GWAS with heterogeneous data: estimating the fraction of phenotypic variation mediated by gene expression data

Eriko Sasaki†,∗, Florian Frommlet‡,∗ and Magnus Nordborg†,1
†Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter (VBC), Vienna, Austria, ‡Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University Vienna, Vienna, Austria

ABSTRACT Intermediate phenotypes such as gene expression values can be used to elucidate the mechanisms by which genetic variation causes phenotypic variation, but jointly analyzing such heterogeneous data is far from trivial. Here we extend a so-called mediation model to handle the confounding effects of genetic background, and use it to analyze flowering time variation in Arabidopsis thaliana, focusing in particular on the central role played by the key regulator FLOWERING TIME LOCUS C (FLC). FLC polymorphism and FLC expression are both strongly correlated with flowering time variation, but the effect of the former is only partly mediated through the latter. Furthermore, the latter also reflects genetic background effects. We demonstrate that it is possible to partition these effects, shedding light on the complex regulatory network that underlies flowering time variation.

KEYWORDS Mediation analysis, Flowering time; Natural variation; FLC; Genetic architecture; Correlation network

INTRODUCTION A crucial question in genetics is understanding how genetic variation translates into phenotypic variation. DNA sequence polymorphisms influence final phenotypes through intermediate phenotypes such protein structures, epigenetic states, and gene expression levels — many of which can be assayed using modern technologies. Understanding how these intermediate, molecular phenotypes mediate the effects of genetic variation is of fundamental interest, and has enormous applied implications.

Interest has in particular focused on gene expression levels since they dynamically respond to environmental stimuli, developmental transitions, and other physiological states. Mapping studies have shown that eQTL (expression Quantitative Trait Loci) frequently coincide with causal variants identified using GWAS (Genome-Wide Association Studies, see; Nicolae et al. 2010; GTEx Consortium 2017), supporting the notion that a substantial proportion of genetic variants influence the phenotype by regulating expression levels of the corresponding genes (Nicolae et al. 2010; Cubillos et al. 2012; Barfield et al. 2017; Chun et al. 2017; Mancuso et al. 2017). However, even if this were true, the correlation between measured expression variation and phenotypic variation would not necessarily be perfect due to time-, tissue-, and environment-specific regulation. To quantify this, the genetic effect can be decomposed using a “mediation model” (see Materials and Methods for more about mediation models) into an “indirect effect” that can be explained by gene expression levels and a “direct effect” that cannot be (Figure 1; Baron 1986; Valeri and Vanderweele 2013; Huang et al. 2015). A recent study reported that only about 20% of human disease heritability is thus mediated by gene expression (O’Connor et al. 2017).

Here we use GWAS and mediation analysis to study the transcriptional network regulating flowering time in Arabidopsis thaliana. A novel feature of our analysis is that we explicitly model the confounding effects of the genetic background using a linear mixed-model approach that has become standard in GWAS (Vilhjálmsson and Nordborg 2013). That confounding can bias mediation analyses is well known (Richiardi et al. 2013; Yang et al. 2017), as is the fact that genetic background is a major confounder of GWAS in Arabidopsis thaliana — especially of locally adaptive traits (Aranzana et al. 2005; Atwell et al. 2010). To our knowledge this is the first time that mediation analysis including random effects for the genetic background is performed, where we justify our approach using the statistical theory of counterfactuals.
Flowering time in *A. thaliana* is well-suited for the development of mediation models for at least three reasons. First, systematically collected multi-layer data is available. Since *A. thaliana* is highly selfing and naturally exists as inbred lines, multiple phenotypes, including intermediate ones such as gene expression, have been collected for the same genotype. Second, sampling and growth conditions are controllable and uniform, unlike in human studies, making modeling easier. Finally, flowering time is one of the best understood traits in plants. More than one hundred genes in several major pathways have been described: the photoperiod, ambient temperature, autonomous, integrator, gibberellin and vernalization pathways combine to regulate flowering (Simpson and Dean 2002; Kim et al. 2009; Wellmer and Riechmann 2010; Srikanth and Schmid 2011; Andres and Coupland 2012).

Our primary goal in this study is to use flowering time as an example to explore how best to combine heterogeneous, multi-layer data in order to improve our understanding of the genotype-phenotype map. Our results illustrate well the complexities inherent even in a very simple network structure.

**Materials and Methods**

Data sets

We used published *A. thaliana* data sets containing genotypes (Long et al. 2013), RNA-seq transcriptome data (Dubin et al. 2015), as well as flowering time phenotypes (Sasaki et al. 2015, Table S1). All plants were grown under constant 10°C (132 lines) and 16°C (154 lines) in 16 h day length condition. For RNA-seq analysis, RNA was extracted from whole rosettes collected at 11-12 h after dawn at nine-leaf stage (Dubin et al. 2015). In addition, we used a dataset for flowering time and *FLC* expression including global populations (Shindo et al. 2005; Atwell et al. 2010, 101 lines). Plants were grown under natural light conditions in the greenhouse (22-23°C) from October 2002 to March 2003. *FLC* expression was measured by q-RT-PCR using RNA extracted from young leaves after 4 weeks of growth (nearly nine-leaf stage). With respect to genotypes the genome-wide SNP information in the 1001 genome project was used (1001 Genomes Consortium 2016). The dataset of 10°C were used for model building, it of 16°C and greenhouse were used for prediction of flowering time by the model.

**Correlation analysis**

Both Spearman’s (ρ) and Pearson’s (r) correlation coefficient between flowering time and expression levels were calculated for 20,285 genes for which more than 10% lines showed detectable expression levels. The Benjamini Hochberg procedure (Benjamini 1995) was applied to the p-values corresponding to ρ to obtain genes with the most highly correlated expression levels while controlling FDR at 5%. For the resulting genes a correlation network (Figure 2) was visualized with Cytoscape (Shannon et al. 2003) using a Bonferroni corrected threshold of 1.35e-05 for p-values (741 = C38 tests for pairs of 38 genes + flowering time at a family wise error rate of 0.01).

**GO analysis**

Enrichment of known flowering time genes was estimated using BiNGO as a plugin of Cytoscape (Maere et al. 2005) with Benjamini-Hochberg FDR correction (Benjamini 1995). The GO term “regulation of flower development” (TAIR; Berardini et al. 2015) was used for the analysis of flowering time genes. FDR was calculated based on the GO list as described in Sasaki et al. (2015).

**Genome-Wide Association Studies (GWAS)**

GWAS analysis for flowering time and *FLC* expression was performed using LIMIX (Lippert et al. 2014), and the following linear mixed-model (LMM):

\[
Y = Xβ + u + e
\]

\[
\text{var}(Y) = σ^2_ε K + σ^2_e I
\]

where \(Y\) is the \(n \times 1\) vector of a phenotype (either the standardized flowering time or the standardized *FLC* expression), \(X\) is the \(n \times 1\) vector of the standardized genotype to be tested (SNP), and \(β\) is the parameter of the corresponding fixed effect. \(u \sim N(0, σ^2_u K)\) is the random effect including the kinship matrix \(K\) representing genetic relatedness (IBS) (Yu et al. 2006; Kang et al. 2008) and \(e \sim N(0, σ^2_e I)\) refers to the residual. Bonferroni-correction was used for multiple-testing correction (using a family wise error rate of 5% with 3,401,897 SNPs after excluding all SNPs with MAF ≤ 0.1).

**Variance component analysis**

Cis-genetic effects of loci on an expression level \(Y\) was estimated using local_vs.global.mmm() function in mixmogam (https://github.com/bvilhjal/mixmogam) with the model

\[
Y = U_{local} + U_{global} + ε
\]

\(U_{local} \sim N(0, σ^2_{local} K_{local})\)

\(U_{global} \sim N(0, σ^2_{global} K_{global})\)

\(ε \sim N(0, σ^2_e I)\)

Here \(U_{local}\) and \(U_{global}\) are random effects corresponding to local and global relatedness, respectively, and \(ε\) is noise. The local region is defined as ± 15 Kb coding region of each gene, and the global region is defined as the entire genome. With mixmogam the local and global IBS matrices were calculated as genetic relatedness using all SNPs in local and global regions, respectively. Significance of the variance components was estimated by permutation tests (1000 times) with maintaining the chromosomal order of all observations but shuffling the relative positions of the two variables.

**Figure 1** A genotype-phenotype model that includes gene expression. The phenotype is affected by a genetic polymorphism that is partly mediated by the expression of a nearby gene, resulting in a direct and indirect genetic effect. Both gene expression and phenotype are also affected by confounding genetic background.
Mediation analysis

The idea of using pathway analysis to dissect biological effects into direct and indirect causal relationships was developed already about 100 years ago by Wright (1921). However, his methods were by and large ignored in the biological sciences (see Shipley 2016, for a discussion of why this was the case), and it was rather in the social sciences that similar ideas were developed almost half a century later. For example, Baron and Kenny (1986) discussed questions of mediation analysis in the context of pathway models. The modern approach to causal inference however relies upon the counterfactual framework (see, for example: Pearl 2009; Imbens and Rubin 2015).

To develop these ideas denote by \(X\) some input variable, by \(M\) the mediator and by \(Y\) the outcome. In the counterfactual framework one conceptualizes for each individual different potential outcomes depending on the state of other variables. For example one would denote by \(Y_u(u)\) the state of \(Y\) for individual \(u\) when \(X\) would be equal to \(x\). Although in practice never observable, one contemplates the potential outcomes depending on different values of \(x\) as mathematically existing entities — the counterfactual variables. The average causal effect (or total effect) of a change from \(x\) to \(x'\) is then defined by \(E(Y_u - Y_u)\). Under certain assumptions it is then possible to estimate this causal effect from observational data (Pearl 2009; VanderWeele 2015).

If one is interested in the effect that changing \(x\) has on the outcome \(Y\) directly (i.e., effects that are not mediated by other variables), then the first idea is to look at counterfactual outcomes when keeping the levels \(m\) of the mediator \(M\) fixed. This leads to the so-called Controlled Direct Effect CDE \(E(Y|X=x,M)\). The CDE is not very appealing in our case for two reasons. First, the gene expression levels (our mediator) can certainly not be controlled at all, and, second, controlling does not provide the definition of an indirect effect. In contrast, the concepts of natural direct effect (NDE) and natural indirect effect (NIE) introduced by Pearl (2001) are directly applicable. The natural direct effect defined as \(NDE = E(Y|m = -Y|x)\) compares (at an individual level) the change in the outcome \(Y\) between input \(x\) and input \(x\) but assuming the mediator level would take the counterfactual value \(m\). So NDE measures the change in outcome when the mediator level is kept fixed while changing input \(x\). In contrast the natural indirect effect is defined as \(NIE = E(Y|X=x,M)\), so here the input is kept fixed at \(x\) and one measures the change in outcome that would occur by changing the mediator according to the counterfactual \(m\). Identifiability assumptions for NDE and IDE for observational data are given in Pearl (2001) and VanderWeele and Vansteelandt (2009).

Now the concepts of CDE, NDE and IDE make it possible to obtain clear definitions of direct and indirect effects for rather general classes of regression models. We want to illustrate this first in the context of the simplest possible linear regression model for mediation analysis. To this end consider data from a sample of size \(n\) with input \(X\) in \(\mathbb{R}^n\), a mediator \(M\) in \(\mathbb{R}^n\) and a trait variable \(Y\) in \(\mathbb{R}^n\) and consider the model

\[
M = X\beta_1 + \epsilon_M, \quad \epsilon_M \sim N(0, \sigma_M^2 I_n)
\]

\[
Y = X\beta_2 + M\theta_1 + \epsilon_Y, \quad \epsilon_Y \sim N(0, \sigma_Y^2 I_n)
\]

where \(I_n\) is the \(n \times n\) identity matrix, \(\beta_1\) and \(\beta_2\) are parameters for the fixed effects, and \(\theta_1\) is the parameter for the effect of the mediator \(M\). In our case, \(X\) corresponds to SNP \(G\), \(M\) to the gene expression levels of the corresponding gene and \(Y\) denotes the flowering time. The classical approach of pathway analysis simply consists of plugging in the model for \(M\) in the second equation for \(Y\),

\[
Y = X\beta_2 + (X\beta_1 + \epsilon_M)\theta_1 + \epsilon_Y,
\]

with corresponding expectation

\[
E(Y) = X(\beta_2 + \beta_1\theta_1).
\]

Then \(\beta_2\) would be referred to as the direct effect and \(\beta_1\theta_1\) as the indirect effect which is mediated through \(M\). It turns out that for the simple model (1) both CDE and NDE coincide with the direct effect from pathway analysis \(CDE = NDE = \beta_2\) and also \(NIE = \beta_1\theta_1\). This follows for example from the analysis given in VanderWeele and Vansteelandt (2009), where a slightly more general model including interactions between the input \(X\) and the mediator \(M\) is considered. For the linear model (1) standard software for regression can then be used to obtain estimates of NDE and NIE and VanderWeele and Vansteelandt (2009) also show how to compute the corresponding standard deviations. A simple SAS macro to perform these computations is described in Valeri and Vanderweele (2013).

One shortcoming of this extremely simple mediation approach is that it does not take into account at all the polygenic effect from other SNPs. The customary mixed model approach to GWAS analysis uses a random effect to model that polygenic effect and we would like to incorporate such random effects into the mediation analysis. Thus consider the following generalization of (1)

\[
M = X\beta_1 + Z\gamma_1 + \epsilon_M, \quad \epsilon_M \sim N(0, \sigma_M^2 I_n)
\]

\[
Y = X\beta_2 + M\theta_1 + Z\gamma_2 + \epsilon_Y, \quad \epsilon_Y \sim N(0, \sigma_Y^2 I_n),
\]

where \(\gamma_1\) and \(\gamma_2\) act as random effects for the polygenic effects by all other SNPs, say \(Z = (Z_1, \ldots, Z_p)\) (where the SNP genotypes have been standardized) and \(\gamma_i \sim N_p(0, \sigma^2 I_p)\). Now the error terms \(\epsilon_M\) and \(\epsilon_Y\) might be seen specifically to model environmental effects and measurement errors of \(M\) and \(Y\), respectively. According to the pathway approach plugging in the model for \(M\) in the second equation for \(Y\) now yields

\[
Y = X\beta_2 + (X\beta_1 + Z\gamma_1 + \epsilon_M)\theta_1 + Z\gamma_2 + \epsilon_Y,
\]

and taking expectations again results in (2). Therefore, according to pathway analysis the definitions of direct and indirect effects remain exactly the same as in case of the standard mediation model (1) without random effects. In terms of counterfactuals it is straightforward to see that

\[
NDE = E(Y|m = -Y|x) = E(Y|X=x,M)
\]

\[
= \beta_2 x + \theta_1 m - \beta_2 x^* - \theta_1 m^* = \beta_2 (x - x^*)
\]

and denoting by \(f_{M|X}(m|x)\) the conditional density function of \(M\) given \(X = x\) we obtain

\[
NIE = E(Y|X=x, M) - E(Y|X=x, M)
\]

\[
= \int_m E(Y|x, m) f_{M|X}(m|x) \, dm
\]

\[
- \int_m E(Y|x, m) f_{M|X}(m|x) \, dm
\]

\[
= \int_m (x^* \beta_2 + m \theta_1) f_{M|X}(m|x) \, dm
\]

\[
- \int_m (x^* \beta_2 + m \theta_1) f_{M|X}(m|x) \, dm
\]

\[
= \theta_1 E(M|X = x) - \theta_1 E(M|X = x^*) = \theta_1 \beta_1 (x - x^*)
\]

where we used (4) for the fourth equality and (3) for the last equality. In summary, it follows that also in case of the mixed model, the
We began by asking whether gene expression, as measured in whole plants (above-ground tissue only) at a few weeks of age (the nine-leaf stage) was correlated with flowering time (Table 1). This represents a highly significant enrichment, which persists at higher FDR cut-offs (Figure 2A; see Methods). The amount of flowering time variation explained by SNP FLCL (0.2%) showed significant correlation with flowering time (Table 1). This represents a highly significant enrichment, which persists at higher FDR cut-offs (Figure 2A; see Methods).

**RESULTS**

**Correlation between gene expression and flowering time**

We began by asking whether gene expression, as measured in whole plants (above-ground tissue only) at a few weeks of age (the nine-leaf stage) was correlated with the eventual flowering of the same genotype (at 10°C under long-day conditions; see Methods) across 132 inbred lines (Table S1). According to the Benjamini Hochberg procedure at an FDR level of 5%, 38 out of 20,285 genes (0.2%) showed significant correlation with flowering time (Table 1). Of these, 9 were annotated as being related to flowering, and 5 were also part of a more conservative list of a priori candidates (Srikant and Schmid 2011). This represents a highly significant enrichment, which persists at higher FDR cut-offs (Figure 2A; see Methods).

### Table 1 List of genes whose expression is significantly correlated with flowering time. Spearman correlation coefficient ρ with its corresponding p-value, as well as the squared Pearson correlation coefficient r² which quantifies the explained variation of a simple linear model.

| Gene ID      | ρ   | p-value | r²   | Description |
|--------------|-----|---------|------|-------------|
| AT5G10140    | 0.63| 3.05E-16| 0.53 | FLC*        |
| AT1G65480    | -0.54| 2.64E-11| 0.37 | FT*         |
| AT2G45660    | -0.47| 1.35E-08| 0.22 | SOCI*       |
| AT2G41640    | -0.42| 7.03E-07| 0.20 | Glycosyl-transferase |
| AT3G57920    | -0.39| 3.28E-06| 0.17 | SPL15       |
| AT1G04400    | -0.38| 5.24E-06| 0.15 | CRY2*       |
| AT3G52310    | -0.38| 5.39E-06| 0.14 | RD29A       |
| AT1G69440    | -0.38| 5.53E-06| 0.18 | AGO7        |
| AT3G13100    | -0.38| 7.71E-06| 0.10 | ATP-BINDING CASSETTE C7 |
| AT1G23870    | -0.38| 8.98E-06| 0.16 | TPS9        |
| AT5G4630    | -0.39| 9.65E-06| 0.13 | Terpenoid cyclases |
| AT3G09100    | -0.37| 9.74E-06| 0.11 | protein coding |
| AT5G51720    | 0.37 | 9.90E-06| 0.07 | AT-NEET |
| AT4G33040    | -0.37| 1.02E-05| 0.12 | protein coding |
| AT3G04485    | 0.37 | 1.51E-05| 0.13 | other RNA |
| AT1G77810    | -0.37| 1.62E-05| 0.10 | Galactosyl-transferase |
| AT2G13560    | -0.36| 1.70E-05| 0.11 | NAD-MEI |
| AT3G08990    | 0.36 | 1.73E-05| 0.08 | protein coding |
| AT1G17020    | -0.36| 1.78E-05| 0.07 | SRG1 |
| AT1G06160    | 0.36 | 2.26E-05| 0.07 | ORA59 |
| AT3G19860    | -0.36| 2.35E-05| 0.11 | BHLH121 |
| AT5G48400    | -0.36| 2.60E-05| 0.10 | ATGLR1.2 |
| AT3G19500    | 0.36 | 2.76E-05| 0.14 | protein coding |
| AT3G05660    | -0.36| 2.80E-05| 0.11 | ARLP33 |
| AT4G24540    | -0.35| 3.33E-05| 0.11 | AGL24* |
| AT4G52120    | -0.35| 3.42E-05| 0.15 | CYP71B11 |
| AT3G18840    | -0.35| 4.03E-05| 0.08 | TPR-like-superfamily protein |
| AT2G18196    | 0.35 | 4.67E-05| 0.11 | protein coding |
| AT5G46210    | -0.35| 4.78E-05| 0.10 | ATCULA |
| AT1G53165    | -0.35| 5.01E-05| 0.09 | ATMAP4K ALPHA1 |
| AT3G20250    | -0.34| 5.12E-05| 0.09 | APUM5 |
| AT5G45900    | 0.34 | 5.68E-05| 0.12 | protein coding |
| AT3G55610    | -0.34| 6.47E-05| 0.12 | P5CS2 |
| AT4G18130    | -0.34| 6.63E-05| 0.13 | PHYE |
| AT1G78050    | -0.34| 6.82E-05| 0.12 | PGM |
| AT5G10490    | -0.34| 6.94E-05| 0.12 | MSL2 |
| AT5G58900    | 0.34 | 7.22E-05| 0.10 | protein coding |
| AT2G46500    | -0.34| 7.92E-05| 0.11 | ATP4K |

* Genes in bold have flowering-related mutant phenotypes; *denotes genes that are also part of a more conservative list of a priori candidates (Srikant and Schmid 2011).
Figure 2  Correlation between flowering time and gene expression levels in the Swedish population. (A) The significance of the GO enrichment for flowering time genes (and implied FDR; see Methods) as function of the significance threshold for the flower-expression correlation. (B) Outline of the flowering pathways in A. thaliana (reviewed in, e.g., Kim et al. 2009; Wellmer and Riechmann 2010; Srikanth and Schmid 2011). FLC represses the floral integrator genes FD, FT, and SOC1. FT is induced by the photoperiod pathway through CONSTANS (CO), which is induced by CRYPTOCLOROMES (CRYs); the FT protein is a mobile flowering signal that works with FD to induce SOC1 and floral meristem genes including APETALA1 (AP1), FRUITFUL (FUL), and SEPALATA (SPL3). AGL24 and SOC1 regulate each other in positive feedback loops and induce transcription of LFY. The gibberellin pathway promotes flowering by inducing SOC1 and the floral meristem-identity gene LEAFY (LFY). (C) A correlation network based on gene expression levels. Nodes show flowering time (yellow) and the genes in Table 1 (blue, or orange for the a priori gene set). Edges show significant correlations between nodes (with Bonferroni correction to control FWER at $\alpha = 0.01$) in pink or blue (for positive and negative correlations, respectively).

The top three genes (Table 1) were all a priori flowering time genes: FLOWERING LOCUS C (FLC; Michaels and Amasino 1999; Sheldon et al. 1999) in the vernalization pathway, and FLOWERING LOCUS T (FT; Kardailsky et al. 1999; Kobayashi et al. 1999) and SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1; Samach et al. 2000) in the “integrator” pathway (Table 1; Figure 2B). In agreement with previous work, FLC expression was clearly most strongly correlated: the explained variance, $r^2 = 0.53$, is strikingly similar to what was seen by Lempe et al. (2005) using a sample under different environmental conditions. The expression of the integrator loci FT and SOC1 is less strongly correlated with flowering, which is interesting given that these loci are thought to act downstream of FLC, and are in this sense closer to the phenotype (Figure 2B; Schmid et al. 2003; Wellmer and Riechmann 2010).

The correlation network connecting the genes in Table 1 with flowering was consistent with the known flowering-time pathways (Figure 2C). The integrator pathway connected FT and SOC1 with another strong a priori candidate, AGL24, a known inducer of SOC1 (Yu et al. 2002, 2004; Michaels et al. 2003). The photoperiod pathway was not connected with the integrator pathway but included CRY2 (Toth et al. 2001) as a hub gene in a network containing 19 other genes. The vernalization pathway, via FLC, clearly plays a central role, connecting the integrator pathway and the photoperiod pathways via FT and CRY2.

The genetic basis of expression and flowering variation

The correlation network between gene expression and flowering time (Figure 2C) inherently undirected and tells us little about causation, but insight can be gained by identifying the genetic causes of the expression variation (Schadt et al. 2005). We used variance component analysis (Lippert et al. 2014; Meng et al. 2016) to estimate the effect of local (cis-acting) genetic variation on gene expression, using a 30 kb window surrounding each gene. Based on permutation tests, ($p < 0.05$), roughly one third of the genes in Table 1 were significantly cis-regulated (Figure 3 and Table S2). FLC stood out in that not only it was strongly cis-regulated, but genetic variation at the gene was also significantly associated with expression of almost half of the other genes in Table 1. Thus genetic variation at FLC is affecting the expression of these loci in trans, almost certainly through its effect on FLC expression. In contrast, the expression level of several other genes highly correlated with flowering time, including FT, SOC1, and CRY2 showed no evidence of cis-regulation, but strong evidence for being regulated by genetic variation at FLC. This suggests that FLC is the key determinant of flowering time under our conditions.

To further study the effect of FLC, we carried out genome-wide
We are thus faced with a seemingly paradoxical result. How can (Figure 4)? We note that there is no non-synonymous variation in
expression, when of associations explaining more than 10% of the variation (cf
effects are along the diagonal), and the top panel shows number
of associations explaining more than 10% of the variation (cf. Table S2).

association studies (GWAS) for flowering time and FLC expression (Figures 4 and S1). In agreement with our previous results (Sasaki
et al. 2015), GWAS for flowering time identified a genome-wide sig-
nificant association with a single nucleotide polymorphism (SNP) in the promoter region of FLC (Chr5:3,180,721; p-value = 1.14E-08,
MAF = 0.38) in addition to weaker associations in two other
significance thresholds and orange arrows in panel A show a priori
flowering time genes (from Sasaki et al. 2015); the arrow in B
shows the SNP in the FLC region identified in A.

Figure 4 GWAS for flowering time (A) and FLC expression (B).
Gray horizontal lines indicate Bonferroni-corrected 5% signif-
icance thresholds and orange arrows in panel A show a priori
flowering time genes (from Sasaki et al. 2015); the arrow in B
shows the SNP in the FLC region identified in A.

A mediation model of flowering time variation

We are thus faced with a seemingly paradoxical result. How can
a SNP at FLC (SNP_{FLC}) predict flowering time but not FLC
expression, when FLC expression strongly predicts flowering time
(Figure 4)? We note that there is no non-synonymous variation in
this gene (Li et al. 2014), so the effect of local genetic variation must
be regulatory — and indeed the variance component analysis con-
irms the existence of massive cis-regulatory variation (Figure 3).

This suggests two things: first, SNP_{FLC} must affect flowering
through some aspect of FLC expression that is not captured by
our expression data; second, the expression variation we measure
must partly be caused by genetic variation not tagged by SNP_{FLC}
in cis or in trans. As just noted, the variance-component analysis
clearly supports cis-regulation of FLC (Figure 3).

To estimate the extent to which the effect of SNP_{FLC} on flowering
captured by FLC expression, we performed a statistical me-
diation analysis (Baron 1986; Valeri and Vanderweele 2013; Palmer
et al. 2017). Specifically, we modeled a trait Y under the regu-
lation of a causal factor G that partly acts through an intermediate
mediator M in the context of a confounding background factor C
(Figures 1 and 5A). In the present context, we assumed that the
SNP_{FLC} (G) regulates flowering time (Y) and that this effect is
partly captured by the measured FLC expression (M). Because FLC
expression was measured at the vegetative stage, many weeks
before flowering, we assume that M affects Y and not the other way
around. We used a linear mixed-model approach (LMM; see Meth-
ods) to extend the mediation model to allow genetic background
loci to affect both M and Y (Figure 5A).

Using this model, we estimate that 40.8% of the total effect of
SNP_{FLC} is mediated by (measured) FLC expression. As argued
above, the remaining 59.2% must thus be due to unmeasured
effects on FLC regulation, as it is hard to see how SNP_{FLC} could
affect flowering any other way.

Furthermore, the model explained nearly half of the phenotypic
variation (r^2 = 0.48), and both SNP_{FLC} and FLC expression
contributed significantly (p < 0.01). Interestingly, the latter explained
more of the variation (r^2 = 0.29) than the former (r^2 = 0.19),
presumably reflecting cis-genetic variation at FLC not tagged by
SNP_{FLC} as well as the effect of trans-acting background genetic
loci (Figure 5B).

The importance of the genetic background can readily be seen
by comparing the result above to those obtained using a model that
does not control for confounding genetic background (Figure 5A).
Under this model, SNP_{FLC} explained a much higher proportion
of the phenotypic variance (r^2 = 0.34 vs r^2 = 0.19 above), as
Figure 5  Mediation analysis of flowering time regulation by FLC. (A) Models used. The full model correcting for genetic background is shown on top (LMM, linear mixed-model), and the model without such a correction is shown below (LM, linear model). For details see text. Estimates are shown in blue. (B) Proportion of flowering time variation ($r^2$) explained by SNP $\text{FLC}$ and $\text{FLC}$ expression under the two models (see text). (C) QQ plots of genome-wide association for flowering time and $\text{FLC}$ expression with (blue line) and without (red line) correcting for population structure. (D) The SNP $\text{FLC}$ effect that is mediated by expression of each of the genes in Table 1. Red bars indicate that effect is significant ($p < 0.05$).

Observed in the presence of confounding (Figure 5B). The effect of confounding can also be seen in a genome-wide inflation of $p$-values (Figure 5C). Finally, we investigated the extent to which the effect of SNP $\text{FLC}$ might be mediated by the expression levels of other genes by simply replacing $\text{FLC}$ expression with that of another gene in the model (Figure 5A). Of the 38 genes in Table 1, 16 showed significant mediation of SNP $\text{FLC}$ at $p \leq 0.05$ (Figure 5D). Among those were most of the genes having flowering-related mutant phenotypes. Correlation with $\text{FLC}$ expression was not a strong predictor for mediating the SNP $\text{FLC}$ effect. For example, genes related to the integrator pathway, including $\text{FT}$, $\text{SOC1}$, $\text{AGL24}$, and $\text{SPL15}$ (Figure 2), all mediated SNP $\text{FLC}$ regardless of the correlation with $\text{FLC}$ expression. On the other hand, $\text{CRY2}$ in the photoperiod pathway did not mediate SNP $\text{FLC}$ although its expression is significantly correlated with that of $\text{FLC}$. In contradiction to this result, the variance component analysis shows trans regulation of $\text{FLC}$ on $\text{CRY2}$ expression (Figure 3). These suggest that $\text{CRY2}$ might be regulated by $\text{FLC}$ polymorphisms not tagged by SNP $\text{FLC}$ (or epistasis).

**Prediction of flowering time using the FLC model**

As described in the previous section, we explain almost half of flowering time variation ($r^2 = 0.48$) using SNP $\text{FLC}$ ($r^2 = 0.19$) and SNP $\text{FLC}$-independent $\text{FLC}$ expression ($r^2 = 0.29$). Thus a single SNP and a single expression measurement allows us to predict flowering time rather well (Figure 6A).

To investigate the limits of this “single gene” model we tried to predict flowering using flowering time and expression data generated for the same population, but at a higher growth temperature, namely 16°C (Dubin et al. 2015). Higher temperature generally accelerates flowering, but also prevents vernalization (Duncan et al. 2015), thus significantly delaying flowering for some genotypes (Sasaki et al. 2015).

We predicted flowering time at 16°C using SNP $\text{FLC}$ and $\text{FLC}$ ex-
Prediction at 16°C with effects of genetic background. We applied parameters estimated using the 10°C data to the model (see Methods for details). SNP_{FLC} was significantly associated with flowering time in the 16°C data as well ($p$-value = 3.31E-07; Figure S1A-B), but a correlation between FLC expression and flowering time was only seen for early-flowering lines that have no requirement of vernalization (cf. Figures 6A and B). Nonetheless the performance of the model changed surprisingly little (the explained variation of flowering time decreased from 48% to 43% (Figure 6B).

We also tested the model on a different population for which greenhouse (approximately 23°C) and FLC expression data were available (Shindo et al. 2005). In these data, SNP_{FLC} was not significantly associated with flowering time, but FLC expression still showed a weak correlation with flowering, and the model predicted 29% of flowering time variation (Figure 6C). This decreased prediction accuracy might be due to unknown genes that affect FLC action as suggested by Shindo et al. (2005).

**DISCUSSION**

Our primary goal in this study was to use flowering time in *A. thaliana* as a test case for understanding the connection between genotype and phenotype. Specifically, we built a statistical model to understand how genetic variation and gene expression variation at the central flowering regulator FLC combine to cause phenotypic variation. Both variables are significantly correlated with flowering time, but not with each other. We resolve this apparent paradox by demonstrating that genetic variation at FLC is only partly captured by measuring FLC expression, and that FLC expression also captures the effect of genetic background loci. The complexity apparent in even such a simple network has broader implications for our ability to understanding the genotype-phenotype map. We also demonstrate that it is essential to control for genetic background in these kinds of studies. Using a classical linear mixed model (LMM) approach commonly used in GWAS studies, we developed a simple mediation model that takes genetic background into account, and showed that it dramatically reduced overestimation of the effect of FLC. Although Principal Component Analysis (PCA) can, in principle, also handle complex confounding (Yang et al. 2017),

---

**Figure 6** Prediction of flowering time. (A) Top: A scatter plot between flowering time and the expression level of FLC, both at 10°C, with histograms for each phenotype illustrating the effect of SNP_{FLC}. Reference and non-reference alleles are shown in blue and red, respectively. The dashed lines are regression lines for each allele. Bottom: predicted vs observed flowering time. (B) The 10°C model applied to the same population grown at 16°C. (C) The 10°C model applied to a different population grown in the greenhouse. Dashed lines in model fits show 95% confidence intervals.
the LMM-based approach is simple, and has a clear theoretical justification and interpretation (VihlJalmsson and Nordborg 2013).

According to our estimates, less than half of the effect of the main SNP at FLC is captured by FLC expression (Figure 5). Given that there is no non-synonymous variation at FLC, the missing variation must reflect aspects of FLC expression we did not measure (e.g., tissue- or time-specific expression). Conversely, the fact that FLC expression only partly reflects the main SNP almost certainly reflects both allelic heterogeneity at FLC (Hagenblad et al. 2004; Shindo et al. 2005; Li et al. 2014) and background genetic loci. Integration analyses have reported weak connection in genetic regulation between intermediate and final phenotypes in both Arabidopsis and humans (Zhang et al. 2011; GTExConsortium 2017). Although the observation has been attributed to noise and other confounding effects (Leek and Storey 2007; Fusi et al. 2012), genetic complexity likely also contributes. Mediation analyses like those carried out here should help resolve this.

Our results also shed some light on the network regulating flowering time. Our correlation and variance component analyses (Figures 2–3), support the considerable experimental evidence that pathway (Hepworth et al. 2004; Shindo et al. 2005; Li et al. 2014) and background genetic loci. Integration analyses have reported weak connection in genetic regulation between intermediate and final phenotypes in both Arabidopsis and humans (Zhang et al. 2011; GTExConsortium 2017). Although the observation has been attributed to noise and other confounding effects (Leek and Storey 2007; Fusi et al. 2012), genetic complexity likely also contributes. Mediation analyses like those carried out here should help resolve this.

In conclusion, our results illustrate how genetic variation and intermediate phenotypes such as gene expression may be combined to understand the genotype-phenotype map, while at the same time illustrating the complexity of even an extremely simple network dominated by a single locus.

LITERATURE CITED

1001 Genomes Consortium, 2016 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell 166: 481–491.

Andres, F. and G. Coupland, 2012 The genetic basis of flowering responses to seasonal cues. Nature Rev. Genet. 13: 627–39.

Aranzana, M. J., S. Kim, K. Zhao, E. Bakker, M. Horton, et al., 2005 Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 1: e60.

Atwell, S., Y. S. Huang, B. J. VilhJalmsson, G. Willems, M. Horton, et al., 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627–31.

Barfield, R., H. Feng, A. Gusev, W. L., W. Zheng, et al., 2017 Assessing the genetic effect mediated through gene expression from summary eQTL and GWAS data. bioRxiv.

Baron, D., R.M.; Kenny, 1986 he moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Jornal of Personality and Social Psychology 51: 1173–1182.

Baron, R. M. and D. A. Kenny, 1986 The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51: 1173–82.

Benjamini, Y., Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 57: 289–300.

Berardini, T. Z., L. Reiser, D. Li, Y. Mezheritsky, R. Muller, et al., 2015 The Arabidopsis Information Resource: Making and mining the “gold standard” annotated reference plant genome. Genesis 53: 474–85.

Chun, S., A. Casparino, N. A. Patsopoulos, D. C. Croteau-Chonka, B. A. Raby, et al., 2017 Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nature Genet. 49: 600–605.

Cubillos, F. A., V. Coustham, and O. Loudet, 2012 Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. Curr. Opin. Plant Biol. 15: 192–198.

Dubin, M. J., P. Zhang, D. Z. Meng, M. S. Remigereau, E. J. Osborne, et al., 2015 DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife 4.

Duncan, S., S. Holm, J. Questa, J. Irwin, A. Grant, et al., 2015 Seasonal shift in timing of vernalization as an adaptation to extreme winter. Elife 4.

El-Assal, E. S., C. Alonso-Blanco, A. J. Peeters, C. Wagemaker, J. L. Weller, et al. 2003 The role of Cryptochrome 2 in flowering in Arabidopsis. Plant Physiol. 133: 1504–1516.

Fusi, N., O. Stegle, and N. D. Lawrence, 2012 Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. PLoS Comput. Biol. 8: e1002330.

GTExConsortium, 2017 Genetic effects on gene expression across human tissues. Nature 550: 204–213.

Hagenblad, J., C. Tang, J. Molitor, J. Werner, K. Zhao, et al., 2004 Haplotype structure and phenotypic associations in the chromosomal regions surrounding two Arabidopsis thaliana flowering time loci. Genetics 168: 1627–1638.

Hepworth, S. R., F. Valverde, D. Ravenscroft, A. Mouradov, and G. Coupland, 2002 Antagonistic regulation of flowering-time gene SOC1 by CONSTANS and FLC via separate promoter motifs. EMBO J 21: 4327–37.

Huang, Y. T., L. Liang, M. F. Moffatt, W. O. Cookson, and X. Lin, 2015 igwas: Integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. Genet Epidemiol 39: 347–56.

Imbens, G. and D. Rubin, 2015 Causal inference for statistics, social, and biomedical sciences. an introduction. Cambridge University Press.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, et al., 2008 Efficient control of population structure in model organism association mapping. Genetics 178: 1709–23.

Kardailsky, I., V. K. Shukla, J. H. Ahn, N. Dagenais, S. K. Christiansen, et al., 1999 Activation tagging of the floral inducer FT. Science 286: 1962–5.

Kim, D. H., M. R. Doyle, S. Sung, and R. M. Amasino, 2005 DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife 4.

Lempe, J., S. Balasubramanian, S. Sureshkumar, A. Singh, M. Schmid, et al., 2005 Diversity of flowering responses in wild
Arabidopsis thaliana strains. PLoS Genet. 1: 109–18.
Li, P., D. Filiault, M. S. Box, E. Kerdaffrec, C. van Oosterhout, et al., 2014 Multiple FLC haplotypes defined by independent cis-regulatory variation underpin life history diversity in Arabidopsis thaliana. Genes Dev 28: 1635–40.
Lippert, C., P. F. Casale, B. Rakitsch, and O. Stegle, 2014 Limix: genetic analysis of multiple traits. BioRxiv.
Long, Q., F. A. Rabanal, D. Z. Meng, C. D. Huber, A. Farlow, et al., 2013 Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nature Genet. 45: 884–890.
Maere, S., K. Heymans, and M. Kuiper, 2005 Bingo: a cytoscape plug-in to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21: 3448–9.
Mancuso, N., H. Shi, P. Goddard, G. Kichaev, A. Gusev, et al., 2017 Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am J Hum Genet 100: 473–487.
Meng, D., M. Dubin, P. Zhang, E. J. Osborne, O. Stegle, et al., 2016 Limited contribution of DNA methylation variation to expression regulation in Arabidopsis thaliana. PLoS Genet. 12: e1006141.
Michaels, S. D. and R. M. Amasino, 1999 Flowering locus c encodes a novel mads domain protein that acts as a repressor of flowering. Plant Cell 11: 949–56.
Michaels, S. D., G. Ditta, C. Gustafson-Brown, S. Pelaz, M. Yanofsky, et al., 2003 AGL24 acts as a promoter of flowering in Arabidopsis and is positively regulated by vernalization. Plant Journal 33: 867–874.
Michaels, S. D., E. Himelblau, S. Y. Kim, F. M. Schomburg, and R. M. Amasino, 2005 Integration of flowering signals in winter-annual Arabidopsis. Plant Physiol. 137: 149–156.
Nakagawa, S. and H. Schielzeth, 2013 A general and simple method for obtaining r2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4: 133–142.
Nicolaie, D. L., E. Gamazon, W. Zhang, S. W. Duan, M. E. Dolan, et al., 2010 Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from gwas. Plos Genet. 6.
O’Connor, L. J., A. Gusev, X. Liu, R. R. Loh, H. K. Finucane, et al., 2017 Estimating the proportion of disease heritability mediated by gene expression levels. BioRxiv.
Palmer, W. H., H. Jarrod, and O. D. J., 2017 Rna interference pathways display high rates of adaptive protein evolution across multiple invertebrates. bioRxiv.
Pearl, J., 2001 Direct and indirect effects. Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann pp. 411–420.
Pearl, J., 2009 Causality: second edition. Cambridge University Press.
Richardi, L., R. Bellocco, and D. Zugna, 2013 Mediation analysis in epidemiology: methods, interpretation and bias. Int J Epidemiol 42: 1511–1519.
Samach, A., H. Onouchi, S. E. Gold, G. S. Ditta, Z. Schwarz-Sommer, et al., 2000 Distinct roles of CONSTANS target genes in reproductive development of Arabidopsis. Science 288: 1613–1616.
Sasaki, E., P. Zhang, S. Atwell, D. Meng, and M. Nordborg, 2015 “Missing” G x E variation controls flowering time in Arabidopsis thaliana. PLoS Genet. 11: e1005997.
Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, et al., 2005 An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genet. 37: 710–717.