Deeply Diverged Alleles in the *Arabidopsis* AREB1 Transcription Factor Drive Genome-Wide Differences in Transcriptional Response to the Environment

David L. Des Marais,*†,† Wesley D. Skillern,1 and Thomas E. Juenger*†

1Department of Integrative Biology and Institute for Cellular and Molecular Biology, The University of Texas at Austin

†Present address: Arnold Arboretum and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA

*Corresponding author: E-mail: desmarais@fas.harvard.edu; tjuenger@austin.texas.edu.

Abstract

Gene regulatory variation is an important driver of the evolution of physiological and developmental responses to the environment. The abscisic acid (ABA) signaling pathway has long been studied as a key component of the cellular response to abiotic stresses in plants. We identify two haplotypes in an *Arabidopsis thaliana* transcription factor, AREB1, which plays a central role in ABA-mediated response to osmotic stress. These two haplotypes show the sequence signature of long-term maintenance of genetic diversity, suggesting a role for a diversifying selection process such as balancing selection. We find that the two haplotypes, distinguished by a large number of single nucleotide polymorphisms and the presence or absence of four small insertion/deletions in AREB1 intron 1 and exon 2, are at roughly equal frequencies in *Arabidopsis*, and show high linkage disequilibrium and deep sequence divergence. We use a transgenic approach, along with mRNA Sequencing-based assay of genome-wide expression levels, and find considerable functional divergence between alleles representing the two haplotype groups. Specifically, we find that, under benign soil–water conditions, transgenic lines containing different AREB1 alleles differ in the expression of a large number of genes associated with pathogen response. There are relatively modest gene expression differences between the two transgenic lines under restricted soil water content. Our finding of pathogen-related activity expands the known roles of AREB1 in *A. thaliana* and reveals the molecular basis of gene-by-environment interaction in a putatively adaptive plant regulatory protein.

Key words: gene-by-environment interaction, gene expression, signaling network, abscisic acid, balancing selection.

Introduction

Natural populations are characterized by abundant segregating genetic variation in gene expression phenotypes. Although debate remains as to the relative importance of regulatory versus protein-coding mechanisms in driving phenotypic divergence within and between populations (Hoekstra and Coyne 2007; Carroll 2008), it is clear that gene expression is an important component of evolutionary change (Wray 2007; Stern and Orgogozo 2008). Biologically relevant variation in gene expression among individuals can be attributed to differences in their constitutive gene expression, to differences in the genomic or external environment, and to genetic differences in response to the environment (Gibson 2008; Hodgins-Davis and Townsend 2009; Des Marais et al. 2012). The molecular genetic variation in response to environmental cues is termed gene by environment interaction (GEI). GEI in higher-order phenotypes such as fitness has been studied for decades and is recognized as a necessary component for the evolution of phenotypically plastic responses (Bradshaw 1965). GEI can shape the fitness of populations across a species range and therefore likely plays an important role in the evolution of locally adapted genotypes (Hereford 2009). An important goal for molecular and evolutionary genetics, then, is to identify what sorts of gene regulatory variants segregate in natural populations, how these variants interact with the environment, and whether and how they are shaped by evolutionary processes such as selection and neutral genetic drift.

Plants have evolved many strategies which allow them to adjust their physiological and developmental states to match the unpredictable environment in which they grow. In many cases, natural diversity in these physiological and developmental responses is attributed to differences in gene expression, referred to as expression GEI (eGEI; Hannah et al. 2006; Des Marais et al. 2012; Rengel et al. 2012). The molecular variants that drive the patterns of expression variation across different environments are often unknown. Linkage mapping of environmental response expression phenotypes (expression Quantitative Trait Loci; eQTL) can distinguish between cis-acting (local) and trans-acting (distant) regulatory control (e.g., Smith and Kruglyak 2008; Lowry et al. 2013). This strategy has limitations; for example, because the mapping resolution of eQTL analyses is often coarse and complex epistatic interactions can complicate the detection of small-effect mutations. Moreover, because eQTL studies are
generally limited to a small number of genotypes, such studies cannot directly assess the frequency of mapped variants in natural populations. An alternate approach for studying the evolutionary cause and consequence of eGEI is to contrast directly the allelic effects of candidate genes in known environmentally responsive regulatory networks (Des Marais et al. 2013).

The plant hormone abscisic acid (ABA) plays diverse roles in both intra- and intercellular signaling response to exogenous and developmental cues (Finkelstein and Rock 2002). During periods of osmotic stress (e.g., low soil water availability, reduced relative humidity, or high salinity), ABA induces numerous cellular and whole-plant physiological and developmental changes (Chaves et al. 2003). Principal among these ABA-induced changes are alteration of growth patterns and biomass partitioning, stomatal closure to reduce transpiration and water loss, and biosynthesis of metabolites to maintain cellular osmotic status and protect cellular components from further damage. Growing evidence also points to a role for ABA signaling in response to pathogenic stress, suggesting a complex network of interaction between ABA and other plant hormones; transcription factors are likely candidates for a focus of such signal integration (Adie et al. 2007; Ton et al. 2009). The molecular mechanisms of ABA signal transduction in plants are well-understood, from the biosynthesis and mobilization of ABA itself, to the cellular perception of ABA, through a complex set of cellular components which involve both pre- and posttranslational regulatory dynamics (Cutler et al. 2010).

Exogenous application of ABA results in considerable alteration of the genome’s transcriptional landscape (Seki et al. 2002). Early work on ABA regulation identified a conserved sequence motif, the ABA-responsive element (ABRE), in the proximal promoters of ABA- and osmotic stress-responsive genes in wheat (Guitian et al. 1990), rice (Mundy et al. 1990), and Arabidopsis thaliana (hereafter “Arabidopsis”; Yamaguchi-Shinozaki and Shinozaki 1994). ABRE motifs are found in the proximal promoters (1,000 bp from the transcription initiation site) of an estimated 2,500 genes in the reference genome of Arabidopsis (Geisler et al. 2006). ABREs promote transcriptional activation through recruitment of bZIP transcription factors in the ABRE element binding (AREB; also called ABRE binding factor, ABF), gene family (Choi et al. 2000; Uno et al. 2000). The first functional work on this gene family focused on the AREB1 protein, which requires ABA-mediated phosphorylation for full protein activity (Kim et al. 2004; Fujita et al. 2005; Fujita, Nakashima, et al. 2009). Subsequent work has revealed complex signaling relationships between three AREB family members, their upstream activators, and their downstream transcriptional targets (Furihata et al. 2006; Yoshida et al. 2010). This AREB-ABRE transcriptional network plays an essential role in putatively adaptive response to myriad environmental cues across diverse plant species (Yamaguchi-Shinozaki and Shinozaki 2006), and so natural variation in this network could affect GEI of thousands of genes and their associated phenotypes.

In this study, we demonstrate a role for natural genetic diversity in the AREB1 transcription factor in affecting environment-responsive transcriptional diversity. We identify two intermediate-frequency haplotype variants at AREB1, characterized by the presence or absence of several small sequence motifs in the AREB1 intronic and coding sequence. These haplotypes bare the sequence signature of balanced polymorphism. We further show that these alternate transcription factor variants drive differences in gene regulatory network response to environmental cues and we generate hypotheses for new functional roles for AREB1. Our results support the hypothesis that variation at AREB1 is a molecular driver of GEI in Arabidopsis and may play an important role in environmental adaptation.

Results

Natural Diversity at AREB1

In order to infer the past evolutionary history of the AREB1 gene, we first characterize sequence diversity in AREB1 among natural accessions of Arabidopsis. We identified two haplotype groups at AREB1 among 16 Sanger-generated sequences, distinguished by the insertion of four short DNA motifs as compared with the reference (Col-0) genome sequence (fig. 1). Two of these insertions, of 16 and 18 bp in length, are in intron 1 of the AREB1 gene. We refer to these as insertions 1 and 2, respectively. A third insertion of 9 bp is found in the beginning of exon 2, which is annotated as 5’s-untranslated region (UTR) in the reference genome. A fourth insertion results in five extra amino acids after the sixth amino acid position in the Col-0 sequence. Finally, a glycine to alanine substitution at amino acid position 20 (using Col-0 numbering) is in nearly perfect association with the four insertions. We refer to the haplotypes defined by these differences as “Col-type” (lacking insertions, as in the Col-0 reference genome sequence) and “Ws2-type” (containing the insertions, with the Ws-2 accession as a representative example of this condition). We used a cleaved amplified polymorphic sequence assay to estimate the frequency of these two haplotypes in natural Arabidopsis populations. Among 238 sampled accessions from a broad geographic distribution, we found a cleavage pattern consistent with the Col-type haplotype in 124 accessions and cleavage consistent with the Ws2-type haplotype in 114 accessions. These frequencies are remarkably close to equal proportions of the two allele classes (52.1% Col-type: 48% Ws2-type).

We next asked whether the two haplotypes show associations with life history patterns in Arabidopsis, are associated with climatic variation, or are structured spatially across the native range of Arabidopsis. Arabidopsis generally displays an annual life history strategy, wherein diverse internal and environmental cues including day length, light quality, temperature, and moisture regulate the transition from vegetative to reproductive growth (Bailey and Dean 2006). Many natural accessions require prolonged periods of cold (vernalization) to stimulate this transition while other accession flower readily without such stimuli. We find no association between vernalization requirement, as scored by Lasky et al. (2012), and AREB1 haplotype (likelihood ratio test, $\chi^2 = 2.36; P = 0.12$). One single nucleotide polymorphism (SNP) from
the panel of 1307 accessions genotyped by Kim et al. (2007) falls in intron 1 of AREB1 (at position 17167636 of TAIR10 chromosome 1); this SNP shows perfect association with AREB1 haplotype in our sequencing panel and is used in the following two analyses. We find no association between AREB1 haplotypes and climate parameters based on the analysis by Lasky et al. (2014). There is no apparent geographical structure to the distribution of AREB1 haplotypes, as assessed by Fst metrics estimated by Horton et al. (2012). We cannot exclude the hypothesis, however, that the AREB1 haplotypes show differential distribution at very small local geographic and climatic scales.

We investigated patterns of nucleotide diversity in a DNA alignment spanning approximately 2,500 bp upstream and downstream of the in/dels observed in AREB1. We used Sanger-generated sequences and Illumina sequences from a de novo-assembled genome resequencing project (Gan et al. 2011) for this and subsequent analyses because we found that many of the accessions available through community resequencing projects did not accurately call SNPs adjacent to the small in/dels in AREB1 (see Materials and Methods). Nucleotide diversity through exon 1 and intron 1 within each of the two haplotype groups is quite moderate (fig. 2A, broken lines; the apparent span of low diversity located upstream of AREB1 is due to a sequence alignment gap introduced by transposable element which is missing in some of our sampled accession.) In contrast, sequence divergence between the two haplotype groups is highly elevated through intron 1 (fig. 2A, solid line).

To test the null hypothesis that the AREB1 exon 1 and intron 1 sequences are evolving neutrally, we compared the Tajima’s $D$ statistic to a sample of 876 Sanger-sequenced regions of the genome, generated previously (Nordborg et al. 2005). Tajima’s $D$ in AREB1 exon 1 and intron 1 is 2.46. Due to the demographic history of Arabidopsis, Tajima’s $D$ values are skewed toward values less than zero and nonnormally distributed; this pattern makes a simple test of a significant deviation from zero an unreliable test of neutrally evolving sequences. We therefore used a bootstrap resampling procedure to create a null distribution from the 876 sequenced regions, and find that Tajima’s $D$ in AREB1 is outside the upper 95% confidence interval [CI] for this sample ($P = 0.0127$; see supplementary fig. S1, Supplementary Material online). Significantly elevated values of $D$ are classically interpreted as evidence for the long-term maintenance of sequence diversity, for example, via balancing selection or spatially structured populations.

We compared within-species diversity through this region to sequence divergence from A. lyrata using the sliding $G$-statistic of McDonald (McDonald 1998). The sliding $G$ test asks whether “runs” of sites with segregating polymorphisms exceed runs of divergence polymorphisms (or vice versa). We find that several short stretches of sequence through AREB1 intron1 have a elevated proportion of polymorphism relative to divergence from A. lyrata, though none of the associated $G$ test statistics is significant at $\alpha = 0.05$.

**Linkage Disequilibrium around Intron 1 of AREB1**

The elevated diversity which we observed between haplotypes at AREB1 suggests that recombination may be reduced in this region, leading to a pattern of linkage disequilibrium (LD) among SNPs. We investigated the extent of possible LD around AREB1 at two scales. First, we extracted all 60 SNPs within 25 kb of the AREB1 start codon identified in the 1307 accessions genotyped by Kim et al. (2007). The single SNP from this panel that falls in intron 1 of AREB1 (at position 17167636 of TAIR10 chromosome 1) shows no significant association with the other 59 SNPs in the analysis after Bonferonni correction for multiple tests (all $r^2 < 0.1$; supplementary fig. S2, Supplementary Material online). This result suggests that the evolutionary processes causing elevated diversity in intron 1 region are restricted to AREB1 itself. There is, however, a block of significant LD which extends from a SNP located 3′ of the coding region of AREB1 (this is proximal to the centromere of chromosome 1, as AREB1 is encoded on the minus strand) to a SNP located in intron 2 of At1g45229, approximately 6.4 kb downstream of AREB1.

To investigate the pattern of LD in AREB1 more locally, we estimated LD among all 68 unambiguous diallelic SNPs in 34 Sanger-sequenced accessions spanning 2,500 bp on either side of the AREB1 start codon. This analysis reveals high LD between all SNPs occurring in AREB1 exon 1, intron 1, and the proximal part of exon 2 (all $r^2 > 0.6$ and significant at $\alpha = 0.05$ after Bonferonni correction; fig. 2C). We find additional SNPs located downstream of the AREB1 3′-UTR which are in high LD with each other, though not in significant LD with the SNPs identified in the proximal part of AREB1. There is likewise no significant LD between any of the AREB1 intron 1 SNPs and SNPs located upstream of AREB1 (in the intergenic region between AREB1 and At1g45230, or in AT1g45230...
Although the values of LD in the vicinity of AREB1 intron 1 are not unusually high compared with other regions of the genome, the decrease in LD near this interval are striking. Genome wide, LD (measured as \( r^2 \)) in Arabidopsis diminishes by approximately 50% within 3–4 kb (Kim et al. 2007; Horton et al. 2012) whereas around the AREB1 intron 1 LD decays approximately 70% within 3 kb toward the stop codon and 100% in less than 1 kb into the intergenic region.

**Fig. 2.** Partitioning of genetic diversity in the genomic region surrounding AREB1. (A) Nucleotide diversity (\( \pi \)) within each haplotype class, and sequence divergence between the two haplotype classes. (B) Schematic of the gene structure surrounding AREB1 (based on TAIR10 annotations). (C) LD among all sequenced SNPs in the genomic region surrounding AREB1. Black boxes denote significant association between two SNPs at \( \alpha = 0.05 \) after Bonferonni correction. Gray boxes denote correlations (\( r^2 \)) greater than 0.2, but not significantly different from zero correlation at \( \alpha = 0.05 \) after Bonferonni correction.

Historical Relationships of the AREB1 Alleles

The paucity of observed recombination between AREB1 haplotypes could lead to the effectively independent evolution of sequences within each group. Under this scenario, we expect to observe two diverged classes of gene sequences. A gene tree reconstructed from the sequence of AREB1 exon 1, intron 1, and the beginning of exon 2, reveals very deep divergence between the two haplotype classes (fig. 3). (In this analysis, we excluded the inserted sequence motifs characteristic of Ws2-like alleles and reconstructed the gene tree based solely on the SNPs identified throughout this region.) We infer that the two haplotypes originated subsequent to the divergence of A. thaliana from A. lyrata. Among the 63 accessions included in this analysis, we identified two accessions that departed from the canonical Ws2-type or Col-type patterns of in/del polymorphism. The Oy-0 accession has insertion 1 (like Ws2), but otherwise has AREB1 sequence characteristic of the Col-type haplotype. The gene tree analysis places this sequence...
with the Col-type alleles, albeit with low statistical support (bootstrap value of 52%). Conversely, Mt-0 lacks insertion 1, like the other Col-type alleles, but has the other three insertions, similar to the Ws2-type alleles. The gene tree analysis places this sequence as a lineage sister to the Ws2-type alleles.

**The Transcription of Both AREB1 Variants Is Strongly Induced by Soil Water Drying**

Natural variation in the AREB1 transcription factor could affect its own expression and protein function, as well as the transcriptional activity of the many downstream genes which AREB1 is known to regulate (e.g., Fujita et al. 2005). In order to directly measure the functional effects of natural variation in AREB1, we cloned the complete genomic fragment of AREB1 from Col-0, and from Ws-2, as representatives of their respective haplotype classes. These fragments were introduced transgenically into a common genetic background with substantially reduced AREB1 function due to a synthetic T-DNA insertion in the immediate upstream sequence of AREB1 (hereafter “AREB1 knockout”). We then assessed the transcriptional activity of AREB1 itself, along with AREB1 knockout.
with the transcriptional activity of all annotated genes in the Arabidopsis genome (TAIR10 annotation) using mRNA Sequencing on the Illumina HiSeq platform. We grew replicate plants representing multiple independent transgenic insertions with either of the two alleles in two imposed environments: A well-watered treatment and a treatment intended to simulate an end-of-season soil drying by gradually reducing soil water content by approximately 60% (hereafter “drought”). As a transgenic negative control we also scored transcriptional activity in replicate transgenic lines carrying an empty transformation vector in the AREB1 knockout. We performed statistical contrasts in a mixed model framework for specific hypotheses related to differential expression between the genotypes under study.

We find that the empty transformation vector has a negligible effect on gene expression when compared with the common transgenic host line (AREB1 knock-out; supplementary fig. S3A, Supplementary Material online). At False Discovery Rate (FDR) = 0.05, only seven genes (out of 12,700 in the analysis) are differentially expressed between the two lines in the well-watered environment (supplementary table S1A, Supplementary Material online; we did not score expression for these two genotypes in the dry environment). We conclude that the transformation vector had a minimal effect on gene expression in our transgenic experiments.

The expression of AREB1 itself is affected by the drought treatment, but shows nonsignificant differences between alleles (supplementary fig. S4, Supplementary Material online). Both alleles are upregulated under our simulated drought treatment ($F_{1,41} = 9.66, P = 0.0034$), which is consistent with earlier work in the Columbia wild-type background (Uno et al. 2000; Fujita et al. 2005). The Ws-2 allele is more weakly expressed than the Col-0 allele in each environment, though this response difference is not significant (allele × treatment $F_{4,41} = 0.12, P = 0.9736$).

Both AREB1 Haplotypes Are Associated with the Transcription of Genes Involved in Response to Soil Drying

Past work on AREB1 and its paralogs in the Col-0 background showed that the AREB transcription factors play a central role in regulating ABA-mediated transcriptional response to osmotic stress, including drought (Fujita et al. 2005; Yoshida et al. 2010). We were therefore interested in addressing whether the natural genetic variation in AREB1 affects the transcriptional response of the ABA signaling network under restricted soil water availability. We find that both alleles confer strong transcriptional response of target genes to our simulated drought treatment (supplementary fig. S3C and D, Supplementary Material online): 1,877 genes show a significant treatment response in both alleles (FDR = 0.05; see supplementary table S1B, Supplementary Material online, for gene lists). In this context, “response” refers to the difference in transcript abundance between the drought and control treatment which may reflect either up- or downregulation when comparing the two treatments. As expected, the list of responsive transcripts is strongly enriched for genes involved in abiotic stress responses (supplementary table S2A, Supplementary Material online), and the promoters of responsive genes are enriched for sequence motifs commonly associated with abiotic stress responses (supplementary table S2B, Supplementary Material online). One thousand four hundred eighty-three genes show a significant response to the treatment in the Col-0 transgenic lines but not Ws-2 lines, and 952 genes show the reverse pattern. In 87.4% of the genes showing significant response in one accession but not the other accession (2129/2435), the contrasting genes show the same direction of response. This observation suggests that many of the apparent differences in response between the alleles are due to lack of statistical power to detect a treatment effect in one or both of the populations of samples. Accordingly, we found only 37 genes differentially expressed between the alleles when we directly contrasted transcript abundances between transgenic lines in the simulated drought environment (supplementary fig. S3B and table S1C, Supplementary Material online). Of these genes, 12 are more highly expressed in the Ws-2 lines, whereas 25 are more highly expressed in the Col lines. These results suggest that the AREB1 alleles have largely similar functions in the transcriptional regulatory network under the soil drying conditions assayed here.

The AREB1 Haplotypes Drive Distinct Transcriptional Networks under Well-Watered Conditions

We next contrasted allelic effects on genome-wide expression levels in the control (well-watered) environment. We find that 578 genes are differentially expressed between transgenic lines harboring the two AREB1 alleles; 377 of these genes are more highly expressed in the Ws-2 lines and 201 are more highly expressed in the Col-0 lines (FDR = 0.05; see fig. 4 and supplementary table S1D–F, Supplementary Material online). The set of genes more highly expressed in the Ws-2 lines is strongly enriched for genes with functional annotations involving response to external stimuli, specifically defense responses and response to biotic stimuli (tables 1 and 2). We further find that the proximal promoters (−1,000 bp from the transcription start site) for genes in this set are enriched, relative to genomic background, for the ABRE sequence motif (hypergeometric test $P < 10^{-16}$), which is directly involved in mediating transcriptional regulation by AREB proteins (Choi et al. 2000; Uno et al. 2000). This gene set is also enriched for TATA box elements ($P < 10^{-10}$) and the W-box motif ($P < 10^{-2})$. TATA boxes are generally associated with highly inducible genes, and W-box motifs are involved in transcriptional regulation in the presence of pathogens (Yu et al. 2001).

In contrast, the set of genes more highly expressed in Col-0 lines is strongly enriched for genes related to primary metabolism, including biosynthetic processes and genes involved in photosynthesis (tables 1 and 2). The only overrepresented promoter sequence motif in this gene list, as compared with genomic background, is the I-box ($P < 10^{-2})$. The I-box motif is frequently found associated with light-regulated
genes involved in photosynthesis (e.g., Vandepoele et al. 2009).

**Discussion**

Extensive empirical work has shown that gene expression is a key determinant of the physiological and developmental response of organisms to the external environment. This plasticity is particularly important in plant species, which are sessile and must therefore acclimate to the unpredictable environment in which they germinate. We have directly demonstrated the role of allelic diversity in a regulatory gene in determining expression levels of downstream transcriptional targets. Moreover, the pattern of nucleotide diversity in this regulatory gene is consistent with the process of balancing selection maintaining diversity over long periods of time.

We discuss the implications of our results in the context of the evolution of gene regulatory networks and the role played by these networks in GEI in natural populations.

**Balancing Selection in AREB1**

Natural diversity in intron 1 of AREB1 is structured into two intermediate-frequency haplotypes. This pattern of divergent alleles leading to elevated diversity in local regions of the genome is frequently observed in Arabidopsis (Stahl et al. 1999; Aguade 2001; Tian et al. 2002; Kroymann et al. 2003; Mauricio et al. 2003; Shepard and Purugganan 2003; Kroymann and Mitchell-Olds 2005; Bakker et al. 2006; Du et al. 2007, 2008; Caldwell and Michelmore 2009; Reininga et al. 2009; Huard-Chauveau et al. 2013). Very often the selective forces—if they exist—maintaining such elevated diversity are unknown. In other cases the observed sequence diversity is consistent with the long-term maintenance of multiple alleles under the diversifying effects of balancing selection. Classically, balancing selection has been conceived of as a within-population process reflecting two or more fitness optima alternating, for example, in a temporally fluctuating manner (Charlesworth 2006). In practice, using patterns of sequence diversity to distinguish between balancing selection and other kinds of diversifying selection, such as local adaptation, can be challenging if precise boundaries between metapopulations are unknown (Charlesworth et al. 1997). In the absence of such geographical information, an understanding of the possible selective agent(s) favoring the alleles can provide confidence in assessing the role of balancing selection. Due to the highly selfing life-history and resulting high levels of homozygosity in natural populations of A. thaliana,
overdominance is not believed to play a role in maintaining diversity in populations of this species. Most published studies of balancing selection in *A. thaliana* invoke fluctuating occurrence of a pathogen: One allele confers tolerance to the selective agent but is disfavored in that absence of that agent, presumably due to some cost of expressing the trait constitutively (Tian et al. 2002; Mauricio et al. 2003; Karasov et al. 2014).

In the present case, the patterns of sequence diversity around *AREB1* strongly suggest the long-term persistence of two classes of alleles with very little genetic exchange between them. The gene tree of sampled intron 1 sequences from diverse natural accessions show two deeply diverged haplotypes (figs. 2A and 3), there is LD among SNPs in intron 1 (fig. 2C), there is elevated Tajima’s *D* in intron 1, and the two haplotype classes are segregating at nearly equal frequencies. Based on measures of *F*<sub>st</sub> reported previously (Horton et al. 2012), we find no evidence for geographical structure among the *AREB1* alleles, though we cannot rule out the interesting hypothesis that fine-scale geographical variation has structured haplotype diversity.

One equivocal piece of evidence relating to the effects of balancing selection is the ratio of within-species polymorphism to divergence from *A. lyrata*. If selection acts to maintain diversity within *A. thaliana*, we would expect an elevated ratio of within-to-between species diversity; we do not observe this pattern in *AREB1*. One possible explanation for this result is that there is elevated diversity within *A. lyrata* as well, thus inflating observed between-species divergence. A second, more mundane, explanation is that alignment gaps introduced by the multiple segregating in/dels within *A. thaliana*—and relative to *A. lyrata*—prohibit us from accurately counting polymorphic and divergent sites.

A second line of evidence that natural selection is maintaining diversity at *AREB1* is the divergent pattern of gene expression which we observe between transgenic lines expressing alleles representative of the two haplotypes. We find relatively similar gene expression responses of the two alleles to our simulated drought environment: Both alleles are strongly upregulated in response to drought (supplementary fig. S3C and D, Supplementary Material online) and are associated with the expression of many abiotic-stress related genes (supplementary table S2, Supplementary Material online); this suggests that the Ws-2 allele likely functions in a similar manner to the reference Col allele under the soil drying condition which we assayed. In contrast, in the “control” (well-watered) environment we find enrichment of genes related to pathogen response which are expressed at a higher level in lines carrying the Ws-2 *AREB1* allele (tables 1 and 2). A suite of genes involved in primary metabolism are coordinately more lowly expressed in the Ws-2 lines relative to the Col lines, suggesting that the pathogen-associated expression driven by the Ws-2 allele may be associated with slower growth in plants carrying the Ws-2 allele. Karasov et al. (2014) recently demonstrated a selective cost in natural settings for expressing a protein responsible for pathogen response in the absence of the pathogen. They further showed that the presence or absence of this protein in natural

### Table 2. Higher Control Environment Expression in Col-0 lines (197 genes in GO analysis).

<table>
<thead>
<tr>
<th>GO Biological Process</th>
<th>Number of genes in Process</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolic process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biosynthetic process</td>
<td>56</td>
<td>0.0006</td>
</tr>
<tr>
<td>Macromolecule biosynthesis</td>
<td>39</td>
<td>0.002</td>
</tr>
<tr>
<td>Cellular macromolecule biosynthesis*</td>
<td>39</td>
<td>0.002</td>
</tr>
<tr>
<td>Translation*</td>
<td>32</td>
<td>2.00E-12</td>
</tr>
<tr>
<td>Cellular biosynthesis</td>
<td>53</td>
<td>0.0006</td>
</tr>
<tr>
<td>Cellular macromolecule biosynthesis*</td>
<td>39</td>
<td>0.002</td>
</tr>
<tr>
<td>Translation*</td>
<td>32</td>
<td>2.00E-12</td>
</tr>
<tr>
<td>Macromolecule</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>43</td>
<td>0.002</td>
</tr>
<tr>
<td>Cellular</td>
<td>42</td>
<td>0.0005</td>
</tr>
<tr>
<td>Protein folding</td>
<td>8</td>
<td>0.017</td>
</tr>
<tr>
<td>Translation*</td>
<td>32</td>
<td>2.00E-12</td>
</tr>
<tr>
<td>Gene expression</td>
<td>38</td>
<td>0.004</td>
</tr>
<tr>
<td>Cellular metabolic process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generation of precursor metabolites and energy</td>
<td>8</td>
<td>0.025</td>
</tr>
<tr>
<td>Light reaction photosynthesis*</td>
<td>6</td>
<td>0.003</td>
</tr>
<tr>
<td>Photosynthesis</td>
<td>7</td>
<td>0.008</td>
</tr>
<tr>
<td>Light reaction photosynthesis*</td>
<td>6</td>
<td>0.003</td>
</tr>
<tr>
<td>Cellular component biogenesis</td>
<td>14</td>
<td>0.002</td>
</tr>
<tr>
<td>Ribonucleotide complex</td>
<td>9</td>
<td>0.0006</td>
</tr>
<tr>
<td>Ribosome biogenesis</td>
<td>9</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

*Occurs in multiple GO hierarchies.
populations is likely due to diffuse interactions with a community of pathogens across time and space; this pattern of a cost of resistance has been demonstrated in other pathosystems, as well (Korves and Bergelson 2004; Todesco et al. 2010). The apparent dual function of AREB1—regulation of abiotic stress-responsive gene expression under water limitation and, in the case of Ws-2, expression patterns consistent with pathogen resistance and slower primary metabolism under seemingly benign soil water content—suggests that the selective forces maintaining genetic diversity at AREB1 are likewise complex and diffuse. We return to this interesting case of GEI below.

**AREB1 Function**

AREB1/ABF2 and its paralogs, AREB2/ABF4 and AREB3, have been studied extensively in the context of ABA-mediated plant response to osmotic stress and glucose signaling (Uno et al. 2000; Kang et al. 2002; Kim et al. 2004; Fujita et al. 2005; Furhata et al. 2006; Yoshida et al. 2010). All of this research was performed in the Col-0 and Ler backgrounds (Ler has indels and SNPs typical of the Col-type alleles in AREB1 intron 1). In Col-0, the first 60 amino acids of the coding sequence are essential for transcriptional activity in vivo (Fujita et al. 2005). This N-terminal domain contains a conserved phosphorylation site at amino acid residue 23 (numbered in the Col-0 sequence; see fig. 1) suggesting that either insertion 4 of the Ws-2 haplotype or the associated Gly20Ala substitution may affect posttranslational activity of AREB1 protein. These structural variants may also affect interaction between AREB1 and the other AREB proteins, which form heterodimers with one another in vivo (Yoshida et al. 2010). Any of the three insertions in the intron 1 sequence, and any of the associated SNPs, may likewise drive gene expression differences between the alleles, as this region is 5’ of the translational start codon. Indeed, we do observe differences in AREB1 expression between alleles, though these differences are not statistically significant at the time point and tissue (leaf) sampled here. Because the coding and intronic features are in tight LD among natural accessions (fig. 2C), determining which combination of variants is responsible for the observed functional differences will require construction of numerous domain-swapping or site-directed mutation lines, which is beyond the scope of this study. The two accessions containing apparent recombinants between AREB1 haplotypes, Mt-0 and Oy-0, could be used for such an analysis. It should be noted that *A. lyrata* AREB1 sequence has all four insertions, with several single nucleotide variants in the intronic insertions relative to the Ws-2 sequence, but lacks the Gly20Ala substitution characteristic of the insertion haplotype. This finding suggests that the Col-0 haplotype may represent the derived condition of AREB1.

The effect of the two AREB1 haplotypes on genome-wide gene expression could be due either to relative upregulation of target genes in Ws-2 or relative downregulation of target genes in Col-0. Because extensive research either on pathogen signaling or on AREB1 in the Col-0 accession has not identified a role for AREB1 in response to pathogens, we prefer a model in which Ws-2 allele drives pathogen-related gene expression in our control condition, whereas the Col-0 allele does not. Regardless of the evolutionary polarity of this difference, there is a strong and significant difference in function between the alleles. Considerable evidence now points to extensive information “cross-talk” between environmental signaling pathways in plants; ABA-responsive transcription factors appear to serve as key players in this signal integration (Fujita Y, Fujita M, et al. 2009). Further research may reveal precisely how and to what degree the Ws2-type AREB1 participates in ABA-mediated pathogen response signaling.

**GEI and Signaling Network Topology**

More broadly, our work provides a mechanistic explanation for the frequent observation of high covariance in genome-wide gene expression profiles (e.g., Ayroles et al. 2009; Des Marais et al. 2012; Innocenti and Chenoweth 2013). For example, in a recent study of the expression of over 20,000 genes in ten natural accessions of *Arabidopsis*, we found that over 70% of the variance in gene expression could be explained by just five principal components (Des Marais et al. 2012). This high covariance among the expression levels of genes is evidence of a hierarchical arrangement of regulatory factors in gene expression networks. It is now well appreciated that environmental perturbation can rewire genetic networks: The activity of a single regulatory gene can affect the expression of hundreds or thousands of downstream transcripts (Luscombe et al. 2004).

The hierarchical structure of signaling networks has important implications for the evolution of GEIs. Classically, GEI was assumed to be dominated by strong allele-driven trade-offs (antagonistic pleiotropy) between environments. Under this architecture, an allele confers a relatively higher trait value in one environment, and a relatively lower trait value in a second environment compared with an alternative allele. The pathogen-driven balanced polymorphisms in plants, described earlier, are cases of antagonistic pleiotropy: Alleles conferring resistance have higher trait value (in this case, relative fitness) in the presence of a pathogen, while those conferring susceptibility have higher trait values in the absence of a pathogen. Recently, we showed that rank-changing functional effects are detected infrequently among studies of plant interactions with the abiotic environment (Des Marais et al. 2013). Instead, a pattern of conditional functional neutrality is much more commonly identified. Conditional genetic variants are characterized by strong difference in trait expression in one environment, with little or no difference in expression in a second environment.

A recent study linking nucleotide variation in *A. thaliana* accessions with local climate at their place of origin found a genomic signature of conditional neutrality (Fournier-Level et al. 2011). In that study, genetic variants were found to segregate in a selectively neutral fashion across much of the range of *A. thaliana*—presumably because they have little or no functional effect in those environments—but were apparently removed from populations by negative selection where they are deleterious. In the two environments we studied here,
variation in AREB1 appears to represent functional conditional neutrality. In the dry environment, there are few genes showing expression differences under the control of divergent AREB1 alleles while there are strong transcriptional differences in allelic effects in the wet environment. Interestingly, our functional enrichment analysis of gene expression in this wet environment suggests that the mechanism of selection maintaining the haplotypes may involve another environmental variable: Presence of a pathogen. We hypothesize that there may be little selective difference between the haplotypes in the dry environment, but that presence or absence of a pathogen in the wet environment could lead to selective advantage for one haplotype. We suspect that such conditionally neutral alleles play an important role in the evolution of environmentally responsive traits in wide-ranging species such as A. thaliana. Conditionally neutral alleles may drift randomly throughout much of the range and be either purged or maintained by selection when environmental conditions expose their functional differences. Identifying functional variants such as AREB1 is an important first step in quantifying the frequency and mechanisms of such processes.

Materials and Methods
Plant Growth and DNA Extraction
All seed stocks were obtained from the Arabidopsis Biological Resource Center. Names and stock numbers for accessions used in this study are reported in supplementary table S3, Supplementary Material online. To obtain genomic DNA, seeds were cold-stratified in water for 3 days and then sown directly onto Sunshine MGP potting mix (SunGro Horticulture). Plants were grown in growth chambers under long-day conditions (16 h) until rosettes were large enough to harvest for DNA extraction. DNA was extracted using a modified CTAB protocol (Kelly and Willis 1998). Genomic DNA of AREB1 (At1g45249) and surrounding regions were polymerase chain reaction (PCR)-amplified using a series of oligonucleotide primers listed in supplementary table S4, Supplementary Material online. PCR amplicons were visualized by electrophoresis and then cleaned using a 5:1 mixture of Exonuclease 1 and Shrimp Alkaline Phosphatase (New England Biolabs). Sanger sequencing was accomplished using combinations of the primers listed in supplementary table S4, Supplementary Material online, on AB3730 DNA analyzers (Applied Biosystems).

DNA Alignments and Sequence Diversity
All alignments were generated manually in Sequencher (Version 4.6) or Mesquite (Maddison WP and Maddison DR 2007). We compared our Sanger-sequenced genomic fragments with publicly available sequences generated as part of the Arabidopsis 1001 genomes project (Weigel and Mott 2009) and identified many SNPs in our data set, particularly those flanking sequences inserted relative to the reference Columbia sequence, which were not present or were miscalled in the resequencing data. In contrast, the genome sequences of 18 accessions (downloaded from http://signal.salk.edu/atg1001/index.php on 1 March 2014) were assembled by Gan et al. (2011) without relying on a reference genome. Based on overlap with our Sanger sequencing efforts, the genome sequence of these 18 accessions appear to accurately report SNP variants in the in/del-rich regions of AREB1. We combined our de novo Sanger-generated sequences with the Gan et al. sequences to construct two alignments for subsequent analysis. The “long” alignment comprises 6,146 bp of aligned sequence across 34 natural accessions, spanning approximately 2,500 bp upstream and downstream of the AREB1 start codon (accessions denoted “Complete Genomic Fragment” in supplementary table S3, Supplementary Material online). The “deep” alignment comprises 571 bp across 63 accessions, spanning AREB1 exon 1, intron 1, and the proximal part of exon 2 (all accessions in supplementary table S3, Supplementary Material online).

We used DNASP to calculate nucleotide diversity within each haplotype group, and divergence between the two groups, using a sliding window analysis on the long alignment with 100 bp windows and 25 bp steps. We also used the sliding G statistic of McDonald (1998) to test the hypothesis that there is an excess of within-species nucleotide diversity in the long alignment compared with divergence from A. lyrata. This analysis measures the frequency of runs of polymorphic SNPs versus fixed differences in sliding windows along the sequence of interest. Initial analyses of 100 Monte Carlo coalescent simulations implemented in DNASlider across a range of recombination parameters identified the recombination parameter of 8 as the most conservative for setting a statistical threshold of significance for the G-statistic. To test for significant excess (or deficit) of polymorphism relative to divergence we then performed 10,000 replicates of the coalescent simulation.

We used DNASP to calculate the Tajima’s D statistic using the deep alignment, based on the total number of mutations. The population history of A. thaliana causes most regions of the genome to show strongly negative values of Tajima’s D, making a simple test for neutrality at values which differ from zero to be unreliable. To test for departure from neutrality in the AREB1 intron 1 region, we resampled the data set of Nordborg et al. (2005), which comprises approximately 500–600 bp of sequence from 876 regions across the Arabidopsis genome. To create a null distribution of values of D, we performed 10,000 random bootstrap samples, and estimated 95% CI from the resulting distribution.

Estimating LD around AREB1
We investigated patterns of LD around AREB1 at two genomic scales. First, we extracted the 60 SNPs within 25 kb of the start site of AREB1 from a panel of 1307 accessions genotyped at 214,051 SNPs (Kim et al. 2007). We used TASSEL version 4.0 (Bradbury et al. 2007) to estimate the correlation between all pairs of SNPs and to calculate significant deviation from the null hypothesis of no correlation using a two-tailed Fisher’s Exact test. Because of the very large number of tests being made, we corrected for multiple comparisons using a strict Bonferroni correction. The one SNP from this panel found in intron 1 of AREB1 is in very weak association ($r^2 < 0.1$) with
other SNPs. In order to assess the patterns of LD in the immediate genomic vicinity of AREB1, we used TASSEL to estimate LD in the long alignment and corrected for multiple tests, as above.

Gene Tree of AREB1 Alleles

To estimate the historical relationships among the alleles identified in the intron 1 region of AREB1, we estimated a gene tree using the deep alignment. We estimated the Hasegawa, Kishino and Yano (HKY) model of sequence evolution, with a distribution of rate classes and estimates empirically, as the most likely model for these data using jModelTest. We then estimated a gene tree using the Neighbor joining criterion as implemented in PAUP (ver 4.0). Branch support was estimated from 1,000 bootstrap replicates each with 100 random taxon additions and Tree Bisect and Reconnection (TBR) branch swapping.

Gene Expression Response of Alleles

We PCR-amplified the complete genomic fragment of AREB1, spanning the 3'-UTR of At1g45249 through the 3'-UTR of At1g45230, from Col-0 and Ws-2. The Ws-2 genomic sequence has four small insertions in AREB1 intron 1 and exon 2 relative to the Col-0 reference sequence, along with many associated SNPs. PCR amplicons were cloned into the Gateway LR reaction (Invitrogen) and then recombinated via the Gateway LR reaction (Invitrogen) into a pMDC162 plant binary vector. We used Agrobacterium-mediated floral-dip transformation (Clough and Bent 1998) to introduce the resulting binary plasmids into the AREB1 KO line (SALK_138855C). SALK_138855C contains a synthetic T-DNA insertion in exon 1 of AREB1 in the Col-0 background; this insertion lowers the inducibility of AREB1 expression under drought (supplementary fig. S4, Supplementary Material online). Transgenic plants were advanced to the T3 generation and then used in an experiment to measure gene expression effects of the two alleles. We also created an empty vector control transgenic line. All subsequent experiments employed three transgenic lines (representing independent DNA insertion events of each construct) for both the Col-0 and Ws-2 constructs, and two transgenic lines for the empty vector control.

We grew three experimental replicates of the Col-0 transgenics, Ws-2 transgenics, Col-0, SALK_138855C, and the vector control in a randomized block design in growth chambers. Seeds were sown directly onto potting media as described previously (Des Marais et al. 2012) and grown under short days (10 h light/14 h dark) at 24°C, with 18°C nights. On day 21 following germination, all pots were brought to field capacity for water content. Control plants were subsequently watered each morning to field capacity. Starting on day 22, water was withheld from drought treatment plants, although each morning the pots were weighed and water readded to ensure that all pots dried down at a constant rate (5–10% soil water content per day). On day 30, drought-treatment pots were estimated to contain 35% of soil water content relative to field capacity; this level of soil water drying has been in used in past studies of drought physiology and gene expression response in Arabidopsis (Des Marais et al. 2012). At harvest, plants were carefully inspected for signs of senescence or pathogen response; no aberrant leaves were identified. Rosettes from both treatments were excised within the first hour after lights came on in the morning; two leaves were flash-frozen on liquid nitrogen and stored at −80°C for RNA extraction. Two additional leaves were reserved to measure leaf relative water content (RWC) as described previously (Des Marais et al. 2012). The dry-down resulted in a significant decrease in leaf RWC by approximately 20% (Analysis of Variance (ANOVA) F_{1,38.8} = 39.8, P < 0.0001). In total there were 51 plants across ten genotypes: Three transgenic replicates for each AREB1 allele, Col-0, SALK_138855C, and two transgenic replicates of the empty vector control. We imposed two treatments on all plants, with no dry treatment imposed on SALK_138855C and the transgenic empty vector. There were three biological replicates of each genotype by treatment combination described.

RNA was extracted using Spectrum Plant Total RNA Kits (Sigma) according to the manufacturer’s specifications. We also performed an on-column DNase treatment (Qiagen). To estimate genome-wide transcript abundances, each RNA sample was processed for short-read sequencing using an mRNA tag-seq approach developed by Meyer et al. (2011). We adapted the Meyer et al. protocol for use on the Illumina platform by substituting the oligonucleotide primers reported in their protocol with Illumina-specific primers listed in supplementary table S5, Supplementary Material online. Additional details on the library preparation protocol can be found at http://www.bio.utexas.edu/research/matz_lab/matzlab/Methods.html (last accessed October 2013). In brief, this protocol enriches the cDNA pool for 300–500 bp fragments at the 3'-end of transcripts, and therefore focuses sequencing effort on a smaller portion of transcripts than traditional total-RNA Illumina preparations. The protocol has been used previously to measure plant responses to soil drying, and estimated read counts show very high correspondence with estimates of transcript abundance from Quantitative/Real-Time PCR (Q/RT-PCR) (Meyer et al. 2014). Fifty-one prepared, barcoded, libraries were quantified on a Qubit platform (Invitrogen) and split into two pools. Each pool was loaded into a single lane of an Illumina HiSeq 2500 analyzer. We recovered between 2 and 10 million raw single-end 100 bp reads per sample. Raw reads were trimmed and filtered based on read quality, and then mapped to the reference Col-0 genome sequence (version TAIR10) using BWA Mem (Li and Durbin 2009). In total, we identified 19,207 unique gene models in our data set. Prior to statistical analysis, we removed gene models with very low sampling (where > 40% of samples had zero counts; this filtering removed 6,507 genes) and accounted for variation among samples in library sequencing depth using KDM normalization implemented in JMPGenomics 6.1 (SAS Institute). The final data set used in statistical analysis therefore included 12,700 gene models. Note that measures of differential expression using mRNASequencing counts are often corrected for the
length of sampled transcripts, to account for differences in target sample size during library preparation. We elected not to perform such a correction because a size-selection step during our library preparation restricts sampling to fragments approximately 300–500 bp, all from the 3’-end of transcripts. All sequencing targets are therefore of similar size and should not bias final transcript counts.

We used a mixed model approach to test the hypothesis that the two natural alleles at AREB1 affect the gene expression networks of plants under different environmental conditions. We fit a model in JMPGenomics with Genotype (Col allele,Ws-2 allele, knock-out, empty vector control) and treatment (control or dry soil) as fixed effects, and transgenic replicate as a random effect nested within genotype. Incorporating this random effect allowed us to account for variance in expression due to random insertion of the transgene in the genome of the host genotype. The data were modeled as a negative binomial distribution, and we employed a log-link function in the mixed model. The observed fit of this distribution to the data, estimated for each genes as $\chi^2$ divided by the degrees of freedom, was very good and centered around 1. We then performed specific contrasts to test several hypotheses at an experiment-wide FDR of 0.05. We first asked whether the empty vector control transgenic line differed from the host genotype, the AREB1 knock-out SALK_138855C. We identified seven genes (out of 12,700 in the analysis) which differ in expression between these genotypes in the control environment. Based on this result, we concluded that the transformation vector did not have an appreciable effect on gene expression in our experiments. We next contrasted the expression levels of the two natural alleles in the control environment and the soil-drying environment. We also contrasted the treatment response (expression level in dry minus expression level in control environment) of each allele.

Functional Analysis of Expression Responses

We used the method of Zhou et al. (2010), implemented in the online program AgriGO (http://bioinfo.cau.edu.cn/agriGO/analysis.php, last accessed October, 2013), to test the hypothesis that the two natural alleles at AREB1 affect the expression of distinct functional classes of genes under the two environments. This gene enrichment tool uses Fisher’s exact test, with FDR = 0.05, to test the hypothesis that a query set of genes is overenriched in particular Gene Ontology (GO) category terms relative to a control set of genes. We used a custom gene set, including only the genes for which we detected transcripts in our experiments, as the control set in these analyses. We also tested the hypothesis that the two natural alleles at AREB1 affect the expression of genes that have particular sequence motifs in their proximal promoter regions. These tests were performed using the ATHENA software (O’Connor et al. 2005), at a conservative significance cutoff of $\alpha < 10^{-4}$.

Supplementary Material

Supplementary material is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Chakri Gava, Michelle Hartnett, Emeline Sukamtoh, and Tierney Logan for assistance with data collection. Kyle Hernandez developed the bioinformatics pipeline for processing raw RNASequencing reads. Jesse Lasky and Dan Bolnick gave excellent advice on statistical analyses. Members of the Juenger Lab, Robin Hopkins, and Federico Roda provided feedback on the manuscript. Jim Richards, John K. McKay, and Saunak Sen provided stimulating discussion and guidance during the early stages of the project. This work was supported by the National Science Foundation (IOS-0922457 to T.E.J., DEB-0618347 to T.E.J.) and a United States Department of Agriculture NIFA-AFRI postdoctoral fellowship (2011-67012-30663 to D.L.D.).

References


