Genetic Regulation of Transcriptional Variation in Natural *Arabidopsis thaliana* Accessions

Yanjun Zan,*,† Xia Shen,†,‡,§, Simon K. G. Forsberg,*,† and Örjan Carlborg*,†,1**

*Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 23 Sweden, †Department of Clinical Sciences, Division of Computational Genetics, Swedish University of Agricultural Sciences, SE-756 51 Uppsala, Sweden, ‡Usher Institute for Population Health Sciences and Informatics, and **MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, EH16 4UX United Kingdom, and §Department of Medical Epidemiology and Biostatistics, Karolinska Institute, SE-171 77 Stockholm, Sweden

ORCID IDs: 0000-0001-5571-4578 (Y.Z.); 0000-0003-4390-1979 (X.S.); 0000-0002-7451-9222 (S.K.G.F.); 0000-0002-2722-5264 (Ö.C.)

ABSTRACT An increased knowledge of the genetic regulation of expression in *Arabidopsis thaliana* is likely to provide important insights about the basis of the plant's extensive phenotypic variation. Here, we reanalyzed two publicly available datasets with genome-wide data on genetic and transcript variation in large collections of natural *A. thaliana* accessions. Transcripts from more than half of all genes were detected in the leaves of all accessions, and from nearly all annotated genes in at least one accession. Thousands of genes had high transcript levels in some accessions, but no transcripts at all in others, and this pattern was correlated with the genome-wide genotype. In total, 2669 eQTL were mapped in the largest population, and 717 of them were replicated in the other population. A total of 646 *cis*-eQTL-regulated genes that lacked detectable transcripts in some accessions was found, and for 159 of these we identified one, or several, common structural variants in the populations that were shown to be likely contributors to the lack of detectable RNA transcripts for these genes. This study thus provides new insights into the overall genetic regulation of global gene expression diversity in the leaf of natural *A. thaliana* accessions. Further, it also shows that strong *cis*-acting polymorphisms, many of which are likely to be structural variations, make important contributions to the transcriptional variation in the worldwide *A. thaliana* population.

KEYWORDS
eQTL mapping
RNA sequencing
gene expression
Arabidopsis thaliana
structural variation

Several earlier studies have utilized genome-wide data to explore the link between genetic and phenotypic diversity in natural *Arabidopsis thaliana* populations (Atwell *et al.* 2010; Cao *et al.* 2011; Chao *et al.* 2014; Sanchez-Bermejo *et al.* 2014; Johanson *et al.* 2000; Baxter *et al.* 2010; Forsberg *et al.* 2015; Shen *et al.* 2014; Nemri *et al.* 2010; Weigel 2012). Some of this phenotypic variation was found to be caused by regulatory genetic variants that lead to differences in gene expression [Expression Level Polymorphisms (ELPs)], underlying, for example, semi-dwarfism (Barboza *et al.* 2013), changes in flowering time (Schwartz *et al.* 2009; Johanson *et al.* 2000), changes in seed flotation (Saez-Aguayo *et al.* 2014), and changes in self-incompatibility (Nasrallah *et al.* 2004). Using smaller collections of natural *A. thaliana* accessions, large expression variation has been observed both at individual gene (Richards *et al.* 2012; Gan *et al.* 2011) and gene-network (Kliebenstein *et al.* 2006; van Leeuwen *et al.* 2007) response levels. Thus, extended efforts to scan for ELPs on a whole genome and transcriptome level in large collections of natural *A. thaliana* accessions, are expected to provide useful insights about the link between genetic and expression level variation in the worldwide population. This may ultimately reveal interesting candidate genes and mechanisms that have contributed to adaptation to the natural environment.

Expression quantitative trait loci (eQTL) mapping is a useful approach to link genetic and expression variation. It has been applied successfully in many organisms, including yeast (Brem *et al.* 2002) and plants (West *et al.* 2007), as well as in animals and humans (Schadt *et al.* 2003). Most eQTLs are detected in the close vicinity of the gene itself (*cis*-eQTL), and these often explain a large proportion of the observed expression variation (Li *et al.* 2006; Petretto *et al.* 2006; Wentzell *et al.*

doi: 10.1534/g3.116.030874
Manuscript received April 7, 2016; accepted for publication May 21, 2016; published Early Online May 24, 2016.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.116.030874/-/DC1
[1]Corresponding author: Örjan Carlborg, Uppsala University, Department of Medical Biochemistry and Microbiology, Box 582, SE-751 23 Uppsala, Sweden. E-mail: orjan.carlborg@imbim.uu.se

2007). Fewer eQTL [from 20% to 50% reported in various organisms (Morley *et al.* 2004; Li *et al.* 2006; Keurentjes *et al.* 2007)] are located in the remainder of the genome (*trans*-eQTL). Using eQTL mapping, the genetic regulation of expression variation has been dissected in *A. thaliana* using recombinant inbred lines (RIL) (Keurentjes *et al.* 2007; West *et al.* 2007; Plantegenet *et al.* 2009) and other experimental crosses (Zhang *et al.* 2011). These initial studies have revealed that the majority of the expression variation in *A. thaliana* is heritable, that most of the detected eQTLs are located *in cis*, and that structural variations are a common mechanism contributing to this variation (Plantegenet *et al.* 2009). The data generated within the *A. thaliana* 1001 genomes projects now provides an opportunity to explore this expression variation also in larger collections of natural *A. thaliana* accessions that better cover the worldwide distribution of the plant (Cao *et al.* 2011; Schmitz *et al.* 2013; Dubin *et al.* 2015).

Here, we explored the genome-wide expression variation, and the genetic regulation of gene expression, using a dataset generated from 144 widely distributed natural *A. thaliana* accessions (*SCHMITZ-data*; Supplemental Material, Figure S1; Schmitz *et al.* 2013), and replicated some of our findings in a second dataset with 107 natural *A. thaliana* accessions (*DUBIN-data*; Dubin *et al.* 2015). Transcripts from a core-set of genes were present in the leaf in all accessions, but thousands of genes showed a different pattern with high transcript levels in some accessions, but no transcripts at all in others. RNA sequencing bias was a concern for low-expressed genes, but, focusing on the highly expressed genes we found a large overall contribution by genetics to this variation. Hundreds of *cis*-eQTL contributing to the lack of transcripts in some accessions were mapped in the larger dataset (Schmitz *et al.* 2013), and many of these replicated in the independent smaller dataset (Dubin *et al.* 2015). For about one-quarter of the *cis*-eQTL genes one, or in many cases several, common structural variants were significantly associated with the lack of reads in the transcriptome analyses. This indicates that the lack of transcripts observed for many accessions in the worldwide *A.* thaliana population is often due to common deletions of transcription start sites or the whole genes. Our results thus provide an overall perspective on the transcript-level variation in natural *A. thaliana* accessions, and dissect the genetics underlying the presence or absence of transcripts for individual genes in individual accessions. Overall, our results confirm that loss-of-function alleles (Grant *et al.* 1995; Aukerman *et al.* 1997; Johanson *et al.* 2000; Kliebenstein *et al.* 2001; Kroymann *et al.* 2003; Mouchel *et al.* 2004; Werner *et al.* 2005), often due to structural variations (Plantegenet *et al.* 2009; Gan *et al.* 2011), are important contributors to the overall transcriptional variation also in the worldwide *A. thaliana* population.

MATERIALS AND METHODS

Whole genome resequencing and RNA-seq data for a population of 144 natural A. thaliana accessions

In an earlier study, Schmitz *et al.* (2013) performed genome resequencing and RNA-sequencing in a collection of 144 natural *A. thaliana* accessions. We downloaded this RNA sequencing data (NCBI GEO database access ID: GSE43858) together with their corresponding whole-genome single nucleotide polymorphism (SNP) genotypes (http://signal.salk.edu/atg1001/download.php). According to the author's description and the original publication (Schmitz *et al.* 2013), the plants were grown at 22°, and leaves had been collected from rosettes prior to flowering. Further, RNA reads had been aligned to SNP-substituted reference genomes for each accession using Bioscope version 1.3 with default parameters. Cufflinks version 1.1 had been used to quantify Transcript levels using the following parameters: '–F' 0; '–b'; '–N'; '–library-type' fr-secondstrand; '–G' TAIR10.gtf. Raw Fragment Per Kilobase of exon per Million fragments mapped (FPKM) values were quantile normalized by the 75th percentile and multiplied by 1,000,000. We removed two accessions from the data (Alst_1 and Ws_2) due to missing genotype data, and two accessions (Ann_1 and Got_7) due to their low transcript Call Rate (16,861 and 18,693 genes with transcripts as compared to the range of 22,574 to 26,967 for the other accessions). The final dataset used for eQTL mapping (*SCHMITZ-data*) included 1,347,036 SNPs with MAF above 0.05, a call-rate above 0.95, and RNA-seq derived FPKM-values for 33,554 genes.

Whole genome resequencing and RNA-seq data for a population of 107 Swedish natural A. thaliana accessions

We downloaded a second public dataset of 107 Swedish *A. thaliana* lines with fully sequenced genomes and transcriptomes (Dubin *et al.* 2015) from plants grown at 10° and 16°. Here, RNA had been prepared from whole rosettes collected at the 9-true-leaf stage. RNA reads had been aligned using PALMapper aligner using a variant-aware alignment strategy. Reads that were longer than 24 bp and uniquely mapped into the exonic regions had been used to quantify expression. Further details about read filtering and transcript quantification can be found in Dubin *et al.* (2015). We used the same quality control procedures for this dataset as for the larger dataset described above. We compared the data from the experiments done at 10° and 16°, and found that they were quantitatively similar; therefore, we used only the data generated at 10° (*DUBIN-data*) for further analysis. The same QC procedure as for the *SCHMITZ-data* described above was applied to this data, and here no accessions were removed. In total, the final data contained 1,785,214 SNPs with MAF above 0.05, a call-rate above 0.95, and RNA-seq derived Reads Per Kilobase per Million mapped reads (RPKM)-values for 33,322 genes.

cis-eQTL mapping to detect polymorphisms contributing to the accession specific presence of gene transcripts

First, we selected the 4317 genes in the *SCHMITZ-data* that i) had transcripts in $>$ 14 (10%), but $<$ 126 (90%), of the accessions; and ii) had transcript-levels higher than the 2nd lowest expressed gene with transcripts in all accessions. Then, we binarized the transcript-level phenotype by assigning values zero or one for each accession depending on whether transcripts for the gene were present (normalized FPKM $>$ 0) or not (normalized FPKM $<$ 0). Then, a logistic regression approach was used to perform an analysis across the SNP markers in a $\pm$ 1 Mb region around each of the tested genes using the *qtsore* (*family* = *binomial*) function in GenABEL (Aulchenko *et al.* 2007).

For significance testing, we first applied a Bonferroni corrected significance threshold correcting for the number of SNPs tested in the $\pm$ 1 Mb region around the gene. Second, we accounted for the potential noise in the transcript measurements by applying a second filtering based on a permutation test performed as follows. Under the assumption that all the 0 values (lack of transcripts) resulted from nonbiological noise, the binarized phenotypes were randomly shuffled with respect to the *cis*-SNP genotypes 1000 times, and an association scan performed in each of these datasets as described above. In each scan, the minimum *P*-value was saved to provide an empirical null-distribution for every trait. Trait-specific significance-thresholds were obtained by taking the 1% cut off from these distributions. To account for multiple testing across all tested traits (genes), the FDR was

calculated using the empirical *P*-values as the expected number of significant traits with eQTL at the given significance-threshold under the null hypothesis, divided by the number of traits where significant eQTL were detected.

Mapping of cis- and trans eQTLs for genes with transcripts in most accessions

For the 20,610 genes in the *SCHMITZ-data* where transcripts were detected in > 90% of the accessions, a standard eQTL mapping approach was used by performing a genome-wide association (GWA) analysis fitting inverse-Gaussian transformed expression values to the genome-wide SNP genotypes in a linear mixed model (1) using the *polygenic* and *mmscore* functions in GenABEL package (Aulchenko *et al.* 2007).

$$Y = X\beta + Zu + e \quad (1)$$

Here, Y is the transformed expression phenotype, which is normally distributed with mean 0. X is the design matrix with one column containing the allele-count of the tested SNP (0, 1, and 2 for minor-allele homozygous, heterozygous, and major-allele homozygous genotypes, respectively). β is a vector of the additive allele-substitution effect for the SNP. Z is the design matrix obtained from a Cholesky decomposition of the kinship matrix G estimated from the whole-genome, MAF-filtered SNP data with the *ibs* function (option weight = 'freq') in the GenABEL package (Aulchenko *et al.* 2007). The Z matrix satisfies $ZZ' = G$, and, thus, the random effect vector u is normally distributed, $u \sim N(0, I\sigma_g^2)$. e is the normally distributed residual variance with $e \sim N(0, I\sigma_e^2)$. The GWA results were visualized using the *cgmisc* R-package (Kierczak *et al.* 2015).

A permutation test was used to set the significance threshold in the analysis (Nelson *et al.* 2013; Smith and Kruglyak 2008). As the number of traits was too large to derive a trait-specific threshold, a GWA-scan was performed across all traits to identify those with at least one SNP with $P < 1 \times 10^{-6}$. Among the genes that passed this threshold, a random set of 200 traits was selected. For each of these traits, GWA were performed in 200 permuted datasets where GRAMMAR+ transformed residuals (Belonogova *et al.* 2013; Forsberg *et al.* 2015) were used as phenotype, resulting in 40,000 permutations in total. Based on this total distribution, we derived a 1% significance threshold [$-\log_{10}$ (*P*-value) = 6.84]. The false discovery rate (FDR) among the obtained results was calculated across the traits as $Q_e = E(Q) = E(V/R)$, where V is the number of significant results when the null hypotheses is true (*i.e.*, α*number of tested trait, $\alpha = 0.05$) and R is the total number of trait with significant eQTL (Benjamini and Hochberg 1995).

Replication of detected eQTLs

As 26/38% the SNPs in the *SCHMITZ-data* were missing from the *DUBIN-data* (before/after filtering for MAF), we replicated our findings by testing for associations to SNPs in the *DUBIN-data* that were physically close to the detected top SNP in the *SCHMITZ-data*. The average expected LD-block size in Arabidopsis is about 10 kb (Kim *et al.* 2007), and therefore we performed the replication analysis to SNPs located in a ± 10 kb region around the eQTL peaks in the *SCHMITZ-data* and associate their genotypes to the transcript level of the corresponding genes in the *DUBIN-data*. We considered a *cis*-eQTL replicated if i) the significance for any of the SNPs tested in this region passed a significance threshold corrected for the number of tested SNPs in this region using Bonferroni correction, and ii) the overlapping SNPs show effects in the same direction. The FDR among the obtained results was calculated across the traits as $Q_e = E(Q) = E(V/R)$, where V is the number of significant results when the null hypotheses is true (*i.e.*, α*number of tested regions, $\alpha = 0.05$) and R is the total number of regions with significant eQTL (Benjamini and Hochberg 1995).

Calculating covariances between transcriptome variation and genome variation

Based on the binarized expression values of the 4317 selected genes (selection procedure described above), we created a relationship matrix following the same approach as when calculating the genomic IBS matrix (see details above). The correlation between the transcriptome and genome variation was calculated as the correlation between elements in these two relationship matrixes.

Associating putative structural variants with loss of expression in individual genes

Using the genome resequencing data, we first identified a set of putative structural variants by screening for individual genes in the accessions where no reads were mapped to either the transcription start site (TSS), or the entire gene body of these eQTL genes. The information about the genomic locations where individual accessions lacked mapped reads are available for download from the SALK webpage (http://signal.salk.edu/atg1001/download.php). Since the probability of observing no mapped reads at the exact same extended region in multiple accessions by chance is very low, we compiled a high-confidence set of common structural variants by retaining only those where reads were lacking in the TSS or gene body of the same gene in more than five accessions. To quantify the contribution of these structural variants to the expression variation observed in the RNA-seq analysis, we tested for association between the presence/absence of the structural variation and the presence/absence of RNA-seq reads using a Fisher exact test.

Definition of cis and trans eQTL peaks

For each gene with a significant GWA, the SNP with the lowest *P*-value was selected as the peak location for the eQTL. When the leading SNPs had been defined for each trait, SNPs were considered to represent the same eQTL if they were located within 1 Mb of each other in the TAIR10 reference genome. eQTL peaks located within 1 Mb up or downstream of the genes whose expression was used as phenotype were classified as *cis*-eQTL, while the remaining eQTL peaks were classified as *trans*-eQTL.

Data availability

The public data we used are available as follows. The genome-wide RNA sequencing data for the 144 natural *A. thaliana* accessions (*SCHMITZ-data*) are available in the NCBI GEO database with access ID: GSE43858, and the genome wide DNA sequencing data for the same set of 144 accessions are available for download from http://signal.salk.edu/atg1001/download.php. The genome-wide DNA sequencing data for the 107 Swedish *A. thaliana* accessions (*DUBIN-data*) are available for download from https://github.com/Gregor-Mendel-Institute/swedish-genomes. and the genome wide RNA sequencing data for the same set of 107 accessions are available in the NCBI GEO database with access ID: GSE54680. Genome wide RNA sequencing data for 144 natural *A. thaliana* accessions (*SCHMITZ-data*): NCBI GEO database access ID: GSE43858. Genome wide DNA sequencing data for 144 natural *A. thaliana* accessions: http://signal.salk.edu/atg1001/download.php. Genome wide DNA sequencing data for

107 *A. thaliana* accessions (*DUBIN-data*): https://github.com/Gregor-Mendel-Institute/swedish-genomes. Genome wide RNA sequencing data for 107 *A. thaliana* accessions (*DUBIN-data*): NCBI GEO database access ID: GSE54680.

RESULTS

RNA sequencing detects transcripts from nearly all TAIR10 annotated genes in the leaf of natural A. thaliana accessions

We downloaded publicly available RNA-seq (Schmitz *et al.* 2013) and whole-genome SNP genotype data (Cao *et al.* 2011) for a population of 144 widely distributed natural *A. thaliana* accessions (*SCHMITZ-data*; Figure S1). We then compared the overlap to all 33,602 annotated genes in the TAIR10, including 27,416 protein coding genes, 4827 pseudogenes or transposable element genes, and 1359 ncRNAs. In this data, we first explored the variability in RNA-seq scored expression-values across 33,554 genes and 140 accessions that passed quality control (see *Materials and Methods*). A gene was considered as expressed if it had a normalized FPKM value greater than zero. The available RNA-seq data were generated from a single tissue (leaf) and we therefore expected that a considerable proportion of the genes would be transcriptionally inactive due to tissue specific expression. The data, however, showed that among the 33,554 genes in the *SCHMITZ-data*, only 289 lacked transcripts across all the accessions in the population (*i.e.*, had a normalized FPKM = 0 in all accessions).

A core-set of genes has detectable transcripts in the leaf of all the evaluated natural A. thaliana accessions

In the *SCHMIZ-data*, transcripts were detected (Normalized FPKM > 0) in all accessions for a large set of genes (18,289; Figure 1). In the *DUBIN-data*, 12,927 genes had detectable transcripts in all accessions and all of these transcripts, except 79, had detectable transcripts in all of the accession in the *SCHMITZ-data* (Schmitz *et al.* 2013) (Figure 1). Of the 10,549 genes for which transcripts were detected (RPKM > 0) only in some accessions in the *DUBIN-data*, transcripts were detected for 4129 in all and 6393 in some (RPKM > 0) accessions in the *SCHMITZ-data* (Figure 1). Hence, for only 27 genes with transcripts detected in at least one of the accessions in the *DUBIN-data*, no transcripts were detected in any of the accessions in the *SCHMITZ-data* (Figure 1). The lower sequencing coverage, and more stringent filtering of the sequence reads in the *DUBIN-data* (Dubin *et al.* 2015), likely explains why transcripts were detected for more genes in more accessions in the *SCHMITZ-data*. Few (27) genes were uniquely expressed in the *DUBIN-data*, 13 of these are tRNA genes, nine of which code for proline (Table S1).

The number of genes with detected transcripts in the leaf of individual A. thaliana accessions is highly variable

Within the individual accessions of the *SCHMITZ-data* (Schmitz *et al.* 2013) we found transcripts (Normalized FPKM > 0) for a considerably lower number of genes than the total number of genes with detected transcripts across the entire population. The number of genes with RNA-seq detected transcripts in the individual accessions (Normalized FPKM > 0) varied from 22,574 to 26,967 with an average of 24,565 (Figure 2A). The proportion of genes with detected transcripts was thus higher in this dataset than that was reported earlier for shoots in two natural accessions (Richards *et al.* 2012), but similar to that in the study of 18 natural accessions (Gan *et al.* 2011) [73% in the *SCHMITZ-data*, 64% in (Richards *et al.* 2012) and about 70% in (Gan *et al.* 2011)]. A similar pattern of variability in the number of genes with detected transcripts was also found in the *DUBIN-data* (Figure 2C).

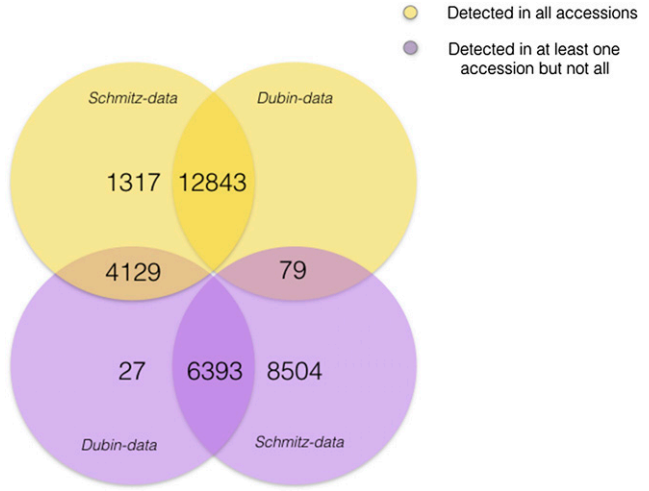

Figure 1 Overlap of RNA-seq scored transcripts in the leaf of 140 natural *A. thaliana* accessions (Schmitz *et al.* 2013) (*SCHMITZ-data*; Normalized FPKM > 0) and 107 Swedish natural *A. thaliana* accessions (Dubin *et al.* 2015) (*DUBIN-data*; RPKM > 0). The numbers of detected transcripts in all accessions of the respective datasets are shown in yellow. The numbers of detected transcripts in at least one, but not all, of the accessions in the respective datasets are shown in purple.

RNA-seq bias likely explains part of the variability in detected transcripts of the individual A. thaliana accessions for low-expressed genes

RNA-seq detects transcripts from nearly all genes in the genome in at least one of the accessions in the *SCHMITZ-data*. In addition to a core set of genes with transcripts in all accessions, there is a large set of genes (on average 6256 per accession) for which transcripts are detected only in some of the accessions. A possible explanation for this might be that RNA-seq is unable to reliably detect low levels of transcripts in the samples. Such RNA-seq introduced bias would result in a random detection of transcripts for individual accessions (*i.e.*, produce a score = 0) for low-expressed genes, and lead to a similar variability in the presence or absence of transcripts among the accessions as when there is a true accession-specific expression. To evaluate the potential contribution of RNA-seq bias to the results, we first evaluated the relationship between the rank of the transcript-levels across the genes and the number of accessions in which transcripts were detected (Figure 2B). There was a clear overall trend in the data that transcripts were detected in fewer accessions for genes with lower overall transcript-levels. A similar trend was observed also in the *DUBIN-data* (Figure 2D). Based on this, we conclude that RNA-seq bias contributes to the observed variation in presence or absence of transcripts among the accessions. However, transcripts were missing in a large fraction of the studied accessions also for many of the genes with high rank/overall transcript level. As RNA-seq bias is an unlikely explanation in these cases, it suggests that at least part of the variability in the presence or absence of transcripts might be due to accession-specific expression, or structural variations leading to different sets of genes in divergent accessions (Gan *et al.* 2011). Thus, our overall results agree across the two datasets in that i) the number of genes expressed in the leaf of an individual plant is large, and ii) that the actual set of genes that are present or expressed varies between the individual accessions.

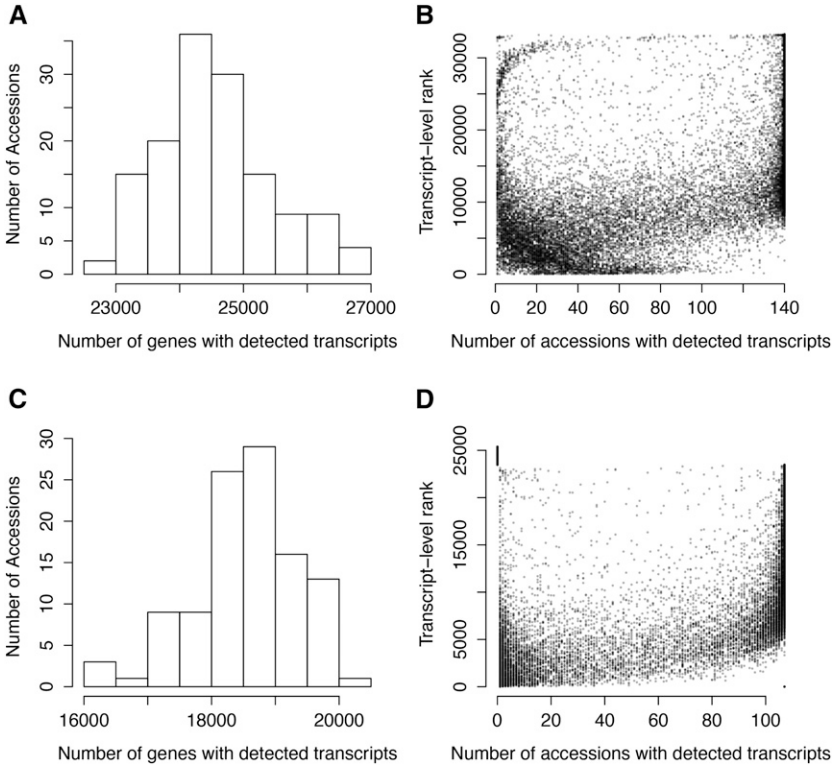

Figure 2 (A) Distribution of the number of genes with transcripts in the leaf of 140 natural *A. thaliana* accessions (Schmitz *et al.* 2013) scored by RNA-seq (Normalized FPKM $>$ 0). (B) Relationship between the ranks of the average transcript levels for all genes with transcripts detected in at least one accession (*y*-axis), and the number of accessions in (Schmitz *et al.* 2013) where transcripts for the gene is found (*x*-axis). Each dot in the plot represent one of the 33,265 genes with FPKM $>$ 0 in at least one accession of (Schmitz *et al.* 2013). The transcript-level rank is based on average transcript levels in the accessions where transcripts for a particular gene are detected. Due to this, the ranks are less precise for transcripts present in fewer accessions. (C) Distribution of the number of genes with detected transcripts in the leaf of 107 Swedish natural *A. thaliana* accessions (Dubin *et al.* 2015) scored by RNA-sequencing (RPKM $>$ 0). (D) Relationship between the ranks of the average transcript levels for all genes with detected transcripts in at least one accession and the number of accessions in (Dubin *et al.* 2015) where the gene is expressed. Each dot in the plot represent one of the 25,382 genes with RPKM $>$ 0 in at least one accession of Dubin *et al.* (2015).

Genetics contributes to the variability in detected transcripts of the individual A. thaliana accessions

The observed variability in the RNA-seq scored presence or absence of transcripts for individual genes between accessions could also be caused by experiment specific environmental factors affecting *e.g.*, plant growth or sample treatments. As transcript variability due to such effects on plants, and samples are expected to be experiment-specific, they should not replicate across experiments. We therefore used the data from a second publicly available RNA-seq dataset from 107 natural Swedish accessions, with fully sequenced genomes and transcriptomes (*DUBIN-data*; Dubin *et al.* 2015), to explore whether a similar pattern of accession-specific transcripts was present also in this data among the genes with high transcript levels. In the *DUBIN-data*, the number of genes with detected transcripts were lower both when measured within any accession (23,478 with RPKM $>$ 0), or within the individual accessions (16,136 to 20,109 with an average of 18,663; Figure 2C). The core set of genes expressed in all accessions contained 12,927 genes. The lower number of genes with transcripts is likely a result of the lower sequencing depth (∼one-quarter of the *SCHMITZ-data*), and the more stringent filtering of the reads. However, the overall trend in the results is the same as in the *SCHMITZ-data*: genes with lower transcript-levels had transcripts detected in fewer accessions, but transcripts for many genes with high overall transcript levels were also found only in a few accessions (Figure 2D). This overlap between the results from the two datasets suggests that the variation in which transcripts are highly expressed in the leaf of several, but not all, accessions is unlikely to be due to either technical bias, or experimental specific errors as described above.

To explore whether there was a genetic basis for the presence or absence of transcripts for individual genes, we studied a set of 4317 genes in the *SCHMITZ-data* in more detail. These genes were selected to i) have detected transcripts in $>$ 14 (10%), but $<$ 126 (90%), of the accessions; and ii) have transcript levels that were higher than the 2nd lowest expressed gene with transcripts in all accessions (*Materials and Methods*; Figure 2B). This subset was chosen to remove most of the genes influenced by the RNA-seq bias for low-expressed genes (as discussed above). To test whether there is a genetic basis underlying the observed variation for these genes, we estimated the pairwise relationships among 140 individual accessions based on expression and genome wide genetic covariance separately (*Materials and Methods*). As illustrated in Figure 3, there was a highly significant correlation between the genetic and transcriptome covariances for this set of genes (r = 0.035, P = 1.0 × 10^{-16}; Figure 3). Although the correlation is low, it indicates that plants that are genetically identical (*i.e.*, the same accession), on average share 4317 × 0.035 = 151 transcripts more than they would with a genetically unrelated accession. This shows that there is a significant genetic contribution to the sharing of transcripts between accessions, suggesting that the accession-specificity of the transcripts has a genetic basis.

cis-eQTL contribute to the accession-specific patterns in the transcriptome

Loss-of-function alleles are known to be important contributors to natural trait-variation in *A. thaliana* (Forsberg *et al.* 2015; Shen *et al.* 2014; Michaels *et al.* 2003; Barboza *et al.* 2013; Grant *et al.* 1995; Aukerman *et al.* 1997; Johanson *et al.* 2000; Kliebenstein *et al.* 2001; Kroymann *et al.* 2003; Mouchel *et al.* 2004; Werner *et al.* 2005). Earlier works on the genetic regulation of expression-variation in an *A. thaliana* have also shown that *cis*-eQTL have a larger effect on expression than *trans*-eQTL (Keurentjes *et al.* 2007; West *et al.* 2007). It is also known that many such ELP genes carry deletions in the regulatory region (Plantegenet *et al.* 2009), but in some cases the transcriptional variation is also due to large structural variations leading to, for example, loss of entire genes (Gan *et al.* 2011). To maximize eQTL-mapping power, we therefore screened the 4317 genes that are most likely to have a genetically controlled accession-specific expression in the *SCHMITZ-data* (*i.e.*, that are the least likely to be affected by RNA-seq bias as

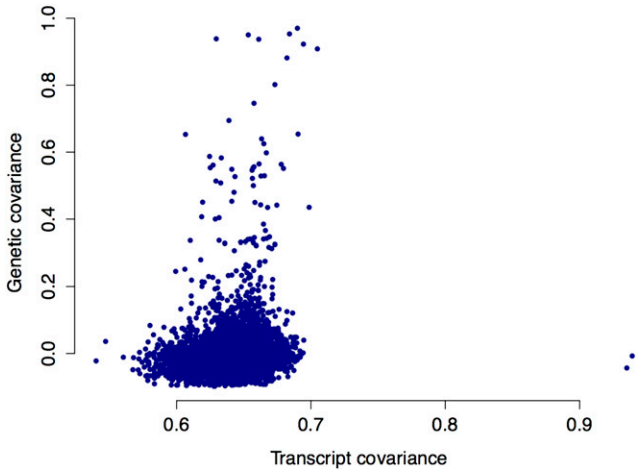

Figure 3 Correlation between the genetic and transcriptome covariances among 4317 genes with transcripts detected in between 14 and 126 of the accessions in the *SCHMITZ-data*, and that are expressed above a level where transcripts RNA-seq have been able to detect transcripts for a gene in all accessions. Each dot in the figure represents a pairwise relationship between two accessions, with the transcript covariance on the *y*-axis and the genetic covariance on the *y*-axis.

discussed above) for *cis*-eQTL alleles affecting the presence or absence of transcripts in the accessions. A customized eQTL-mapping approach designed for this scenario was developed and used for this analysis (see *Materials and Methods* for details).

In total, 349 *cis*-eQTLs were detected (FDR = 12.3%, Table S2) in the *SCHMITZ-data*. For 172 of the 349 genes, whose transcript-levels were affected by these *cis*-eQTL, transcripts were also detected in some, but not all, of the accessions of the *DUBIN-data* (Dubin *et al.* 2015). By performing the same *cis*-eQTL analysis for the presence or absence of transcripts for these 172 genes, 81 of the *cis*-eQTL (FDR = 0.1) could be replicated (Table S3). Given that the collections of accessions in the *SCHMITZ-* and *DUBIN-data* were obtained from nonoverlapping geographical locations, it is striking that so many of the *cis*-eQTL with the ability of almost shutting off the expression are present in, and could be replicated across, such diverse datasets.

Mapping of eQTL for genes with transcripts in most accessions

The majority of the genes in the *SCHMITZ-data* (20,610) (Schmitz *et al.* 2013) were expressed in > 90% of the accessions. The transcript-levels (Normalized FPKM) for these genes were quantile-transformed and used as phenotypes in linear mixed model based, kinship corrected genome-wide eQTL scans. In total, these analyses revealed 2320 eQTL (FDR = 0.09), with 1844 (79.5%) of the associations *in cis* (within a ± 1 Mb window around the mapped gene, actual distance to TSS given in Figure S2), and 476 (20.5%) eQTL *in trans* affecting the expression of 2240 genes (Table S4). All eQTLs were significant after correction for genome-wide analyses across multiple expression traits. Out of the 2320 genes with eQTL in the *SCHMITZ-data*, 2006 had transcripts in all accessions of the *DUBIN-data* (Dubin *et al.* 2015), and 649 of the eQTL affecting 636 genes could be replicated at a 0.01 Bonferroni threshold correcting for the number of tested markers in the replication region (FDR = 0.031; Table S5). Using the list of genes that have earlier been shown to have an altered phenotype from loss-of function alleles from an earlier review (Lloyd and Meinke 2012), we found that among the 2240 genes with eQTL in the *SCHMITZ-data*, 175 have been shown to have such effects and 38 of these were among those replicated in the *DUBIN-data*. As many of these genes have strong effects on potentially adaptive traits, including development, hormone pathways and stress responses, they are plausible functional candidate adaptive genes in *A. thaliana*. These genes are listed in Table S6.

Significant contribution by common structural variants to many of the eQTL

Structural variations, including regulatory unit deletions (Plantegenet *et al.* 2009) and larger genome rearrangements (Gan *et al.* 2011) have been found to contribute to the transcriptional variation in natural *A. thaliana* population both for individual genes (Nordborg *et al.* 2005; Weigel and Mott 2009; Gan *et al.* 2011; Cao *et al.* 2011), and on a whole transcriptome level (Gan *et al.* 2011; Plantegenet *et al.* 2009). We evaluated the contribution of common structural variations to the presence or absence of transcripts in individual accessions by quantifying the overlap of mapped reads in the genome and transcriptome sequencing to the 646 detected eQTL genes that completely lacked transcripts across the gene body of at least one accession. In total, we found that 155 of the 349 *cis*-eQTL genes had putative structural variations (see *Materials and Methods*). On average, each accession carried 58 genes with such structural variations and each gene was disrupted in 52 accessions (Figure S3). To quantify the contribution of these structural variants to the expression variation observed in the RNA-seq analysis, we tested for association between the presence/absence of the structural variation and the presence/absence of RNA-seq reads in the accessions using a Fisher exact test. This association was significant on a nominal level ($P < 0.05$) for 122 of these genes, and 94 of them passed a significance-threshold that was Bonferroni corrected for testing 155 genes. This suggests that the accession-specific transcript pattern observed is often due to structural variations (Table S2) and our eQTL approach is efficient in detecting such segregating variants. In the standard eQTL analysis that was used to map genes that had transcripts in > 90% of the accessions, 297 genes lacked transcripts in at least one accession. By conducting the same structural-variant analysis for these genes, we identified an additional 93 genes with common high-confidence structural variations. In total, 37 of these were significantly associated with the transcript-levels in the accessions on a nominal level ($P < 0.05$) and 17 passed a multiple-testing corrected significance threshold. Thus, in total 248 genes were found to carry at least one common structural variation and for 159 of these, the variants were associated with the transcript-levels in the accessions. Of the 248 genes with common structural variations, 85 contained at least two common structural variants and 60 of the genes with two variants were significantly associated with the presence/absence of transcripts in the accessions. Our results show that structural variations are common, and often multi-allelic, in natural *A. thaliana* accessions and make important contributions to the observed transcriptome variation.

Identification of functional candidate genes with cis-eQTL contributing to the accession specific presence or absence of transcripts

Of the 349 genes affected by the *cis*-eQTL mapped in our study, 12 had earlier been subjected to functional studies in which the gene had a distinct phenotypic effect (Table 1). The *cis*-eQTL-mapping result for one of these genes (*AT2G21045*; High Arsenic Content 1; *HAC1*) is illustrated in Figure 4. *HAC1* has a skewed distribution of RNA-seq scored transcript levels (Figure 4, A and C), and a highly significant

Table 1 Genes with *cis*-eQTL detected in the population of 140 natural *A. thaliana* accessions (SCHMITZ-data) contributing to the accession specific presence or absence of transcripts and earlier reported biological function

Locus	Gene	SNP[a]	MAF[b]	Log odds ratio ± SE[c]	*P*-value[d]	Replicated[e]	Reference
AT2G21045	*HAC1*	chr2_9028685	0.14	4.84 ± 0.76	1.75×10^{-10}	Yes-GBF	Chao *et al.* 2014
AT5G59340	*WOX2*	chr5_23935224	0.43	2.82 ± 0.54	1.83×10^{-7}	Yes	Lie *et al.* 2012
AT1G77235	*MIR402*	chr1_29007464	0.46	7.07 ± 1.02	5.18×10^{-12}	No.e	Kim *et al.* 2010
AT2G38220	*APD3*	chr2_16017043	0.29	0.37 ± 0.07	3.64×10^{-7}	No.e	Luo *et al.* 2012
AT1G13430	*ATST4C*	chr1_4601762	0.51	0.36 ± 0.06	3.87×10^{-10}	No	Marsolais *et al.* 2007
AT1G66600	*WRKY63*	chr1_24217798	0.06	0.14 ± 0.02	3.22×10^{-10}	No	Van Aken *et al.* 2013
AT2G19500	*ATCKX2*	chr2_8168512	0.07	0.18 ± 0.03	6.13×10^{-9}	No	von Schwartzenberg *et al.* 2007
AT3G27620	*AOX1C*	chr3_10230473	0.39	0.4 ± 0.07	6.45×10^{-8}	No	Wang *et al.* 2012
AT3G57270	*BG1*	chr3_22031771	0.06	0.15 ± 0.02	1.09×10^{-9}	No	Zavaliev *et al.* 2013
AT4G01420	*CBL5*	chr4_583422	0.12	0.2 ± 0.03	7.48×10^{-12}	No	Cheong *et al.* 2010
AT4G02850	*DAAR1*	chr4_597373	0.06	0.15 ± 0.02	1.41×10^{-9}	No	Strauch *et al.* 2015
AT4G29340	*PRF4*	chr4_14639984	0.21	0.36 ± 0.07	2.32×10^{-7}	No	Deeks *et al.* 2005

[a] Top SNP in association analysis.
[b] Minor allele frequency of the top associated SNP in the *SCHMITZ-data*.
[c] Log odds ratio of the top associated SNP ± SE.
[d] Nominal *P*-value for the top associated SNP.
[e] Could the original association in the *SCHMITZ-data* be replicated in the *DUBIN-data*: Yes-GBF, Replicated at Genome Wide Bonferroni threshold; Yes, replicated at Bonferroni threshold correcting for number of markers in ± 1 Mb window of the peak SNP in *Dubin-data*; No.e, not expressed in *DUBIN-data* so could not be tested for replication; No, nonsignificant in replication analysis.

cis-eQTL signal in both the *SCHMITZ* and *DUBIN-data* (Figure 2, B and D; $P = 1.75 \times 10^{-10}/P = 9.32 \times 10^{-11}$, respectively). It has been shown that loss of *HAC1* expression increases the amount of Arsenic in the leaf and that several functional polymorphisms might be present in natural *A. thaliana* population (Chao *et al.* 2014; Sanchez-Bermejo *et al.* 2014). Here, we detected a *cis*-eQTL for the expression of *HAC1* with the top SNP located within an exon of *HAC1* (Figure 4B; top SNP Chromosome 2 at 9,028,685 bp) in the *SCHMITZ-data*. This signal was replicated with a high significance also in the *DUBIN-data* (Figure 4D). This gene is thus a highly interesting functional candidate adaptive gene as i) it has earlier confirmed effects on Arsenic-levels in the plant, ii) there is a strong *cis*-eQTL regulating presence or absence of transcript for the gene segregating in natural *A. thaliana* accessions, and iii) effect of the polymorphisms could be replicated at high significance in both analyzed populations. Further experimental studies are, however, needed to functionally replicate the effect of the remaining 11 polymorphisms on expression and the indicated phenotype.

DISCUSSION

We studied transcript variation in the leaf of natural *A. thaliana* accessions by analyzing data from two large sets of natural accessions from the global *A. thaliana* population. In this data, there was a significant

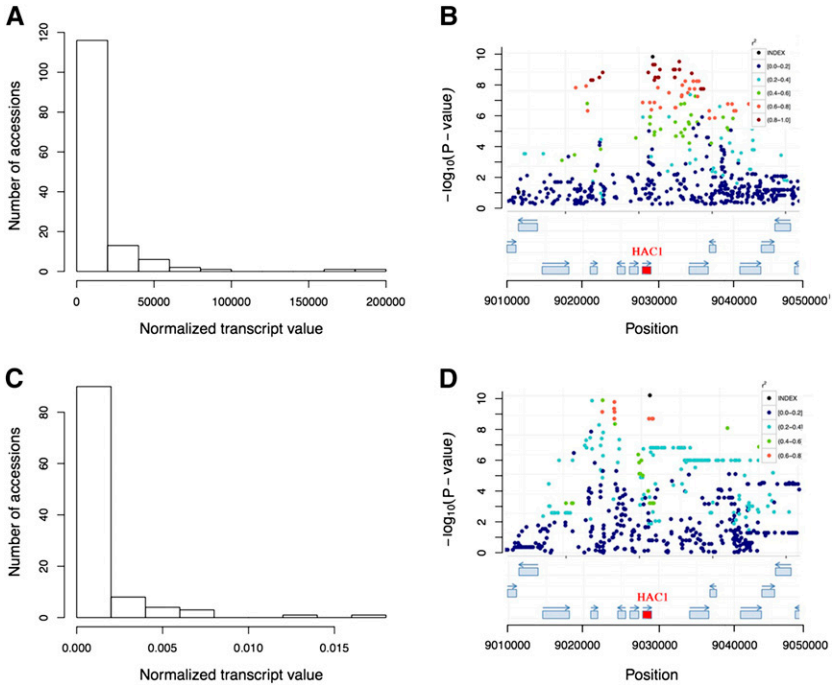

Figure 4 The eQTL analysis detects a highly significant, replicable association for the expression of the gene HAC1 (AT2G21045). The peak SNP is located in an exon of HAC1. (A/C) Distributions of transcript-levels for the 140/107 accessions (FPKM/RPKM-values from RNA-sequencing) in the *SCHMITZ-data* (A) and *DUBIN-data* (C) (Schmitz *et al.* 2013; Dubin *et al.* 2015), respectively. (B, D) Illustrations of the association-profiles (Kierczak *et al.* 2015) for expression of the gene AT2G21045 (HAC1) in the *SCHMITZ-data* (B) and the *DUBIN-data* (D) (Schmitz *et al.* 2013)/(Dubin *et al.* 2015), respectively. There is a highly significant *cis*-eQTL to a SNP located in an exon of HAC1.

correlation between transcript variation and genome divergence among the accessions and thousands of individual eQTL that contribute to this variation in leaf transcript-levels across the accessions could be mapped. These extensive and genetically controlled differences in expression-levels controlled by alleles that segregate in the wild are a useful resource to explore how selection might have acted on different regulatory variants of genes important for the adaptation of *A. thaliana* to diverse living conditions. Several functional candidate adaptive genes were identified among the eQTL, for which further experimental validation would be highly motivated.

The average proportion of genes with detected transcripts in the leaf within the individual accessions was slightly higher in this study than in earlier studies based on smaller numbers of accessions (73% here *vs.* 64% in Richards *et al.* 2012). In the seedling, Gan *et al.* (2011) report a slightly higher estimate for protein-coding genes, but lower if all genes are considered. A large core-set of genes (61% of all with scored transcripts) was detected with transcripts in nearly all ($>$ 90%) of the accessions. The genes in this core-set overlapped to a great extent between the two independent collections of natural *A. thaliana* accessions studied here, suggesting that these genes include both basal, and leaf-specifically, expressed genes.

Here, we observed that thousands of genes only have transcripts in a subset of the accessions in the two studied collections. The majority of the accession-specific transcripts were, however, found at low levels in only a few accessions. As these are likely the result of the high sensitivity of the RNA-seq approach, it is difficult to separate true accession specific expression from bias in the RNA-seq analysis. For example, some genes are known to vary their transcripts levels in a circadian manner (2% according to Schaffer *et al.* 2001; or 6% according to Harmer *et al.* 2000), which might reduce the expression of these genes to a lower level in all accessions and lead to random detection in RNA sequencing. However, transcripts from a relatively large number of genes were abundant in some accessions, and completely absent in others. This suggests that these genes are more likely to be expressed, or lost, in an accession-specific manner, as illustrated by the correlation between the genetic and transcriptome covariances, which suggests that genetically identical individuals would share many (~150; see *Results*) more transcripts than unrelated ones.

Here, we developed an approach to map *cis*-eQTL affecting the presence or absence of transcripts in the accessions motivated by earlier findings that illustrated the importance of strong loss-of-function alleles for adaptation in *A. thaliana* (Forsberg *et al.* 2015; Shen *et al.* 2014; Barboza *et al.* 2013; Kliebenstein *et al.* 2001; Grant *et al.* 1995; Aukerman *et al.* 1997; Johanson *et al.* 2000; Mouchel *et al.* 2004; Werner *et al.* 2005). Given that phenotypes of many mutant alleles are already described in the literature, our aim was to identify new functional candidate adaptive genes by screening for *cis*-eQTL where alleles contributing to the presence or absence of transcripts for the studied genes segregated in populations of natural *A. thaliana* accessions. If these *cis*-eQTL affect a gene with a known phenotype, it can then be considered as a strong candidate adaptive gene. Among our results, we find a particularly interesting example of where our approach is able to detect functional genetic variants that have been studied in detail. *HAC1* (Chao *et al.* 2014; Sanchez-Bermejo *et al.* 2014) was detected by our approach in the larger collection (Schmitz *et al.* 2013) and replicated in the smaller (Dubin *et al.* 2015), and the functional natural variation in this gene and its role in the reduction of Arsenik has earlier been described in the literature.

Earlier studies have shown that utilization of several natural *A. thaliana* accessions, in addition to *Col-0*, is useful for uncovering biologically important genetic and phenotypic variations that are not present in this reference accession (Forsberg *et al.* 2015; Shen *et al.* 2014; Barboza *et al.* 2013; Kliebenstein *et al.* 2001; Grant *et al.* 1995; Aukerman *et al.* 1997; Johanson *et al.* 2000; Gan *et al.* 2011; Kroymann *et al.* 2003). Our results further support this as 5% (111) of the *cis*-eQTL genes lacked transcripts in the leaf of *Col-0*. The fact that half of them (53) also lacked transcripts in *Col-0* seedling and pollen (Loraine *et al.* 2013) (Table S7) suggests that these genes are not expressed at all in this accession (Gan *et al.* 2011).

Structural variations have earlier been found to be common in the genome, and to make a significant contribution to the transcriptional variation in *A. thaliana* genes (see, for example, Nordborg *et al.* 2005; Weigel and Mott 2009; Gan *et al.* 2011; Cao *et al.* 2011; Kroymann *et al.* 2003; Plantegenet *et al.* 2009). We found that 248 of the 646 genes that lacked transcripts in one or more accessions, and for which eQTL were mapped, also lacked reads that covered the transcription start site, or the entire gene, in the available whole-genome sequencing data. For about one-third of the genes (80), both types of structural variants were common. For a majority (159) of these genes, the structural variants were significantly associated with the presence or absence of the transcript. This result confirms that structural variations regulated transcript variation are relatively common in *A. thaliana* (Plantegenet *et al.* 2009; Gan *et al.* 2011), that they are often multi-allelic, and are likely to make significant contributions to the transcriptome variation in the worldwide population.

When there is a leap in technology, it is important to adapt the existing analytical methods to better utilize the additional information. One of the concerns about the earlier microarray-based eQTL studies is that, in the context of eQTL mapping, the limited detection range of the microarray measurements (*i.e.*, the upper and lower bounds for detection of expression) are unlikely to capture the full range of expression differences present in nature. Power might therefore have been lost in cases when the biological range of expression exceeded the detection boundaries of the array. Here, we made use of the information about expression-traits where the transcript distributions for the assayed genes are heavily zero-inflated (*i.e.*, contain many accessions with no detectable transcripts). We show that these distributions, in many cases, are influenced by *cis*-eQTL with strong effects on the transcript levels. These distributions are too skewed to be appropriately transformed or modeled using other distributions such as a negative binomial (Sun 2012). Earlier eQTL-mapping studies based on RNA-seq data have therefore either removed genes whose transcript levels were zero-inflated (Brown *et al.* 2014), utilized nonparametric testing to avoid the assumption of normality (Montgomery *et al.* 2010), or utilized regression without first addressing the potential issues arising from the non-normality of the transcript-levels (Pickrell *et al.* 2010). Our approach, based on the observed distribution properties, to propose and test for a particular type of genetic effects causing accession-specific expression, revealed many new *cis*-eQTL that, in many cases, are likely to be due to structural variants that either delete important regulatory regions for the genes, or the entire gene. Many of these associations replicated in an independent dataset and several genes were promising functional candidate adaptive genes. Although the proposed *cis*-eQTL-mapping approach is not a general framework for mapping all eQTL, it illustrates the value of developing and utilizing analysis methods for genome and transcriptome data to test well-defined biological hypotheses.

ACKNOWLEDGMENTS

We thank Lars Hennig, David Salt, Daniel Kliebenstein, and Detlef Weigel for valuable comments and suggestions during the analyses.

LITERATURE CITED

Atwell, S., Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton *et al.*, 2010 Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465(7298): 627–631.

Aukerman, M. J., M. Hirschfeld, L. Wester, M. Weaver, T. Clack *et al.*, 1997 A deletion in the PHYD gene of the *Arabidopsis* Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. Plant Cell 9(8): 1317–1326.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. Van Duijn, 2007 GenABEL: an R library for genome-wide association analysis. Bioinformatics 23(10): 1294–1296.

Barboza, L., S. Effgen, C. Alonso-Blanco, R. Kooke, J. J. Keurentjes *et al.*, 2013 *Arabidopsis* semidwarfs evolved from independent mutations in GA20ox1, ortholog to green revolution dwarf alleles in rice and barley. Proc. Natl. Acad. Sci. USA 110(39): 15818–15823.

Baxter, I., J. N. Brazelton, D. Yu, Y. S. Huang, B. Lahner *et al.*, 2010 A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1;1. PLoS Genet. 6(11): e1001193.

Belonogova, N. M., G. R. Svishcheva, C. M. van Duijn, Y. S. Aulchenko, and T. I. Axenovich, 2013 Region-based association analysis of human quantitative traits in related individuals. PLoS One 8(6): e65395.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289–300.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science 296(5568): 752–755.

Brown, A. A., A. Buil, A. Viñuela, T. Lappalainen, H.-F. Zheng *et al.*, 2014 Genetic interactions affecting human gene expression identified by variance association mapping. eLife 3: e01381.

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43(10): 956–963.

Chao, D.-Y., Y. Chen, J. Chen, S. Shi, Z. Chen *et al.*, 2014 Genome-wide association mapping identifies a new arsenate reductase enzyme critical for limiting arsenic accumulation in plants. PLoS Biol. 12(12): e1002009.

Cheong, Y. H., S. J. Sung, B.-G. Kim, G. K. Pandey, J.-S. Cho *et al.*, 2010 Constitutive overexpression of the calcium sensor CBL5 confers osmotic or drought stress tolerance in *Arabidopsis*. Mol. Cells 29(2): 159–165.

Deeks, M. J., F. Cvrcková, L. M. Machesky, V. Mikitová, T. Ketelaar *et al.*, 2005 *Arabidopsis* group Ie formins localize to specific cell membrane domains, interact with actin-binding proteins and cause defects in cell expansion upon aberrant expression. New Phytol. 168(3): 529–540.

Dubin, M. J., P. Zhang, D. Meng, M.-S. Remigereau, E. J. Osborne *et al.*, 2015 DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. eLife 4: e05255.

Forsberg, S. K., M. E. Andreatta, X.-Y. Huang, J. Danku, D. E. Salt *et al.*, 2015 The Multi-allelic genetic architecture of a variance-heterogeneity locus for molybdenum concentration in leaves acts as a source of unexplained additive genetic variance. PLoS Genet. 11(11): e1005648.

Gan, X., O. Stegle, J. Behr, J. G. Steffen, P. Drewe *et al.*, 2011 Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477(7365): 419–423.

Grant, M. R., L. Godiard, E. Straube, T. Ashfield, J. Lewald *et al.*, 1995 Structure of the *Arabidopsis* RPM1 gene enabling dual specificity disease resistance. Science 269(5225): 843–846.

Harmer, S. L., J. B. Hogenesch, M. Straume, H. S. Chang, B. Han *et al.*, 2000 Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. Science 290(5499): 2110–2113.

Johanson, U., J. West, C. Lister, S. Michaels, R. Amasino *et al.*, 2000 Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. Science 290(5490): 344–347.

Keurentjes, J. J., J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken *et al.*, 2007 Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. Proc. Natl. Acad. Sci. USA 104(5): 1708–1713.

Kierczak, M., J. Jabłońska, S. K. Forsberg, M. Bianchi, K. Tengvall *et al.*, 2015 cgmisc: enhanced genome-wide association analyses and visualization. Bioinformatics 31(23): 3830–3831.

Kim, J. Y., K. J. Kwak, H. J. Jung, H. J. Lee, and H. Kang, 2010 MicroRNA402 affects seed germination of *Arabidopsis thaliana* under stress conditions via targeting DEMETER-LIKE Protein3 mRNA. Plant Cell Physiol. 51(6): 1079–1083.

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark *et al.*, 2007 Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39(9): 1151–1155.

Kliebenstein, D. J., V. M. Lambrix, M. Reichelt, J. Gershenzon, and T. Mitchell-Olds, 2001 Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. Plant Cell 13(3): 681–693.

Kliebenstein, D. J., M. A. West, H. Van Leeuwen, K. Kim, R. Doerge *et al.*, 2006 Genomic survey of gene expression diversity in *Arabidopsis thaliana*. Genetics 172(2): 1179–1189.

Kroymann, J., S. Donnerhacke, D. Schnabelrauch, and T. Mitchell-Olds, 2003 Evolutionary dynamics of an *Arabidopsis* insect resistance quantitative trait locus. Proc. Natl. Acad. Sci. USA 100(suppl 2): 14587–14592.

Li, Y., O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu *et al.*, 2006 Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. PLoS Genet. 2(12): e222.

Lie, C., C. Kelsom, and X. Wu, 2012 WOX2 and STIMPY-LIKE/WOX8 promote cotyledon boundary formation in *Arabidopsis*. Plant J. 72(4): 674–682.

Lloyd, J., and D. Meinke, 2012 A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*. Plant Physiol. 158(3): 1115–1129.

Loraine, A. E., S. McCormick, A. Estrada, K. Patel, and P. Qin, 2013 RNA-seq of *Arabidopsis* pollen uncovers novel transcription and alternative splicing. Plant Physiol. 162(2): 1092–1109.

Luo, G., H. Gu, J. Liu, and L. J. Qu, 2012 Four closely-related RING-type E3 ligases, APD1–4, are involved in pollen mitosis II regulation in *Arabidopsis*. J. Integr. Plant Biol. 54(10): 814–827.

Marsolais, F., J. Boyd, Y. Paredes, A.-M. Schinas, M. Garcia *et al.*, 2007 Molecular and biochemical characterization of two brassinosteroid sulfotransferases from *Arabidopsis*, AtST4a (At2g14920) and AtST1 (At2g03760). Planta 225(5): 1233–1244.

Michaels, S. D., Y. He, K. C. Scortecci, and R. M. Amasino, 2003 Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. Proc. Natl. Acad. Sci. USA 100(17): 10102–10107.

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle *et al.*, 2010 Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464(7289): 773–777.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. Nature 430(7001): 743–747.

Mouchel, C. F., G. C. Briggs, and C. S. Hardtke, 2004 Natural genetic variation in *Arabidopsis* identifies BREVIS RADIX, a novel regulator of cell proliferation and elongation in the root. Genes Dev. 18(6): 700–714.

Nasrallah, M., P. Liu, S. Sherman-Broyles, N. Boggs, and J. Nasrallah, 2004 Natural variation in expression of self-incompatibility in *Arabidopsis thaliana*: implications for the evolution of selfing. Proc. Natl. Acad. Sci. USA 101(45): 16070–16074.

Nelson, R. M., M. E. Pettersson, X. Li, and Ö. Carlborg, 2013 Variance heterogeneity in *Saccharomyces cerevisiae* expression data: trans-regulation and epistasis. PLoS One 8(11): e79507.

Nemri, A., S. Atwell, A. M. Tarone, Y. S. Huang, K. Zhao *et al.*, 2010 Genome-wide survey of *Arabidopsis* natural variation in downy mildew resistance using combined association and linkage mapping. Proc. Natl. Acad. Sci. USA 107(22): 10302–10307.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian *et al.*, 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3(7): 1289.

Petretto, E., J. Mangion, N. J. Dickens, S. A. Cook, M. K. Kumaran *et al.*, 2006 Heritability and tissue specificity of expression quantitative trait loci. PLoS Genet. 2(10): e172.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt *et al.*, 2010 Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464(7289): 768–772.

Plantegenet, S., J. Weber, D. R. Goldstein, G. Zeller, C. Nussbaumer *et al.*, 2009 Comprehensive analysis of *Arabidopsis* expression level polymorphisms with simple inheritance. Mol. Syst. Biol. 5(1): 242.

Richards, C. L., U. Rosas, J. Banta, N. Bhambhra, and M. D. Purugganan, 2012 Genome-wide patterns of *Arabidopsis* gene expression in nature. PLoS Genet. 8(4): e1002662.

Saez-Aguayo, S., C. Rondeau-Mouro, A. Macquet, I. Kronholm, M.-C. Ralet *et al.*, 2014 Local evolution of seed flotation in *Arabidopsis*. PLoS Genet. 10(3): e1004221.

Sanchez-Bermejo, E., G. Castrillo, B. del Llano, C. Navarro, S. Zarco-Fernandez *et al.*, 2014 Natural variation in arsenate tolerance identifies an arsenate reductase in *Arabidopsis thaliana*. Nat. Commun. 5: 4617.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature 422(6929): 297–302.

Schaffer, R., J. Landgraf, M. Accerbi, V. Simon, M. Larson *et al.*, 2001 Microarray analysis of diurnal and circadian-regulated genes in *Arabidopsis*. Plant Cell 13(1): 113–123.

Schmitz, R. J., M. D. Schultz, M. A. Urich, J. R. Nery, M. Pelizzola *et al.*, 2013 Patterns of population epigenomic diversity. Nature 495(7440): 193–198.

Schwartz, C., S. Balasubramanian, N. Warthmann, T. P. Michael, J. Lempe *et al.*, 2009 Cis-regulatory changes at FLOWERING LOCUS T mediate natural variation in flowering responses of *Arabidopsis thaliana*. Genetics 183(2): 723–732.

Shen, X., J. De Jonge, S. K. Forsberg, M. E. Pettersson, Z. Sheng *et al.*, 2014 Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. PLoS Genet. 10(12): e1004842.

Smith, E. N., and L. Kruglyak, 2008 Gene-environment interaction in yeast gene expression. PLoS Biol. 6(4): e83.

Strauch, R. C., E. Svedin, B. Dilkes, C. Chapple, and X. Li, 2015 Discovery of a novel amino acid racemase through exploration of natural variation in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA 112(37): 11726–11731.

Sun, W., 2012 A statistical framework for eQTL mapping using RNA-seq data. Biometrics 68(1): 1–11.

Van Aken, O., B. Zhang, S. Law, R. Narsai, and J. Whelan, 2013 AtWRKY40 and AtWRKY63 modulate the expression of stress-responsive nuclear genes encoding mitochondrial and chloroplast proteins. Plant Physiol. 162(1): 254–271.

van Leeuwen, H., D. J. Kliebenstein, M. A. West, K. Kim, R. van Poecke *et al.*, 2007 Natural variation among *Arabidopsis thaliana* accessions for transcriptome response to exogenous salicylic acid. Plant Cell 19(7): 2099–2110.

von Schwartzenberg, K., M. F. Núñez, H. Blaschke, P. I. Dobrev, O. Novák *et al.*, 2007 Cytokinins in the bryophyte *Physcomitrella patens*: analyses of activity, distribution, and cytokinin oxidase/dehydrogenase overexpression reveal the role of extracellular cytokinins. Plant Physiol. 145(3): 786–800.

Wang, H., J. Huang, X. Liang, and Y. Bi, 2012 Involvement of hydrogen peroxide, calcium, and ethylene in the induction of the alternative pathway in chilling-stressed *Arabidopsis* callus. Planta 235(1): 53–67.

Weigel, D., 2012 Natural variation in *Arabidopsis*: from molecular genetics to ecological genomics. Plant Physiol. 158(1): 2–22.

Weigel, D., and R. Mott, 2009 The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. 10(5): 107.

Wentzell, A. M., H. C. Rowe, B. G. Hansen, C. Ticconi, B. A. Halkier *et al.*, 2007 Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet. 3(9): 1687–1701.

Werner, J. D., J. O. Borevitz, N. Warthmann, G. T. Trainer, J. R. Ecker *et al.*, 2005 Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc. Natl. Acad. Sci. USA 102(7): 2460–2465.

West, M. A., K. Kim, D. J. Kliebenstein, H. van Leeuwen, R. W. Michelmore *et al.*, 2007 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. Genetics 175(3): 1441–1450.

Zavaliev, R., A. Levy, A. Gera, and B. L. Epel, 2013 Subcellular dynamics and role of *Arabidopsis* β-1, 3-glucanases in cell-to-cell movement of tobamoviruses. Mol. Plant Microbe Interact. 26(9): 1016–1030.

Zhang, X., A. J. Cal, and J. O. Borevitz, 2011 Genetic architecture of regulatory variation in *Arabidopsis thaliana*. Genome Res. 21(5): 725–733.

Communicating editor: D. J. de Koning