasy kits were used to extract the DNA. Genotyping was done using the Sequenom MassArray technology and was conducted by Sequenom (http://www.sequenom.com). Overall, ~87% of the htSNPs were successfully genotyped in ≥ 375 accessions. We also genotyped an additional 150 common SNPs throughout the genome at random loci distinct from those used for htSNP genotyping. 131 of these SNPs (87%) were successfully genotyped. These SNPs were used primarily for the assessment of population structure.

Population Structure Assessment. Two programs – STRUCTURE (FALUSH *et al.* 2003; PRITCHARD *et al.* 2000a) and InStruct (GAO *et al.* 2007) – were used to determine the extent of population structure in our panel of accessions. These programs are very similar, with the primary difference being that InStruct explicitly estimates selfing rates along with population structure. In these analyses, the 131 genomewide SNPs for population structure in addition to 182 htSNPs (one randomly selected from each successfully genotyped 2010 Background locus) were used. Only accessions with unique multi-locus genotypes across all background markers were included and in cases where accessions were identical across all background loci, only one accession was included as the representative of that multi-locus genotype. Both programs were run three times across a range of K values starting at $K = 1$ and ending at $K = 30$. In STRUCTURE, the correlated frequencies with admixture model was used. In

InStruct, mode 2 was used, which infers population structure and selfing rates at the population level.

Calculation of Haplotype Sharing. Haplotype sharing (**K**) was computed from a data matrix including all accessions and their genotypes at both the population structure SNPs (SNP genotypes) and the 2010 Background loci (haplotypes based on multiple SNPs). As in Zhao et al. (2007), haplotype sharing was computed between every possible pairwise combination of accessions as the total number of loci in common between the accessions divided by the total number of loci with present data for both individuals. This provides a measure of the proportion of loci that are identical in state between any pair of accessions.

Phenotyping of the Natural Accessions. Phenotype data used for association mapping with the natural accessions are from growth chamber experiments conducted at North Carolina State University's Phytotron facility and are previously published (OLSEN *et al.* 2004). Phenotyping of the HSRILs was done in growth chambers at New York University. In brief, the phenotyping experiments for both the accessions and the HSRILs were conducted using long day conditions (14 hrs light: 10 hrs dark) and short day conditions (10 hrs light: 14 hrs dark). Days to flowering were measured as the number of julien days after which the primary inflorescence had extended more than 1 mm above the rosette, whereas rosette leaf number was the number of total rosette leaves on a plant at bolting.

Association Tests. 275 phenotyped accessions with non-redundant multi-locus genotypes were used for association mapping. Several models were used for testing for genotype-phenotype association. The most complex model used was of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Q}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

with **Y** a vector of phenotypes, **X** a vector of single locus genotypes that were considered as fixed effects, $\boldsymbol{\alpha}$ a vector of fixed effects of the n – 1 genotype classes, **Q** a matrix of the $K - 1$ subpopulation ancestry estimates for each individual from STRUCTURE, $\boldsymbol{\beta}$ a vector of the fixed effects for each of the $K - 1$ subpopulations, **Z** an identity matrix, **u** a matrix of random

deviates due to genomewide relatedness (as inferred from **K**), and **ε** a vector of residual errors. This full model we refer to in the text as the mixed model with **Q** + **K** (YU *et al.* 2006; ZHAO *et al.* 2007). In addition, to this model, we consider models without the **Q** or the **K** terms (the mixed model with **K** only (YU *et al.* 2006; ZHAO *et al.* 2007) and the structured association [ANOVA with **Q**] (PRITCHARD *et al.* 2000b; THORNSBERRY *et al.* 2001), respectively), as well as the model without either source of population structure information (naïve association [standard one-way ANOVA]). PROC MIXED was used for all tests and was run in SAS v9.1.3. For the HSRILs, one-way ANOVAs were conducted in JMP v5 to test if SNPs were associated with differences in flowering time across the lines. Note that HSRILs have not been formally reported (Paula Kover, unpublished), although the inercrossing phase of line construction has been described elsewhere (SCARCELLI *et al.* 2007).

Calculation of F_{ST}. 861 SNPs from our dataset met the criteria for inclusion in this analysis (≥ 50% complete data and minor allele frequency ≥ 0.05 in the group of 201 accessions with ≥ 0.9 membership to one subpopulation). We used a previously described formula for calculating F_{ST} that accounts for differences in sample sizes across subpopulations (WEIR 1996).

Acknowledgments

We thank Johanna Schmitt, Steve Welch, Amity Wilczek and members of the Purugganan lab for insight into this project or comments on this manuscript. I.M.E was supported by both a Department of Education Graduate Assistance in Areas of National Need Biotechnology Fellowship and by a National Science Foundation (NSF) Department of Education Graduate Research Fellowship. M.D.P. was supported by grants from the Department of Defense and the NSF's Frontiers in Biological Research program.

References

ALONSO-BLANCO, C., B. MENDEZ-VIGO and M. KOORNNEEF, 2005 From phenotypic to molecular polymorphisms involved in naturally occurring variation of plant development. Int J Dev Biol **49:** 717-732.

ARANZANA, M. J., S. KIM, K. ZHAO, E. BAKKER, M. HORTON *et al.*, 2005 Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes. PLoS Genet **1:** e60.

AUKERMAN, M. J., M. HIRSCHFELD, L. WESTER, M. WEAVER, T. CLACK *et al.*, 1997 A deletion in the PHYD gene of the Arabidopsis Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. Plant Cell **9:** 1317-1326.

BALASUBRAMANIAN, S., S. SURESHKUMAR, M. AGRAWAL, T. P. MICHAEL, C. WESSINGER *et al.*, 2006 The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana. Nat Genet **38:** 711-715.

CAICEDO, A. L., J. R. STINCHCOMBE, K. M. OLSEN, J. SCHMITT and M. D. PURUGGANAN, 2004 Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci U S A **101:** 15670-15675.

CARLSON, C. S., M. A. EBERLE, M. J. RIEDER, Q. YI, L. KRUGLYAK *et al.*, 2004 Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet **74:** 106-120.

CHURCHILL, G. A., D. C. AIREY, H. ALLAYEE, J. M. ANGEL, A. D. ATTIE *et al.*, 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet **36:** 1133-1137.

EHRENREICH, I. M., P. A. STAFFORD and M. D. PURUGGANAN, 2007 The genetic architecture of shoot branching in Arabidopsis thaliana: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping. Genetics **176:** 1223-1236.

EL-ASSAL, S. E.-D., C. ALONSO-BLANCO, A. J. 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.

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.

FILIAULT, D. L., C. A. WESSINGER, J. R. DINNENY, J. LUTES, J. O. BOREVITZ *et al.*, 2008 Amino acid polymorphisms in Arabidopsis phytochrome B cause differential responses to light. Proc Natl Acad Sci U S A **105:** 3157-3162.

GAO, H., S. WILLIAMSON and C. D. BUSTAMANTE, 2007 A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics **176:** 1635-1651.

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.

KIM, S., V. PLAGNOL, T. T. HU, C. TOOMAJIAN, R. M. CLARK *et al.*, 2007 Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat Genet **39:** 1151-1155.

KOORNNEEF, M., C. ALONSO-BLANCO and D. VREUGDENHIL, 2004 Naturally occurring genetic variation in Arabidopsis thaliana. Annu Rev Plant Biol **55:** 141-172.

MACDONALD, S. J., and A. D. LONG, 2007 Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of Drosophila melanogaster. Genetics **176:** 1261-1281.

MALOOF, J. N., J. O. BOREVITZ, T. DABI, J. LUTES, R. B. NEHRING *et al.*, 2001 Natural variation in light sensitivity of Arabidopsis. Nat Genet **29:** 441-446.

MITCHELL-OLDS, T., and J. SCHMITT, 2006 Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. Nature **441:** 947-952.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol **3:** e196.

OLSEN, K. M., S. S. HALLDORSDOTTIR, J. R. STINCHCOMBE, C. WEINIG, J. SCHMITT *et al.*, 2004 Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics **167:** 1361-1369.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000a Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000b Association mapping in structured populations. Am J Hum Genet **67:** 170-181.

SCARCELLI, N., J. M. CHEVERUD, B. A. SCHAAL and P. X. KOVER, 2007 Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. Proc Natl Acad Sci U S A **104:** 16986-16991.

SHARBEL, T., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Genetic distance by isolation in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Molecular Ecology **9:** 2109-2118.

SIMPSON, G. G., and C. DEAN, 2002 Arabidopsis, the Rosetta stone of flowering time? Science **296:** 285-289.

STINCHCOMBE, J. R., C. WEINIG, M. UNGERER, K. M. OLSEN, C. MAYS *et al.*, 2004 A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci U S A **101:** 4712-4717.

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.

TOOMAJIAN, C., T. T. HU, M. J. ARANZANA, C. LISTER, C. TANG *et al.*, 2006 A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. PLoS Biol **4:** e137.

WEIGEL, D., and M. NORDBORG, 2005 Natural variation in Arabidopsis. How do we find the causal genes? Plant Physiol **138:** 567-568.

WEIR, B., 1996 *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

WERNER, J. D., J. O. BOREVITZ, N. H. UHLENHAUT, J. R. ECKER, J. CHORY *et al.*, 2005a FRIGIDA-independent variation in flowering time of natural Arabidopsis thaliana accessions. Genetics **170:** 1197-1207.

WERNER, J. D., J. O. BOREVITZ, N. WARTHMANN, G. T. TRAINER, J. R. ECKER *et al.*, 2005b Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci U S A **102:** 2460-2465.

YU, J., and E. S. BUCKLER, 2006 Genetic association mapping and genome organization of maize. Curr Opin Biotechnol **17:** 155-160.

YU, J., J. B. HOLLAND, M. D. MCMULLEN and E. S. BUCKLER, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics **178:** 539-551.

YU, J., G. PRESSOIR, W. H. BRIGGS, I. VROH BI, M. YAMASAKI *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet **38:** 203-208.

ZHAO, K., M. J. ARANZANA, S. KIM, C. LISTER, C. SHINDO *et al.*, 2007 An Arabidopsis example of association mapping in structured samples. PLoS Genet **3:** e4.

Table 1. Numbers of nominally significant background loci by model[1]

Model	LD – FT	SD - FT	LD – RLN	SD – RLN
Naïve	27	16	23	29
K	27	19	21	29
Q_2	27	16	25	29
Q_{10}	21	7	22	15
K & Q_2	27	16	22	26
K & Q_{10}	20	9	14	13

[1] 5% of all background loci is equal to ~9 loci.

Table 2. Candidate gene core associations[1]

Gene	Core			
	LD-FT	**SD-FT**	**LD-RLN**	**SD-RLN**
CO	*	*	*	+
FES1	*	+	*	
FLC	*	+	*	
FRI	+		*	*
HOS1		*		+
PHYD	+	*		
PIE1			*	
TFL1		+		*
TFL2		+		*
VIN3	+	*	+	
VIN3-L		*		

[1] Empirical significance is indicated by an "*", whereas nominal significance is indicated by a "+". "-" indicates no significance.

Table 3. Counts of nominally significant loci per trait

Trait	Candidate Gene Cores		2010 Background	
	***P* ≤ 0.05**	***P* > 0.05**	***P* ≤ 0.05**	***P* > 0.05**
LD-FT	6	45	20	162
SD-FT	9*	42	9	173
LD-RLN	6	45	14	168
SD-RLN	6	45	13	169

* Excess of nominally significant candidate gene cores. Fisher's Exact Test $P = 0.005$.

Table 4. Counts of empirically significant loci per trait

Trait	Candidate Gene Cores		2010 Background	
	$P \leq 0.05$	$P > 0.05$	$P \leq 0.05$	$P > 0.05$
LD-FT	3	48	7	175
SD-FT	4	47	7	175
LD-RLN	5	46	6	176
SD-RLN	3	48	7	175

Table 5. Comparison of associations in accessions and HSRILs[1]

HtSNP	Accessions				HSRILs			
	LD-FT	SD-FT	LD-RLN	SD-RLN	LD-FT	SD-FT	LD-RLN	SD-RLN
ATMYB33 p119				+	+			
CO p347	+	+	+	+		+		
FD p2772			+	+				
FES1 p1877	+	+	+					
FLC p2775 & p3312	+				+	+	+	+
FLC p6809	+							
FRI p725			+	+				
GA1 p7762			+		+	+	+	
GA1 p8429		+	+				+	
GI p5241	+							
HOS1 p1176 & p5516	+							
HOS1 p1788		+		+				
LD p258		+						
PHYD p3094	+	+		+	+			+
PIE p898			+					
RGL2 p2115		+		+				
TFL2 p1199		+						
TFL2 p1346				+				
TFL2 p2993		+		+	+	+		+
VIN3 p2942	+		+	+			+	
VIN3-L p50 & p5026	+		+					
VIN3-L 4961	+	+						

[1] Light grey htSNPs were replicated once and dark grey htSNPs were replicated twice.

Table 6. Locations of replicated htSNPs and polymorphisms that are in LD with them

HtSNP	Location	Polymorphisms in Perfect LD[1]
CO p347	5' Intergenic/Putative Promoter	3: 5' UTR (1), Intronic (1), 3' Intergenic (1)
FLC p2775	Intronic	None
FLC p3312	Intronic	None
GA1 p7762	Intronic	2: Intronic (1), 3' Intergenic (1)
GA1 p8429	Exonic-Replacement	None
PHYD p3094	Exonic-Replacement	43: Exonic-Replacement (9), Exonic-Silent (11), Intronic (3 SNPs, 1 indel), 3' Intergenic (15 SNPs, 4 indels)
TFL2 p2993	3' UTR	None
VIN3 p2942	3' Intergenic	None

[1] Unless specified, stated polymorphisms are SNPs.

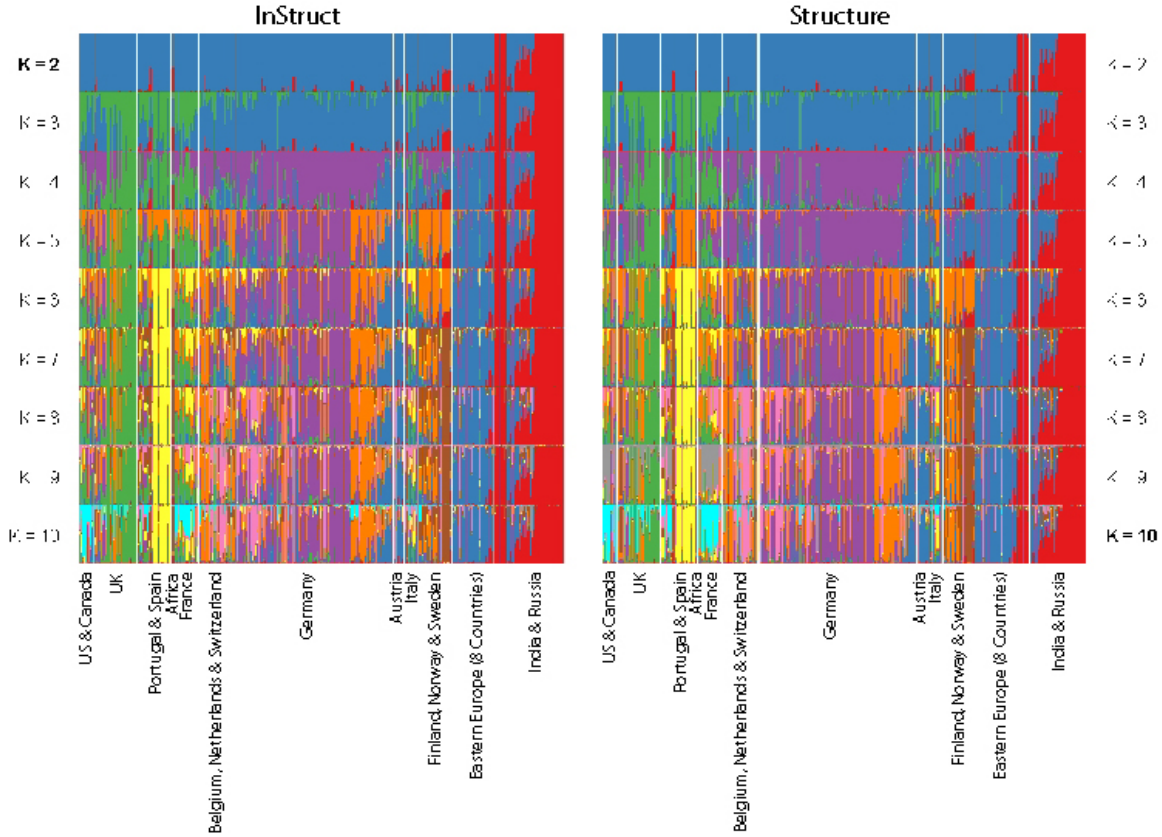


Figure 1. Population structure in the genotyped accessions

A plot of ancestry estimated from Instruct and Structure for the 403 unique genotypes in the dataset. Runs from $K = 2$ through $K = 10$ are presented. White vertical lines separate countries or geographic regions. Within each region, accessions are sorted by latitude with the lowest and highest reported latitudes of sampling within a region on the left and right sides, respectively.

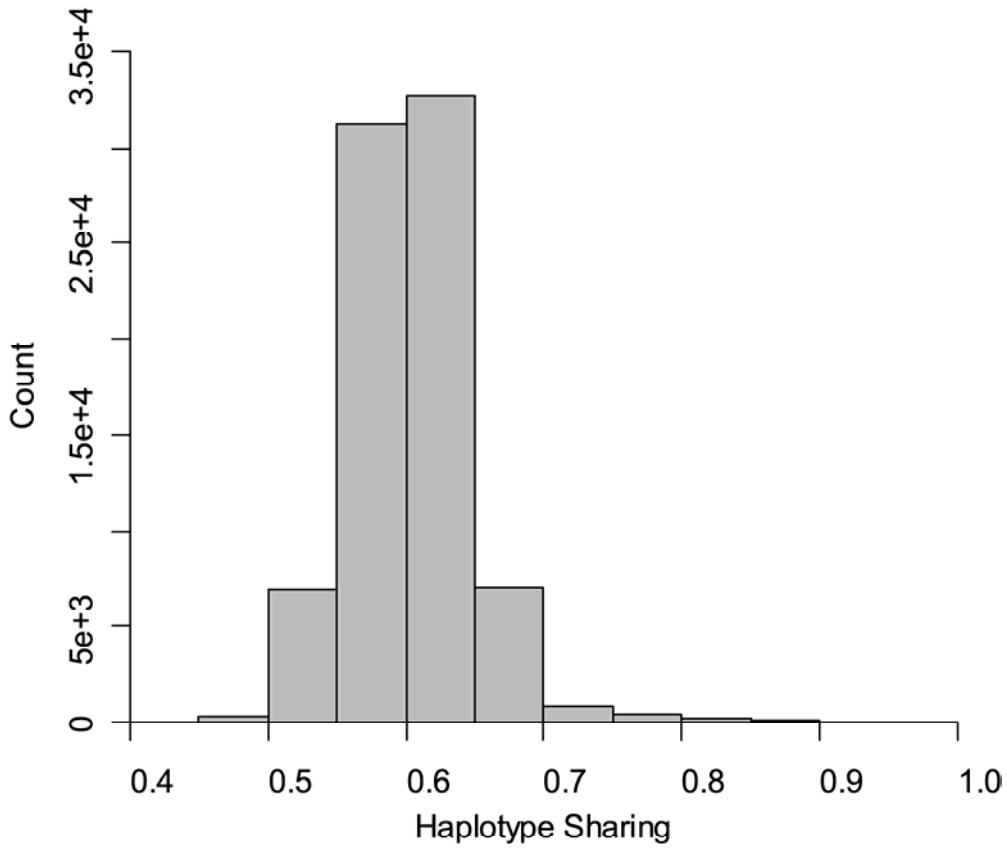


Figure 2. Haplotype sharing across the unique genotypes

The proportion of alleles shared between any two individuals are plotted as a histogram.

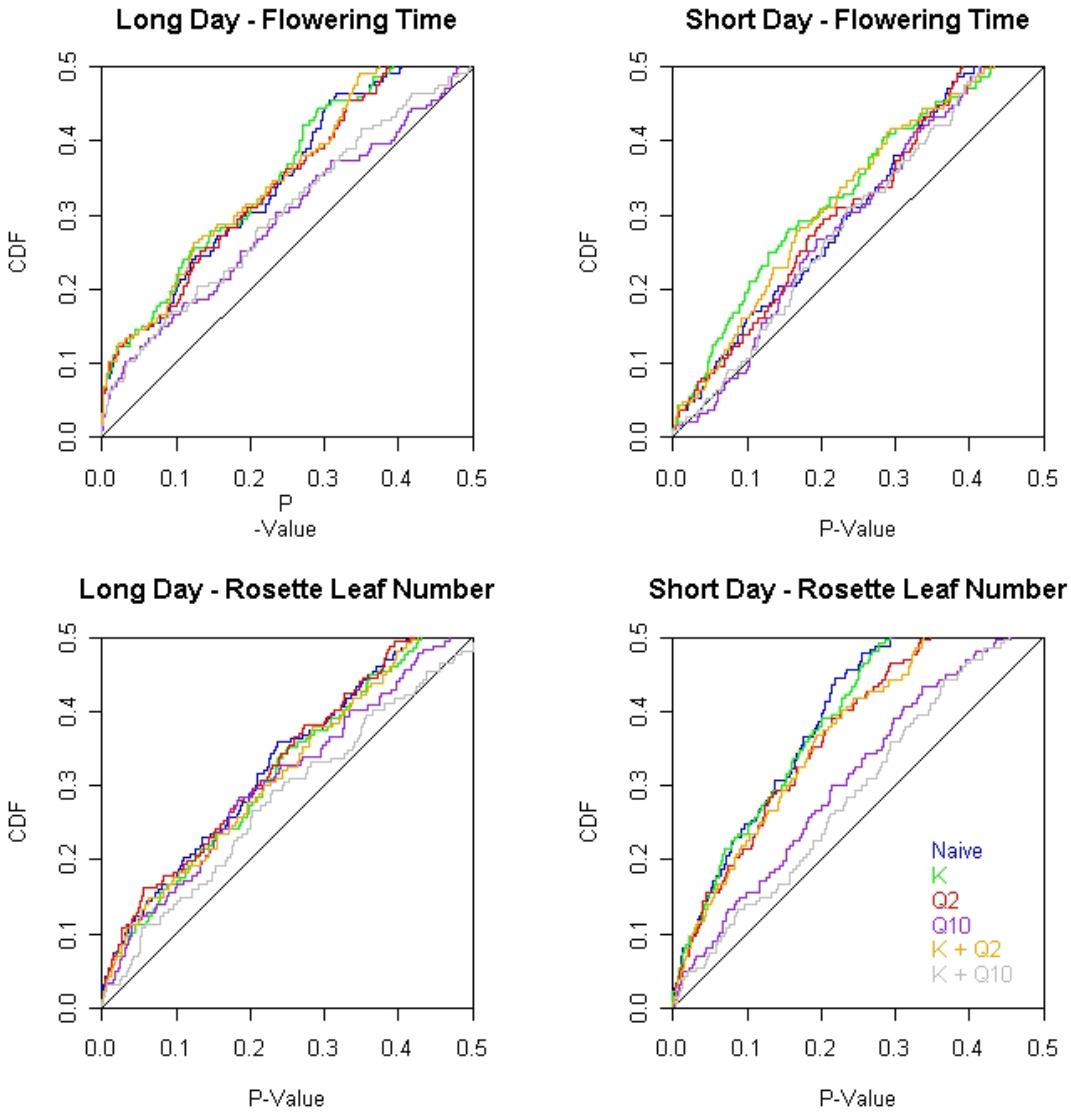


Figure 3. Cumulative density functions (cdfs) for the 2010 Background loci using several alternative models

The Naïve association is a one-way ANOVA, whereas the models including **K** (i.e. haplotype sharing) and/or **Q** (i.e. STRUCTURE ancestry estimates) are variants of the full model described in the Methods. The axes are restricted to a maximum of 0.5 to facilitate comparison of the different models.

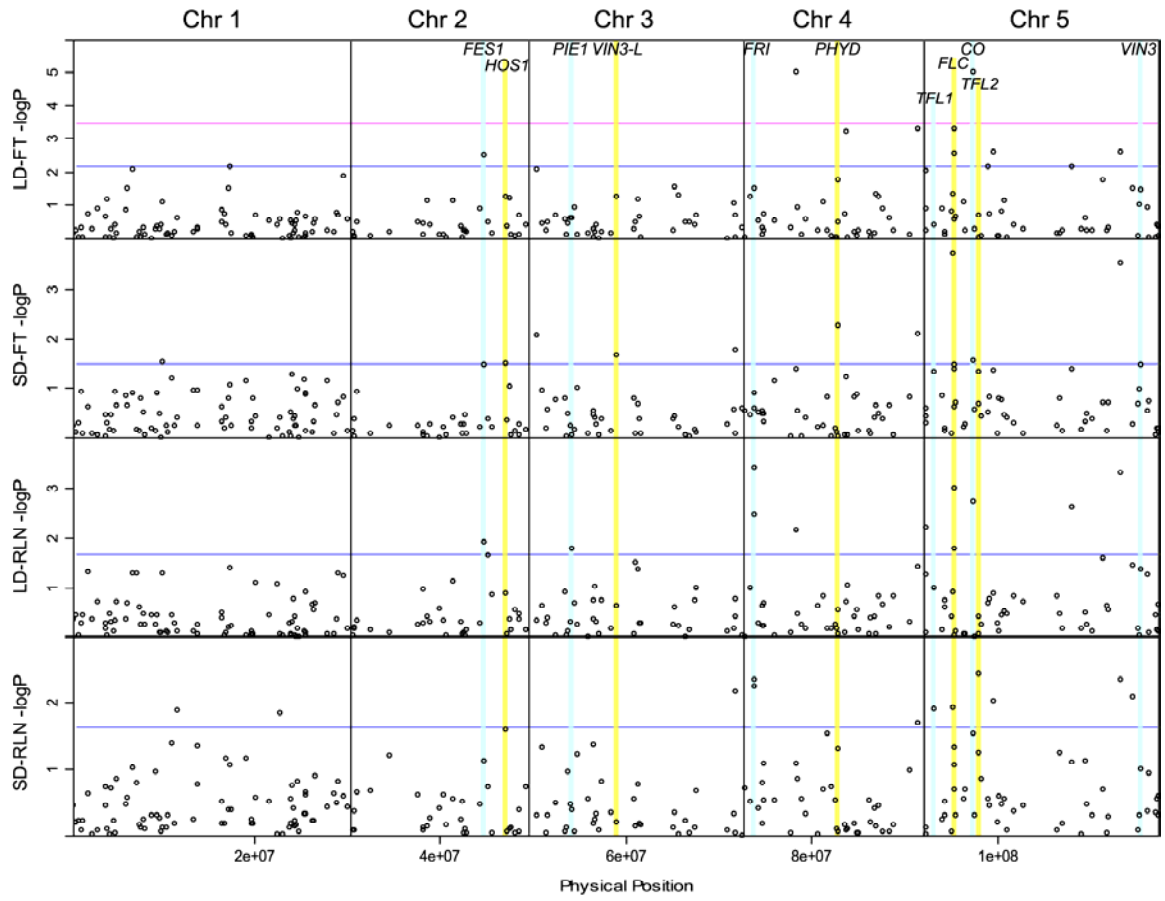


Figure 4. Associations across all haplotypes

Results for the **K** + $\mathbf{Q_{10}}$ model at each 2010 locus and flowering time gene core are plotted as –log (*P*-value) by physical position in the genome. Candidate gene cores that are empirically significant are highlighted by light blue or yellow horizontal lines. Empirical and Bonferroni-corrected multiple-testing thresholds (α = 0.05) for significance are plotted by trait as blue and red horizontal lines, respectively. For SD-FT, LD-RLN, and SD-RLN, the Bonnferroni-corrected threshold is not shown as it exceeds the most significant locus.

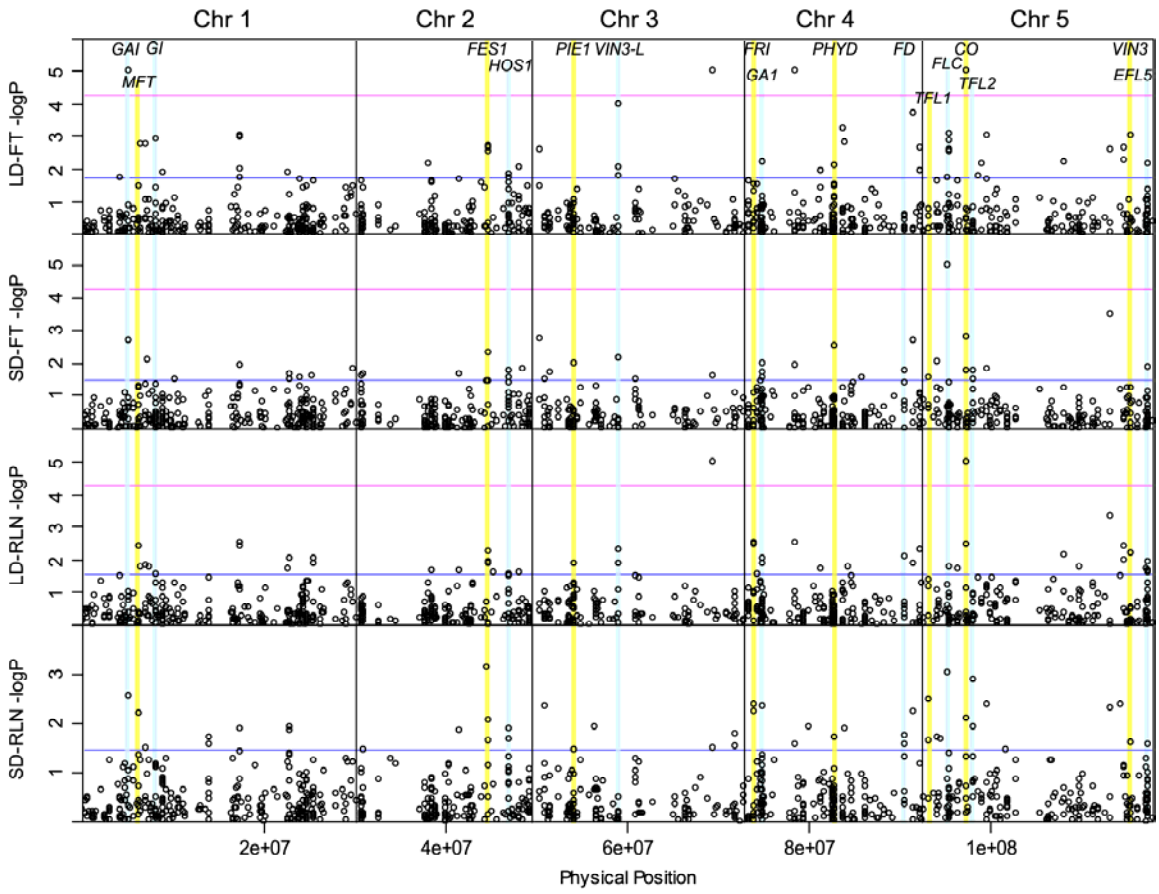


Figure 5. Associations across all SNPs

Results for the **K** + $\mathbf{Q_{10}}$ model at each genotyped SNP are plotted as –log (*P*-value) by physical position in the genome. Candidate gene htSNPs that are empirically significant are highlighted by light blue or yellow horizontal lines. Empirical and Bonferroni-corrected multiple-testing thresholds (α = 0.05) for significance are plotted by trait as blue and red horizontal lines, respectively. For SD-RLN, the Bonnferroni-corrected threshold is not shown as it exceeds the most significant locus.

Table S1. Genes included in this study

Gene	Abbreviation	Gene ID	Annotation
AGAMOUS-LIKE 24	*AGL24*	At4g24540	MADS-box protein
Arabidopsis thaliana CENTRORADIALIS	*ATC*	At2g27550	*TFL1* homolog
MYB DOMAIN PROTEIN 33	*ATMYB33*	At5g06100	Myb transcription factor 33
BROTHER OF FT AND TFL1	*BFT*	At5g62040	*FT* homolog
CYCLING DOF FACTOR 1	*CDF1*	At5g62430	Dof-type zinc finger
CONSTANS	*CO*	At5g15840	Similar to zinc finger
CRYPTOCHROME 1	*CRY1*	At4g08920	Blue-light photoreceptor
CRYPTOCHROME 2	*CRY2*	At1g04400	Blue-light photoreceptor
EARLY BOLTING IN SHORT DAYS	*EBS*	At4g22140	Putative plant chromatin remodeling factor
EARLY FLOWERING 5	*ELF5*	At5g62640	Nuclear targeted protein
EARLY IN SHORT DAYS 4	*ESD4*	At4g15880	SUMO-specific protease
FD	*FD*	At4g35900	bZIP transcription factor
FD PARALOG	*FDP*	At2g17770	bZIP transcription factor
FRIGIDA-ESSENTIAL 1	*FES1*	At2g33835	Zinc finger
FLAVIN-BINDING KELCH DOMAIN F BOX PROTEIN 1	*FKF1*	At1g68050	F-box protein
FLOWERING LOCUS C	*FLC*	At5g10140	MADS-box protein
FLOWERING LOCUS KH DOMAIN	*FLK*	At3g04610	Nucleic acid binding
FLOWERING PROMOTING FACTOR 1	*FPF1*	At5g24860	Small, 12.6 kDa protein
FRIGIDA	*FRI*	At4g00650	Vernalization response factor
FRIGIDA-LIKE 1	*FRL1*	At5g16320	*FRI*-related gene
FRIGIDA-LIKE 2	*FRL2*	At1g31814	*FRI*-related gene
FLOWERING LOCUS T	*FT*	At1g65480	*TFL1* homolog; antagonist of *TFL1*
FVE	*FVE*	At2g19520	Unknown
GA REQUIRING 1	*GA1*	At4g02780	Gibberellin biosynthesis
GA INSENSITIVE	*GAI*	At1g14920	Repressor of GA responses
GAI AN REVERTANT 1; *GA INSENSITIVE DWARF 1C*	*GAr1*	At5g27320	GA receptor homolog
GAI AN REVERTANT 2; *GA INSENSITIVE DWARF 1A*	*GAr2*	At3g05120	GA receptor homolog
GAI AN REVERTANT 3; *GA INSENSITIVE DWARF 1B*	*GAr3*	At3g63010	GA receptor homolog
GIGANTEA	*GI*	At1g22770	Circadian clock gene
HIGH EXPRESSION OF OSMOTICALLY RESPONSIVE GENES 1	*HOS1*	At2g39810	RING finger E3 ligase

Table S1 (continued)

Gene	Abbreviation	Gene ID	Annotation
ENHANCER OF AG-4 2	*HUA2*	At5g23150	Transcription factor
LUMINIDEPENDENS	*LD*	At4g02560	Transcription factor
MOTHER OF FT AND TFL1	*MFT*	At1g18090	Nuclease
PHYTOCHROME AND FLOWERING TIME 1	*PFT1*	At1g25540	Transcription coactivator
PHYTOCHROME A	*PHYA*	At1g09570	G-protein coupled red/far red photoreceptor
PHYTOCHROME B	*PHYB*	At2g18790	G-protein coupled red/far red photoreceptor
PHYTOCHROME D	*PHYD*	At4g16250	G-protein coupled red/far red photoreceptor
PHYTOCHROME E	*PHYE*	At4g18130	G-protein coupled photoreceptor
PHOTOPERIOD-INDEPENDENT EARLY FLOWERING 1	*PIE1*	At3g12810	ATP-dependent chromatin remodeling protein
REPRESSOR OF GA1-3	*RGA*	At2g01570	VHIID/DELLA transcription factor
RGA-LIKE 1	*RGL1*	At1g66350	RGA homolog
RGA-LIKE 2	*RGL2*	At3g03450	RGA homolog
SLEEPY 1	*SLY1*	At4g24210	F-box protein involved in GA signaling
SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1	*SOC1*	At2g45660	Transcription factor
SPINDLY	*SPY*	At3g11540	Glucosamine transferase
SHORT VEGETATIVE PHASE	*SVP*	At2g22540	Transcription factor
TERMINAL FLOWER 1	*TFL1*	At5g03840	Phosphatidylethanolamine binding
TERMINAL FLOWER 2	*TFL2*	At5g17690	Chromatin maintenance protein
TWIN SISTER OF FT	*TSF*	At4g20370	FT homolog
VERNALIZATION INSENSITIVE 3	*VIN3*	At5g57380	Homeodomain protein
VERNALIZATION INSENSITIVE 3-LIKE 1	*VIN3-L*	At3g24440	Chromatin modification

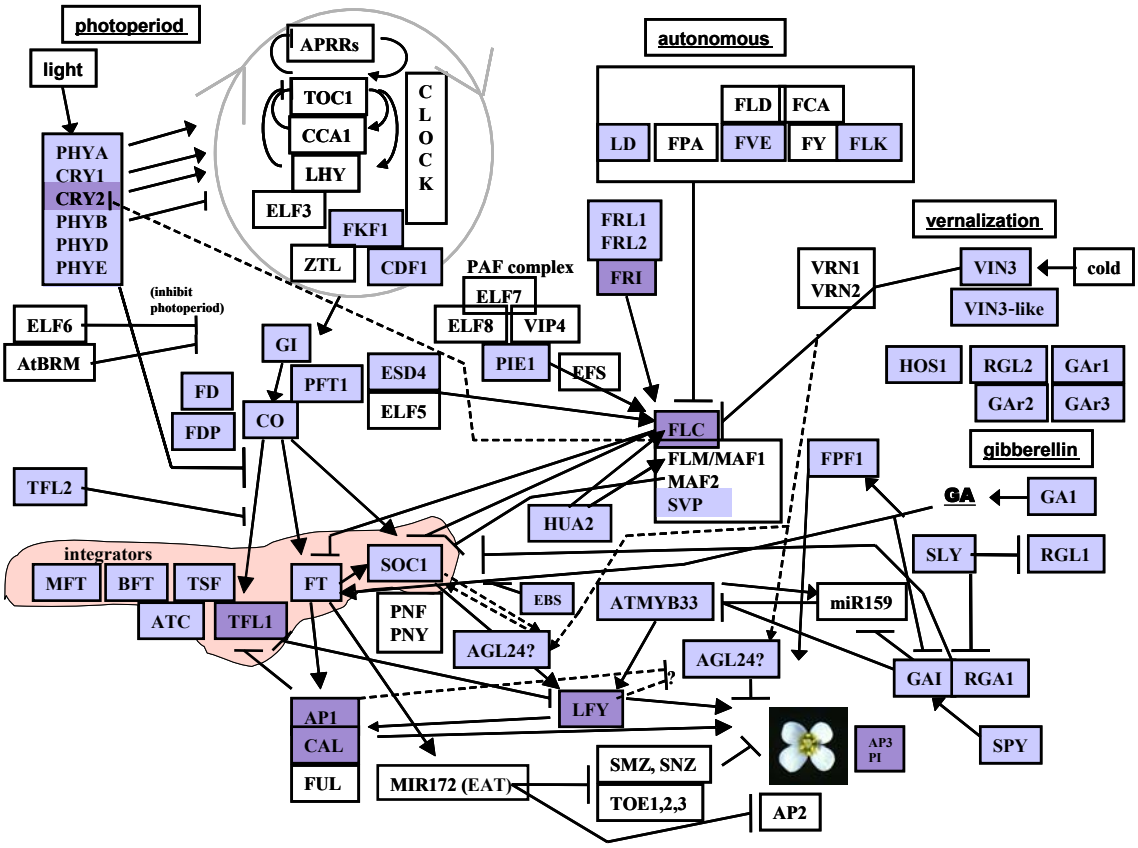


Figure S1. The known flowering time genetic network

Numerous genes have been characterized with a role in the flowering of *A. thaliana* based on forward genetic screens. The interactions of these genes are known in many cases based on genetic interaction studies, as shown here in this literature-based network (Judy Roe, Kansas State University, unpublished). Genes in blue or purple were included in this association study. Genes in blue have soon-to-be published resequencing data, whereas genes in purple have published resequencing data.

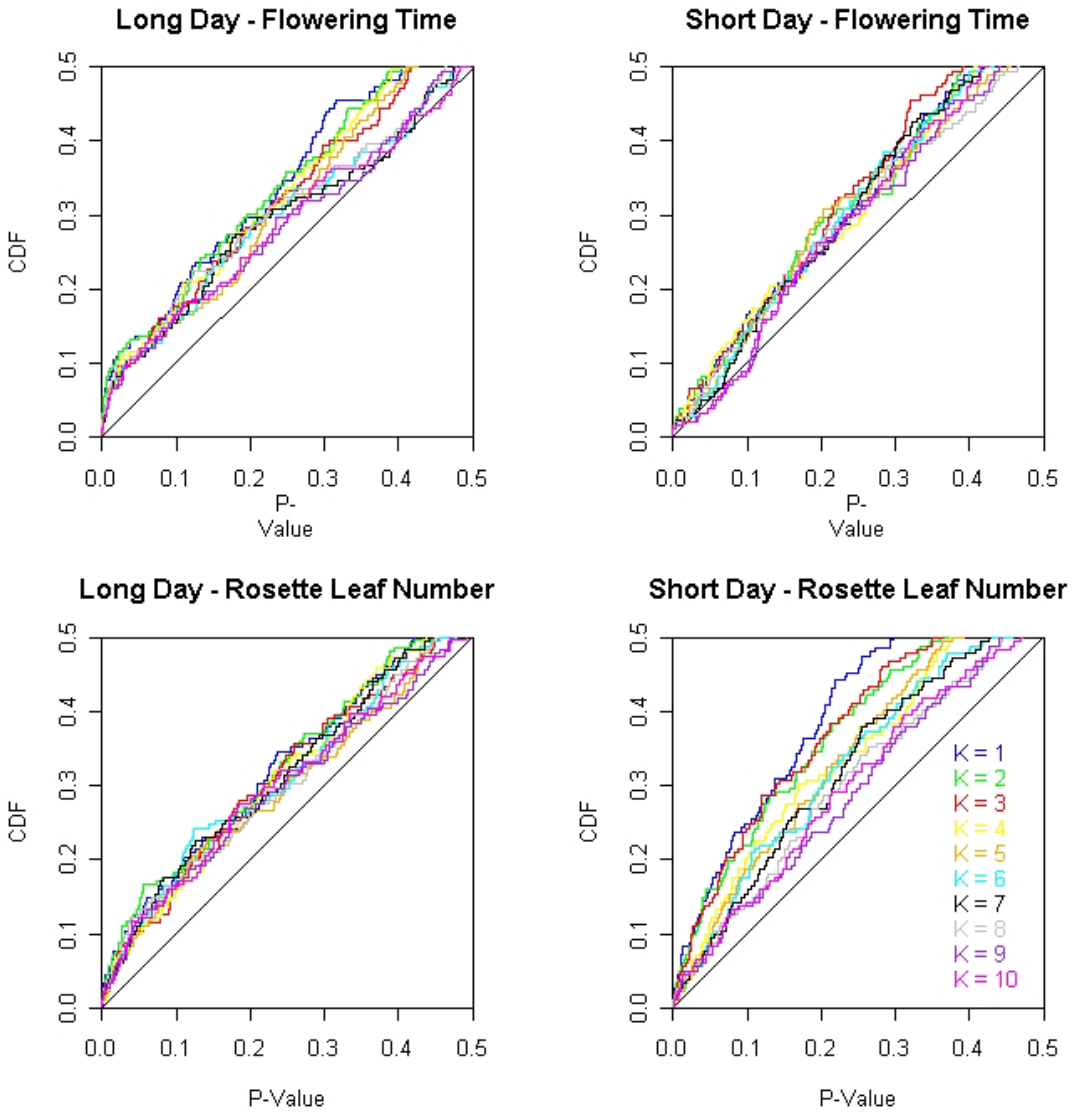


Figure S2. Cdfs for the 2010 Background loci using structured association models

Results for models of the form $\mathbf{Y} = \mathbf{X\alpha} + \mathbf{Q\beta} + \boldsymbol{\varepsilon}$ are presented going from $K = 1$, which is the naïve association, to $K = 10$. The axes are restricted to a maximum of 0.5 to facilitate comparison of the different models.

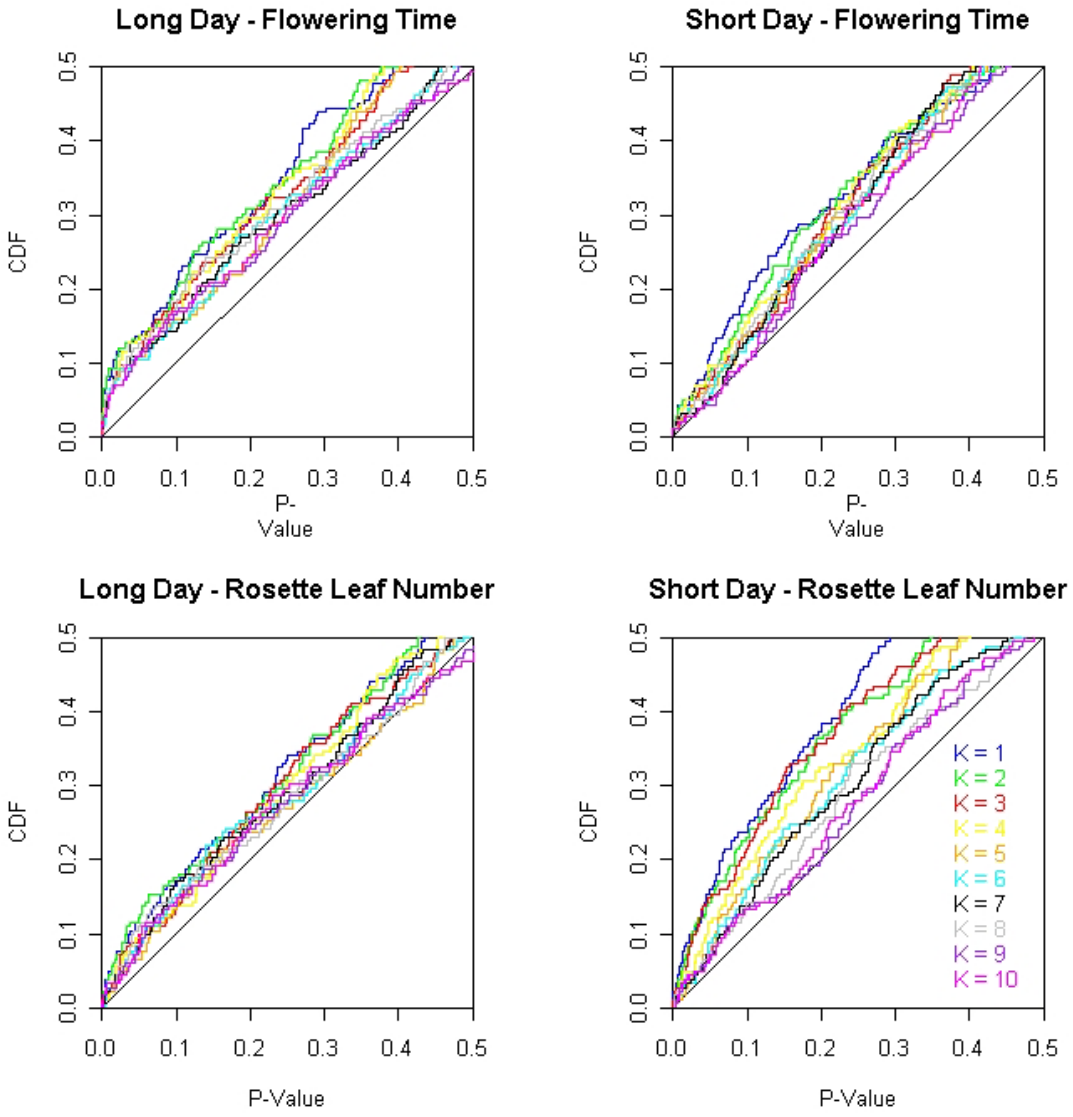


Figure S3. Cdfs for the 2010 Background loci using mixed models

Results for models of the form $\mathbf{Y} = \mathbf{X\alpha} + \mathbf{Q\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$ are presented going from $K = 1$, which is the mixed model with **K** only, to $K = 10$. The axes are restricted to a maximum of 0.5 to facilitate comparison of the different models.

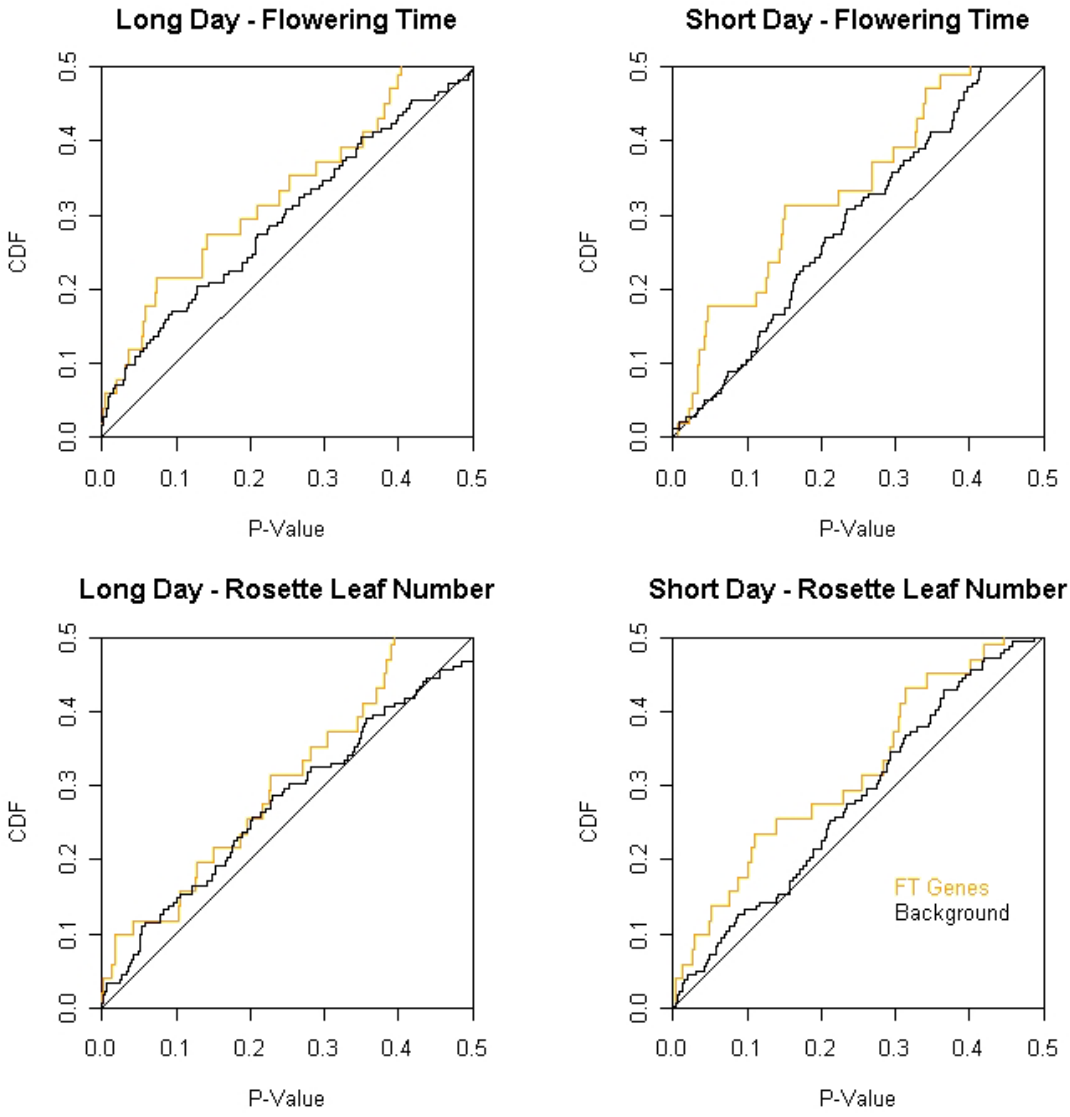


Figure S4. Cdfs for the 2010 Background loci and candidate gene cores using the K + Q_{10} mixed model

The axes are restricted to a maximum of 0.5 to facilitate comparison of the different models.

CHAPTER FOUR:

Sequence variation of miRNAs and their binding sites in *Arabidopsis thaliana*

Sequence variation of miRNAs and their binding sites in *Arabidopsis thaliana*

Ian M. Ehrenreich[1,2] and Michael D. Purugganan[2]

[1]Department of Genetics, Box 7614, North Carolina State University, Raleigh, North Carolina 27695 USA
[2]Department of Biology and Center for Genomics and Systems Biology, New York University, 100 Washington Square East, New York, New York 10003 USA

Corresponding author: Ian M. Ehrenreich, Telephone: (212) 998-8465, Email: ehrenreich@ncsu.edu.

Contributions: IME and MDP designed this study and cowrote this manuscript.

This chapter is a modified version of previously published work.

Reference: Ehrenreich, I.M. and M.D. Purugganan. 2008. Sequence variation of miRNAs and their binding sites in *Arabidopsis thaliana*. *Plant Physiology* 146:1974-1982.

Abstract

Major differences exist between plants and animals both in the extent of miRNA-based gene regulation and the sequence complementarity requirements for miRNA-mRNA pairing. Whether these differences affect how these sites evolve at the molecular level is unknown. To determine the extent of sequence variation at miRNAs and their targets in a plant species, we resequenced 16 miRNA families (66 miRNAs in total) and all 52 of the characterized binding sites for these miRNAs in the plant model *Arabidopsis thaliana*, accounting for around 50 percent of the known miRNAs and binding sites in this species. As has been shown previously in humans, we find that both miRNAs and their target binding sites have very low nucleotide variation and divergence compared to their flanking sequences in *A. thaliana*, indicating strong purifying selection on these sites in this species. Sequence data flanking the mature miRNAs, however, exhibit normal levels of polymorphism for the accessions in this study and, in some cases, non-neutral evolution or subtle effects on predicted pre-miRNA secondary structure, suggesting that there is raw material for the differential function of miRNA alleles. Overall, our results show that despite differences in the architecture of miRNA-based regulation, miRNAs and their targets are similarly constrained in both plants and animals.

Introduction

Changes in gene regulation have long been thought to be important to evolutionary diversification (KING and WILSON 1975). Extensive variation in gene expression has been documented both within and across many species (e.g. in primates (ENARD *et al.* 2002; MORLEY *et al.* 2004; WHITNEY *et al.* 2003), *Fundulus* (OLEKSIAK *et al.* 2002; OLEKSIAK *et al.* 2005), and *Drosophila* (JIN *et al.* 2001; RIFKIN *et al.* 2003)), though in most cases the regulatory mechanisms and phenotypic consequences of this diversity are unknown. Although changes in transcriptional regulation are likely a major source of gene expression

diversity, variability in posttranscriptional regulation may also contribute (CHEN and RAJEWSKY 2007).

MicroRNAs (miRNAs) are small RNAs ~21 nucleotides long with complementarity to specific regions in messenger RNAs (mRNAs) and are important posttranscriptional regulators of gene expression in eukaryotes (CARRINGTON and AMBROS 2003). Binding of miRNAs to target mRNAs triggers the cleavage, translational repression, or deadenylation of these targets (ZHANG *et al.* 2007). Once transcribed, miRNAs are processed into small, stem-loop precursors (pre-miRNAs) that are further processed by RNaseIII-type endonucleases (Drosha and Dicer in animals, and Dicer-like 1 [DCL1] in plants) and methylated to form mature miRNAs (JONES-RHOADES *et al.* 2006). Mature miRNAs join with Argonaute to form RNA-induced silencing complexes (RISCs) that can subsequently target specific mRNAs (JONES-RHOADES *et al.* 2006).

The number of miRNAs per eukaryotic genome varies by species. For instance, miRBase presently lists 114, 117, and 326 miRNA genes in *Arabidopsis thaliana*, *Caenorhabditis elegans,* and humans, respectively (GRIFFITHS-JONES 2004; GRIFFITHS-JONES *et al.* 2006). Whereas animals have many unique mature miRNA sequences and small miRNA families, the situation is reversed in plants, which typically have a small number of unique miRNA sequences and large miRNA families (LI and MAO 2007). Though many miRNAs have been identified across eukaryotes, a recent high-throughput sequencing experiment suggests that additional, lowly expressed miRNAs may exist that have escaped previous molecular and bioinformatics approaches (FAHLGREN *et al.* 2007).

The total number of genes targeted by miRNAs is also highly variable and genome-specific. Only 1% of all protein coding genes appear to be miRNA targets in *A. thaliana* (JONES-RHOADES and BARTEL 2004; RHOADES *et al.* 2002), while at least 20% of all protein coding genes are likely miRNA targets in animals (GRUN *et al.* 2005; KREK *et al.* 2005; LALL *et al.* 2006; LEWIS *et al.* 2005; LEWIS *et al.* 2003). Furthermore, differences exist between

plants and animals in the number of genes targeted by each miRNA. For example, while each *Drosophila* miRNA has on average over 50 predicted targets (GRUN *et al.* 2005), most *A. thaliana* miRNAs have six targets or fewer (JONES-RHOADES *et al.* 2006).

Differences in miRNA function also exist between plants and animals. In animals, the complementarity between the first six to eight bases of a target to a miRNA are most important to binding (RAJEWSKY 2006). In contrast, plants require, with exception, complementarity across the entire miRNA and binding target (SCHWAB *et al.* 2006; SCHWAB *et al.* 2005). Also, while in animals miRNA binding sites are almost exclusively within 3' untranslated regions (UTRs), most plant miRNAs have binding sites within coding exons (CHEN and RAJEWSKY 2007).

The extent to which miRNAs contribute to phenotypic evolution is unclear, but evidence suggests they could play an important role. While essential miRNA-target site interactions have been conserved for > 400 million years in plants and animals (e.g. miR165/166 and Class III HD-ZIP genes in land plants (FLOYD and BOWMAN 2004) and the *let-7* miRNA and the *lin-41* mRNA in metazoans (PASQUINELLI *et al.* 2000)), others appear to be species or clade-specific and may function in evolutionarily-derived traits (AXTELL and BARTEL 2005; BONNET *et al.* 2004). Little is known about the microevolution of miRNAs-target site interactions, though a few studies have documented functional polymorphisms at these sites. A single nucleotide polymorphism (SNP) that results in a *de novo* miRNA binding site has been shown to underlie a quantitative trait locus for muscularity in sheep (CLOP *et al.* 2006), while a SNP in an existing miRNA binding site may cause Tourette's syndrome in humans (ABELSON *et al.* 2005). Additionally, a SNP has been identified in human herpesvirus that affects Drosha processing of a miRNA precursor (GOTTWEIN *et al.* 2006). Genomewide surveys of miRNA and miRNA binding site polymorphism have only been conducted in humans. These studies have shown levels of polymorphism at miRNAs and their targets are lower than at coding or neutral regions, and the mutations at these sites exhibit a general signature of purifying selection (CHEN and RAJEWSKY 2006; SAUNDERS *et*

al. 2007). One of these studies has also shown evidence for positive selection on some miRNA binding sites based on long range haplotype signatures (SAUNDERS *et al.* 2007), implying that beneficial miRNA target site polymorphisms may exist.

Predictions can be made about the expected level of miRNA and miRNA binding site sequence variation in plants relative to humans based on the functional differences between plant and animal miRNAs that are described above. Plant miRNAs typically have fewer mRNA targets and more miRNA family members than animal miRNAs, which may lead to reduced constraint on and higher sequence diversity in miRNA sequences in plants. However, as most plant miRNAs perform important functions in development and physiology and spatiotemporal functional differences may exist among plant miRNA family members making each independently essential, constraint on plant miRNAs may parallel that observed in humans. As for miRNA binding sites, constraint is likely to be strong across the entire miRNA binding site in plants due to the importance of the entire binding site in miRNA-mRNA pairing. Additionally, plant miRNA binding sites may experience additional constraint due to the presence of these sites largely in coding exons in plants.

We assess the levels and patterns of nucleotide polymorphism in miRNAs and their binding sites in the model plant *A. thaliana* by resequencing more than half of the characterized miRNAs and binding targets in this species from 24 diverse accessions. We find significantly reduced genetic variation at these sites relative to flanking sequence, with only four SNPs and an indel present in our sample. However, we do find substantial variation flanking miRNAs both within *A. thaliana* and between it and the closely related outgroup *A. lyrata*. Interestingly, four miRNAs exhibit non-neutral patterns of molecular variation and numerous SNPs are predicted to have subtle effects on pre-miRNA secondary structure. Our results suggest that mutations within mature miRNAs and their binding sites do not contribute substantially to gene expression and phenotypic variation in this model plant species, but that ample variation flanks mature miRNAs that could contribute to the evolutionary diversification of these key regulatory genes.

Results

Single nucleotide polymorphisms (SNPs) and nucleotide divergence in A. thaliana miRNAs. We investigated the sequence variation of 66 miRNAs belonging to 16 miRNA families, as well as 52 mRNA binding site targets that represented all the validated targets for these miRNAs (Table 1). On average, we resequenced four miRNAs per family and three target sites per miRNA family in a set of 24 accessions (Supplemental Table 1). These miRNAs were selected because (i) their interactions with mRNA targets have been functionally characterized and/or (ii) they target transcripts of genes with known roles in development. Altogether, these comprise over 55 percent of the presently described miRNAs and 40 percent of the validated binding sites in *A. thaliana* (based on data in (JONES-RHOADES *et al.* 2006)).

For each miRNA, we sequenced on average 489 bps, with ~ 133 bps of pre-miRNA (based on pre-miRNA predictions in miRBase), as well as about 180 and 176 bps of upstream and downstream flanking sequence, respectively. The average level of single nucleotide polymorphism per site (θ) (WATTERSON 1975) at these miRNAs is 0.0004 ± 0.0003 (mean ± standard error of the mean), or one SNP every 2.5 kb of miRNA sequence. Underscoring this low sequence diversity, only two miRNAs, miR156d and miR395f, were actually found to be polymorphic. The microRNA miR156d has a single SNP segregating at 10 percent frequency (found in both the Ei-2 and Ll-0 accessions), while there is a single SNP in miR395f found in the Cvi-0 accession (4 percent frequency) [Figure 1]. The miR156d SNP is at miRNA-mRNA mismatch position and is unlikely to affect binding, while the miR395f minor allele disrupts a complementary position relative to the major allele (Figure 1). Sequence comparisons with the closely related outgroup *A. lyrata* show that both of these polymorphisms are derived within *A. thaliana*. No fixed differences exist between *A. thaliana* and *A. lyrata* at the examined miRNAs (i.e. $K = 0$). Levels of both nucleotide polymorphism and divergence at these sites are significantly below background levels of $\theta =$

0.0055 ± 0.0002 and K = 0.085 ± 0.005 as assessed using 1,213 previously resequenced genomewide loci (Wilcoxon Rank Sum test, $P < 0.0001$ for both polymorphism and divergence).

We also estimated levels of nucleotide diversity for the sequences flanking the miRNAs, including the pre-miRNAs and the upstream and downstream flanking sequence. Nucleotide polymorphism levels at these sites are substantially higher than those observed in the miRNAs themselves, with a mean $\theta = 0.0025 \pm 0.0003$ for the pre-miRNA, and mean θ = 0.0051 ± 0.0006 and 0.0054 ± 0.0006 for the upstream and downstream flanking sequences, respectively (Figure 2). Although no insertion/deletion polymorphisms (indels) were observed in the mature miRNA sequences, numerous indels were detected in the pre-miRNAs and flanking sites (0.7 per kb at pre-miRNAs, 1.7 per kb for upstream sequence and 2 per kb for downstream flanking sequence). Levels of nucleotide divergence are also higher at these sites, with mean K = 0.052 ± 0.007 for pre-miRNA, 0.11 ± 0.014 for upstream sequence, and 0.2 ± 0.026 at downstream sites (Figure 2). The dramatically reduced intraspecific polymorphism and interspecific divergence at mature miRNAs, and to a lesser extent, pre-miRNAs suggests that purifying selection is the predominant evolutionary force that acts on miRNAs in *A. thaliana*.

Levels and patterns of nucleotide polymorphism and divergence in miRNA target binding sites. We also estimated polymorphism and divergence at the target binding sites of the miRNAs we examined (Table 1). In all cases, these sites had been previously validated as the target sites of specific miRNAs (see (JONES-RHOADES *et al.* 2006) for references). Of these binding sites, 47 are in exons, two are in 5' UTRs and three are in 3' UTRs. Six of the exonic binding sites and one of the 5' UTR binding sites are interrupted by introns and, consequently, require splicing to bind with their complementary miRNAs. For each binding site, we sequenced on average 476 bps, with the miRNA binding site at the center of the sequenced region.

Like their cognate miRNAs, we also observe significantly low polymorphism levels at the miRNA binding sites relative to background polymorphism, with mean nucleotide diversity equal to 0.0005 ± 0.0003 (Wilcoxon Rank Sum test, $P < 0.0001$). Only two binding sites of the 52 we studied – in the *AUXIN SIGNALING F-BOX 1* (*AFB1*) and *TARGET OF EAT 3* (*TOE3*) genes – are polymorphic (Figure 1). The *AFB1* binding site, which is targeted by miR393, has a single SNP segregating at 12 percent frequency (in the Edi-0, Ga-0 and Ll-0 accessions). The *AFB1* minor allele converts a miRNA-mRNA match position to a mismatch position relative to the major allele (Figure 1). *TOE3* is targeted by miR172 and this binding site has a seven bp deletion and a SNP that co-segregate at 4 percent frequency in our sample (found in the Gy-0 accession). The *TOE3* binding site deletion, however, is partially recovered in the mRNA due to upstream sequence similarity, resulting in only a single bp deletion and a SNP in the mature transcript (Figure 1). Although the low-frequency *TOE3* polymorphisms are derived mutations, the derived mutation at *AFB1* is the common SNP allele. Nucleotide divergence at target binding sites is $K = 0.003 \pm 0.002$, which is significantly lower than the genomewide average (Wilcoxon Rank Sum test, $P < 0.0001$). Only one binding site – in *AUXIN RESPONSE FACTOR 10 (ARF10)*, which is targeted by miR160 – exhibits a fixed sequence differences between species. This substitution occurs at a mismatch position in the miRNA-mRNA pairing sequence (Figure 3).

Levels of nucleotide variation were also reduced at miRNA binding sites relative to their flanking sequences (mean $\theta = 0.0028 \pm 0.0005$ and 0.0033 ± 0.0005 for upstream and downstream flanking sequences, respectively) (see Figure 4). These flanking nucleotide diversity values are low in comparison to data surrounding mature miRNAs, and are likely due to the location of many of these binding sites in coding exons. To correct for this, we also calculated silent site nucleotide diversity (θ_{silent}) for miRNA binding sites and their flanking sequences. These estimates ($\theta_{silent} = 0.0015 \pm 0.0007$, 0.0069 ± 0.0026, and 0.0053 ± 0.0009 for miRNA binding sites, upstream, and downstream sequence, respectively) are higher than those for uncorrected nucleotide diversity estimates. The relative levels of variation across the site classes remain similar to uncorrected values, however, as nucleotide

diversity at the binding sites is still much lower than at flanking sites. Additionally, divergence was much lower at binding sites than at upstream (K = 0.062 ± 0.005) or downstream (K = 0.067 ± 0.001) sites (Figure 4).

Summary statistics of the nucleotide site-frequency spectrum. Although miRNAs and their target binding sites have little variation, we observe normal levels of sequence variation in regions flanking miRNAs. Selection on these polymorphisms or those linked to them could generate non-neutral patterns of linked sequence variation. To examine this possibility, we calculated Tajima's D (TAJIMA 1989) and Fay and Wu's H (FAY and WU 2000) for the entire sequence fragment containing each miRNA, as well as for the Nordborg et al. (2005) genomewide fragments. Using the Nordborg data as an empirical distribution, these tests identify four fragments that possess extreme values for either Tajima's D or Fay and Wu's H. MiR393a has a high Tajima's D value (D = 3.39; empirical $P < .001$) [Figure 5]. Significant Tajima's D values can be indicative of balancing selection or extreme population stratification at a locus. MiR166f, miR167d, and miR395c have low Fay and Wu's H values (H = -10.03, -11.44, and -7.22 for miR166f, miR167d, and miR395c, respectively; empirical $P < .05$ for each miRNA) [Figure 5] when compared to the empirical distribution, which can be indicative of positive selection. These results suggest that polymorphisms at or linked to miRNA genes may be targets of selection, though these findings must be regarded cautiously given the complex demography and pattern of linkage disequilibrium (LD) of this species (NORDBORG *et al.* 2005; SCHMID *et al.* 2005).

Secondary structure predictions of pre-miRNA haplotypes using biologically relevant temperatures. To evaluate the possible impacts of SNPs on pre-miRNA secondary structure, we computationally predicted the secondary structure and Gibbs free energy (ΔG) of the pre-miRNA from each observed haplotype using the mfold program (WALTER *et al.* 1994; ZUKER 2003). Due to both seasonal variation in temperature and geographic differences in climate, *A. thaliana* experiences a broad range of temperatures in the wild that may be important to consider when predicting secondary structures for organic macromolecules. We

selected for analysis two temperatures – 5 and 20° Celsius (C) – that are tolerable extremes of the range of temperatures that *A. thaliana* experiences naturally.

Of the 66 miRNAs we looked at overall, only 35 had SNPs segregating in their predicted pre-miRNAs; 62 total SNPs were identified across these pre-miRNAs. Using the predicted pre-miRNA secondary structure for the Col-0 haplotype of each miRNA, we determined the structural context of each SNP within its respective pre-miRNA. The vast majority of the SNPs were located in double-stranded stem regions – 40 in the general stem and 7 in the miRNA or miRNA* (Figure 6). Of the remaining 15 SNPs, 9 were located in the primary loop at the top of the pre-miRNA stem-loop molecule and the remaining 6 were in secondary loops occurring along the stem of the molecule (Figure 6).

We predicted pre-miRNA secondary structure at both 5 and 20° C, which represent a sampling of the temperature extremes *A. thaliana* might be expected to experience during its lifecycle. Of the pre-miRNA SNPs, 33 (53%) were predicted to alter pre-miRNA secondary structure at both temperatures relative to the Col-0 pre-miRNA allele (Figure 6). All predicted secondary structure changes were subtle (i.e. addition or subtraction of small loops along the stem; two nucleotide enlargement or shrinking of primary or secondary stem loops) and appeared to maintain the general integrity of the pre-miRNA stem-loop molecule. SNPs disrupting secondary structure occurred in all structural contexts of pre-miRNAs (Figure 6). For 26 SNPs (42% of all pre-miRNA SNPs), pre-miRNA secondary structure was entirely maintained across pre-miRNA alleles. Ten of these SNPs had no structural effect because they occurred within loops. The other 16 SNPs that did not affect secondary structure were all located along the pre-miRNA stem and fell into one of five classes (SNP counts in parentheses): i) occurring within mismatch positions (2 SNPs), ii) creating a non-disruptive mismatch from a match (7 SNPs), iii) creating a match from a non-disruptive mismatch (1 SNP), iv) a purine transition (A ↔ G) with the pairing base a U (3 SNPs), and v) a pyrimidine transition (C ↔ U) with the pairing base a G (3 SNPs). 3 SNPs (5% of all SNPs) had predicted structural effects at 5° C, but not at 20° C. These SNPs occur in a loop in

miR156d, at an A ↔ G with U pairing site (class iv) in miR157c, and at a C ↔ U with G pairing site (class v) in miR164a.

To more quantitatively assess the effects of SNPs on pre-miRNA stability, we next measured ΔG for each pre-miRNA haplotype and calculated the difference in ΔG (ΔΔG) between the Col-0 allele and the non-Col-0 alleles. For loci with more than two alleles, the mean ΔΔG was calculated across all values for the locus. 91% of the loci (32 of 35) had a mean ΔΔG that fell within the range of -6 to 4 kcal/mol at both temperatures (Figure 7). On average, mean ΔΔG was -1.64 kcal/mol and -1.53 kcal/mol at 5 and 20° C, respectively, suggesting that most SNPs destabilize pre-miRNAs in relation to the Col-0 allele. Interestingly, temperature has a clear effect on ΔΔG as values are more dispersed around the mean at 5° C relative to 20° C, suggesting that the destabilizing effects of polymorphisms on RNA secondary structure are enhanced by cold temperature.

Discussion

Posttranscriptional regulation of gene expression is a common phenomenon across eukaryotes, but the extent to which variability in this process contributes to diversity in gene expression and phenotype is unclear (CHEN and RAJEWSKY 2007). Examples in humans (ABELSON *et al.* 2005), sheep (CLOP *et al.* 2006), and herpesvirus (GOTTWEIN *et al.* 2006) suggest that functional polymorphisms at miRNAs and miRNA binding sites do exist. Genomic surveys of polymorphism at miRNAs and their targets can determine the prevalence of such variants across a species. To date, genomewide levels of nucleotide variation at miRNAs and miRNA binding sites have only been assessed in humans (CHEN and RAJEWSKY 2006; SAUNDERS *et al.* 2007), and the low levels of polymorphism at these sites indicates strong purifying selection on these regulatory RNAs and their targets.

MiRNAs in the model plant *A. thaliana* have been implicated in several developmental processes, including flowering time (ACHARD *et al.* 2004; AUKERMAN and

SAKAI 2003), juvenile/adult transition (WU and POETHIG 2006), leaf shape (NIKOVICS *et al.* 2006), and adaxial/abaxial polarity (EMERY *et al.* 2003). Our results indicate that these *A. thaliana* miRNAs and their binding sites evolve under strong sequence constraint. Indeed, in all these genes, only two miRNA SNPs and two target site SNPs and an indel were detected in our sample, with most of these polymorphisms being low frequency derived polymorphisms. Additionally, only one substitutional difference exists between *A. thaliana* and *A. lyrata* at these sites.

Overall, our results support that, like in humans, the predominant force acting on *A. thaliana* miRNAs and their targets is purifying selection. Despite the presence of more copies of each mature miRNA sequence in the *A. thaliana* genome than in the human genome and the smaller number of mRNA targets per miRNA in plants, plant miRNAs exhibit very strong purifying selection comparable to that observed in humans (SAUNDERS *et al.* 2007). The prediction that miRNA sequences should be conserved across their entirety in plants, which is contrary to patterns of sequence variation in human miRNAs, does hold true. This corroborates the findings of molecular biology experiments that document plants' requirements for sequence complementarity across the miRNA-mRNA pairing sequence (SCHWAB *et al.* 2006; SCHWAB *et al.* 2005). Unlike in humans (SAUNDERS *et al.* 2007), no moderate frequency binding site polymorphisms segregate in our sample, suggesting that miRNA binding site variation is unlikely to contribute to phenotypic diversity in *A. thaliana.*

The degree of purifying selection on miRNAs and their target binding sites can be assessed by comparing levels of variation at miRNAs and their targets to levels of amino acid-changing variation in protein coding genes. The mean level of nonsynonymous polymorphism (θ_{nsyn}) for the miRNA target gene exons resequenced in this study is 0.002 ± 0.0003, which is over four-fold higher than nucleotide diversity values at miRNAs and their binding sites. This is also observed at the interspecific level; the mean rate of nonsynonymous substitution (K_a) between *A. thaliana* and *A. lyrata* is 0.026 ± 0.003, which is nearly thirty-fold higher than mean *K* for miRNAs and approximately ten-fold higher than

K values at miRNA binding sites. These comparisons indicate that purifying selection on miRNAs and their binding sites is stronger than it is for amino acid changes in protein-coding genes.

The strong sequence constraint of miRNAs and their binding sites suggest that evolutionary changes in these sequences are unlikely to be major contributors to natural variation in *A. thaliana*. We have, however, identified a small number of rare miRNA and target site polymorphisms that may have functional effects, and have shown that substantial flanking variation exists both within *A. thaliana* and between it and *A. lyrata*. Overall, our results imply that the roles of miRNA-target interactions in plant function are essential and are subject to strong purifying selection, but that variation flanking these sites could contribute to regulatory diversity at these genes and their downstream targets.

Empirical and computational approaches have shown that pre-miRNA secondary structure is important to the processing and maturation of miRNAs (RITCHIE *et al.* 2007; ZENG *et al.* 2005). Our analyses of the effects of SNPs on predicted pre-miRNA secondary structure suggests that another element contributing to sequence constraint at and near miRNAs may be selection for the maintenance of the pre-miRNA stem-loop. 42% of all detected SNPs were predicted to have no effect on pre-miRNA secondary structure. Of the numerous SNPs that were identified with structural effects, all had subtle effects, maintaining the general integrity of the stem-loop molecule. These results together imply that purifying selection culls mutations with strong effects on pre-miRNA secondary structure. This finding is similar to results from diverse organisms, such as bacteria (KATZ and BURGE 2003), flies (KIRBY *et al.* 1995), and mammals (CHAMARY and HURST 2005), showing constraint on other types of RNA molecules.

Of note is that most detected pre-miRNA polymorphisms appear to have a destabilizing effect on RNA secondary structure since the majority of the studied loci have non-zero $\Delta\Delta G$ values. These destabilizing effects appear to be partially mediated by

temperature, a point supported both by the increased dispersion of ΔΔG at 5° C relative to 20° C and the identification of three polymorphisms with structural effects at 5° C and not 20° C. These findings suggest that the use of biologically relevant temperature, which in *A. thaliana* represents the range of environmental temperatures a plant might experience during a growing season, may be an important consideration when predicting RNA or protein secondary structure. Indeed, temperature could mediate gene regulation in nature through its effects on secondary structure.

MiRNAs comprise a key class of regulatory loci in eukaryotic systems, and we are beginning to understand the evolutionary forces that govern the diversification of these genes. Our work suggests that despite fundamental differences in miRNA-based regulation, miRNAs and their targets are similarly constrained in both plants and animals. The possibility of variation in cis-regulation or processing of miRNAs in A. thaliana and other species merits further attention, though documenting such functional variation, if it exists, will be technically challenging due to the presence of multiple copies of many mature miRNAs. We have shown, however, that the raw material for such variation does exist in A. thaliana and that it may be responsive to temperature, laying the groundwork for future experiments focused on potential molecular functional variation at miRNAs in A. thaliana.

Materials and methods

PCR and DNA sequencing. MiRNAs and binding targets were chosen as a subset of those listed in (JONES-RHOADES *et al.* 2006). MiRNA precursors were determined based on references provided at the miRBase website (http://microrna.sanger.ac.uk/). The *A. thaliana* accessions used in this study were chosen to span the geographic range of the species. All primers were designed from the Columbia-0 (Col-0) genome sequence using Primer3 (ROZEN and SKALETSKY 2000). Primer pairs were designed to amplify products between 400 and 600 bps in length. The amplified regions were centered on the miRNA precursor listed in miRBase or the miRNA binding site. All PCR primers were blasted against the Col-0

genome sequence on the TAIR website (The Arabidopsis Information Resource – www.arabidopsis.org) to ensure that only the targeted genomic region would be amplified. PCR and sequencing was done as previously described (OLSEN *et al.* 2006) by Cogenics (New Haven, CT). On average, 23 individuals were successfully sequenced per miRNA and miRNA binding site.

Sequence analysis. Sequences were initially aligned and edited using the Phred and Phrap programs (Codon Code, Dedham, MA) and BioLign Version 2.09.1 (Tom Hall, Ibis Therapeutics, Carlsbad, CA). Additional manual alignment and polymorphism identification were conducted in BioEdit Version 7.0.5 (Tom Hall, Ibis Therapeutics, Carlsbad, CA). Reported summary statistics were calculated in either Microsoft Excel (Redmond, WA), DnaSP Version 4.1.0 (ROZAS *et al.* 2003), or Variscan (VILELLA *et al.* 2005). Site classifications (i.e. as miRNA, pre-miRNA, upstream, or downstream) were made based on information in miRBase. Nucleotide diversity was calculated as θ, the population mutation rate per locus based on the number of segregating sites (WATTERSON 1975). Nucleotide substitution rates (K) were calculated in Variscan based on the Jukes-Cantor model. Silent site variation was calculated based on all mutations that did not affect the amino acid sequence of the protein encoded by a target mRNA, including intronic and UTR sequence. The empirical distribution for Tajima's D (TAJIMA 1989), as well as the background level of polymorphism for the accessions in this study, was generated using 1,213 previously published, genomewide resequencing fragments (NORDBORG *et al.* 2005). Tajima's D was calculated since it can be useful in detecting both positive selection and balancing selection. We also calculated Fay and Wu's H (FAY and WU 2000) because it previously was shown to be less biased by demography in *A. thaliana* than other tests evaluating the site frequency spectrum (SCHMID *et al.* 2005). *A. lyrata* sequence data was obtained for all fragments in this study, as well as for 100 randomly selected fragments from the Nordborg et al. (2005) study to generate the background distribution of K and Fay and Wu's H. The *A. lyrata* data was acquired by using the Trace Archive database Mega BLAST search function at NCBI (http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml). The top one or two hits per fragment

were assembled and aligned to the *A. thaliana* multiple alignments. No significant hits were found corresponding to miR395a, although there was a region of sequence similarity found for the upstream region of this fragment. Any sites initially found to be diverged between *A. thaliana* and *A. lyrata* at the miRNA binding sites were re-examined by additional BLASTing. It should be noted that the study of Nordborg et al. (2005) used 96 accessions, 24 of which are included in this study. To account for this, we only used sequence data corresponding to the accessions in this study to generate empirical distributions. Wilcoxon Rank Sum tests, ANOVAs, and regressions were conducted in JMP Version 5 (SAS, Cary, NC).

Secondary structure prediction. The program mfold v2.3 was used to predict the pre-miRNA secondary structure and the ΔG for each naturally occurring pre-miRNA haplotype (excluding those differentiated from the Col-0 allele by indels). Comparison of SNP locations to the predicted Col-0 structure at 5° C was used to identify the structural context of each SNP. In cases where multiple structures were predicted for a particular pre-miRNA haplotype, ΔG was calculated as the average of all these predictions. ΔΔG was then calculated for each locus by subtracting the non-Col-0 allele's ΔG from the Col-0 allele's ΔG. Since the number of haplotypes per pre-miRNA was variable, we calculated a mean ΔΔG per pre-miRNA, which was simply the average of all ΔΔG values for that locus.

Acknowledgments

We thank Daisuke Saisho and members of the Purugganan laboratory for assistance with this project and manuscript. We also thank Kevin Chen for reading of a draft of this manuscript. This work was supported by a Department of Education Graduate Assistance in Areas of National Need Fellowship and a National Science Foundation (NSF) Graduate Research Fellowship to I. M. E., and by grants from the NSF's Frontiers in Integrated Biological Research and Plant Genome Research Programs to M. D. P.

References

ABELSON, J. F., K. Y. KWAN, B. J. O'ROAK, D. Y. BAEK, A. A. STILLMAN *et al.*, 2005 Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science 310: 317-320.

ACHARD, P., A. HERR, D. C. BAULCOMBE and N. P. HARBERD, 2004 Modulation of floral development by a gibberellin-regulated microRNA. Development 131: 3357-3365.

AUKERMAN, M. J., and H. SAKAI, 2003 Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. Plant Cell 15: 2730-2741.

AXTELL, M. J., and D. P. BARTEL, 2005 Antiquity of microRNAs and their targets in land plants. Plant Cell 17: 1658-1673.

BONNET, E., J. WUYTS, P. ROUZE and Y. VAN DE PEER, 2004 Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. Proc Natl Acad Sci U S A 101: 11511-11516.

CARRINGTON, J. C., and V. AMBROS, 2003 Role of microRNAs in plant and animal development. Science 301: 336-338.

CHAMARY, J. V., and L. D. HURST, 2005 Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6: R75.

CHEN, K., and N. RAJEWSKY, 2006 Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet 38: 1452-1456.

CHEN, K., and N. RAJEWSKY, 2007 The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8: 93-103.

CLOP, A., F. MARCQ, H. TAKEDA, D. PIROTTIN, X. TORDOIR *et al.*, 2006 A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet 38: 813-818.

EMERY, J. F., S. K. FLOYD, J. ALVAREZ, Y. ESHED, N. P. HAWKER *et al.*, 2003 Radial patterning of Arabidopsis shoots by class III HD-ZIP and KANADI genes. Curr Biol 13: 1768-1774.

ENARD, W., P. KHAITOVICH, J. KLOSE, S. ZOLLNER, F. HEISSIG *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science 296: 340-343.

FAHLGREN, N., M. D. HOWELL, K. D. KASSCHAU, E. J. CHAPMAN, C. M. SULLIVAN *et al.*, 2007 High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. PLoS ONE 2: e219.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

FLOYD, S. K., and J. L. BOWMAN, 2004 Gene regulation: ancient microRNA target sequences in plants. Nature 428: 485-486.

GOTTWEIN, E., X. CAI and B. R. CULLEN, 2006 A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects Drosha processing. J Virol 80: 5321-5326.

GRIFFITHS-JONES, S., 2004 The microRNA Registry. Nucleic Acids Res 32: D109-111.

GRIFFITHS-JONES, S., R. J. GROCOCK, S. VAN DONGEN, A. BATEMAN and A. J. ENRIGHT, 2006 miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34: D140-144.

GRUN, D., Y. L. WANG, D. LANGENBERGER, K. C. GUNSALUS and N. RAJEWSKY, 2005 microRNA target predictions across seven Drosophila species and comparison to mammalian targets. PLoS Comput Biol 1: e13.

JIN, W., R. M. RILEY, R. D. WOLFINGER, K. P. WHITE, G. PASSADOR-GURGEL *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 29: 389-395.

JONES-RHOADES, M. W., and D. P. BARTEL, 2004 Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell 14: 787-799.

JONES-RHOADES, M. W., D. P. BARTEL and B. BARTEL, 2006 MicroRNAs and their regulatory roles in plants. Annu Rev Plant Biol 57: 19-53.

KATZ, L., and C. B. BURGE, 2003 Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res 13: 2042-2051.

KING, M. C., and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science 188: 107-116.

KIRBY, D. A., S. V. MUSE and W. STEPHAN, 1995 Maintenance of pre-mRNA secondary structure by epistatic selection. Proc Natl Acad Sci U S A 92: 9047-9051.

KREK, A., D. GRUN, M. N. POY, R. WOLF, L. ROSENBERG *et al.*, 2005 Combinatorial microRNA target predictions. Nat Genet 37: 495-500.

LALL, S., D. GRUN, A. KREK, K. CHEN, Y. L. WANG *et al.*, 2006 A genomewide map of conserved microRNA targets in C. elegans. Curr Biol 16: 460-471.

LEWIS, B. P., C. B. BURGE and D. P. BARTEL, 2005 Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15-20.

LEWIS, B. P., I. H. SHIH, M. W. JONES-RHOADES, D. P. BARTEL and C. B. BURGE, 2003 Prediction of mammalian microRNA targets. Cell 115: 787-798.

LI, A., and L. MAO, 2007 Evolution of plant microRNA gene families. Cell Res 17: 212-218.

MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genomewide variation in human gene expression. Nature 430: 743-747.

NIKOVICS, K., T. BLEIN, A. PEAUCELLE, T. ISHIDA, H. MORIN *et al.*, 2006 The balance between the MIR164A and CUC2 genes controls leaf margin serration in Arabidopsis. Plant Cell 18: 2929-2945.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3: e196.

OLEKSIAK, M. F., G. A. CHURCHILL and D. L. CRAWFORD, 2002 Variation in gene expression within and among natural populations. Nat Genet 32: 261-266.

OLEKSIAK, M. F., J. L. ROACH and D. L. CRAWFORD, 2005 Natural variation in cardiac metabolism and gene expression in Fundulus heteroclitus. Nat Genet 37: 67-72.

OLSEN, K. M., A. L. CAICEDO, N. POLATO, A. MCCLUNG, S. MCCOUCH *et al.*, 2006 Selection under domestication: evidence for a sweep in the rice waxy genomic region. Genetics 173: 975-983.

PASQUINELLI, A. E., B. J. REINHART, F. SLACK, M. Q. MARTINDALE, M. I. KURODA *et al.*, 2000 Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408: 86-89.

RAJEWSKY, N., 2006 microRNA target predictions in animals. Nat Genet 38 Suppl: S8-13.

RHOADES, M. W., B. J. REINHART, L. P. LIM, C. B. BURGE, B. BARTEL *et al.*, 2002 Prediction of plant microRNA targets. Cell 110: 513-520.

RIFKIN, S. A., J. KIM and K. P. WHITE, 2003 Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet 33: 138-144.

RITCHIE, W., M. LEGENDRE and D. GAUTHERET, 2007 RNA stem-loops: to be or not to be cleaved by RNAse III. Rna 13: 457-462.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers, pp. 365-386 in *Bioinformatics Methods and Protocols: Methods for Molecular Biology*, edited by S. KRAWETZ and S. MISENER. Humana Press, Totowa, NJ.

SAUNDERS, M. A., H. LIANG and W. H. LI, 2007 Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci U S A 104: 3300-3305.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in Arabidopsis thaliana reveals a genomewide departure from a neutral model of DNA sequence polymorphism. Genetics 169: 1601-1615.

SCHWAB, R., S. OSSOWSKI, M. RIESTER, N. WARTHMANN and D. WEIGEL, 2006 Highly specific gene silencing by artificial microRNAs in Arabidopsis. Plant Cell 18: 1121-1133.

SCHWAB, R., J. F. PALATNIK, M. RIESTER, C. SCHOMMER, M. SCHMID *et al.*, 2005 Specific effects of microRNAs on the plant transcriptome. Dev Cell 8: 517-527.

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

VILELLA, A. J., A. BLANCO-GARCIA, S. HUTTER and J. ROZAS, 2005 VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21: 2791-2793.

WALTER, A. E., D. H. TURNER, J. KIM, M. H. LYTTLE, P. MULLER *et al.*, 1994 Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc Natl Acad Sci U S A 91: 9218-9222.

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

WHITNEY, A. R., M. DIEHN, S. J. POPPER, A. A. ALIZADEH, J. C. BOLDRICK *et al.*, 2003 Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 100: 1896-1901.

WU, G., and R. S. POETHIG, 2006 Temporal regulation of shoot development in Arabidopsis thaliana by miR156 and its target SPL3. Development 133: 3539-3547.

ZENG, Y., R. YI and B. R. CULLEN, 2005 Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. Embo J 24: 138-148.

ZHANG, B., Q. WANG and X. PAN, 2007 MicroRNAs and their regulatory roles in animals and plants. J Cell Physiol 210: 279-289.

ZUKER, M., 2003 Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406-3415.

Table 1. MiRNAs and targets included in this study[a]

MiRNA Family	No. of Copies	Target Family	No. of Targets	Verified Targets
miR156/157	12	SBP	11	*SPL2, SPL3, SPL4, SPL10*
miR159/319	6	MYB and TCP	13	*MYB33, MYB65, TCP2, TCP3, TCP4, TCP10, TCP24*
miR160	3	ARF	3	*ARF10, ARF16, ARF17*
miR162	2	Dicer	1	*DCL1*
miR164	3	NAC	6	*CUC1, CUC2, NAC1, At5g07680, At5g61430*
miR165/166	9	HD-ZIPIII	5	*PHB, PHV, REV, ATHB-8, ATHB-15*
miR167	4	ARF	2	*ARF6, ARF8*
miR168	2	ARGONAUTE	1	*AGO1*
miR170/171	4	SCL	3	*SCL6-III, SCL6-IV*
miR172	5	AP2	6	*AP2, TOE1, TOE2, TOE3*
miR393	2	F-Box	5	*TIR1, AFB1, AFB2, AFB3, At3g23690*
miR394	2	F-box	2	*At1g27340*
miR395	6	APS	3	*APS1, APS4*
miR396	2	GRF	7	*GRL1, GRL2, GRL3, GRL7, GRL8, GRL9*
miR398	3	CSD	3	*CSD1, CSD2, At3g15640*
miR403	1	ARGONAUTE	1	*AGO2*

[a] Table adapted from Rhoades et al. (2006).

Figure 1. Polymorphisms occurring in *A. thaliana* miR156d (A), miR395f (B), the *AFB1* binding site for miR393 (C), and the *TOE3* binding site for miR172

For the miRNA polymorphisms, an example binding site is included for reference. The targeting mature miRNA sequence is included for reference purposes with the binding site polymorphisms. Binding sites are portrayed 5' → 3', while miRNAs are presented 3' → 5'. Polymorphic positions are colored blue and marked by a box. Major and minor describe the high and low frequency alleles, respectively. Complementary positions (i.e. non-mismatches or gaps) are depicted with black lines. G:U mismatches, which may contribute to miRNA-target binding, positions are marked with a black oval. Note pairing symbols are based on major allele states.

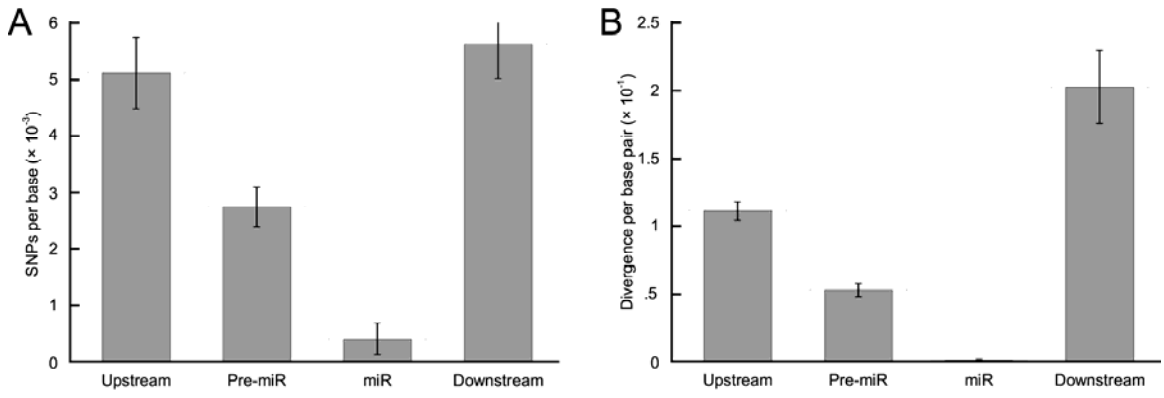


Figure 2. Mean levels of polymorphism (A) and divergence (B) at miRNAs and flanking regions

Standard error bars are included. Non-terminated error bars indicate that the sum of the mean and the error occurs above the maximum of the Y-axis.

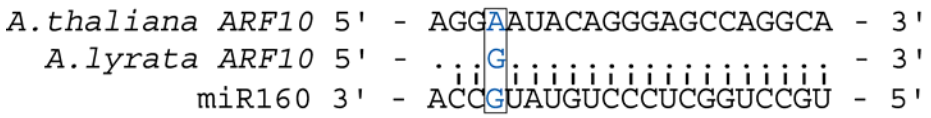


Figure 3. The nucleotide substitution that occurs between *A. thaliana* and *A. lyrata* in *ARF10*

The *ARF10* binding site is portrayed 5' → 3', while the sequence for miR160, which targets *ARF10*, is presented 3' → 5'. The variable position is colored blue and marked by a box. Complementary positions (i.e. non-mismatches or gaps) are depicted with black lines. Note pairing symbols are based on major allele states.

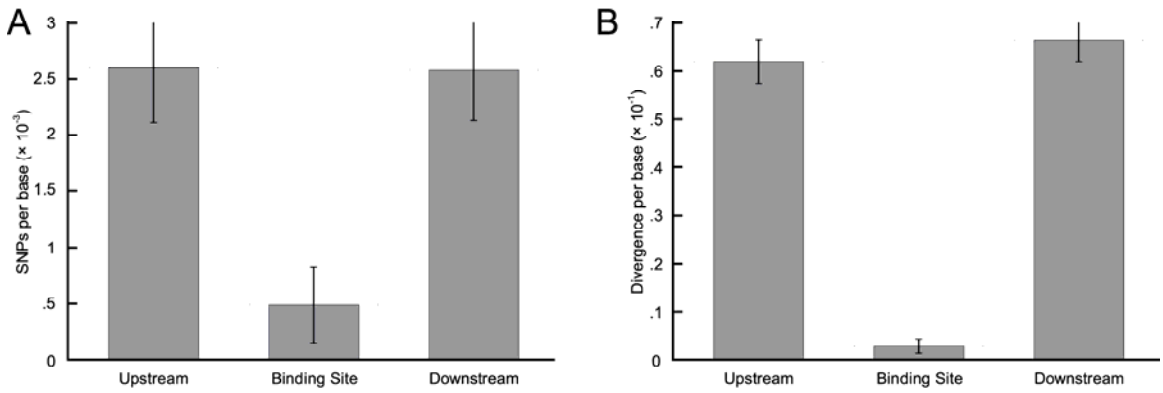


Figure 4. Mean levels of polymorphism (A) and divergence (B) at miRNA binding sites and flanking regions

Standard error bars are included. Non-terminated error bars indicate that the sum of the mean and the error occurs above the maximum of the Y-axis.

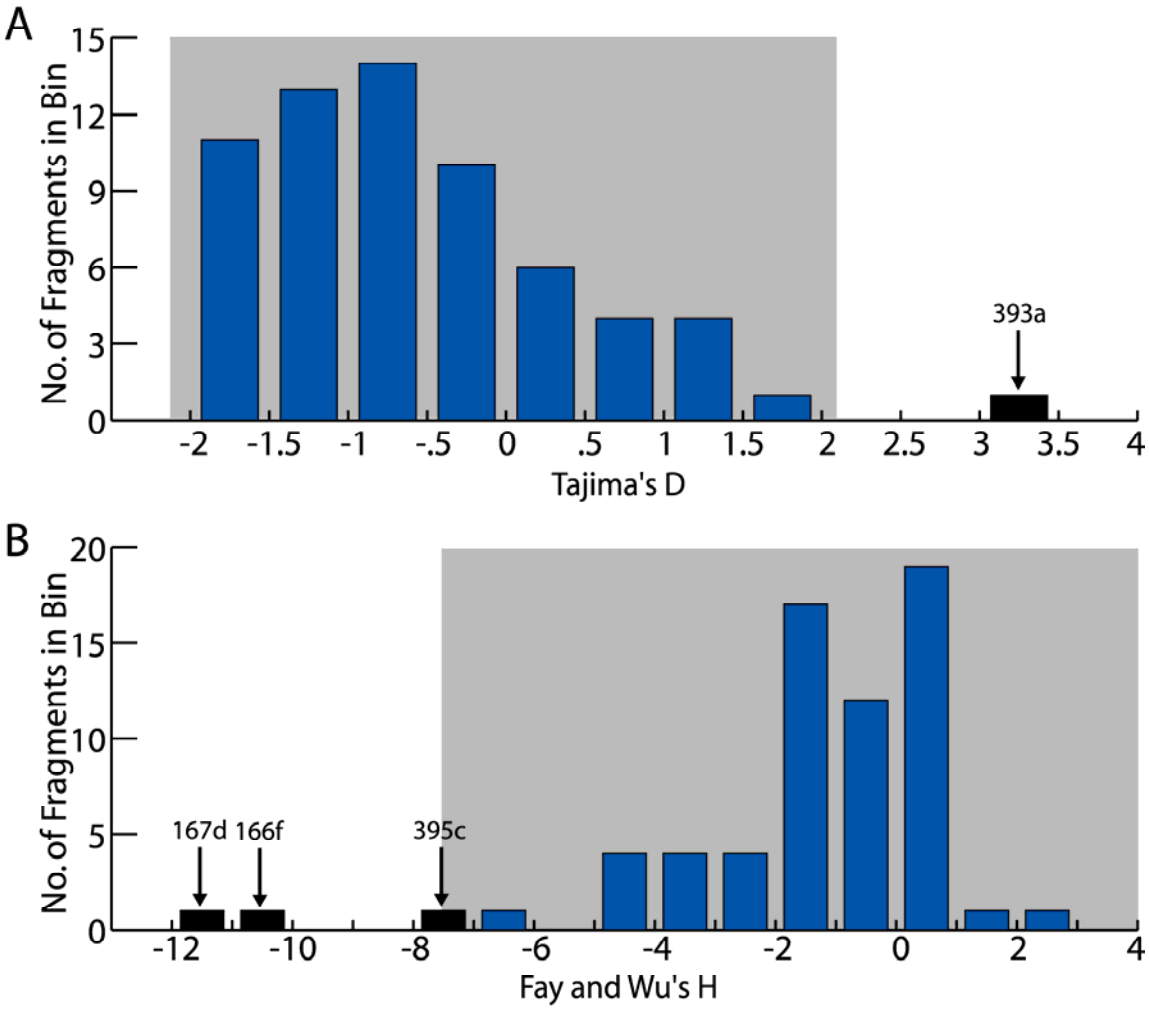
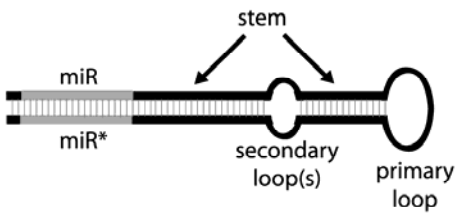


Figure 5. Distribution of Tajima's D (A) and Fay and Wu's H (B) values across all miRNA resequencing fragments

Significant fragments are labeled and colored black. Grey boxes represent the range of the middle 95% or the top 95% of the empirical distribution for Tajima's D and Fay and Wu's H, respectively, determined using genomewide background fragments from Nordborg et al. (2005).

A



B

Site Class	SNPs in Class	SNPs Affecting Structure	T°-Sensitive SNPs
miR	2	1	0
miR*	5	2	1
primary loop	9	4	1
secondary loop(s)	6	2	0
stem	40	27	1
total	62	36 (58%)	3 (5%)

Figure 6. Locations of SNPs within pre-miRNA molecules

Structural context classifications are displayed in (A). In (B), SNP counts per site class are presented. The number of SNPs affecting secondary structure in either 5 or 20° C is listed, as well as the number of SNPs that displayed effects only at 5° C ("T°-sensitive SNPs"). Note that none of the SNPs had dramatic effects on pre-miRNA secondary structure and that in all cases the general integrity of the molecule was maintained. SNP effects on secondary structure were subtle and included the addition or subtraction of a secondary loop or the enlargement or shrinking of a primary or secondary loop.

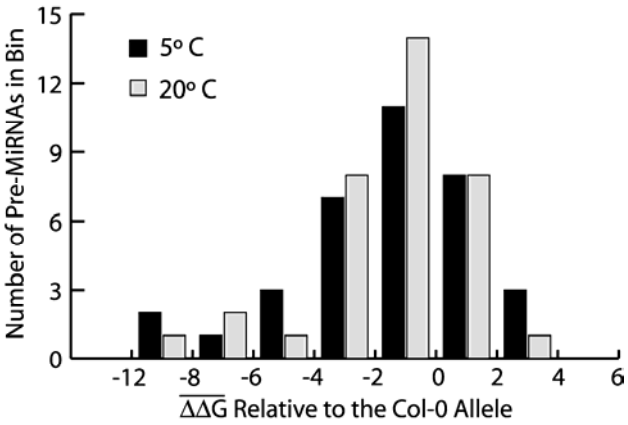


Figure 7. Distribution of mean ΔΔG values

The 35 resequenced loci with SNPs detected in the pre-miRNA were included.

CHAPTER FIVE:

Conclusion

Abstract

Identifying how adaptations arise at the genetic level is critical to improving our understanding of evolution. The research described in this dissertation is an attempt to contribute to this broad research area by trying to map the quantitative trait genes (QTGs) responsible for diversity in ecological important complex traits in *Arabidopsis thaliana*. Here, the results of this work are synthesized and possible future directions for this research are described.

The genetic basis of adaptation in complex traits

Understanding the genetic architecture of adaptation is a central aim of evolutionary biology. To understand how adaptive traits arise, however, it is necessary to clone the genes responsible for these adaptations (STINCHCOMBE and HOEKSTRA 2008). Arguably this is most feasible to do at a microevolutionay scale either within species in which beneficial polymorphisms segregate or between closely related species that have diverged presumably in part due to adaptation to different niches. To date, only a small number of adaptive traits have been characterized genetically, with few, if any, adaptive complex traits completely characterized (HOEKSTRA and COYNE 2007).

Dissecting adaptive complex traits may provide fundamental, mechanistic insights into how adaptation occurs (EHRENREICH and PURUGGANAN 2006). This is because only by elucidating the basis of many adaptive traits can we determine i) the distribution of the number of loci typically involved in an adaptation, ii) the distribution of the effect sizes of loci involved in an adaptation, and iii) the functional mechanisms by which adaptations arise. As we discussed briefly in Chapter 1, theoretical results have produced evidence supporting a wide variety of schemes for how adaptation can occur at complex traits without much agreement between different modeling approaches (FISHER 1930; GILLESPIE 1991; KIMURA 1983; ORR 1998). These discrepancies may reflect an underlying truth that complex trait

adaptation can occur through a variety of paths in terms of numbers and effects of the genes involved, that the process of adaptation might not be generalizable but rather trait- and context-specific, or that these modeling approaches may be inaccurate representations of the process of adaptation (ORR 2005). It is impossible to distinguish between these possibilities without empirical evidence.

A stifling limitation in this area is the small number of adaptations that have been resolved to the level of specific genes (as reviewed in (HOEKSTRA and COYNE 2007) and elsewhere). The number of complex traits that have been dissected to a gene resolution is increasing, yet it is unclear how many of these traits and their underlying causal alleles represent adaptive versus neutral or deleterious variation. For instance, in *A. thaliana*, a handful of natural alleles that cause flowering time variation have been positionally cloned (as reviewed in (EHRENREICH and PURUGGANAN 2006) and elsewhere), and of these, nearly half have been singleton alleles (e.g. *CRYPTOCHROME2* (EL-ASSAL *et al.* 2001), *FLOWERING LOCUS M* (WERNER *et al.* 2005), *PHYTOCHROME D* (AUKERMAN *et al.* 1997)) while the other half have been common (e.g. *FRIGIDA* (JOHANSON *et al.* 2000), *PHYTOCHROME B* (FILIAULT *et al.* 2008), and *PHYTOCHROME C* (BALASUBRAMANIAN *et al.* 2006)). Based on these findings, some have argued that the common variants may be beneficial alleles involved in ongoing adaptation to climatic differences across the species range (e.g. (STINCHCOMBE *et al.* 2004)), whereas the rare ones may represent deleterious alleles or alleles involved in local adaptation (as discussed in (FILIAULT *et al.* 2008)). However, discerning the fitness effects of such segregating alleles is a challenge that is impossible to solve with frequency information alone, especially in a species with a complicated population structure and demographic history like *A. thaliana* (NORDBORG *et al.* 2005; SCHMID *et al.* 2005), and presently only strong evidence for positive selection on any of these alleles exists for the *FRIGIDA* gene (TOOMAJIAN *et al.* 2006).

Motivated by a desire to contribute to this area, we have tried to map genes involved in variation in traits, namely flowering time and shoot architecture, that we think may be

contributing to ongoing adaptive evolution in the plant genetic model *A. thaliana*. By identifying the genetic basis of variation in these traits, it may be possible to understand mechanistically how selection operates in the wild on complex traits from both molecular and population genetic perspectives. However, mapping these genes systematically is a challenge and the development of *A. thaliana* as a model system for these types of questions has increased awareness of just how complicated this challenge is in this species (WEIGEL and NORDBORG 2005). We will attempt to synthesize our findings in a broad fashion – gleaning what we can both about the fundamental question of how the adaptive evolution of complex traits occurs, as well as about strategies for mapping QTGs in a systematic fashion.

Reverse genetics and the mapping of flowering time and shoot branching QTGs in *A. thaliana*

We have primarily used reverse genetics approaches, namely candidate gene resequencing (Chapters 2 and 4) and association mapping screens (Chapters 2 and 3), to identify genes involved in trait variation in *A. thaliana*. Across the three studies included in this dissertation, we have surveyed nearly 200 candidate gene loci and have had several key findings regarding the potential contributions of these genes to trait variation.

We began with a moderately-sized screen of 36 genes with known roles in shoot architecture that is described in Chapter 1 (EHRENREICH *et al.* 2007). This study resulted in the identification of several promising candidates for shoot branching variation in the *MORE AXILLARY GROWTH 2* (*MAX2*), *MAX3*, and *SUPERSHOOT 1* (*SPS1*) genes. However, single and two-locus epistasis linkage mapping results from recombinant inbred lines (RILs) did not conclusively validate or disprove the effects of these loci, leaving the biological significance of genetic variation at these loci unclear. Several possible explanations exist for these results: i) the associations were spurious, despite the use of conservative statistical approaches; ii) the recombinant inbred lines (RILs) used for comparison with the association mapping results did not contain the polymorphisms driving the candidate gene associations;

or iii) the genetic variance for shoot branching was converted from additive in the accessions to epistatic in the RILs during the construction of the RILs. Point (i) is impossible to determine with certainty based on existing data. Point (ii) is an issue because of the two RIL populations we used, one segregated for relevant *MAX2* SNPs, the other segregated for relevant *SPS1* SNPs, and neither segregated for relevant *MAX3* SNPs. As for point (iii), we speculated that if linkage disequilibrium (LD) between functional variants exists due to selfing or population structure, then the intercrossing of individuals containing different allele combinations could disrupt the effects of these loci and cause them to be detectable only through two-locus epistasis scans. Through a genomewide two-locus epistasis scan in the RILs, we did find suggestive evidence that this phenomenon could be possible in *A. thaliana*.

The shoot branching variation study elucidated several areas that needed improvement in terms of conducting a useful candidate gene association mapping study in *A. thaliana*. These were: i) having full gene resequencing data is preferable to having short resequencing fragments, given that LD decays very rapidly; ii) including a substantial number of background markers is necessary to determine genomewide patterns of trait association; and iii) it is necessary to have a mapping population that can permit the replication of any observed genotype-phenotype association found in the natural accessions. We attempted to address these issues in the study presented in Chapter 3, which explored the genetic basis of flowering time variation at 51 loci that were identified primarily through forward genetic screens as being involved in flowering time. In this study we genotyped over 1,000 SNPs both at the candidate genes and at random background fragments across 475 accessions, of which we were able to use 275 accessions for association mapping. We then attempted to replicate our candidate gene associations in a set of 360 recombinant inbred lines from heterogeneous stock (HSRILs) that were produced by intercrossing 19 accessions for six generations (as described in (SCARCELLI *et al.* 2007)), followed by selfing of the lines to homozygosity. This study produced several interesting results, namely i) that candidate genes do not appear to be associated with trait variation more often than random loci; ii) that

even after the implementation of highly conservative statistical approaches to testing for genotype-phenotype association, most remaining associations could not be replicated in the HSRILs; and iii) that the genes *CONSTANS* (*CO*), *GIBBERELIC ACID REQUIRING 1* (*GA1*), and *TERMINAL FLOWER 2* (*TFL2*) are strong novel candidates for flowering time variation in *A. thaliana*. Future work will be necessary to conclusively validate these associations and to determine the specific causal polymorphisms and their molecular effects.

Lastly, Chapter 4 examined the potential contribution of microRNAs (miRNAs) to phenotypic diversity in *A. thaliana* (EHRENREICH and PURUGGANAN 2008). MiRNAs are small RNAs that largely act as translational repressors in *A. thaliana* and other eukaryotes by cleaving or blocking the translation of target messenger RNAs (mRNAs) to which they possess sequence complementarity (CARRINGTON and AMBROS 2003). By resequencing more than 50% of the miRNAs and their binding sites in this species, we were able to determine how much variation exists at these loci and their targets, which is necessary to determine whether these loci might contribute to phenotypic variation. Overall, we found these loci are under intense constraint, identifying only four polymorphisms and one diverged site overall. We also found that the bulk of the polymorphisms that exist at miRNAs or their binding targets are in positions that do not matter for mRNA-miRNA pairing or for the structural integrity of precursor miRNA molecules. However, we did find ample variation in regions flanking miRNAs that could potentially generate functional differences between alleles. A recent study focused on interspecific sequence differences at miR319a showed that such miRNA-flanking variation can have effects on the expression patterns of miRNAs (WARTHMANN *et al.* 2008). The loci in our study have largely been characterized and are conserved across multiple plant clades. Other putative miRNAs that have arisen more recently were surveyed by others using resequencing tiling array data and, interestingly, these miRNAs do have slightly higher polymorphism than the loci we studied (ZELLER *et al.* 2008). Keeping in mind the caveat that some of these loci may not actually be miRNAs, these results are suggestive that the functionality of young miRNAs could segregate in *A. thaliana* and plants in general. However, further work is necessary to

determine if and how miRNAs contribute to adaptive evolution over both micro- and macroevolutionary timescales.

Overall, our results have provided evidence that certain genes are unlikely to contribute to trait variation, either generally speaking or specifically in flowering time and shoot architecture. However, we have developed strong evidence for the involvement of some genes in trait variation, such as *FLOWERING LOCUS C* (*FLC*), *GA1*, and *TFL2*. The evidence for these genes comes from the combined results of association mapping in multiple, independent populations. It will be up to subsequent students and postdocs to experimentally test whether these genes represent true QTGs and whether they matter to adaptive evolution in *A. thaliana*.

New resources and technologies give rise to new ways of conducting quantitative genetics in *A. thaliana*

Over the last few years, *A. thaliana* has matured as a system for a population and quantitative genetics. Recently, information about more than 200,000 SNPs was published as part of an effort to resequence 20 accessions using microarrays (CLARK *et al.* 2007). Soon, genotype data at these SNPs will become available for nearly all stock center accessions (as described in (KIM *et al.* 2007)). Efforts are already underway to entirely resequence every stock center accession with massively parallel sequencing technologies (http://www.1001genomes.org). Essentially, the data collection phase of population genetics in *A. thaliana* will soon be completed, and the challenge will no longer be getting DNA polymorphism information but rather finding which DNA polymorphisms affect trait variation and more specifically contribute to ongoing adaptation in this species. These questions have always been at the forefront of work in this area, but these upcoming resources will facilitate a new type of quantitative genetics where we know every polymorphism and need only focus on defining which ones matter biologically.

To fully exploit these new genotyping and resequencing data, it will be necessary to generate resources that can be used to differentiate the polymorphisms that matter from those that do not on a large-scale. In the absence of gene replacement technology, this effort likely will have to be pursued with traditional mapping approaches. Because linkage disequilibrium (LD) decays rapidly in this species, association mapping has been advocated as the ideal approach to identify this functional genetic variation (KIM *et al.* 2007). The studies we present in Chapter 2 and 3, as well as papers by others (ARANZANA *et al.* 2005; ZHAO *et al.* 2007), call into question whether this approach is ideal though. Indeed, in a system that is amenable to laboratory crossing and in which large numbers of progeny can be obtained with ease, it seems that the construction of improved linkage mapping resources may be worth consideration for the future.

History shows that traditional biparental RILs do not capture a sufficient level of genetic diversity and do not provide a resolution where they will be useful in this effort (as described in (EHRENREICH and PURUGGANAN 2006)). Two options exist for linkage mapping approaches that will be more effective. First, the creation of advanced intercross lines via the intercrossing of two parents and their progeny for a large number of generations will surely provide an improved mapping resolution. Optimal breeding strategies for constructing such lines were recently reported (ROCKMAN and KRUGLYAK 2008) and the construction of advanced intercross lines in *A. thaliana* would certainly be feasible within a several year period. To date, the most advanced intercross population is one derived from two generations of intercrossing (as reported in (LI *et al.* 2006)). However, examples of using advanced intercross lines to achieve substantially higher mapping resolution than would be possible with traditional biparental RILs exist in other species, such as *Caenorhabditis elegans* (e.g. (SEIDEL *et al.* 2008)).

Second, multi-parent RILs (e.g. the HSRILs described earlier) or RIL populations constructed from the crossing of many individuals to a common individual provide an opportunity to survey a large amount of genetic diversity using linkage mapping approaches.

Such lines have been described in fly (MACDONALD and LONG 2007), maize (YU *et al.* 2008), mouse (CHURCHILL *et al.* 2004), and now *A. thaliana* (the intercrossing phase of line construction was described in (SCARCELLI *et al.* 2007) but the final inbred lines used in Chapter 3 are unpublished), and these lines permit the high resolution mapping of linkages with subsequent fine-mapping of putative causal alleles potentially to a nucleotide level by association mapping (as described methodologically in (YALCIN *et al.* 2005; YU *et al.* 2008) and elsewhere). What will be the most useful method to systematically map the QTGs in *A. thaliana* is an open question, but the possibility that it is not association mapping must be considered.

A. thaliana is a potentially excellent model system for studying adaptation because it possesses a tremendous amount of both genetic and phenotypic diversity. However, proceeding into the future, it is worth considering how to optimally use the ample resources in this system to identify the genetic bases of adaptations. We have attempted to contribute to a foundation for the exploration of the mechanisms generating adaptive evolution in plants by using *A. thaliana* as a model. The future in this area is bright and it is likely that a better understanding of complex trait genetics, and more specifically adaptive complex trait genetics, is something that can be achieved within the coming decade.

References

ARANZANA, M. J., S. KIM, K. ZHAO, E. BAKKER, M. HORTON *et al.*, 2005 Genome-Wide Association Mapping in Arabidopsis Identifies Previously Known Flowering Time and Pathogen Resistance Genes. PLoS Genet 1: e60.

AUKERMAN, M. J., M. HIRSCHFELD, L. WESTER, M. WEAVER, T. CLACK *et al.*, 1997 A deletion in the PHYD gene of the Arabidopsis Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. Plant Cell 9: 1317-1326.

BALASUBRAMANIAN, S., S. SURESHKUMAR, M. AGRAWAL, T. P. MICHAEL, C. WESSINGER *et al.*, 2006 The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of Arabidopsis thaliana. Nat Genet 38: 711-715.

CARRINGTON, J. C., and V. AMBROS, 2003 Role of microRNAs in plant and animal development. Science 301: 336-338.

CHURCHILL, G. A., D. C. AIREY, H. ALLAYEE, J. M. ANGEL, A. D. ATTIE *et al.*, 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36: 1133-1137.

CLARK, R. M., G. SCHWEIKERT, C. TOOMAJIAN, S. OSSOWSKI, G. ZELLER *et al.*, 2007 Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 317: 338-342.

EHRENREICH, I. M., and M. D. PURUGGANAN, 2006 The molecular genetic basis of plant adaptation. Am J Bot 93: 953-962.

EHRENREICH, I. M., and M. D. PURUGGANAN, 2008 Sequence variation of MicroRNAs and their binding sites in Arabidopsis. Plant Physiol 146: 1974-1982.

EHRENREICH, I. M., P. A. STAFFORD and M. D. PURUGGANAN, 2007 The genetic architecture of shoot branching in Arabidopsis thaliana: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping. Genetics 176: 1223-1236.

EL-ASSAL, S. E.-D., C. ALONSO-BLANCO, A. J. 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.

FILIAULT, D. L., C. A. WESSINGER, J. R. DINNENY, J. LUTES, J. O. BOREVITZ *et al.*, 2008 Amino acid polymorphisms in Arabidopsis phytochrome B cause differential responses to light. Proc Natl Acad Sci U S A 105: 3157-3162.

FISHER, R., 1930 *The genetical theory of natural selection*. Oxford University Press, Oxford.

GILLESPIE, J., 1991 *The causes of molecular evolution*. Oxford University Press, Oxford.

HOEKSTRA, H. E., and J. A. COYNE, 2007 The locus of evolution: evo devo and the genetics of adaptation. Evolution 61: 995-1016.

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.

KIM, S., V. PLAGNOL, T. T. HU, C. TOOMAJIAN, R. M. CLARK *et al.*, 2007 Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat Genet 39: 1151-1155.

KIMURA, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

LI, Y., P. ROYCEWICZ, E. SMITH and J. O. BOREVITZ, 2006 Genetics of local adaptation in the laboratory: flowering time quantitative trait loci under geographic and seasonal conditions in Arabidopsis. PLoS ONE 1: e105.

MACDONALD, S. J., and A. D. LONG, 2007 Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of Drosophila melanogaster. Genetics 176: 1261-1281.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 3: e196.

ORR, H., 1998 The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52: 935-949.

ORR, H. A., 2005 Theories of adaptation: what they do and don't say. Genetica 123: 3-13.

ROCKMAN, M. V., and L. KRUGLYAK, 2008 Breeding designs for recombinant inbred advanced intercross lines. Genetics 179: 1069-1078.

SCARCELLI, N., J. M. CHEVERUD, B. A. SCHAAL and P. X. KOVER, 2007 Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. Proc Natl Acad Sci U S A 104: 16986-16991.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169: 1601-1615.

SEIDEL, H. S., M. V. ROCKMAN and L. KRUGLYAK, 2008 Widespread genetic incompatibility in C. elegans maintained by balancing selection. Science 319: 589-594.

STINCHCOMBE, J. R., and H. E. HOEKSTRA, 2008 Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100: 158-170.

STINCHCOMBE, J. R., C. WEINIG, M. UNGERER, K. M. OLSEN, C. MAYS *et al.*, 2004 A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci U S A 101: 4712-4717.

TOOMAJIAN, C., T. T. HU, M. J. ARANZANA, C. LISTER, C. TANG *et al.*, 2006 A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. PLoS Biol 4: e137.

WARTHMANN, N., S. DAS, C. LANZ and D. WEIGEL, 2008 Comparative analysis of the MIR319a microRNA locus in Arabidopsis and related Brassicaceae. Mol Biol Evol 25: 892-902.

WEIGEL, D., and M. NORDBORG, 2005 Natural variation in Arabidopsis. How do we find the causal genes? Plant Physiol 138: 567-568.

WERNER, J. D., J. O. BOREVITZ, N. WARTHMANN, G. T. TRAINER, J. R. ECKER *et al.*, 2005 Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci U S A 102: 2460-2465.

YALCIN, B., J. FLINT and R. MOTT, 2005 Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. Genetics 171: 673-681.

YU, J., J. B. HOLLAND, M. D. MCMULLEN and E. S. BUCKLER, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics 178: 539-551.

ZELLER, G., R. M. CLARK, K. SCHNEEBERGER, A. BOHLEN, D. WEIGEL *et al.*, 2008 Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. Genome Res 18: 918-929.

ZHAO, K., M. J. ARANZANA, S. KIM, C. LISTER, C. SHINDO *et al.*, 2007 An Arabidopsis example of association mapping in structured samples. PLoS Genet 3: e4.