A multilocus association analysis method integrating phenotype and expression data reveals multiple novel associations to flowering time variation in wild-collected Arabidopsis thaliana

Yanjun Zan | Örjan Carlborg

Abstract

The adaptation to a new habitat often results in a confounding between genomewide genotype and beneficial alleles. When the confounding is strong, or the allelic effects is weak, it is a major statistical challenge to detect the adaptive polymorphisms. We describe a novel approach to dissect polygenic traits in natural populations. First, candidate adaptive loci are identified by screening for loci directly associated with the adaptive trait or the expression of genes known to affect it. Then, a multilocus genetic architecture is inferred using a backward elimination association analysis across all candidate loci with an adaptive false discovery rate-based threshold. Effects of population stratification are controlled by accounting for genomic kinship in both steps of the analysis and also by simultaneously testing all candidate loci in the multilocus model. We illustrate the method by exploring the polygenic basis of an important adaptive trait, flowering time in Arabidopsis thaliana, using public data from the 1,001 genomes project. We revealed associations between 33 (29) loci and flowering time at 10 (16)°C in this collection of natural accessions, where standard genomewide association analysis methods detected five (3) loci. The 33 (29) loci explained approximately 55.1 (48.7)% of the total phenotypic variance of the respective traits. Our work illustrates how the genetic basis of highly polygenic adaptive traits in natural populations can be explored in much greater detail using new multilocus mapping approaches taking advantage of prior biological information, genome and transcriptome data.

KEYWORDS

Arabidopsis thaliana, expression QTL, flowering, genomewide association analysis, polygenic

1 | INTRODUCTION

The genomewide association study (GWAS) is a powerful approach to dissect the genetic basis of phenotypic variation in populations, but its application in natural populations is statistically challenging. For example, when populations develop local adaptations to their environment, this often results in a confounding between adaptive alleles and the genomewide genotype of the local population. This confounding is particularly challenging when adaptation of highly polygenic traits is a consequence of small changes in the allele...
frequencies of standing variants across many loci (Burke, Liti, & Long, 2014; Orozco-terWengel et al., 2012; Teotónio, Chelo, Bradić, Rose, & Long, 2009). The power to detect loci making small contributions to the trait variation is generally low in GWAS due to corrections for multiple testing. The additional burden of corrections for population structure decreases power further. By making combined use of powerful model populations, and new analytical approaches accounting for the polygenic nature of complex adaptive traits, it is possible to dissect a polygenic genetic architecture with greater sensitivity to facilitate deeper insights into the genetic basis of adaptive traits (Sheng, Pettersson, Honaker, Siegel, & Carlborg, 2015; Zan et al., 2017). These powerful statistical methods have not yet been adapted to analyses of natural populations, but could also be useful for identifying the genetic mechanisms that govern adaptation of polygenic traits in nature.

*Arabidopsis thaliana* is a widely used model species in plant biology. It has colonized a wide range of ecological habitats around the world, and the available genotype, expression and phenotype data on many of these ecotypes provide unique opportunities to study the genetic basis of adaptation. Flowering time is one of the most studied adaptive traits in this species as it has an important role in ecological adaptation and also a potential impact on agronomic production in related species. Studies in the laboratory, often on the reference accession Col-0, have identified many genes that influence flowering time by controlling multiple biological pathways, for example, photoperiod (El-Assal, Alonso-Blanco, Peeters, Raz, & Koornneef, 2001; Filiault et al., 2008), vernalization (Li et al., 2014; Shindo et al., 2005) and plant hormone signalling (Sharma et al., 2016) have been identified. Studies of flowering time variation in nature include both QTL mapping in crosses between divergent accessions and genomewide association (GWA) studies in collections of wild accessions (Alonso-Blanco et al., 2016; Atwell et al., 2010; Brachi et al., 2010; Salomé et al., 2011). The molecular studies suggest that hundreds of genes could potentially influence flowering time whereas the studies of natural populations have only been able to associate genetic polymorphisms in a handful of genes with large variation in worldwide flowering time. Hence, even though flowering time is also potentially highly polygenic in nature, the burden of multiple-testing and population stratification correction in GWAS analyses of natural associations is too high to identify adaptive alleles that are either rare, have smaller effects or are confounded with population structure even in collections of >1,000 inbred accessions (Alonso-Blanco et al., 2016).

Here, we describe an analytical approach to dissect the genetic basis of polygenic adaptive traits in natural populations. Its use is illustrated by dissecting flowering time variation in a publicly available data set of wild-collected *A. thaliana* accessions from the 1,001 genomes project (Alonso-Blanco et al., 2016). By utilizing prior information on known flowering time genes and loci in *A. thaliana* and integrating it in our analysis of genotype, expression and trait variation data from the 1,001 genomes project, a highly polygenic basis of the worldwide variation in flowering time was revealed.

## 2 MATERIALS AND METHODS

### 2.1 Data

We reanalysed a recently released data set on 1,004 wild-collected *A. thaliana* accessions from the 1,001 genomes project (Alonso-Blanco et al., 2016). All genotype, phenotype and transcriptome data are publicly available within the *Arabidopsis thaliana* 1,001 genomes project (Alonso-Blanco et al., 2016; Kawakatsu et al., 2016).

The genotypes for the accessions were downloaded in the form of a whole-genome SNP data matrix from http://1001genomes.org/data/GMIMPI/releases/v3.1/SNP_matrix_imputed_hdf5/1001_SNPs_MATRIX.tar.gz. These genotypes were obtained from whole-genome resequencing data, by calling SNPs using the MPI-SHORE and GMI-GATK pipelines and then imputing the whole-genome SNP genotypes (Alonso-Blanco et al., 2016). We filtered these SNPs for minor allele frequency (MAF > 0.03) and pairwise LD ($r^2 > .99$) to obtain genotypes for 1,396,438 SNPs on 1,004 individuals.

Flowering times were measured as days to flowering for plants grown in the growth chambers with following settings: 6 days of stratification in the dark at 4°C, followed by growth under constant temperature of 10°C or 16°C with 16 hr light/8 hr darkness (Alonso-Blanco et al., 2016). The flowering time phenotypes were downloaded from https://arapheno.1001genomes.org/phenotype/261/ and https://arapheno.1001genomes.org/phenotype/262/. Plants used for RNA sequencing were grown in the growth chambers with 16 hr of light and 8 hr of darkness at 22°C (S. C. Huang, personal communication). Leaves of intermediate ages (not the youngest or the oldest) were sampled before bolting, from which RNA was prepared and sequenced (Kawakatsu et al., 2016). In total, 728 transcriptomes were downloaded from NCBI (GSE80744), and 648 of these were used in the expression QTL mapping analysis as the remaining ones had no matching genotypes in the SNP data set. Untransformed Reads Per Kilobase per Million mapped reads (RPKM) values were used in the association analyses and for 272 (97%) of the evaluated genes, <50 accessions had RPKM = 0. Zero-inflated distributions were therefore not a major concern for the evaluated genes. Expression distributions for significant genes were further visually inspected before accepting them as candidate loci for the polygenic analysis, and none of them showed severe signs of zero inflation. We also downloaded a functional description file from TAIR FTP site (ftp://ftp.arabidopsis.org/home/tair/TAIR_Public_Releases/TAIR_Data_20140331/TAIR10_functional_descriptions_20140331.txt) for annotating relevant genes in our study. Information on TAIR 10 annotation (The Arabidopsis Information Resource version 10) and Gene Ontology (GO) annotation were extracted from this file for genes listed in supporting tables.

### 2.2 An analytical framework for inferring polygenic genetic architectures

Described here is an analytical framework for inferring polygenic architectures in natural populations. A multilocus genetic model is
developed at a predefined false discovery rate significance threshold using a backward elimination approach with an adaptive model selection criterion (Abramovich, Benjamini, Donoho, & Johnstone, 2006; Gavrilov, Benjamini, & Sarkar, 2009; Sheng et al., 2015; Zan et al., 2017). This analysis is performed across a set of candidate loci that, based on prior biological and statistical evidence, are considered likely to influence the trait in the population. Here, stringent preselection is implemented and loci included should either (i) have genomewide significant effects on the trait in a single-locus genomewide association analysis or (ii) have genomewide significant effects on the expression of a gene known to affect the studied trait in earlier experiments, both in the studied population.

2.2.1 | Genomewide association analysis for flowering time

Previous genomewide association (GWA) analysis in this data set reported five genomewide significant associations for the flowering time at 10°C (16°C) using a mixed model analysis to account for population structure (Alonso-Blanco et al., 2016). We here extended these analyses for these traits by implementing a step-wise conditional GWA mapping. At initiation, all loci already detected in the single-locus GWAS were included as cofactors in the analysis model. Then, one or several conditional GWA analyses were performed, where in each round all loci that passed the multiple-testing corrected genomewide significance threshold were added to the model. For FT10, two conditional analyses were performed and for FT16 only one. The significance threshold for each scan was calculated using a Bonferroni correction for the number of SNPs with \( r^2 < .95 \), resulting in a threshold of \( -\log_{10}(p) = 7.33 \). The model used in the GWA analyses was

\[
y = X\beta + Zu + e. \tag{1}
\]

Where is the mean flowering time at 10°C (16°C), \( X \) the design matrix with the number of columns equal to the number of already selected SNPs plus one for the SNP tested for inclusion (coded as 0, 2 for minor allele homozygous, major allele homozygous genotypes, respectively), \( \beta \) is a vector of the effects of substituting two alleles at the corresponding SNPs in \( X \), and \( Z \) is the design matrix obtained from a Cholesky decomposition of the \( G \) matrix (identity by state—IBS) kinship matrix estimated from the whole-genome SNP data with the \( \text{mmscore} \) function (weighted by allele frequencies) in the \text{GENABEL} package (Aulchenko et al., 2007). The \( Z \) matrix satisfies \( ZZ' = G \), thus, the random effect vector \( u \) will be normally distributed with variance \( \sigma_e^2 \), \( u \sim N(0, \sigma_u^2) \), and \( e \) is the normally distributed residual with variance \( \sigma_r^2 \), \( e \sim N(0, \sigma_e^2) \). The analyses were performed using the \text{polygenic} and \text{mmscore} functions in the \text{GENABEL} package (Aulchenko et al., 2007). The \text{polygenic} function was run only once to estimate the \( \sigma_r^2 \) and \( \sigma_u^2 \), and then \text{mmscore} function, implementing iterative scans for each SNP, was performed to obtain the genomewide association statistics. In those iterative scan, these two parameters, \( \sigma_e^2 \) and \( \sigma_{ug}^2 \), were assumed to be constant. Hence, this approach is essentially a linear model associating SNPs to the phenotypic residuals, which excluded the polygenic effects.

2.2.2 | Expression QTL mapping targeting genes in flowering time-related pathways

A list of 282 flowering time genes was obtained from (Brachi et al., 2010). From the available transcriptome sequence data in the 1,001 genomes project, the expression levels were extracted for those genes as described in (Zan, Shen, Forsberg, & Carlborg, 2016). Whole-genome eQTL mapping was then performed for these genes across the pruned SNP set \( (r^2 > .99) \) described in the Data section above. In the analysis, model 1 was used but with a slightly altered design matrix \( X \). Here, \( X \) was a column vector containing only the genotype of the tested SNP, coded as 0, 2 for minor allele homozygous, major allele homozygous genotypes. As in the GWA, a Bonferroni corrected significance threshold was used and calculated based on the number of SNPs in the data set with \( r^2 < .95 \), resulting in a multiple-testing corrected threshold of \( -\log_{10}(p) = 7.33 \). These analyses were also performed using the \text{polygenic} and \text{mmscore} functions in the \text{GENABEL} package (Aulchenko et al., 2007).

2.2.3 | Multilocus backward elimination association analysis

We have previously developed a multilocus approach to explore the genetic architecture of highly polygenic traits, and a detailed description of the approach is available in (Brandt, Ahsan, Honaker, & Siegel, 2017; Lillie et al., 2017; Sheng et al., 2015). In short, the method was designed to study traits where earlier data suggest them to be polygenic in the analysed population. In such cases, standard association analysis approaches to detect individual loci using stringent multiple-testing corrected significance thresholds will have low power to infer the true genetic architecture of the traits. This is because the many loci contributing to a polygenic trait will make small individual contributions to the genetic variance. By implementing a multilocus mapping approach in a backward elimination model selection framework, it is possible to simultaneously account for the joint genetic effects of many loci and identify those that together contribute to the trait variance. Not known in advance is how many loci will contribute to the trait and therefore the number of loci in the final model could vary substantially. To account for this, this analysis was implemented with an adaptive model selection criterion to control the false discovery rate (FDR) in the final model (Abramovich et al., 2006; Gavrilov et al., 2009). Using a 15% FDR as the termination criterion of the analysis, most of the selected loci are expected to be true associations to the trait. This makes it possible to gain overall insights to the polygenic basis of the studied trait from the mapped loci, but as the significance of individual loci is not tested, it is not suitable to use the analysis to make conclusions about individual loci.

The multilocus analysis was here extended to identify a multilocus model across loci with direct associations to flowering time and
loci that might influence flowering time via their transcriptional regulation of known flowering genes. The major focus of the multilocus analysis was here to evaluate the joint contribution of all these loci to flowering time variation. In particular, the analysis would reveal if using information on flowering gene, eQTL was important for dissecting the genetic basis of flowering time variation in this data set.

The first step of the multilocus analysis was a prescreening, where the 150 significant eQTL markers were divided into three sets with approximately 50 markers each. The markers were divided to include markers on the same chromosomes. This step was used since a backward elimination analysis across >100 markers would be unstable due to an overparameterized linear model. For each of these prescreening sets, a backward elimination analysis was performed with termination based on the adaptive FDR criterion. A slightly lenient threshold was applied to include as many markers as possible, hence, markers reaching the 20% FDR in any of these analyses were kept for joint analyses.

The second step of the analysis was a full analysis, where all markers selected in the prescreening of the eQTL loci, and the loci with direct associations to flowering time were analysed jointly. The end result was a set of loci that together contribute to flowering time variation in this population at a 15% FDR. All analyses were implemented in custom scripts in the statistical software R (R Core Team 2015), and these are available in the R package BE on Github (https://github.com/yanjunzan/BE). The fixed additive genetic effects of those 33 loci, modelled as the difference between the two homozygous genotypes, and the corresponding p value of all associations to flowering time were analysed jointly. The proportion of the total variance explained by these sets was estimated by fitting the mixed model

\[ y = \mu + X_8 \beta_8 + Zu + e \]

where \( R^2 \) is the variance explained by set \( S \), and \( X_8 \) and \( \beta_8 \) are the design matrix and estimates for set \( S \) from model (2). The variance explained by the set of 25 eQTL SNPs selected in the multilocus analysis (V25) was calculated as \( V_{25} = V_{33} - V_8 \). We also estimated the variances explained by the sets of markers obtained by incrementally adding each of the 33 SNP markers from the selected multilocus model, in the order from largest to smallest additive effect size obtained when fitting all loci together, to \( S \). In each step, the genotype for the added marker was included in \( X_8 \) in the mixed model (2) and the variance of the set estimated using (3). The obtained estimates for these 33 sets are reported in Figure 4b. All variance estimation analyses were implemented in custom R scripts where the mixed models were fitted using the R package HGLM (Rönnegård et al., 2010).

2.4 Visualization of the results

Figure 3 was made using the R package circlize (Gu, Gu, Eils, Schlesner, & Brors, 2014).

3 Results

We develop a new approach to dissect the polygenic basis of complex traits in natural populations and illustrate its use by analysing flowering time variation in the latest release of the 1,001 genomes A. thaliana data set. First, candidate adaptive loci were identified as those with individual associations to flowering time, or associations with the expression of known flowering time genes, in this population. Then, the polygenic basis of flowering time was revealed using a multilocus association analysis across the selected candidate loci.
3.1 A genomewide association analysis to detect loci associated with flowering time

A publicly available data set with genotypes and phenotypes measured on a collection of 1,004 natural Arabidopsis thaliana accessions was reanalysed. Two flowering time phenotypes measured under greenhouse conditions, flowering time at 10°C (FT10) and 16°C (FT16), were used in a previous genomewide association study by (Alonso-Blanco et al., 2016). Here, we illustrate our new analytical approach by reanalysing these phenotypes and describe the results for FT10 in more detail. A more schematic summary of the results is provided for FT16.

Five loci were associated with FT10 in (Alonso-Blanco et al., 2016) using a mixed model analysis accounting for population structure (Figure 1a; Figure S1). To select the best possible candidates for the multilocus analysis to reveal the polygenic basis of flowering time in this population, we also explore the genetic architecture of FT10 beyond those loci. First, the five earlier reported loci were included as cofactors in the genetic model. Then, the genome was scanned iteratively to identify additional loci directly associated with FT10 (inflation factors provided in Figure S2). In this scan, we eased the burden of multiple-testing correction by only correcting for tests performed on markers in LD < 0.95, resulting in a selection threshold for terminating the forward selection scan of $-\log_{10} p = 7.33$. Two additional loci were associated with FT10 at this significance level after correction for population structure via a mixed model analysis (Figure 1b). One of these associations (peak on chromosome 4, 17,282,047 bp) overlapped with the confidence interval of an earlier reported flowering time QTL in A. thaliana recombination inbred lines (Salomé et al., 2011; Table S1). In addition to these loci, a third locus (peak on chromosome 4, 10,999,188 bp; Figure 1c) had a p-value ($-\log_{10} p = 7.0$) near the selection threshold. Despite not reaching the threshold, it was selected as a candidate for the multilocus analysis motivated by its overlap with the confidence interval of an earlier reported flowering time QTL (Salomé et al., 2011; Table S1) and it being located only 3.8 kb upstream of the known flowering time gene TSF (Twin sister of FT, AT4G20370 (Ando et al., 2013)).

3.2 Loci associated with expression variation in earlier reported flowering time genes

Molecular studies in the reference Arabidopsis thaliana accession Col-0 have revealed many genes that are directly or indirectly involved in the regulation of flowering time (Brachi et al., 2010). Candidate loci for the polygenic analysis of flowering time were identified by screening the 1,001 genomes data for polymorphisms associated with expression variation in 282 known flowering time genes (Brachi et al., 2010) using a mixed model-based expression QTL (eQTL) mapping analysis accounting for population structure. In total, 150 eQTL regulated the expression of 123 flowering time genes at genomewide significance (Table S2; inflation factors in Figure S2D). One locus on chromosome 5 (18,590,327 bp) regulated expression of the well-studied flowering time gene DOG1 (AT5G45830). Another locus on chromosome 5 (22,603,100 bp) acted as a trans-eQTL hotspot regulating the expression of 33 flowering time genes (Figure 2; Table S3). Among the 33 genes regulated by this trans-eQTL, 19 have a TAIR10 annotation/GO term related to flowering time via light-mediated developmental, or photoperiod control, pathways (Table S3). The LD block around the top associated SNP ($r^2 > 0.5$) spans 12 kb on chromosome 5 including five genes (AT5G55830, AT5G55835, AT5G55840, AT5G55850 and AT5G55855). An obvious candidate among those is AT5G55835 as it is known to be expressed as a microRNA regulating many developmental processes, including changes in the vegetative phase, and coping with abiotic stress (Lei, Lin, & An, 2016; Wang, Czech, & Weigel, 2009; Yang, Conway, & Poethig, 2011; Yu et al., 2015).

3.3 Revealing the polygenic basis of flowering time in a multilocus analysis of loci associated with time to flowering at 10°C or the expression of flowering time genes

A backward elimination, model selection analysis with an adaptive false discovery rate (FDR) criterion (Brandt et al., 2017; Lillie et al., 2017; Sheng et al., 2015; Zan et al., 2017) was used to identify a polygenic model of loci associated with FT10 in the 1,001 genomes collection. The inputs for the analysis were the candidate loci selected via their direct associations with FT10 in the genomewide association analysis (Table S1) and/or their associations with the expression of known flowering time genes (Table S2). At 15% FDR, 25 of the eQTL and all eight FT10 associated loci were retained in the final model (Figure 3; Table S4). As at a 15% FDR, 28 of the 33 loci are expected to be true associations, this illustrates the ability of the multilocus association method to reveal the highly polygenic architecture of FT10 in this population.

The additive genetic effects of the associated loci range from 1.8 to 5.8 days with a median of 2.9 (Figure 4a; Table S4). The proportion of the total genetic variance for FT10 that could be explained by fitting the Identity by State genomewide kinship matrix is 82.3%. The 33 FT10 associated loci together explain 67.0% of the genetic variance captured by the kinship matrix, with three-fourths of this contributed by the 13 loci with the largest effects (Figure 4b). The additive variance contributed by the individual locus with the largest effect (chromosome 5 at 18,590,327 bp) is 10.7% of the total genetic variance in flowering time in this population.

3.4 Large overlap between loci associated with flowering times at 10°C and 16°C

To explore the overlap between the polygenic architectures of FT10 and FT16, we also performed an identical three-step analysis for FT16. The two phenotypes have a high phenotypic correlation...
(Pearson correlation = .88, \( p \)-value \(< 2.2 \times 10^{-16} \)), but the standard GWAS analysis only detects three loci for FT16 at the genomewide significance level. Our conditional forward selection GWA analysis for FT16 failed to reveal any additional loci directly associated with this trait (Figure S3). All three loci associated with FT16 are mapped within 20 kb of the peak SNPs associated with FT10 (Tables S1 and S5). The multilocus analysis across the 150 eQTL regulating expression of flowering time genes and the three FT16 associated loci identified 29 loci associated with FT16 at 15% FDR. In total, 15 of the 26 eQTL overlapped with the 25 eQTL associated with FT10. (Table S6). The proportion of the total genetic variance for FT16 that could be explained by fitting the Identity by State genomewide kinship matrix is 83.5%. The 29 FT10 associated loci together explain 48.7% of the total FT16 variation, equivalent to 58.4% of the genetic variance captured by the kinship matrix. The three (26) loci detected in the GWAS (eGWAS) analyses together explained 19.1 (29.7%) of the FT16 variance, corresponding to 22.8 (35.6%) of the genetic variance captured by the kinship matrix.

4 | DISCUSSION

Here, we describe a multilocus method to dissect polygenic genetic architectures of adaptive traits in natural populations. It is a framework to simultaneously evaluate how multiple loci, selected based on solid prior knowledge, contribute to an adaptive trait. Here, candidate loci were selected by identifying (i) loci with strong marker—adaptive trait associations, in the studied population, and (ii) functional candidate genes with expression under genetic control by segregating eQTL. Specifics on how candidate is selected in our study and a more detailed discussion are included below.

4.1 | Utilizing prior information to select candidate loci for multilocus analyses

The model trait in the study, flowering time, is a well-studied adaptive trait. In the standard GWAS, a candidate locus was selected that did not quite reach the stringent multiple-testing corrected
threshold. Here, prior biological information, in the form of colocation with a QTL affecting the adaptive trait and a strong candidate gene, motivated its selection as a candidate locus. This locus was ultimately included in the final polygenic model. Second, the genes tested in the eGWAS were preselected based on known effects on flowering time in earlier studies (Brachi et al., 2010). It was hypothesized that if the expression variation in these genes was under genetic control in the population, this might ultimately also influence the plant phenotype. Nearly half of these genes were regulated by eQTL and ~20 of them were included in the final polygenic model. These results illustrate the value of using biological information to inform statistical analyses of complex adaptive traits.

4.2 | Balancing prior information and statistical significance when selecting candidate loci

Too lenient selection of candidates will decrease the power in the polygenic analysis. This is because the adaptive FDR criterion, which is the basis for the polygenic backward elimination analysis, is most powerful for sets of candidates where many of them actually contribute to the trait. A threshold that is too stringent, however, might remove important loci with weaker effects from the polygenic analysis. Balancing quantitative statistical and experimental biological prior information in an appropriate manner is therefore a key in the candidate selection procedure. Here, the analysed population was large, had genotype, expression and phenotype data of high quality and much prior biological information on the genetics of the adaptive trait from experimental work. A stringent prescreening process was therefore implemented to benefit the power in the polygenic analysis. All preselected candidates from the GWAS and eGWAS were strongly supported by statistics as the LD-pruned Bonferroni corrections used provided very similar selection stringency as genomewide permutation-based thresholds (Text S1). Lower preselection thresholds could be considered in less powerful data sets, by for example using FDR-based thresholds. Here in our study, however, the benefits of including more candidates in preselection analyses are likely to be counteracted by a reduced power in the polygenic analysis. This is because the number of candidates was already large in relation to the number of individuals in the data set.

4.3 | Model selection using a backward elimination strategy with an adaptive FDR criteria

Two key features make the backward elimination strategy with an adaptive FDR criteria particularly well suited to analyse this type of data. First, the adaptive model selection criterion makes it possible to infer the polygenic architecture of a trait at a predefined statistical significance level. This distinguishes it from other model selection strategies, including shrinkage algorithms such as LASSO, that are based on fixed evaluation criteria or penalty parameters imposing implicit assumptions about the rate of true positives amongst the evaluated loci. This rate is unknown in real data. Second, in the backward elimination framework, the models are fitted using ordinary least squares regression. Shrinkage-based methods (such as LASSO) aim to, out of several correlated variables, select the one most strongly associated with the response and shrink the effects of others towards zero. In the context of an association study, this results in an implicit assumption that the evaluated loci contributing to the trait are uncorrelated (i.e., not in LD). While this might be a valid assumption in many situations, the correlations between variables (LD between markers) are not always statistical noise to be removed. On the contrary, in some instances, correlations provide valuable information about genetic phenomena including epistasis, multi-allelic, imperfect LD between markers and causal polymorphisms, and patterns of coselection. The value of not excluding correlated SNPs has earlier been illustrated in analyses of experimental populations, where it revealed both multi-allelic and epistatic loci (Sheng et al., 2015; Zan et al., 2017).

A possible caveat of the backward elimination algorithm is that, if the number of candidates selected in the prescreening is large, not all candidates can be evaluated in a single model selection step. Therefore, sets of candidates were first analysed separately with a lower (20% FDR) threshold. The markers remaining in these analyses were then analysed together to obtain the final polygenic model. The sensitivity of the analysis to the way the markers were divided was explored empirically (Text S1; Table S7) to reveal that the majority of the loci in the polygenic model were insensitive to how the markers were divided in the pre-analysis. A few loci, however, differed between the models but at a 15% FDR threshold, this is to be expected.
4.4 Multilocus mapping reveals the polygenetic architecture underlying flowering time traits

In total, eight (three) large effect loci were detected in the GWA analyses for FT10 (FT16). Together they contributed 40.4 (19.1)% of the phenotypic variance for FT10 (FT16) in this population. This finding is consistent with earlier reports in *Arabidopsis thaliana* that major loci often make important contributions to adaptive traits (Brachi et al., 2015; Forsberg et al., 2015; Rus et al., 2006; Shen et al., 2014; Shindo et al., 2005). Here, we can also show that in addition to the major loci, many others also contribute to the variation in this adaptive phenotype. These are difficult to detect in the GWA, either due to their small individual contributions, or due to their low minor allele frequency. Together, however, they contribute a considerable fraction of the flowering time variation in the 1,001 genomes collection and especially so for FT16. This finding suggests that polygenic adaptation, in addition to major alleles, is likely to make an important contribution to adaptation in *A. thaliana*. This is consistent with earlier experimental findings on studies of artificial selection responses, where it has been shown that selection on many loci can lead to extreme adaptations in, for example, seed protein and kernel oil in maize (Laurie et al., 2004; Lucas, Zhao, Schneerman, & Moose, 2013), and body weight in mice (Allan, Eisen, & Pomp, 2005) and chicken (Besnier et al., 2011; Johansson, Pettersson, Siegel, & Carlberg, 2010; Sheng et al., 2015; Wahlberg et al., 2009).

4.5 Selection of candidate adaptive loci using expression data

The optimal design for the eGWAS used to identify candidate adaptive loci would have been to collect expression and phenotype data from plants grown under the same conditions. The expression profiles of relevant genes would then have been more informative regarding their connection to the phenotype. Here, expression and phenotypes were collected under different conditions. This does not
affect the validity of eQTL as candidate adaptive loci as, regardless under which conditions they influence expression, their genotypes are identical and what is evaluated in the polygenic analysis is how this genotype potentially contributes to flowering. Our results show that they were highly valuable for identifying candidate genes, but it is unknown how many contributing genes were missed due to their environmentally dependent differential expression.

4.6 | Multilocus mapping in other types of populations

The A. thaliana 1,001 genomes data set is much more comprehensive and powerful than what can be expected for most other natural populations. For example, expression data or information on hundreds of experimentally supported candidate genes will likely not be available. Many natural populations are also outbreeding which will decrease the power in the association analysis. However, it is when the power in each separate analysis is low that the potential gain by integrating all available information from multiple data sources is the greatest. By adapting the stringency in the prescreening step to the specific data set, while being aware that in smaller and less powerful populations the Beavis effect might have a strong impact on the outcome of the analysis (Slate, 2013), new small-effect loci can also be revealed in outbreeding populations. Results illustrating this include those from our studies of an outbreeding experimental chicken population, where prior information on mapped QTL and selective-sweeps has guided a polygenic analysis to reveal novel loci and complex genetic mechanisms (Sheng et al., 2015; Zan

FIGURE 4 (a) Histogram illustrating the additive effects for the 33 loci associated with FT10 in our multilocus analysis at a 15% FDR. Loci detected as candidates in the standard (conditional) GWAS in light (dark) green and eGWAS in purple, respectively. (b) Cumulative variance explained by the 33 loci (red line). These estimates were calculated by adding one locus at a time in the model in the order determined by the effect sizes in (a). The 25%, 50% and 75% quartiles of the total variance explained by the loci are highlighted using dashed black lines.
et al., 2017). Further, given the decreasing costs for in-depth characterization of biological systems, it is expected that, for example, RNA sequencing will be more widely used in the future to guide preselection of candidate loci in an increasing number of populations.

5 | CONCLUSIONS

In conclusion, our study demonstrates the value of integrating prior information on the genetic basis of a trait with multiple types of experimental data to dissect polygenic adaptive traits. Using prior information on known flowering time genes, and experimental data on gene expression, flowering time variation and genomewide SNP variation, a highly polygenic basis was revealed for this trait in the 1,001 genomes collection. The results suggest that a combination of major effect alleles and polygenic adaptation has been important for flowering time adaptation in Arabidopsis thaliana.

ACKNOWLEDGEMENTS

This work was supported by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas grant ID 2013-450 to OC).

AUTHOR CONTRIBUTIONS

OC and YZ initiated the study, designed the project and the statistical analyses; YZ wrote the analysis scripts and performed the data analyses; OC and YZ summarized the results and wrote the manuscript.

DATA ACCESSIBILITY

All data used in this publication are publically available as part of earlier publications and the 1,001 genomes A. thaliana project.

DISCLOSURE DECLARATION

The authors declare no competing interest.

ORCID

Örjan Carlberg ID http://orcid.org/0000-0002-2722-5264

REFERENCES

Abramovich, F., Benjamini, Y., Donoho, D. L., & Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. The Annals of Statistics, 34(2), 584–653. https://doi.org/10.1214/009053606000000074
Allan, M. F., Eisen, E. J., & Pomp, D. (2005). Genomic mapping of direct and correlated responses to long-term selection for rapid growth rate in mice. Genetics, 170(4), 1863–1877. https://doi.org/10.1534/genetics.105.041319
Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M. M., … Zhou, X. (2016). 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell, 166(2), 481–491. https://doi.org/10.1016/j.cell.2016.05.063
Ando, E., Ohnishi, M., Wang, Y., Matsuhashi, T., Watanabe, A., Hayashi, Y., … Kinoshita, T. (2013). TWIN SISTER OF FT, GIANTEA, and CONSTANS have a positive but indirect effect on blue light-induced stomatal opening in Arabidopsis. Plant Physiology, 162(3), 1529–1538. https://doi.org/10.1104/pp.113.2177984
Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., … Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature, 465(7298), 627–631. https://doi.org/10.1038/nature08800
Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: An R library for genome-wide association analysis. Bioinformatics, 23(10), 1294–1296. https://doi.org/10.1093/bioinformatics/btm108
Besnier, F., Wahliberg, P., Rönnegård, L., Ek, W., Andersson, L., Siegel, P. B., & Carlberg, O. (2011). Fine mapping and replication of QTL in outbred chicken advanced intercross lines. Genetics Selection Evolution, 43(1). 3. https://doi.org/10.1186/1297-9666-43-3
Brachi, B., Faure, N., Horton, M., Fiauauw, E., Vazquez, A., Nordborg, M., Roux, F. (2010). Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genetics, 6(5), 40. https://doi.org/10.1371/journal.pgen.1000940
Brachi, B., Meyer, C. G., Villoutreix, R., Platt, A., Morton, T. C., Roux, F., & Bergelson, J. (2015). Co-selected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. Proceedings of the National Academy of Sciences, 112(13), 4032–4037. https://doi.org/10.1073/pnas.1421416112
Brandt, M., Ahsan, M., Honaker, C. F., & Siegel, P. B. (2017). Imputation-based fine-mapping suggests that most QTL in an outbred chicken advanced intercross body weight line are due to multiple, linked loci. G3 (Bethesda, Md.), 7(1), 119–128. https://doi.org/10.1534/g3.116.034012
Burke, M. K., Liti, G., & Long, A. D. (2014). Standing genetic variation drives repeatable experimental evolution in outcrossing populations of Saccharomyces cerevisiae. Molecular Biology and Evolution, 31(12), 3228–3239. https://doi.org/10.1093/molbev/msu256
El-Assal, E. S., Alonso-Blanco, C., Peeters, A. J., Raz, V., & Koornneef, M. (2001). A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2. Nature Genetics, 29(4), 435–440. https://doi.org/10.1038/ng767
Filiault, D. L., Wessinger, C. A., Dinneny, J. R., Lutes, J., Borevitz, J. O., Weigel, D., … Maloof, J. N. (2008). Amino acid polymorphisms in Arabidopsis phytochrome B cause differential responses to light. Proceedings of the National Academy of Sciences of the United States of America, 105(8), 3157–3162. https://doi.org/10.1073/pnas.0712174105
Forsberg, S. K. G., Andreatta, M. E., Huang, X. Y., Danku, J., Salt, D. E., & Carlberg, O. (2015). The multi-allelic genetic architecture of a variance-heterogeneity locus for molybdenum concentration in leaves acts as a source of unexplained additive genetic variance. PLoS Genetics, 11(11), e1005648. https://doi.org/10.1371/journal.pgen.1005648
Gavrilov, Y., Benjamini, Y., & Sarkar, S. K. (2009). An adaptive step-down procedure with proven FDR control under independence. Annals of Statistics, 37(2), 619–629. https://doi.org/10.1214/07-AOS556
Gu, Z., Gu, L., Elis, R., Schlesner, M., & Brors, B. (2014). Circize implements and enhances circular visualization in R. Bioinformatics, 30(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393
Johansson, A. M., Pettersson, M. E., Siegel, P. B., & Carlberg, O. (2010). Genome-wide effects of long-term divergent selection. PLoS Genetics, 6(11), e1001188. https://doi.org/10.1371/journal.pgen.1001188
Kawakatsu, T., Huang, S. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., … Ecker, J. R. (2016). Epigenomic diversity in a global collection
of Arabidopsis thaliana accessions. Cell, 166(2), 492–506. https://doi.org/10.1016/j.cell.2016.06.044

Laurie, C. C., Chasalow, S. D., LeDeaux, J. R., McCarron, R., Bush, D., Hauge, B., … Dudley, J. W. (2004). The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics, 168(4), 2141–2155. https://doi.org/10.1534/genetics.104.029686

Lei, K. J., Lin, Y. M., & An, G. Y. (2016). miR156 modulates rhizosphere acidification in response to phosphate limitation in Arabidopsis. Journal of Plant Research, 129(2), 275–284. https://doi.org/10.1007/s10265-015-0778-8

Li, P., Filiault, D., Box, M. S., Kerdaффrec, E., van Oosterhout, C., Wilczek, A. M., … Dean, C. (2014). Multiple FLC haplotypes defined by independent cisregulatory variation underpin life history diversity in Arabidopsis thaliana. Genes and Development, 28(15), 1635–1640. https://doi.org/10.1101/gad.245993.114

Lillie, M., Sheng, Z., Honaker, C. F., Dar, B. J., Ashwell, C. M., Siegel, P. B., & Carlborg, O. (2017). Genome-wide standing variation facilitates long-term response to bidirectional selection for antibody response in chickens. BMC Genetics, 18(1), 99. https://doi.org/10.1186/s12864-016-0414-7

Lucas, C. J., Zhao, H., Schneerman, M., & Moose, S. P. (2013). Genomic changes in response to 110 cycles of selection for seed protein and oil concentration in maize. In P. W. Bercot (Ed.), Seed genomics (pp. 217–236). Oxford, UK: Wiley-Blackwell. https://doi.org/10.1002/9781118852524.ch12

Orozco-terWengel, P., Kapun, M., Nolte, V., Van Oosterhout, C., Wilczek, A. M., … Dean, C. (2016). Multiple FLC haplotypes defined by independent cisregulatory variation underpin life history diversity in Arabidopsis thaliana. Genes and Development, 28(15), 1635–1640. https://doi.org/10.1101/gad.245993.114

Lillie, M., Sheng, Z., Honaker, C. F., Dar, B. J., Ashwell, C. M., Siegel, P. B., & Carlborg, O. (2017). Genome-wide standing variation facilitates long-term response to bidirectional selection for antibody response in chickens. BMC Genetics, 18(1), 99. https://doi.org/10.1186/s12864-016-0414-7

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rønnegård, L., Shen, X., & Alam, M. (2010). hcllC: A package for fitting hierarchical generalized linear models. The R Journal, 2, 20–28. Retrieved from http://journal.r-project.org/archive/2010-2/RJournal_2010-2.Roennegaard-et-al.pdf

Rus, A., Baxter, I., Muthukumar, B., Gustin, J., Lahner, B., Yakubova, E., & Salt, D. E. (2006). Natural variants of AtHKT1 enhance Na+ accumulation in two wild populations of Arabidopsis. PLoS Genetics, 2(12), 1964–1973. https://doi.org/10.1371/journal.pgen.0020210

Salomé, P. A., Bomblies, K., Laitinen, R. A. E., Yant, L., Mott, R., & Weigel, D. (2011). Genetic architecture of flowering-time variation in Arabidopsis thaliana. Genetics, 188(2), 421–433. https://doi.org/10.1534/genetics.111.126607

Sharma, N., Xin, R., Kim, D.-H., Sung, S., Lange, T., & Huq, E. (2016). NO FLOWERING IN SHORT DAY (NFL) is a bHLH transcription factor that promotes flowering specifically under short-day conditions in Arabidopsis. Development, 143(4), 682–690. https://doi.org/10.1242/dev.128595

Shen, X., De Jonge, J., Forsberg, S. K. G., Pettersson, M. E., Sheng, Z., Hennig, L., & Carlborg, O. (2014). Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. PLoS Genetics, 10(12), e1004842. https://doi.org/10.1371/journal.pgen.1004842

Sheng, Z., Pettersson, M. E., Honaker, C. F., Siegel, P. B., & Carlborg, O. (2015). Standing genetic variation as a major contributor to adaptation in the Virginia chicken lines selection experiment. Genome Biology, 16(1), 219. https://doi.org/10.1186/s13059-015-0785-z

Shindo, C., Aranzana, M. J., Lister, C., Baxter, C., Nicholls, C., Nordborg, M., & Dean, C. (2005). Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis. Plant Physiology, 138(June), 1163–1173. https://doi.org/10.1104/pp.105.061309.1

Slate, J. (2013). From beavis to beak color: A simulation study to examine how much QTL mapping can reveal about the genetic architecture of quantitative traits. Evolution, 67(5), 1251–1262. https://doi.org/10.1111/evo.12060

Teotonio, H., Chelo, I. M., Bradič, M., Rose, M. R., & Long, A. D. (2009). Experimental evolution reveals natural selection on standing genetic variation. Nature Genetics, 41(2), 251–257. https://doi.org/10.1038/ng.289

Wahlberg, P., Carlborg, O., Foglio, M., Tordoir, X., Syvanen, A.-C., Lathrop, M., … Andersson, L. (2009). Genetic analysis of an F(2) intercross between two chicken lines divergently selected for body weight. BMC Genomics, 10, 248. https://doi.org/10.1186/1471-2164-10-248

Wang, J. W., Czech, B., & Weigel, D. (2009). miR156-regulated SPL transcription factors define an endogenous flowering pathway in Arabidopsis thaliana. Cell, 138(4), 738–749. https://doi.org/10.1016/j.cell.2009.06.014

Yang, L., Conway, S. R., & Poethig, R. S. (2011). Vegetative phase change is mediated by a leaf-derived signal that represses the transcription of miR156. Development (Cambridge, England), 138(2), 245–249. https://doi.org/10.1242/dev.058578

Yu, Z. X., Wang, L. J., Zhao, B., Shan, C. M., Zhang, Y. H., Chen, D. F., & Chen, X. Y. (2015). Progressive regulation of sesquiterpene biosynthesis in arabidopsis and patchouli (pogostemon cablin) by the MIR156-targeted SPL transcription factors. Molecular Plant, 8(1), 98–110. https://doi.org/10.1007/s13004-014-1100-2

Zan Y, Carlborg O. A multilocus association analysis method integrating phenotype and expression data reveals multiple novel associations to flowering time variation in wild-collected Arabidopsis thaliana accessions. G3: Genes, Genomes, Genetics, 6(8), 2319–2328. https://doi.org/10.1534/g3.116.030874

Zan, Y., Shen, X., Forsberg, S. K. G., & Carlborg, O. (2016). Genetic regulation of transcriptional variation in natural Arabidopsis thaliana accessions. G3: Genes, Genomes, Genetics, 6(8), 2319–2328. https://doi.org/10.1534/g3.116.030874

Zan, Y., Shen, X., Forsberg, S. K. G., & Carlborg, O. (2017). Artificial selection response due to polygenic adaptation from a multilocus, multiallelic genetic architecture. Molecular Biology and Evolution, 2, 7–10. https://doi.org/10.1093/molbev/msx194

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Zan Y, Carlborg Ö. A multilocus association analysis method integrating phenotype and expression data reveals multiple novel associations to flowering time variation in wild-collected Arabidopsis thaliana. Mol Ecol Resour. 2018;18:798–808. https://doi.org/10.1111/1755-0998.12757