Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript Level Variation in Arabidopsis


*University of California-Davis, Department of Plant Sciences, Davis, CA 95616-8780 USA
†Purdue University, Department of Statistics, West Lafayette, IN 47907-2067 USA
‡The Genome Center, University of California-Davis, Davis, CA 95616 USA
Running head: Global eQTL Mapping

Key words: expression QTL (eQTL), functional genomics, quantitative genetics, transcript level variation, recombinant inbred population, Arabidopsis

Corresponding Author:
Dina A. St.Clair
University of California-Davis
Department of Plant Sciences
Mail Stop 3
One Shields Ave
Davis, CA 95616-8780
Fax: 530-752-9659
Phone: 530-752-1740
Email: dastclair@ucdavis.edu
ABSTRACT

The genetic architecture of transcript level variation is largely unknown. The genetic determinants of transcript level variation were characterized in a recombinant inbred line (RIL) population (n = 211) of *Arabidopsis thaliana* using whole genome microarray analysis and expression QTL (eQTL) mapping of transcript levels as expression traits (e-traits). Genetic control of transcription was highly complex: one-third of the quantitatively controlled transcripts/e-traits were regulated by *cis*-eQTLs and many *trans*-eQTLs mapped to ‘hotspots’ that regulated hundreds to thousands of e-traits. Several thousand eQTLs of large phenotypic effect were detected, but almost all (93%) of the 36,871 eQTLs were associated with small phenotypic effects ($R^2 < 0.3$). Many transcripts/e-traits were controlled by multiple eQTLs with opposite allelic effects and exhibited higher heritability in the RILs than their parents, suggesting non-additive genetic variation. To our knowledge this is the first large-scale global eQTL study in a relatively large plant mapping population. It reveals that the genetic control of transcript level is highly variable and multifaceted, and that this complexity may be a general characteristic of eukaryotes.
INTRODUCTION

Transcript levels, when assessed in an experimental or mapping population, can be considered as quantitative traits and their variation used to map expression QTLs (eQTLs) (JANSEN and NAP 2001; DOERGE 2002; SCHADT et al. 2003). Recent studies in a limited number of organisms indicate that the levels of transcripts are heritable and can be under multigenic control (BREM et al. 2002; SCHADT et al. 2003; YVERT et al. 2003; KIRST et al. 2004; BREM and KRUGLYAK 2005; BYSTRYKH et al. 2005; CHESLER et al. 2005; DRAKE et al. 2005; HUBNER et al. 2005; DeCOOK et al. 2006). Global eQTL analyses in yeast, mice and humans have detected significant levels of cis polymorphism controlling individual genes, as well as evidence for clustered trans-eQTLs that simultaneously regulate a large fraction of the transcriptome (BREM et al. 2002; SCHADT et al. 2003; MORLEY et al. 2004). In yeast, the complex inheritance of transcript levels (BREM and KRUGLYAK 2005) has been revealed by detecting significant levels of non-additive genetic variance, epistatic interactions and transgressive segregation. However, comparable studies in plants have yet to be reported. Furthermore, the relationship between transcript level variation and downstream phenotypic trait variation is not well understood (MACKAY 2001; GIBSON and WEIR 2005).

While the identification of cis-eQTLs has led to successful QTL cloning in plants (KLIBENSTEIN et al. 2001; LAMBRIX et al. 2001; KROYMANN et al. 2003; CAICEDO et al. 2004; WERNER et al. 2005; ZHANG et al. 2006), genome-wide eQTL analyses in large mapping populations of plants have been lacking. To our knowledge, the study that we describe here is among the first large global eQTL mapping studies in plants. Our overarching goal was to assess and report on the genetic architecture of transcript level variation in a higher plant species. To address this goal, we employed a sample of 211 recombinant inbred lines (RILs) from a
population derived from a cross between two inbred *Arabidopsis thaliana* accessions (Bayreuth-0 and Shahdara) that were grown in a biologically replicated experiment to quantify genome-wide transcript levels using Affymetrix whole genome microarrays. The 211 RILs represent a sample of individuals from the population of all possible individuals who represent this particular cross. Transcript level variation in the 211 RILs was measured and used as expression quantitative traits (e-traits) to map eQTLs using a framework map of single feature polymorphism (SFP) markers (West et al. 2006) for the 211 RILs. We report on the genomic distributions for eQTLs, as well as the phenotypic effects and numbers of cis- and trans-eQTLs. We also estimate the broad-sense heritability of transcript levels/e-traits in the RILs and parent lines, and assess transgressive segregation and non-additive genetic variation.

**MATERIALS AND METHODS**

**Plant Material and Experimental Conditions:** Seeds for *Arabidopsis thaliana* accessions Bayreuth (Bay-0), Shahdara (Sha), and the F$_8$ generation of a Bay-0 × Sha recombinant inbred line (RIL) population (Loudet et al. 2002) were obtained from TAIR (stock #CS57920; www.arabidopsis.org). The advanced generation RILs were created by single seed descent from the F$_2$ generation (Loudet et al. 2002). The RIL (F$_8$) plants and parental accessions were grown in a single growth chamber at UC Davis and allowed to self-pollinate; seed was harvested from individual plants to produce sufficient seed for each homozygous F$_9$ line for our replicated experiments.

For the Bay-0 × Sha RIL microarray experiment, five plants per biological replicate for each of 211 RILs, plus parental controls, Bay-0 and Sha, were grown as previously described (West et al. 2006). At six weeks post-germination, the plants were treated with 0.02% Silwet
L77, a surfactant (Lehle Seeds, Round Rock, TX, USA) and harvested 28 hrs post treatment, as previously described (WEST et al. 2006).

**Microarray Hybridization and Quality Control:** Total RNA was extracted, converted to cDNA, and labeled with biotin as recommended by the manufacturer (Affymetrix, Santa Clara, CA, USA; www.affymetrix.com). Affymetrix ATH1 GeneChip microarrays were hybridized, washed, scanned, and checked for quality as reported previously (WEST et al. 2006). The RIL haplotypes obtained using SFP markers derived from the same GeneChips were consistent with the haplotypes determined previously by microsatellite analysis of genomic DNA, indicating that each GeneChip dataset was derived from the designated RIL (LOUDET et al. 2002; WEST et al. 2006).

**QTL Analysis:** Raw image data from the RIL GeneChips were converted to numeric data via Bioconductor software (www.bioconductor.org). Due to the non-biological variation that is present in every microarray experiment, we normalized across all arrays for the purpose of addressing any non-linear relationships between arrays. Quantile normalization reduces non-biological variation that is largely due to the technology itself, and when applied at the probe level it has been shown to outperform other normalization methods that are based on what is referred to as a “base-line array” (BOLSTAD et al. 2003). After the quantile normalization, raw intensity values were log₂ transformed. For this analysis of gene expression traits (e-traits), transcript levels of each gene were averaged over the two microarray replicates per RIL to give a single transcript measurement per RIL.

A genetic map was estimated using JoinMap (Kyazma B.V., Wageningen, Netherlands), 540 SFP markers obtained as described previously (WEST et al. 2006), and 38 reference SSR
markers (LOUDET et al. 2002) scored on the 211 RILs. In previous work (WEST et al. 2006) we used a subset of 148 RILs to create a high-density SFP map. When the microarray data for the remaining 63 RILs became available, we re-estimated the genetic map based on a total sample of 211 RILs. The resulting map had an average marker density of 0.78 cM. Composite interval mapping (CIM) analysis (ZENG et al. 1999) was employed in conjunction with a 5 cM framework map (93 SFP and 2 SSR markers; Supplemental Table 1 online). The “zmapqtl” CIM module of QTL-Cartographer Version 1.17 (BASTEN et al. 1999) with a walking speed of 1 cM and a window size of 10 cM was employed to analyze each e-trait. To obtain a threshold criterion for declaring statistically significant eQTL, a Global Permutation Threshold (GPT) was obtained by permuting the e-traits while maintaining the genetic information. Because it was computationally not feasible (i.e., at least 1000 permutations for each of the 22,794 e-traits) to compute an empirical threshold value (CHURCHILL and DOERGE 1994) for each e-trait that we modeled with composite interval mapping, we adopted a GPT approach as follows. For each of 100 randomly selected transcripts or e-traits, the null distribution of the maximum likelihood ratio test (LRT) statistic was empirically estimated using permutation thresholds based on 1,000 permutations (CHURCHILL and DOERGE 1994). Inspection of the means and standard deviations for the 100 null distributions from the permuted transcript values revealed little variation, and the average 95 percentile permutation threshold was 12.0329, with a sample standard deviation of 0.3478. Permutation thresholds were then computed based on each of the 100 representative null distributions. A representative null distribution based on 100,000 maximum LRT statistics (1,000 permutations × 100 randomly selected e-traits) was employed for all 22,794 non-control transcripts represented on the ATH1 microarray. The GPT was calculated as the 95% upper bound of the representative null distribution and equaled 12.0583.
**Identification of eQTL location:** After conducting composite interval mapping for each of the 22,794 e-traits, it was necessary to summarize the resulting eQTL with a certain level of resolution to the known location of the gene features on the map. For this we utilized the Eqtl module of QTL-Cartographer to count, summarize and locate potential eQTLs (BASTEN *et al.* 1999). Since experimental limitations (e.g., genetic map density and the sample size of the mapping population), statistical issues (e.g., limitations of the statistical method), and genetic factors (e.g., ghost QTL, false positive QTL, linked QTL, etc.) come into play when locating QTL, we attempted to summarize our eQTL results in a general manner by using an “exclusionary window” that is available in the Eqtl module to select and count eQTLs. However, we first investigated the effect of varying the exclusionary window size (5, 10 and 20 cM) on the eQTL summary statistics. Changing the exclusionary window size had no effect on the number of transcripts/e-traits with detectable eQTLs, but increasing it decreased the number of eQTLs that were counted for each transcript (Supplemental Fig. 1 online). The average number of eQTLs counted per transcript decreased from 3.7 to 2.3 when results were compared between windows of size 5 and 20 cM (Supplemental Table 2 online), but this did not affect the genetic positions of the large *trans*-effect peaks (data not shown). These results suggest that the smaller exclusionary window sizes are failing to accommodate closely linked eQTLs, ghost QTL (Haley and Knott 1992), and other artifacts that are the result of either experimental limitations or the limitations of composite interval mapping to locate eQTLs. After deciding on an exclusionary window of 20 cM to select and count eQTLs, the following protocol was used in turn for each of the 22,794 transcripts whose levels were measured and analyzed as e-traits. First, all potential eQTLs were sorted by their LRT statistics. The eQTL with the largest LRT statistic was selected,
and all other potential eQTLs within the exclusionary window (i.e., within 10 cM on either side of the chosen eQTL) were disregarded. The next eQTL was considered and the process iterated until all eQTL for the particular e-trait were exhausted.

Our observation that a 5 cM exclusionary window (Eqtl module of QTL-Cartographer) is most likely affected by additional linked eQTLs is supported by the realization that increasing the window size to 20 cM shifted the distribution to more transcripts/e-trait with fewer eQTLs (1 to 3) (Supplemental Fig. 1 online). To investigate this phenomena further, we examined four genes that were map-based cloned and shown to be controlled by single cis-eQTLs in Arabidopsis (KLIEBENSTEIN et al. 2001; LAMBRIX et al. 2001). These four genes showed large LRT statistics in cis- to the physical/genetic location of each gene, as expected (Supplemental Fig. 2 online). When a 5 cM exclusionary window was used for identifying eQTLs, 3 to 5 eQTLs were incorrectly associated with each of the four cloned loci instead of a single cis-eQTL for each locus. Increasing the window size to 20 cM eliminated these extra eQTL peaks in each test case. Consequently, we used a conservative 20 cM exclusionary window for the global eQTL analysis reported here.

**Estimation of Heritability:** We calculated estimates of broad-sense heritability (H) for each e-trait as $H = \frac{\sigma^2_g}{\sigma^2_p}$, where $\sigma^2_g$ is the estimated e-trait (transcript level) genetic variance among different genotypes in this sample of 211 RILs, and $\sigma^2_p$ is the estimated phenotypic variance for an e-trait (Falconer and Mackay 2003). Heritability was estimated for all 22,746 non-control e-trait in two microarray data sets independently. The first dataset included two GeneChips (biological replicates) for each of the 211 RILs, permitting estimation of the heritability for each e-trait within the RIL population. The estimates of $\sigma^2_R$ (variance due to replicate) and $\sigma^2_e$
(experimental error variance) were determined for each e-trait in this dataset. The second dataset included eight GeneChips per parent (Bay-0 and Sha), representing 16 different biological replicates, four replicates of each parent grown concurrently with the two RIL replications. The parental data allowed for the estimation of the broad-sense heritability for the levels of each transcript, which were then compared with the estimated heritabilities of e-trails in the 211 RILs.

**Cis- versus Trans-eQTL Identification:** The SFP markers (West et al. 2006), which are derived from genes with known physical positions, were used to anchor the genetic map to the genome. The CIM eQTL results for each transcript were examined to investigate if there was an eQTL within 3.5 cM of each gene’s physical location. Expanding this exclusionary window to 5 cM had minimal effect on the numbers of cis-eQTLs identified (data not shown; further discussion below). To test the impact of cis-eQTLs on the relationship between the parental and RIL heritability estimates, we limited this analysis to only those 4,933 transcripts represented by a unique probe set to ensure that the probe set was specific to only one genomic location.

**Transgressive Segregation and eQTLs of Opposite Allelic Effect:** An ad hoc approach was devised to test for transgressive segregation (Brem and Kruglyak 2005). For each e-trait, we counted the number of RILs whose average transcript level accumulation was one pooled standard deviation above or below the minimum or maximum parental mean value. The Bay-0 and Sha accessions may be fixed for groups of alleles with opposite effects at two eQTLs (e.g., Bay-0 as +−/+− and Sha as −+/−+) which would either cancel each other out, or generate progeny with a greater phenotypic distribution than the parents due to the segregation of recombinant individuals with ++/++ and −−/−− genotypes for the two eQTLs (Lynch and Walsh 1998). For the 10,084 e-traits with two or more significant eQTLs, the additive effect of
each eQTL was queried to determine if the transcript was associated with eQTLs having opposite allelic effects. An e-trait was defined as exhibiting opposite allelic effect if it had at least one eQTL with a positive Bay-0 allelic effect and at least one eQTL with a positive Sha allelic effect. If all of the eQTLs for an e-trait had only positive Bay-0 or only positive Sha allelic effects, the transcript was classified as not showing transgressive segregation.

**eQTL Hotspot Significance Threshold**: To determine whether a genetic location associated with multiple eQTLs was a significant cluster or ‘hotspot’, we estimated a significance threshold using permutation. The positions of the 36,871 eQTLs at the marker intervals were permuted across the genome 1,000 times, and the maximal number of eQTLs per genetic position per permutation was obtained. Using the distribution of the maximum number of eQTLs, the criterion for declaration of a significant eQTL hotspot was 133 eQTLs per genetic position at alpha= 0.05.

**Gene Ontology (GO) Analysis of trans-eQTL hotspots**: The GO analysis was performed using GOslim terms (a high-level GO term for functional categorization). Analysis was performed using the GO annotation (BERARDINI et al. 2004) from the TAIR web site (download 20051119; [http://www.arabidopsis.org/](http://www.arabidopsis.org/)). The frequency of each classification was obtained for the completely sequenced genome and the gene lists for each of the 17 trans-eQTL hotspots (see Results). Each gene list was tested for significant deviation ($P \leq 0.001$) from the expected frequencies for the complete genome using $\chi^2$ analysis (CHEN et al. 2005).
Accession numbers: The microarray dataset used in this study have been deposited at EBI ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) under IDs xxxx.

RESULTS

Numbers of eQTLs detected: The majority of genes within the Arabidopsis genome exhibited heritable transcriptional variation that is controlled by eQTLs in a Bayreuth (Bay-0) × Shahdara (Sha) RIL population (Table 1). Of the 22,746 Arabidopsis transcripts represented on the Affymetrix ATH1 microarray and whose levels were measured as e-traits in the 211 RILs, 69% were associated with one or more significant eQTLs (likelihood ratio test [LRT] statistics > 12.0583, the 95% Global Permutation Threshold; see Methods). A total of 36,871 distinct eQTLs were detected and counted using the Eqtl module of QTL-Cartographer and an exclusionary window size of 20 cM (for details, see Methods). The number of eQTLs detected per transcript/e-trait varied from 0 to 11. The transcripts/e-traits associated with eQTLs had an average of 2.34 and a median of 2 eQTLs per transcript. Additionally, eQTLs were identified for non-nuclear genes: 69% of the 155 organelle-encoded transcripts identified nuclear-encoded eQTLs (Table 1).

eQTL genomic distribution: One third of the e-traits with detectable eQTLs had a cis-eQTL (i.e., coincident with the physical position of the gene whose transcript level varied) (Table 1). This proportion was strikingly uniform across the five Arabidopsis chromosomes, as evidenced by the diagonal of significant LRT statistics across each chromosome (Fig. 1A). Furthermore, there were approximately equal contributions of parental alleles that increased transcript abundance (Fig. 1A).
The remaining (trans-) eQTLs were non-uniformly distributed across the chromosomes, with the majority of eQTLs localized on chromosome II (Fig. 1, Table 1). eQTLs were clustered into trans-eQTL ‘hotspots’ that are visible as vertical bands of eQTLs (Fig. 1A) and as peaks of transcript numbers with common eQTL map positions (Fig. 1B). Seventeen significant trans-eQTL hotspots were detected across the genome, with a minimum of one per chromosome. Ten of these 17 hotspots contained eQTLs associated with more than 200 different transcripts; the largest hotspot was associated with 2,528 transcripts (Fig. 1B; Table 2). Chromosome II had the greatest number of hotspots (five), as well as the hotspots associated with the greatest number of transcripts (Fig. 1; Table 2). A Gene Ontology analysis of the transcripts associated with the 17 trans-eQTL hotspots did not reveal any significant over- or under-representation of any functional or biological process GO categories (data not shown). In addition, our previous network eQTL analysis in this RIL population using eighteen a priori-defined gene networks indicated that the chromosome II hotspots did not control the majority of these networks (Kliebenstein et al. 2006).

A strong directional bias was evident for each of the trans-eQTL hotspots (Fig. 1A; Table 2), such that the same parental allele up-regulated most of the transcripts associated with a hotspot. At eight trans-eQTL hotspots, the Sha allele increased transcript levels (positive effect); the other nine hotspots were associated with the opposite allelic effect (i.e., the Sha allele decreased accumulation). A global average of 54.7% of the 36,871 eQTLs had a positive Sha allelic effect and 45.3% had a positive Bay-0 allelic effect. In contrast, the effects of each trans-eQTL hotspot significantly deviated from these global averages: for 15 of the 17 hotspots, greater than 80% of the transcripts associated with each trans-eQTL hotspot were affected positively by the same parental allele (Table 2).
**Proportion of phenotypic variation per eQTL:** Most individual eQTLs accounted for only a small proportion of the associated transcript/e-trait’s estimated phenotypic variation ($R^2$): 89% of eQTLs accounted for less than 20% of the $R^2$ for each transcript (Fig. 2). The *cis*-eQTLs typically controlled more of an e-trait’s phenotypic variation than did the *trans*-eQTLs (Fig. 2, inset). This difference in the $R^2$ distributions between *cis*- and *trans*-eQTLs was not due to the presence of single feature polymorphisms (SFPs) because > 91% of all genes with a *cis*-eQTL did not have a SFP (Supplemental Table 3 online). Furthermore, the genes with a SFP and a *cis*-eQTL had an $R^2$ distribution similar to genes with a *cis*-eQTL that lacked a SFP (data not shown).

**eQTL allelic effects and transcript heritabilities:** The opposing directionality of allelic effects in the *trans*-eQTL hotspots suggests the potential for non-additive genetic variation in transcript levels between Bay-0 and Sha. The presence of non-additive variation is also supported by the observation that while the parental microarray data showed a significant difference in transcript levels for only 3,351 of the 22,746 transcripts (false discovery rate 0.05) (data not shown), the RIL microarray data allowed detection of eQTLs for 69% of the e-traits, indicating that the levels of those transcripts are most likely genetically controlled. To further test for non-additive genetic variation in this population, we estimated and compared the broad sense heritability ($H$) of 22,746 transcripts/e-traits with data from parental microarrays and from replicated microarrays for the 211 RILs. The vast majority of the e-traits showed greater heritability in the RILs than in the parents (Fig. 3a), suggesting non-additive genetic variation in the RILs.
Therefore, for the majority of e-traits, the observed variation of transcript levels in the parents was not predictive of the variation present in their progeny.

**Transgressive segregation:** One potential source of non-additive genetic variation is transgressive segregation whereby the progeny have average e-trait values that exceed the range of the parent values. Comparing transcript accumulation for 22,746 transcripts between the parents and RILs identified 14,258 transcripts for which at least 10% of the RILs were one standard deviation beyond the parental range (data not shown). Since transgressive segregation can be due to the presence of eQTLs with opposing parental allelic effects, our investigation revealed that of the 10,084 e-trait associated with two or more eQTLs, 6,911 exhibited eQTLs with opposing additive effects, suggesting that such eQTLs are a significant component of the non-additive genetic variation within the 211 Bay-0 × Sha RILs. Of these 6,911 e-trait, 4,871 show evidence of transgressive segregation as described above, indicating that the use of one standard deviation above and below the parental range is a conservative assessment of non-additive genetic variation. Our statistical analyses and these summaries do not, however, indicate whether epistasis is a component of the non-additive genetic variation that is prevalent in this population. Due to the complexity of the required statistical analyses involving both a large number of tests and a large number of eQTLs, epistasis will be investigated in a subsequent effort.

**Transcript heritabilities in RILs versus parents:** While the majority of e-trait estimated heritabilities showed little correlation between the parents and the RILs, some did exhibit a positive relationship (Fig. 3b, 3c). We found that transcripts/e-trait associated with a *cis*-eQTL
are most likely to exhibit a linear heritability relationship in the parents and RILs, because \textit{cis}-eQTLs tend to control more of the e-trait’s variation ($R^2$, Fig. 2). In fact, the 5,127 e-traits with a \textit{cis}-eQTL exhibited a stronger linear relationship of heritability between the parents and the RILs (Fig. 3c) than did the 10,644 e-traits with only \textit{trans}-eQTLs (i.e., no \textit{cis}-linkages) (Fig. 3d). Even though $R^2$ values are not additive, if used as a relative measure of phenotypic variation for a specific transcript, the totaled $R^2$ for transcripts with a \textit{cis}-eQTL explains more of the relative phenotypic variation than the corresponding totaled $R^2$ for transcripts with only \textit{trans}-eQTLs (Supplemental Fig. 3 online). A small proportion of e-trait had high heritability ($H > 0.7$) in the RILs, but no significant eQTLs, suggesting that these effects are too small to detect in a sample of 211 individuals.

**DISCUSSION**

The genetic architecture of transcript level variation in the \textit{Arabidopsis thaliana} Bay-0 × Sha RIL population was surprisingly highly variable and complex. Single eQTLs were detected for a large proportion (69%) of the 22,746 transcripts/e-trait (Table 1). Many e-trait were also associated with multiple eQTLs. We realize that this experiment and these data represent only a single sample of the life history variation for this species; however, the expression levels of the majority of transcripts showed quantitative variation. These transcript levels are probably quantitatively controlled, given the fact that this RIL population, derived from two inbred accessions, represents only a fraction of the variation present in this diverse species (NORDBORG \textit{et al.} 2005; KLIEBENSTEIN \textit{et al.} 2006). The complexity of this control also implies that transcriptional regulation in other higher plant species is likely to be as multifaceted. Our findings in Arabidopsis, in conjunction with studies in yeast (YVERT \textit{et al.} 2003; BREM and KRUGLYAK 2005), mice (SCHADT \textit{et al.} 2003; CHESLER \textit{et al.} 2005) and humans (MORLEY \textit{et al.}...
2004), support the working hypothesis that global complexity of transcriptional regulation may be a general feature of eukaryotes.

The transcript levels of most genes in this Arabidopsis RIL population were controlled by multiple eQTLs with small $R^2$ values (Fig. 2). The RIL sample size, the observable number of recombinants, and the statistical analysis that we employed can all be considered limiting factors of this study that affect our ability to detect eQTLs of small phenotypic effect (Beavis 1998; Mackay 2001; Doerge 2002; Kim et al. 2005). The genetic variance associated with an eQTL (i.e., the proportion of the phenotypic variance for which it accounts) is larger in a RIL population than in F$_2$ or backcross populations (Beavis 1998; Flint et al. 2005). Therefore, our use of a RIL mapping population structure in conjunction with 211 biologically replicated RILs greatly facilitated the detection of smaller effect eQTLs due to the relatively large number of observed recombination events that, in turn, allowed more precise estimation of phenotypic variances. Nevertheless, it is likely that additional small effect eQTLs remain undetected in this population. The observed skewed (e.g., long right-hand tail) $R^2$ histogram (Fig. 2) suggests that an increase in population size ($n > 211$) and additional biological replication per line would most likely increase detection of eQTLs with $R^2$ values $< 0.1$ (Kim et al. 2005).

This study revealed that small effect eQTLs were prevalent in the 211 RILs, with less than 6% of the 36,871 eQTLs accounting for a large proportion of the estimated phenotypic variation ($R^2 > 0.3$). These results contrast to studies that were based on smaller population sizes in which large phenotypic effect $cis$-eQTLs were prevalent (Schadt et al. 2003; Decook et al. 2006). Studies based on small sizes in combination with population structures that do not maximize recombination events reduces QTL resolution and lead to sampling bias when
detecting eQTLs/QTLs of larger phenotypic effect (Beavis 1998; Bogdan and Doerge 2005; de Koning and Haley 2005; Gibson and Weir 2005).

Our observation that cis- and trans-eQTLs are associated with different distributions of $R^2$ (Fig. 2) was also found in Drosophila (Meiklejohn et al. 2003; Wayne et al. 2004; Hughes et al. 2006). A potential explanation for this observation is that the transcript abundance of most genes is regulated by multiple signal transduction networks at multiple levels (e.g., transcription and transcript stability), thus a polymorphism in any one of these regulatory levels may be functionally limited to only a small change in transcript accumulation for those genes controlled in trans by that polymorphism. However, induced mutations in transcription factors can generate large expression differences in genes regulated in trans by the transcription factor. Alternatively, the polymorphism underlying a trans-eQTL hotspot has greater potential to be pleiotropic than a cis-eQTL with no corresponding hotspot. Large-effect mutations in pleiotropic genes are more likely to be deleterious than mutations in less interconnected genes (Wright 1968; Turelli 1988; Wagner 2000; Jeong et al. 2001; Fraser et al. 2002; Yu et al. 2004). There may be an evolutionary fitness limitation on the potential genetic effect of polymorphisms that generate trans-eQTL hotspots. Additional populations and species need to be evaluated to determine if cis-eQTL effects are larger than trans-eQTL effects in general.

The majority of transcripts/e-traits exhibited higher estimated broad-sense heritabilities (H) within the RILs than in the parents (Fig. 3), indicating that the transcript level variation present in these two homozygous inbred Arabidopsis accessions does not accurately predict transcript variation in their RIL progeny. Furthermore, our results indicate that if parental transcript variation is used to select a restricted set of genes for subsequent study in the progeny, many genes would be overlooked. The difference in estimated transcript level heritabilities for
parents and RILs also suggests that the majority of the genetic variance influencing the levels of most transcripts is non-additive (i.e., includes dominance and epistatic interaction components; Lynch & Walsh 1998). Indeed, approximately 70% of the transcripts with two or more eQTLs showed opposite allelic effects. Since the parents and RILs were grown together in the same environment, it is unlikely that genotype × environment interactions were a major source of differences in phenotypic variance estimates between parents and RILs, supporting the conclusion that non-additive genetic variance was the main contributor to differences in e-trait H between parents and RILs. Our results are consistent with a yeast global transcriptome study for which transgressive segregation and non-additive genetic variation were also prevalent (BREM and KRUGLYAK 2005). The predominance of non-additive genetic variation in our study also suggests that surveys of transcript level variation in parental accessions, at least within Arabidopsis, will significantly underestimate the actual transcript level variation in progeny populations derived from intercrossing these accessions. However, the underestimation will primarily affect trans-eQTLs as the cis-eQTLs typically exhibit larger $R^2$ values (Fig. 2) and a positive correlation of transcript level heritability between the parents and their RILs (Fig. 3c). This latter observation supports our previous suggestion that correlation between transcript level variation and sequence polymorphism within seven Arabidopsis accessions was due to cis control of the majority of the detected transcript level variation (KLIBENSTEIN et al. 2006). Interestingly, a small number of transcripts (41) exhibited high heritability, but no detectable eQTLs. This could be due to control by multiple loci with small additive effects or control by loci that interact non-additively (BREM and KRUGLYAK 2005). Neither of these issues is investigated here, and serves as motivation for further work in this area.
While cis-eQTLs accounted for a greater proportion of the phenotypic variance, the majority of eQTLs within this RIL population are situated in trans to the genes whose transcript levels varied. Furthermore, most trans-eQTLs clustered into regions defined as trans-eQTL “hotspots” that control the levels of a larger number of transcripts but only a small fraction of the variation for each transcript. Trans-eQTL hotspots have been reported in other organisms (BREM et al. 2002; SCHADT et al. 2003). Interestingly, in our study the e-traits affected by a trans-eQTL hotspot also showed directionality such that most transcripts were either up- or down-regulated by the same parental allele at the trans-eQTLs (Fig. 1A, Table 2). However, examination of their genomic positions and GO annotation analysis did not identify any functional categories significantly associated with any of the 17 trans-eQTL hotspots (data not shown), thus the biological function of the trans-eQTL hotspots remains unknown. Lastly, our comparison of the trans-eQTL hotspots to known phenotypic QTL locations did not identify any significant enrichment of phenotypic QTLs in these regions (KLIEBENSTEIN et al. 2006) (data not shown).

Clearly, further genetic dissection of each trans-eQTL hotspot is required to identify the genes that influence so many other genes in trans and to understand the biological function of such hotspots. One approach to finding candidate causal genes is to identify cis-eQTLs underlying a trans-eQTL hotspot; our analysis identified ~100 cis-eQTLs per hotspot. However, due to the potential genetic complexity of e-QTL regions it may not be a straightforward task to identify the specific causal gene or genes. QTLs and e-QTLs detected as single genetic effects may be composed of multiple, physically linked loci, each associated with a small portion of the phenotypic variation (FLINT et al. 2005). Understanding how closely linked loci of small effect interact to control a quantitative phenotype would require a very large mapping population (i.e.,
many recombinants) and appropriate follow-up studies in order to separate and isolate each of the QTLs involved (Kim et al. 2005).

An important question that remains to be addressed is how transcript level variation is related to and controls downstream phenotypic variation (Gibson and Weir 2005). Studies in plants have correlated specific phenotypic QTLs with transcript level variation controlled in cis (Kliebenstein et al. 2001; Lambrix et al. 2001; Zhang et al. 2006). Several of these studies identified cis-regulated transcript level variation in signal transduction genes controlling flowering time or circadian rhythm variation (Johanson et al. 2000; Caicedo et al. 2004; Werner et al. 2005). This cis variation in a signal transduction component potentially controls small changes in transcript accumulation for numerous genes in trans. As mentioned previously, we hypothesize that cis-eQTL variation may underlie some trans-eQTL hotspots whereby the genes affected in trans are the mechanism through which the cis-affected gene influences the downstream phenotype. Testing this hypothesis will require high-throughput analysis of a RIL population for both transcript level and downstream phenotypic variation, then subsequent functional analysis of the causal genes.

A fundamental issue in quantitative genetics is how the genotype determines the quantitative trait phenotype (Mackay 2001). A study of transcriptional variation is merely one component of this question. Determining the actual biological relationship between transcript level variation and other phenotypes, such as protein and metabolite levels, that in turn influence trait phenotypes further downstream, is a highly complex task (Jansen et al. 2002). For example, post-transcriptional regulation can diminish the correlation between transcript levels and protein levels. Furthermore, inaccuracy in measurements of global transcript accumulation can reduce the ability to detect the relationship between mRNA levels and downstream phenotypes.
Reducing experimental error via biological replication of large segregating populations representing many recombination events (JANSEN and NAP 2001) can improve one’s ability to detect meaningful biological associations of transcriptomic variation with proteomic and metabolomic variation. The Bay-0 × Sha RIL population, consisting of hundreds of genetically stable inbred lines and the global eQTL analysis described here provide a basis from which to study how transcript level variation may influence downstream phenotypic trait variation in higher plants. Such studies will enable determination of how the genotype determines quantitative trait phenotypes and contribute towards a deeper understanding of the genotype-phenotype relationship in eukaryotes.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation 2010 Project, grant # MCB-0115109 to DAS, RWM, and RWD. We thank Rebecca Walker and Tanya Tang for technical assistance and Steve Edberg for assistance with data management and dissemination.
LITERATURE CITED


Table 1. Global eQTL Description.

The Affymetrix ATH1 microarray detects transcripts for genes on all five Arabidopsis chromosomes (I-V) as well as additional non-nuclear encoded genes. The control probe sets are not included in this table.

<table>
<thead>
<tr>
<th>Chrom.</th>
<th>Genes(^a)</th>
<th>Genes with an eQTL(^b)</th>
<th>Genes with cis eQTL(^c)</th>
<th>Genes with an eQTL on this chromosome(^d)</th>
<th>Number of eQTLs(^e)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>%</td>
<td>Total</td>
<td>%</td>
<td>Total</td>
</tr>
<tr>
<td>I</td>
<td>5862</td>
<td>4114 70%</td>
<td>1333 32%</td>
<td>5101 32%</td>
<td>6404 17%</td>
</tr>
<tr>
<td>II</td>
<td>3752</td>
<td>2473 66%</td>
<td>794 32%</td>
<td>11174 71%</td>
<td>16699 45%</td>
</tr>
<tr>
<td>III</td>
<td>4490</td>
<td>3135 70%</td>
<td>1024 33%</td>
<td>3096 20%</td>
<td>3664 10%</td>
</tr>
<tr>
<td>IV</td>
<td>3379</td>
<td>2308 68%</td>
<td>787 34%</td>
<td>3004 19%</td>
<td>3591 10%</td>
</tr>
<tr>
<td>V</td>
<td>5108</td>
<td>3634 71%</td>
<td>1189 33%</td>
<td>4988 32%</td>
<td>6513 18%</td>
</tr>
<tr>
<td>Non-Nuclear</td>
<td>155 107 69%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Total</td>
<td>22794</td>
<td>15771 69%</td>
<td>5127 32%</td>
<td>15771 100%</td>
<td>36871 100%</td>
</tr>
</tbody>
</table>

\(^a\) Number of Genes indicates the number of genes per chromosome or group whose transcripts are detected on the Affymetrix ATH1 microarray. Total is the sum for the entire genome.

\(^b\) Genes from this chromosome with eQTL lists the number of genes from each chromosome or group whose transcript accumulation was controlled by at least one eQTL; genes with eQTL (%)
shows the percentage of genes from each chromosome whose transcript accumulation was controlled by at least one eQTL.

c Genes with \textit{cis}-eQTL shows the number of genes per chromosome whose transcripts had a significant eQTL in the \textit{cis} position, defined as within 3.5 cM of the gene’s genetic position; percent of genes with a \textit{cis}-eQTL lists the percent of genes determined by only those genes with at least one detectable eQTL.

d Genes with an eQTL on this chromosome lists the number of genes within the genome whose transcripts were significantly controlled by at least one eQTL on this specific chromosome.

e Number of eQTL detected on this chromosome indicates the number of significant eQTLs detected on this specific chromosome.
Table 2. Assessment of directional bias of allelic effects at trans-eQTL hotspots.

All eQTLs with their LRT statistic maxima mapping to a marker at a trans-eQTL hotspot were analyzed for additive effect estimates contributed by each parental allele. Permutation analysis established 133 eQTLs as the significance threshold (alpha = 0.05) for declaring a significant trans-eQTL hotspot (see Methods). Within each hotspot, the number of eQTLs with a negative additive effect estimate (i.e., Bay-0 allele increased transcript level) was compared with the number of eQTLs with a positive additive effect estimate (i.e., Sha allele increased transcript level).

<table>
<thead>
<tr>
<th>Marker</th>
<th>Chromosome</th>
<th>position (cM)</th>
<th># eQTL</th>
<th>negative</th>
<th>positive</th>
<th>% Bay</th>
<th>% Sha</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.01</td>
<td>164</td>
<td>142</td>
<td>22</td>
<td>86.6%</td>
<td>13.4%</td>
<td>2.37 x 10^{-26}</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>61.51</td>
<td>353</td>
<td>16</td>
<td>337</td>
<td>4.5%</td>
<td>95.5%</td>
<td>2.00 x 10^{-53}</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>99.88</td>
<td>193</td>
<td>51</td>
<td>142</td>
<td>26.4%</td>
<td>73.6%</td>
<td>1.38 x 10^{-7}</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>101.88</td>
<td>186</td>
<td>175</td>
<td>11</td>
<td>94.1%</td>
<td>5.9%</td>
<td>9.52 x 10^{-41}</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>0.01</td>
<td>1202</td>
<td>1028</td>
<td>174</td>
<td>85.5%</td>
<td>14.5%</td>
<td>1.06 x 10^{-172}</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>13.59</td>
<td>1439</td>
<td>1341</td>
<td>98</td>
<td>93.2%</td>
<td>6.8%</td>
<td>1.35 x 10^{-291}</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>22.97</td>
<td>182</td>
<td>164</td>
<td>18</td>
<td>90.1%</td>
<td>9.9%</td>
<td>6.16 x 10^{-34}</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>42.44</td>
<td>2528</td>
<td>5</td>
<td>2523</td>
<td>0.2%</td>
<td>99.8%</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>55.87</td>
<td>710</td>
<td>10</td>
<td>700</td>
<td>1.4%</td>
<td>98.6%</td>
<td>4.64 x 10^{-122}</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>9.14</td>
<td>381</td>
<td>367</td>
<td>14</td>
<td>96.3%</td>
<td>3.7%</td>
<td>4.68 x 10^{-89}</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>72.38</td>
<td>162</td>
<td>26</td>
<td>136</td>
<td>16.0%</td>
<td>84.0%</td>
<td>7.48 x 10^{-14}</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>77.48</td>
<td>163</td>
<td>30</td>
<td>133</td>
<td>18.4%</td>
<td>81.6%</td>
<td>5.27 x 10^{-12}</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>55.40</td>
<td>256</td>
<td>255</td>
<td>1</td>
<td>99.6%</td>
<td>0.4%</td>
<td>3.08 x 10^{-68}</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>63.54</td>
<td>195</td>
<td>123</td>
<td>72</td>
<td>63.1%</td>
<td>36.9%</td>
<td>6.14 x 10^{-7}</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>66.48</td>
<td>261</td>
<td>245</td>
<td>16</td>
<td>93.9%</td>
<td>6.1%</td>
<td>5.58 x 10^{-56}</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>70.70</td>
<td>417</td>
<td>69</td>
<td>348</td>
<td>16.5%</td>
<td>83.5%</td>
<td>$4.12 \times 10^{-32}$</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>75.51</td>
<td>202</td>
<td>37</td>
<td>165</td>
<td>18.3%</td>
<td>81.7%</td>
<td>$1.32 \times 10^{-14}$</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Number of eQTLs with LRT statistic maxima located at the marker.

\(^b\) Number of eQTLs with a negative additive effect estimate.

\(^c\) Number of eQTLs with a positive additive effect estimate.

\(^d\) Percentage of transcripts positively influenced by the Bay-0 allele.

\(^e\) Percentage of transcripts positively influenced by the Sha allele.

\(^f\) \(P\) values for \(\chi^2\) test for significant deviation from expected 50% contribution from each allele.
FIGURE LEGENDS

**Figure 1.** Genomic architecture of eQTLs across five Arabidopsis chromosomes.

**A.** Heat map of likelihood ratio test (LRT) statistics obtained by composite interval mapping (CIM) eQTL analysis for 22,591 nuclear-encoded transcripts (y-axis) plotted against 464 markers (x-axis) across five chromosomes. Colors indicate chromosomal regions where LRT statistics were significantly greater than the global permutation threshold (GPT > 12.0583) at \( P < 0.05 \). Red indicates a positive effect of the presence of the Sha allele, and green indicates a positive effect of the Bay-0 allele. Vertical dotted lines separate the five chromosomes (I–V).

**B.** Numbers of transcripts/e-traits for which eQTLs are detected. Number of transcripts is indicated on the y-axis, plotted against the genetic location of the eQTLs in cM on the x-axis. The permuted threshold \( (P = 0.05) \) for detection of a significant trans-eQTL hotspot is 133 transcripts, indicated by the red horizontal line.

**Figure 2.** Distribution of percent phenotypic effect (\( R^2 \)) for all eQTLs.

A histogram of the distribution of \( R^2 \) values for all 36,871 eQTLs are shown separated into 0.01 bins; the maximum \( R^2 \) was 0.97. The two inset pie graphs illustrate the \( R^2 \) distributions for eQTLs that are *cis* (5127 total *cis*-eQTL) or *trans* (31,777 total *trans*-eQTL) to the gene’s physical position. The color scale to the right indicates the \( R^2 \) bins for the pie graphs.

**Figure 3.** Transcript level/e-trait heritabilities in RILs versus parents.

**A.** Histograms of estimated broad sense heritability (H) values in RILs and parents (Bay-0 and Sha). Black (filled) bars show the histogram of heritability values for all 22,746 transcripts/e-
traits as estimated with the RIL microarray data. White (open) bars show the histogram of H for all 22,746 transcripts/e-traits as estimated with the parental (Bay-0 and Sha) microarray data.

**B – D.** Relationship between estimated transcript/e-trait H in the RILs and in the parents. H values for all 22,746 transcripts are plotted as a hexbin graph for both the RIL (x-axis) and parental (y-axis) H estimates. Gray scale to the right of each figure indicates data point density per bin. The x- and y- axes for panels B – D are identical.

**B.** All transcripts. **C.** Transcripts for which a *cis*-eQTL was mapped. **D.** Transcripts that mapped to only *trans*-eQTLs (no *cis*).