Separating an allele associated with late flowering and slow maturation of *Arabidopsis thaliana* from population structure

Yanjun Zan[1☯], Xiao Feng[2,3☯], Zheng Ning[3], Weilin Xu[3,4], Qianhui Wan[3,5], Dongyu Zeng[6], Ziyi Zeng[7], Yang Liu[6*], Xia Shen[3,8*]

1 Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

2 State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China

3 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

4 Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA

5 Department of Mathematics, University of Wisconsin-Madison, Madison, Wisconsin, USA

6 School of Life Sciences, Sun Yat-Sen University, Guangzhou, China

7 School of Engineering, Sun Yat-Sen University, Guangzhou, China

8 Center for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, Scotland, UK

☯These authors contributed equally to this work.

* Correspondence to xia.shen@ed.ac.uk (X.S.) or liuy353@mail.sysu.edu.cn (Y.L.)

Abstract

Genome-wide association analysis is a powerful tool to identify genomic loci underlying complex traits. However, the application in natural populations comes with challenges, especially power loss due to population stratification. Here, we introduce a bivariate analysis approach to a GWAS dataset of *Arabidopsis thaliana*. A common allele,

strongly confounded with population structure, is discovered to be associated with late flowering and slow maturation of the plant. The discovered genetic effect on flowering time is further replicated in independent datasets. Using Mendelian randomization analysis based on summary statistics from our GWAS and expression QTL scans, we predicted and replicated a candidate gene *AT1G11560* that potentially causes this association. Further analysis indicates that this locus is co-selected with flowering-time-related genes. We demonstrate the efficiency of multi-phenotype analysis to uncover hidden genetic loci masked by population structure. The discovered pleiotropic genotype-phenotype map provides new insights into understanding the genetic correlation of complex traits.

Author Summary

Joint-analyzing multiple phenotypes is of increasing interest in this post-GWAS era, because of its potential power to reveal more discoveries and its potential insights into pleiotropic genetic architecture. Here, using publicly available *A. thaliana* data, we provide a "textbook" empirical evidence showing how a novel allele, highly confounded with population structure but carries a large genetic effect, can be detected via a double-trait analysis. The allele postpones the flowering time and maturation endpoint of the plant at the same time. The discovered genetic effect can be replicated. We illustrate the bivariate genotype-phenotype map that produces such statistical power. Combining with gene expression genomic scans, we also predict candidate genes using summary-level Mendelian randomization analysis. The results indicate that multi-phenotype analysis is a powerful and reliable strategy to uncover additional value in the established GWAS data.

Introduction

Evolution has resulted in the speciation and adaptation of various organisms. Although natural selection applies to all kinds of species, the resulted natural population structures have dramatic difference. Especially, due to their lack of mobility, plants, comparing to humans and most animals, have established much stronger population

structure adaptive to specific climate conditions (Ch. 11 in [1]). This makes it difficult, for instance in modern genomic studies, to distinguish genotypic effects on plants' phenotypes from geographical stratification [2].

Fast-developing genotyping techniques have made genome-wide association study (GWAS) one of the most useful approaches for discovering genomic loci that regulate phenotypes in various organisms [2–4]. In human GWAS, we learnt that most of the discovered loci associated with complex traits or disease have very small effects [5]. The detected single nucleotide polymorphisms (SNPs) need to have sufficiently high minor allele frequencies (MAFs) for the statistical tests to gain enough power, while high-MAF variants tend to have small effects on the studied phenotypes as these variants were under weak selection pressure. Alleles that have high penetrance on a phenotype are normally under strong selection, resulting in low MAFs of the corresponding SNPs. Thus, a major challenge in human GWAS appears to be the trade-off between statistical power and the effect size of the variant to detect [6–8].

Although similar trade-off also applies to GWAS in plant populations, e.g. in the natural population of *Arabidopsis thaliana*, in terms of discovery power, the major challenge is different. As each individual plant accession is sampled from a specific geographical location in the world, accessions with different genotypes normally have much greater phenotypic differences compared to those in humans. It appears that the genome can explain a large proportion of variation in the plant phenotype, however, the population structure in nature makes such a genomic effect heavily confounded with the environmental effect due to geographical stratification. Therefore, there can be a number of alleles, who have large genetic effects on a certain phenotype, but masked by the population structure.

As a community based effort, over 1000 natural *A. thaliana* accessions have been collected from worldwide geographical locations [9, 10]. Most of those plants have been sequenced for genome, transcriptome, and even methylome, and these datasets have been made publicly available for worldwide researchers. Many accessions in this collection have been phenotyped for developmental, metabolic, inomics, stress resistance traits [2], and more and more phenotypes are gradually releasing. Previous analysis in those datasets have revealed substantial connections between genotypic and phenotypic variations in this species. The application of association mapping have provided insights

to the genetic basis of complex traits [2, 11, 12], adaptation [13] and evolutionary process. Nevertheless, many essential genotype-phenotype links are still difficult to establish based on the current GWAS data, due to the substantial population stratification highly correlated with the sampling origins of the plants. Therefore, novel powerful analyses are required to further uncover hidden genetic regulation.

Based on publicly available *A. thaliana* datasets [2, 9, 10, 14], here, we aim to use a bivariate analysis method that combines the discovery power of two correlated phenotypes [15], in order to map novel pleiotropic loci that simultaneously regulate both traits. We interpret the statistical significance with a double-trait genotype-phenotype map. We try to replicate and *in silico* functionally investigate the candidate genes that may drive such associations.

RESULTS

Bivariate genomic scan identifies a hidden locus simultaneously associated with flowering and maturation periods

We re-analyzed a public dataset of a natural *A. thaliana* collection, where 43 developmental phenotypes and 23 flowering-time-related phenotypes were previously published [2]. The number of accessions with measured phenotypes varies from 93 to 193 with a median of 147 (Table S1). We first excluded all variants with minor allele frequencies (MAF) less than 0.1 and performed single-trait GWA analysis for all these traits based on a linear mixed model, so that the confounded genetic effects due to population stratification is adjusted. We then applied our recently developed multi-trait GWAS method [15] to all pairwise combination of the phenotypes (Materials & Methods). One novel locus, in one of the pairwise test, reached the most stringent 5% Bonferroni-corrected genome-wide significance threshold for the 2,145 pairs of traits and 173,220 variants, i.e. $p < 1.35 \times 10^{-10}$ (Table 1, Figure 1a). This signal also reaches single-trait genome-wide significance in other six pairs of traits highly correlated with the top pair (Figure S1), without Bonferroni-correction for the number of tested trait pairs (Table 1, Figure S3-S8).

For the most significant trait combination, 2W (days to flowering time under long

day with vernalized for 2 weeks) and MT GH (maturation period), the linkage disequilibrium (LD) block of this locus (LD $r > 0.7$) covers about a 260 kb interval on chromosome 1, with a top variant at 3,906,923 bp (double-trait $p = 9.9 \times 10^{-12}$, Figure 1b, Table 1). The detected locus shows joint effects on flowering and maturation, where the effect on flowering time (2W) is notably large (15.3 days), and that on maturation period (MT GH) is 2.5 days (Table 1). These correspond to narrow-sense heritability values of 24% and 10% of the two phenotypes, respectively.

[TABLE 1 ABOUT HERE]

[FIGURE 1 ABOUT HERE]

Double-trait analysis is sufficiently powerful to overcome the confounding population structure

The detected joint-effect locus was missed in the corresponding single-trait GWA analysis of 2W (effect = 15.3, $p = 2.26 \times 10^{-5}$ after correcting for population stratification) and that of MT GH (effect = 2.5, $p = 3.70 \times 10^{-5}$). Notably, this locus was not even detectable at the genome-wide significance level in a much larger population of more than 1,000 *A. thaliana* accessions [9, 10] due to its severe confounding with the natural population structure. The statistical significance can only be identified when considering the joint distribution of the bivariate statistic. According to the genome-wide Z-scores (student t-statistics), these two phenotypes are negatively correlated, as the plant's lifespan is relatively stable (estimated and observed phenotypic correlation = -0.55 and -0.68, respectively). However, the observed effects on the two traits are both substantially positive, showing sufficient deviation from the joint distribution that led to bivariate statistical significance (Figure 2).

[FIGURE 2 ABOUT HERE]

The strong confounding with the population structure can also be visualized by the allele frequency distribution of the top associated SNP across different *A. thaliana* sub-populations based on the genome re-sequencing data from the *A. thaliana* 1001-genomes project [9] (Figure 3). The sub-populations were divided by admixture analysis using ADMIXTURE [9, 16]. The plus allele increasing flowering time was

predominantly found in Sweden and almost fixed in the Northern Sweden population (Figure 3b; allele frequency = 0.97 in Northern Sweden and 0.51 in Southern Sweden). Overall, the phenotype, e.g. flowering time at 10 °C, highly correlates with the frequency of the plus allele (Figure 3). The genotype at this locus follows a latitude decline, where the northern accessions are enriched with the plus allele and the southern accessions are enriched with the minus allele (Figure 3). This spatially imbalanced enrichment shows strong confounding with the population structure, which is why standard single-trait GWAS loses power substantially.

[FIGURE 3 ABOUT HERE]

Replication of the detected genetic effect on flowering time

Although we are lack of an independent dataset of *A. thaliana* maturation duration to replicate the bivariate statistical test, datasets containing additional independent *A. thaliana* flowering time measurements are available. We downloaded a flowering time GWAS dataset measured in 1,135 natural accessions from the 1001-genomes project collection [9] and performed a single-trait association analysis of our discovered top SNP with linear mixed model correction for the population structure. The genetic effect was significantly replicated for flowering time at 10 °C (effect = 1.7 days, $p = 0.037$) and flowering time at 16 °C (effect = 3.6 days, $p = 0.003$). The effects on flowering time in the replication sample appear to be smaller than in the discovery population, possibly due to Winner's curse in the discovery phase.

We also screened literature for conventional quantitative trait loci (QTL) studies in intercrosses using natural *A. thaliana* accessions. Our detected signal is underneath a reported QTL peak for flowering time from an intercross between a Swedish and an Italian accession [17] (Figure S2). This, together with the replication above, justifies the detected association. Although the discovered genetic effect on maturation period is not directly replicated, the effect does exists when the effect on flowering is justified, as the pleiotropic signal must be driven by both phenotypes.

Prediction and replication of candidate genes using summary-level Mendelian randomization

As a community-based effort, all the natural *A. thaliana* accessions from the 1001-genomes project were measured for their transcriptome [9,10]. Such a public gene expression dataset allows us to predict candidate genes underlying the association signal. We extracted the expression levels of 19 genes within a ± 20kb window around the top associated SNP using RNA-seq gene expression measurements from 140 accessions [14]. Among these, the distributions of 14 gene expression phenotypes significantly deviate from normality (Kolmogorov-Smirnov test statistic > 0.8), and these genes were filtered out due to potential unreliable measurements [18]. The remaining 5 genes were passed onto eQTL mapping at the discovered locus (Materials & Methods).

Based on the locus-specific eQTL mapping summary statistics, we applied the recently developed Summary-level Mendelian Randomization (SMR) method [19] to predict potential candidate genes among these five genes. The analysis integrates summary association statistics from GWAS and eQTL scan to predict functional candidate genes using multiple-instrument Mendelian randomization [20], where the complementary HEterogeneity In Dependent Instruments (HEIDI) test checks that the gene expression and flowering time share the same underlying causal variant. One significant candidate *AT1G11560* was detected after Bonferroni correction for five tests (Figure 4, Table 2). This candidate gene prediction result was also replicated using an independent eQTL mapping dataset [10].

[TABLE 2 ABOUT HERE]

[FIGURE 4 ABOUT HERE]

Indication of co-selection with genes in flowering-related pathways

As flowering time is a well-known polygenic trait, we expect multiple loci to be involved and possibly co-selected as a result of parallel evolution. Therefore, we explored the evidence of co-selection by associating the expression values of 288 known genes in flowering-time-related pathways and 1 gene in the maturation related pathway with our top SNP using transcriptome data from 648 *A. thaliana* accessions [9] (Materials &

Methods). In total, six genes (*NF-YA8*, *AT5G53360*, *SPL15*, *AGL42*, *FLC*, *AGL20*) were associated with our top SNP (false discovery rate < 0.05), where, conservatively, four genes (*AT5G53360*, *AGL42*, *FLC*, *AGL20*) were replicated after Bonferroni correction for six tests using data from an independent collection of 140 *A. thaliana* [14] (Table 3). This indicates that co-selected genes in multiple pathways determine the flowering time variation in nature, and our detected locus contributes to a part of that.

[TABLE 3 ABOUT HERE]

DISCUSSION

A serious issue of GWAS in natural population is the confounding between true underlying genetic effects and the population structure, which can lead to spurious associations between genotypes and phenotypes if population stratification is not properly adjusted [6–8]. Incorporation of the random polygenic effect using linear mixed models can effectively control the population structure, but such correction often compromises the true signals. Here, we applied a bivariate analysis to a classic dataset and successfully separated a locus from strong population structure. The detected allele is associated with late flowering and slow maturation of *A. thaliana*, which was corrected away by the linear mixed model in standard single-trait analysis. The replication of the genetic effect on flowering time in an old intercross linkage analysis and another independent dataset improves the confidence of this association. The discovered association is a typical example that jointly modeling phenotypes that share genetic basis can boost discovery power and reveal pleiotropic genotype-phenotype map at the same time.

Together with our recent application of multivariate analysis in human isolated populations [15], the results further indicate that multi-phenotype analysis is an effective approach to detect hidden loci that are lack of discovery power in single-phenotype analysis thus is worth testing in broader applications. Multivariate analysis appears to have the greatest power when the locus-specific genetic correlation does not agree with the natural phenotypic correlation. For instance, like the discovery here, for two traits that are negatively correlated, loci that generate positive genetic correlation between the traits tend to have good chance to be detected in a joint analysis.

In GWAS, phenotypes are usually chosen based on morphological, physiological or economical features. Those features are usually feasible and simple to quantify; however, they might not be directly representative for the underlying genetic or biological factor that we try to detect. Fortunately, a certain degree of biological pathway sharing among complex traits is common, i.e. pleiotropy [21]. Nowadays, it is very common that multiple phenotypes are measured for same individuals in many GWAS datasets, especially in omics study where thousands of phenotypes are measured. Instead of focusing on one phenotype at a time, it is of essential value to jointly model multiple phenotypes, attempting to detect pleiotropic loci that affect multiple traits with biological relevance.

In this study, all the pairs of traits that are associated with the detected locus contain at least one flowering-time trait, and nearly all of them have maturation duration involved. Detection of the novel locus in a bivariate analysis indicates shared genetic basis for the two types of developmental traits, which measure the lengths of two important period during the plant's life time. By integrating the expression level information and GWAS result using SMR/HEIDI test, we were able to predict candidate genes in this region. However, further work beyond the scope of this paper is still required to establish the molecular biological basis underlying the replicate association.

Many genetic variants affecting flowering time have been mapped and many genes promoting flowering times have been well characterized using standard lab accession, Col-0 [22]. Unlike simple traits, where only one or a few alleles are driving the trait's variation, there are many more variants throughout the genome that contribute to the variation of flowering time. The associations between our top SNP and the expression of many flowering-time-related genes serve as evidence of co-selection or parallel adaptation.

In conclusion, our study demonstrates the efficiency of joint modeling multiple-phenotypes which overcomes severe power loss due to population stratification in association studies. We discover and replicate a pleiotropic allele that regulate flowering and maturation periods simultaneously, providing novel insights in understanding the plant's development over life time. By integrating gene expression information with the GWAS results, we predict a functional candidate underneath the associated genomic region. Analysis of gene expression with other flowering-time-related

genes show evidence of co-selection of the predicted candidate with many genes in flowering-time pathways. We encourage wider applications of such a multivariate framework in future analyses of genomic data.

Acknowledgements

X.S. was in receipt of a Swedish Research Council (VR) grant (No. 537-2014-371). International collaboration within this work was partly supported by the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) initiation grant to X.S. (No. IB2015-6000) and Karolinska Institutet travel grant (No. 2017-00534). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

X.S. initiated and coordinated the study. Y.Z. and X.F. performed the main data analysis. Z.N. and X.S. contributed to statistical modeling and interpretation. W.X., Q.W., D.Z. and Z.Z. contributed to data processing. Y.Z., X.F. and X.S. wrote the manuscript. Y.L. and X.S. supervised the study.

Competing interests statement

The authors have declared that no competing interests exist.

Tables

Table 1: Discovery and replication analyses results for the novel pleiotropic locus. Reported association statistics are for the top variant at the locus for each pair of traits. [1]LD: Days to flowering time under Long Day. [2]0W: Days to flowering time under long day without vernalization. [3]2W: Days to flowering time under long day with vernalized for 2 weeks at 5°C, 8hrs daylight. [4]4W: Days to flowering time under long day with vernalized for 4 weeks at 5°C, 8 hrs daylight. [5]0W GH FT: Days to flowering time (greenhouse). [6]FT GH: Days to flowering (greenhouse). [7]MT GH: Maturation period (greenhouse), 20°C, 16 hrs daylight. [8]RP GH: Reproduction period (greenhouse), 20°C, 16 hrs daylight. [9]RA: Reference allele. [10]EA: Effect allele. [11]MAF: Minor allele frequency. [12]Correlation refers to observed phenotypic correlation. [13]FT: Flowering time.

Double-trait Analysis

Trait 1	Trait 2	Chr	Position	RA[9]	EA[10]	MAF[11]	P	Correlation[12]
LD[1]	MT GH[7]	1	3895353	C	T	0.20	6.3×10^{-9}	-0.39
0W[2]	MT GH[7]	1	3896072	G	T	0.20	8.4×10^{-9}	-0.58
2W[3]	MT GH[7]	1	3906923	T	C	0.22	9.9×10^{-12}	-0.68
2W[3]	RP GH[8]	1	3978064	A	C	0.27	1.3×10^{-8}	-0.17
4W[4]	MT GH[7]	1	3906923	T	C	0.22	3.1×10^{-9}	-0.64
0W GH FT[5]	MT GH[7]	1	3906923	T	C	0.22	1.8×10^{-8}	-0.36
FT GH[6]	MT GH[7]	1	3896072	G	T	0.20	1.5×10^{-8}	-0.60

Single-trait Analysis						*Replication*			
Trait 1			Trait 2			FT[13] 10°C		FT[13] 16°C	
Effect	P	h^2	Effect	P	h^2	Effect	P	Effect	P
33.5	5.6×10^{-6}	0.22	2.42	6.0×10^{-4}	0.07	1.95	2.3×10^{-2}	3.96	1.5×10^{-3}
17.3	1.6×10^{-4}	0.17	2.59	2.1×10^{-4}	0.09	1.95	2.3×10^{-2}	3.96	1.5×10^{-3}
15.3	2.3×10^{-5}	0.24	2.47	3.7×10^{-5}	0.10	1.72	3.7×10^{-2}	3.56	3.0×10^{-3}
19.7	6.8×10^{-7}	0.26	2.65	1.6×10^{-3}	0.06	1.57	5.6×10^{-2}	2.57	3.4×10^{-2}
11.6	1.7×10^{-3}	0.16	2.47	3.7×10^{-5}	0.10	1.72	3.7×10^{-2}	3.56	3.0×10^{-3}
25.8	3.8×10^{-5}	0.21	2.47	3.7×10^{-5}	0.10	1.72	3.7×10^{-2}	3.56	3.0×10^{-3}
14.9	1.8×10^{-3}	0.11	2.59	2.1×10^{-4}	0.09	1.95	2.3×10^{-2}	3.96	1.5×10^{-3}

Table 2: Summary of the SMR/HEIDI analysis results. [1]Top SNP: The top SNP in expression QTL analysis. [2]MAF: Minor allele frequency of the top associated SNP. [3]P_{SMR}: p-value from SMR using a collection of 140 *A. thaliana* accessions. [4]P_{HEIDI}: p-value from HEIDI test using a collection of 140 *A. thaliana*. [5]P_{SMR}: p-value from SMR using a second collection of 648 accessions. [6]P_{HEIDI}: p-value from HEIDI test using a second collection of 648 accessions.

Gene	Top SNP[1]	MAF[2]	P^3_{SMR}	P^4_{HEIDI}	P^5_{SMR}	P^6_{HEIDI}
AT1G11560	Chr1:3881093	0.34	6.8×10^{-3}	4.8×10^{-1}	3.2×10^{-2}	2.6×10^{-1}
AT1G11655	Chr1:3874970	0.39	4.1×10^{-2}	9.7×10^{-2}	5.9×10^{-1}	NA
AT1G11690	Chr1:4299126	0.04	3.7×10^{-1}	NA	9.4×10^{-1}	NA
AT1G11590	Chr1:3716355	0.11	5.0×10^{-1}	NA	2.2×10^{-2}	1.5×10^{-1}
AT1G11482	Chr1:3830013	0.63	8.2×10^{-1}	NA	1.5×10^{-1}	NA

Table 3: Genes in flowering-time pathways whose expression are associated with the detected locus. [1]p-value from a expression dataset generated from 648 accessions in the *A. thaliana* 1001-genomes project [10]. [2]FDR value computed from p-value. [3]Replication p-value from another subset of 140 accessions [14].

Locus ID	Gene Name	p-value[1]	q-value[2]	Replication p-value[3]
AT1G17590	*NF-YA8*	1.6×10^{-7}	2.3×10^{-5}	1.7×10^{-2}
AT5G53360	*AT5G53360*	5.8×10^{-7}	5.7×10^{-5}	3.2×10^{-4}
AT3G57920	*SPL15*	7.9×10^{-4}	7.8×10^{-3}	1.7×10^{-2}
AT5G62165	*AGL42*	1.2×10^{-3}	1.1×10^{-2}	6.3×10^{-3}
AT5G10140	*FLC*	1.5×10^{-3}	1.3×10^{-2}	5.7×10^{-4}
AT2G45660	*AGL20*	1.8×10^{-3}	1.4×10^{-2}	1.2×10^{-3}

Figures

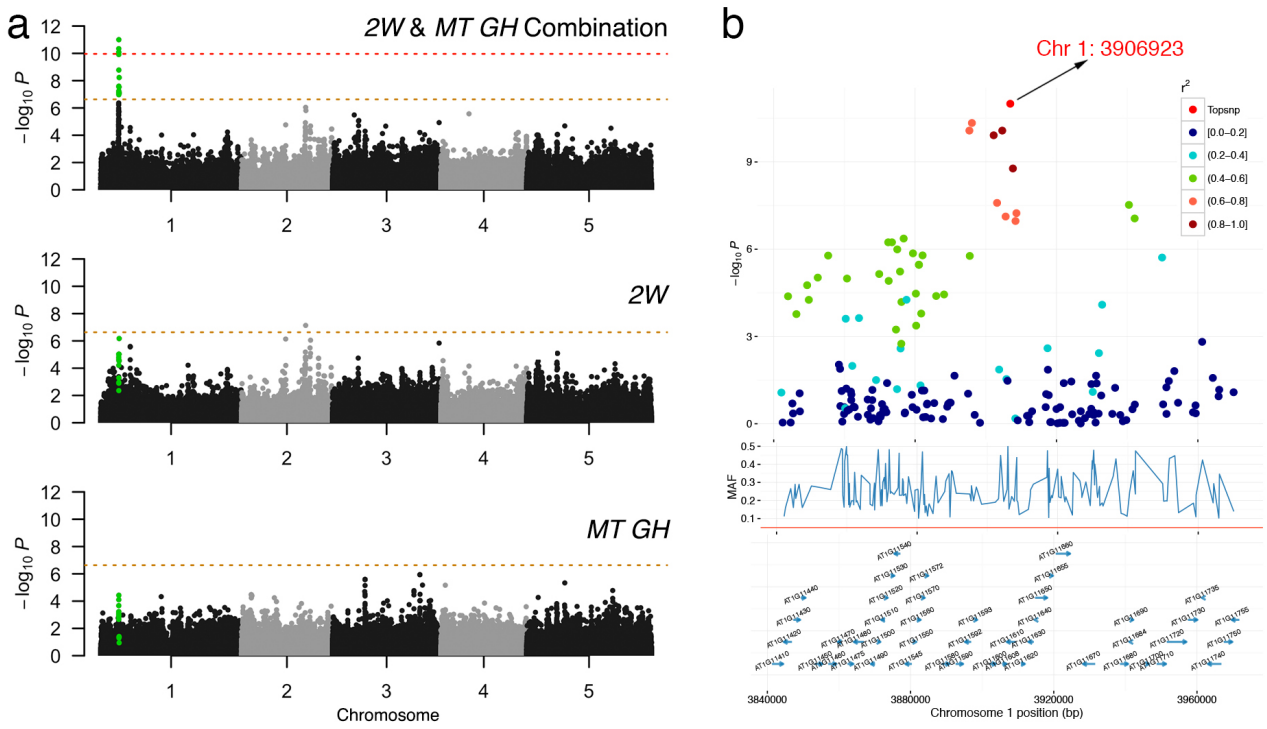

Figure 1: Bivariate genome-wide association analysis of two developmental trait. 2W: Days to flowering time (FT) under long day (LD) with vernalized for 2 wks at 5°C, 8hrs daylight, MT GH: Maturation period. (a) Manhattan plots comparison of bivariate and univariate analysis results, where the novel variants only discoverable when combining two phenotypes are shown in green. The horizontal dashed line represents a 5% Bonferroni-corrected genome-wide significant threshold for the number of variants and also the number of tested trait pairs, respectively. (b) Zooming in the novel locus detected using bivariate analysis. r: linkage disequilibrium measured as correlation coefficient between the top variant and each variant in the region.

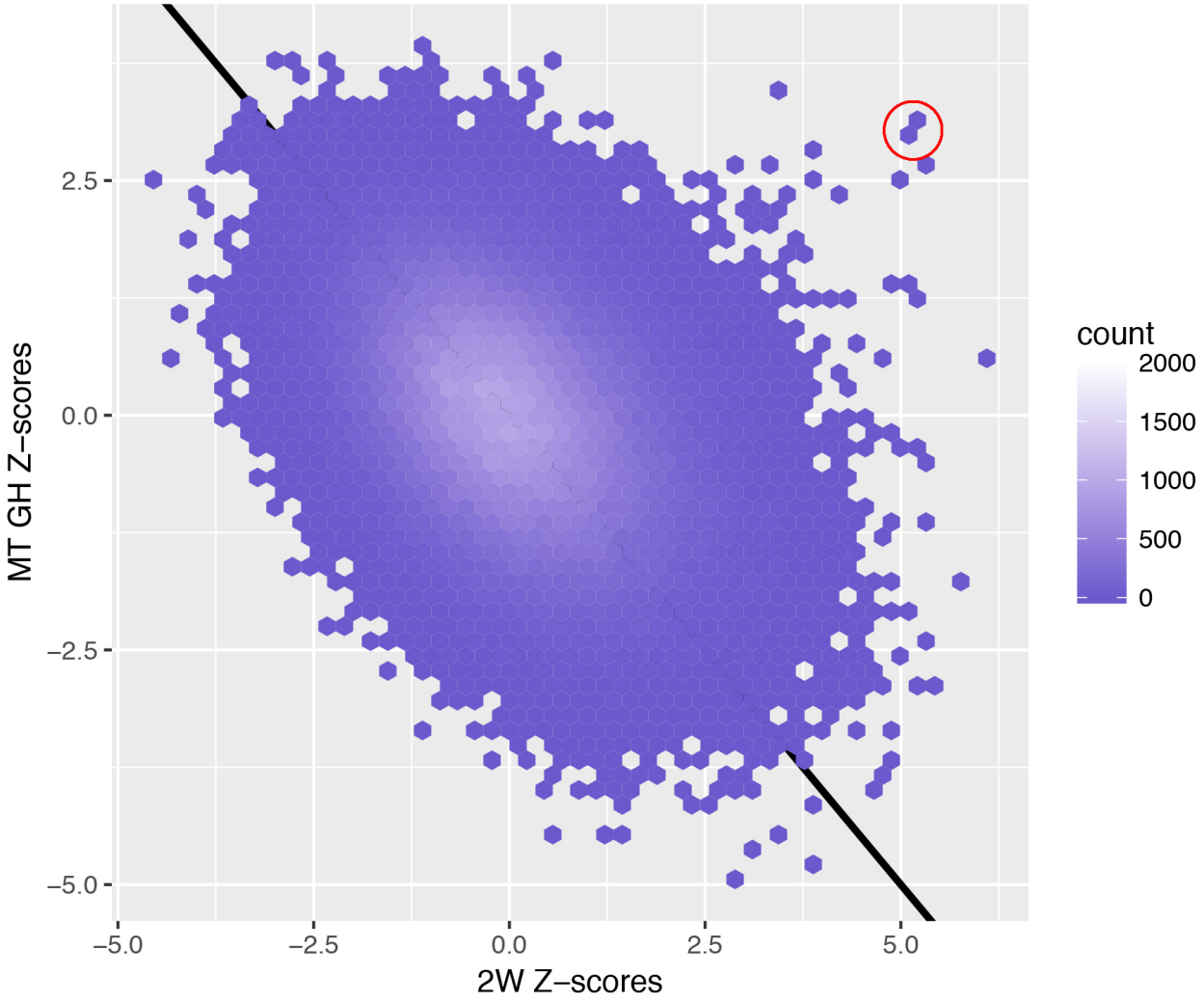

Figure 2: Hexbin scatter plot comparing all Z-scores of the two traits across the genome, showing the bivariate statistical significance of the detected locus. The top variants of the locus is marked on the edge of the empirical bivariate normal distribution with a red circle. The black line with a slope of -1 is provided as a visual guide.

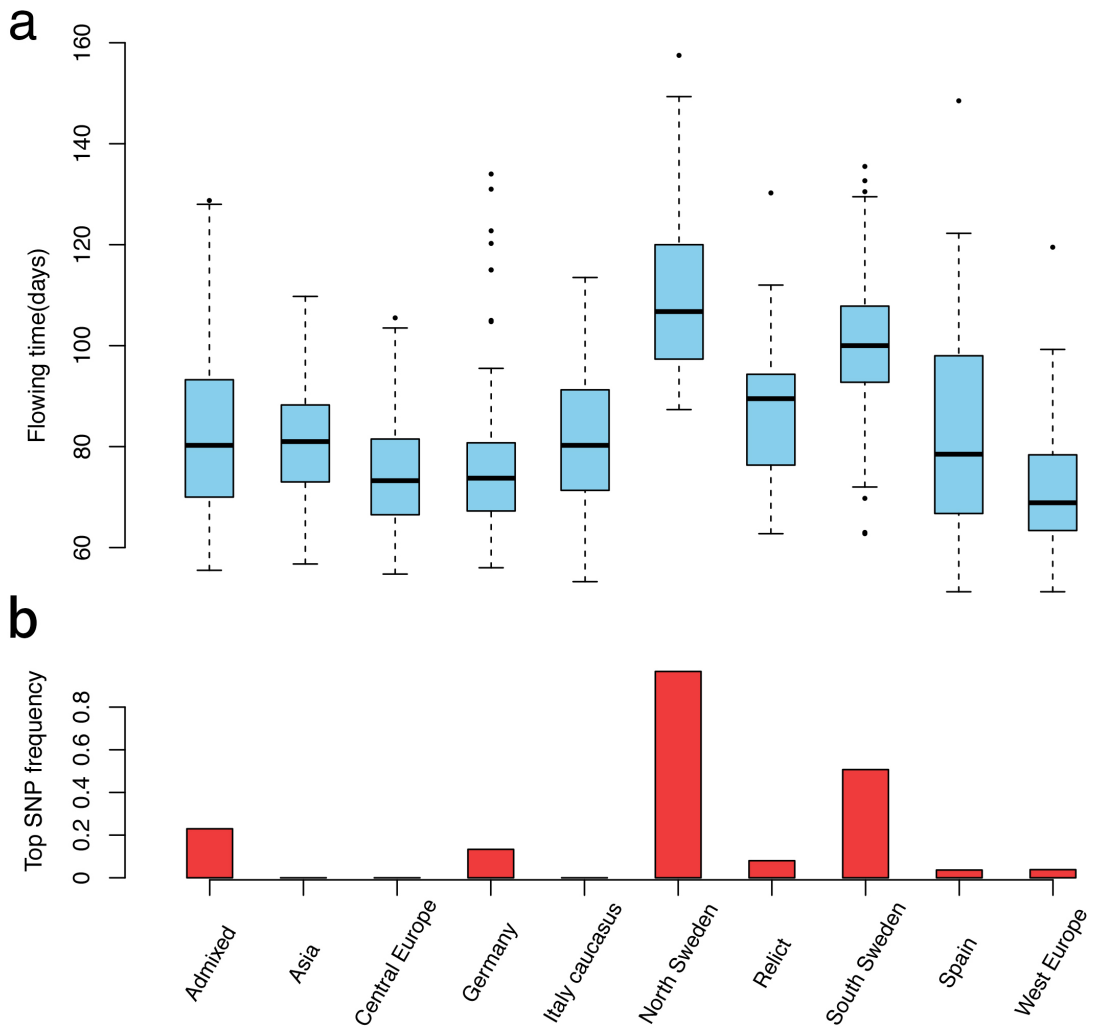

Figure 3: The discovered locus is highly confounded with population structure. a) Flowering time variation (10°C) among different sub-populations of *Arabidopsis thaliana*. These populations are divided by admixture analysis [9]; b) Frequency of the top associated SNP at chromosome 1, 3,906,923 bp in different sub-populations. The association between the structure of the phenotype and that of the allele frequency shows the population confounding at this locus.

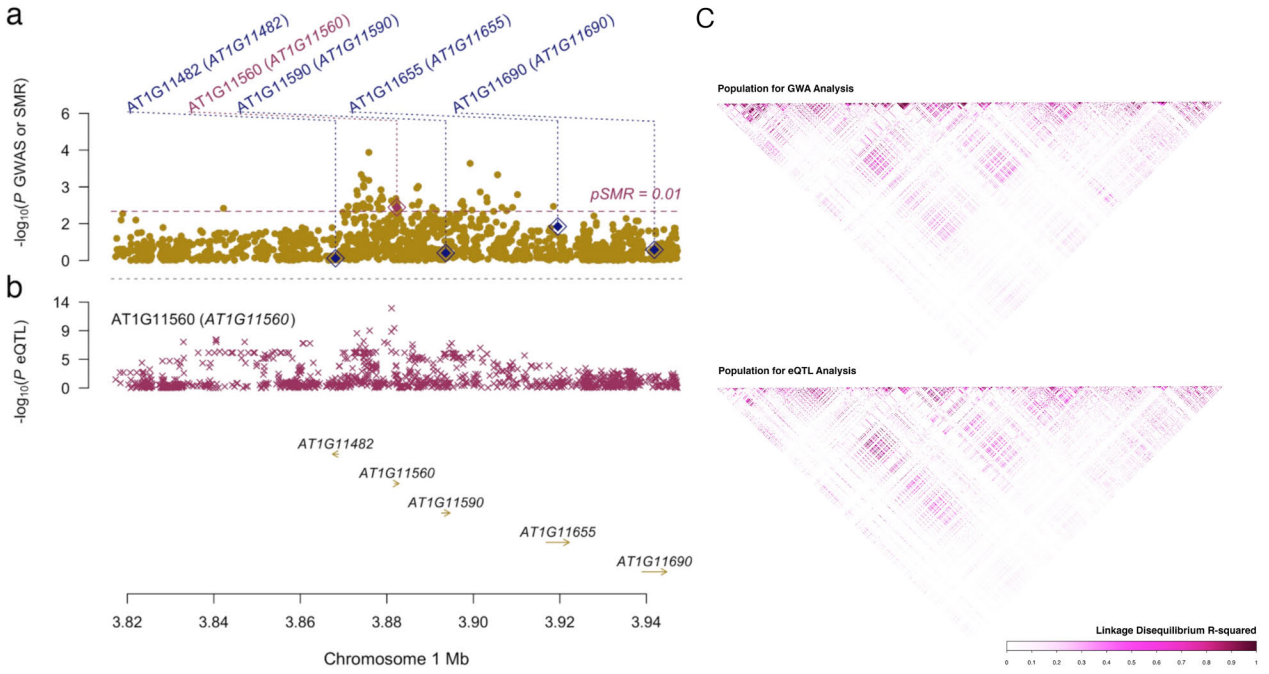

Figure 4: Prioritized candidate genes at the detected locus for flowering time using SMR analysis. a) Manhattan plot of association between flowering time at 10°C and SNPs around 40kb of top associated SNP in bivariate analysis. The diamonds highlight top eQTL for individual genes; b) Manhattan plot of association between expression of *AT1G11560* and SNPs around 40kb of top associated SNP in bivariate analysis. Genes tested in SMR analysis are highlighted using arrows; c) Similar linkage-disequilibrium structure at the locus for the corresponding populations of GWA and eQTL analyses.

MATERIALS & METHODS

Genome-wide 250k SNP array genotype data and phenotype data for 199 natural Arabidopsis thaliana accessions

We downloaded a public dataset on collection of 199 natural *Arabidopsis thaliana* inbred lines contains 107 phenotypes and corresponding genotypes [2].Those files are publicly available at `https://github.com/Gregor-Mendel-Institute/atpolydb/blob/master/miscellaneous_data/phenotype_published_raw.tsv`, and `https://github.com/Gregor-Mendel-Institute/atpolydb/blob/master/250k_snp_data/call_method_75.tar.gz`. 214,051 SNPs were available. After filtering out the variants with minor allele frequency less than 0.10, 173,220 SNPs remained.

Whole genome re-sequencing and RNA-seq data for a population of 1,135 natural A. thaliana accessions

1,135 natural *Arabidopsis thaliana* accessions have been collected and sequenced for the whole genome and transcriptome [9, 10]. We downloaded this sequencing dataset and removed the accessions with no measured phenotype and SNPs with minor allele frequency below 0.05 and a call-rate below 0.95. The final dataset includes 1001 individuals with 2,222,379 SNPs and measured flowing time at 10°C. To scan for candidate genes, we also downloaded the transcriptome dataset of a subset of this collection ($n = 728$) [10]. The final eQTL scan dataset contains RNA-seq derived RPKM-values for 24,150 genes in 648 accessions whose phenotypic and genotypic data are both available.

Whole genome re-sequencing derived SNP genotype and RNA-sequencing derived transcriptome data for a population of 144 natural A. thaliana accessions

In an earlier study, Schmitz et al. [14] RNA-sequenced a collection of 144 natural *A. thaliana* accessions. We downloaded this data together with their corresponding whole-genome SNP genotypes available as a part of the 1001 Genomes project [9, 10] to replicate our SMR findings. Following the quality control procedure in [18], we removed

two accessions from the data (Alst_1 and Ws_2) due to missing genotype data and two accessions (Ann_1 and Got_7) due to their low transcript call rate (16,861 and 18,693 genes with transcripts as compared to the range of 22,574 to 26,967 for the other the accessions). The final dataset used for eQTL mapping included 1,347,036 SNPs with MAF above 0.05 and call-rate above 0.95 for 140 accessions, and corresponding RNA-seq derived FPKM-values for 33,554 genes.

Single-trait analysis for flowering time trait

For all available traits in this dataset, we first performed a mixed model based single trait genome wide association analysis to generate single trait summaries statistics. Those summaries statistics were used as input for double trait analysis described in the following section. To replicate our signal, we also performed a single trait genome wide association analysis using a collection generated in 1001-genomes project [9]. To correct for the population structure in these *A. thaliana* accessions, single-trait genome wide scan was performed based on linear mixed models, using the polygenic and mmscore procedure in GenABEL [23].

Double-trait genome-wide association analysis

We performed double-trait genome scans using our recently developed multivariate analysis method implemented in the MultiABEL package [15]. The method takes the whole-genome summary statistics to infer phenotypic correlation coefficients and conducts MANOVA analysis. The phenotypic correlation coefficient of two traits can be unbiasedly estimated by the correlation of genome-wide Z-scores, which is proportional to the phenotypic correlation on the liability scale. In this way, the bivariate MANOVA analysis is carried out on the liability scale, with bivariate p-values reported.

eQTL and SMR analysis

We screened for candidate genes by analyzing the expression data in a subset of the 1001-genomes collection containing 140 accessions. Expression values for 19 genes around 20kb up/downstream of the top associated SNP were extracted from [14]. 14 genes did not pass Kolmogorov-Smirnov test (ks test statistics ¡ 0.8) were filtered out

due to potential unreliable measurement mentioned in [18]. The remaining five genes were subsequently passed onto eQTL mapping using qtscore procedure in GenABEL [23]. Output were reformatted according to the description in [19]. Together with the flowering time single-trait scan results [9], these were further passed onto SMR analysis scanning for association between individual gene expression and flowering time. The SMR analysis were repeated for 5 top candidates, in an independent gene expression dataset containing 648 accessions [10] following the same procedure.

References

1. Crawley MJ. Plant Ecology. John Wiley & Sons; 2009.

2. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature. 2010;465(7298):627–631.

3. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature reviews Genetics. 2005;6(2):95–108.

4. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nature genetics. 2010;42(11):961–967.

5. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature genetics. 2010;42(7):565–569.

6. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant methods. 2013;9:29.

7. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature genetics. 2014;46(2):100–106.

8. Wellenreuther M, Hansson B. Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. Trends in genetics : TIG. 2016;32(3):155–164.

9. Consortium G. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016;166(2):481–491.

10. Kawakatsu T, Huang SSC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 2016;166(2):492–505.

11. Shen X, Pettersson M, Rönnegård L, Carlborg O. Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana. PLoS genetics. 2012;8(8):e1002839.

12. Wang B, Li Z, Xu W, Feng X, Wan Q, Zan Y, et al. Bivariate genomic analysis identifies a hidden locus associated with bacteria hypersensitive response in Arabidopsis thaliana. Scientific reports. 2017;7:45281.

13. Shen X, De Jonge J, Forsberg SKG, Pettersson ME, Sheng Z, Hennig L, et al. Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. PLoS genetics. 2014;10(12):e1004842.

14. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of population epigenomic diversity. Nature. 2013;495(7440):193–198.

15. Shen X, Klarić L, Sharapov S, Mangino M, Ning Z, Wu D, et al. Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. Nature communications. 2017;8(1):447.

16. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009;19(9):1655–1664.

17. Dittmar EL, Oakley CG, Ågren J, Schemske DW. Flowering time QTL in natural populations of Arabidopsis thaliana and implications for their adaptive value. Molecular ecology. 2014;23(17):4291–4303.

18. Zan Y, Shen X, Forsberg SKG, Carlborg O. Genetic Regulation of Transcriptional Variation in Natural Arabidopsis thaliana Accessions. G3 (Bethesda, Md). 2016;6(8):2319–2328.

19. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics. 2016;48(5):481–487.

20. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG, EPIC-InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. European journal of epidemiology. 2015;30:543.

21. Visscher PM, Yang J. A plethora of pleiotropy across complex traits. Nature genetics. 2016;48(7):707–708.

22. Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, et al. Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS genetics. 2010;6(5):e1000940.

23. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics (Oxford, England). 2007;23(10):1294–1296.

Table S1: Phenotypes included in the bivariate analyses. Details about phenotyping can be referred to [2].

Phenotype	Description	Number of Accessions
LD	Days to flowering time (FT) under Long Day (LD)	167
LDV	Days to flowering time (FT) under Long Day (LD) (5 wks vernalization)	168
SD	Days to flowering time (FT) under Short Day (SD)	162
SDV	Days to flowering time (FT) under Short Day (SD) (5 wks vernalization)	159
0W	Days to FT under LD without vernalization	137
2W	Days to FT under LD with 2wks vernalization	152
4W	Days to FT under LD with 4wks vernalization	119
8W	Days to FT under LD with 8wks vernalization	155
FLC	FLC gene expression	167
FRI	FRI gene expression	164
FT10	Flowering time (FT), 10°C	194
FT16	Flowering time (FT), 16°C	193
FT22	Flowering time (FT), 22°C	193
LN10	leaf number at flowering time (LN), 10°C	177
LN16	leaf number at flowering time (LN), 16°C	176
LN22	leaf number at flowering time (LN), 22°C	176
8W GH FT	Days to FT with 8 wks vernalization	162
8W GH LN	LN at FT with 8 wks vernalization	163
0W GH FT	Days to FT without vernalization	153
0W GH LN	LN at FT without vernalization	135
FT Field	Days to flowering of plants grown in the field	180
FT Diameter Field	Plant diameter at flowering (field)	180
FT GH	Days to flowering (greenhouse)	166
LES	Presence or absence of lesioning	95
YEL	Presence or absence of yellowing	95
LY	Presence or absence of either lesioning or yellowing	95
FW	Fresh weight of plants	95
DW	Dry weight of plants	95
Chlorosis 10	Visual chlorosis presence, 10°C	177
Chlorosis 16	Visual chlorosis presence, 16°C	176
Chlorosis 22	Visual chlorosis presence, 22°C	176

Anthocyanin 10	Visual anthocyanin presence, 10°C	177
Anthocyanin 16	Visual anthocyanin presence, 16°C	176
Anthocyanin 22	Visual anthocyanin presence, 22°C	177
Seed Dormancy	Seed dormancy level	83
Germ 10	Days to germination, 10°C	177
Germ 16	Days to germination, 16°C	176
Germ 22	Days to germination, 22°C	177
Seedling Growth	Seedling growth rate	100
Vern Growth	Vegetative growth rate during vernalization	110
After Vern Growth	Vegetative growth rate after vernalization	110
Secondary Dormancy	Decrease in germination rate after prolonged exposure to cold temperature	93
Germ in dark	Germination in the dark	93
DSDS50	Duration of seed dry storage required for 50% of the seeds to germinate	109
Seed bank 133-91	Non-monotonous dynamic of dormancy release	110
Storage 7 days	Primary dormancy, 7 days dry storage	110
Storage 28 days	Primary dormancy, 28 days dry storage	110
Storage 56 days	Primary dormancy, 56 days dry storage	110
Hypocotyl length	Hypocotyl length	89
Width 10	Plant diameter, 10°C	176
Width 16	Plant diameter, 16°C	175
Width 22	Plant diameter, 22°C	175
Leaf serr 10	Level of leaf serration, 10°C	174
Leaf serr 16	Level of leaf serration, 16°C	176
Leaf serr 22	Level of leaf serration, 22°C	176
Leaf roll 10	Leaf roll presence, 10°C	177
Leaf roll 16	Leaf roll presence, 16°C	176
Leaf roll 22	Leaf roll presence, 22°C	176
Rosette Erect 22	Presence of rosette errectness, 22°C	176
Silique 16	Silique length, 16°C	95
Silique 22	Silique length, 22°C	95
FT Duration GH	Flowering period duration	147
LC Duration GH	Life cycle period	147
LFS GH	Last flower senescence	148
MT GH	Maturation period	147
RP GH	Reproduction period	147

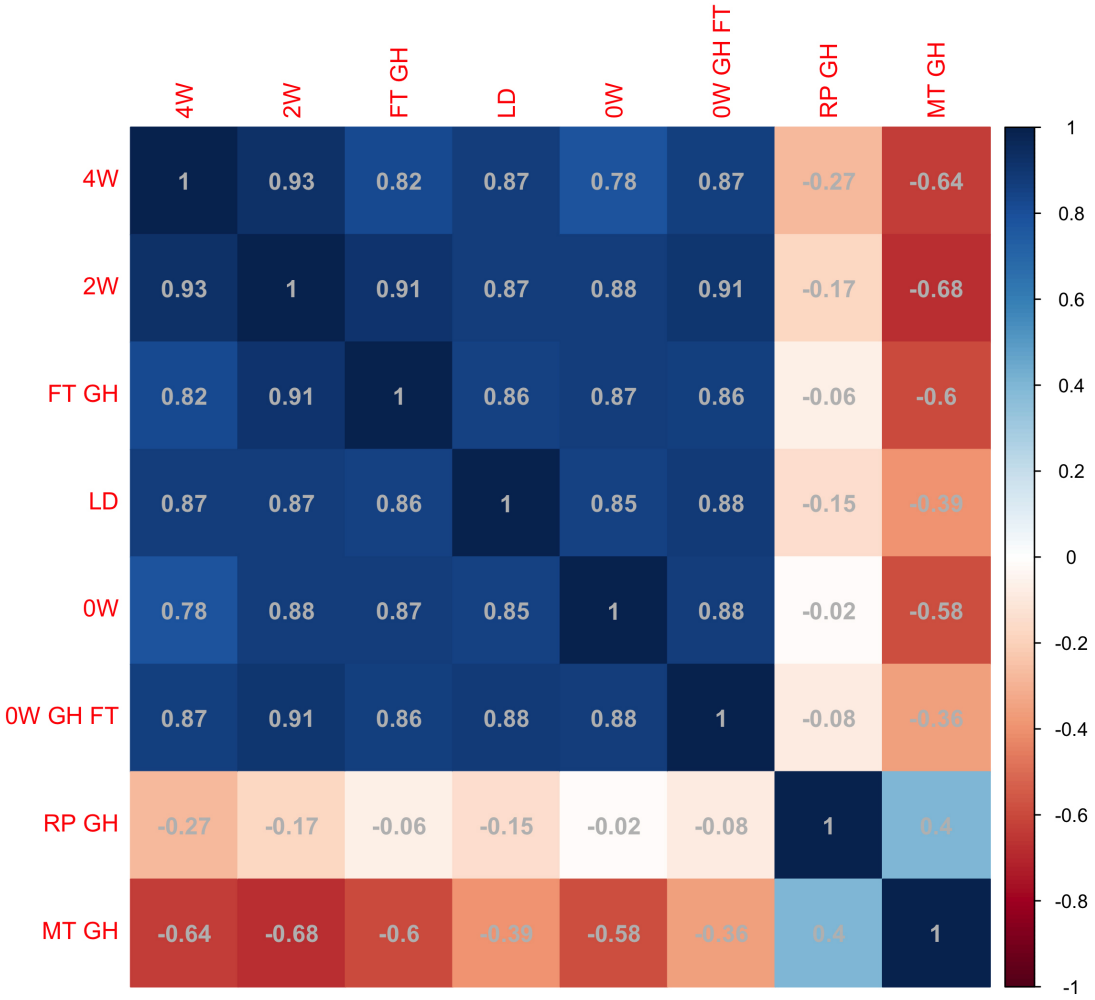

Figure S1: Phenotypic correlations among flowering time related traits, maturation period and reproduction period phenotypes. The flowering time related traits are: 4W: Days to flowering time (FT) under long day (LD) with vernalized for 4 wks at 5°C, 8hrs daylight; 2W: Days to flowering time (FT) under long day (LD) with vernalized for 2 wks at 5°C, 8hrs daylight; FT GH: Days to flowering (greenhouse); LD: Days to flowering time (FT) under Long Day (LD); 0W: Days to flowering time (FT) under Long Day (LD) without vernalization; 0W GH FT: Days to flowering time (FT).

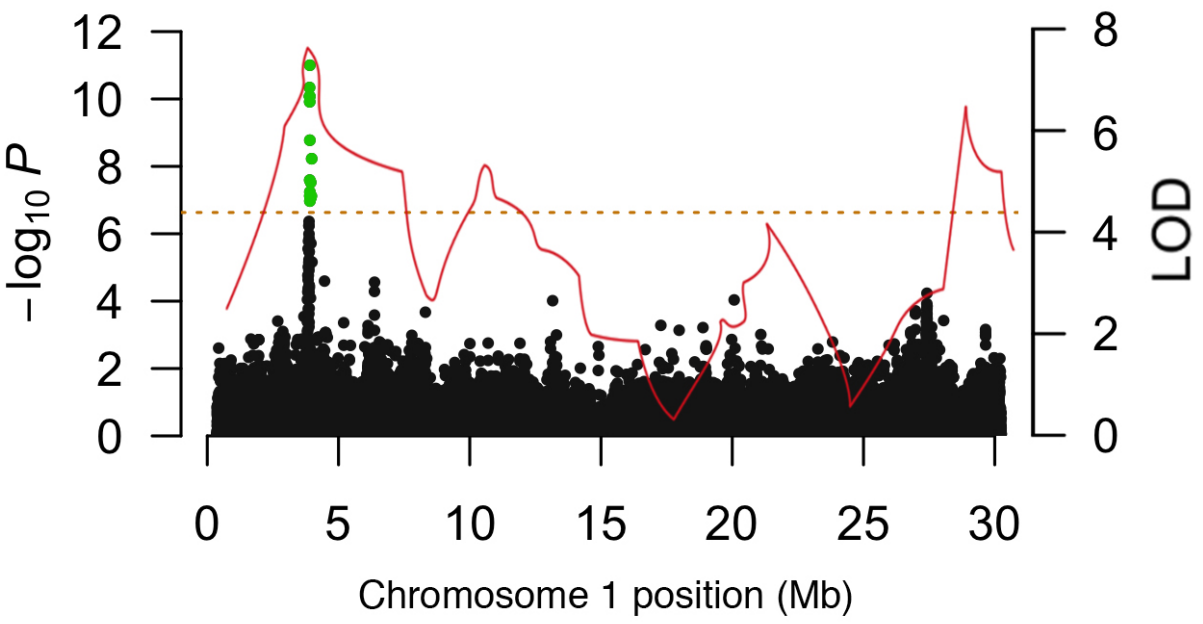

Figure S2: Overlapping between QTL mapping and double-trait GWAS result. The curve shows stepwise LOD profiles in chromosome 1 that are generated from a QTL mapping study using a cross between Italy and Sweden population analyzed by [17] (reproduced by depicting the curvature of Figure 3a therein). The Manhattan plot shows chromosome 1 signal in our bivariate analysis.

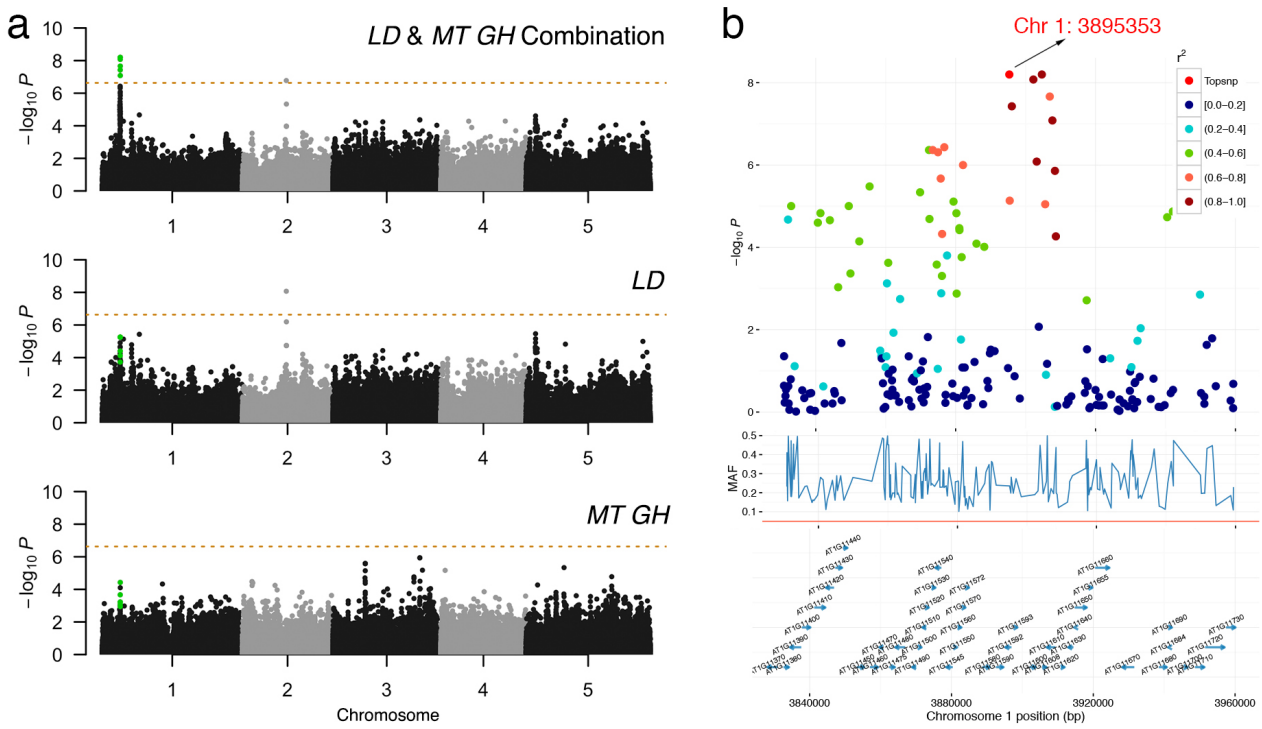

Figure S3: Bivariate genome-wide association analysis of two developmental trait, LD: Days to flowering time (FT) under Long Day (LD), MT GH: Maturation period. (a) Manhattan plots comparison of bivariate and univariate analysis results, where the novel variants only discoverable when combining two phenotypes are shown in green. The horizontal dashed line represents a 5% Bonferroni-corrected genome-wide significant threshold. (b) Zooming in the novel locus detected using bivariate analysis. r: linkage disequilibrium measured as correlation coefficient between the top variant and each variant in the region.

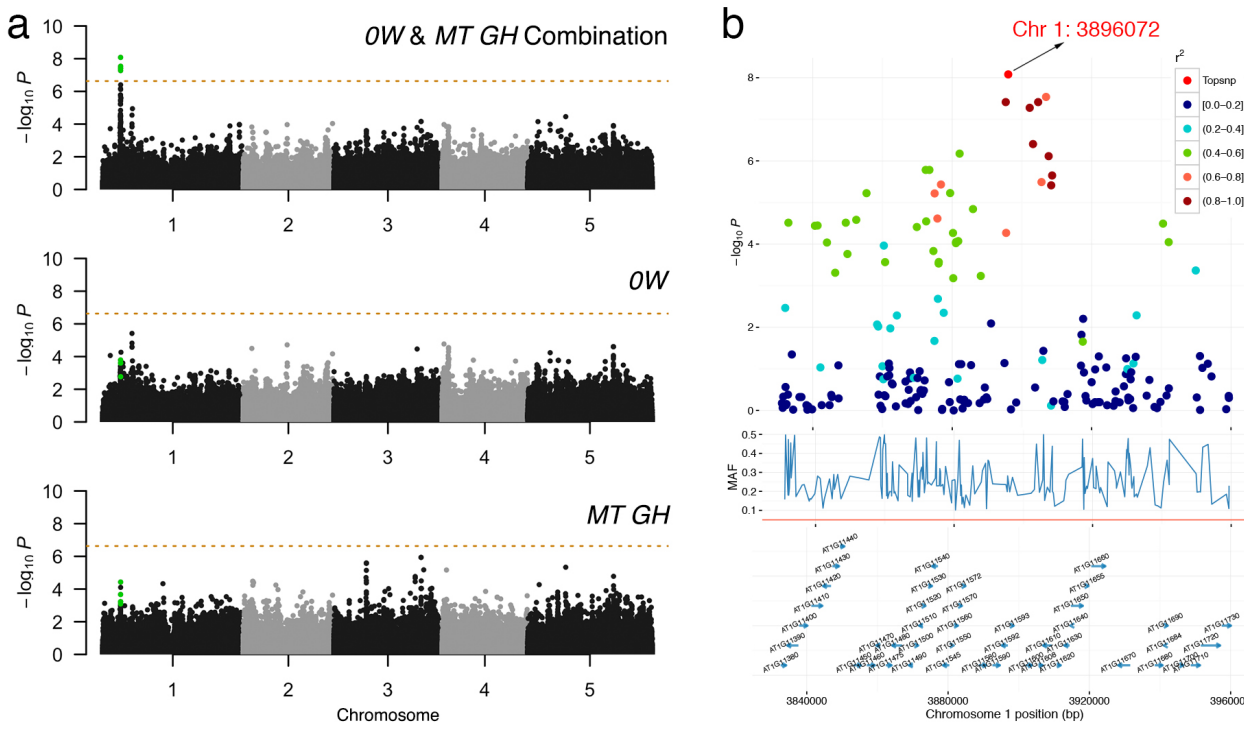

Figure S4: Bivariate genome-wide association analysis of two developmental trait, 0W: Days to flowering time (FT) under Long Day (LD) without vernalization, MT GH: Maturation period. (a) Manhattan plots comparison of bivariate and univariate analysis results, where the novel variants only discoverable when combining two phenotypes are shown in green. The horizontal dashed line represents a 5% Bonferroni-corrected genome-wide significant threshold. (b) Zooming in the novel locus detected using bivariate analysis. r: linkage disequilibrium measured as correlation coefficient between the top variant and each variant in the region.

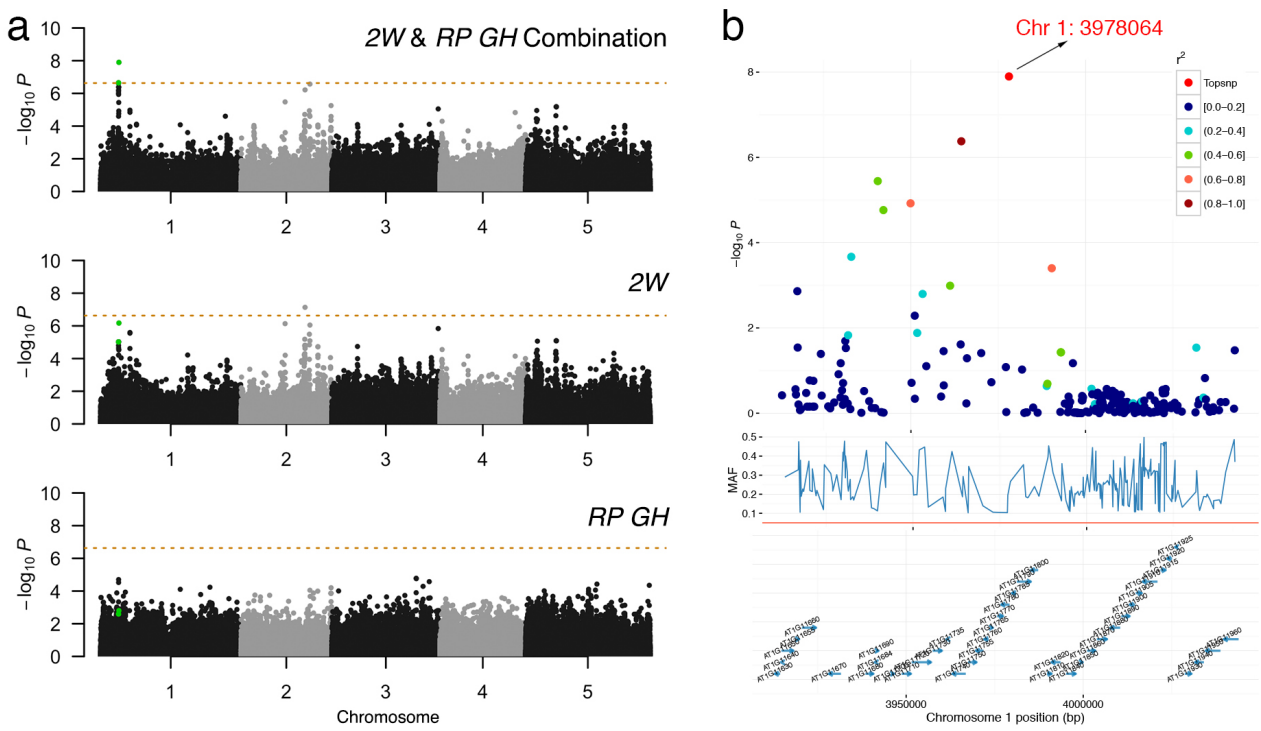

Figure S5: Bivariate genome-wide association analysis of two developmental trait, 2W: Days to flowering time (FT) under long day (LD) with vernalized for 2 wks at 5°C, 8 hrs daylight, RP GH: Reproduction period. (a) Manhattan plots comparison of bivariate and univariate analysis results, where the novel variants only discoverable when combining two phenotypes are shown in green. The horizontal dashed line represents a 5% Bonferroni-corrected genome-wide significant threshold. (b) Zooming in the novel locus detected using bivariate analysis. r: linkage disequilibrium measured as correlation coefficient between the top variant and each variant in the region.

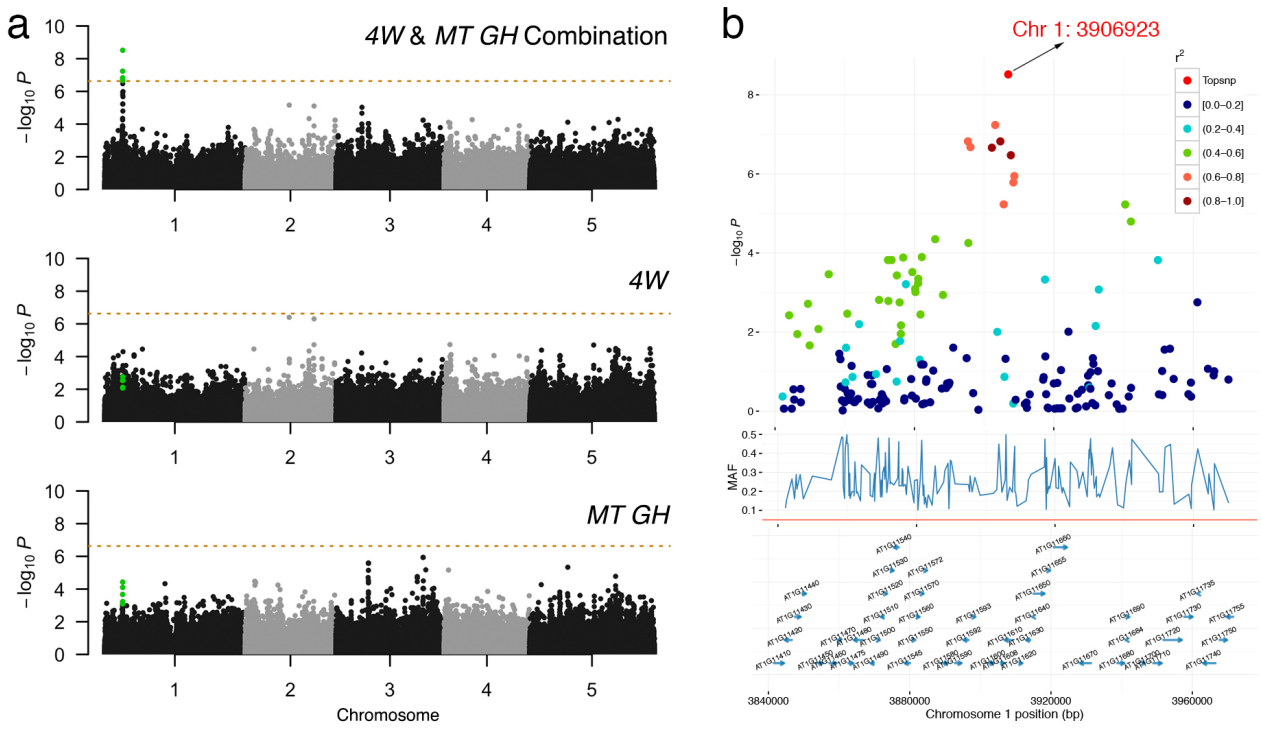

Figure S6: Bivariate genome-wide association analysis of two developmental trait, 4W: Days to flowering time (FT) under long day (LD) with vernalized for 4 wks at 5°C, 8hrs daylight, MT GH: Maturation period. (a) Manhattan plots comparison of bivariate and univariate analysis results, where the novel variants only discoverable when combining two phenotypes are shown in green. The horizontal dashed line represents a 5% Bonferroni-corrected genome-wide significant threshold. (b) Zooming in the novel locus detected using bivariate analysis. r: linkage disequilibrium measured as correlation coefficient between the top variant and each variant in the region.

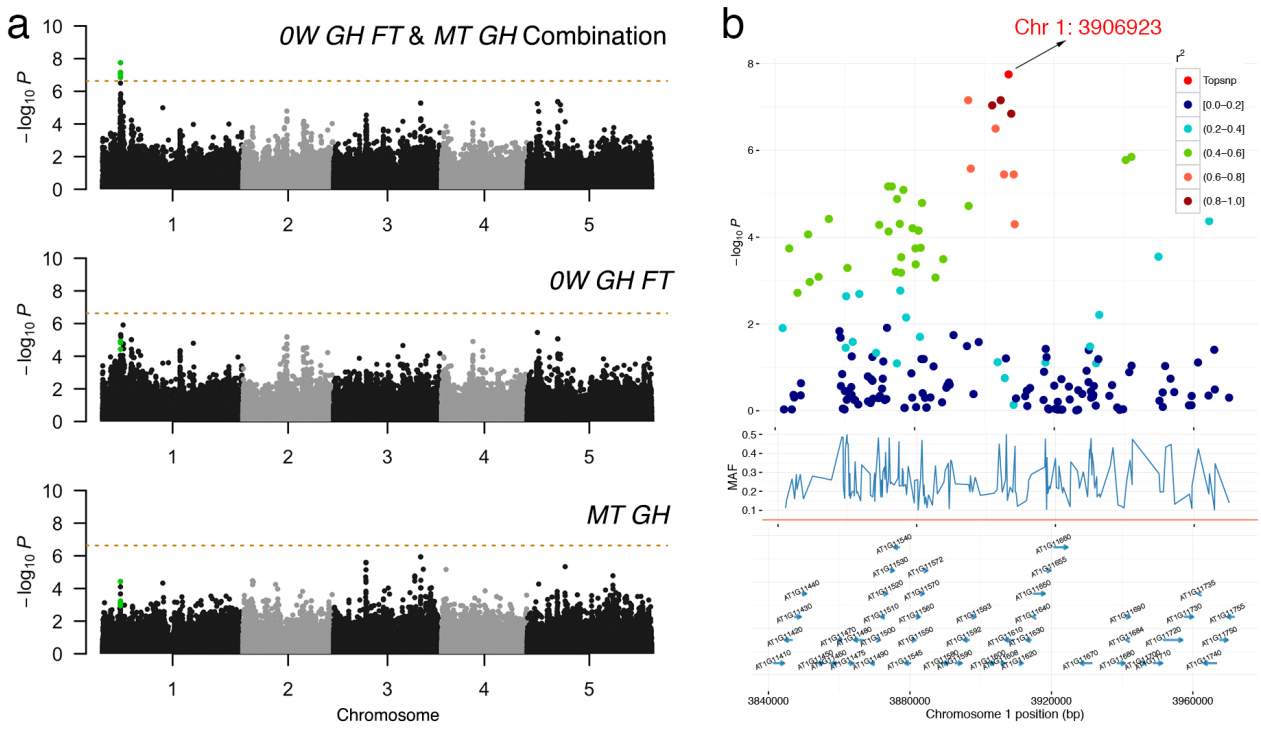

Figure S7: Bivariate genome-wide association analysis of two developmental trait, 0W GH FT: Days to flowering time (FT), MT GH: Maturation period. (a) Manhattan plots comparison of bivariate and univariate analysis results, where the novel variants only discoverable when combining two phenotypes are shown in green. The horizontal dashed line represents a 5% Bonferroni-corrected genome-wide significant threshold. (b) Zooming in the novel locus detected using bivariate analysis. r: linkage disequilibrium measured as correlation coefficient between the top variant and each variant in the region.

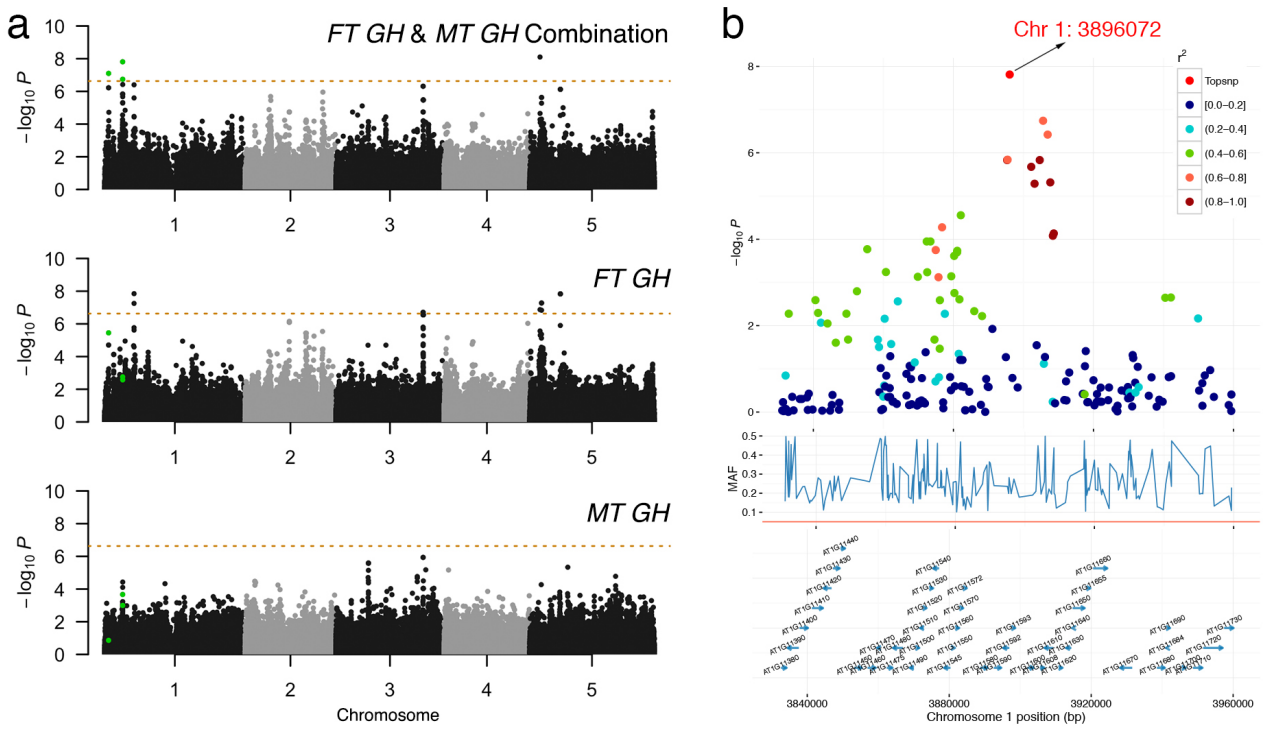

Figure S8: Bivariate genome-wide association analysis of two developmental trait, FT GH: Days to flowering (greenhouse), MT GH: Maturation period. (a) Manhattan plots comparison of bivariate and univariate analysis results, where the novel variants only discoverable when combining two phenotypes are shown in green. The horizontal dashed line represents a 5% Bonferroni-corrected genome-wide significant threshold. (b) Zooming in the novel locus detected using bivariate analysis. r: linkage disequilibrium measured as correlation coefficient between the top variant and each variant in the region.