Extensive cis-Regulatory Variation Robust to Environmental Perturbation in Arabidopsis

Francisco A. Cubillos, Oliver Stegle, Cécile Grondin, Matthieu Canut, Sébastien Tisné, Isabelle Gy, and Olivier Loudet

INTRODUCTION

Quantitative traits vary as a function of the genetic makeup of the individual and environmental factors. Large-scale linkage mapping and genome-wide association studies have uncovered an abundance of associations between specific genetic loci and phenotypes, explaining substantial proportions of the variability in a trait (Alonso-Blanco et al., 2009; Trontin et al., 2011). Ultimately, the phenotypic response to such quantitative trait loci is mediated at the molecular level, motivating efforts to investigate the genetic regulation of transcriptional modules and their effect on complex traits (Cubillos et al., 2012b). However, the inherent differences in statistical power to detect cis-acting versus trans-acting regulators preclude an objective assessment of their relative contributions to expression variability. First, each gene has many more genome-wide variants to be considered when testing for trans effects; moreover, for any individual gene, several trans factors likely act together (including indirectly), which renders an explicit identification of all causal variants difficult. Addressing these challenges is important, as accurate quantification of the cis and trans components of gene expression traits can yield new insights into the genetic architecture of molecular traits and their interactions in networks (Kliebenstein, 2009).
The extent to which cis- and trans-acting factors respond to environmental perturbations in plants is still unclear. Genes can coordinate react to environmental stimuli altering the cis and trans response (Kliebenstein, 2009). Evidence in human cohorts suggests that cis-eQTLs are predominantly robust with respect to cellular environments such as tissue type, where genetic effects on transcription were rarely found to be specific to an environment (Grundberg et al., 2011). Similarly, local putative cis eQTLs in mice and Arabidopsis thaliana recombinant inbred lines were largely conserved between different treatments (Ziebarth et al., 2012; Lowry et al., 2013). However, owing to the statistical limitations mentioned above, most studies have focused on putative cis effects and, hence, a balanced assessment of how regulatory variants, both in cis and trans, respond to environmental changes is lacking. In a single environmental background, the utilization of F1 hybrids has become an established tool to estimate allele-specific expression (ASE), which allows direct quantification of cis-regulatory effects. First, this approach provides unambiguous proof that proximal cis-acting polymorphisms cause expression differences. The principle for establishing the extent of cis variation in F1 hybrids is based on the fact that trans-acting variants will affect both parental alleles equally, hence cancelling their respective contribution, while cis effects will remain allele-specific and stand out. In the F1 hybrids, the trans component of genetic effects can be assessed by contrasting the F1 hybrid to pools of parental RNA with balanced contributions from both parents. In expression estimates from these pooled samples, the trans effect is not cancelled and therefore can be accounted for (as one factor summing all individual trans effects and potentially some more complex features), allowing the dissection of cis- and trans-regulatory variation for every transcript (Wittkopp et al., 2004; McManus et al., 2010). The deep assessment of ASE in different species has revealed wide divergence within regulatory regions between and within species ranging from fruitflies (Drosophila melanogaster; McManus et al., 2010), Arabidopsis (Zhang and Borevitz, 2009), yeast (Emerson et al., 2010; Skelly et al., 2011; Martin et al., 2012), and mouse (Goncalves et al., 2012) to maize (Zea mays; Paschold et al., 2012). Consequently, integration of these data into a single model allows estimation of and statistically testing for the contribution of cis and trans factors (Goncalves et al., 2012).

Here, to determine the extent to which genotype and environmental factors determine allelic expression differences in A. thaliana, we extended the principles of ASE analysis to study genotype-environment interactions. We developed a sensitive approach to estimate the contribution of cis- and trans-regulation altering transcript abundance between divergent accessions Columbia-0 (Col-0) and Cape Verde Island (Cvi-0), grown under contrasting environments. RNA-seq-based expression estimates were used to quantify genome-wide patterns of ASE in F1 hybrids and their parental plants. This approach allowed for high-resolution eQTL mapping despite a small sample, allowing us to detect more than 4700 genes exhibiting ASE. Moreover, the estimation of transcript abundance in two conditions enabled the assessment of additive environmental effects as well as the extent to which cis and trans genetic effects respond to this external perturbation (genotype x environment [GxE]). Differences between conditions could largely be attributed to ASE magnitude changes (cis x E) and trans-interactions due to changes in the direction of the effect between conditions. Overall, these results reveal the prevalence of ASE and natural variation in expression levels between two divergent accessions, providing direct evidence of the transcriptional effects of allelic changes between environments and the robust response of cis-acting genetic regulators across conditions.

RESULTS

Cvi-0 Resequencing

Accurate estimation of allele-specific expression depends on a detailed map of segregating alleles between the parental lines. To this end, we resequenced the Col-0 and Cvi-0 genomes at ~23× coverage to detect single nucleotide polymorphisms (SNPs) and polymorphic regions that could allow for the quantification of ASE. We surveyed the variation in SNPs, gene content, and insertions and deletions (indels). Genome-wide, we identified 726,112 high-quality SNPs (see Methods) and 116,129 nucleotide indels in the Cvi-0 nuclear genome with respect to the Col-0 reference (Supplemental Data Set 1), highlighting the genetic distance of these accessions in comparison to previous large-scale surveys of natural A. thaliana accessions (Cao et al., 2011). Many of these polymorphisms were found in intergenic regions, which is in line with previous variation studies (Gan et al., 2011; Schneeberger et al., 2011). Within coding regions, we found more exonic SNPs than intronic SNPs (250,040 versus 207,324), which stands in striking contrast with indels, which were substantially more prevalent in introns (5-fold enrichment; Figure 1). Overall, SNP variants were found in exonic regions of 26,698 protein-coding genes, covering 79.4% of the annotated elements in the TAIR10 A. thaliana genome. In order to validate our SNP calling in Cvi-0, we performed Sanger sequencing of 29 genes previously identified as local-eQTLs for the same cross.

Figure 1. Distribution of Polymorphisms from the Assembly of the Cvi-0 Genome.

Classification and percentage of SNPs and indels based on the Col-0 reference genome.
(Cubillos et al., 2012b). This orthogonal technology identified a total of 155 SNPs, which were in complete agreement with the genotype calling based on the Illumina short read data (no false-positive SNP call), suggesting sufficient accuracy of our SNP-calling pipeline to reliably estimate ASE. In addition, we resequenced the Col-0 accession used in this study to a similar coverage as Cvi-0. Interestingly, our Col-0 line differed from the reference TAIR10 sequence at 711 unambiguously identified SNPs (Supplemental Data Set 1) with high confidence. From this set, 33.3% represent SNPs within coding exon regions; however, only 25% of them are nonsynonymous changes (affecting 59 genes).

**Quantification of ASE in Parental Pools and F1 Hybrids**

To study the effect of the environment, as well as cis- and trans-acting factors, on gene expression, we examined the extent of expression changes across environments in F1 hybrids and parental pools between Col-0 and Cvi-0 accessions in *A. thaliana*. Both the hybrids and the parental strains were grown in two environmental conditions: well-watered (four F1 individuals, three Col-0 x Cvi-0 and one Cvi-0 x Col-0) and water-deficit (three F1 individuals, Col-0 x Cvi-0) conditions (hereafter referred to as control and drought stress conditions, respectively; see Methods and Figure 2). For every F1 individual, RNA-seq generated Illumina single-end reads of 100 bp from cDNA libraries. Likewise, to account for trans variation, we considered four parental pools between accessions (two pools for each growth condition), which were sequenced using equal amounts of RNA from each parent (Figure 2) (Wittkopp et al., 2004; McManus et al., 2010). Incorporating the parental pools enabled us to (1) discriminate expression divergence due to segregating purely cis-acting variants (observed in F1 hybrids), trans-acting factors (summed altogether), or a combination of both genetic effects (Figure 3A) (Wittkopp et al., 2004) and (2) estimate the response of these factors to environmental perturbations. Here, trans effects include distal genetic effects, interactions between cis- and trans-acting variants, dominance effects, and higher order genetic effects that cannot be explained by the cis haplotype.

RNA-seq reads for each sample were aligned to the Col-0 reference genome, using a variant-aware mapping approach to avoid biases in the mapping procedure. Reads that spanned segregating variants between Col-0 and Cvi-0 were annotated as originating from one of the two haplotypes. The genetic distance between Cvi-0 and Col-0 permitted us to assign a large fraction of RNA-seq reads to either genome (on average 13% of the reads), and we used these counts to quantify ASE (Supplemental Figure 1). Haplotype-specific expression estimates were obtained for each gene, counting the number of allele-specific reads stemming from either genomic background. To maximize the sensitivity to detect ASE events, reads from replicates were pooled into four groups, the parental pool and F1 hybrids, each under control and drought stress conditions (Figure 2). Genes with fewer than five allele-specific reads for each parental and hybrid group were removed from the analysis, resulting in 11,003 genes (hereafter referred to as ASEq genes) for which we could quantify ASE, and representing 63% of the genes for which reads were found.

Initially, to estimate the number of genes exhibiting significant allele-specific imbalance in each condition, a basic binomial exact test with a genome-wide false discovery rate (FDR) of 5% was employed in hybrids, using all reads per gene that could be unequivocally assigned to an allele. To ensure that these are regulatory genetic effects, we first tested for genomic imprinting, which refers to a mechanism in plants and animals, crucial for normal development, where inherited alleles are differentially expressed depending on the sex of the parents (Feil and Berger, 2007). In *A. thaliana*, imprinting effects have been observed in endosperm during embryo development (Scott et al., 1998; Gehring et al., 2011). A test for imprinting was performed by pairwise comparison of ASE levels between the three F1 Col x Cvi individuals and the single reciprocal F1 Cvi x Col plant grown under control conditions. As expected for individuals at late growth stages, we did not detect genes with consistent opposite ASE levels between reciprocal crosses (q < 0.05), rejecting the hypothesis of imprinting as a cause for the observed allelic imbalance.

Genome wide, we detected 4706 and 4424 genes exhibiting allele-specific imbalance in F1 hybrids for control and drought stress at a 5% FDR (Figure 3B), representing 42.7 and 40.2%, respectively, of all the ASEq (Figure 3C; Supplemental Data Set 2). Across genes, the ASE ratios were highly correlated between environments (Spearman correlation test, rho = 0.96, P < 0.00001), with 3318 genes exhibiting significant allelic imbalance in both environments (q < 0.05, binomial exact test; green dots in Figure 3B), suggesting widespread cis-regulatory variation sustained across conditions. Moreover, 47.4% (46.8%) of the significant ASE genes exhibited greater expression levels from the Cvi allele for the control (stress) condition, in agreement with an expected symmetrical divergence of cis-regulatory polymorphisms. We considered possible reference biases (Supplemental Figure 2), finding that our alignment approach that accounts for Cvi variants (see Methods) was close to unbiased. This suggests that our results are unlikely to be severely affected by reference bias. To evaluate the accuracy of the RNA-seq-based detection of ASE, we validated our results in the same hybrid samples by pyrosequencing (Wittkopp et al., 2004), estimating ASE in 34 genes previously mapped as local-eQTLs (Cubillos et al., 2012b) (Supplemental Figure 3). In both conditions, the RNA-seq-based estimates and the quantification by pyrosequencing were highly correlated (P < 0.0001, Spearman correlation test: $R^2 = 0.81$ and $R^2 = 0.87$ for control and drought stress, respectively) (Supplemental Figure 3). We further compared the consistency between our estimates for genes with significant ASE and eQTLs identified in the Cvi-0 x Col-0 recombinant inbred lines (Cubillos et al., 2012b). Out of the 2010 genes for which local-eQTLs were previously detected (Figure 3C), we were able to evaluate ASE (under control conditions) for 1475, and 954 (64.6%) showed evidence of variation attributable to cis factors with a consistent directional effect in 85.6% of the cases. Moreover, we found a significant correlation between both studies in the extent of the differential allelic regulation for those genes sharing directionality (P < 2e-16, $R = 0.56$ Pearson correlation coefficient; Supplemental Figure 4), indicating the overall consistency of these regulations across independent genetic designs and experiments. Nevertheless, our current ASE
Figure 2. Overview of the ASE Approach.
strategy detected over four thousand genes with a significant cis-contribution not previously mapped in the RILs, demonstrating its powerful complementation to previous eQTLs mapping studies.

**Contribution of cis- and trans-Regulatory Variants over ASE**

Next, to obtain quantitative estimates of the effect of individual genetic factors on expression differences between accessions, we employed a variance component model, initially independently of the environment, to estimate the variance in contributions of cis- and trans-regulatory polymorphisms to variation in expression of each ASEq gene. Conceptually, this approach is similar to an ANOVA model, estimating the proportion of variance that is attributable to cis and trans effects. However, importantly, this approach models jointly the quantification of RNA-seq reads specific for each allele, thereby ascertaining the effect of cis-regulatory variants (see Methods). We arbitrarily fixed a minimal explained variance of 10% to classify genes as either cis- or trans-regulated. Under this strict criterion, analysis of the 11,003 ASEq genes showed that variation in allelic expression is predominantly cis-regulated, rather than trans-regulated (P < 2e-16, t test; Supplemental Data Set 2). We detected 5137 genes predominantly driven by cis factors, where just the cis component explained more than 10% of the trait’s variation (this was consistent in 83% of the cases with ASE levels as tested in the previous section, under either control and/or stress conditions, 5% FDR, binomial test). In addition, we found 1155 genes with a predominant trans variance effect, which is four times lower than the analogous cis-contribution. Among the genes with pronounced ASE (>10% variance), we found a higher average proportion of variance explained by cis effects compared with trans (45.3% versus 21.5%, P < 0.0001 Kolmogorov-Smirnov test). For 338 genes, the model identified both cis and trans components explaining more than 10% of the gene expression variance, representing 3% of all genes with quantifiable ASE (Figure 4A). In line with the overall preferences for cis effects, also in this set of genes sharing cis- and trans-regulations, the average variance explained in cis was significantly greater than the variance explained by trans factors (34.2% versus 18.8%, P < 0.0001, Kolmogorov-Smirnov test). To understand the frequency at which cis- and trans-regulatory variants converge in the direction of their effect, we analyzed the set of 338 genes exhibiting both regulatory mechanisms. We categorized the direction of the cis and trans effect as either “convergent” (same direction) or “compensatory” (opposite directions). In 49.7% (51.4%) of the cases under control (stress) conditions, allelic effects were convergent between cis and trans (Figure 4B), indicating a nearly uniform occurrence of multiple regulatory variants acting in different directions.

We next examined the set of 1475 ASEq genes for which local-eQTLs were identified in our previous study of the same cross, but in a RIL design (Cubillos et al., 2012b). ASE analysis revealed that 994 of these genes indeed showed evidence of variation attributable to cis factors explaining over 10% of the expression trait. This represents 67.4% consistency in cis effects between these experiments, in agreement with similar findings in mice studies (Lagarrigue et al., 2013). By contrast, out of the 2294 distant-eQTLs previously detected, the overlap with ASEq genes associated with variable trans-acting regulation in this study was only 6.1%, suggesting a less robust expression pattern for trans factors between these experiments performed at different times in distinct conditions. This could also be affected by power issues or inherent limitations resulting from the difference in genetic design to resolve specific cases (e.g., compensating trans-eQTLs controlling the expression of one gene).

**Environmental Effects on Expression Variation**

To assess the extent to which environment affects gene expression levels, we estimated the number of genes in F1 hybrids differentially expressed between drought and control conditions, independently of the parental origin of each allele, using DESeq (Anders and Huber, 2010). We found 678 genes exhibiting significant differences in expression between the two environments (FDR 5%), with fold changes ranging from 0.4 to 29 (Supplemental Figure 5A and Supplemental Data Set 2). This limited transcriptional response reflects our intention to perform physiologically relevant mild water deficit treatments resulting in quantitative decreases in growth rates (Tisné et al., 2013). Using quantitative RT-PCR, we confirmed differential transcript accumulation for six genes, finding good levels of agreement (Supplemental Figure 5B). The direction of changes for the 678 stress-responsive genes was approximately balanced, with 342 upregulated and 336 downregulated genes under drought stress compared with the control condition. The analogous analysis performed in the parental pools yielded 207 differentially expressed genes (Supplemental Figure 5C), out of which only 112 were also detected in F1 hybrids. Differences between hybrids and parental pools indicate that the genotype configuration can affect the response of gene expression to the environment. Moreover, analysis of the F1 gene set responding to the environment by Singular Enrichment Analysis showed a significant enrichment for Gene Ontology biological processes responding to stimulus (e.g., “response to chemical stimulus” and “response
Comparisons between DESeq and ASE results identified a shared set of 302 genes exhibiting ASE under both environments and 138 exhibiting ASE only in a single condition, suggesting a specific response in the latter group. For example, VITAMIN C DEFECTIVE5 (VTC5), which encodes a GDP-L-galactose phosphorylase and has a role in response to water deprivation (Heyndrickx and Vandepoele, 2012), showed significant overexpression of the Col allele and a lower increase of the Cvi allele in response to drought stress, as opposed to control conditions where VTC5 alleles showed no differences (Figure 5A).

Environmental differences can affect gene expression independently of the genotype or alter the specific response of an allelic variant. To determine the extent to which our different treatments modulates the effect of cis- and trans-regulatory factors to influence gene expression, we extended the variance component analysis to include direct environmental effects and genotype x environment (GxE) interactions involving cis- (Gc x E) or trans-regulatory factors (Gt x E). Genome wide, this analysis identified 2348 genes that were subject to significant environmental regulation, out of which 1618 genes had significant variation explained by GxE interactions (>10% variance in either cis or trans). Out of the 1618 genes with a significant GxE effect, 919 had a significant contribution from Gc x E, 524 from Gt x E and 175 had both types of cis and trans interactions Gct x E (Figure 5B; Supplemental Data Set 2). Among genes with a significant GxE effect, the median variance explained by cis factors (Gc x E) was larger than for trans factors (Gt x E) (24.7% versus 20.1%, respectively, P < 0.0005, t test), which is consistent with the relative importance of direct cis and trans effects. However, notably, genes with a significant Gt x E component were frequent (43.2% of the genes under GxE) (Figure 5B), which contrasts with direct genetic effects, which were dominated by cis-regulation (Figure 4A).
Detection of the Directional Effect of the cis/trans-Regulation across Environments

To understand how cis- and trans-regulatory elements interact to shape gene expression between drought stress and control conditions, we determined whether the change between environments was solely a magnitude effect or a directional change in genes under Gc x E and Gt x E. Overall, for Gc x E genes, we observed a significant majority of cases displaying magnitude changes rather than direction changes (Binomial test $P < 2 \times 10^{-16}$) in the effect of the environment on allele specificity (Figure 6A). For genes subject to Gt x E, we observed a corresponding tendency, with a significant number of genes with magnitude changes across environments (Binomial test $P < 2 \times 10^{-16}$), but the percentage of genes exhibiting direction changes was greater compared with Gc x E genes (34% versus 11%, respectively). These results indicate that, although robust cis- and trans-regulation occurs across environments, drought stress would especially affect the direction of the trans-regulation, but show less of an effect on cis regulation, which tended to have consistent directions in 89% of the cases across conditions.

We then compared parental sequence divergence within gene bodies (number of SNPs/kb) and regulatory regions (up to 1 kb upstream the transcribed sequence) in the 175 genes with both cis and trans factors interacting with the environment (Gct x E), including genes with a stronger trans than cis interaction variance (Gt>c x E, trans x environment > cis x environment) and genes with a stronger cis than trans interaction variance (Gc>t x E, cis x environment > trans x environment). A SNP enrichment comparison between Gc>t x E and Gt>c x E genes did not show evidence of a significant difference between groups within gene bodies (9.7 versus 7.9, respectively; $P < 0.16$, t test) or regulatory regions (10.4 versus 8.6, respectively; $P < 0.2$, t test; Figure 6B). A similar comparison between genes with changes in expression divergence across environments purely in cis (Gc x E > 0.1 > Gt x E) and purely in trans (Gt x E > 0.1 > Gc x E) amplified the previous tendency in regulatory regions (9.54 versus 8.71, respectively; $P = 0.06$, t test) or within gene bodies (8.5 versus 7.7, respectively; $P = 0.04$, t test), with a marginally significant difference in both cases. Moreover, when we analyzed SNP
density independently of the environment, we found in genes with an exclusively cis component (Gc > 0.1 > Gt) a significant SNP enrichment within gene bodies (10.21 versus 7.24, respectively, P < 0.0001, t test) and regulatory regions (10.25 versus 8.78, respectively, P < 0.0001, t test) over those genes solely under trans-regulation. Together, these results indicate an excess of polymorphisms within regulatory and coding regions in genes under cis-regulatory variation and, to a lower extent, those cis variants interacting with the environment.

**DISCUSSION**

Here, we have shown how allele-specific expression profiling in two divergent A. thaliana accessions and across different environments allows for estimation of the contribution that cis- and trans-regulatory variants make to variability in gene expression. Notably, we extended previous experimental designs by explicitly controlling for a stress-related environmental factor to study the direct contribution on gene expression and interactions with the genotype. Genome-wide estimates of allele-specific transcript abundance were obtained using RNA-seq read counts from individual F1 hybrids and pools of parents between Col-0 and Cvi-0 accessions grown under finely controlled environmental conditions.

In order to perform an exhaustive genome-wide ASE analysis, it is important to build on comprehensive and accurate genomic information from both parental accessions. To this end, resequencing using Illumina short read technologies allowed us to obtain a catalog of Cvi-0 genetic variants and identify high-quality unambiguous SNPs between our Col-0 and the TAIR10 reference Col-0 accession (Figure 1). Likely, all these SNPs differing between Col-0 stocks correspond to spontaneous mutations fixed in our Col-0 line after it was split from the reference Col-0 (Ossowski et al., 2010), rather than representing errors in the TAIR10 reference sequence. We used the catalog of Col/Cvi SNPs mapping to 26,698 genes to assign individual RNA-seq reads to either genome. After applying a stringent filter for genes with low expression and genes where we are not able to quantify properly the expression level for one of the two alleles, we were able to quantify ASE for 11,003 genes, which represent ~60% of the genes expressed in 3-week-old rosettes grown under our conditions (Figure 2). We analyzed the ASE by two complementary strategies. First, we applied a well-established binomial test (FDR < 5%) to hybrid data and found that, under both conditions, over 40% of the interrogated genes showed significant evidence of being alternatively regulated due to polymorphisms acting in cis (Figure 3C). Strikingly, a majority of these genes (3318) showed ASE under both conditions, indicating a prevalent cis-effect across these two environments. Second, we used a variance component model to dissect the contribution of the variation in transcript accumulation from cis (Gc), trans (Gt), and environmental (E) factors. In our model, we assumed that changes in the absolute abundance of a particular transcript may be due to: (1) genetic factors (the Gc, Gt, or Gct...
A. thaliana found in a Col-0 x Vancouver hybrid, in experiments using and independently of the environment (>10% variance), in the moderate number of genes with significant correlation estimated from population-level analysis (Pickrell et al., 2010). The reduction in transcript abundance from distant factors (Cubillos et al., 2012) suggests a quantitatively minor contribution of cis-regulatory effects toward natural variation in gene expression (Figure 4A). These results are also in agreement with our recent eQTL survey in Cvi-0 x Col-0 RILs, which showed more local- than distant-eQTLs at a conservative FDR, with milder effect contributions on transcript abundance from distant factors (Cubillos et al., 2012b). The remarkable difference between the number of genes under cis or trans regulation suggests that variation in transcript abundance between accessions seems to be governed mostly by polymorphisms near the gene, as previously suggested in a study of a large panel of A. thaliana accessions (Gan et al., 2011). Nevertheless, part of the imbalance between the number of cis- and trans-regulatory variants in this study might be due to statistical limitations in the detection of complex cis x trans genetic interactions in only two parental pool replicates per condition, versus three replicates for F1 hybrids. Furthermore, differences in statistical power to identify cis and trans regulators could be mainly due to larger effects in cis factors, which are then easier to detect (Kliebenstein, 2009). Indeed, when applying less stringent arbitrary cutoffs to the variance component analysis, a number of genes previously thought to be controlled solely in cis also appear to be significantly controlled in trans (Figure 4A), which is in line with polygenic and weak trans regulation for most genes. Finally, complex compensating factors (that may be more frequent among trans than cis elements) will be difficult to detect when comparing F1 and parental pools, since these will be summed in any case.

Furthermore, we attempted to validate our ASE and trans estimations by pyrosequencing using a large set of genes exhibiting variable levels of either form of differential regulation. This approach demonstrated the accuracy of our quantitative estimations with significantly correlated ratios between both approaches (Supplemental Figure 3). We successfully validated the 27 genes predicted to have a significant cis variance from our binomial test. Also, 10 of 22 genes expected to be under trans-acting variation were validated by pyrosequencing. Inconsistencies between both techniques for weaker trans effects might arise because pools of parents used on each measurement were derived from different individuals, while the same RNA samples (= same F1 individuals) where used in RNA-seq and pyrosequencing validation on hybrids.

Lately, the incidence of genes under both types of regulation in opposite directions has been extensively analyzed (Smith and Kruglyak, 2008; Goncalves et al., 2012). In our study, we observed that the occurrence of compensatory cis- and trans-regulatory variation is not rare, as nearly half of the interrogated genes showed this type of regulation, independently of the condition (Figure 4B). To add to our understanding of the major equilibria behind the regulation of transcript accumulation, we studied a challenging environmental contrast: Our controlled mild drought stress treatment mimics a major abiotic limitation to growth, which plants face in the wild and have developed diverse strategies to cope with (Bouchabke et al., 2008). We used DESeq to evaluate the response to stress and detect genes differentially expressed between conditions, independently of the parental origin of each read (Supplemental Figure 5). Comparison of ASE between environments showed that cis-regulatory effects tend to be particularly stable between environments (Figure 3C). A recent study in human cell lines and primary tissue showed that tissue-specific cis-regulatory effects are rare (Grundberg et al., 2011), which is in line with our results between environmental conditions. Yet, the variance component model allowed us to identify complex GxE interactions (Figure 6A): We observed that Gc x E was still the predominant type of interaction, explaining a greater expression variance with respect to other types of GxE. Nevertheless, trans-acting factors were significantly enriched among those involved in GxE (Figure 5B) compared with their prevalence in either environment (Figure 4A), suggesting that the stress treatment preferentially impacts trans regulation profiles. Overall, the variance component model allowed us to detect more genes responding to environmental perturbations compared with the well established DESeq tool, likely due to the power to dissociate the specific allelic response for each GxE interaction. A previous study similarly using two growth conditions in yeast established that local linkages are generally more stable across environments, as opposed to distant linkages, which are more likely to vary depending on the condition (Smith and Kruglyak, 2008). Moreover, work in yeast showed that the identification of cis variants could represent a workhorse for the identification of directional regulatory changes and natural phenotypic differences between individuals (Martin et al., 2012). Therefore, our results can serve as a test for future approaches aiming to tackle for instance drought stress adaptation differences between Col-0 and Cvi-0 accessions.

Interestingly, in our study, the majority of genes with Gc x E effects showed an ASE magnitude change between conditions rather than a change in the direction of the effect. Thus, cis-regulation can have an effect in both conditions and the magnitude of the ASE variation might be influenced by the abundance of the transcript in a specific condition. Overall, our results provide quantitative evidence that genes might be additionally influenced by environment-specific cis-regulatory elements demonstrating the extent of cis variation across conditions. A similar analysis based on the directionality of the trans effect on Gt x E genes showed a less dramatic trend, with 34% of the genes exhibiting a change in the direction of the trans effect between environments (Figure 6A). Distinct trans regulators influencing a gene’s expression depending on the environment seems a plausible explanation for some of the
divergent effects detected in our study. The lower levels of polymorphisms and greater sequence constraint within regulatory regions and gene body in Gt × E genes reaffirms that in A. thaliana, trans regulation plays an important part in response to distinct environmental conditions. An acceptable explanation for the appearance of compensatory effects is the coevolution of cis-regulatory changes and transcription factors able to buffer the response of gene expression upon changes in environmental conditions. Coevolutionary changes and epistatic interactions between cis and trans regulators have been described in a single environment for worms (Barrière et al., 2012) and flies (Landry et al., 2005). Thus, conserved expression profiles could also be maintained across conditions due to the coordinated response of coevolved alleles.

We note that our analysis is not free of limitation. The ability to detect ASE is intrinsically linked to the overall expression levels of individual genes and the availability of heterozygous sites to quantify ASE. In addition to this challenge of lower limits of detection, the effect size of cis and trans genetic variants will determine the detection power and false discovery rates. In particular, in the context of population level association mapping studies, it has repeatedly been reported that cis effects have larger effect sizes than trans effects, which may bias the ability to identify causal loci. We note that the ASE approach does not, in principle, suffer from this limitation as the effect of all cis or all trans factors jointly is ascertained.

Our RNA-seq results provide broad evidence of the extent of ASE in A. thaliana. The significantly greater incidence of ASE found in this study could be due to the low heterogeneity between the examined replicate individuals (Tisné et al., 2013) and to the great resolution provided by RNA-seq to quantify allelic ratios (McManus et al., 2010). Both factors converge toward more testing power; in addition, the chosen parental accessions are particularly genetically and epigenetically distant (Simon et al., 2012; Schmitz et al., 2013). Use of the Phenoscope platform has proven to finely control environmental variations that could introduce noise and variance between individuals, aiding the resolution in this study (Tisné et al., 2013). Moreover, studying regulatory variation using RNA-seq provides greater depth coverage per transcript than previous array-based approaches by increasing the ability to quantify and distinguish between allelic variants (McManus et al., 2010). Ratios (McManus et al., 2010). Both factors converge toward more testing power; in addition, the chosen parental accessions are particularly genetically and epigenetically distant (Simon et al., 2012; Schmitz et al., 2013). Use of the Phenoscope platform has proven to finely control environmental variations that could introduce noise and variance between individuals, aiding the resolution in this study (Tisné et al., 2013). Moreover, studying regulatory variation using RNA-seq provides greater depth coverage per transcript than previous array-based approaches by increasing the ability to quantify and distinguish between allelic variants (McManus et al., 2010; Ozsolak and Milos, 2011). Furthermore, we focused on transcript level variation to delineate the response of cis- and trans-regulatory elements across environments, providing evidence of a large panel of loci showing GxE interactions. Even though the number of genes exhibiting ASE is significantly greater than previously reported, we believe that a larger number of individuals together with tissue-specific sampling across a wider spectrum of environments could still widely extend the repertoire of genes displaying expression differences. The question of whether the pattern herein described is transposable to other plant organisms remains; nevertheless, our findings lay the ground for other similar studies in more complex plant genomes. Characterizing the molecular mechanisms by which cis- and trans-regulatory elements affect allelic expression in natural populations will represent a crucial step toward understanding how regulatory variants modulate transcript accumulation and, ultimately, macroscopic phenotypes.

METHODS

Plant Material and Growth Conditions

Seeds from the INRA-Versailles resource center (http://publiclinies. versailles.inra.fr/) were obtained to perform reciprocal crosses between Arabidopsis thaliana accessions Cvi-0 (166AV) and Col-0 (186AV). These accessions were chosen based on the large number of polymorphisms between them and the wide distribution of local-eQTLs (Cubillos et al., 2012b) (http://signal.salk.edu/atg1001/). We stratified F1 seeds and parental accession for 4 d at 4°C and seeds were germinated. Plants were grown under strictly controlled and monitored environmental conditions as previously described (Tisné et al., 2013). Briefly, individual plants were cultivated in the Phenoscope automated phenotyping setup for 21 d under short days (8 h photoperiod) and well watered (60% soil water content) or mild water-deficit (30% soil water content) conditions, controlled and maintained twice a day by automatically adjusted watering. All plants were continuously rotated during the experiment to homogenize environmental conditions and rosette images were recorded every day and stored in the Phenoscope database for further analyses. For each genotype, replicates were collected above the root 21 d after sowing and stored at −80°C. Images were analyzed and growth profiles representing the rate of expansion of the projected rosette area were studied to identify remaining outliers and discard them from further analysis and RNA-sequencing.

RNA Extraction, cDNA Preparation, and RNA-Sequencing

Frozen rosette tissue was pulverized in liquid nitrogen with a mortar and total RNA was extracted using the RNaseasy Plant mini kit (Qiagen kit 74904) according to the supplier’s instructions. Parental pools were generated by mixing 10 μg of RNA from each of two Col-0 and two Cvi-0 individuals. RNA samples were then treated with DNase I to remove genomic DNA traces. For library preparation, beads with oligo(dT) were used to isolate poly(A) mRNA. cDNA samples were prepared using hexamer primers and the Illumina mRNA-seq sample prep kit according to the manufacturer’s instructions (Illumina). Briefly, samples were subjected to fragmentation and these short fragments were used as templates for cDNA first-strand preparation using random-hexamer primers. The second-cDNA strand was synthesized using 1× buffer (500 mM Tris-HCl, pH 7.8, 50 mM MgCl₂, and 10 mM DTT), 25 mM deoxynucleotide triphosphate, 2 units of RNaseH, and 50 units of DNA polymerase I. Short fragments were purified with the QiaQuick PCR extraction kit and re-suspended in EB buffer for end repair, dA-tailing, and barcode ligation. Suitable fragments for PCR amplification were size-selected by gel electrophoresis on 2% agarose gels. cDNA was PCR amplified with Illumina primers for 12 cycles. Parental and F1 hybrids libraries were subjected to single-end sequencing using Illumina HiSequation 2000 following the manufacturer’s instructions, where every genotype replicate was sequenced in an independent lane to avoid technical biases. The RNA-seq data from this study are available from the NCBI’s Gene Expression Omnibus under accession number GSE43560.

Cvi-0 and Col-0 DNA Isolation, Library Preparation, Sequencing, and Data Analysis

Seeds from Cvi-0 and Col-0 were obtained from the same set of seeds used to generate the F1 hybrids for RNA-seq. Plants were grown in a greenhouse in typical long-day conditions (16-h photoperiod) at 20°C, and whole plants were collected above the roots 21 d after sowing. DNA was extracted with the CTAB method, and sequencing libraries were prepared using the Illumina Genomic DNA Sample Prep protocol. Paired-end libraries with average insert size of 400 nucleotides and 100-nucleotide read length were generated. Sequencing was performed using Illumina HiSequation 2000.
DNA-seq reads were mapped to the Col-0 reference genome (TAIR10 assembly) obtained from the TAIR website (http://www.Arabidopsis.org/) using Bowtie2 (Langmead and Salzberg, 2012) with default settings and BWA allowing up to four mismatches. SNPs were called using the variant caller UnifiedGenotyper tool from GATK, which allows to call differences from the reference sequence (DePristo et al., 2011), with options `-max_alternate_alleles 1-genotype_likelihoods_model BOTH -stand_emit_conf 500.` Heterozygous sites were filtered out to avoid copy number variation or nonunique sequences. These data were used to generate a Cvi consensus sequence.

RNA-seq Alignments and Allele-Specific Quantification of Gene Expression

RNA-seq reads were aligned to the Col-0 reference genome using PALMapper, a variant-aware spliced aligner (Jean et al., 2010). The set of detected Cvi-SNPs and indels was accounted for during the alignment process to minimize read mapping biases between both genomes. Across the 10 samples, between 25.5 million and 27.6 million reads could be mapped, with the fraction of mappable reads varying between 92.8 and 93.8%.

In each sample and for each gene, allele-specific gene expression levels were estimated from all segregating exonic SNPs. Owing to the extent of divergence between Cvi-0 and Col-0, individual RNA-seq reads tended to span multiple segregating alleles. To avoid double-counting, allele-specific counts were created on the basis of individual reads summed across the gene, ensuring consistency of the reads originating either from the Cvi-0 or Col-0 haplotype. Reads that could not be consistently assigned to either haplotype were removed (on average <0.0005% of all reads). For the binomial testing, raw allele-specific read counts in each sample were used. For the variance component modeling, read counts from both genomes were separately normalized, adjusting for differences in library size (using the count median as in Anders and Huber, 2010) and treated as separate RNA quantification experiments (Supplemental Figure 2).

ASE Tests

For every F1 hybrid replicate (Col-0 x Cvi-0 and Cvi-0 x Col-0), we initially summed all the read counts per allele (either Col or Cvi) per gene and performed a binomial exact test against the null hypothesis of no allele-specific expression to test for cis-regulatory divergence. Genes with either no count for any of the alleles or less than five allele-specific reads for each group were discarded from the analysis. Tests were performed in the R statistical environment using the binom.test function (http://www.R-project.org). Resulting P values were corrected for multiple testing using the q-value package (Storey and Tibshirani, 2003). The obtained q-value represented the minimum FDR at which each individual ASE event remains significant.

Variance Component Model

Estimation of variance was performed by jointly fitting a mixed linear model to the allele-specific read count estimates of a particular gene across all samples. To estimate variance parameters, the count information was modeled using a linear mixed model applied to variance stabilized expression levels (Supplemental Text 1). Overdispersion parameters were fit using the approach used in DESeq (Anders and Huber, 2010), employing a separate variance function for parental pools, and hybrids in control and stress conditions. The full regression model contained covariates that correspond to the environmental state (stress/control), the allele at the quantification site for ASE (cis effect), and an indicator variable whether the sample is a parental pool or an F1 (trans effect). To ascertain GxSE effects, the cis and trans covariates were split to model differences of the effect in the control and stress condition, resulting in five covariates. We fit this model to the allele-specific expression estimates of each gene, using the read depth information for the Col and Cvi allele (20 observations per gene). Briefly, the difference between alleles (Cvi/Col) is used to estimate the contribution of cis genetic effects, whereas the deviation between the pools and the F1 lines can be used to determine the trans genetic component of gene expression variation. Further details of the variance component model are provided in Supplemental Text 1.

Pyrosequencing and Real-Time PCR for RNA-seq Validation

Pyrosequencing (PyroMark24; Qiagen) was used to estimate relative allele-specific expression and validate results obtained by RNA-seq. RNA from sequenced samples in hybrids and new replicates of pool of parents were used to generate cDNA. cDNA was prepared using the Fermentas first-strand cDNA synthesis kit with poly-T primers and diluted five times for PCR. F1 genomic DNA was used for technical control as previously indicated (Landry et al., 2005). PCR of cDNA was performed for 34 genes with variable levels of ASE. Pyrosequencing reactions were performed as previously described (Pineau et al., 2012). Briefly, pyrosequencing reactions were performed using SNPs representative of the ASE between Col/Cvi obtained by RNA-seq to independently assess the relative contribution of each allele. Pyrosequencing was performed on cDNA from six F1 individuals (three from each condition), four pools of parents (two from each condition), and on F1 genomic DNA as a control to normalize the ratios against possible pyrosequencing or PCR biases.

Real-time PCR was performed with 3 μL of the diluted cDNA and 1 μL the MESA FAST qPCR Plus for SYBR Green (Eurogentec) kit in a 25-μL reaction volume. Primers sequences are indicated in Supplemental Table 2. Primer amplification efficiency was checked in all cases. The UBC21 gene (At1g25760) was used as control and relative expression ratios were estimated as previously described using the 2^(-ΔΔCT) method (Livak and Schmittgen, 2001).

Accession Numbers

Sequence data from this article can be found in the Arabidopsis Genome Initiative or GenBank/EMBL databases under the following accession numbers: VTC5, At1g55120; At2g38870; PRR5, At1g75040; HSP70, At3g12580; MOT1, At2g25680; QQS, At3g30720; At5g55450; and UBC21, At5g25760.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. RPKM Scatterplot for Reads Spanning SNPs.

Supplemental Figure 2. Percentage of Reads Specifying Aligning to Col (Red) and Cvi (Blue) Genomes before and after the Normalization Procedure.

Supplemental Figure 3. Comparison of Allele-Specific Expression Ratios Obtained from RNA-seq or Pyrosequencing in Different Samples and Conditions.

Supplemental Figure 4. Overlap Comparison for the Additive Effect Estimated from Local-eQTL and Genes under Significant Allele-Specific Expression.

Supplemental Figure 5. Expression Changes across Conditions in F1 and Parental Pools Estimated by DESeq.

Supplemental Table 1. GO Term Enrichment for Genes Differentially Expressed between Conditions in F1 Hybrids.

Supplemental Table 2. Sequencing Primers Used in the qPCR Validation.

ACKNOWLEDGMENTS

We thank Elodie Gilbault and Lién Bach for technical help with the Phenoscope platform. This work was supported by funding from the European Commission Framework Programme 7, ERC Starting Grant DECODE/ERC-2009-StG-243359 to O.L., and ANR Grant 2Complex/ANR-09-BLAN-0368 to O.L. O.S. was supported by a Marie Curie FP7 fellowship.

AUTHOR CONTRIBUTIONS


Received July 22, 2014; revised October 22, 2014; accepted November 8, 2014; published November 26, 2014.

REFERENCES


Extensive cis-Regulatory Variation Robust to Environmental Perturbation in Arabidopsis
Francisco A. Cubillos, Oliver Stegle, Cécile Grondin, Matthieu Canut, Sébastien Tisné, Isabelle Gy and Olivier Loudet
Plant Cell 2014;26;4298-4310; originally published online November 26, 2014; DOI 10.1105/tpc.114.130310

This information is current as of June 19, 2017