

Genomic Survey of Gene Expression Diversity in *Arabidopsis thaliana*

Daniel J. Kliebenstein,^{*,1} Marilyn A. L. West,^{*,1} Hans van Leeuwen,^{*} Kyunga Kim,[†]
R. W. Doerge,[†] Richard W. Michelmore^{*,‡} and Dina A. St. Clair^{*,2}

^{*}Department of Plant Sciences, University of California, Davis, California 95616-8780, [†]Department of Statistics, Purdue University, West Lafayette, Indiana 47907-2067 and [‡]The Genome Center, University of California, Davis, California 95616

Manuscript received August 12, 2005
Accepted for publication September 27, 2005

ABSTRACT

Differential gene expression controls variation in numerous plant traits, such as flowering time and plant/pest interactions, but little is known about the genomic distribution of the determinants of transcript levels and their associated variation. Affymetrix ATH1 GeneChip microarrays representing 22,810 genes were used to survey the transcriptome of seven *Arabidopsis thaliana* accessions in the presence and absence of exogenously applied salicylic acid (SA). These accessions encompassed ~80% of the moderate- to high-frequency nucleotide polymorphisms in *Arabidopsis*. A factorial design, consisting of three biological replicates per accession for the two treatments at three time points (4, 28, and 52 hr post-treatment), and a total of 126 microarrays were used. Between any pair of *Arabidopsis* accessions, we detected on average 2234 genes (ranging from 1428 to 3334) that were significantly differentially expressed under the conditions of this experiment, using a split-plot analysis of variance. Upward of 6433 genes were differentially expressed between at least one pair of accessions. These results suggest that analysis of additional genetic, developmental, and environmental conditions may show that a significant fraction of the *Arabidopsis* genome is differentially expressed. Examination of sequence diversity demonstrated a significant positive association with diversity in gene expression.

NATURAL phenotypic variation is often quantitatively distributed and occurs in both simple and complex traits. DNA sequence polymorphisms can be a primary genetic cause of phenotypic variation in populations. Some recent studies have assessed genomewide DNA sequence variation within and between different species (GIBBS *et al.* 2003; SCHMID *et al.* 2003, 2005; STEIN *et al.* 2003; BENTLEY and PARKHILL 2004; NORDBORG *et al.*, 2005). These studies identified variation throughout the genome, including single-nucleotide and INDEL polymorphisms in all gene components: promoter, UTRs, ORFs, and introns. However, cataloging of sequence variation is only the first step in understanding how DNA variation contributes to phenotypic variation. Recent advances in technology have enabled comparative studies that focus on the variation present in the transcriptome, metabolome, and proteome. Recent work comparing sequence divergence to expression divergence between two *Drosophila* species found little evidence for a correlation between the two measures of divergence (NUZHIDIN *et al.* 2004). Understanding how variation in these intermediate processes is correlated will aid in linking DNA sequence polymorphism to its phenotypic consequences.

Differences in gene expression, termed expression level polymorphisms (ELPs) (DOERGE 2002), have been shown to control inter- and intraspecific phenotypic variation in various organisms (WANG *et al.* 1999; CARROL 2000; BREM *et al.* 2002). In plants, naturally occurring ELPs have been associated with changes during maize domestication (WANG *et al.* 1999), flowering-time control in *Arabidopsis* (JOHANSON *et al.* 2000; CAICEDO *et al.* 2004; LEMPE *et al.* 2005; WERNER *et al.* 2005), pathogen resistance (GRANT *et al.* 1995; GASSMANN *et al.* 1999), and *Arabidopsis* insect resistance and secondary metabolism (KLIEBENSTEIN *et al.* 2001, 2002; LAMBRIX *et al.* 2001). ELPs can be caused by various types of DNA sequence polymorphisms, including *trans*-acting factors (CAICEDO *et al.* 2004), *cis*-acting promoter polymorphisms (FRARY *et al.* 2000; CONG *et al.* 2002), *cis*-acting processing variants, such as splice-site polymorphisms (LAMBRIX *et al.* 2001), and whole or partial gene deletions that functionally act to nullify the gene's expression (KLIEBENSTEIN *et al.* 2001; LAMBRIX *et al.* 2001). In contrast, the more drastic mRNA processing and INDEL polymorphisms may be overlooked even though they can be the basis of variation in plant/environment interactions (LAMBRIX *et al.* 2001; KLIEBENSTEIN *et al.* 2002). To fully study the sequence level polymorphisms that alter gene expression patterns, we include the full range of polymorphism types in our definition of ELPs.

In spite of the potential importance of ELPs in controlling phenotypic variation, little is known about

¹These authors contributed equally to this work.

²Corresponding author: Department of Plant Sciences, University of California, Mail Stop 3, 1 Shields Ave., Davis, CA 95616-8780.
E-mail: dastclair@ucdavis.edu

the extent of ELPs within a species or the distribution of ELPs within the genome. ELP variation within a species should be described at several levels. First, which genes are differentially expressed among accessions or genotypes? Such genes may be under selective pressure for altered expression patterns or the expression difference may be the result of neutral sequence variation. Second, how does variation in gene expression relate to the underlying nucleotide polymorphism level? Comparing the nucleotide polymorphism level between individuals to their ELP divergence can help assess if these two types of genetic diversity are related. By comparing the pattern of nucleotide polymorphism to ELP variation, it may be possible to determine if nucleotide polymorphisms and expression level polymorphism show similar variation patterns, suggesting comparable levels of neutrality. Third, how does gene expression variation differ in response to changes in the environment and/or treatments? These questions, as well as many others, can be addressed by quantitative analyses of natural variation in gene expression within a species.

In this study, we assay natural variation in *Arabidopsis* gene expression on a genome-wide basis in terms of which genes exhibit ELPs, how frequently these genes vary in their expression patterns between pairs of accessions, and the consequences of varying a treatment on ELP detection. The Affymetrix ATH1 GeneChip microarray representing 22,810 genes was employed to query gene expression in seven *Arabidopsis thaliana* accessions. These accessions represent ~80% of the moderate-to high-frequency polymorphisms within *Arabidopsis* (NORDBORG *et al.* 2005). All statistically significant ELPs in 5-week-old foliar tissue were used to obtain an overview of species-wide variation. Our analysis revealed between 1428 and 3344 differentially expressed genes for any pair of the *Arabidopsis* accessions surveyed. The genes exhibiting differential expression between at least one pair of these seven accessions composed more than one-third of the *Arabidopsis* genome. The frequency of differential gene expression between accessions was significantly positively correlated with the genomewide sequence divergence among the same accessions. Furthermore, this sequence/expression divergence correlation was also evident in terms of chromosome organization/physical localization. This study provides a first description of a species-level sampling of gene expression variation within *Arabidopsis*, as well as an assessment of how ELP variation relates to nucleotide polymorphism levels.

MATERIALS AND METHODS

Plant material and experimental conditions: Seeds of seven *Arabidopsis thaliana* accessions (Col-0, Cvi-1, Est, Kin-0, Mt-0, Tsu-1, and Van-0) were provided by J. Chory, Salk Institute, San Diego. These accessions can be obtained from The European Arabidopsis Stock Centre (<http://arabidopsis.info>) and The Arabidopsis Information Resource (TAIR) (www.arabidopsis.org).

To identify a level of exogenous salicylic acid (SA) application that caused gene induction without phytotoxicity in all accessions, a preliminary experiment was conducted. Multiple plants of each accession were grown in a growth chamber under short-day conditions (8 hr light at 100–120 μ Ei, 20° day/20° night) to maintain plants in the vegetative phase for 5–6 weeks. The accessions were treated with SA at approximately the same developmental stage (as defined by rosette diameter and leaf number). Plants were selected for treatment when they either reached a rosette diameter of 50–60 mm or had ~15–20 leaves. Subsequently, the plants were sprayed with a series of SA (Sigma-Aldrich, St. Louis) concentrations [0, 0.1, 0.2, 0.3, and 1 mM, all in 0.02% Silwet L77, a surfactant (Lehle Seeds, Round Rock, TX)] to runoff and then harvested at four time points (4, 28, 52, and 76 hr) post-SA treatment and quick-frozen in liquid nitrogen. Three plants per accession were bulked for each concentration/time point to generate sufficient quantities of RNA. RNA was extracted from the bulked plants using TRIzol as per the manufacturer's protocol (Invitrogen, Carlsbad, CA) and applied to slot blots as described (AUSUBEL *et al.* 2004). Genes that had been reported previously as responsive to SA induction were used as probes on the slot blots (*PRI*, *PR2*, *PR5*, *NPRI*, *EDS1*, *EDS5*, and *PAD4*) (GLAZEBOOK 2001), and β -tubulin served as a negative control. DNA sequences were obtained from GenBank (www.ncbi.nlm.nih.gov) and TAIR (www.arabidopsis.org) to design primers for PCR amplification and cloning. *PRI*, the most informative marker gene, indicated that across the seven accessions, plants responded to 0.30 mM SA during the first three time points (4, 28, and 52 hr) without exhibiting phytotoxicity (data not shown). Therefore, two treatments, 0.30 mM SA plus a zero SA control, and sampling at three time points, 4, 28, and 52 hr post-SA application, were chosen to conduct global assays for ELPs among these seven accessions.

A factorial experiment was conducted with the seven accessions, using three biological replications for each of the two treatments and the three time points. Due to space limitations in the growth chambers, these replications were performed on different dates. The accessions were seeded out on four dates (10/8/02, 10/29/02, 12/19/02, and 2/6/03). If plants of an accession did not grow properly in one of the first three planting dates, all samples from that accession for the affected date were discarded. Subsequently, plants from the fourth planting date for that accession were used as a replacement third replicate to maintain a balanced factorial consisting of three biological replications per accession. For each biological replicate, multiple plants (~30) per accession were grown in a growth chamber under short-day conditions (8 hr light at 100–120 μ Ei, 20° day/20° night) to maintain plants in the vegetative phase for 6–7 weeks. Plants with 25–30 rosette leaves (rosette diameter 60–70 mm) were selected and sprayed with 0.30 mM SA in 0.02% Silwet L77 to runoff, and control plants were sprayed with 0.02% Silwet L77 alone. Plants from each combination of accession and treatment were harvested at 4, 28, and 52 hr post-treatment and quick-frozen in liquid nitrogen for subsequent RNA extractions. Each plant was harvested only once. To avoid diurnal variation in gene expression, the time points were 24 hr apart and harvesting was always performed over a 30-min period at the same time of day (1:00 PM).

RNA isolation and microarray analysis of transcriptome: RNA was purified from the 126 samples from the factorial experiment using the TRIzol extraction procedure as above, followed by purification on RNeasy columns (QIAGEN, Valencia, CA). Labeled cRNA was prepared and hybridized to Affymetrix ATH1 GeneChips, containing 22,810 *Arabidopsis* genes, according to the manufacturer's guidelines (Affymetrix, Santa Clara, CA). The GeneChips were scanned with an Affymetrix GeneArray 2500 scanner and data acquired via the Microarray

Suite software MAS 5.0. The MAS 5.0 algorithm with default scaling was used to obtain gene expression levels for analysis.

Quality control: A preliminary analysis was conducted with GeneSpring v. 4.2 software (Agilent Technologies, Palo Alto, CA) for the purpose of selecting diagnostic genes that could be used to identify the accession hybridized to each GeneChip. Thirty-one genes were chosen whose expression level allowed each accession to be uniquely identified, regardless of treatment or time point. Using these 31 diagnostic genes, we fingerprinted all the chips. This procedure detected 11 GeneChips that did not fit the expected expression profile for a given accession, but matched the profile of a different accession, suggesting that an error had occurred during sample processing. These chips had all been run on the same 2 days, suggesting a single procedural error. When the 11 original RNA samples were reanalyzed on replacement GeneChips, the expression profiles fit their expected accession profile. This confirmed that a sample switch had occurred during initial chip processing, so the data from the incorrect GeneChip were removed and replaced with the correct replacement GeneChip.

Statistical analysis: We next proceeded to conduct a statistical analysis with the following goals: to detect differential gene expression among accessions, which may represent potential ELPs; to identify accessions with gene expression changes resulting from SA application; and to account for array (chip-to-chip) variation. A mixed linear model analysis of variance (ANOVA) was used to analyze the GeneChip data. This approach is more powerful than nonparametric methods to detect differentially expressed genes and allows partitioning of the sources of variation in gene expression (*e.g.*, accession, treatment, time points, array, gene, and their various interaction terms) that in turn improves accuracy and enables experimental interpretation (CRAIG *et al.* 2003). The mixed linear model also takes into account the array (chip-to-chip) variation, which enables better estimation of the “true” underlying gene effects, especially when the array variation is large. The mixed model, as employed here, does not require data “normalization” procedures since the components of variation are incorporated in the model. Furthermore, this model permits analysis of expression level data from all genes represented on the GeneChip without relying on Affymetrix software to call the presence or absence of expression.

For each pair of accessions (21 in total), an ANOVA was performed using a split-plot mixed linear model with a random array effect. In this model y_{ijkgr} denotes the expression level of gene g , measured from the accession i under SA treatment j at the time point k for the chip replication r . The ANOVA model for the log-transformed expression levels is

$$\begin{aligned} \log(y_{ijkgr}) = & \mu + P_i + S_j + T_k + (PS)_{ij} + (PT)_{ik} + (ST)_{jk} \\ & + (PST)_{ijk} + A_{(ijk)r} + G_g + (PG)_{ig} + (SG)_{jg} \\ & + (TG)_{kg} + (PSG)_{ijg} + (PTG)_{ikg} + (STG)_{jkg} \\ & + (PSTG)_{ijkgr} + \epsilon_{ijkgr}, \end{aligned}$$

where $i = 1, 2; j = 1, 2; k = 1, 2, 3; g = 1, \dots, 22,810$; and $r = 1, 2, 3$. The main effects are denoted as P, S, T , and G and represent accession, SA treatment, time point, and gene, respectively. The array effect (A) is assumed to be distributed as a normal random variable with mean 0 and variance σ_A^2 ; ϵ_{ijkgr} represents the subplot error and is assumed to be normally distributed with mean 0 and variance σ_ϵ^2 .

Two null hypotheses were employed to test for gene expression differences both within and between accessions. The first null hypothesis was used to statistically identify genes whose expression level significantly changed between treatments within an accession. For each gene of each accession at each time point, $H_{01}: S_j + (SG)_{jg} = S_{j'} + (SG)_{j'g}$ was tested.

A second hypothesis was tested to detect potential ELPs. Specifically, for each gene under each treatment condition at each time point, there is no difference in expression levels between two accessions: $H_{02}: P_i + (PG)_{ig} = P_{i'} + (PG)_{i'g}$. For each of 21 accession pairs, t -tests were used with the type I error adjusted for multiple comparisons, Holm's correction (HOLM 1979), and false discovery rate (HOLM 1979; BENJAMINI and HOCHBERG 1995). The accession analysis was based on the total number of genes identified as statistically significant by at least one of the two hypotheses H_{01} and H_{02} , the number of potential ELPs detected by t -tests for H_{02} , and the number of genes with possible ELPs related to SA induction as identified by t -tests for both H_{01} and H_{02} . The numbers of such genes for all 21 parental pairs were compared to determine the maximum and minimum number of differences.

Quantitative RT-PCR: To validate gene expression data obtained from the GeneChips, we chose 15 single-copy genes for quantitative RT-PCR analysis (QRT-PCR). These genes were scattered throughout the genome and represented the range of ELP frequencies: 5 genes polymorphic in >50% of pairwise comparisons, 5 genes polymorphic in ~50% of pairwise comparisons, and 5 genes polymorphic in <50% of pairwise comparisons. Total RNA from the same samples used for hybridization to the GeneChips was treated with deoxyribonuclease I (Sigma, St. Louis) according to the manufacturer's instructions. The DNase-treated RNA was then used for first-strand cDNA synthesis with the Advantage RT-for-PCR kit (Clontech, Mountain View, CA), according to the manufacturer's instructions. The PCR primers were designed with Primer3 software (ROZEN and SKALETSKY 2000) to amplify fragments of ~140 bp (supplemental Table 1 at <http://www.genetics.org/supplemental/>). The *Actin7* gene, At5g09810, was used as a control gene for QRT-PCR since it exhibited invariant expression among all seven accessions. The optimal annealing temperature was determined with gradient PCR using the Robocycler Gradient 96 (Stratagene, La Jolla, CA). Real-time quantitative PCR was performed with the AB7300 Real Time PCR system, using the SYBR Green PCR Master Mix, the RT-PCR kit, and 96-well optical reaction plates [Applied Biosystems (Foster City, CA) 7300 Real Time PCR system]. PCR cycling conditions were 10 min at 95°, 40 cycles of 15 sec at 95°, and 1 min at 57°. Forward and reverse primer concentrations of 50, 300, and 900 nM were used to determine optimal conditions for each gene. The quality of the amplification was determined using the dissociation curve analysis of the AB7300 Real Time PCR system. Next, real-time PCR was performed with the control gene plus two of our chosen genes per plate, with seven cDNA samples and two no-template controls, each with two biological replicates and two technical replicates. For genes without an intron in the genomic DNA corresponding to the amplicon, control cDNA samples were used for which the reverse transcriptase had been omitted during cDNA synthesis. Threshold cycle (C_T) values reported by the AB7300 Real Time PCR system were used for data analysis. Three genes did not amplify under any condition utilized and were not considered in the analysis.

Single-feature polymorphism analysis: The Affymetrix ATH1 GeneChip contains oligomer probes designed from the DNA sequence of accession Col-0. When these arrays are hybridized with other genotypes, differences in hybridization signal may be due to true ELPs or due to sequence polymorphisms that result in altered probe affinity. The latter can be detected as single-feature polymorphisms (SFPs) between genotypes (BOREVITZ *et al.* 2003; WINZELER *et al.* 2003). Most Arabidopsis genes on the ATH1 GeneChip that contain a SFP are due to single-nucleotide polymorphisms (BOREVITZ *et al.* 2003). In total, 11 probe pairs represent each gene on the ATH1 GeneChip. To test whether SFPs had a major impact on

perceived expression differences, we utilized the individual probe data from the ATH1 GeneChips to identify SFPs between Col-0 and Cvi-1. These two accessions were used as they showed the largest number of potential ELPs. Every probe from 2000 genes was tested for an SFP using the following algorithm: $(\text{probe}^{\text{Cvi-1}}/\text{probeset average}^{\text{Cvi-1}})/(\text{probe}^{\text{Col-0}}/\text{probeset average}^{\text{Col-0}})$. A value of 0.40 was used as the cutoff for declaring a significant SFP as this was the empirically determined 0.05 threshold derived from a histogram of all ~22,000 tests: 5% of the tests were between the values of 0.1 and 0.4, and 95% of the tests were between 0.4 and 1. The genes were classified by the number of SFPs that were identified and whether the gene showed differential expression in our specific pairwise comparisons of accessions. We randomly selected 400 genes that had one or more SFPs per gene and calculated the average expression level for each gene from each GeneChip in two ways: using all 11 perfect-match (PM) probes that included one or more with an SFP and using only the probes lacking any SFPs. An ANOVA was performed for each gene, using a split-split-plot mixed linear model with a random array effect, identical to the previous model above except that the sub-subplot was defined as the probes within a gene and the subplot was the gene. By comparing the per-gene expression values with and without SFP-containing probes, we could test whether SFPs were a significant source of variation in a gene's overall expression value.

ELP sliding-window analysis: All 22,810 genes represented on the ATH1 GeneChip and all of the 875 sequenced loci (NORDBORG *et al.* 2005) (walnut.usc.edu/2010/) were positioned on the Arabidopsis physical map (TAIR website; www.arabidopsis.org), using the ATG start codon as the start of each gene or the first base pair as the start of each sequenced locus. Each gene was classified according to whether it showed differential expression between any of our six accessions that were also a part of the published sequence diversity study (Col-0, Cvi-1, Kin-0, Mt-0, Tsu-1, and Van-0) (NORDBORG *et al.* 2005). Microsoft Excel was used to determine the frequency of gene expression divergence along each chromosome, in a 500-gene sliding window. The values were expressed as the ELP/gene ratio by dividing the number of ELPs per 500-gene sliding window by 500 to place these values on a similar scale as the nucleotide polymorphism values for ease of comparison. This measure of divergence was plotted at the physical position for the median of the 500 genes within the window. The window was moved to the next adjacent gene, and the process was repeated until the end of the chromosome was reached. We utilized a fixed 500-gene window to minimize the effects of variation in gene density upon the estimate. Windows of 100 and 250 genes provided similar results, although the traces were less smooth (data not shown).

To estimate the threshold limit at a significance level of 0.05 for the frequency of differential gene expression, the positions of the differentially expressed genes were randomly permuted across the genome 500 times. The sliding-window analysis was repeated for each permutation, and the maximal and minimal values of differentially expressed genes per window per randomly permuted genome were obtained. The upper and lower $P = 0.05$ limits were then empirically set, using these distributions from the permutations (0.056 ELP/gene for the lower bound and 0.132 ELP/gene for the upper bound). The extreme lower and upper values identified in the permutation analysis were 0.048 ELP/gene and 0.150 ELP/gene, respectively.

Sequence divergence analysis: Sequence alignments for 96 *A. thaliana* accessions at 875 loci were downloaded from the public website (NORDBORG *et al.* 2005) (walnut.usc.edu/2010/). Six of the seven accessions (Col-0, Cvi-1, Kin-0, Mt-0, Tsu-1, and Van-0) in our study (Est was not included) were represented in

this sequence collection. Alignments were then filtered to identify and count the polymorphisms across all 96 accessions, across the 6 accessions in common between the two studies, as well as for each possible pairwise combination among our 6 accessions. This information was utilized to measure the frequency of each polymorphism across all 96 accessions and to classify whether it was polymorphic in our 6 accessions. The polymorphism information from all possible pairwise alignments of the 6 accessions was used to calculate the per-locus nucleotide diversity (π) for each pairwise alignment (TAJIMA 1983; NEI 1987). A linear regression analysis was utilized to determine the relationship between sequence polymorphism and level of ELPs. The number of ELPs per accession pair was regressed upon the average π -measure for the same accession pair. Each accession pair was given a categorical variable indicating whether that pair contained the Col-0 accession or not. This variable was included in the analysis to test if the linear relationship was dependent upon the presence of Col-0; the presence of Col-0 did affect the ELP estimate, but it did not alter the positive association between ELP number and the average π -measure. A 5-Mbp sliding window of the average π -measure for the same 6 accessions using all available 875 sequenced loci was generated in Excel as described previously. The average value was plotted at the median physical position. A 5-Mbp window size was selected to allow for sufficient averaging across the sequenced loci.

GO annotation analysis: The complete genome gene ontology (GO) biological classification list was obtained and used to classify each gene (www.arabidopsis.org). The frequency of each first-level classification was obtained for the complete genome as well as for each list of ELPs from the pairwise accession ANOVAs. Each ELP list was tested for significant deviations from the expected frequencies for the complete genome using χ^2 -analysis (CHEN *et al.* 2005).

RESULTS

Diversity sampling: Six of our 7 accessions are included in a set of 96 Arabidopsis accessions from which 875 loci have been sequenced (NORDBORG *et al.* 2005) (walnut.usc.edu/2010/), allowing us to estimate the amount of species-wide DNA polymorphism contained in our sample of accessions. We identified the polymorphisms present in 875 loci distributed across the genome and scored the frequency of each polymorphism among the 96 accessions and if that polymorphism was variable in our 6 accessions. Our 6 accessions contained ~80% of the species-wide medium- to high-frequency polymorphisms (minor allele frequency >10%) (Figure 1). In the rare- to low-frequency class, our accessions contained only 20–30% of the species diversity (Figure 1). Therefore, our accessions were representative of most of the medium- to high-frequency sequence variation in Arabidopsis.

ANOVA of microarray data for differentially expressed genes: The split-plot ANOVA revealed that four effects were significant for all 21 pairs of accessions: array, gene, interaction of gene with accession, and gene with time point (data not shown and Table 1). Since expression levels vary across different genes, a significant gene effect was expected. The significant

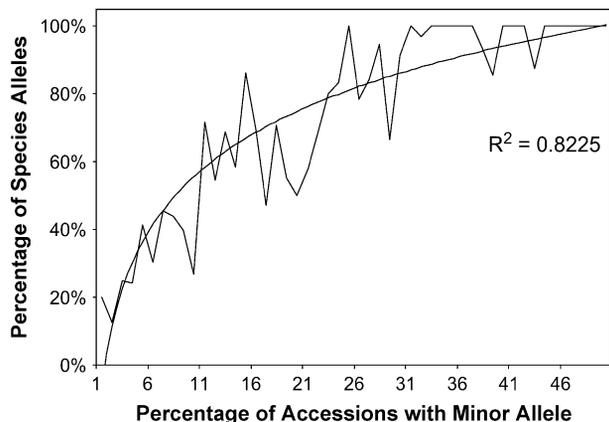


FIGURE 1.—Sampling of species variation. The percentage of polymorphisms in 96 *Arabidopsis* accessions that are polymorphic in the six accessions in common with this data set (Col-0, Cvi-1, Kin-0, Mt-0, Tsu-1, and Van-0) are shown (NORDBORG *et al.* 2005). The alleles are binned on the basis of the minor allele percentage in the 96-accession data set (shown on the *x*-axis). The smooth line represents the expected random-sampling distribution for 6 accessions for each allele percentage class.

gene \times accession interaction effect indicates that gene expression behaved differently in the various accessions. Since these differences may be due to potential ELPs or sequence polymorphisms, the statistical analysis cannot distinguish between these two alternatives. A significant gene \times time point interaction effect indicates that different genes exhibit distinct expression patterns over time. The significant array effect shows that there is chip-to-chip variation, but the split-plot model properly accounted for this source of variation. Over the 21 pairs of accessions, there were no significant effects for SA treatment, interaction of accession \times treatment, and accession \times treatment \times gene (three-way interaction), suggesting that the SA treatment did not strongly affect the overall expression levels across all 22,810 genes, which is expected given that only a subset of genes is induced by SA. However, the regulation of expression over time was definitely affected by SA treatment, as evidenced by significant effects for time points and all time-related interactions.

The pairwise ANOVA identified on average 2234 genes showing differential expression signal (potential ELPs) between any two accessions across treatments, using false discovery rate (FDR) multiple-testing correction (BENJAMINI and HOCHBERG 1995), and 331 genes when using Holm's multiple testing correction (HOLM 1979) (Figure 2). Accession Cvi-1 had more potential ELPs in all of its pairwise comparisons than did any of the other accessions (Figure 2). To assess the variability of potential ELPs, we added the number of pairwise comparisons for which any given gene showed a differential expression signal. The resulting distribution was heavy tailed and skewed toward low- to moderate-frequency ELPs and was similar for gene lists obtained

TABLE 1
ANOVA of pairwise expression differences

Effect ^a	d.f.	SS	MS	Fvalue	P-value
<i>P</i>	1	1,843.1	1,843.139	7.118	0.0135
<i>S</i>	1	328.6	328.687	1.269	0.2710
<i>T</i>	2	1,361.9	680.990	2.630	0.0928
<i>PS</i>	1	4.0	4.036	0.016	0.9017
<i>PT</i>	2	227.1	113.554	0.439	0.6501
<i>ST</i>	2	518.9	259.460	1.002	0.3820
<i>PST</i>	2	1,408.7	704.392	2.720	0.0861
Array	24	6,215.0	258.960	5,232.731	<0.0001
<i>G</i>	22,809	564,328.8	24.742	499.944	<0.0001
<i>PG</i>	22,809	7,418.9	0.325	6.573	<0.0001
<i>SG</i>	22,809	1,149.2	0.050	1.018	0.0294
<i>TG</i>	45,618	6,144.5	0.135	2.722	<0.0001
<i>PSG</i>	22,809	991.3	0.044	0.878	>0.9999
<i>PTG</i>	45,618	2,426.3	0.053	1.075	<0.0000
<i>STG</i>	45,618	3,218.0	0.071	1.426	<0.0000
<i>PSTG</i>	45,618	2,264.6	0.050	1.003	0.3243
Error	547,416	27,090.8	0.050		
Corrected total	821,159	626,940.5			

Pairwise ANOVA comparing accessions Col-0 and Van-0 as an example of results from a split-plot analysis of variance is shown. A similar analysis was performed for all 21 pairwise combinations of the seven accessions in our study.

^aEffects are denoted as follows: *P*, accession; *S*, SA treatment condition; *T*, time point; *G*, gene.

using either the FDR or Holm's procedure (Figure 3). Interestingly, several hundred genes showed differential expression signals in over half of the pairwise tests, suggesting that these genes have multiple alleles that can confer distinct expression levels.

We included a SA treatment as we are interested in the natural variation in responses to SA application. The use of two treatments (SA and control) provided an opportunity to examine the consequences of experimental treatment variation on the set of genes showing differential expression. SA is a major signaling molecule involved in plant defense responses (GAFFNEY *et al.* 1993; DELANEY *et al.* 1995). We utilized a 0.30-mM SA treatment, as higher SA levels caused phytotoxicity in some accessions (data not shown). Statistical analysis identified an average of 660 genes with FDR and 25 genes with Holm's procedure showing differential SA response between at least one pair of accessions. These numbers represent ~10–20% of the genes showing a difference in expression among accessions that did not exhibit a treatment interaction. This suggests there is considerable variation in responses to this key signaling molecule, but how this relates to variation in disease resistance is currently unknown.

Control genes: To test the ability of our ANOVA to detect genes that had previously been reported as varying

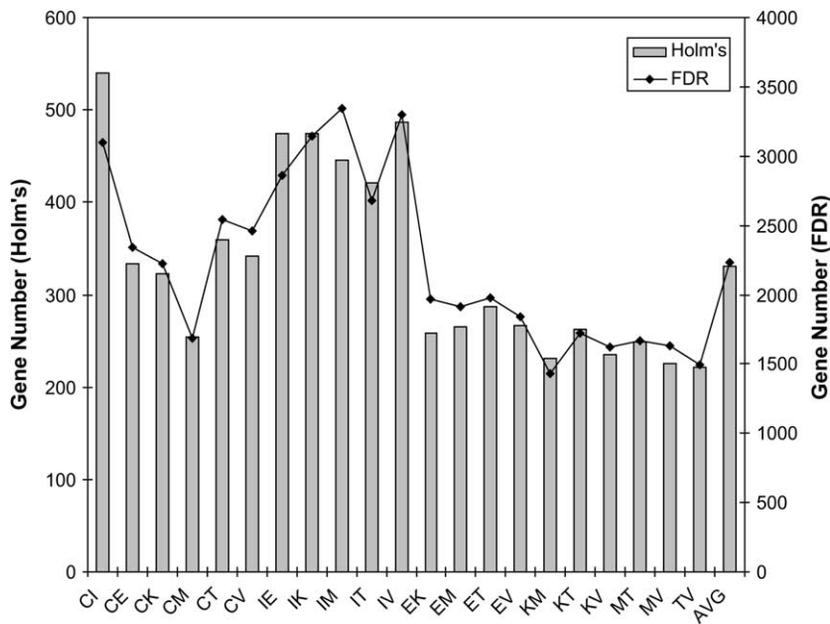


FIGURE 2.—Differential gene expression per pairs of accessions. The number of genes showing differential expression per accession pair is shown. The left y-axis and bars show the number of genes showing polymorphic expression, as determined by a significant H_{01} (differentially expressed between two accessions), with significance determined by Holm's test (see MATERIALS AND METHODS for details). The right y-axis and diamonds show the number of genes with expression polymorphisms that satisfy H_{01} with the significance level determined by the FDR test. The accession pair is listed on the x-axis: C, Col-0; I, Cvi-1; E, Est; K, Kin-0; M, Mt-0; T, Tsu-1; and V, Van-0. Avg is the average across all pairwise comparisons.

between accessions, we studied a set of genes that were known to display natural variation in DNA sequence, mRNA accumulation, and phenotype. This comparison included genes involved in the production of the plant-defense-related glucosinolates (*ESP*, *AOP2*, *AOP3*, *GS-OH*, *Mam1*, and *MamL*) (KLIBENSTEIN *et al.* 2001; LAMBRIX *et al.* 2001; KROYMANN *et al.* 2003), flowering time (*FLC*, *FRI*, *FLM*, and *HUA2*) (JOHANSON *et al.* 2000; CAICEDO *et al.* 2004; DOYLE *et al.* 2005; LEMPE *et al.* 2005; WERNER *et al.* 2005), hypocotyl light responses (*PHYA* and *CRY2*) (EL-ASSAL *et al.* 2001; MALOOF *et al.* 2001), and elicitation of resistance (*RPM1*, *RPS2*, *RPS4*, *RPS5*, and *RPP5*) (GRANT *et al.* 1995; PARKER *et al.* 1997;

GASSMANN *et al.* 1999; HENK *et al.* 1999; MAURICIO *et al.* 2003). These genes were classified on the basis of the nature of the molecular polymorphisms (Table 2). Any polymorphism (INDEL, splice site polymorphism, or observed mRNA difference) that might cause a change in observed gene expression was listed as a potential ELP. Our statistical analysis identified 12 of the 13 genes previously known or predicted as being differentially expressed (Table 2). One gene (*FLM*) that we did not identify as polymorphic has an INDEL in an accession that was not included in our study (WERNER *et al.* 2005). Genes with nonsense or missense polymorphisms served as controls and, as expected, were not detected as differentially expressed. These results suggest that our experiment has a low false negative rate for detecting ELPs.

To check if the above analysis was biased by using genes known to cause phenotypic polymorphisms, we used the 12 genes showing differential expression in the seven accessions for QRT-PCR analysis for differential gene expression. The average correlation for the expression values of these 12 genes between QRT-PCR and ATH1 GeneChip values was 0.69, indicating that the microarray analysis accurately detects genes with ELPs.

SFPs and differential expression: To test if the presence of SFPs significantly alters estimates of gene expression, we calculated the apparent expression levels for 400 genes that contained SFPs between Col-0 and Cvi-1 with and without signal from the SFP-containing probes and employed a split-split-plot ANOVA. Of the 400 SFP-containing genes, the SFP was responsible for $\sim 0.13\%$ of the variance in signal among these genes. Furthermore, inclusion of the probes with SFPs in the estimation of a gene's expression value led to only a 7% decrease in the estimate. Therefore, SFPs are not a significant

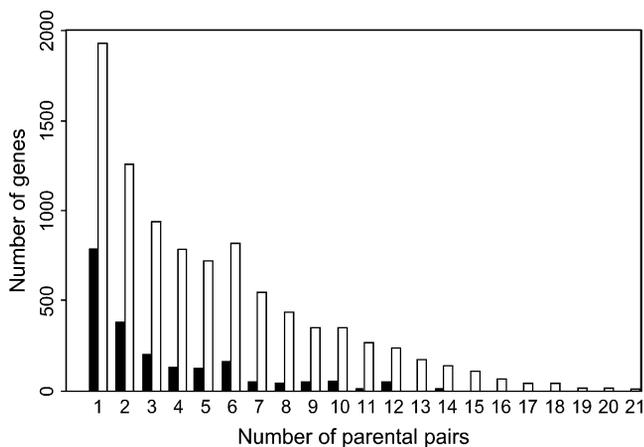


FIGURE 3.—Frequency of pairwise accession gene expression differences. The horizontal axis indicates the number of accession pairs ($n = 21$ pairs) for which the genes are detected as differentially expressed. Solid bars denote the number of genes identified with Holm's procedure; open bars denote the number of genes identified with the FDR procedure.

TABLE 2
Identification of known genes with expected variation in expression among accessions

Class ^a	Gene name	Gene ID	Polymorphism ^b	Gene identified? ^c	
				Holm's	FDR
Glucosinolate	<i>ESP</i>	AT1G54040	Expression	Yes	Yes
	<i>AOP2</i>	AT4G03060	Expression	Yes	Yes
	<i>AOP3</i>	AT4G03050	Expression	Yes	Yes
	<i>GS-OH</i>	AT2G25450	Expression	Yes	Yes
	<i>MAMI</i>	AT5G23010	Expression	Yes	Yes
	<i>MAML</i>	AT5G23020	Expression	Yes	Yes
Flowering time	<i>FLC</i>	AT5G10140	Expression	Yes	Yes
	<i>FRI</i>	AT4G00650	Expression	No	Yes
	<i>FLM</i>	AT1G77080	Expression	No	No
	<i>HUA2</i>	AT5G23150	Nonsense	No	No
Light response	<i>PHYA</i>	AT1G09570	Missense	No	No
	<i>CRY2</i>	AT1G04400	Missense	No	No
Defense	<i>RPM1</i>	AT3G07040	Expression	Yes	Yes
	<i>RPS2</i>	AT4G26090	Missense	No	No
	<i>RPS4</i>	AT5G45250	Expression	Yes	Yes
	<i>RPS5</i>	AT1G12220	Expression	Yes	Yes
	<i>RPP5</i>	AT4G16860	Expression	Yes	Yes

Shown are genes previously reported to be naturally variable among Arabidopsis accessions in foliar tissue and that generate a known phenotype.

^aDenotes the phenotype altered by the variation.

^bLists whether the variation is expression based or not, as described in the literature (see text for references). Polymorphisms labeled as nonsense or missense have the corresponding sequence polymorphism and are not known to alter expression in plants.

^cIndicates if the gene was identified as being naturally variable (satisfying H_{02}) by either Holm's procedure or FDR in our analyses.

source of variation in gene expression estimates and the differences among accessions were predominantly due to actual ELPs.

Gene function bias: Previous work had suggested a bias in the functional classification of genes that show differential expression among Arabidopsis accessions (CHEN *et al.* 2005). When the GO biological annotations were used to classify the genes, the differentially expressed genes were significantly enriched for genes classified as controlling biotic and abiotic responses, stress responses, and signal transduction (Figure 4). These differences in classification frequencies were significant by χ^2 -tests in all pairwise comparisons per category as well as for the average of the pairwise comparisons after adjusting for multiple comparisons with Bonferroni correction (data not shown). In addition, there was a significant decrease in genes classified as being involved in transport (Figure 4). This result was similar to those previously reported by CHEN *et al.* (2005). A similar comparison using the GO molecular annotations identified a significant difference from the whole-genome expectation (in over half of the pairwise comparisons) only for genes classified as receptor binding or activity. No classes in GO cellular annotations were significant in the pairwise comparisons (data not shown).

Relationship of pairwise expression differences to sequence divergence: If there is no major filtering mechanism, such as differential selection pressures upon sequence and expression polymorphisms or massive *trans*-acting effects, to bias the impact of DNA sequence variation on gene expression variation, a genome-wide relationship between DNA sequence variation and ELP variation is expected. To assess this relationship, we determined the pairwise π , average pairwise nucleotide divergence (TAJIMA 1983; NEI 1987) between the six accessions for which there were sequence data, and then compared π with the number of genes we detected as differentially expressed between the same accession pair (as determined by FDR). There was a significant positive relationship between these values (Figure 5). This linear regression analysis accounted for whether the pair included Col-0, from which the oligomer probes on the ATH1 GeneChip were designed ($N = 15$, $P < 0.0001$, $r = 0.83$; Col-0 presence in accession pair, d.f. = 1, $P = 0.0483$) (Figure 5). The estimates of gene expression divergence provided by Holm's multiple testing correction generated a nearly identical relationship between sequence and gene expression divergence ($N = 15$, $P < 0.0001$, $r = 0.86$; Col-0 presence in accession pair, d.f. = 1, $P = 0.0097$). This association was

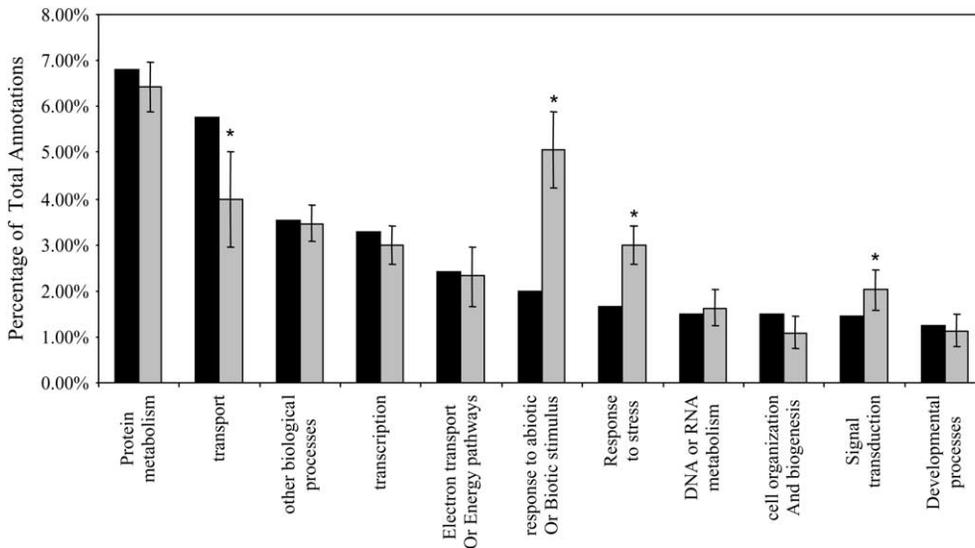


FIGURE 4.—GO biological annotation comparison for differentially expressed genes. Solid bars show the percentage of total annotations in each GO biological annotation classification using the completely sequenced Arabidopsis genome from Col-0. Shaded bars (with standard errors) show the average percentage of total annotations for each of the above GO biological classifications using all 21 comparisons between pairs of accessions. Asterisks indicate those classifications that are significantly different from the whole-genome analysis as determined by χ^2 -tests, with Bonferroni correction. GO biological classifications that describe unknown proteins were not included (*i.e.*, other physiological processes, other metabolic processes, other cellular processes, and biological processes unknown) since they were uninformative.

also significant when comparing π against the number of genes showing an expression difference dependent upon an accession \times treatment interaction ($N = 15$, $P < 0.0001$, $r = 0.95$; Col-0 presence in accession pair, d.f. = 1, $P = 0.001$). These genes were not likely to have been affected by hybridization differences, such as those caused by INDEL polymorphisms, since these genes exhibited nonvariable gene expression in one treatment across two accessions but showed differential gene expression under the other treatment. This relationship between π and the genes significant for accession \times treatment interactions further supports the hypothesis that SFPs did not significantly influence estimates of

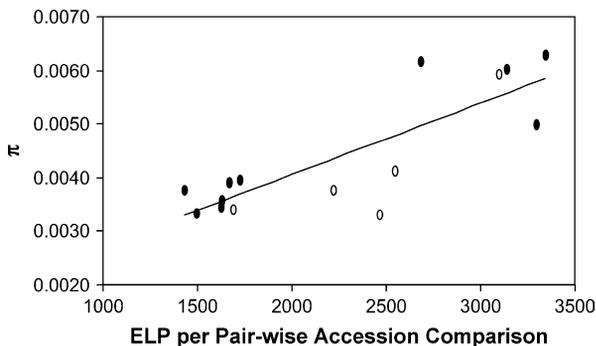


FIGURE 5.—Sequence divergence and gene expression differences are associated. The relationship between π and number of ELPs between pairs of six accessions (Col-0, Cvi-1, Kin-0, Mt-0, Tsu-1, and Van-0), as determined by linear regression and FDR gene list ($n = 15$, $P < 0.001$, $r^2 = 0.83$), is shown. Pairwise comparisons containing Col-0 are shown as open circles; solid circles denote pairs that do not include Col-0. A similar linear relationship was identified with Holm's gene list ($n = 15$, $P < 0.001$, $r = 0.80$).

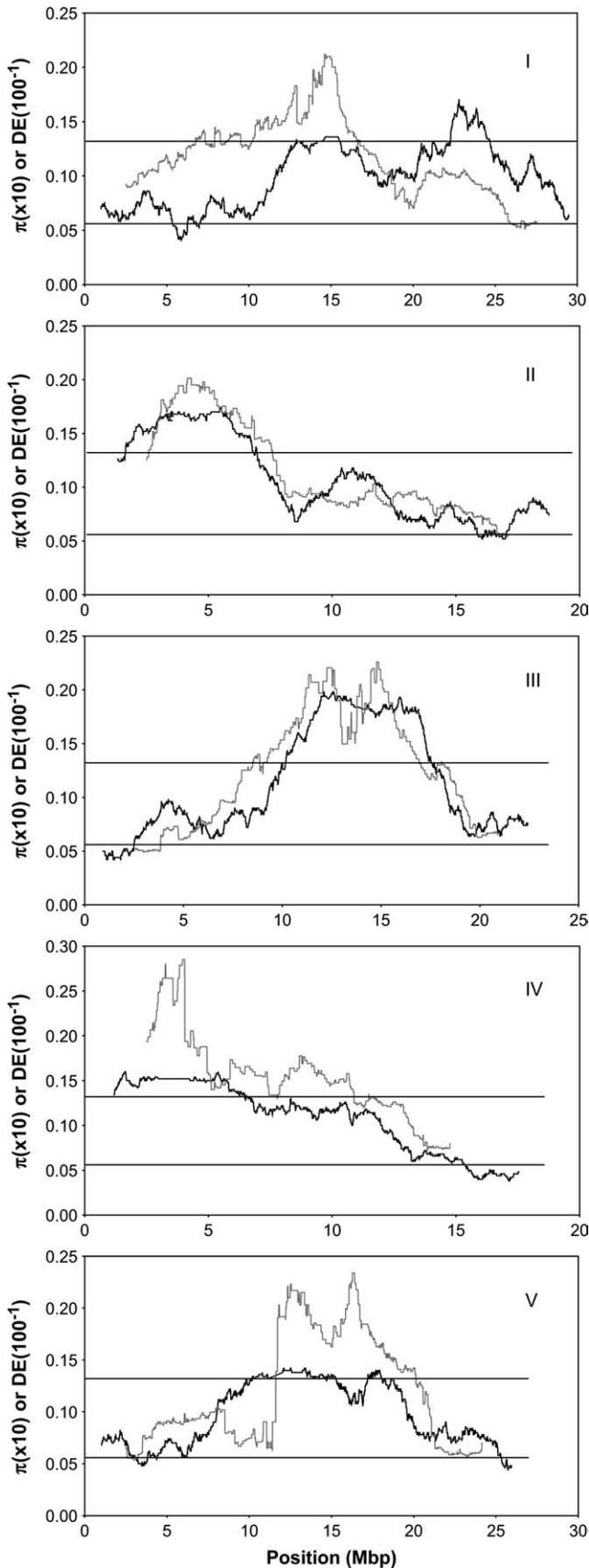
gene expression variation and that we indeed measured ELPs among accessions.

Genomic distribution of expression differences and sequence divergence: The relationship between π and gene expression variation (previous section) is a genome-wide average comparison. To assess if there was a relationship between local DNA sequence variation and gene expression variation, we conducted a sliding-window analysis along the Arabidopsis chromosomes (Figure 6). A distinct relationship in local sequence variation and gene expression variation was detected, as the two traces were very similar across the chromosomes (Figure 6). Local peaks of DNA sequence divergence among the accessions typically colocalized with local enrichments in the number of genes showing polymorphic gene expression. In particular, there is a region near the middle of chromosome III that has a significant elevation in gene expression polymorphism and is also enriched in DNA sequence variation (Figure 6). Two regions with a lack of local correlation between sequence diversity and polymorphic gene expression can be explained by technical or biological reasons. The flat region for differential gene expression at the top of chromosome IV is an area of low gene density due to the presence of the nucleolar organizing region. The ~ 10 -Mbp region on chromosome V that had low estimates of π also has a paucity of sequenced loci and may represent a local sampling bias.

DISCUSSION

We identified a total of 2066 genes in the seven accessions (Col-0, Cvi-1, Est, Kin-0, Mt-0, Tsu-1, and Van-0)

as differentially expressed under Holm's multiple testing procedure and 6433 genes with FDR, suggesting that 10–30% of the transcriptome showed natural variation



in expression in foliar tissue. Considering that the limited number of accessions used here efficiently sampled only moderate- to high-frequency DNA polymorphisms, it is highly probable that these are lower bounds on the frequencies of ELPs. There are likely many rare ELPs that were not represented in our accessions. Furthermore, we sampled only one tissue, at one developmental stage, for one set of environmental conditions, and in the absence of biotic pests. Therefore, even the conservative estimate of differentially expressed genes from Holm's multiple testing procedure suggests that a significant fraction of the Arabidopsis genome shows differential expression under one circumstance or another. This will be particularly the case if genotype \times environment or genotype \times development interactions play a significant role in differential gene expression. The frequency distribution of ELPs in our accessions also suggests that our estimate of ELPs is conservative. Over 70% of the differentially expressed genes were present in <3 of the 21 pairwise comparisons of seven accessions. Thus, like DNA sequence polymorphisms in Arabidopsis, most gene expression polymorphisms appear to be rare- to low-frequency events. An analysis of a much larger collection of accessions and more extensive variation in experimental conditions would be required to generate an estimate of the higher bounds of gene expression variation even for a single tissue at a single developmental age.

The ELPs detected in our study showed a bias in their categorization by the GO biological annotations, with an overrepresentation of genes involved in signaling and plant/stress responses compared to the base distribution from the Col-0 sequence annotation (Figure 4). These biases were the same as those found by CHEN *et al.* (2005) in their analysis of variable gene expression among accessions sampled at different developmental stages. As suggested by these categorical differences, the greatest opportunities for the evolution of differences in gene expression levels in Arabidopsis are in responses to the environment. This idea is consistent with the number of flowering-time and light-response quantitative trait loci (QTL) that have been mapped to known signaling loci (JOHANSON *et al.* 2000; EL-ASSAL *et al.* 2001;

FIGURE 6.—Local chromosomal association between nucleotide and differential gene expression variation. Sliding-window comparison of nucleotide variation (π) vs. differential gene expression in the 6 accessions (Col-0, Cvi-1, Kin-0, Mt-0, Tsu-1, and Van-0) in common with the 96 accessions from the sequence database is shown (NORDBORG *et al.* 2005). The black line is the average number of ELPs per locus measured in a 500-gene-wide sliding window. The x-axis of each graph is in megabase pairs as marked. The horizontal lines represent the $P=0.05$ thresholds for the differentially expressed sliding window as obtained via the permutation analysis (0.056 is the $P=0.05$ minimum and 0.132 is the $P=0.05$ maximum). The gray line represents a 5-Mbp sliding-window analysis of average π per loci.

MALOOF *et al.* 2001; CAICEDO *et al.* 2004; DOYLE *et al.* 2005; WERNER *et al.* 2005). However, it is also possible that even though the plants were grown under carefully controlled conditions, there were physiological differences between the accessions that led to differential stress responses.

We identified a positive relationship between ELPs and DNA sequence divergence at both the whole-genome and local subchromosomal levels (Figures 5 and 6). This relationship implies that there is a direct association between DNA sequence polymorphism and ELP. Such a relationship at the local level suggests that there is a significant impact of *cis*- (or linked *trans*-) DNA sequence polymorphism on gene expression. If the majority of expression differences were due to unlinked *trans*-polymorphisms, the global estimates of sequence and expression polymorphism would still be correlated but at the local level there would not be a relationship between ELP and DNA sequence polymorphism. However, there was heterogeneity in the local relationship between sequence and gene expression variation. For example, chromosome I appeared to be split, with the top half of the chromosome showing an association while the bottom half of chromosome I showed a decrease in π in relation to gene expression variation (Figure 6). Thus, there seemed to be an enhanced frequency of ELPs at the distal end of chromosome I. It will require genetic analysis of a segregating population to provide more definitive data on the relative importance of *cis*- vs. *trans*-acting DNA polymorphisms in regulating ELPs in Arabidopsis.

This work provides a baseline description of the level and patterns of ELP polymorphism within Arabidopsis. As a number of Arabidopsis QTL have been shown to be controlled by an ELP within a causative gene, our data may enhance QTL cloning by providing rapid candidate gene identification. Furthermore, these data provide a foundation from which to explore the impacts of different environmental conditions and developmental stages upon ELPs. Additionally, ELPs can be mapped in a segregating population as expression QTL to characterize how *cis*- and *trans*-acting loci coordinate to generate ELP patterns within the species (BREM *et al.* 2002; SCHADT *et al.* 2003; YVERT *et al.* 2003; BREM and KRUGLYAK 2005).

Technical assistance was provided by Carol Lam and Rebecca Walker. This research was supported by the National Science Foundation (NSF) 2010 Project grant MCB-0115109 to D.A.S., R.W.M., and R.W.D. and by NSF grant MCB-0323759 to D.J.K.

LITERATURE CITED

- AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN *et al.*, 2004 *Current Protocols in Molecular Biology*. John Wiley & Sons, New York.
- BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**: 289–300.
- BENTLEY, S. D., and J. PARKHILL, 2004 Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.* **38**: 771–792.
- BOREVITZ, J. O., D. LIANG, D. PLOUFFE, H. S. CHANG, T. ZHU *et al.*, 2003 Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13**: 513–523.
- BREM, R. B., and L. KRUGLYAK, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* **102**: 1572–1577.
- BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.
- CAICEDO, A. L., J. R. STINGHCOMBE, K. M. OLSEN, J. SCHMITT and M. D. PURUGGANAN, 2004 Epistatic interaction between Arabidopsis *FRJ* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. *Proc. Natl. Acad. Sci. USA* **101**: 15670–15675.
- CARROL, S. B., 2000 Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**: 577–580.
- CHEN, W. J., S. H. CHANG, M. E. HUDSON, W.-K. KWAN, J. LI *et al.*, 2005 Contribution of transcriptional regulation to natural variations in Arabidopsis. *Genome Biol.* **6**: R32.
- CONG, B., J. P. LIU and S. D. TANKLEY, 2002 Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations. *Proc. Natl. Acad. Sci. USA* **99**: 13606–13611.
- CRAIG, B. A., M. A. BLACK and R. W. DOERGE, 2003 Gene expression data: the technology and statistical analysis. *J. Agric. Biol. Environ. Stat.* **8**: 1–28.
- DELANEY, T. P., L. FRIEDRICH and J. A. RYALS, 1995 Arabidopsis signal transduction mutant defective in chemically and biologically induced disease resistance. *Proc. Natl. Acad. Sci. USA* **92**: 6602–6606.
- DOERGE, R. W., 2002 Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.* **3**: 43–52.
- DOYLE, M. R., C. M. BIZZELL, M. R. KELLER, S. D. MICHAELS, J. D. SONG *et al.*, 2005 HUA2 is required for the expression of floral repressors in Arabidopsis thaliana. *Plant J.* **41**: 376–385.
- EL-ASSAL, S. E. D., C. ALONSO-BLANCO, A. J. M. PEETERS, V. RAZ and M. KOORNNEEF, 2001 A QTL for flowering time in Arabidopsis reveals a novel allele of *CRY2*. *Nat. Genet.* **29**: 435–440.
- FRARY, A., T. C. NESBITT, A. FRARY, S. GRANDILLO, E. VAN DER KNAAP *et al.*, 2000 *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**: 85–88.
- GAFFNEY, T., L. FRIEDRICH, B. VERNOOIJ, D. NEGROTTO, G. NYE *et al.*, 1993 Requirement of salicylic acid for the induction of systemic acquired-resistance. *Science* **261**: 754–756.
- GASSMANN, W., M. E. HINSCH and B. J. STASKAWICZ, 1999 The Arabidopsis *RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant J.* **20**: 265–277.
- GIBBS, R. A., J. W. BELMONT, P. HARDENBOL, T. D. WILLIS, F. L. YU *et al.*, 2003 The International HapMap Project. *Nature* **426**: 789–796.
- GLAZEBROOK, J., 2001 Genes controlling expression of defense responses in Arabidopsis—2001 status. *Curr. Opin. Plant Biol.* **4**: 301–308.
- GRANT, M. R., L. GODIARD, E. STRAUPE, T. ASHFIELD, J. LEWALD *et al.*, 1995 Structure of the Arabidopsis *Rpm1* gene enabling dual-specificity disease resistance. *Science* **269**: 843–846.
- HENK, A. D., R. F. WARREN and R. W. INNES, 1999 A new *Ae*-like transposon of Arabidopsis is associated with a deletion of the *RPS5* disease resistance gene. *Genetics* **151**: 1581–1589.
- HOLM, S., 1979 A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**: 65–70.
- JOHANSON, U., J. WEST, C. LISTER, S. MICHAELS, R. AMASINO *et al.*, 2000 Molecular analysis of *FRIGIDA*, a major determinant of natural variation in Arabidopsis flowering time. *Science* **290**: 344–347.
- KLIEBENSTEIN, D., V. LAMBRIX, M. REICHEL, J. GERSHENZON and T. MITCHELL-OLDS, 2001 Gene duplication and the diversification of secondary metabolism: side chain modification of glucosinolates in Arabidopsis thaliana. *Plant Cell* **13**: 681–693.
- KLIEBENSTEIN, D. J., D. PEDERSEN and T. MITCHELL-OLDS, 2002 Comparative analysis of insect resistance QTL and QTL controlling the myrosinase/glucosinolate system in Arabidopsis thaliana. *Genetics* **161**: 325–332.
- KROYMANN, J., S. DONNERHACKE, D. SCHNABELRAUCH and T. MITCHELL-OLDS, 2003 Evolutionary dynamics of an Arabidopsis insect

- resistance quantitative trait locus. *Proc. Natl. Acad. Sci. USA* **100**: 14587–14592.
- LAMBRIX, V., M. REICHELDT, T. MITCHELL-OLDS, D. KLIEBENSTEIN and J. GERSHENZON, 2001 The *Arabidopsis* epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences *Trichoplusia ni* herbivory. *Plant Cell* **13**: 2793–2807.
- LEMPE, J., S. BALASUBRAMANIAN, S. SURESHKUMAR, A. SINGH, M. SCHMID *et al.*, 2005 Diversity of flowering responses in wild *Arabidopsis thaliana* strains. *PLoS Genet.* **1**: 109–118.
- MALOOF, J. N., J. O. BOREVITZ, T. DABI, J. LUTES, R. B. NEHRING *et al.*, 2001 Natural variation in light sensitivity of *Arabidopsis*. *Nat. Genet.* **29**: 441–446.
- MAURICIO, R., E. A. STAHL, T. KORVES, D. C. TIAN, M. KREITMAN *et al.*, 2003 Natural selection for polymorphism in the disease resistance gene *Rps2* of *Arabidopsis thaliana*. *Genetics* **163**: 735–746.
- NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**: e196.
- NUZHIDIN, S. V., M. L. WAYNE, K. L. HARMON and L. M. MCINTYRE, 2004 Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol. Biol. Evol.* **21**: 1308–1317.
- PARKER, J. E., M. J. COLEMAN, V. SZABO, L. N. FROST, R. SCHMIDT *et al.*, 1997 The *Arabidopsis* downy mildew resistance gene *RPP5* shares similarity to the toll and interleukin-1 receptors with N and L6. *Plant Cell* **9**: 879–894.
- ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers, pp. 365–386 in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by S. KRAWETZ and S. MISENER. Humana Press, Totowa, NJ.
- SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
- SCHMID, K. J., T. R. SORENSEN, R. STRACKE, O. TORJEK, T. ALTMANN *et al.*, 2003 Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* **13**: 1250–1257.
- SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169**: 1601–1615.
- STEIN, L. D., Z. R. BAO, D. BLASAR, T. BLUMENTHAL, M. R. BRENT *et al.*, 2003 The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**: 166.
- TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- WANG, R. L., A. STEC, J. HEY, L. LUKENS and J. DOEBLEY, 1999 The limits of selection during maize domestication. *Nature* **398**: 236–239.
- WERNER, J. D., J. O. BOREVITZ, N. WARTHMAN, G. T. TRAINER, J. R. ECKER *et al.*, 2005 Quantitative trait locus mapping and DNA array hybridization identify an *FLM* deletion as a cause for natural flowering-time variation. *Proc. Natl. Acad. Sci. USA* **102**: 2460–2465.
- WINZELER, E. A., C. I. CASTILLO-DAVIS, G. OSHIRO, D. LIANG, D. R. RICHARDS *et al.*, 2003 Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**: 79–89.
- YVERT, G., R. B. BREM, J. WHITTLE, J. M. AKEY, E. FOSS *et al.*, 2003 Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* **35**: 57–64.

Communicating editor: J. A. BIRCHLER