Common genetic variants shape broad patterns of within-population variation in gene expression.

Emily B. Josephs[1], Young Wha Lee[2], Corlett W. Wood[3], Daniel J. Schoen[4], Stephen I. Wright[5], John R. Stinchcombe[6]

[1] Dept. of Plant Biology, Michigan State University, East Lansing, MI, USA.
[2] Indigo Ag, Boston, MA, USA
[3] Dept. of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA
[4] Dept. of Biology, McGill University, Montreal, QC, Canada
[5] Dept. of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

Corresponding Author: Emily Josephs, josep993@msu.edu

Abstract

Understanding the persistence of genetic variation within populations has long been a goal of evolutionary biology. One promising route towards achieving this goal is using population genetic approaches to describe how selection acts on the loci associated with trait variation. In particular, gene expression provides a model trait for addressing the challenge of the maintenance of variation because it can be measured genome-wide without information about how gene expression affects traits. Previous work has shown that loci affecting the expression of nearby genes (cis-eQTL) tend to be under purifying selection, but we lack a clear understanding of the selective forces acting on variants that affect the expression of large numbers of genes across the genome (large-effect trans-eQTL). Here, we identify loci that affect the expression of coexpression networks using genomic and transcriptomic data from one population of the obligately outcrossing plant, *Capsella grandiflora*. We identify nine loci associated with the expression of 10s to 1000s of genes. One of these loci is also associated with trait variation, but we do not detect evidence of balancing selection acting on sequence variation surrounding these loci.

Introduction

Understanding why genetic variation persists in populations has long been a goal of evolutionary biology (Mitchell-Olds et al. 2007). Variation within populations may be 1) neutral and maintained by mutation-drift balance, 2) deleterious and maintained by mutation-selection balance, or 3) conditionally beneficial and maintained by balancing selection (Johnson and Barton 2005). The availability of large genomic and phenotypic datasets offer the potential of evaluating the relative importance of these three hypotheses by identifying the genetic loci that are associated with a trait and use population genetic approaches to describe how selection acts on these loci (Josephs et al. 2017a; Sella and Barton 2019). Alongside this research program, gene expression has emerged as a powerful model trait for addressing the challenge of the maintenance of variation (Kliebenstein 2009). Gene expression is a crucial aspect of the genotype to phenotype map and expression studies provide a large set of traits that can be easily measured without prior information about how these traits might relate to fitness (Rockman and Kruglyak 2006). Examining a large set of gene expression traits thus offers the potential of understanding the evolutionary forces acting on traits in general, rather than a few or a handful of traits chosen for specific reasons. Here, we ask whether we can

detect loci that affect the expression of large numbers of genes and, if so, whether we can determine the evolutionary forces maintaining variation at these loci.

The genetic variation that shapes expression can be partitioned into two categories: cis-regulatory variants that only affect the allele they are linked to and trans-regulatory variants, that affect both alleles equally and can be located near or far from the gene they regulate (Wittkopp et al. 2004; Emerson and Li 2010). Previous work has mapped the genetic variants that affect expression (eQTLs) and shown that eQTLs that act in cis are generally under negative selection (Battle et al. 2014; Josephs et al. 2015; Glassberg et al. 2019; Hernandez et al. 2019). These studies have been possible using small samples (100-1000) because cis-eQTLs will be located near the genes they regulate, so fewer tests are needed to find them than would be needed for trans-eQTLs, which could be anywhere in the genome (Albert et al. 2018). However, a focus on cis-regulatory variants necessarily misses trans-eQTLs, which may be under different selection pressures than cis-eQTLs.

Since trans-regulatory variation affects the expression of multiple genes, trans-regulatory elements may have greater pleiotropic effects on phenotype and be subject to stronger purifying selection than cis-regulatory variants (McGuigan et al. 2014). This prediction is supported by evidence of greater trans-regulatory variation within species compared to between species (Wittkopp et al. 2004; Wittkopp et al. 2008), reduced population frequencies of distant eQTLs compared to local eQTLs (Zhang et al. 2011), and greater effect sizes of standing cis-regulatory variation than trans regulation (Kliebenstein 2009; Liu et al. 2016; Mähler et al. 2017), although these effect size differences may also be caused by differences in mutational input (Metzger et al. 2016).

However, despite the expectation that purifying selection will reduce trans-acting regulatory variation within species, linkage mapping from crossing experiments and population-based association mapping have often found trans-regulatory hotspots, where genetic variation at a locus affects expression of numerous genes (Keurentjes et al. 2007; West et al. 2007; Rockman et al. 2010; Lowry et al. 2013; Battle et al. 2014; Liu et al. 2016; Albert et al. 2018) but see (Mähler et al. 2017). Segregating trans-variation is more likely to be tissue-specific than cis-regulatory variation in humans (GTEx Consortium et al. 2017) and, in *Arabidopsis thaliana*, trans-eQTLs are particularly important for expression changes in response to drought (Lowry et al. 2013; Clauw et al. 2016). These findings suggest that trans-eQTLs contribute to standing variation, especially in specific tissues and environments.

While well-powered studies in crosses or large population samples have been able to detect significant trans-regulatory variation, the large number of tests required to map genome-wide trans eQTLs for every gene make identifying trans-eQTLs challenging in many experiments with even moderate sample sizes. One potential approach to detecting trans eQTLs is to look for loci associated with the expression of many genes (Kliebenstein et al. 2006; Hore et al. 2016; Brynedal et al. 2017). Coexpression networks can be used to summarize expression across many genes and test for associations between genetic variants and the expression of network modules. Coexpression networks are being increasingly appreciated as a powerful way to find patterns in large transcriptomic datasets (Saha et al. 2016; Josephs et al. 2017b; Mähler et al. 2017; Wisecaver et al. 2017; Palakurty et al. 2018; Mack et al. 2019). For example, coexpression networks made across conditions, tissues, and developmental time can successfully identify specialized metabolic pathways (Wisecaver et al. 2017) and coexpression modules made with a diverse panel of mouse lines correlate with phenotype (Mack et al. 2018). In addition, changes in coexpression module expression have been linked to adaptation (Campbell-Staton et al. 2017) and changing ecological conditions (Palakurty et al. 2018).

In this study we map eQTLs associated with variation of coexpression networks in a single population of the plant *Capsella grandiflora*. *Capsella grandiflora* is an obligately outcrossing member of the Brassicaceae family with large effective population size and high levels of genetic sequence diversity (Slotte et al. 2010; Williamson et al. 2014). We show how these coexpression eQTLs correspond to some extent to the eQTLs detected in standard approaches, link eQTLs and coexpression networks to phenotypic variation, and use population genomic data to test for evidence of selection on eQTLs.

Results

Coexpression module GWAS

We identified 24 coexpression modules ranging in size from 73 to 4005 genes **(Figure S1)**. We summarized expression level across modules, which we will refer to as 'module expression', for each individual using eigengenes. Module expression values showed varying distributions: some modules had normal distributions, some were bimodal, and some showed strong skews where a few individuals had very high module expression compared to other individuals **(Figure S2)**. Because the skewed distribution of module expression values could lead to false positives during association mapping, we quantile-normalized expression level for association mapping. Of the 24 *C. grandiflora* coexpression modules detected, 10 modules showed a high degree of preservation in *A. thaliana*, and another 11 showed a moderate degree of preservation **(Table S1)**.

Genome-wide association mapping identified nine SNPs associated with eight modules (FDR < 0.1, **Table 1)**. We refer to these SNPs as 'coexpression-eQTLs'. Two coexpression eQTLs were located in coding regions, three in exons, one in a conserved noncoding sequence, and four in intergenic sequence **(Figure 1, Table 1)**.

One intergenic SNP that controlled expression of the 'lightcyan', 'magenta', and 'mediumpurple' modules was also associated with the expression of a nearby gene Carubv10025200m ($p = 2.24x10^{-6}$, n=144) and heterozygotes at this SNP had stronger allele-specific expression than homozygotes ($p = 0.00283$, $n = 89$), consistent with this locus acting in cis (**Figure 2A, B)**. The gene Carubv100252000m is also in the 'lightcyan' module and it has the 18th highest intramodular connectivity out of the 2,886 genes in the module, putting it in the top 1% of connected genes. Additionally, there is transposable element (TE) located near the eQTL present in three individuals (Uzunović et al. 2019). All three of these individuals carrying the TE were homozygous for the reference allele of the eQTL, but TE presence was not associated with gene expression, allele-specific expression, or flowering time **(Fig. S3)**. Carubv10025200m is an ortholog of the *Arabidopsis thaliana* gene AT2G35040.1 and has a number of predicted functions including catalysis of a reaction involving formyltetrahydrofolate, a chemical involved in regulating flowering time (Wang et al. 2017).

How coexpression-eQTLs relate to all-by-all eQTLs

eQTL studies typically test for associations between all SNPs and all genes, so it is useful to know how coexpression-eQTLs relate to eQTLs that would be found using standard analyses (referred here to as 'all-by-all'). We tested for associations between leaf expression at all genes (n=20,792) with tag SNPs and identified 18 associations between 17 SNPs and the expression of 15 genes ($p < 4.55 \times 10^{-11}$, FDR < 0.1, **Fig. S4, Fig. S5**).

There were 14 all-by-all eQTLs located within 5kb of the genes whose expression they were associated with and all but one of the 17 all-by-all eQTLs were located within 15kb of the gene they regulated.

There were no SNPs shared between the 18 all-by-all eQTLs and the 9 coexpression module eQTLs. However, in the all-by-all eQTL analysis with a relaxed significance threshold ($p < 0.001$) all of the coexpression eQTLs were associated with the expression of at least one gene (median = 64 genes, range = 12 - 389). Within the set of associations between coexpression modules and individual genes, six out of the nine coexpression QTLs showed enrichments for associations with individual genes within the correct modules (chi squared test $p < 0.05$, **Fig. 3)**. This finding suggests that coexpression eQTLs and all-by-all eQTLs are picking up on similar associations even though there are not overlaps between significantly associated SNPs from the two methods.

Relating coexpression modules to traits

We also conducted GWAS on phenotypic traits (days to bolting, days to flower, leaf nitrogen content, leaf carbon content and leaf shape traits) following the same procedures described above for coexpression modules. No associations were significant at a FDR < 0.1 or even at an FDR < 0.25.

Module expression was correlated with a number of trait measurements. For example, expression of the 'lightcyan' module was correlated with both days to bolting (rho = 0.32, $p < 0.0005$, **Fig. 4A**) and leaf nitrogen content (rho = 0.61, $p < 0.0001$, **Fig. 4B**). We plot the relationship between all measured traits and the seven modules that had at least one significant eQTL in **Fig. 5.** We saw a number of patterns, where expression in some modules, like 'green' and 'royalblue', only had significant correlations with one trait, while other modules, like 'lightcyan' and 'turquoise', showed a significant correlation with multiple traits. In addition, while life history and leaf traits were significantly correlated with a number of modules, leaf shape traits dissection index and alpha shape dissection index were not correlated with any of these modules. As might be expected from these correlations between expression module and traits, the eQTL discovered for the 'lightcyan' module was significantly correlated with days to bolt ($p = 0.00487$, n = 139, **Figure 2C**). Individuals with one alternate copy at this eQTL flowered an average of 3 days later than individuals homozygous for the reference allele.

Signatures of selection on eQTLs.

While we have evidence that local cis-regulatory eQTLs are in general under negative selection in this population (Josephs et al. 2015), we were curious if we could detect evidence of selection on eQTLs detected in the coexpression eQTL analysis as well as in the all-by-all analysis. We measured $\boldsymbol{\pi}$ and Tajima's D at putatively neutral sites across the genome in 500 bp windows and used SweeD to test for evidence of selective sweeps in 50 SNP windows. None of the coexpression eQTLs were located in windows that were outliers (top 2.5% of windows) for $\boldsymbol{\pi}$, Tajima's D, or sweep likelihood (**Fig. S6, Fig. S7, Fig. S8**). All-by-all eQTLs were slightly more likely to be found in windows with high $\boldsymbol{\pi}$ and/or Tajima's D: Specifically, out of 50 SNPs identified to be associated with the expression of at least one gene in the all-by-all analysis, 2 were in windows that were in the top 2.5% of the distribution for $\boldsymbol{\pi}$ and 3 were in windows in the top 2.5% distribution for Tajima's D (1 expected by chance). However, this result may not be that surprising in that we have better power to detect associations with high frequency alleles, and these alleles may be preferentially found in regions with many other high frequency alleles. None of the all-by-all eQTLs were in the top 2.5% of the distribution for sweep likelihood. We also compared local (within 5kb of gene) and trans eQTL identified in the all-by-all analysis but there were no significant differences in $\boldsymbol{\pi}$, Tajima's D or selective sweep likelihood in the two groups.

Discussion

We have mapped the genetic basis of genome-wide regulatory variation within a single population of an outcrossing plant. We used coexpression modules to summarize the expression of multiple genes and identified a number of associations between module expression and genotype. These associations included coding SNPs, local eQTL for genes within the module, and noncoding SNPs. Coexpression eQTLs are different than eQTLs found in all-by-all mapping but the association signals are shared across the two methods. One of the coexpression eQTLs related to phenotypic differences in flowering time and nitrogen content. Overall, we show that relatively common trans-eQTLs are present within this single population, with consequences for expression and trait variation.

Mapping the expression level (eigengene) of coexpression modules is a powerful alternative to mapping the expression of genes individually and mapping trait values. Previous work has shown that mapping trait summaries is more successful at finding associations that mapping individual traits (Kliebenstein et al. 2006; Angelovici et al. 2017). One of the potential explanations for the success of coexpression mapping is that, if errors in measuring expression are uncorrelated across genes, summarizing expression using modules will have less error than looking at each gene independently (Kliebenstein 2009). Additionally, trans-eQTLs that affect the expression of many genes could themselves shape coexpression modules, increasing our power to detect these trans-eQTLs. However, it is also important to consider that by focussing on coexpression eQTLs, we may not be finding a representative sample of all trans-eQTLs, as is shown in comparisons of our results from the coexpression eQTL analysis and the all-by-all analysis. It is important for researchers to keep the differences in methods in mind as they decide how to do their own eQTL analyses.

While we detected a number of coexpression eQTLs that segregate in the population, it is still unclear what evolutionary forces have allowed these large-effect alleles to approach intermediate or high frequency. These alleles have appreciable frequencies (6% to 36%) and effect sizes (absolute values ranging from 0.68 to 1.25; Table 1) while affecting the expression of numerous genes (ranging from 73 to 4005). Expression of these modules, in turn, is significantly associated with flowering time, a trait closely related to fitness and expected to be under fluctuating selection in variable climates. Despite their frequency, effect size, and potential link to traits expected to be under fluctuating selection, we failed to detect population genetic evidence of balancing selection around coexpression eQTLs. The signatures of within-population balancing selection are difficult to detect in population genomic data, so it may be that these alleles are under balancing selection but this is not detectable (Charlesworth 2006). Alternatively, these alleles could be selectively neutral, perhaps because they have larger effects in lab conditions than in the wild. Or they may be at high frequency due to recent immigration from other populations, as has been seen in other plants (Monnahan et al. 2015), or from other species like the sympatric selfing species *Capsella rubella*.

One important aspect of our use of coexpression modules in the eQTL analysis is that we used "genotype networks' generated from expression data measured in the same tissue type at the same time in a set of genetically distinct individuals. Therefore, the coexpression modules we observed were shaped by genetic perturbations, not tissue or developmental differences. While coexpression measured across multiple timepoints ('developmental networks') has been linked to functional relationships (Eisen et al. 1998; Stuart et al. 2003), coexpression modules generated from genetically distinct individuals have different properties than those generated from different tissue types (Mähler et al. 2017; Schaefer et al. 2018). In some cases, this

difference is helpful: analyses combining GWAS and coexpression networks have the most power when using coexpression networks made from genetically distinct samples (Schaefer et al. 2018). However, it is important to keep in mind that the expression datasets used will affect coexpression modules.

Mapping eQTLs has furthered our understanding of the nature of genetic variation maintained within natural populations. Analyses combining genomic and transcriptomic data from natural populations are relevant in the context of models using transcriptomic data to build a mechanistic understanding of the evolutionary forces maintaining variation within populations (Boyle et al. 2017; Wray et al. 2018; Liu et al. 2019). In addition, since gene expression is important for adaptive divergence (Shapiro et al. 2004; Whitehead and Crawford 2006; Fraser 2013), understanding the maintenance of genetic variation for expression is important for understanding how organisms will adapt to new environments.

Materials and Methods

Genomic, transcriptomic, and phenotypic data

All genomic and transcriptomic sequence data was previously published in Josephs et al. (2015) and Josephs et al. (2017b). We collected individuals from a single population of *C. grandiflora* individuals located near Monodendri, Greece. We conducted a generation of random crosses in the greenhouse, and then grew 146 individuals descended from these random crosses in a growth chamber with 16 hours of daylight at 22°C. We measured traits on these individuals, extracted RNA from leaf tissue collected and flash frozen 39 days after planting using Qiagen RNAeasy kits. We extracted DNA from leaf material using a CTAB procedure. Both RNA and DNA was sequenced at the Genome Quebec facility with Hiseq 2000 with Truseq libraries with 100bp long reads. DNA was mapped to the standard *C. rubella* reference genome (Slotte et al. 2013) with Stampy (Lunter and Goodson 2011) and RNA was mapped to an exon-only reference genome using Stampy as well. SNPs were called from the genomic sequence data using GATK Unified Genotyper (Van der Auwera et al. 2002) and expression levels were measured with HTseq (Anders et al. 2015).

In addition to collecting RNA and DNA for sequencing on these 146 individuals, we measured a number of phenotypes. We measured days to bolting and days to flowering daily (measured since planting date). We collected leaves at day 49 after planting, scanned leaves, and measured leaf shape as reported in (Sicard et al. 2014). Briefly, dissection index was calculated as DI = (perimeter^2)/(4π*area), so that a circle of the same area would have a value of 1.0 and increasing values indicate increasing complexity and alpha shape dissection index is a similar parameter, but for alpha shapes. We measured leaf carbon and nitrogen content in one leaf per individual. Leaves were collected at day 49 after planting, dried, and ground to powder for elemental analysis by the Ecosystems Analysis Lab at the University of Nebraska. We note that both shape and elemental data came from different leaves than the RNAseq data. We estimated Pearson and spearman correlations between module expression and trait values with the cor.test function in R (R Core Team 2018).

Building coexpression networks

We used the program WGCNA (Langfelder and Horvath 2008)(version 1.34, running under R version 2.15.1) to identify coexpression modules present within the 146 transcriptomes using the expression level of all genes with median expression greater than 0 (n = 20,792). The coexpression analysis groups together genes with similar patterns of pairwise correlation of expression. We were interested in retaining the information embodied in the sign of the gene expression correlations, so we conducted a signed network

analysis using the following adjacency function: $a_{ij} = |0.5 + 0.5 \times \text{cor}(x_i,x_j)|^{\beta}$, where $\text{cor}(x_i,x_j)$ is the correlation of gene expression of the *i*th and *j*th gene, and β is the soft thresholding value. We used a soft thresholding value of 12, as suggested by the authors of the WGCNA package for signed networks. Genes that exhibited similar patterns of connectivity (i.e., genes showing high "topological overlap") were grouped together in the same coexpression modules, based on hierarchical clustering of topological overlap values, in which a dynamic branch-cutting algorithm was used to define initial gene co-expression modules. Module eigengenes (the first principal component of the gene expression values of modules) were calculated, and modules whose eigengenes were highly correlated were merged to arrive at the final set of co-expression modules. The resulting modules were labeled with different colors for ease of referencing. Total connectivity for these genes was previously reported in (Josephs et al. 2017b).

We used module preservation analysis to determine whether and to what extent co-expression modules detected in *C. grandiflora* (the "reference data set") are conserved in a second (or "test") data set, published RNAseq data for 19,706 genes obtained from seedling tissue of 20 ecotypes of *A. thaliana* (Gan et al. 2011). We applied the *C. grandiflora* modules identified above to the test *A. thaliana* dataset, and calculated network-based preservation statistics. We combined these statistics into a single composite preservation measure that reflects preservation of both module density and module connectivity patterns (Langfelder and Horvath 2008). We calculated a Z-statistic for the composite preservation measure by randomly permuting the module labels assigned to the test data and re-calculating the network-based preservation statistics 200 times. The Z-summary statistic asymptotically follows the normal distribution with mean 0 and SD = 1, and can be converted to a p-value under the standard normal distribution. Simulations conducted by (Langfelder and Horvath 2008)) show that Z-summary values > 10 correspond with strong evidence of module preservation in the test data, Z-summary values between 2 and 10 correspond with moderate evidence of module preservation in the test data.

Many of the modules had expression levels that were very skewed, such that a few individuals showed extremely high module expression compared to the rest of the individuals **(Fig S1)**. To reduce potential false-positives due to skewed expression levels, we quantile normalized module expression levels using the qqnorm function in R (R Core Team 2018).

Association mapping

We tested for associations between SNP genotype and individual gene expression, phenotypes, and module expression. For all association mapping analyses, we filtered out SNPs with a minor allele frequency below 0.01 and more than 0.05 missing data, leaving 5,560,798 SNPs. We used Haploview to identify 1,873,867 tag SNPs with minor allele frequency > 0.05 that described the dataset.

We tested for associations between tag SNPs with the expression of 20,794 genes using the linear model in Matrix eQTL (Shabalin 2012). While all samples came from the same population, we controlled for residual population structure by generating a centered kinship matrix with GEMMA (Zhou and Stephens 2012) and including the first five principal components of the kinship matrix as covariates. Gene expression levels were quantile normalized following the same procedure used for coexpression module eigengenes. Since all tag SNPs were tested against all genes, we conducted 39,348,132,582 tests. Matrix eQTL estimates false discovery rates using a Benjamini–Hochberg procedure.

We did association mapping with GEMMA (Zhou and Stephens 2012) on module eigengenes (PC1 of expression values of a module), morphological, and life history traits as our phenotypes. We controlled for residual population structure using the standardized kinship matrix and SNPs with minor allele frequency >

0.01 and missing data < 0.05. We used the likelihood ratio p values (Xing et al. 2012) and calculated the p-value cutoffs corresponding to a false discovery rate of 0.1 for each trait and module expression level using QValue (Dabney and Storey).

After trans-eQTLs were identified using GWAS, we conducted post-hoc tests to see if these loci were associated with total expression level. Expression levels and trait values were quantile normalized with the qqnorm function in R and then we used the lm function in R to test for a linear relationship between the number of alternate alleles in an individual and normalized expression or normalized trait values (R Core Team 2018). We used a similar procedure to test for associations between trans-eQTL genotype and allele specific expression of nearby genes, which was measured for 99 individuals in (Josephs et al. 2015), except that this time we conducted a t-test for differences in quantile-normalized allele-specific-expression between heterozygous and homozygous individuals. We also used t-tests to compare expression traits and flowering time for individuals that did or did not carry a transposable element insertion (Uzunović et al. 2019). We identified orthologs between genes in the *C. rubella* reference genome and *A. thaliana* as described in (Williamson et al. 2014).

Population genetic signatures of selection

We used genomic sequence from 188 individuals published in (Josephs et al. 2015). We downsampled all sites to 320 chromosomes per site and then calculated pi and Tajima's D in 500 bp windows across the genome at non-coding (excluding conserved non-coding sites from (Haudry et al. 2013)), intronic, and 4-fold degenerate sites. We used SweeD to calculate the likelihood of a selective sweep occuring on every 50th SNP (windows were ~600 bp wide on average) using non-conserved intergenic, intronic, and 4-fold degenerate sites for 182 individuals from the focal population (Pavlidis et al. 2013).

Acknowledgements

We thank Niroshini Epitawalage, Amanda Gorton, Robert Williamson and Jasmina Uzunović for research assistance, as well as Chad Niederhuth, Graham Coop, Jeff Ross-Ibarra, and members of the Stinchcombe, Wright, Coop and Ross-Ibarra labs for helpful comments and advice. This work was supported in part by Michigan State University through computational resources provided by the Institute for Cyber-Enabled Research. This work was funded by a National Science Foundation Graduate Research Fellowship (DGE-1048376) and a National Science Foundation Plant Genome Postdoctoral Fellowship (IOS- 1523733) to EBJ, Natural Sciences and Engineering Research Council of Canada Discovery Grants to DJS, SIW, and JRS and a Value-directed Evolutionary Genomics Initiative grant (Genome Quebec/Genome Canada) to DJS, SIW, and JRS.

Table 1: Information about significant coexpression eQTLs (FDR < 0.1). 'CNS' stands for 'conserved noncoding sequence'.

Module	SNP	Allele Frequency	Effect size	P value	Site Type
green	scaffold_1:4926298	0.14	0.8402709	5.48×10^{-8}	CNS
green	scaffold_7:3258417	0.127	0.9443182	$2.02 \times 10_{-8}$	0-fold degenerate
lightcyan	scaffold_4:9693971	0.13	0.8587179	2.19×10^{-8}	Intergene
mediumpurple3	scaffold_4:9693971	0.13	0.8562629	2.47×10^{-8}	Intergene
magenta	scaffold_4:9693971	0.13	-0.8297452	5.19×10^{-8}	Intergene
royalblue	scaffold_2:9078261	0.178	-0.8757908	5.77×10^{-9}	Intergene
saddlebrown	scaffold_5:3468364	0.065	1.284273	2.65×10^{-8}	Intron
sienna3	scaffold_6:9131753	0.164	0.8564962	4.61×10^{-8}	Intergene
sienna3	scaffold_8:9343275	0.223	-0.7306257	2.71×10^{-8}	Intron
turquoise	scaffold_6:5791152	0.271	0.7710614	6.06×10^{-9}	Exon (not 0-fold or 4-fold)
turquoise	scaffold_6:5791770	0.356	0.6880041	4.29×10^{-8}	Intron

Figures

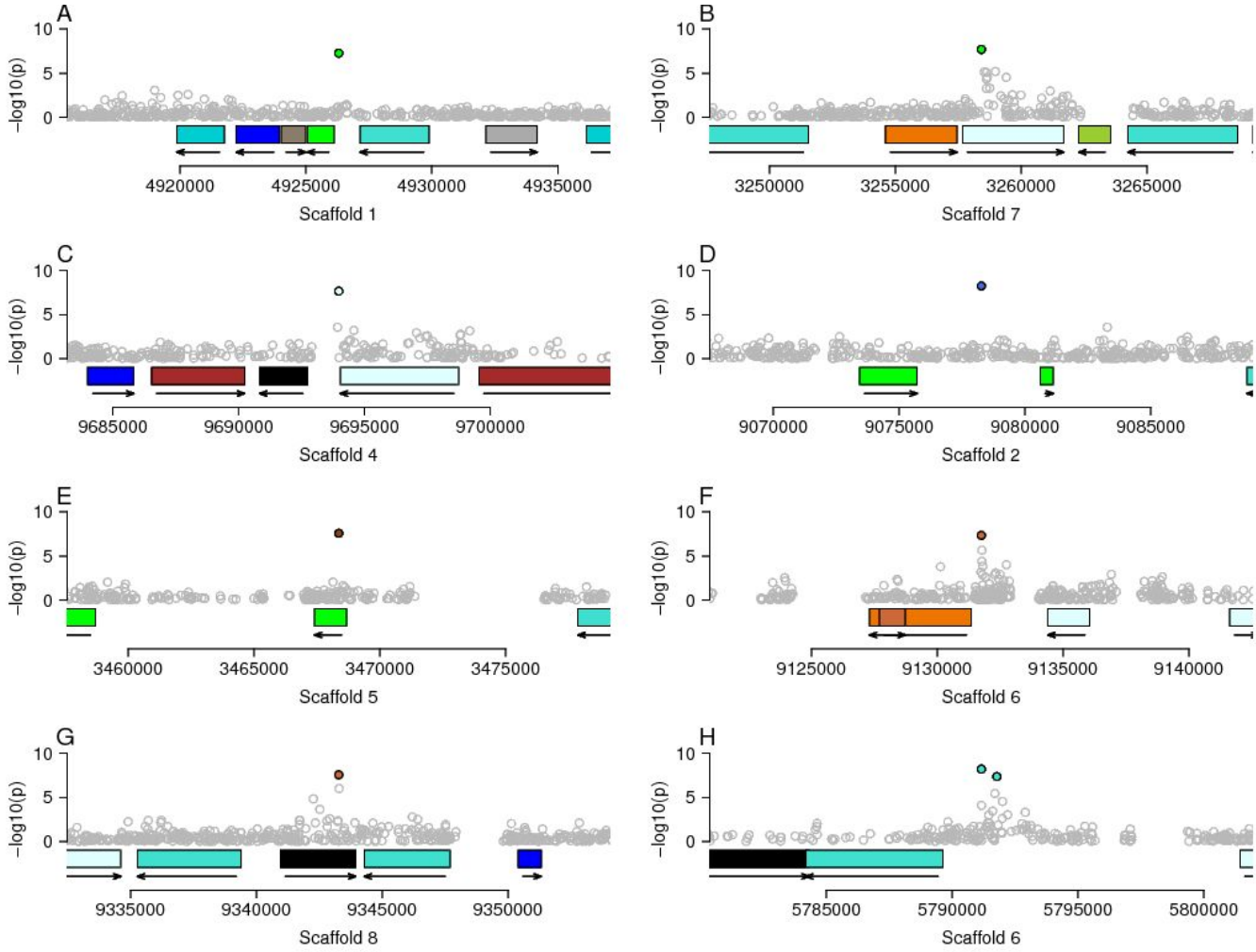

Figure 1: Physical locations of coexpression QTLs. Coexpression eQTLs are represented by points colored by the modules they are associated with. All other SNPs are plotted in gray. All SNPs are plotted by location, on the x axis, and the significance of association with the module indicated by color on the y axis. A) scaffold_1:4926298, B) scaffold_7:3258417, C) scaffold_4:9693971 (associated with the light cyan, magenta, and mediumpurple3 modules), D) scaffold_2:9078261, E) scaffold_5:3468364, F) scaffold_6:9131753, G) scaffold_8:9343275 and H) scaffold_6:5791152 and scaffold_6:5791770. Nearby genes are indicated by rectangles colored by module membership with black lines indicating the direction of transcription.

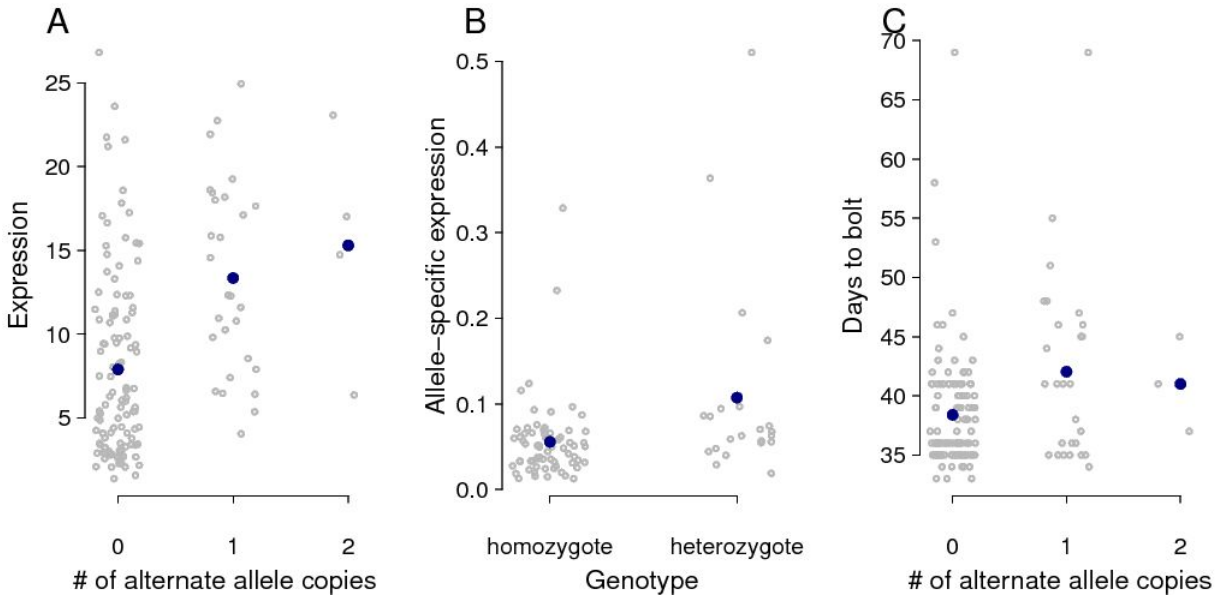

Figure 2: Effects of genotype at the eQTL for the light cyan module on A) expression of the nearby gene Carubv10025200m (n=144, p=2.24x10^{-6}), B) allele-specific expression (n=89, p=2.83x10^{-}), and C) days to bolt (n=139, p=0.00487)

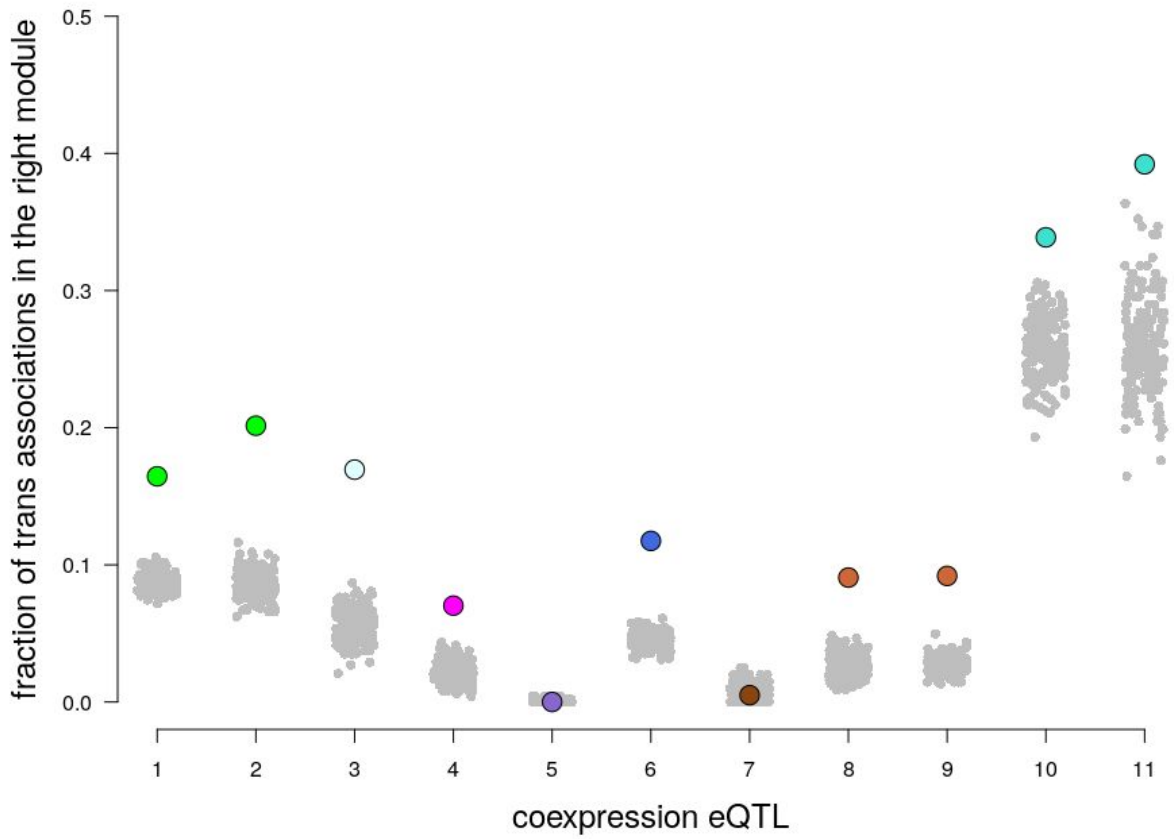

Figure 3: Shared associations between coexpression eQTL and all-by-all eQTL. This plot shows the 9 SNPs associated with coexpression modules on the x axis (the same SNP is shown in positions 3-5 because it is associated with the expression of three modules) and, on the y-axis, the colored dots show proportion of associations between that SNP and all other genes ($p < 0.01$) that are with genes in the same module whose expression is associated with the coexpression eQTL. Gray dots show the proportions observed in 200 permutations where the module identity of each gene was randomly sampled from the distribution of all modules. The order of coexpression eQTL is the same as presented in Table 1.

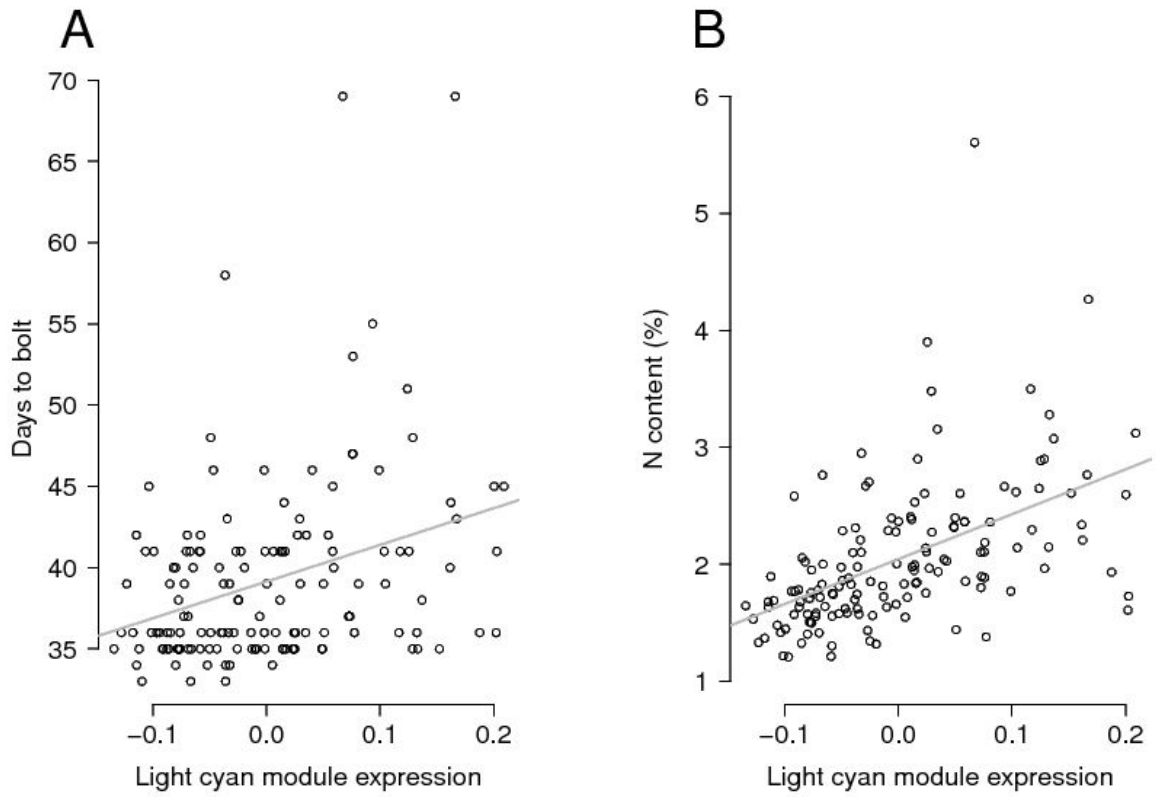

Figure 4: Correlations between module expression level and traits. A) the correlation between expression of the light cyan module and days to bolt. Dots represent individual plants, gray lines are linear regressions. B) The same figure as A, except that the y axis now shows nitrogen content in leaves.

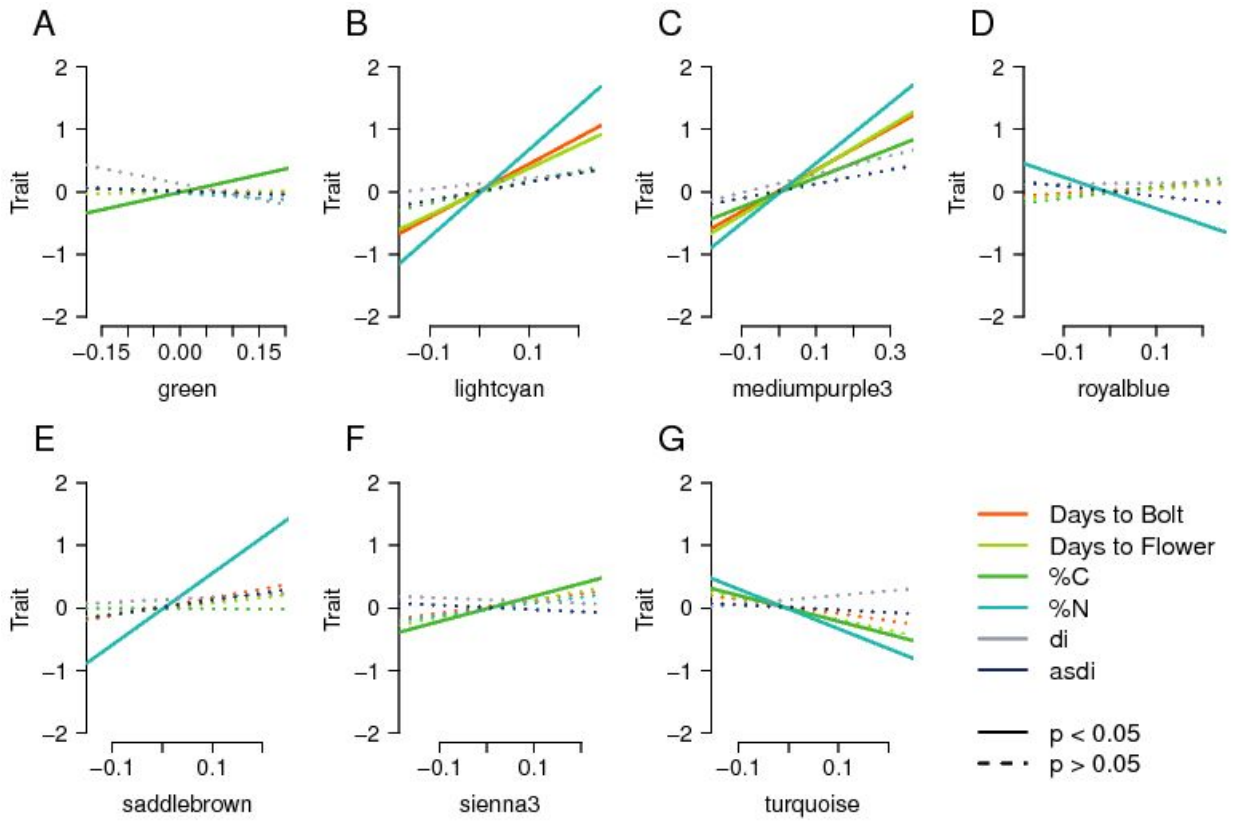

Figure 5: Correlations between module expression level and traits. Each line is a Pearson correlation between the quantile-normalized trait value and module expression. "Di" refers to dissection index and "asdi" to the alpha shape dissection index.

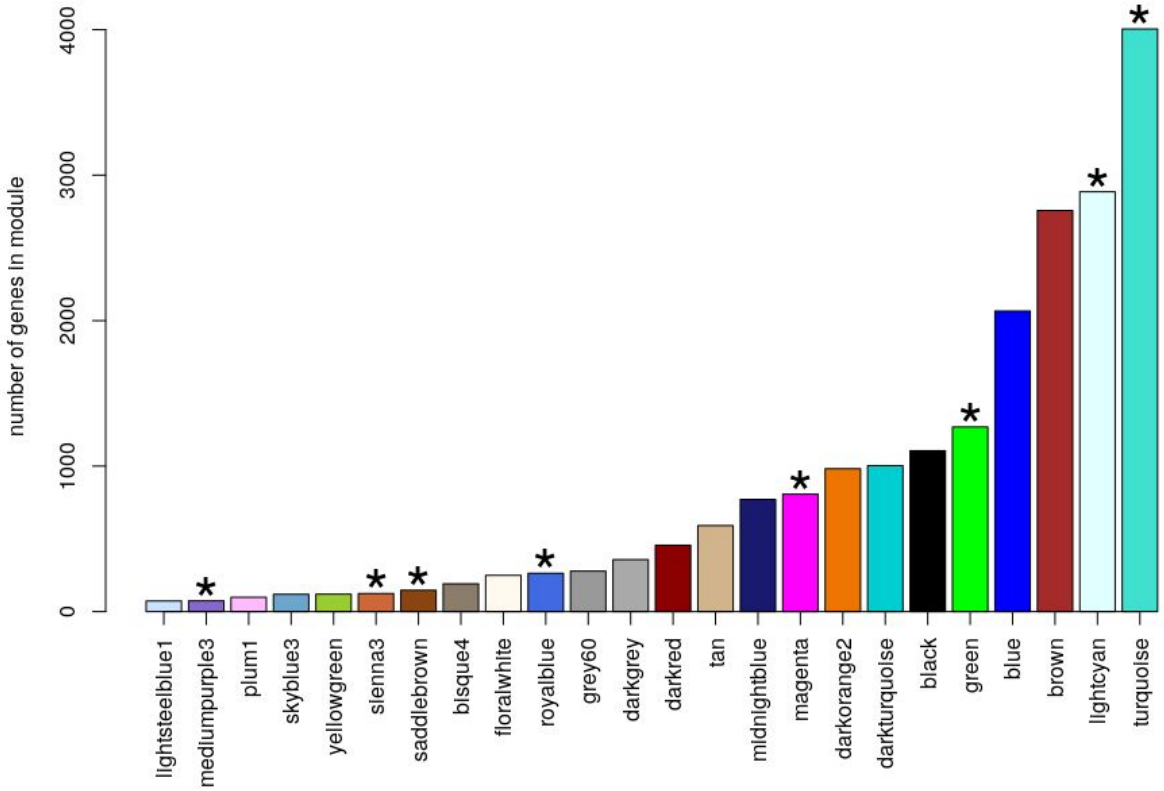

Figure S1: The number of genes in each module. There is an asterisk above modules if this module had a significant eQTL at FDR < 0.1.

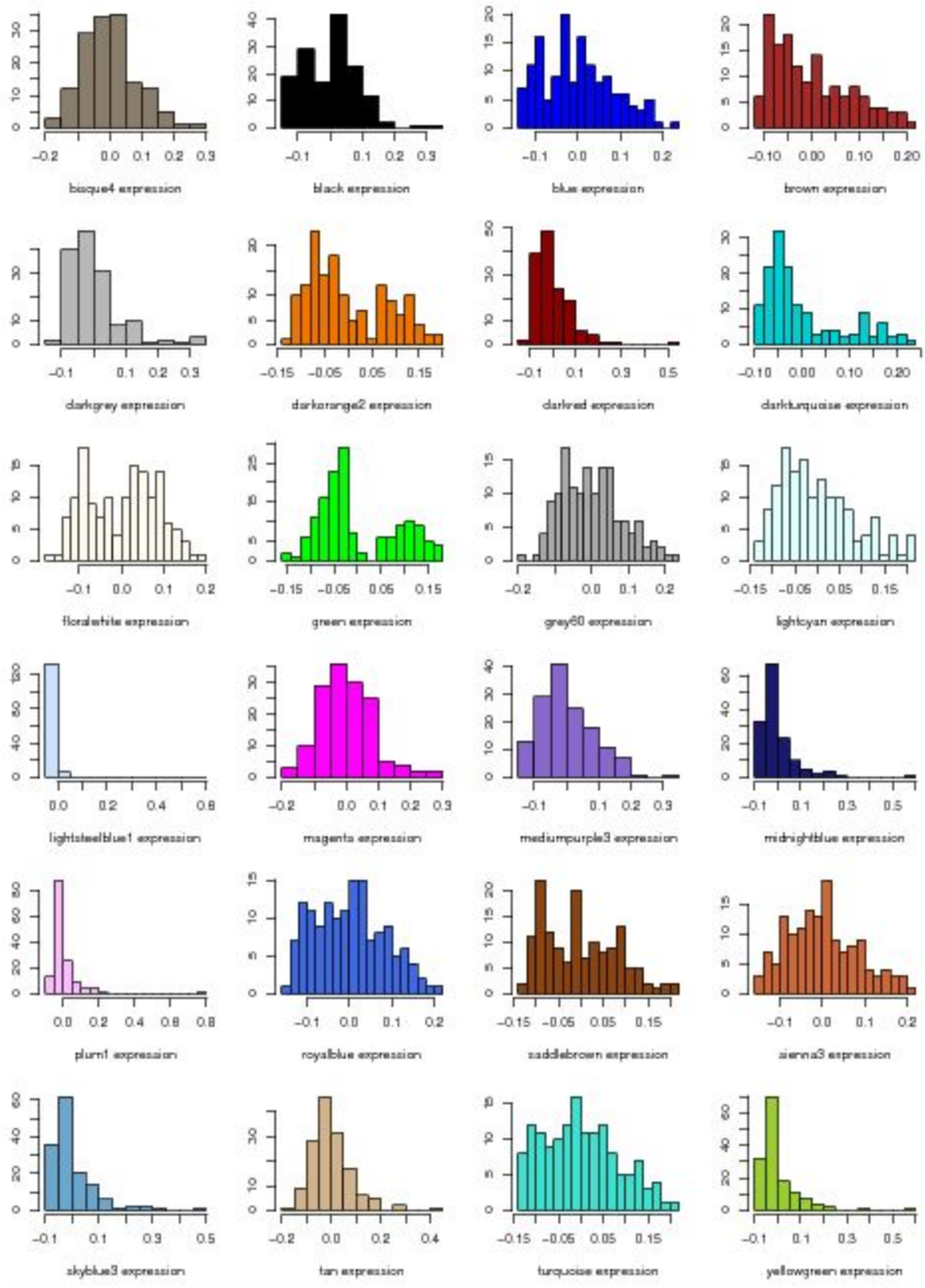

Figure S2: Distribution of un-normalized eigengenes. Each plot shows a histogram of eigengene expression for a specific coexpression module. The module name is labeled below the x axis.

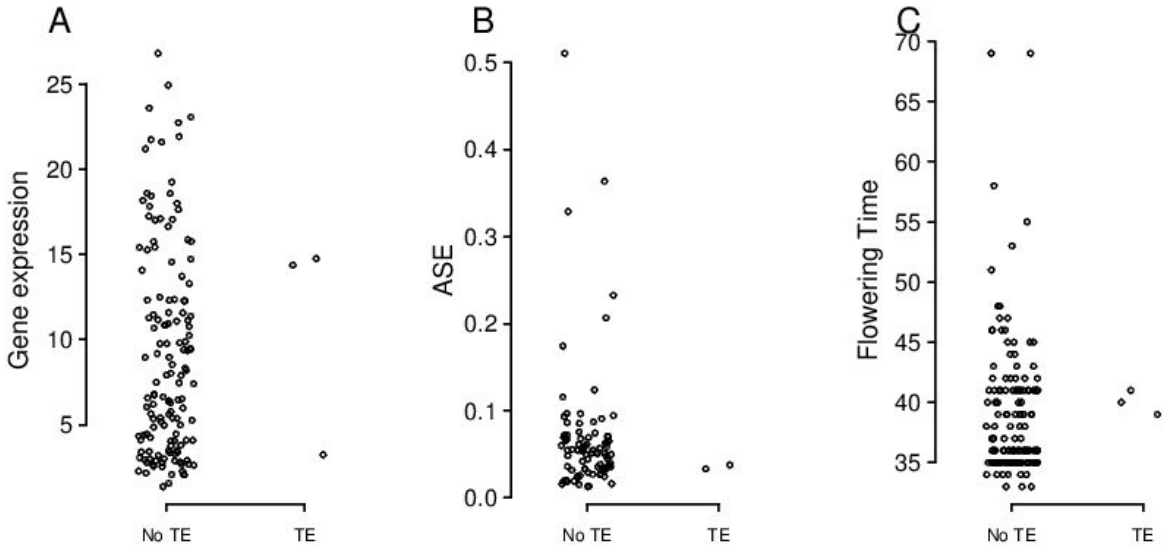

Figure S3: Effects of presence absence of TE "rnd-1_family-413" on A) expression of the nearby gene Carubv10025200m (n=144, p = 0.638) allele-specific expression (n=89, p=0.524), and C) days to bolt (n=139, p=0.805).

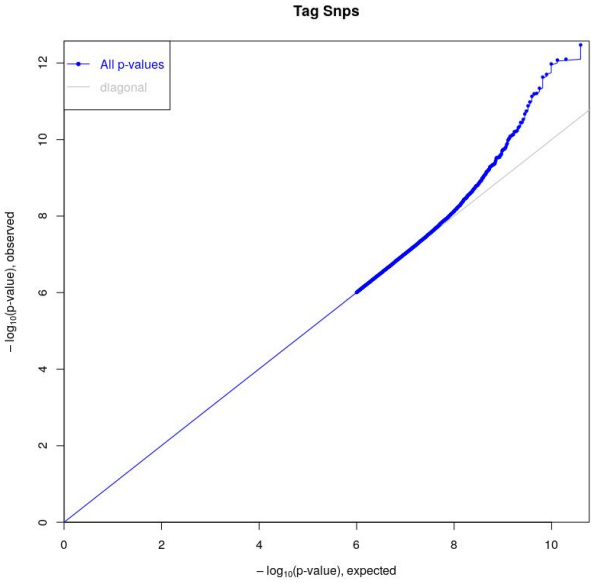

Figure S4: QQ plot for all-by-all analysis.

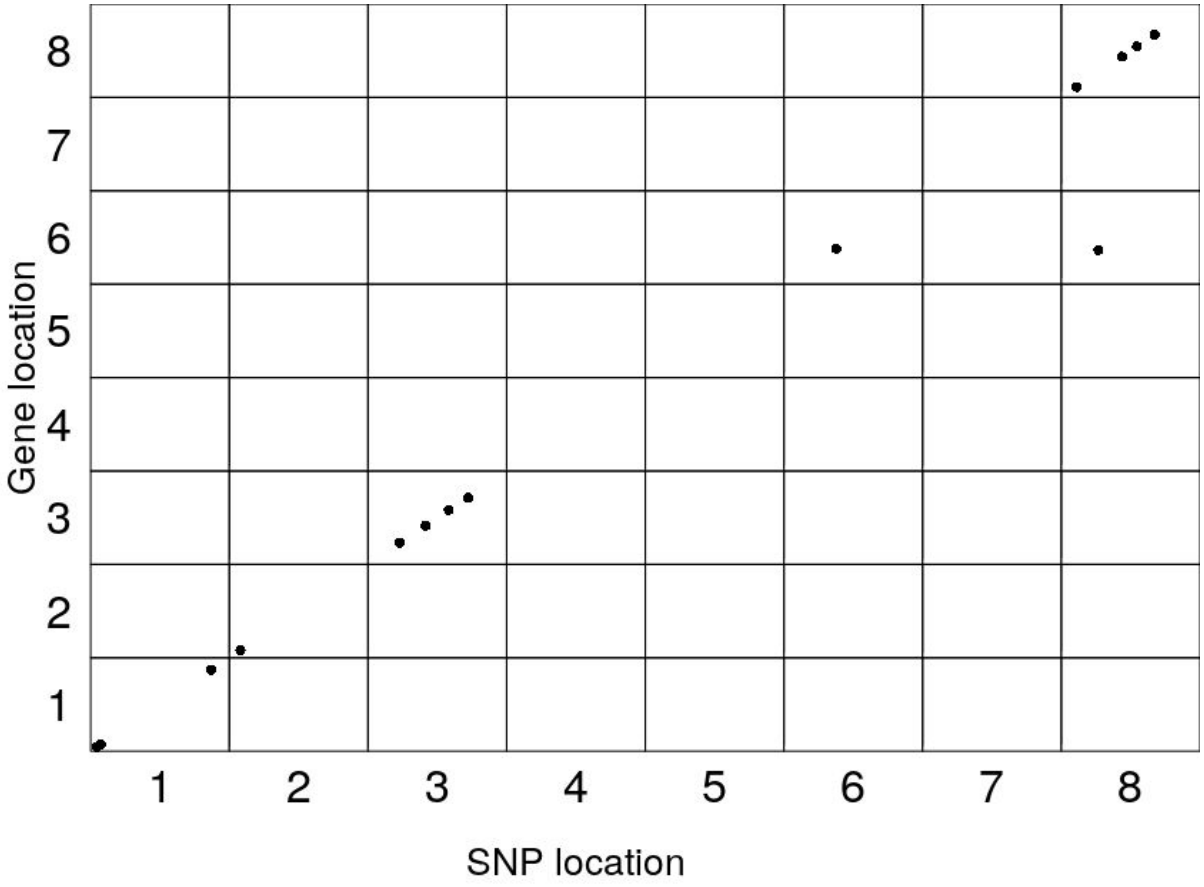

Figure S5: Associations between SNPs and gene expression level plotted for each chromosome. Black dots show associations where FDR < 0.1.

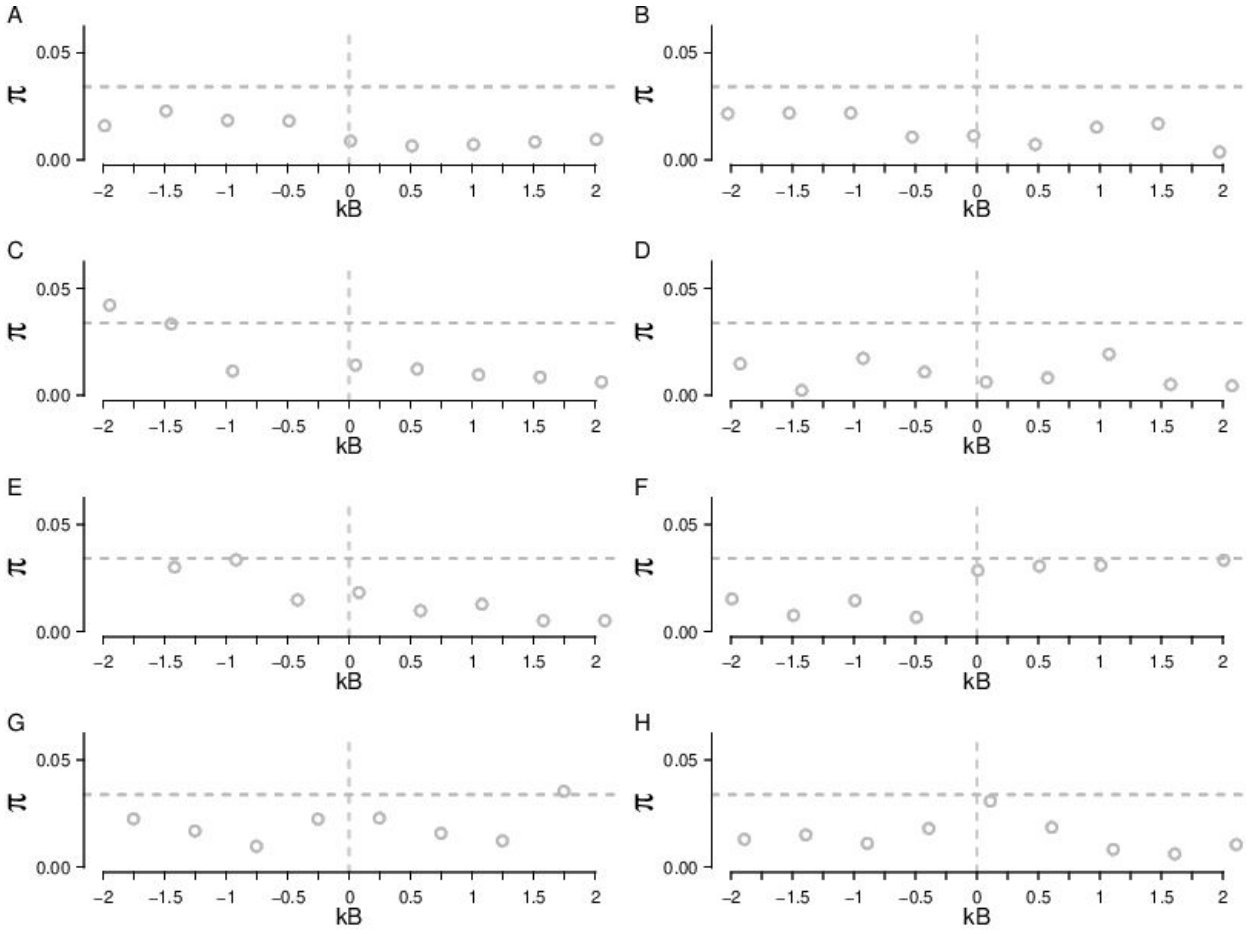

Figure S6: Diversity (π) at putatively neutral sites in 500 bp windows around coexpression-eQTLs. Dotted lines show 95% cutoffs of the observed distribution. Each panel corresponds to the eQTL shown in **Fig. 1.**

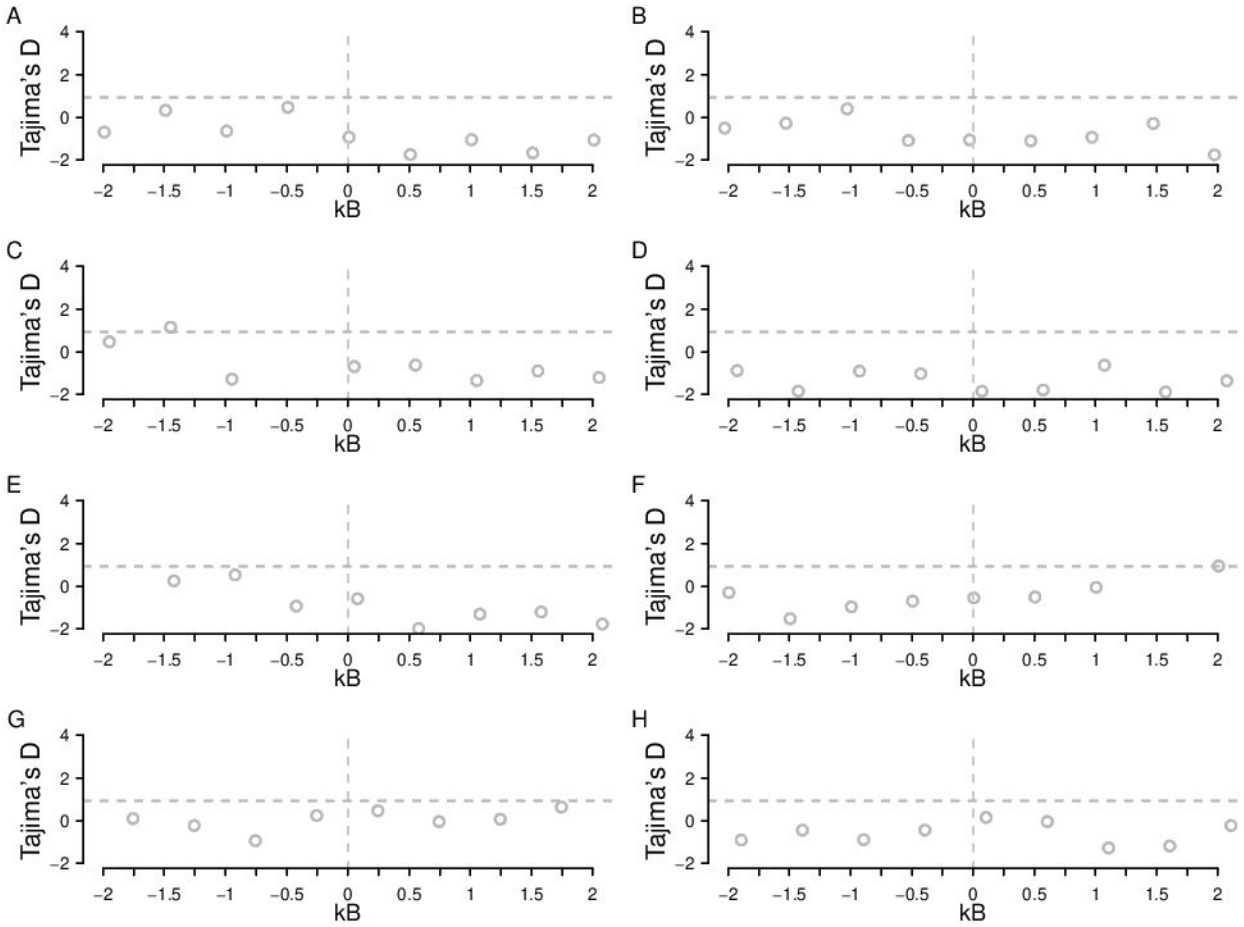

Figure S7: Tajima's D at putatively neutral sites in 500 bp windows around coexpression-eQTLs. Dotted lines show 95% cutoffs of the observed distribution. Each panel corresponds to the eQTL shown in **Fig. 1.**

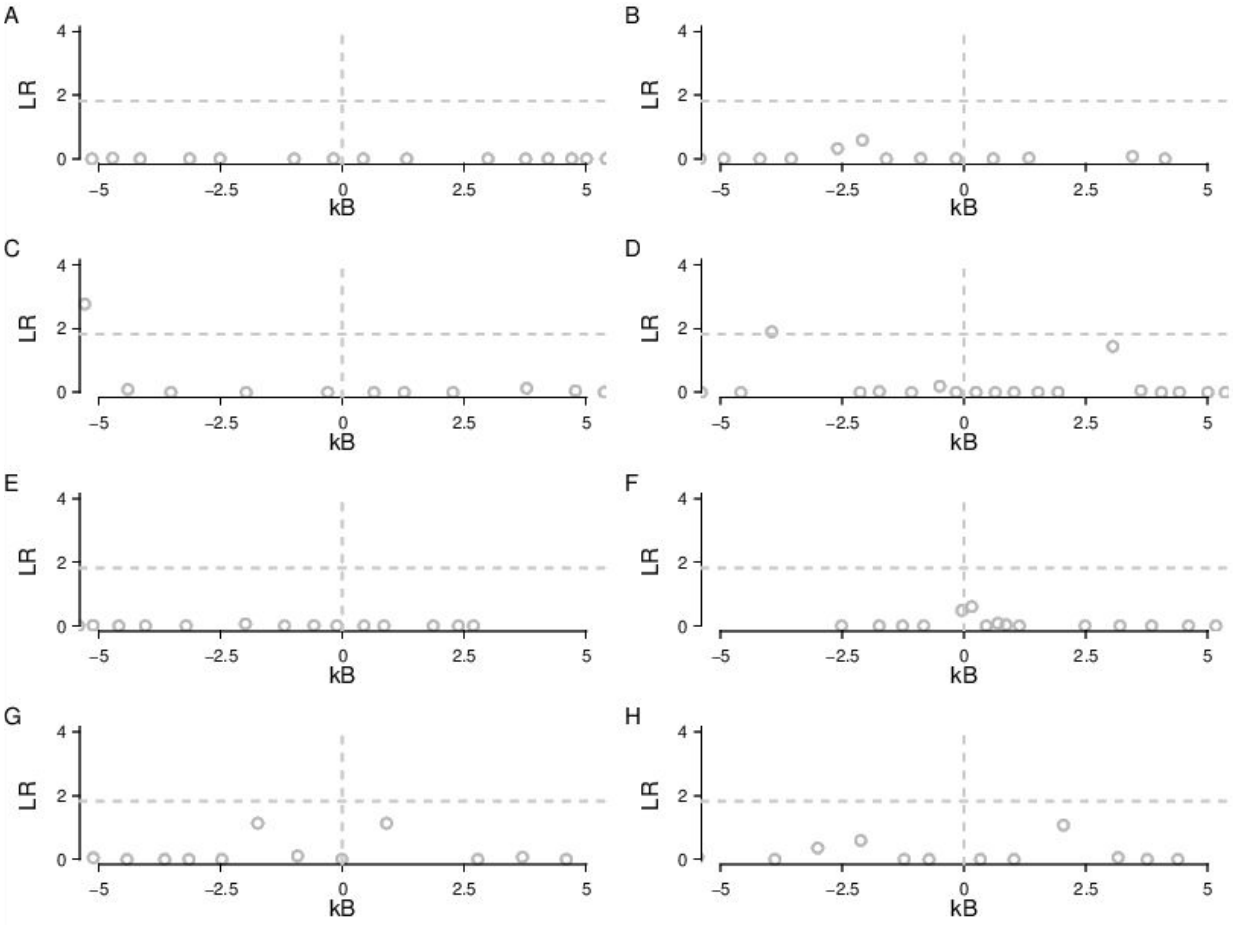

Figure S8: SweeD sweep likelihoods in windows around coexpression-eQTLs. Dotted lines show 95% cutoffs of the observed distribution of likelihoods. Each panel corresponds to the eQTL shown in **Fig. 1.**

Module	Z score	Preservation
bisque4	3.827004515	Moderate
black	8.104335146	Moderate
blue	12.66389021	High
brown	-1.149348995	Low
darkgrey	1.194865185	Low
darkorange2	4.59774833	Moderate
darkred	3.276554588	Moderate
darkturquoise	6.151745241	Moderate
floralwhite	5.314340775	Moderate
gold	9.440480564	Moderate
green	3.10372058	Moderate
grey60	3.873707795	Moderate
lightcyan	18.41008199	High
lightsteelblue1	-0.078025751	Low
magenta	0.084835506	Low
mediumpurple3	13.53882312	High
midnightblue	24.00151223	High
plum1	10.28628803	High
royalblue	14.43400336	High
saddlebrown	10.66207142	High
sienna3	11.23932229	High
skyblue3	21.61021517	High
tan	2.090374819	Moderate
turquoise	6.410564238	Moderate
yellowgreen	10.23190232	HIgh

Table S1: Module preservation Z statistics. Following (Langfelder and Horvath 2008), we define modules with a Z score > 10 as highly preserved and modules with a Z score between 2 and 10 as moderately preserved.

References

Albert FW, Bloom JS, Siegel J, Day L, Kruglyak L. 2018. Genetics of trans-regulatory variation in gene expression. Elife [Internet] 7. Available from: http://dx.doi.org/10.7554/eLife.35471

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169.

Angelovici R, Batushansky A, Deason N, Gonzalez-Jorge S, Gore MA, Fait A, DellaPenna D. 2017. Network-Guided GWAS Improves Identification of Genes Affecting Free Amino Acids. Plant Physiol. 173:872–886.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, et al. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24:14–24.

Boyle EA, Li YI, Pritchard JK. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169:1177–1186.

Brynedal B, Choi J, Raj T, Bjornson R, Stranger BE, Neale BM, Voight BF, Cotsapas C. 2017. Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. Am. J. Hum. Genet. 100:581–591.

Campbell-Staton SC, Cheviron ZA, Rochette N, Catchen J, Losos JB, Edwards SV. 2017. Winter storms drive rapid phenotypic, regulatory, and genomic shifts in the green anole lizard. Science 357:495–498.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2:e64.

Clauw P, Coppens F, Korte A, Herman D, Slabbinck B, Dhondt S, Van Daele T, De Milde L, Vermeersch M, Maleux K, et al. 2016. Leaf Growth Response to Mild Drought: Natural Variation in Arabidopsis Sheds Light on Trait Architecture. Plant Cell 28:2417–2434.

Dabney A, Storey JD. qvalue: Qvalue estimation for false discovery rate control. Available from: http://www.bioconductor.org/packages/release/bioc/manuals/qvalue/man/qvalue.pdf

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. 95:14863–14868.

Emerson JJ, Li W-H. 2010. The genetic basis of evolutionary change in gene expression levels. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365:2581–2590.

Fraser HB. 2013. Gene expression drives local adaptation in humans. Genome Res. 23:1089–1096.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana.

Nature 477:419–423.

Glassberg EC, Gao Z, Harpak A, Lan X, Pritchard JK. 2019. Evidence for Weak Selective Constraint on Human Gene Expression. Genetics 211:757–772.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. 2017. Genetic effects on gene expression across human tissues. Nature 550:204–213.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. 45:891–898.

Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. 2019. Ultrarare variants drive substantial cis heritability of human gene expression. Nat. Genet. 51:1349–1355.

Hore V, Viñuela A, Buil A, Knight J, McCarthy MI, Small K, Marchini J. 2016. Tensor decomposition for multiple-tissue gene expression experiments. Nat. Genet. [Internet]. Available from: http://dx.doi.org/10.1038/ng.3624

Johnson T, Barton N. 2005. Theoretical models of selection and mutation on quantitative traits. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360:1411–1425.

Josephs EB, Lee YW, Stinchcombe JR, Wright SI. 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proc. Natl. Acad. Sci. U. S. A. 112:15390–15395.

Josephs EB, Stinchcombe JR, Wright SI. 2017a. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? New Phytol. [Internet]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.14410

Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. 2017b. The relationship between selection, network connectivity, and regulatory variation within a population of Capsella grandiora. Genome Biol. Evol. [Internet]. Available from: http://dx.doi.org/10.1093/gbe/evx068

Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJM, Vreugdenhil D, Koornneef M, Jansen RC. 2007. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc. Natl. Acad. Sci. U. S. A. 104:1708–1713.

Kliebenstein D. 2009. Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. Annu. Rev. Plant Biol. 60:93–114.

Kliebenstein DJ, West MAL, van Leeuwen H, Loudet O, Doerge RW, St Clair DA. 2006. Identification of QTLs controlling gene expression networks defined a priori. BMC Bioinformatics 7:308.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.

Liu H, Luo X, Niu L, Xiao Y, Chen L, Liu J, Wang X, Jin M, Li W, Zhang Q, et al. 2016. Distant eQTLs and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. Mol. Plant [Internet]. Available from: http://dx.doi.org/10.1016/j.molp.2016.06.016

Liu X, Li YI, Pritchard JK. 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177:1022–1034.e6.

Lowry DB, Logan TL, Santuari L, Hardtke CS, Richards JH, DeRose-Wilson LJ, McKay JK, Sen S, Juenger TE. 2013. Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis. Plant Cell 25:3266–3279.

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

Mack KL, Ballinger MA, Phifer-Rixey M, Nachman MW. 2018. Gene regulation underlies environmental adaptation in house mice. Genome Res. 28:1636–1645.

Mack KL, Phifer-Rixey M, Harr B, Nachman MW. 2019. Gene Expression Networks Across Multiple Tissues Are Associated with Rates of Molecular Evolution in Wild House Mice. Genes [Internet] 10. Available from: http://dx.doi.org/10.3390/genes10030225

Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. 2017. Gene co-expression network connectivity is an important determinant of selective constraint. PLoS Genet. 13:e1006402.

McGuigan K, Collet JM, Allen SL, Chenoweth SF, Blows MW. 2014. Pleiotropic mutations are subject to strong stabilizing selection. Genetics 197:1051–1062.

Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B, Wittkopp PJ. 2016. Contrasting frequencies and effects of cis- and trans-regulatory mutations affecting gene expression. Mol. Biol. Evol. [Internet]. Available from: http://dx.doi.org/10.1093/molbev/msw011

Mitchell-Olds T, Willis JH, Goldstein DB. 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? Nat. Rev. Genet. 8:845–856.

Monnahan PJ, Colicchio J, Kelly JK. 2015. A genomic selection component analysis characterizes migration-selection balance. Evolution [Internet]. Available from: http://dx.doi.org/10.1111/evo.12698

Palakurty SX, Stinchcombe JR, Afkhami ME. 2018. Cooperation and coexpression: How coexpression networks shift in response to multiple mutualists. Mol. Ecol. 27:1860–1873.

Pavlidis P, Živkovic D, Stamatakis A, Alachiotis N. 2013. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30:2224–2234.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. Available from: https://www.R-project.org/

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. Nat. Rev. Genet. 7:862–872.

Rockman MV, Skrovanek SS, Kruglyak L. 2010. Selection at linked sites shapes heritable phenotypic variation in C. elegans. Science 330:372–376.

Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, GTEx Consortium, Engelhardt BE, Battle A. 2016. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. bioRxiv [Internet]:078741. Available from: http://biorxiv.org/content/early/2016/10/02/078741

Schaefer R, Michno J-M, Jeffers J, Hoekenga OA, Dilkes BP, Baxter IR, Myers C. 2018. Integrating co-expression networks with GWAS to prioritize causal genes in maize. Plant Cell [Internet]. Available from: http://dx.doi.org/10.1105/tpc.18.00299

Sella G, Barton NH. 2019. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. Annu. Rev. Genomics Hum. Genet. [Internet]. Available from: http://dx.doi.org/10.1146/annurev-genom-083115-022316

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28:1353–1358.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717–723.

Sicard A, Thamm A, Marona C, Lee YW, Wahl V, Stinchcombe JR, Wright SI, Kappel C, Lenhard M. 2014. Repeated evolutionary changes of leaf morphology caused by mutations to a homeobox gene. Curr. Biol. 24:1880–1886.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in Capsella grandiflora, a plant species with a large effective population size. Mol. Biol. Evol. 27:1813–1821.

Slotte T, Hazzouri KM, Arvid Ågren J, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Killian Newman L, et al. 2013. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat. Genet. 45:831–835.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science 302:249–255.

Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. 2019. Transposable elements are important contributors to standing variation in gene expression in Capsella grandiflora. Mol. Biol. Evol. [Internet]. Available from: http://dx.doi.org/10.1093/molbev/msz098

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2002. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc.

Wang L, Kong D, Lv Q, Niu G, Han T, Zhao X, Meng S, Cheng Q, Guo S, Du J, et al. 2017. Tetrahydrofolate Modulates Floral Transition through Epigenetic Silencing. Plant Physiol. 174:1274–1284.

West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St Clair DA. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics 175:1441–1450.

Whitehead A, Crawford DL. 2006. Neutral and adaptive variation in gene expression. Proc. Natl. Acad. Sci. U. S. A. 103:5425–5430.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of Capsella grandiflora. PLoS Genet. 10:e1004622.

Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. 2017. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell 29:944–959.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. Nature 430:85–88.

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between Drosophila species. Nat. Genet. 40:346–350.

Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. 2018. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell 173:1573–1580.

Xing G, Lin C-Y, Wooding SP, Xing C. 2012. Blindly using Wald's test can miss rare disease-causal variants in case-control association studies. Ann. Hum. Genet. 76:168–177.

Zhang X, Cal AJ, Borevitz JO. 2011. Genetic architecture of regulatory variation in Arabidopsis thaliana. Genome Res. 21:725–733.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44:821–824.