DOI: 10.1534/genetics.105.042275

Genetic Regulation of Gene Expression During Shoot Development in Arabidopsis

Rhonda DeCook,* Sonia Lall,† Dan Nettleton* and Stephen H. Howell†,[1]

**Department of Statistics, Iowa State University, Ames, Iowa 50011 and †Plant Sciences Institute, Iowa State University, Ames, Iowa 50011*

Manuscript received February 21, 2005
Accepted for publication September 28, 2005

ABSTRACT

The genetic control of gene expression during shoot development in *Arabidopsis thaliana* was analyzed by combining quantitative trait loci (QTL) and microarray analysis. Using oligonucleotide array data from 30 recombinant inbred lines derived from a cross of Columbia and Landsberg *erecta* ecotypes, the Arabidopsis genome was scanned for marker-by-gene linkages or so-called expression QTL (eQTL). Single-feature polymorphisms (SFPs) associated with sequence disparities between ecotypes were purged from the data. SFPs may alter the hybridization efficiency between cDNAs from one ecotype with probes of another ecotype. In genome scans, five eQTL hot spots were found with significant marker-by-gene linkages. Two of the hot spots coincided with classical QTL conditioning shoot regeneration, suggesting that some of the heritable gene expression changes observed in this study are related to differences in shoot regeneration efficiency between ecotypes. Some of the most significant eQTL, particularly those at the shoot regeneration QTL sites, tended to show *cis*-chromosomal linkages in that the target genes were located at or near markers to which their expression was linked. However, many linkages of lesser significance showed expected "*trans*-effects," whereby a marker affects the expression of a target gene located elsewhere on the genome. Some of these eQTL were significantly linked to numerous genes throughout the genome, suggesting the occurrence of large groups of coregulated genes controlled by single markers.

SHOOTS develop from shoot apical meristems formed during zygotic embryogenesis in plants (Takada and Tasaka 2002; Baurle and Laux 2003). Shoots can also be produced adventitiously or regenerated in tissue culture by organogenesis or through somatic embryogenesis. Shoot regeneration in tissue culture is a means by which plants can be propagated and transgenic plants generated (Preil 2003). In addition, shoot regeneration in tissue culture makes possible the study of shoot development under controlled conditions.

Shoot regeneration in tissue culture is a trait that often varies between plant species and within a plant species among subspecies, varieties, cultivars, or ecotypes. Variation in shoot regeneration can be troublesome especially when elite lines are difficult to regenerate. Shoot regeneration efficiency is a quantitative trait, and quantitative loci (QTL) associated with variation in shoot regeneration efficiency have been identified in Arabidopsis (Schiantarelli *et al.* 2001) and other plants (Komatsuda *et al.* 1993; Taguchi-Shiobara *et al.* 1997; Holme *et al.* 2004). Shoot regeneration QTL in Arabidopsis were identified in recombinant inbred lines (RIL) that differ in shoot regeneration efficiency (Lall *et al.* 2004). Three significant QTL associated with shoot regeneration efficiency were found: a major QTL on chromosome 5 in which the superior allele was derived from the parent of the Columbia ecotype and two minor loci on chromosomes 1 and 4 in which the Landsberg *erecta* ecotype parent contributed the superior alleles (Lall *et al.* 2004). Because superior alleles are distributed between the two parents, the recombinant inbred (RI) population exhibits transgressive segregation of the shoot regeneration trait in that some lines are more or less robust shoot regenerators than either parent (Lall *et al.* 2004).

Gene expression profiling during shoot regeneration in Arabidopsis has revealed a complex gene expression program with hundreds of significant expression changes (Che *et al.* 2002). The most significant components of variation contributing to the overall pattern of gene expression changes during shoot development are waves of genes that turn on at one developmental stage and off at the next (Che *et al.* 2002). One stage with significant gene expression changes occurs about the time of shoot commitment when shoot formation in root explants becomes independent of added plant hormones (Cary *et al.* 2002) and when an abundance of genes encoding transcription factors and signaling components are upregulated (Che *et al.* 2002).

An effort has been undertaken by others to explore the genetic control of gene expression programs in a variety of organisms, such as yeast, maize, mouse, rat, and humans, by combining microarray and QTL analyses (Brem *et al.* 2002; Schadt *et al.* 2003; Bystrykh

[1]*Corresponding author:* Plant Sciences Institute, 1073 Roy J. Carver Co-Laboratory, Plant Sciences Institute, Iowa State University, Ames, IA 50011. E-mail: shh@iastate.edu

et al. 2005; Chesler *et al.* 2005; Hubner *et al.* 2005). In doing so, gene expression levels are considered as metric traits and genetic linkages between genes with heritable expression levels and so-called expression QTL (eQTL) have been sought (Jansen and Nap 2001; Doerge 2002; Jansen 2003). Both *cis*- and *trans*-acting eQTL have been described for regulatory loci that either do or do not colocalize with the regulated genes (Brem *et al.* 2002; Schadt *et al.* 2003; Bystrykh *et al.* 2005; Chesler *et al.* 2005; Hubner *et al.* 2005). True *cis*-acting eQTL are thought to represent genes with polymorphisms that affect their own expression (Schadt *et al.* 2003). In addition, genome scans conducted in populations segregating for heritable gene expression variation in these organisms have revealed eQTL hot spots. Such hot spots are thought to represent key regulatory loci controlling multiple transcripts—hundreds of transcripts, as in the case of mouse brain gene expression (Chesler *et al.* 2005). In several cases analyzed so far, some eQTL with multiple linkages tended to locate at classical QTL associated with traits segregating in the population under study (Schadt *et al.* 2003; Hubner *et al.* 2005).

In this study, we scanned the Arabidopsis genome for eQTL that control gene expression at the time of shoot commitment. We attempted to distinguish *cis*- from *trans*-chromosomal effects and to determine whether the eQTL were coincident with classical QTL associated with shoot regeneration.

MATERIALS AND METHODS

Plant materials and tissue culture procedures: Thirty recombinant inbred lines generated from a cross of the *Arabidopsis thaliana* (L.) ecotypes Landsberg *erecta* × Columbia (L*er* × Col), were used in this study (Lister and Dean 1993). The RI lines were chosen as having the greatest number of recombination breakpoints across the genome. Seeds were obtained from the Arabidopsis Biological Resource Center. Shoots were regenerated in tissue culture through a two-step regeneration procedure according to Valvekens *et al.* (1988) as described in Lall *et al.* (2004).

RNA extraction and DNA chip analysis: Plant material for RNA extraction was collected 6 days after transferring root segments to shoot induction medium (Lall *et al.* 2004). Root explants from several hundred seedlings in each line (1 g total) were pooled for RNA extraction. RNA extraction and hybridization to Affymetrix ATH1 (Affymetrix, Santa Clara, CA) oligonucleotide arrays were carried out as described in Che *et al.* (2002) except the GeneChip Scanner 3000 was used for scanning the chips and the image data generated from the scans was converted to numerical data using the GeneChip operating system v 1.0 (Affymetrix).

Single-feature-polymorphism-affected probe pair removal and gene expression: The gene chip scans provided probe intensity readings for all probe pairs in the 22,810 probe sets on the Affymetrix ATH1 array (Affymetrix 2002). Of the 22,775 noncontrol probe sets, almost all (99.78%) are composed of 11 probe pairs. The few remaining noncontrol probe sets are composed of 8, 9, or 10 probe pairs. We removed any probe pair from the data set identified as a SFP according to Borevitz *et al.* (2003) where the reference ecotype (Col) hybridized with significantly greater intensity than the L*er* ecotype [false-discovery rate (FDR) < 8%]. Using data and scripts from Borevitz (http://www.naturalvariation.org/methods), R software (http://www.r-project.org), and the affy package available through Bioconductor (http://www.bioconductor.org), we investigated the presence of single-feature polymorphisms (SFPs) with a higher intensity in the L*er* ecotype. These SFPs were not removed because we found them to be much less prevalent and of lesser effect than their counterparts, making them more difficult to detect at a high level of confidence. Removal of probe pairs was accomplished by defining an alternative chip description file (.cdf) environment using the altcdfenvs package (http://www.bioconductor.org) and scripts available in the supplemental materials at http://www.genetics.org/supplemental/. After SFP-affected probe pair removal, we were left with 22,787 probe sets (35 of which were control probe sets) because 23 probe sets had all probe pairs removed. Using the remaining probe pairs in each probe set, we computed the MAS 5.0 signal intensities for all 22,787 probe sets via the affy bioconductor package. The MAS 5.0 values were logged and mean centered for each of the 30 oligonucleotide arrays as a normalization procedure so that expression measures would be comparable across slides. Raw .cel files are available at the Plant Expression Database website (http://www.barleybase.org/plexdb/html/index.php).

Data analysis comparing expression and shoot regeneration phenotype: The shoot regeneration phenotype was determined from the number of shoots per root explant using data and methods described by Lall *et al.* (2004). Briefly, mixed linear model analysis of shoot counts on the square-root scale recommended by Anscombe (1948) was used to obtain measures of shoot regeneration efficiency for each line.

At each gene (probe set), we tested the null hypothesis of no correlation between expression and phenotype using Spearman's rank correlation coefficient, a nonparametric test. For sample size >10, the null distribution of the test statistic can be approximated by a t distribution with $n - 2$ d.f. (degrees of freedom). The resulting P-value was used to determine the significance of the relationship. For various thresholds, we estimated a FDR using the method described by Storey and Tibshirani (2003). For example, the threshold P-value of 8.83×10^{-5} was associated with a significance set containing 20 genes and a FDR of 7.8%.

Data analysis identifying eQTL: A filtered set of 288 markers positioned approximately every 2 cM was chosen to represent the genome. The majority of these markers were nonredundant in that at least 1 of the 30 lines had a recombination event between consecutive markers at 240 of the possible 287 consecutive-marker pairings. Markers with a Col allele were coded as 1 and markers with a L*er* allele were coded as 0. Missing genotypes were replaced with the estimated probability of a Col allele based on flanking markers. For every marker-by-gene combination, a least-squares linear regression was fitted using the coded genotype as the independent variable and expression as the dependent variable. The P-value associated with testing the hypothesis of slope equal to zero was used to determine the significance of the relationship. Using a linear regression with the coded genotypes instead of a two-sample t-test with the original genotype groupings to evaluate the strength of the relationship allowed us to include genotype information from all 30 lines in every test. The linear regression is equivalent to the two-sample t-test when no genotypes are missing.

Under the required assumptions for a t-test, we could apply a Bonferroni adjustment to control the genomewide error rate for each trait (gene expression) at the 0.05 level by choosing a threshold P-value of 1.7361×10^{-4}, but this adjustment does

not account for multiple testing over 22,787 probe sets. To estimate an error rate for the full experiment, we instead used a permutation approach to investigate the FDR associated with various significance thresholds for marker-by-gene linkages. We created 1000 permuted data sets by permuting the RIL labels on the 30 microarrays. For each of five *P*-value thresholds coinciding with Bonferroni genomewide error rates between 0.0035 and 0.05 (*i.e.*, five *P*-value thresholds between 1.2153×10^{-5} and 1.7361×10^{-4}), we counted the number of significant marker-by-gene linkages in the set of 6.56×10^6 tests for each permuted data set. A FDR for each threshold was estimated by applying the method of Storey and Tibshirani (2001) using the region from 0.99 to 1.00 to estimate the proportion of true null hypotheses π_0 (estimated $\pi_0 = 60{,}748/65{,}063.01 \approx 0.934$). The FDRs for these thresholds ranged from 2.3 to 10.2%. The density map of significant linkages was shown for the two extreme thresholds, and thresholds between these values showed density maps with similar patterns.

TABLE 1

Frequency of probe sets with SFP-free probe pairs

No. of SFP-free probe pairs	Frequency in 22,775 probe sets[a]
11	16,047
10	3,321
9	1,483
8	755
7	473
6	236
5	157
4	94
3	90
2	54
1	42
0	23

[a] Does not include 35 control probe sets.

RESULTS

Single-feature polymorphisms: We used RI lines derived from a cross between the two standard ecotypes, Col × L*er* (Lister and Dean 1993), as a mapping population to identify eQTL. A problem with RI lines derived from two different ecotypes is the presence of SFPs, small sequence differences between the ecotypes (Borevitz *et al.* 2003). SFPs have the potential to confound oligonucleotide chip analysis because differences in hybridization efficiency caused by SFPs can be interpreted as differences in gene expression levels.

To circumvent this problem, we utilized information from Borevitz *et al.* (2003) (http://www.naturalvariation.org/sfp) to identify and eliminate from our data analysis SFP-affected probe pairs (a probe pair being a set of two probes with one a perfect match to the gene sequence and the other a mismatch) where the reference ecotype (Col) hybridized with greater intensity than the L*er* ecotype. Removing these probe pairs dispenses with the problem of hybridization artifacts due to ecotype-specific sequence differences within probe pairs. We did not, however, address other, less frequent issues such as possible hybridization differences due to ecotypic differences in gene copy numbers. Of 22,775 noncontrol probe sets on the Affymetrix ATH1 chip, 16,047 had no significant SFPs, and at the other extreme, 23 had significant SFPs in all probe pairs of the probe set (Table 1). The elimination of SFP-affected probe pair data from our data set is not without some impact on the reliability of our data. However, Gautier *et al.* (2004) suggested that the number of probe pairs needed for reliable gene expression is probably <11, the number used by Affymetrix for most noncontrol probes, but the minimal number is not known. To check the sensitivity of our results to the inclusion of less reliable probe sets, we repeated our analysis after removing all probe sets with fewer than five probe pairs. All aspects of the results showed very little change. For example, the 34 genes found to be linked to the major shoot regeneration QTL remained the same, and only one gene containing more than five SFPs was removed from the list of 100 genes showing the most evidence of association between expression and phenotype.

Gene expression pattern signatures: Thirty of the most informative Arabidopsis RI lines (with the most recombinant breakpoints) from Lister and Dean (1993) were analyzed in this study. RNA was extracted from root explants that had been preincubated on callus induction medium for 4 days, transferred to shoot induction medium, and incubated for 6 more days. There are no obvious morphological differences between the two ecotypes at this stage, which precedes shoot emergence. However, about this time, root explants become "committed" to shoot development; that is, they continue to form shoots even when transferred to basal medium without hormones (Cary *et al.* 2002). This stage is characterized by abundant expression changes in genes encoding transcription factors and signaling pathway components (Che *et al.* 2002). Labeled cRNAs were generated from the 30 RNA samples and hybridized to Affymetrix Arabidopsis gene chips with 22,810 genes (probe sets).

Genes were identified with expression patterns that significantly correlate with the shoot regeneration efficiency phenotype. Expression levels for individual genes were plotted against a shoot regeneration phenotype computed from the number of shoots per root explant for each RI line. The genes with the strongest correlation between gene expression and shoot regeneration phenotype at a FDR) of 7.8% were identified (Table 2). At5g48330, a putative regulator of chromosome condensation (cell cycle regulatory protein), showed the most significant correlation between gene expression and shoot regeneration phenotype (*t*-value = 6.6912) (Figure 1). Other genes with high correlation between gene expression and shoot regeneration phenotype include those encoding a VAMP membrane

TABLE 2

Genes with significant relationship between expression and shoot regeneration phenotype and an assessment of their association with the major shoot regeneration QTL

Gene	Probe	*t*-value[a]	Gene function[b]	Linkage *P*-value[c]
At5g48330	248693_at	6.6912	Regulator of chromosome condensation family protein	0.0004187
At4g17870	254705_at	−5.9651	Expressed protein	0.0463669
At5g49840	248575_at	−5.9250	Clp protease	0.0000004
At5g47180	248796_at	−5.6123	VAMP membrane protein	0.0291548
At1g22910	257413_at	−5.5869	RNA recognition motif protein	0.0074835
At1g44800	261335_at	5.4033	Nodulin MtN21 family protein	0.0063793
At3g60390	251374_at	5.1479	Leucine zipper protein	0.0242114
At3g03150	258845_at	5.1256	Expressed protein	0.0000358
At5g54970	248139_at	5.1098	Expressed protein	0.0003867
At2g35605	266641_at	5.0940	SWIB complex BAF60b domain-containing protein	0.0018222
At5g48360	248696_at	5.0192	Formin homology 2 domain-containing protein	0.0002446
At5g47760	248780_at	−4.9368	Putative 4-nitrophenylphosphatase	0.0172426
At5g45650	248961_at	−4.8329	Subtilisin-like protease	0.0003827
At4g39860	252821_at	4.8270	Expressed protein	0.0001439
At5g56970	247956_at	4.7951	Cytokinin oxidase family protein	0.0000326
At5g53850	248234_at	−4.7576	Haloacid dehalogenase-like hydrolase family protein	0.0000049
At2g45510	267500_s_at	−4.6979	Cytochrome P450	0.0125426
At2g42840	263979_at	4.6615	Protodermal factor 1 (L1 layer protein)	0.0034488
At5g46510	248847_at	4.6337	Leucine-rich-repeat class protein	0.0010187
At5g58350	247819_at	−4.5761	WNK family protein kinase	0.0000059

[a] Using Storey and Tibshirani's (2001) false discovery method, column represents genes associated with a 7.8% FDR.
[b] Gene functions according to TAIR.
[c] *P*-value for testing eQTL linkage to major shoot regeneration QTL site (marker 270).

protein, RNA recognition motif protein, SWIB complex BAF60b domain-containing protein, two proteases (Clp and subtilisin-like), protodermal factor (located in the L1 embryonic layer), and so forth.

Genome scan: To identify loci controlling heritable gene expression patterns, such as those described above, the Arabidopsis genome was scanned for marker-by-gene expression linkages. Expression signals (corrected for SFP-affected probe pairs) in the 30 RIL data set were treated as quantitative traits and subjected to linkage analysis using a filtered set of 288 markers that were uniformly positioned about every 2 cM. In doing so, a test statistic evaluating each marker-by-gene association was computed. The resulting 6.56×10^6 *P*-values were subjected to a significance threshold and the proportion of significant linkages was plotted across the Arabidopsis genome (Figure 2). Using a permutation approach, a FDR was estimated for various significance thresholds coinciding with genomewide error rates for a single trait (gene expression) between 0.0035 and 0.05. These thresholds equate to comparison-wise *P*-values of 1.2153×10^{-5} and 1.7361×10^{-4}. The more stringent threshold is associated with 3525 significant linkages distributed over 958 genes and a FDR of 2.3%. The less-stringent threshold is associated with 10,521 significant linkages distributed over 2637 genes and a FDR of 10.2%. Due to correlation in markers, many of these linkages can be considered redundant in that they represent single marker-by-gene linkages. Limiting each gene to link to only one marker (chosen as the marker with the highest test statistic) would remove these redundancies, but the peak of a marker regulating a large number of genes (Figure 2) may appear falsely low due to a "spreading-out" of its regulated genes to nearby markers. [For purposes of illustrating where target genes were located on the genome relative to their linked markers, we did remove these redundancies later (see Figure 6) to produce a clearer plot while maintaining the existing relationship.]

At all threshold levels, the genome scan of markers with significant linkages revealed peaks with higher densities of significant linkages (Figure 2). These peaks are similar to linkage hot spots found in the yeast or mouse genomes (Brem *et al.* 2002; Schadt *et al.* 2003; Chesler *et al.* 2005). It was of interest that two of the hot spots corresponded to two of three shoot regeneration QTL identified in a prior study (Lall *et al.* 2004). The major shoot regeneration QTL is located on the lower arm of chromosome 5, centered on marker 270, and one of two minor QTL is located on chromosome 4 and centered on marker 190 (Figure 2). At the threshold associated with a FDR of 2.3%, marker 270 links to 34 genes: 23 genes were upregulated in association with the Col allele at the marker site, and 11 were downregulated (Table 3). A sampling of the single marker-by-gene associations at this site (selected for genes upregulated in association with the Col allele) clearly shows that the genes are expressed at a much higher level when Col

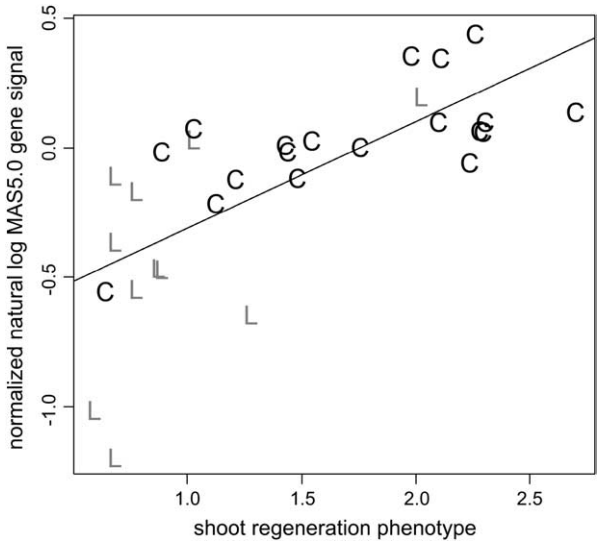

FIGURE 1.—Plot of gene expression *vs.* shoot regeneration phenotype (shoots per explant on square-root scale) for At5g48330 in the 30 RI lines used in this study. At5g48330, encoding a regulator of chromosome condensation family protein, shows the strongest relationship between gene expression and phenotype. Linear regression line and parental allele at the shoot regeneration QTL are indicated by C (Col) or L (L*er*). Expression values represent MAS 5.0 signals that were logged and mean centered for each gene chip. A negative expression reflects a gene with a logged MAS 5.0 value below the average logged expression of genes on a gene chip.

alleles, rather than the L*er* alleles, are present at the marker site (Figure 3, A–C). Similar plots for genes downregulated in association with the Col allele show lower levels of expression when Col alleles are present (Figure 3D). In general, genes with a strong relationship between expression and shoot regeneration phenotype also showed a strong association with the major shoot regeneration QTL (Table 2), although declaration of linkage significance depends on the chosen threshold.

Could the linkage hot spots be artifacts due to allele frequency differences across the genome? To examine this, the proportion of Col and L*er* at each of the 288 markers was estimated. Known allele genotypes were coded as the probability of a Col allele, and unknown allele genotypes were also coded as the probability of a Col allele based on flanking markers. The estimated proportion of Col alleles was based on the expected count of such alleles and was computed as the average of the coded values at each marker. This estimate reduces to a straightforward proportion when all genotypes are known. The out-of-balance group size or proportion of alleles in the larger group (Col or L*er*) was plotted across the genome (Figure 4A). If equal numbers of Col and L*er* alleles were present at each marker, then a flat line centered on 0.5 would be expected. However, allele imbalances were observed across the genome for this set of RI lines. The comparison-wise power of detecting a 1.5 standard deviation difference in the average gene expression for the Col and L*er* groups at each marker was then calculated on the basis of a type I error rate of 0.05 (Figure 4B). By fixing the difference in the means and the type I error rate, power becomes dependent only on the sample size of the two groups and on the accuracy of genotyping. Power is greatest when all genotypes are known and there are equal numbers in each group (15 in each group for this data set). In plotting out-of-balance group size and power across the genome, it can be seen that there are regions, especially in chromosome 1, where there are out-of-balance group sizes and low power markers.

However, the question is whether the linkage hot spots derive from markers with a high power of detection. It can be seen in scans of the genome that clusters of significant linkages and marker power do not necessarily align (Figure 4, B and C). This becomes clearer when comparing numbers of significant linkages to power of detection on a chromosome-by-chromosome basis. None of the chromosomes show a strong linear association between power at marker and number of significant associations (data not shown).

***Cis*- and *trans*-chromosomal effects:** We did not expect to find that the most significant marker-by-gene associations at the FDR of 2.3% involved linkages to genes located in the region of the markers. For example, all of the significant associations with marker 44 were linked to genes in the vicinity of the marker (Figure 5). Of the 23 upregulated genes linked to marker 270 at the shoot regeneration QTL site, all but 1 were located in the region of the major shoot regeneration QTL itself, and of the 11 downregulated genes, all were located near the QTL site. Thus, some of

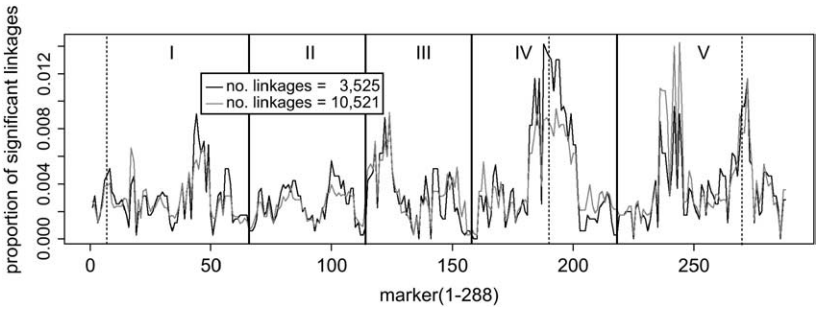

FIGURE 2.—Density map of significant linkages in a genomewide scan. The Arabidopsis genome was scanned for marker-by-gene associations for 288 evenly spaced markers. Scans were conducted at different thresholds as described in the text. Results from two thresholds and the corresponding number of significant linkages are shown here. Plot is corrected by elimination of data for SFP-affected probe pairs.

TABLE 3

Genes with the strongest linkage to marker 270 up- or downregulated in association with the presence of the Col allele at the marker site

AGI name	SFP[a]	SNP[b]
Genes upregulated in association with the Col allele		
At2g18540		
At5g47800	×	
At5g47940	×	
At5g48240		
At5g50410		×
At5g50550		
At5g50560		
At5g50565		
At5g50570		
At5g50580		
At5g50630		
At5g51670		
At5g51820		
At5g51960	×	
At5g51980	×	
At5g53000	×	
At5g53050	×	
At5g53070		
At5g53120	×	×
At5g53950	×	
At5g58730		
At5g59140		
At5g59290		
Genes downregulated in association with the Col allele		
At5g48110	×	
At5g49840	×	×
At5g50230		
At5g51390		
At5g52540		
At5g53360		
At5g53370		
At5g53420		
At5g53760		
At5g53850		
At5g58350	×	

[a] Single-feature polymorphism is a sequence polymorphism detected as a difference in hybridization intensity of randomly labeled genomic DNA to a single feature on a high-density oligonucleotide array (Borevitz *et al.* 2003).

[b] Single-nucleotide polymorphism is a single-base polymorphism usually detected through DNA sequencing analysis.

the most significant linkages were "neighborhood effects" in which the marker was in the vicinity of genes to which the marker was expression linked.

Cis-effects have been reported in yeast and mouse, and markers linked to genes in *cis* have more significant associations than those in *trans* (Brem *et al.* 2002; Schadt *et al.* 2003). In these systems, *cis*-effects appear to have a simple genetic explanation in that they result from polymorphisms that affect the expression level of the genes in which the polymorphisms occur (Brem *et al.* 2002; Schadt *et al.* 2003). In this study, we eliminated from consideration SFPs that may give rise by artifact to apparent expression differences. These are SFPs within the probe sets that may alter the hybridization efficiency of cDNA made from the RNA of one ecotype to the probe of another ecotype. However, genes purged of SFPs in their probe sets are still included in our data (only the probe sets composed completely of SFPs have been eliminated).

Therefore, we asked whether SFPs and *cis*-effects of the kind described in yeast and mice account for the observed variation in expression in the 23 upregulated and 11 downregulated genes linked to the major shoot regeneration marker 270. Of 23 upregulated genes linked to marker 270, 9 had reported single-nucleotide polymorphisms (SNPs) or SFPs (Table 3). [A genome-wide set of SNP markers distinguishing Col and L*er* ecotypes is available through SeqViewer at The Arabidopsis Information Resource (TAIR) http://www.arabidopsis.org/servlets/sv.] In comparing Col or L*er*, the remaining 14 upregulated genes had no reported SNPs or SFPs. Of the 11 downregulated genes, 3 had reported SNPs or SFPs, and 8 did not. Even in the target genes with SFPs or SNPs, there is a low probability that any polymorphism affects gene expression. Thus, the neighborhood effects associated with marker 270 are probably not attributed to "*cis*-effects" of the sort proposed in yeast and mice, *i.e.*, polymorphisms in the target gene (Brem *et al.* 2002; Schadt *et al.* 2003).

It is conceivable that the neighborhood effects that we observe might not result from polymorphisms in genes acting on their own expression, but from polymorphisms in nearby genes. Such neighborhood effects might be revealed in higher-resolution studies involving more RI lines. Nonetheless, it does appear that several genes within a neighborhood are commonly controlled in our study. For example, the four genes most strongly associated with the major shoot regeneration QTL site are located within 37 kb of each other and the correlation in expression between any two of the four genes is positive and >0.81, suggesting that these genes might be commonly controlled.

Significant marker-by-gene linkages across the genome were also illustrated by plotting the position of markers against the position of their corresponding linked genes. Specifically, each gene in the list of significant linkages was plotted against its best controlling marker for the two thresholds associated with FDRs of 2.3 and 10.2% (Figure 6). As reported in previous studies (Schadt *et al.* 2003; Bystrykh *et al.* 2005), we found that the most significant marker-by-gene linkages tended to represent *cis*-chromosomal effects (a marker affects the expression of a target gene in close proximity) as seen in the many solid dots falling along the diagonal in Figure 6. As the number of significant linkages increases, more dots appear off the diagonal, suggesting that *trans*-chromosomal effects (a marker affects the

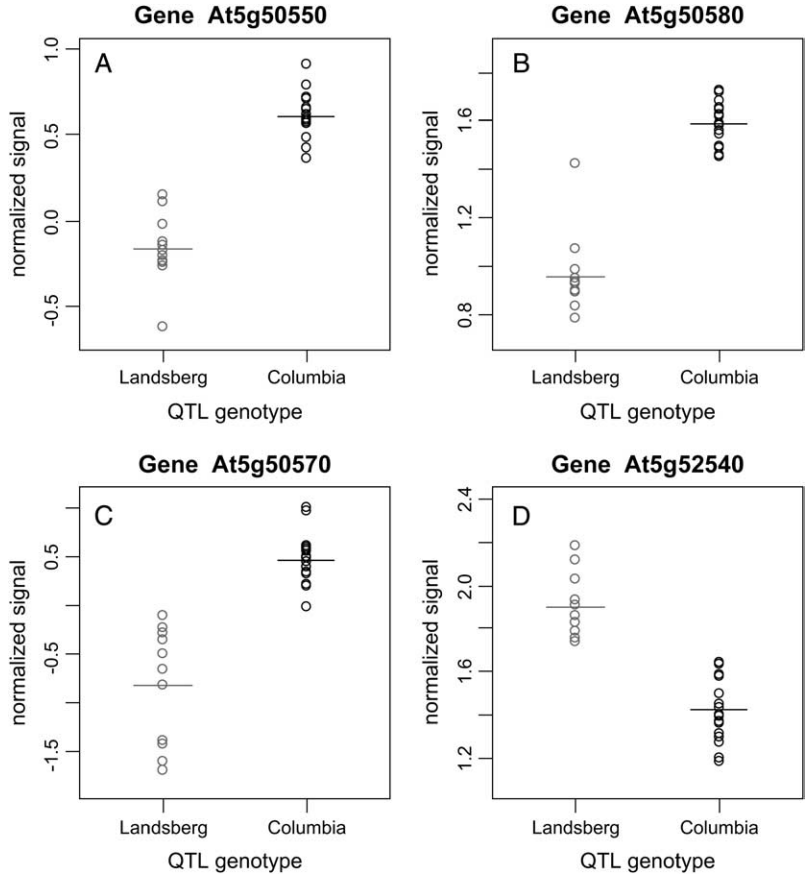

FIGURE 3.—Examples of the effect of the genotype at marker 270 on the expression of various genes to which the marker is significantly linked. Expression levels in the 30 RI lines are grouped according to the presence of the Landsberg *erecta* or Columbia allele at marker 270. Horizontal lines represent QTL genotype group means.

expression of a target gene elsewhere on the genome) are present, but not as strong as the *cis*-effects.

Also apparent in the plot are dense vertical bands indicating regions of the genome that regulate a large number of genes in *trans* during shoot development (Figure 6). Two of these regions are located on chromosome 5 and coincide with previously described hot spots for the 3525 most significant linkages (Figure 2). Genes linked in *trans* to these hot spots show no obvious pattern in that the targeted genes are scattered throughout the genome. The *trans*-chromosomal hot spot located on the lower arm of chromosome 5 is correlated with the major shoot regeneration QTL. This finding was expected because this QTL likely controls many genes throughout the genome associated with shoot regeneration. However, *trans*-effects were not concentrated at the chromosome 4 hot spot described above, suggesting that most of the strong linkages in this region associated with a minor shoot regeneration QTL are *cis*-effects.

DISCUSSION

In scanning the Arabidopsis genome for eQTL associated with heritable changes in gene expression during shoot development, it was found that significant marker-by-gene linkages tended to cluster in hot spots as they do in the yeast or mouse genomes (BREM *et al.* 2002; SCHADT *et al.* 2003). Why they tend to do so is not clear. A hot spot could be due to a single gene at the hot spot that influences the expression of many other genes or it could be a cluster of several genes at the hot spot, each of which act on a few genes. In any case, two of the eQTL hot spots coincided with two of the three QTL associated with the efficiency of shoot regeneration—the major shoot regeneration QTL on chromosome 5 and a minor QTL on chromosome 4 (LALL *et al.* 2004). It was expected that eQTL and shoot regeneration QTL might coincide because the gene expression data were collected during the process of shoot regeneration. Furthermore, QTL that condition the efficiency of shoot regeneration undoubtedly affect the expression of many other genes. It was noted in other studies that eQTL hot spots correspond to QTL or sites of marker genes involving a phenotype that segregated in the mapping population (SCHADT *et al.* 2003; HUBNER *et al.* 2005).

However, it was unexpected to find that the markers with the most significant associations are linked to the expression of genes in the same vicinity of the chromosome as the marker. We refer to these effects as "neighborhood effects," and neighborhoods are very large in molecular terms. The genes with the most significant linkages to marker 270 at the shoot regeneration QTL site cover nearly 1.5 Mbp of DNA. Neighborhood effects might be due to mechanisms similar to those that

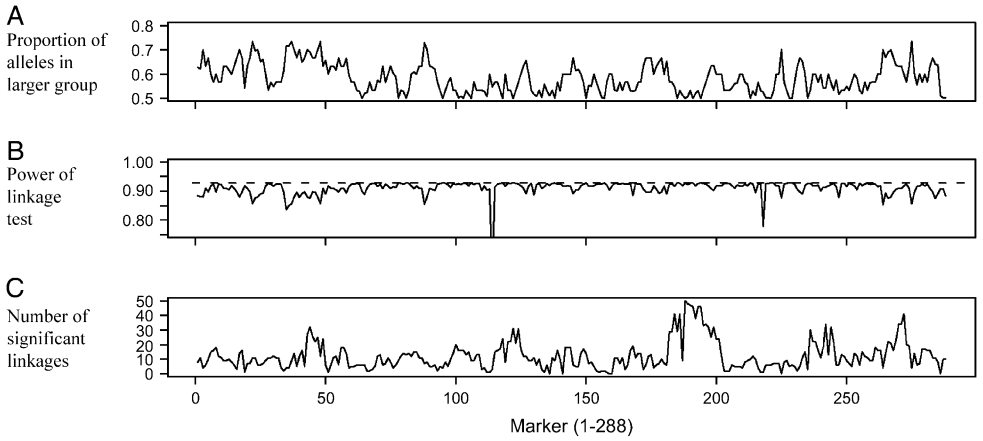

Figure 4.—Allele frequency distribution in a genomewide scan. (A) Out-of-balance group size or the proportion of Col or L*er* alleles in the larger group at each of the 288 markers used in genome scans. (B) The comparison-wise power of detecting a 1.5 standard deviation in the average gene expression at each marker based on a type I error rate of 0.05. Power is greatest when genotypes are known and equal numbers of the two different parental alleles are present at a given marker. The two spikes associated with lower power are due to missing genotypes at the given markers. (C) Distribution of a number of significant linkages at a threshold associated with a FDR of 2.3%.

regulate genes in operons. That, however, seems unlikely, given the large size of the neighborhoods and the distant spacing of some of the affected genes.

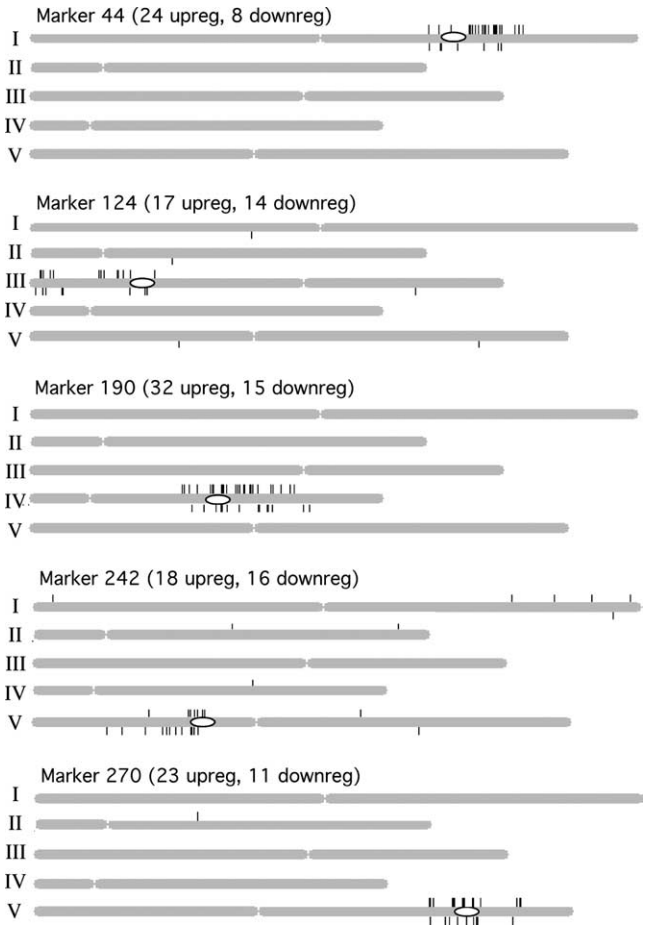

Figure 5.—Location of genes linked to markers at the eQTL hot spots in the Arabidopsis genome at a threshold associated with a FDR of 2.3%. Markers 190 and 270 are centered on the shoot regeneration QTLs. Ticks pointing upward show the location of upregulated genes and downward pointing ticks are downregulated genes. Markers are located about every 2 cM and position of markers are indicated by an oval.

Another possibility is that genes are regulated by epigenetic mechanisms, such as chromatin effects, acting at the chromosome level. Chromatin structure is known to influence gene regulation locally and globally and to specify functional differentiation of chromosomal domains during development in a number of organisms (Weiler and Wakimoto 1995). Chromatin features have been described for some of the Arabidopsis chromosomes on which eQTL hot spots were found (Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center and PE Biosystems Arabidopsis Sequencing Consortium 2000; Lippman *et al.* 2004). In particular, one of the eQTL hot spots in

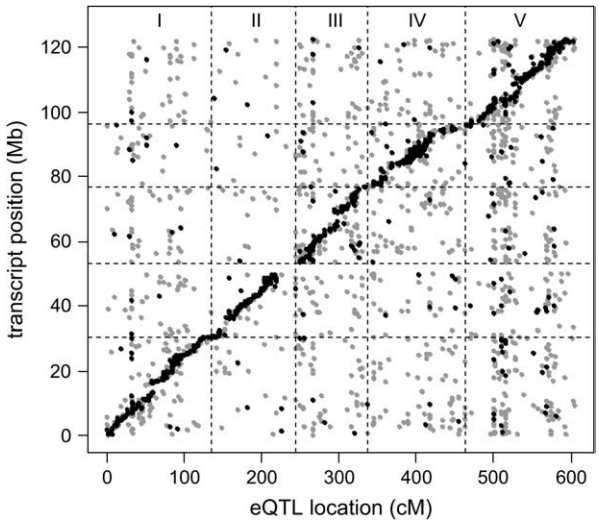

Figure 6.—Significant marker-by-gene linkages plotted as (*x*, *y*) coordinates with the *x*-axis representing the genome location of the marker and the *y*-axis representing the genome location of the linked gene. Each dot represents a single gene plotted against its best controlling marker. Significant linkages at two thresholds (see text) are shown. Solid dots represent significant linkages at a threshold associated with a FDR of 2.3%, and shaded dots represent significant linkages at a less stringent threshold associated with a FDR of 10.2%. Chromosome endpoints are indicated by dashed lines.

our study was located in a euchromatic region on the long arm of chromosome 4, coincident with a shoot regeneration QTL. This chromosome is thought to be organized in euchromatic loops, which emanate from condensed heterochromatic chromacenters (Fransz *et al.* 2002). The chromacenters are composed of heterochromatin from pericentric and nucleolus organizing regions (NOR) and the loops extending from these chromacenters are estimated to be 0.2–2.0 Mbp (Fransz *et al.* 2002), consistent with the dimensions of the neighborhood effects. It is possible that some of the *cis*-chromosomal gene regulation effects that we see on chromosome 4 may involve chromatin or chromosome loops. It would be of interest to compare chromatin structure around some of the up- or downregulated genes in the parental ecotypes.

Two other unexpected findings in our study are also of note. One is that the eQTL hot spot on chromosome 4 is the site of a minor shoot regeneration QTL that has significant epistatic effects on the major shoot regeneration QTL on chromosome 5 (Weiler and Wakimoto 1995). Given this interaction, one might expect significant *trans*-chromosomal linkages between markers and target genes at the QTL on chromosomes 4 and 5. Such linkages were found for the significance thresholds that we investigated at the QTL on chromosome 5, but not at the QTL on chromosome 4. Another unexpected finding was the presence of eQTL hot spots at sites other than the QTL sites. The eQTL hot spot on the upper arm of chromosome 5 is particularly prominent and is not associated with a shoot regeneration QTL. This must mean that although the locus is associated with ecotype-specific gene expression changes, those expression changes have little impact on shoot regeneration.

This study would not have been possible without information on SFPs. Probe-set SFPs resulting in a higher hybridization affinity for the Col ecotype had the potential to alter this analysis significantly. Performing similar analyses on the data before and after SFP probe pair removal allowed us to determine the impact of SFPs on our results. After SFP removal, 34 genes were significantly linked to marker 270 in the eQTL analysis (Table 3). Using the same *P*-value threshold applied to the data before SFP removal, we found 40 significant linkages to marker 270. The number of genes significantly upregulated in Col decreased from 31 to 23 after SFP removal, while the number significantly downregulated in Col increased from 9 to 11. This implies many apparent strong marker-by-gene expression linkages were due to SFP probes. All genes that were eliminated from the significance list contained at least one SFP probe pair and most contained numerous SFPs. The genes that were dropped may still have strong relationships with marker 270, but the relationship was not strong enough to be considered significant at the given level once SFPs were removed. No genes were dropped from the downregulated gene list, but 2 genes that contained SFPs in the original data were added. Although many of the same genes appeared on both significance lists (before and after SFP removal), the fact that some did not suggests that it is important to consider SFPs in data analysis. Study results often guide ongoing research and the removal of SFP probe pairs in data analysis may help researchers avoid inefficient use of resources.

The genetic basis for the major shoot regeneration QTL on chromosome 5 has not yet been determined; however, a number of candidate genes are under study. It will be interesting to know whether the genetic entity that conditions shoot regeneration at this site is also responsible for controlling the target genes in the eQTL analysis.

This work was supported by the National Science Foundation (IBN-0236060 and DMS-0091953) and by the Plant Sciences Institute at Iowa State University.

LITERATURE CITED

Affymetrix, 2002 *Affymetrix Microarray Suite User Guide.* Affymetrix, Santa Clara, CA.

Anscombe, F. J., 1948 The transformation of Poisson, binomial and negative-binomial data. Biometrika **35:** 246–254.

Baurle, I., and T. Laux, 2003 Apical meristems: the plant's fountain of youth. BioEssays **25:** 961–970.

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., G. Yvert, R. Clinton and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752–755.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher *et al.*, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics.' Nat. Genet. **37:** 225–232.

Cary, A. J., P. Che and S. H. Howell, 2002 Developmental events and shoot meristem gene expression patterns during shoot development in *Arabidopsis thaliana.* Plant J. **32:** 867–877.

Che, P., D. J. Gingerich, S. Lall and S. H. Howell, 2002 Global and cytokinin-related gene expression changes during shoot development in Arabidopsis. Plant Cell **14:** 2771–2785.

Chesler, E. J., L. Lu, S. Shou, Y. Qu, J. Gu *et al.*, 2005 Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat. Genet. **37:** 233–242.

Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center and PE Biosystems Arabidopsis Sequencing Consortium, 2000 The complete sequence of a heterochromatic island from a higher eukaryote. Cell **100:** 377–386.

Doerge, R. W., 2002 Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. **3:** 43–52.

Fransz, P., J. H. De Jong, M. Lysak, M. R. Castiglione and I. Schubert, 2002 Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. Proc. Natl. Acad. Sci. USA **99:** 14584–14589.

Gautier, L., M. Moller, L. Friis-Hansen and S. Knudsen, 2004 Alternative mapping of probes to genes for Affymetrix chips. BMC Bioinformatics **5:** 111.

Holme, I. B., A. M. Torp, L. N. Hansen and S. B. Andersen, 2004 Quantitative trait loci affecting plant regeneration from protoplasts of Brassica oleracea. Theor. Appl. Genet. **108:** 1513–1520.

Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz *et al.*, 2005 Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet. **37:** 243–253.

Jansen, R. C., 2003 Studying complex biological systems using multifactorial perturbation. Nat. Rev. Genet. **4:** 145–151.

Jansen, R. C., and J. Nap, 2001 Genetical genomics: the added value from segregation. Trends Genet. **17:** 388–391.

Komatsuda, T., T. Annaka and S. Oka, 1993 Genetic mapping of quantitative trait loci (QTLs) that enhance the shoot differentiation rate in *Hordeum valgare L.* Theor. Appl. Genet. **86:** 713–720.

Lall, S., D. Nettleton, R. DeCook, P. Che and S. H. Howell, 2004 Quantitative trait loci associated with adventitious shoot formation in tissue culture and the program of shoot development in Arabidopsis. Genetics **167:** 1883–1892.

Lippman, Z., A. V. Gendrel, M. Black, M. W. Vaughn, N. Dedhia *et al.*, 2004 Role of transposable elements in heterochromatin and epigenetic control. Nature **430:** 471–476.

Lister, C., and C. Dean, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana.* Plant J. **4:** 745–750.

Preil, W., 2003 Micropropagation of ornamental plants, pp. 115–133 in *Plant Tissue Culture: 100 Years Since Gottlieb Haberlandt,* edited by M. Laimer and W. Ruecker. Springer-Verlag, Berlin.

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.

Schiantarelli, E., A. de la Pena and M. Candela, 2001 Use of recombinant inbred lines (RILs) to identify, locate and map major genes and quantitative trait loci involved with *in vitro* regeneration ability in *Arabidopsis thaliana.* Theor. Appl. Genet. **102:** 335–341.

Storey, J., and R. Tibshirani, 2001 Estimating false discovery rates under dependence. Technical Report 2001–28. Department of Statistics, Stanford University, Palo Alto, CA.

Storey, J., and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA **100:** 9440–9445.

Taguchi-Shiobara, F., S. Y. Lin, K. Tanno, T. Komatsuda, M. Yano *et al.*, 1997 Mapping quantitative trait loci associated with the regeneration ability of seed callus in rice, *Oryza sativa* L. Theor. Appl. Genet. **95:** 828–833.

Takada, S., and M. Tasaka, 2002 Embryonic shoot apical meristem formation in higher plants. J. Plant Res. **115:** 411–417.

Valvekens, D., M. Van Montagu and M. V. Lijsebettens, 1988 *Agrobacterium tumefaciens*-mediated transformation of *Arabidopsis thaliana* root explants by using kanamycin selection. Proc. Natl. Acad. Sci. USA **85:** 5536–5540.

Weiler, K. S., and B. T. Wakimoto, 1995 Heterochromatin and gene expression in Drosophila. Annu. Rev. Genet. **29:** 577–605.

Communicating editor: D. Weigel