MOLECULAR ECOLOGY

Molecular Ecology (2017) **26**, 92–106 doi: 10.1111/mec.13643

SPECIAL ISSUE: THE MOLECULAR MECHANISMS OF ADAPTATION AND SPECIATION: INTEGRATING GENOMIC AND MOLECULAR APPROACHES

Population genomic scans suggest novel genes underlie convergent flowering time evolution in the introduced range of *Arabidopsis thaliana*

BILLIE A. GOULD[1] and JOHN R. STINCHCOMBE[2]
Department of Plant Biology, Michigan State University, 612 Wilson Rd, East Lansing, MI 48823, USA

Abstract

A long-standing question in evolutionary biology is whether the evolution of convergent phenotypes results from selection on the same heritable genetic components. Using whole-genome sequencing and genome scans, we tested whether the evolution of parallel longitudinal flowering time clines in the native and introduced ranges of *Arabidopsis thaliana* has a similar genetic basis. We found that common variants of large effect on flowering time in the native range do not appear to have been under recent strong selection in the introduced range. We identified a set of 38 new candidate genes that are putatively linked to the evolution of flowering time. A high degree of conditional neutrality of flowering time variants between the native and introduced range may preclude parallel evolution at the level of genes. Overall, neither gene pleiotropy nor available standing genetic variation appears to have restricted the evolution of flowering time to high-frequency variants from the native range or to known flowering time pathway genes.

Keywords: cline, convergent evolution, flowering time, genome scan, introduced species, selective sweep

Received 15 January 2016; revision received 6 April 2016; accepted 8 April 2016

Introduction

A long-standing question in evolutionary biology is whether the evolution of convergent phenotypes results from selection on the same heritable genetic components. Are there multiple genetic mechanisms to produce adaptive phenotypes? Or, do correlations between traits and the pleiotropic constraint imposed by some genes and complex pathways confine evolution to using the same genetic mechanisms or loci? Answering these questions provides direct evidence of the role genetic constraint or flexibility in evolutionary responses. A fundamental goal in evolutionary biology is to understand which genes are subject to selection and how often, and to evaluate the general predictability of the evolutionary process at the phenotypic and genetic levels (Wood *et al.* 2005; Arendt & Reznick 2008; Elmer & Meyer 2011; Lee *et al.* 2014). Here, we show how parallel evolution in a key life history trait, flowering time, in the introduced range of the mouse-ear cress (*Arabidopsis thaliana*) utilizes novel loci, suggesting that numerous independent genetic mechanisms can be used to produce adaptive phenotypes.

Convergent evolution of locally adaptive traits can play a significant role in the successful establishment of plants outside their native ranges (Lee 2002; Bossdorf *et al.* 2005; Buswell *et al.* 2011). The existence of parallel geographic clines in the native and introduced ranges is often an indicator that similar climate-based selective pressures have promoted convergent local adaptation in both ranges. For example, North American populations of St. John's Wort (*Hypericum perforatum*) exhibit clines in physiological traits, biomass and fecundity that

Correspondence: Billie Gould, Fax: +1 (517) 353-1926;
E-mail: bgould@msu.edu
[1]Present address: Plant Biology Laboratories, Department of Plant Biology, Michigan State University, 612 Wilson Road, East Lansing, MI 48824-1312, USA
[2]Present address: Department of Ecology and Evolutionary Biology, Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON, Canada

parallel patterns found in its native European range (Maron *et al.* 2004, 2007). Similarly, the invasive purple loosestrife has rapidly evolved geographic differences in time to flowering similar to those found in Europe (Montague *et al.* 2008; Colautti & Barrett 2013). As natural replicates of the evolutionary process, convergently evolved, native and introduced populations are often used as evidence of natural selection shaping phenotypic trends. They also provide a valuable scenario for testing to what extent selection acts repeatedly on the same traits, genes and mutations. Theory (Orr 2005) and intuition suggest that parallel evolution at the genetic level will be common when there are relatively few possible or viable genetic mechanisms to produce adaptive phenotypes (e.g. when there is high gene pleiotropy) and rare when there are many genetic routes to adaptation. Adaptation from standing genetic variation in the introduced range is also predicted to occur first through selection on variants that are already high in frequency and are (or were) adaptive (Barrett & Schluter 2008). To date, however, there have been few studies where the genetic basis of variation in convergently evolved traits has been compared between native and introduced populations.

For plants, flowering when environmental conditions are optimal is one of the most important traits for maximizing fitness, and much of what we know about the genetic pathways controlling flowering time has resulted from molecular work on laboratory strains of *Arabidopsis* (reviewed (Simpson & Dean 2002)). More recently, studies of naturally occurring *Arabidopsis* accessions have revealed a suite of both rare and common genetic variants that influence flowering time variation in the species native European range. Some mutations are common and control a large portion of the rangewide variation in phenotype while others are regionally restricted in their distribution. Null mutations affecting the interacting loci *FRIGIDA (FRI)* and *FLOWERING LOCUS C* (FLC) explain 50–70% of flowering time variation under overwintering conditions in some studies (Caicedo *et al.* 2004; Shindo *et al.* 2005). Common, high-impact variants also occur in the light-sensing protein *PHYTOCHROME C* (PHYC) (Balasubramanian *et al.* 2006; Samis *et al.* 2008) and in the transcription factor MADS affecting flowering 2 (MAF2) (Caicedo *et al.* 2009; Rosloski *et al.* 2010). Rare variants in FRI, FLC and other loci also contribute to differences between specific lines (Gazzani *et al.* 2003; Shindo *et al.* 2005; Li *et al.* 2014). Multiple cross-QTL experiments have yielded similar results as to the number of genomic regions that control flowering time variation between divergent European ecotypes. In a study of 13 recombinant inbred line families generated from mainly non-North American parental genotypes, Brachi *et al.* (2010) found five genomic regions containing upwards of 60 QTL implicated in flowering time variation, many of which were unique to a single cross. Salomé *et al.* (2011) found flowering time variation was also linked to as few as five genomic regions using 17 F2 populations. A recent genomewide association study that included the analysis of 23 separate flowering time-related phenotypes identified significant SNPs in seven loci, some of which affected known flowering time pathway genes (including *FRI* and *FLC*) and some of which did not (Aranzana *et al.* 2005; Atwell *et al.* 2010; Horton *et al.* 2012).

Mutations affecting flowering time variation in *Arabidopsis thaliana* have been largely generated from studies that included few or no genotypes from the species introduced range in North America (but see Atwell *et al.* (2010)). Only one study of which we are aware has directly compared flowering time differences between genotypes from the native and introduced ranges. Samis *et al.* (2012) found similar longitudinal flowering time clines in lines from both Europe and North America. Under natural overwintering conditions in a common garden, plants from more coastal populations flowered later in both Europe and North America. The cline was significant on both continents, but stronger in Europe (−0.20 days/degree longitude) than in North America (−0.09 days/degree longitude, Fig S1A). In the introduced range, the cline is strongly correlated with total annual precipitation, and in these lines, selection acts more strongly in favour of early flowering under low water conditions (Stock *et al.* 2015). As an initial test for a parallel genetic basis underlying flowering time clines in the native and introduced range, Samis and colleagues genotyped populations for eight common polymorphisms across *FRI*, *FLC* and *PHYC*. Some common alleles were present in the introduced range, but none were strongly associated with flowering time phenotype. It remains unknown whether other less common flowering time alleles, novel mutations affecting flowering time pathway genes or variants in other parts of the genome have been co-opted during local adaptation in the introduced range.

In the present study, we used a whole-genome sequencing approach to search for evidence of selection associated with flowering time differences in introduced *Arabidopsis* genotypes. First, we identified SNPs genomewide that show high levels of genetic differentiation between early- and late-flowering individuals, while controlling for population structure. We paired this with diversity-based scans to identify potential regions of recent selection associated with flowering time variation. We then compared the results of these analyses to the functional categorization of the implicated regions and known flowering time genes and mutations from the native range. Collectively, we find little or no evidence of genetic parallelism which suggests that neither pleiotropic constraint nor available standing genetic variation has strongly constrained the evolution of flowering time in the species introduced range to selection on high-frequency variants from the native range or to known flowering time pathway genes.

Methods

Plant material

Thirty-four North American *Arabidopsis* lines were selected for sequencing from among natural accessions previously phenotyped by Samis *et al.* (2012). The lines were obtained from the Arabidopsis Biological Resource Center (ABRC) and from recent collections from natural populations (Table S1). The lines were phenotyped in a rooftop common garden at the University of Toronto after overwintering under natural conditions (see Samis *et al.* (2012)). We chose 17 late-flowering and 17 early-flowering lines for whole-genome sequencing, using previous SNP data to choose lines with the maximum level of genetic diversity. The average flowering time difference between the two phenotypic groups was 6.7 days. In total, we sequenced lines from 14 different populations in the eastern United States, between one and five lines per population. Some of the early-flowering lines and the late-flowering lines came from the same population (see Fig. S1 and Table S1). The presence of early- and late-flowering lines at the same locality raises two questions: Why is there within-population variation, and what effect will it have on our population genomic analyses? As described by Stock *et al.* (2015); also see Brachi *et al.* (2011), within-population variation along a broad cline could be due to recent geneflow, nonequilibrium conditions, seasonally varying selection, diversifying selection pressures on the trait or microgeographic variation in selection. One consequence of early- and late-flowering lines sometimes occurring in the same locality in our sample is that population history is shared across phenotypes, allowing us to control for the effects of population structuring on genetic variation across the genome. In practice, we also use a differentiation statistic that controls for population history and uneven sampling among localities through a reference covariance matrix of SNP allele frequencies.

To generate plant material for sequencing, seeds were stratified in 0.15% agarose at 4 °C for 3 days and then germinated in soil in a growth chamber at 20 °C under 16-h days. DNA was extracted from young leaves using the Qiagen DNeasy Plant Mini kit. One barcoded sequencing library was constructed per line (34 lines total) using the Illumina TruSeq DNA sample prep kit. Library preparation and sequencing were performed at the McGill University and Genome Quebec Innovation Center (Montréal, Québec).

Sequencing and variant detection

Whole-genome sequencing was performed on two lanes of the Illumina HiSeq2000 platform using 100-bp paired end reads. We obtained an average of 24.4 million high-quality reads per line. We aligned reads back to the TAIR10 reference genome using STAMPY (Lunter & Goodson 2011) with a predicted substitution rate of 0.007 (Mitchell-Olds & Schmitt 2006). On average, 92.5% of reads could be mapped back to the reference per line, generating an average of 14.6X coverage across the autosomes. We removed PCR duplicates from each alignment using PICARD TOOLS v1.98 ('http://broadinstitute.github.io/picard' 2015) and used the Genome Analysis Toolkit (GATK, v2.7-2) for downstream filtering and variant calling. We identified insertions and deletions using GATK RealignerTargetCreator and realigned reads around them using INDELREALIGNER. We created reduced alignment files using REDUCEREADS and called variants using HAPLOTYPECALLER. Variants were called using the set of all alignments in discovery mode with a strand call confidence of 30 and an emit confidence of 10. We filtered the initial set of called variants following guidelines provided by GATK best practices documentation. SNPs were excluded that had quality divided by read depth (QD) <2.0, Fisher strand bias (FS) > 60.0 and mapping quality (MQ) < 40.0. Indels were removed that had QD < 2.0, FS > 200.0 and MQ < 40. Filtering removed about 6.7% of the original variants, leaving 2 453 132 autosomal variants in the data set. Downstream analyses were conducted using VCFTOOLS (0.1.11), PYTHON (v.2.7.5) and R (R Core Team 2012).

We compared the SNPs in our lines with results from two previous resequencing studies. We identified all SNP sites with nonmissing allele calls in at least half of the lines (n = 662 892 sites). Our goal was to compare only high-confidence SNPs in North America with SNPs in other data sets. These were compared with filtered SNP calls published from the 1001 genomes project (Cao *et al.* 2011) http://1001genomes.org/data/MPI/MPICao2010/releases/2012_03_13/strains/ and SNPs from 19 genomes that make up the MAGIC nested association panel (Gan *et al.* 2011) (downloaded from http://mus.well.ox.ac.uk/19genomes/variants.tables/).

We validated the filtered set of variants against genotype calls at 135 SNPs previously genotyped in the same lines using Sanger sequencing (Samis *et al.* 2012). Two of the lines had a low validation rate (<60% of markers had matching calls) and we excluded these from further analysis (both were early-flowering genotypes). Among the remaining 32 lines, concordance at homozygous sites between genotype calls using the two methods was 96.4%.

From the initial data set, we isolated a set of high-confidence variants to use for analysis of population structure and differentiation. We chose to retain only variants on the autosomes with no missing data, a minor allele frequency above 0.0625 (four out of 64 possible calls at each site), a minimum mean individual depth of coverage of three reads and a maximum of 80, and excluded all nonbiallelic variants (1.7% of all variants). We used the program SNPEFF (Cingolani *et al.* 2012) to annotate the high-confidence variants and to predict the functional impact of each variant on protein coding genes and noncoding RNAs. High-impact variants are defined as those predicted to cause premature start and stop codons, destruction of splice sites and frameshift mutations in protein coding regions or noncoding RNAs. We used VCFTOOLS to estimate linkage disequilibrium between variants separated by at most 25 kb.

In the high-confidence variant set, we considered variants in difficult to sequence regions as missing data. Because *Arabidopsis* is highly inbreeding and expected to have high homozygosity throughout the genome, we identified genomic regions with low-quality sequence by scanning for regions of unusually high heterozygosity indicative either of sequencing error or recent rare outcrossing events. Per cent heterozygosity was calculated as the percentage of heterozygous genotype calls out of total calls at variant sites in 20 kb, nonoverlapping windows. We calculated both within-individual and across-individual per cent heterozygosity. In a single individual (line PA-DT1-12), we found large regions of high heterozygosity indicative of recent outcrossing and considered genotype calls in these regions for this one individual as missing data. We also considered variants in windows with >20% heterozygous calls across all individuals as missing data. As predicted, heterozygosity was increased in highly repetitive centromeric regions of each chromosome and we excluded all variants in these regions (in total about 10.6 Mbp) from further analysis (equal to about 8.5% of the genome).

We used the program PINDEL (Ye *et al.* 2009) to detect structural variants. We searched for insertions and deletions greater than 79 bp (the largest indel size detected in our genomes by the GATK). We also tested for reads supporting the presence of large replacements and tandem duplications. Structural variants (SVs) were considered high confidence if they were supported in at least four individuals, and had a maximum mean depth of coverage below 80. We used a Fisher's exact test to test whether each high-confidence structural variant was overdifferentiated between early- and late-flowering plant groups. We interpret the results of computational SV detection with caution, and SVs of interest will need to be validated using PCR in future work.

Population structure analysis

We used a subset of the high-confidence SNPs to analyse population structure using the program STRUCTURE (Pritchard *et al.* 2000). We used only SNPs with a mean individual depth of coverage between 8 and 20 reads, and selected a subset spaced at least 10 kb apart throughout the genome to control for linkage disequilibrium (n = 9852 SNPs). We used PGDSpider (Lischer & Excoffier 2012) to convert the SNP data to the correct format and then ran STRUCTURE using an admixture model with a burn-in of 10 000 iterations, 10 000 sampling iterations and K = 2 or 3 groups, chosen based on previous studies of population structure in North American *Arabidopsis* (Beck *et al.* 2008; Platt *et al.* 2010). The STRUCTURE model incorporated a priori information on sample population of origin. Our goals with the STRUCTURE analysis were twofold: (i) to determine whether population structure in our sample followed a longitudinal cline, such that clinical variation in flowering time would be confounded with population structure, and (ii) to determine whether population structure was confounded with the early- and late-flowering groups that we examined with XtX statistics.

To see how our sequenced lines were related to other accessions worldwide, we compared them to a subset of 90 geographically diverse lines genotyped by Platt *et al.* (2010). We used genotype information from whole-genome sequencing calls at 135 SNP marker locations (Samis *et al.* 2012) to calculate distances between individuals and generate neighbour-joining trees in the program TASSEL (Bradbury *et al.* 2007). We included data for 65 lines scattered throughout Eurasia, 25 North American lines and genotype calls from our 32 sequenced lines. To verify genetic diversity of our sample, we also screened our samples for the presence of a common haplotype found in over 1000 lines from North America reported by Platt *et al.* (2010). That haplotype was represented in only 13% of our sequenced lines (n = 4).

Signatures of selection

We used three tests in genomewide scans to detect regions of the genome with potential signatures of selection for flowering time.

Differentiation as X^TX (~F_{ST}). First, we identified variants that were overdifferentiated between early- and late-flowering plants and also predicted to have deleterious impacts on protein coding genes and noncoding RNAs (as annotated by SNPEFF, above). To measure differentiation, we calculated the statistic X^TX for high-confidence variants using the program BAYENV2 (Günther & Coop 2013). The X^TX statistic is analogous to F_{ST}, but is based on allele frequencies that are normalized for population structure and uneven sampling between groups. X^TX controls for residual population structure through the use of a kinship matrix generated from a subsample of genomewide SNPs. The expected value of X^TX is the number of groups being compared (n = 2). We generated a kinship matrix used to normalize allele frequencies using a set of 3000 high-confidence synonymous coding SNPs spaced evenly throughout the genome. We then calculated X^TX for two subsets of markers: (i) 20 213 high-confidence synonymous SNPs spaced no closer than 2500 bp from each other throughout the genome and (ii) the set of all variants predicted to have strong deleterious impacts on genes (n = 17 324). X^TX values were calculated as the average value from three separate BAYENV2 runs using a burn-in of 30 000 iterations and 100 000 sampling iterations each. The distribution of values from synonymous SNPs was used to gauge background levels of differentiation between the early- and late-flowering plant groups. We considered SNPs overdifferentiated if they had average X^TX values above the average of the highest synonymous X^TX values from across three runs. We thus attempted to mitigate the potential problems of population structure in our identification of SNPs as overdifferentiated in two ways: (i) X^TX, by its construction, identifies individual loci with reference to a covariance matrix of allele frequencies, which we estimated with 3000 synonymous SNPs, and (ii) we compare the distribution of X^TX values for high-impact variants to a different set of markers expected to be neutral (the 20 213 synonymous SNPs), as suggested by Lotterhos & Whitlock (2015).

F_{ST} can be inflated by processes that reduce within-population genetic diversity (e.g. inbreeding depression, stronger background selection in some regions of the genome, differences in effective population size) that do not universally reduce variation in the entire sample (Charlesworth 1998). We do not expect this to be a substantial problem with our implementation of X^TX because our comparison of early- and late-flowering genotypes incorporates samples from several populations, with different relationships to native samples, into the categories being compared. Consequently, the forces that would reduce within-sample genetic diversity in one flowering group but not the entire sample are unlikely to be occurring independently in several localities that make up the early- or late-flowering groups.

Cross-population composite likelihood ratio test. We used the cross-population composite likelihood ratio (XP-CLR) test (Nielsen *et al.* 2005) to search for genomic regions indicative of selective sweeps. We analysed high-confidence variants including heterozygous sites (to avoid biases introduced due to missing data). We obtained physical and genetic map distances between 455 markers in the *Arabidopsis* genome from The Arabidopsis Information Resource (Rhee *et al.* 2003) (map files; RIdataGeneticMap.txt and TAIR9_AGI_Map_data). We used the genetic distance information from the markers to predict genetic distances between all markers in our populations using a fourth-order polynomial. We calculated both the XP-CLR of selection in the early-flowering plant group using the late-flowering group as a reference and vice versa. Input genotypes were left unphased, as haplotype information is not required. The presence vs. absence of a selective sweep was tested in 0.5 cM windows around grid points placed every 5 kb along each chromosome. We excluded grid points within centromeric regions (due to poor genotyping accuracy in these repetitive areas). SNPs were downsampled to a maximum of 100 SNPs per window. We calculated the difference between the XP-CLR statistic in the early group and the late group at each grid point and considered as potential sweep regions those with difference values below the first quantile or above the 99th quantile of values.

Nucleotide diversity. We looked genomewide for regions of unusually low nucleotide diversity (π) in either the early- or late-flowering plants. We used the set of all called sites (including invariant sites) to calculate diversity in windows across the genome (described below). To generate call confidence information for all bases throughout the genome, we used an alternative GATK pipeline to call variants. Sequence alignments were used to call variants separately for each individual against the reference genome using GATK HAPLOTYPECALLER. A combined variant set was then generated using GATK CombineGVCFs followed by GenotypeGVCFs, emitting genotype and read depth information for every base in the genome.

We calculated π in nonoverlapping windows across the genome using only genotype calls with a minimum depth of four reads. To keep the number of individuals sampled at each site constant, we randomly subsampled genotype data for 12 individuals only at sites where at least 12 individuals had nonmissing calls. Π was calculated for windows containing 10 000 bases meeting the read depth and missing data cut-offs separately within each phenotypic group. Using the window data, we found the difference between π values in the early- and late-flowering lines at each grid point previously used to calculate XP-CLRs. We considered grid points with difference values below the first quantile and above the 99th quantile of all difference values as potential sweep regions in the late- and early-flowering groups, respectively. Windows of interest around these significant grid points were defined using the boundaries from the original group-specific window calculations.

Identification of genes potentially under selection

To detect genes potentially under selection for flowering time, we identified genes that fell in regions of the genome characterized by a large difference in diversity between groups or a high XP-CLR. Nucleotide diversity (π) was calculated in windows as described above and XP-CLR windows of interest were defined as 2500 bp upstream and downstream of significant grid points. Candidate genes related to flowering time adaptation overlap genomic windows with one of these signatures of selection and contain overdifferentiated F_{ST} (X^TX) outliers. We conducted functional term enrichment analyses of different gene sets of interest using the online tools DAVID (Huang *et al.* 2009) and GOrilla (Eden *et al.* 2009). The set of all genes affected by overdifferentiated (but nondeleterious) SNPs was used as the background set for enrichment analyses of high-impact outliers. The set of all *Arabidopsis* genes was used as the background set for enrichment comparisons of genes in regions with significant signatures of selection.

We also directly examined a predetermined list of 284 candidate genes, identified by Brachi *et al.* (2010) and Atwell *et al.* (2010), which are associated with plant development and responses to endogenous and exogenous environmental cues. We included three genes identified as associated with flowering time by Atwell *et al.* (2010). Genes in this list were termed '*a priori*' candidate genes for flowering time. For each candidate locus, within each phenotypic group, we calculated π using nonmissing genotype calls with data for at least seven individuals. When more individuals had valid calls, the data were randomly downsampled to seven within each phenotypic group. We then compared π values for each gene between the early- and late-flowering plant groups and compared the differences against the distribution of differences calculated similarly for all other genes in the genome.

Extended analysis of candidate loci FLC, FRI, PHYC and MAF2

Indel and SNP polymorphisms in the loci FLC, FRI, PHYC and MAF2 have been previously linked to geographic variation in flowering time in *Arabidopsis* (Caicedo *et al.* 2004, 2009; Balasubramanian *et al.* 2006; Samis *et al.* 2008; Rosloski *et al.* 2010; Li *et al.* 2014). We compared the sequences of these four loci and the surrounding regions with sequences published in other studies to determine whether (i) polymorphisms associated with flowering time identified in other studies are present and similarly associated in the 32 North American lines and (ii) the genes or surrounding sequence haplotypes show any structuring between early- and late-flowering lines. Polymorphisms in +/− 50-kb regions surrounding each locus, or genic sequences were aligned to the reference and used to create neighbour-joining trees (model = 'raw') using the 'ape' package in R, with pairwise.deletion = TRUE (Paradis *et al.* 2014).

Data availability

Arabidopsis lines are available on request. Table S1 contains geographic origin information on the sequenced lines. Short-read sequence data for each line will be deposited in the NCBI short-read archive (SRA) under BioProject number PRJNA288374. A VCF file containing genotype information on the sequenced lines is available on request.

Results

Unique genetic diversity and evidence of outcrossing in North American populations

We found introduced *Arabidopsis* populations have substantial genetic diversity, much of which is unique to the introduced range or at least uncommon in Eurasia. We compared approximately 662 000 SNPs (those with nonmissing genotype calls in at least 50% of individuals) with available large polymorphism data sets for *Arabidopsis* including data from the 18 Eurasian parental accessions used to generate the MAGIC nested association reference panel (Gan *et al.* 2011), the first 80 genomes from the 1001 genomes project (Cao *et al.* 2011) and ~215 000 SNPs generated from the worldwide RegMap panel (Horton *et al.* 2012) (Fig. 1). We found about 81 000 high-confidence SNPs in our North American lines were only present in our data set and not in any other. Only a small number of lines cluster with a common haplogroup (haplogroup 1, Fig. 2) observed in a large portion of North American samples in previous work (Platt *et al.* 2010). The presence of a substantial number of unique SNPs indicates introduced populations are diverse and contain an abundance of new and/or rare mutations that have not been observed in previously sequenced North American or Eurasian natural accessions.

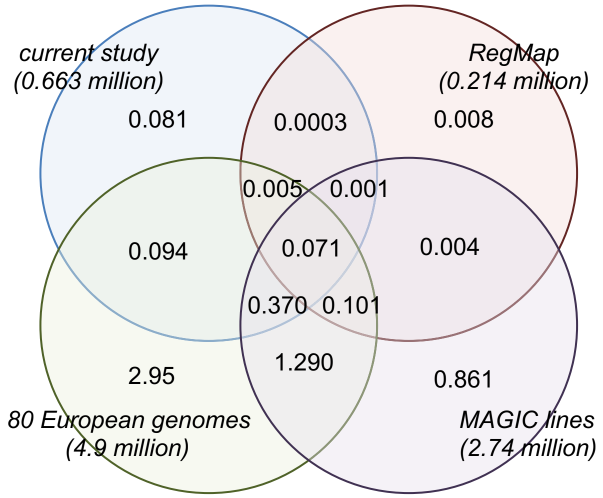


Fig. 1 Overlap between SNPs identified in the current study and three other studies (Cao *et al.* 2011; Gan *et al.* 2011; Horton *et al.* 2012).

On average, short-range LD dropped off at approximately 7 kb on each autosome which is similar to LD estimates of 10 kb in the native range (Kim *et al.* 2007). Based on average pairwise diversity of our sample, the effective population size (N_e) for eastern North America is about 215 000 which is comparable with regional estimates of N_e in Europe (Cao *et al.* 2011). Fixed-window analysis of individual genomes revealed wide regions of unusually high heterozygosity in a single individual from Pennsylvania (PA:DT1-12, Fig. S2) on chromosomes 1, 4 and 5, which is indicative of a recent rare outcrossing event.

Population structure consistent with multiple introductions and mixing among early- and late-flowering groups

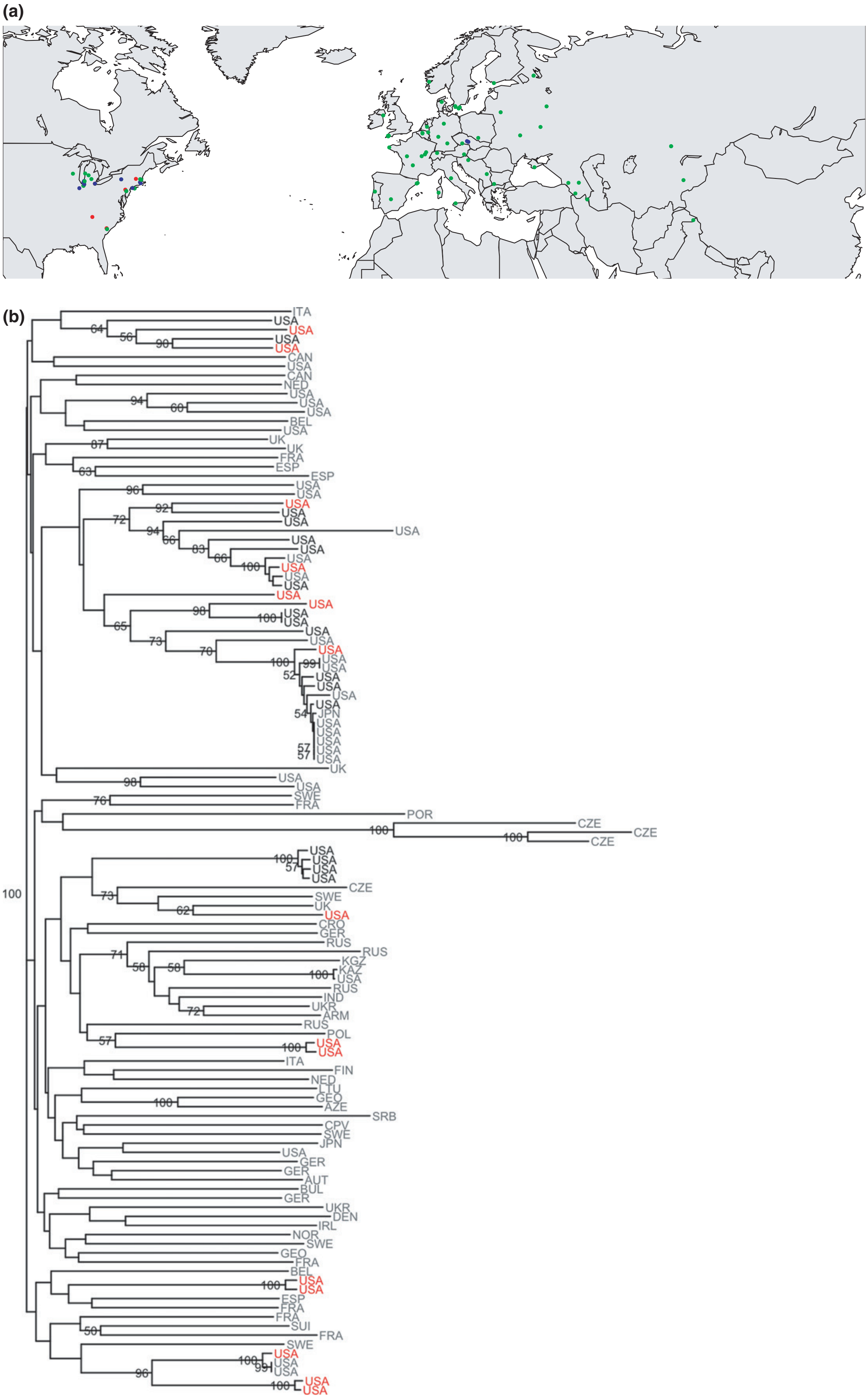
Genomewide data on ~9 800 SNPs showed that natural accessions in the introduced range tend to cluster within individual collection localities, although there is no clear evidence of longitudinal trends in structure assignments (Figs 2A and S3). We also did not observe consistent differences between early- and late-flowering groups in structure ancestry assignments in a way that would suggest that flowering time and population structure covary. Some individuals did not cluster with their nearest geographic neighbours and others were likely to be admixed. For example, two lines from a single population in Massachusetts (HVSQ) each clustered strongly with a different ancestral genetic group (Fig. S3). Some lines from the Midwest (populations Knox and Kin) and from New York (population CSH) also had mixed ancestry.

Comparison of North American lines and a subset of natural accessions from around the world (compared at 126 SNP markers genotyped by Platt *et al.* [34]) showed North American populations cluster with lines from many different parts of Eurasia (Fig. 2, Fig. S4). Clustering was consistent with previous analyses of smaller SNP sets which indicate patterns of isolation by distance at the scale of ~250 km (Kim *et al.* 2007; Beck *et al.* 2008; Platt *et al.* 2010; Samis *et al.* 2012). Genomewide polymorphisms indicated at least three genetic clusters in eastern North America (K = 3) (Fig. S3). There was no indication that early- and late-flowering genotypes stem from independent lineages (Fig S1 B).

Outlier analysis reveals deleterious variants associated with flowering time

We tested whether any variants with predicted deleterious impacts on protein coding genes were overdifferentiated between the early- and late-flowering plant groups and thus potentially linked to adaptation. We considered the impact of a variant potentially deleterious if it caused a frameshift mutation, a gain or loss of a stop or start codon, or disrupted splice sites. As a measure of differentiation between phenotypic groups, we used X^TX (an F_{ST} -like statistic) implemented in the program BAYENV2 (Günther & Coop 2013). X^TX controls for residual population structure through the use of a kinship matrix generated from a subsample of genomewide SNPs. In our simulations, the modal X^TX value for unlinked synonymous coding SNPs was 1.96, close to the expected value for neutral variants ($X^TX = 2.0$), demonstrating adequate control for population structure. In contrast, SNPs predicted to have deleterious impacts on proteins had a right-skewed distribution with modal $X^TX = 2.8$ (Fig. 3). The maximum X^TX value for synonymous SNPs was 2.8, but high-impact variants had values up to 8.1, indicating some are likely to be under selection. There were 996 deleterious variants with differentiation values above the maximum observed at any synonymous SNP and these had putative impacts on 416 protein coding

Fig. 2 Population structure and location of sequenced lines. (A) The color of each marker represents the STRUCTURE group to which each line most likely belongs ($\underline{K}$ = 3 groups; see Fig. S4). (B) Neighbor-joining tree based on 126 SNP markers. Early flowering lines are in red, late flowering lines are in black, lines genotyped by Platt *et al.* are in gray. Lines are labeled with their country of origin. The bracket shows individuals with the most common North American haplotype detected by Platt *et al.* Bootstrap percentages are shown for nodes >50%.


(a)
(b)
ITA
USA
64
USA
56
USA
90
USA
CAN
USA
CAN
NED
94
USA
60
USA
USA
BEL
USA
87
UK
UK
FRA
63
ESP
ESP
96
USA
USA
92
USA
72
USA
USA
94
USA
66
USA
83
USA
66
USA
100
USA
USA
USA
USA
98
USA
100
USA
USA
65
USA
73
USA
70
USA
100
99
USA
USA
52
USA
USA
USA
54
USA
JPN
USA
USA
USA
57
USA
57
USA
UK
98
USA
USA
76
SWE
FRA
POR
100
CZE
CZE
100
CZE
100
100
USA
USA
57
USA
USA
73
CZE
SWE
62
UK
USA
CRO
GER
RUS
71
RUS
58
58
KGZ
100
KAZ
USA
RUS
IND
72
UKR
ARM
RUS
57
POL
100
USA
USA
ITA
FIN
NED
LTU
100
GEO
AZE
SRB
CPV
SWE
JPN
USA
GER
GER
AUT
BUL
GER
UKR
DEN
IRL
NOR
SWE
GEO
FRA
BEL
100
USA
USA
ESP
FRA
FRA
50
SUI
FRA
SWE
100
USA
99
USA
USA
96
100
USA
USA

genes. Among a list of 284 genes with functions related to flowering time, development or environmental sensing (termed *'a priori'* flowering time genes, see Methods), we also scanned for overdifferentiated deleterious variants. Five a priori flowering time candidates contained such variants; however, this number is not more than expected by chance ($\chi^2 = 1.647$, $P = 0.10$). Among the genes containing the most differentiated nonsynonymous SNPs ($n = 156$), there was also no significant enrichment for a priori flowering time candidate genes ($P = 1.0$) or GO terms related to development or environmental sensing.

We also scanned the genome for large structural variants and tested whether any were associated with flowering phenotype. We used the program PINDEL to search for variants larger than the largest insertion or deletion detected using the GATK variant calling pipeline (72 bp). There was strong support for the presence of 3062 large structural variants in the genome, almost 75% of which were deletions with respect to the reference (Fig. S5). Forty-three structural variants were highly differentiated between early- and late-flowering genotypes (Fisher's exact test, lower 1% quantile of *P*-values, $P < 0.0002$) and these were within 5 kb of 60 genes. One affected gene was an a priori flowering time candidate gene, the floral organ identity gene enhancer of HUA2 (HEN2).

Targeted analysis of flowering time candidate genes shows little evidence of selection or association with flowering variation

To identify loci potentially under selection, we compared gene-level nucleotide diversity (π) between the early- and late-flowering plant groups. A reduction in diversity in one group relative to the other may indicate recent selection on a particular locus. Only one out of 284 a priori flowering time candidate genes was within the top 1% of genes with reduced diversity (π) in late-flowering lines, a subunit of the transcription factor *nuclear factor Y* (NF-YA3), also known as HAP2C. Similarly, a single candidate gene fell within the top 1% of genes with reduced diversity in early-flowering lines, the transcription factor *tiny* (TNY) (Fig. S6 A).

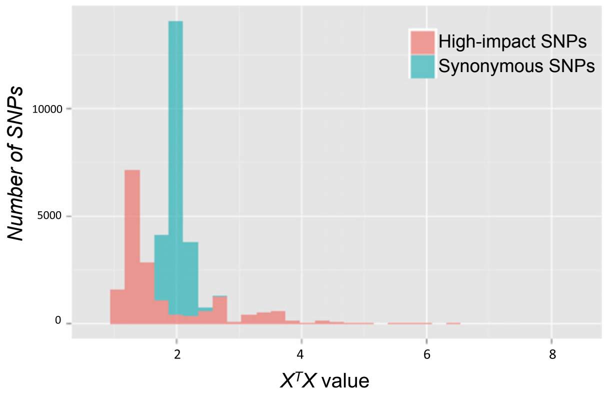


Fig. 3 Distribution of average *XTX* (F_{ST}) values, for synonymous SNPs (blue) and variants with predicted deleterious impacts on protein coding genes (pink).

In the native range of Arabidopsis, at least four genes have alleles that affect flowering time and also occur at moderately high frequencies across the range: FLC, FRI, PHYC and MAF2. We evaluated whether the known common variants at these loci are present in North America and associated with flowering time differences. We also tested for haplotype structuring at these loci indicative that nearby variation is associated with phenotype. We did this by constructing phylogenetic neighbour-joining trees based on alignments of the coding regions and regulatory regions (50 kb up- and downstream) of these genes (Fig. S7). We chose a 50-kb interval based on the method of Li ET AL. 2014 who detected segregating haplotypes near FLC at this scale.

Full sequences of the FRI and FLC genic regions confirmed the results of previous PCR-based genotyping (Samis *et al.* 2012;). FRI contains a ~376-bp deletion that was present in eight lines and a 16-bp insertion present in all but two lines. Neither deletion was strongly associated with flowering time phenotype and analysis of structural variation using PINDEL (Ye *et al.* 2009) did not detect any additional large variants in FRI. We observed several other small putative deletions affecting FRI, but all occurred at low frequency (in one or two individuals).

Within FLC, we did not detect any of the large transposable element insertions that occur within the FLC^A group (or any other large structural variants within the gene) and confer early flowering in the Eurasian range (Gazzani *et al.* 2003; Michaels *et al.* 2003). Haplotypes in the 100-kb region surrounding FLC did not closely match any of the European rapid or slow vernalization haplotypes detected by Li *et al.* (2014) with the exception of a single early-flowering North American line that had strong similarity to the early-flowering rapid-vernalization-2 haplotype (RV2, Fig. S8).

The photoreceptor PHYC contains alleles that vary longitudinally in frequency across Europe and affect flowering in response to day length in some mapping populations (Balasubramanian *et al.* 2006; Samis *et al.* 2008). PHYC is characterized by Ler-type and Col-type haplotypes that are distinguished by a large deletion associated with early-flowering and sensitivity to day length. Analysis of PHYC haplotypes in North America showed that lines can also be distinguished as Ler and Col types, although neither haplotype is strongly associated with early or late flowering (Fig. S9). Four lines from Rhode Island form a unique group outside of the Ler and Col PHYC types, and all are late flowering; however, the association may be due to population structure rather than the presence of a late-flowering functional variant near PHYC. The 100-kb region surrounding PHYC is highly polymorphic containing seven or more haplotype clusters, none of which are characterized by a single flowering time phenotype (B. Gould and J. R. Stinchcombe, unpublished).

Lastly, we examined variation in *MADS AFFECTING FLOWERING 2* (MAF2). Large insertions in MAF2 originating from an adjacent tandemly duplicated paralog, MAF3, are common throughout Eurasia and associated with delayed flowering (Caicedo *et al.* 2009; Rosloski *et al.* 2010). Sequence alignments showed no evidence of MAF2 insertions in any of the North American lines, outside of two regions of microsatellite variation. Neither genic nor upstream MAF2 haplotypes clustered by flowering time phenotype, and PINDEL analysis did not detect any large structural variation within the locus.

Intersection of selection scan results identifies genes in low diversity genomic regions

We calculated nucleotide diversity (π) and the cross-population composite likelihood ratio (XP-CLR) in nonoverlapping windows across the genome to detect regions potentially affected by recent selective sweeps for flowering time. The presence of a recent sweep is indicated where there is a large difference in nucleotide diversity or the XP-CLR between the early- and late-flowering groups relative to the distribution of difference values in other regions in the rest of the genome (Fig. S10). The top 1% of genomewide windows where signatures of selection had the greatest difference between early- and late-flowering lines contained 1210 genes (Fig. 4). These genes are potential candidates for involvement in flowering time evolution, but the set was not enriched for a priori flowering time candidates (n = 13, Fisher's exact test, P = 0.88). There was also no significant enrichment for gene ontology (GO) terms related to development or environmental sensing.

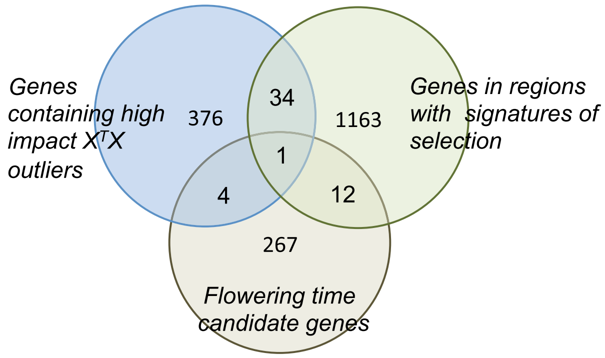

Genes containing high impact X^TX outliers
376
34
1163
Genes in regions with signatures of selection
1
4
12
267
Flowering time candidate genes

Fig. 4 Genes with evidence of involvement in flowering time evolution. The overlap is shown between genes containing highly differentiated variants with predicted deleterious impacts (*XTX* outliers), genes in regions with a signature of selection for flowering time, and genes identified as a priori candidates.

Of the genes potentially under selection, 35 contained one or more variants predicted to have deleterious impacts on protein function and were also overdifferentiated between early- and late-flowering plants ($X^TX > 3.48$; Fig. 3, Table S2). Three additional genes from the set were in close proximity to overdifferentiated structural variants (Table S2). These 38 genes are candidates for involvement in flowering time differentiation in the introduced range.

For the 35 candidate genes with overdifferentiated, deleterious variants, we measured the frequency of the alternate allele both in our sample and in Europe (using data generated by Cao *et al.* 2011 (30)). Most of the deleterious indels had zero frequency in the European data set, but this may be due to low repeatability between current indel prediction methods. Among the overdifferentiated deleterious SNPs in the two ranges (n = 15), the frequency in Europe was predictive of the frequency in North America (Fig. S11). However, the SNPs were not more or less common in the native than the introduced range.

Discussion

Whole-genome sequencing in natural *Arabidopsis* accessions from eastern North America has revealed little evidence that the evolution of flowering time differences in the introduced range has involved selection on known genetic variants from the native range. Some mutations known to have strong influence on flowering time do occur in the introduced range, but we found no evidence that these mutations are associated with differences in flowering time in introduced populations or have been subject to recent hard selective sweeps. Further, regions of the genome containing the strongest signatures of selection do not contain genes that are enriched for functions in development or sensing of environmental cues. Thus, it is unlikely that the evolution of flowering time in the two parts of the range is strongly parallel at the level of genes. Our results therefore suggest that neither pleiotropic constraint nor available standing genetic variation has constrained the evolution of flowering time to selection on high-frequency functional mutations from the native range or known flowering time pathway genes.

We identified regions of the genome with large differences in diversity (π) between phenotypic groups;

however, on average, differences in the top 1% of regions were small and consistent with only putatively weak selection. Also, few a priori flowering time candidate genes showed strong signatures of selection. Given the diversity of introduced populations and evidence of multiple introductions from this and other studies, standing genetic variation has most likely played a central role in the evolution of traits in the introduced range. The absence of a signal of strong selective sweeps at flowering time genes in the introduced range thus might be explained by several nonmutually exclusive mechanisms: (i) evolution through soft rather than hard sweeps; (ii) polygenic adaptation involving many genes of small effect; and (iii) selection on rare alleles present at low frequency in our sample. While soft sweeps, polygenic adaptation and selection on rare alleles can explain why we did not find evidence of selection and differentiation in known flowering time genes, we did find evidence of selection and differentiation in novel genes. We consider both aspects below.

Soft sweeps, G × E and polygenic adaptation

Soft selection on standing genetic variation is probable in introduced *Arabidopsis* which has a history of multiple introductions and is likely to have experienced introduction of multiple adaptive alleles at many loci. Signatures of soft selection may not be strong enough to differentiate from patterns produced by drift using diversity scans. However, haplotype structuring around loci under soft selection may reveal phylogenetic grouping by phenotype (Messer & Petrov 2013). Haplotype structuring around candidate genes FRI, FLC, PHYC and MAF2 did not reveal any evidence of haplotype clustering by phenotype; however, some structuring was apparent at two developmental transcription factor genes found in regions of the genome with the strongest selection signatures (TFIIIA and HAP2C, Fig S12). Soft sweeps at these and other loci may have contributed to flowering time adaptation in the introduced range and warrant further investigation.

Another possibility is that the planting season used in generating our original flowering phenotypes (fall planting, overwintering rosettes, intended to simulate vernalizing, ecologically realistic conditions) altered the expression of variation due to these candidate genes (e.g. Wilczek *et al.* 2009). Under alternative experimental conditions (fall vs. spring generations, earlier or later planting in the fall, etc.), different flowering phenotypes may be expressed, with these genes potentially playing a role in those environments. Stock *et al.* (2015) found a significant across-environment genetic correlation between flowering time in the common garden and a growth chamber and significant genotype by environment interaction for flowering time, suggesting that environment-specific genetic variation contributes to the flowering phenotype.

Likewise, allele frequency changes in many genes of small effect on flowering time may go undetected in outlier scans. The importance of polygenic adaptation in evolution in response to new environments is currently unknown, but may be the predominant form of quantitative trait evolution from standing genetic variation (Hancock *et al.* 2010; Pritchard & Di Rienzo 2010; Pritchard *et al.* 2010). Studies of human height (Lango Allen *et al.* 2010; Turchin *et al.* 2012) reveal how allele frequency differences observed across many loci are often quite small. For instance, allele frequency differences between northern and southern human populations ranged from 0.0044 to 0.016 with sample sizes ranging from 58 to 257 in each population. Our study likely lacked statistical power to detect subtle allele frequency changes across many genes characteristic of polygenic adaptation. A lack of strong signatures of selection in candidate genes may mean that adaptive polymorphisms are abundant but independently relatively weak, each contributing little to overall adaptive differentiation in flowering time.

In addition to power considerations, polygenic adaptation may contribute in ways that are difficult to discern from genomic portraits of polymorphism and divergence. Theoretical work by Latta (1998, 2004) and LeCorre and colleagues (Kremer & Le Corre 2012; Le Corre & Kremer 2012) has demonstrated that trait divergence can evolve even with limited differentiation of the underlying causal loci. In brief, because selection generates covariances among different loci, changes in genetic variances, and hence quantitative genetic differentiation, can far outpace allele frequency changes in individual loci. Consequently, outlier-based approaches, such as ours, are likely to represent a conservative lower bound of the loci involved in adaptation.

Selection on locally common but regionally rare alleles

In the absence of any strong evidence for a parallel genetic basis of flowering time clines, selection on locally common but regionally rare alleles may explain a portion of the evolution of flowering time in the introduced range. Adaptive alleles and haplotypes that occur only in small areas of the introduced range would not occur often enough within our set of introduced lines to show a signal of overdifferentiation or reduced diversity. In the native range, previous studies have shown that alleles that are regionally restricted in their distribution explain up to 24% of the variation in some fitness-related traits (Fournier-Level *et al.* 2011; Li *et al.* 2014). In these studies, the alleles are not associated with signatures of selection and are rarely associated with fitness in more than one region, and their distribution is also often spatially correlated with climate factors. In the introduced range, the evolution of flowering time may follow a similar pattern and would be facilitated by the introduction of adaptive alleles from throughout the native range. For example, we found some evidence of uncommon deletions in *FRI* in our sample, some of which could be at high frequency at a local scale and contribute to flowering time phenotype only in certain populations. Rare mutations in both FRI and FLC have often been associated with flowering time phenotype in *Arabidopsis* accessions from Eurasia (e.g. Le Corre 2005; Sánchez-Bermejo *et al.* 2012). Also, in a study of *Arabidopsis* lines from France, Brachi *et al.* 2013 found genomic regions associated with phenology varied at a local scale. In the future, greater within-population sampling of *Arabidopsis* in North America will be able to test the hypothesis that adaptive alleles are regionally rare but locally abundant.

New candidates for flowering time evolution in the introduced range

Despite the limitations of outlier-based genome scans, many of the loci detected in the present study are likely to represent true recent targets of selection. Outlier-based selection scans are most powerful at detecting selection that results from newly arising mutations or similarly, those that exist within bottlenecked populations prior to demographic expansion (Teshima *et al.* 2006; Jensen *et al.* 2007; Pritchard & Di Rienzo 2010; Lotterhos & Whitlock 2015). The signatures of full or partial selective sweeps in the genome are expected to persist on the order of N_e generations (Przeworski 2002). Introduced *Arabidopsis* has a relatively large effective population size in eastern North America (comparable to regional diversity in Europe, N_e ~215 000 in this study) and a history of very recent introduction over the past ~250 years (Jørgensen & Mauricio 2004). Prior to population expansion, sweeps should be detectable but rare in the introduced range. Results of outlier tests are not likely to detect all targets of selection for flowering time evolution in the introduced range, but the use of multiple test statistics, as in this study, encourages a high rate of true positives among outlier loci (Teshima *et al.* 2006).

We identified a set of 38 candidate genes impacted by large or small overdifferentiated variants that also show signatures of selection (Table S2) for flowering time. The list includes one a priori candidate gene: NF-YA (HAP2C). Nuclear factor Y is a transcription factor that has been shown to influence flowering time through interaction with the major flowering time gene *Constans* (Kumimoto *et al.* 2010). The candidate genes also include two heat-shock proteins, a thigmomorphogenesis locus (involved in response to mechanical stimuli) and a MADS-box transcription factor. The candidate genes generated from this work provide targets for future investigation and functional validation in the study of the genetic basis of flowering time adaptation in introduced *Arabidopsis*. Importantly, the genetic tools available for Arabidopsis will facilitate functional verification of these candidates through knockout studies and other approaches.

Nonparallelism at the level of genes in the evolution of flowering time clines in the native and introduced range of *Arabidopsis* is consistent with the results of a recent study by Hamilton *et al.* (2015). In a transplant of Eurasian genotypes to a common garden in Rhode Island, they found that adaptive variants in the native range differed from adaptive variants in Rhode Island, despite similar climate conditions. Most of the adaptive alleles in Europe and Rhode Island exhibited conditional neutrality: beneficial in one site, with little or no effect in the other. The prevalence of conditional neutrality underlying local adaptation has important consequences for evolution in introduced species. If loci underlying local adaptation in the native range frequently exhibit conditional neutrality (Anderson *et al.* 2011, 2013; e.g. Fournier-Level *et al.* 2011), one would predict that they would be locally abundant, but possibly not have spread rangewide. For species with a history of multiple introductions, like *Arabidopsis*, a complex mosaic of previously locally adapted yet globally rare alleles may underlie any postintroduction evolutionary responses. In these scenarios, evolution of similar broad geographic patterns through parallel genetic mechanisms may be unlikely.

Collectively, our results show that postintroduction evolution of flowering time has been neither confined by pleiotropic constraint to a handful of loci previously implicated in flowering time evolution, nor constrained by limited standing genetic variation. Further studies on the genetic basis of adaptation in complex, quantitative traits will be necessary to determine the prevalence of this pattern.

Acknowledgements

We would like to thank Adriana Salcedo, Stephen Wright, Wei Wang, Robert Williams, Emily Josephs and Maggie Bartkowska for help and advice with analyses. We would also like to thank three anonymous reviewers for helpful comments. This project was funded by NSERC Canada.

References

Anderson JT, Willis JH, Mitchell-Olds T (2011) Evolutionary genetics of plant adaptation. *Trends in Genetics*, **27**, 258–266.

Anderson JT, Lee C-R, Rushworth CA, Colautti RI, Mitchell-Olds T (2013) Genetic trade-offs and conditional neutrality contribute to local adaptation. *Molecular Ecology*, **22**, 699–708.

Aranzana MJ, Kim S, Zhao K *et al.* (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics*, **1**, e60.

Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation?. *Trends in Ecology & Evolution*, **23**, 26–32.

Atwell S, Huang YS, Vilhjálmsson BJ *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Balasubramanian S, Sureshkumar S, Agrawal M *et al.* (2006) The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*. *Nature Genetics*, **38**, 711–715.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38–44.

Beck JB, Schmuths H, Schaal BA (2008) Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Molecular Ecology*, **17**, 902–915.

Bossdorf O, Auge H, Lafuma L, Rogers WE, Siemann E, Prati D (2005) Phenotypic and genetic differentiation between native and introduced plant populations. *Oecologia*, **144**, 1–11.

Brachi B, Faure N, Horton M *et al.* (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genetics*, **6**, e1000940.

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, **12**, 232.

Brachi B, Villoutreix R, Faure N *et al.* (2013) Investigation of the geographical scale of adaptive phenological variation and its underlying genetics in *Arabidopsis thaliana*. *Molecular Ecology*, **22**, 4222–4240.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdos Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Buswell JM, Moles AT, Hartley S (2011) Is rapid evolution common in introduced plant species? *Journal of Ecology*, **99**, 214–224.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD (2004) Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 15670–15675.

Caicedo AL, Richards C, Ehrenreich IM, Purugganan MD (2009) Complex rearrangements lead to novel chimeric gene fusion polymorphisms at the *Arabidopsis thaliana* MAF2-5 flowering time gene cluster. *Molecular Biology and Evolution*, **26**, 699–711.

Cao J, Schneeberger K, Ossowski S *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **10**, 956–965.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.

Cingolani P, Platts A, Wang LL *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 1–13.

Colautti RI, Barrett SCH (2013) Rapid adaptation to climate facilitates range expansion of an invasive plant. *Science*, **342**, 364–366.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, **26**, 298–306.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.

Gan X, Stegle O, Behr J *et al.* (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, **477**, 419–423.

Gazzani S, Gendall AR, Lister C, Dean C (2003) Analysis of the molecular basis of flowering time variation in Arabidopsis accessions. *Plant Physiology*, **132**, 1107–1114.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Hamilton JA, Okada M, Korves T, Schmitt J (2015) The role of climate adaptation in colonization success in *Arabidopsis thaliana*. *Molecular Ecology*, **9**, 2253–2263.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A (2010) Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 2459–2468.

Horton MW, Hancock AM, Huang YS *et al.* (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics*, **44**, 212–216.

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44–57.

Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. *Trends in Genetics*, **23**, 568–577.

Jørgensen S, Mauricio R (2004) Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured. *Molecular Ecology*, **13**, 3403–3413.

Kim S, Plagnol V, Hu TT *et al.* (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, **39**, 1151–1155.

Kremer A, Le Corre V (2012) Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity*, **108**, 375–385.

Kumimoto RW, Zhang Y, Siefers N, Holt BF (2010) NF-YC3, NF-YC4 and NF-YC9 are required for CONSTANS-mediated, photoperiod-dependent flowering in *Arabidopsis thaliana*. *The Plant Journal*, **63**, 379–391.

Lango Allen H, Estrada K, Lettre G *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *The American Naturalist*, **151**, 283–292.

Latta RG (2004) Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist*, **161**, 51–58.

Le Corre V (2005) Variation at two flowering time genes within and among populations of *Arabidopsis thaliana*: comparison with markers and traits. *Molecular Ecology*, **14**, 4181–4192.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

Lee CE (2002) Evolutionary genetics of invasive species. *Trends in Ecology & Evolution*, **17**, 386–391.

Lee YW, Gould BA, Stinchcombe JR (2014) Identifying the genes underlying quantitative traits: a rationale for the QTN programme. *AoB Plants*, **6**, plu004.

Li P, Filiault D, Box MS, Kerdaffrec E, van Oosterhout C *et al.* (2014) Multiple FLC haplotypes defined by independent cis-regulatory variation underpin life history diversity in *Arabidopsis thaliana*. *Genes & Development*, **28**, 1635–1640.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Maron J, ViIá MO, Bommarco R, Elmendorf S, Beardsley P (2004) Rapid evolution of an invasive plant. *Ecological Monographs*, **74**, 261–280.

Maron JL, Elmendorf SC, Vilá M (2007) Contrasting plant physiological adaptation to climate in the native and introduced range of *Hypericum perforatum*. *Evolution*, **61**, 1912–1924.

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, **28**, 659–669.

Michaels SD, He Y, Scortecci KC, Amasino RM (2003) Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. *PNAS*, **17**, 10102–10107.

Mitchell-Olds T, Schmitt J (2006) Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature*, **441**, 947–952.

Montague JL, Barrett SCH, Eckert CG (2008) Re-establishment of clinal variation in flowering time among introduced populations of purple loosestrife (Lythrum salicaria, Lythraceae). *Journal of Evolutionary Biology*, **21**, 234–245.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Orr HA (2005) The probability of parallel evolution. *Evolution*, **59**, 216–220.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Picardtools (2015) http://broadinstitute.github.io/picard.

Platt A, Horton M, Huang YS *et al.* (2010) The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, **6**, e1000843.

Pritchard JK, Di Rienzo A (2010) Adaptation – not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *American Journal of Human Genetics*, **67**, 170–181.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. *Genetics*, **160**, 1179–1189.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rhee SY, Beavis W, Berardini TZ *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, **31**, 224–228.

Rosloski SM, Jali SS, Balasubramanian S, Weigel D, Grbic V (2010) Natural diversity in flowering responses of *Arabidopsis thaliana* caused by variation in a tandem gene array. *Genetics*, **186**, 263–276.

Salomé PA, Bomblies K, Laitinen RAE, Yant L, Mott R, Weigel D (2011) Genetic architecture of flowering-time variation in *Arabidopsis thaliana*. *Genetics*, **188**, 421–433.

Samis KE, Heath KD, Stinchcombe JR (2008) Discordant longitudinal clines in flowering time and phytochrome C in *Arabidopsis thaliana*. *Evolution*, **62**, 2971–2983.

Samis KE, Murren CJ, Bossdorf O *et al.* (2012) Longitudinal trends in climate drive flowering time clines in North American *Arabidopsis thaliana*. *Ecology and Evolution*, **2**, 1162–1180.

Sánchez-Bermejo E, Méndez-Vigo B, Picó FX, Martinez-Zapater JM, Alonso-Blanco C (2012) Novel natural alleles at FLC and LVR loci account for enhanced vernalization responses in *Arabidopsis thaliana*. *Plant, Cell & Environment*, **35**, 1672–1684.

Shindo C, Aranzana MJ, Lister C *et al.* (2005) Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time. *Plant Physiology*, **138**, 1163–1173.

Simpson GG, Dean C (2002) Arabidopsis, the Rosetta stone of flowering time? *Science*, **296**, 285–289.

Stock AJ, McGoey BV, Stinchcombe JR (2015) Water availability as an agent of selection in introduced populations of *Arabidopsis thaliana*: impacts on flowering time evolution. *PeerJ*, **3**, e898.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.

Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN (2012) Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics*, **44**, 1015–1019.

Wilczek AM, Roe JL, Knapp MC *et al.* (2009) Effects of genetic perturbation on seasonal life history plasticity. *Science*, **323**, 930–934.

Wood TE, Burke JM, Rieseberg LH (2005) Parallel genotypic adaptation : when evolution repeats itself. *Genetica*, **123**, 157–170.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

J.S. and B.G. conceived and designed the experiments. B.G. conducted the data analysis. J.S. and B.G. authored the manuscript.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Sequenced lines.

Fig. S2 Heterozygosity across each chromosome in 4 lines from Pennsylvania.

Fig. S3 STRUCTURE group memberships by population.

Fig. S4 STRUCTURE plot of 25 sequenced North American lines and 95 lines genotyped by Platt *et al*. (2010).

Fig. S5 Structural variants.

Fig. S6 Distribution of differences in diversity between early and late flowering lines for genes across the genome.

Fig. S7 Neighbor-joining trees based on the sequences at three candidate genes.

Fig. S8 Neighbor-joining tree based on variation in the 100 kb region surrounding FLC.

Fig. S9 Neighbor-joining tree based on genic sequences of PHYC.

Fig. S10 Signatures of selection.

Fig. S11 Alternate allele substitution frequencies of XTX outlier SNPs in 35 candidate genes.

Fig. S12 Neighbor-joining trees based on the genic sequence of three new candidate genes.

Table S1. Origin and flowering time of sequenced lines.

Table S2. 38 candidate genes and their putative high impact, over-differentiated variants.