Published online August 2, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 13 ***3677–3686***
doi:10.1093/nar/gkl510

Genome-wide screening for *cis*-regulatory variation using a classical diallel crossing scheme

Raphaël Kiekens, Annelies Vercauteren, Beatrijs Moerkerke[1], Els Goetghebeur[1], Hilde Van Den Daele, Roel Sterken, Martin Kuiper, Fred van Eeuwijk[2] and Marnik Vuylsteke*

Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Gent, Belgium, [1]Department of Applied Mathematics and Computer Science, Ghent University, B-9000 Gent, Belgium and [2]Laboratory of Plant Breeding, Wageningen University and Research Centre, NL-6700 AJ Wageningen, The Netherlands

Received March 23, 2006; Revised June 21, 2006; Accepted July 5, 2006

ABSTRACT

Large-scale screening studies carried out to date for genetic variants that affect gene regulation are generally limited to descriptions of differences in allele-specific expression (ASE) detected *in vivo*. Allele-specific differences in gene expression provide evidence for a model whereby *cis*-acting genetic variation results in differential expression between alleles. Such gene surveys for regulatory variation are a first step in identifying the specific nucleotide changes that govern gene expression differences, but they leave the underlying mechanisms unexplored. Here, we propose a quantitative genetics approach to perform a genome-wide analysis of ASE differences (GASED). The GASED approach is based on a diallel design that is often used in plant breeding programs to estimate general combining abilities (GCA) of specific inbred lines and to identify high-yielding hybrid combinations of parents based on their specific combining abilities (SCAs). In a context of gene expression, the values of GCA and SCA parameters allow *cis*- and *trans*-regulatory changes to be distinguished and imbalances in gene expression to be ascribed to *cis*-regulatory variation. With this approach, a total of 715 genes could be identified that are likely to carry allelic polymorphisms responsible for at least a 1.5-fold allelic expression difference in a total of 10 diploid *Arabidopsis thaliana* hybrids. The major strength of the GASED approach, compared to other ASE detection methods, is that it is not restricted to genes with allelic transcript variants. Although a false-positive rate of 9/41 was observed, the GASED approach is a valuable pre-screening method that can accelerate systematic surveys of naturally occurring *cis*-regulatory variation among inbred lines for laboratory species, such as *Arabidopsis*, mouse, rat and fruitfly, and economically important crop species, such as corn.

INTRODUCTION

The detection of allele-specific expression (ASE) and the subsequent identification of the regulatory variants are of increasing interest. In contrast to coding variants, which are relatively easy to detect by re-sequencing exonic sequences across individuals, regulatory variants are practically impossible to discern, even from complete analysis of sequence variation through a gene locus. A tempting approach is to use existing bioinformatics tools to identify functional regulatory variants, but despite advances in the field, these computer predictions have relatively poor specificity (1). *In vitro* methods may offer some help to identify functional polymorphisms, but their role is limited by the inability of plasmid constructs to mimic the role of the natural genomic context in establishing ASE (2).

Current high-throughput approaches to identify genetic polymorphisms that alter gene expression are generally limited to the identification of differences in allelic expression found *in vivo* and to associate these differences with *cis*- and *trans*-regulatory changes. Examples include large-scale correlations of marker genotypes to gene expression levels modeled as quantitative traits in a number of organisms, such as yeast (3), rodents (4–7), human (4,8) and *Arabidopsis*

*To whom correspondence should be addressed. Tel: +32 9 3313860; Fax: +32 9 3313809; Email: marnik.vuylsteke@psb.ugent.be
Present address:
Annelies Vercauteren, Instituut voor Landbouw-en Visserij Onderzoek, Eenheid Plant/Gewasbescherming, Burg. Van Gansberghelaan 96, B-9820 Merelbeke, Belgium

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

thaliana (9,10). These recent studies on 'genetics of gene expression' identified two types of marker-expression trait correlations: those in which a transcript level maps to a genomic region containing the structural gene producing the transcript ('auto-linkages'), and those in which the expression level is associated with a distinct locus elsewhere in the genome. Although auto-linkage implies local *cis*-acting variation, local variation acting *in trans* through a feedback loop (11) or polymorphisms in a *trans*-acting modulator in linkage disequilibrium (LD) with the target gene expression level (12) are also likely to be responsible for auto-linkages.

Another approach to differentiate between *cis*- and *trans*-control involves the quantification of allele-specific transcripts in a heterozygous diploid individual (13–18), such as an F_1 hybrid. This strategy ensures that the observed differences in transcript levels can only be due to *cis*-acting sequence variation because the alleles in the F_1 hybrids are exposed to common *trans*-acting factors and environmental influences. This *cis–trans* test was elegantly extended by Wittkopp *et al*. (19) who included the parental expression ratio. In this way, different patterns of gene regulation could be distinguished: (i) genes that showed the same allelic ratios in the parents and hybrids were determined to be affected by *cis*-regulatory variants; (ii) genes that showed allelic bias in the parents, but equal proportions in the hybrid were determined to be strongly affected by *trans*-regulatory variants; and (iii) genes in which hybrid allelic proportions matched neither parental nor equal proportions were determined to be regulated by a combination of *cis*- and *trans*-variants. Such *in vivo* ASE detection involves the use of a polymorphism in the transcript itself as a marker to differentiate the two allele-specific transcripts in the hybrid. This, however, limits such an analysis to genes with allelic variants that are distinguishable by a genetic polymorphism.

Here, a new strategy to detect and assign imbalance in allelic expression to *cis*-regulatory variation is presented and described below. This novel approach to perform a genome-wide analysis of ASE differences, which we call GASED, is based on the method of Wittkopp *et al*. (19), but differs in that the detection of ASE is not restricted to genes with allelic transcript variants. For this purpose, a classical diallel crossing scheme was chosen as experimental design and the multiple F_1 hybrids were analyzed using a mixed model framework (20). We show how partitioning between-F_1 hybrid genetic variance for mRNA abundance into the additive and non-additive variance components allows to differentiate between *cis*- and *trans*-regulatory changes and to assign imbalances in allelic expression to *cis*-regulatory variation. With this GASED approach we identified a total of 715 and 236 genes that are subject to allelic polymorphisms responsible for at least a 1.5- and 2-fold allelic expression difference, respectively, in a total of 10 diploid *Arabidopsis* hybrids.

MATERIALS AND METHODS

Rationale

Despite gene expression is controlled by multiple *cis*- and *trans*-elements, it is, nevertheless, useful to consider the transcript abundance of a gene as the summation of the individual *cis*- and *trans*-effects. With the easiest case of a diploid individual with only two alleles at each *cis*- and *trans*-locus, the expression value of a gene in an F_1 hybrid resulting from the cross $i \times j$ can be modeled as follows:

$$y_{ijk} = \mu + c_i + ct_{ii} + c_j + ct_{jj} + ct_{ij} + ct_{ji} + \varepsilon_{ijk}, \quad \mathbf{1}$$

where y_{ijk} is the expression phenotype of the k-th offspring from cross $i \times j$, μ is the mean of the expression values obtained in all crosses considered, c_i and c_j are the effects of the *cis*-elements of the i-th and j-th gamete, respectively, ct_{ii} and ct_{jj} represent the interaction between *cis*- and *trans*-elements of the i-th and j-th gamete, respectively, and ct_{ij} and ct_{ji} represent the interaction between *cis*- and *trans*-elements in one gamete with those of the other. Unless *trans*-acting factors bind with the *cis*-regulatory element directly or indirectly by way of a transcription complex, there will be no effect from the *trans*-acting factors *per se*. In cases of homozygosity, i.e. parental lines, $i = j$ and Equation 1 becomes

$$y_{iik} = \mu + 2c_i + 2ct_{ii} + \varepsilon_{iik}, \quad \mathbf{2}$$

where i $(= 1, \ldots, p)$ specifies the parental line.

As individual effects of *cis* and *cis–trans* interactions of the same gamete on the allelic expression cannot be distinguished, Equation 1 can be rewritten as follows:

$$y_{ijk} = \mu + g_i + g_j + s_{ij} + \varepsilon_{ijk}, \quad \mathbf{3}$$

where y_{ijk} is the expression phenotype of the k-th offspring from cross $i \times j$, μ is the population mean effect, g_i and g_j correspond to the $c_i + ct_{ii}$ and $c_j + ct_{jj}$ effects, respectively, and s_{ij} corresponds to the $ct_{ij} + ct_{ji}$ effect. To estimate the parameters g_i, g_j and s_{ij}, a diallel design can be used. Diallel designs are often used in quantitative genetics for analogous situations where additive effects of parental gametes [referred to as general combining abilities (GCAs)] and non-additive effects of the hybrids [referred to as specific combining abilities (SCAs)] need to be estimated. In such a quantitative genetics context, g_i, g_j refer then to the GCAs of parents i and j, whereas s_{ij} is the SCA of $i \times j$ matings. Diallel designs differ from factorial mating designs by the fact that paternal and maternal sets consist of the same genotypes. Hence, with p parents, there are p^2 potential crosses in such an experiment: the p parental lines, one set of $p(p-1)/2$ F_1 lines, and the set of $p(p-1)/2$ reciprocal F_1 lines (21).

Unequal expression of the alleles in a hybrid genetic background indicates the presence of *cis*-regulatory variants (13). In addition, if these *cis*-regulatory variants completely explain the expression difference between parents, the allelic (H) and parental (P) expression ratios will be the same ($P = H \neq 1$) (19). Writing the P and H ratios as functions of the *cis*- and *trans*-acting elements, the equality in P and H, implying the absence of *trans*-regulatory variants, can be expressed as follows:

$$\frac{2(c_i + ct_{ii})}{2(c_i + ct_{jj})} = \frac{c_i + ct_{ii} + ct_{ij}}{c_i + ct_{jj} + ct_{ji}}, \quad \mathbf{4}$$

where the left-hand term equals P and the right-hand term H. It follows from Equation 4 that the absence of *trans*-regulatory variants implies that $ct_{ij} = ct_{ji} = 0$. It also follows

from Equation 4 that screening for ASE differences arising primarily from *cis*-regulatory variants in a particular hybrid cross $i \times j$ implies testing for imbalance in allelic expression, i.e. $c_i+ct_{ii} \neq c_j+ct_{jj}$, in the absence of *trans*-regulatory variants, i.e. $ct_{ij} = ct_{ji} = 0$. In terms of g_i, g_j and s_{ij}, screening for allelic expression differences arising primarily from *cis*-regulatory polymorphisms in a particular hybrid cross $i \times j$, implies testing for ASE differences, i.e. $g_i \neq g_j$, in the absence of *trans*-variants, i.e. $s_{ij} = 0$.

As microarray expression data are believed to be multiplicative, a $\log_2$ transformation is a natural method for analyzing expression data. As a consequence, y_{ijk} in Equations 1 and 3 becomes the $\log_2$ expression phenotype of the k-th offspring from cross $i \times j$ and y_{iik} in Equation 2 becomes the $\log_2$ expression phenotype of the k-th parental line i. In the absence of *trans*-regulatory variation, the equality of P and H ratios on the original scale can then be approximated by an equality of differences in means of expression values on the $\log_2$ scale as follows:

$$2(c_i + ct_{ii}) - 2(c_j + ct_{jj}) = 2g_i - 2g_j, \quad \mathbf{5}$$

where the left-hand term equals the difference in parental mean $\log_2$ values and the right-hand term equals the difference in allelic mean $\log_2$ values in the $i \times j$ cross. From Equation 5, it is obvious that, when $\log_2$ expression values are analyzed, differences between g_i and g_j effects measure half the difference in allelic mean $\log_2$ expression phenotypes.

PLANT MATERIAL AND EXPERIMENTAL DESIGN

GASED analysis

The *A.thaliana* (L.) Heyhn. accessions Columbia (Col-4; N933), Landsberg *erecta* (L*er*; N8581), Cape Verde Islands (Cvi; N8580), Wassilewskija (Ws-4; N5390) and C24 (N906) were derived from seeds obtained from the Nottingham *Arabidopsis* Stock Centre (http://arabidopsis.info/). The 5 accessions were used as parental lines to produce by hand pollination the 10 pairs of reciprocal F_1 hybrids. The 5 parental lines and 10 F_1 hybrids were each represented by 4 hybridizations on 30 microarrays (Figure 1), involving two technical and two biological replicates (referred to as Cy3 and Cy5, and B1 and B2, respectively). This design ensured that each genotype was contrasted directly against four other genotypes, but not directly to both its parents. Dyes were balanced with respect to genotypes.

Verification of identified ASE differences

An independent verification of a reasonable proportion of the identified ASE differences was done by a reduced genetical genomics experiment using 18 L*er*/Cvi recombinant inbred lines (RILs). These RILs were at the F_8 generation (22), and seeds were obtained from the Nottingham *Arabidopsis* Stock Centre (http://arabidopsis.info/). Informative RILs were chosen based on the 99 (mainly AFLP) framework markers covering the 475 cM L*er*/Cvi linkage map (http://arabidopsis.info/): RILs with >4 recombinations per chromosome or >5% scoring error for the framework markers were considered ambiguous for selective mapping and therefore

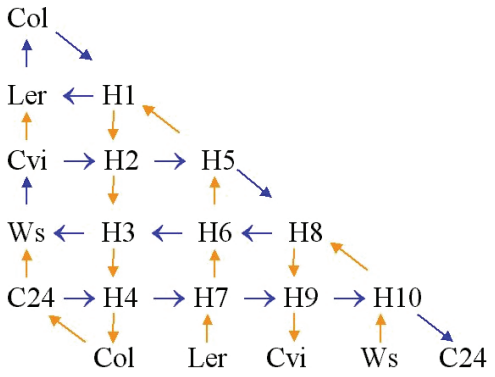

Figure 1. Experimental design, consisting of 30 two-dye CATMA v2.2 microarray (24) experiments to examine transcript levels in RNA samples collected from a diallel experiment in *A.thaliana* with 5 parental lines and 10 F_1 hybrids. The five hybridized parental lines were the homozygous accessions Col, L*er*, Cvi, Ws and C24. F_1 hybrids were obtained by making all pairwise crosses between the five parental lines. The 10 hybrid samples hybridized consisted of pooled progeny from reciprocal crosses. The microarray contained 22 494 unique GSTs from *Arabidopsis* and is represented by an arrow, connecting the two sampled genotypes hybridized to it. The samples at the tail and head of each arrow were labeled with Cy3 and Cy5, respectively. The two replicates are represented by differently colored arrows.

excluded. From the remaining RILs, a set of 20 highly informative RILs (N22089, N22090, N22091, N22093, N22106, N22111, N22112, N22116, N22120, N22124, N22127, N22131, N22132, N22134, N22136, N22137, N22145, N22152, N22155 and N22159) was selected with minimal expected maximum bin length of 6.78 map units, with the software program MapPop version 1.0 (23). A total of 18 RILs (after failure of N22090 and N22124) were each represented by 2 hybridizations on 18 microarrays according to a one-way loop design (the 18 samples were hybridized together in consecutive pairs, each labeled once in red and once in green). Because each allelic state within each gene is expected to be replicated several times across the 18 RIL individuals, albeit in different genetic backgrounds caused by variation by all other genes, introducing replicates into the experiment for further noise reduction was not deemed necessary. Dyes were balanced with respect to genotypes.

Growth conditions

Seeds were sown in pots with standard soil. After 4 days of a cold long-night/warm short-day treatment [16 h dark at 4°C/8 h at 22°C with cool-white light (tube code 840) of 65 mE m^{-2} s^{-1} photosynthetically active radiation], they were transferred to a short-day regime (8 h light/16 h dark at 22°C). To avoid position effects, trays were rotated randomly every 2 days.

Microarrays

The CATMA v2.2 array used for this study contained 23 688 features, including 22 494 unique gene-specific tags (GSTs) from *Arabidopsis* (24), 768 positive and negative control spots (GE-Healthcare, Little Chalfont, UK) and 426 blank spots. Design and synthesis of primary and secondary GST amplicons have been described elsewhere (24,25) The GSTs

primarily match (3′) exons or 3′-untranslated region sequences but occasionally (2.9%) they include matches to intron sequences. The GST amplicons were purified and arrayed as described previously (26). The CATMA GST array was printed at the VIB microarray facility (http://www.microarray.be) and consists of 2 mega-columns and 12 mega-rows, resulting in 24 blocks. Each block represents a set of spots printed with a single and identical print tip. The array design can be accessed via the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) with accession number E-TABM-67 or via the VIB microarray facility website (http://www.microarrays.be).

Sampling, target labeling and hybridizations

To avoid developmental variation in gene expression, whole shoots of parental lines, F_1 hybrids and RILs were harvested at growth stage 1.04 corresponding to a fourth leaf length of ~1 mm (27), 6 h after dawn, and immediately frozen in liquid nitrogen. Total RNA was extracted from pools of 10 plants. For the F_1 hybrids, shoots of reciprocal hybrids were pooled in a 1:1 mixture only after the hybrids had been proven heterogeneous at marker loci. RILs and hybrids/parents were grown under identical growth conditions, but harvested separately in time.

Total RNA was prepared with TRIzol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol. Total RNA (5 µg) was reverse transcribed to double-stranded cDNA and further amplified. Subsequent Cy3 and Cy5 labeling, hybridization, post-hybridization washing and scanning were performed as described previously (24). All protocols for Cy3 and Cy5 labeling, hybridization and scanning can be accessed through the VIB microarray facility website (http://www.microarrays.be) and at the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress). The diallel and RIL transcript profiling data have been submitted to ArrayExpress under E-TABM-67 and E-CAGE-112, respectively.

Pre-processing of expression data

GASED analysis. All analyses were performed on the log base 2 foreground fluorescence intensity measurements. The expression data were analyzed in two steps: (i) a within-slide analysis aimed at removing variation associated with spatial (for instance grid layout on the slide) and structural components (e.g. print order, differential dye responses to binding and scanning) as noise; and (ii) a between-slide analysis aimed at estimating the mean differences between treatments and their standard error. For the within-slide analysis, a spatial linear fixed model of the form given below was applied:

$$\text{response} = \mu + \text{pin} + \text{row} + \text{column} + \text{spline}(\text{intensity}) + \text{residual}, \quad \mathbf{6}$$

where the response variable is the $\log_2$ ratio of the foreground fluorescence intensities (M) measured at the 23 262 spots. Within the model (6), the dye bias was represented by a cubic smoothing spline curve [spline(intensity)] (28), as implemented in the GenStat menu (29) for microarray data analysis. Other potential effects added to the model as fixed terms were the 24-pin effects that printed the slides, and the 252-row and 94-column effects of the microarray layout. Once the adjusted $\log_2$ ratios (M') for each gene were obtained, adjusted $\log_2 R$ and $\log_2 G$ signal intensities were calculated. Positive signals were selected as described previously (30) based on the 48 adjusted $\log_2 R$ and $\log_2 G$ signal intensities of the APB rYR1 negative control spotted 24 times on a single array. At 10 877 GSTs (48%), at least 2 of the 4 observations for each of the 15 genotypes had a signal above the threshold. All values below the signal threshold were reset to the median value of the APB rYR1 negative control intensities.

For the between-slide analysis, a two-step mixed model analysis of variance (31) was used and performed with GenStat (29). Each of the 30 hybridization samples was subjected to a linear normalization model of the form given below:

$$\text{response} = \mu + \text{array} + \text{pin} + (\text{array.pin}) + \text{residual}, \quad \mathbf{7}$$

where the response variable represents the corrected $\log_2$-transformed Cy3 and Cy5 fluorescence intensity measurements of the 10 877 GSTs with a positive signal. Array (modeling the hybridization effects of each of the 30 microarrays), pin and pin by array effects were added as random terms.

Verification of identified ASE differences. All analyses were performed on the log base 2 foreground fluorescence intensity measurements. Within- and between-slide normalization was performed as described for the GASED procedure.

GASED analysis

The GASED strategy is outlined in Figure 2 and aspects of the individual components are discussed below. First, we aimed at estimating the proportion of expressed genes for which a significant part of their variation can be attributed to genotypic differences, i.e. is genetic. The residuals from

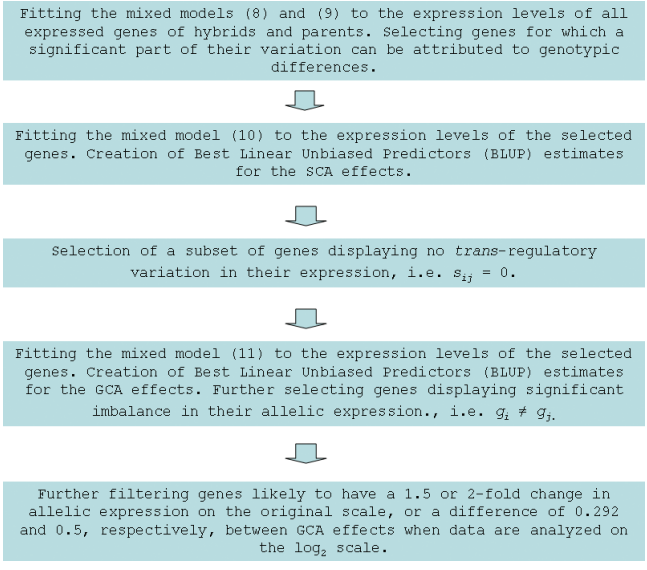

Figure 2. GASED strategy to differentiate between *cis*- and *trans*-regulatory changes on a large scale and to assign imbalances in allelic expression to *cis*-regulatory variation using a classical diallel crossing scheme.

the model (7) were analyzed for each of the 10 877 GSTs separately by a mixed model of the following form:

$$\text{residual} = \mu + \text{dye} + \text{replicate} + \text{genotype}_{ij} + \text{array} + \text{error}, \quad \mathbf{8}$$

partitioning gene-specific variation into gene-specific fixed dye (Cy3 and Cy5) and replicate (B1 and B2) effects, and random genotypic and spot effects. The random genotypic effect, genotype_{ij}, refers to both parental lines ($i = j$) and hybrids ($i \neq j$). The array term models the effects for each spot and equals the (GST.array) interaction effect. Random effects in the model were assumed to be independent and normally distributed with means zero and variance σ_t^2, where t = G (genotype), A (array) and E (error).

The linear mixed model (8) was fitted and the genetic variance component σ_G^2 was estimated by restricted maximum likelihood (REML) as implemented in GenStat (29). The REML directive also calculates the likelihood value, which is required for the test H_0: $\sigma_G^2 = 0$ versus H_1: $\sigma_G^2 > 0$. For that test, the likelihood value under the full model (L_1) given by (8) was compared to the likelihood under the reduced model (L_0) of the form given below:

$$\text{residual} = \mu + \text{dye} + \text{replicate} + \text{array} + \text{error}. \quad \mathbf{9}$$

Twice the difference in log-likelihoods $\lambda = -2(L_1 - L_0)$ is a likelihood ratio test (LRT) statistic, which is usually asymptotically χ^2 distributed with degrees of freedom (df) equal to the difference in the number of parameters included in the full and reduced models (in this case, 1). Because the variance parameter σ_G^2 was constrained to be positive, changes in the log-likelihood based on independence of random effects followed a mixture of χ_0^2 and χ_1^2 distributions in a ratio 50:50 under H_0, rather than the χ^2 distribution. Therefore, we adjusted the P-values for the LRT statistics different from zero and P-value = 1 otherwise, by taking half the P-value from the usual χ_1^2 test.

The false discovery rates (FDRs) were subsequently estimated by modeling the adjusted P-values as a 2-component mixture of Uniform and Beta densities (32), as implemented in GenStat (29); default parameter settings were used to estimate π_0, the proportion of features that are truly null.

In a second step, we wanted to estimate the SCA effect for each gene with a significant σ_G^2 variance component in each of the 10 hybrids, and to calculate its significance. The following linear mixed model was fitted:

$$\text{residual} = \mu + \text{dye} + \text{replicate} + g_i + g_j + s_{ij} + \text{array} + \text{error}, \quad \mathbf{10}$$

which is characteristic for a diallel crossing method in which parents are included and reciprocal F_1 hybrids are pooled [$p(p-1)/2$ combinations]. In this model, gene-specific variation is partitioned into gene-specific fixed dye (Cy3 and Cy5) and replicate (B1 and B2) effects, random GCA effects of the i-th and j-th parent ($i = 1, \ldots, 5$; $j = 1, \ldots, 5$) parameterized as a single matrix of indicator variables for the parents, random SCA effect for the cross between the i-th and the j-th parent ($i \neq j$), and random spot effects modeled by the array term (for a more detailed overview of data structure for a single gene, showing how the GCA and SCA values are parameterized with dummy variables see also Supplementary Table 1). In model (10), the variance due to genotypic differences is, in a way, partitioned between variation due to GCA differences and due to SCA differences. In line with the choice made with respect to the genotypic effects in model (8), we chose to take both GCA and SCA terms random in model (10), and assumed their effects in the model to be normally and independently distributed with means zero and variance σ_{GCA}^2 and σ_{SCA}^2, respectively. Best linear unbiased predictors (BLUP) estimates for the SCA effects were created together with an estimate for the corresponding variance–covariance matrix. Test statistics, i.e. BLUP divided by prediction error, were calculated for each gene in each of the 10 hybrids to test for the SCA effect differing from zero. These ratios were supposed to follow approximately a t-distribution with the df equal to the df for the error term in the gene-specific model (10). For each transcript and for each hybrid, P-values were calculated based on the t-approximation to the test statistics for the contrasts. To derive a list of genes with a non-significant SCA effect, it is important to estimate the false non-discovery rate (FNR) (33,34) defined as the fraction of false negatives among those declared non-significant, and to calculate a Q_a-value that measures the FNR when calling a gene non-significant for SCA effects (and hence all genes with a higher P-value). The Q_a-values were estimated by modeling the P-values as a 2-component mixture of Uniform and Beta densities (32), as implemented in Genstat (29); default parameter settings were used to estimate π_0, the proportion of features that are truly null.

In a third and final step, to estimate the GCA effects for each gene with a non-significant SCA effect in a particular hybrid, and to test H_0: $g_i = g_j$, the following linear mixed model was fitted:

$$\text{residual} = \mu + \text{dye} + \text{replicate} + g_i + g_j + \text{array} + \text{error}. \quad \mathbf{11}$$

From the REML analysis, we saved for each gene a vector of BLUP estimates for the GCA effects and the corresponding estimate for the variance–covariance matrix. The test statistics for the 10 pairwise contrasts were supposed to follow approximately a t-distribution with the df equal to the df for the error term in the gene-specific model (11), and corresponding P-values were calculated. The FDR was subsequently estimated from the obtained P-values.

To further filter genes likely to have a certain minimal difference in GCA effects in absolute value on the $\log_2$-scale, which we called Δ^1, we used a newly developed testing methodology that allowed to focus on only those differences that were likely to be at least the chosen threshold value Δ^1 (35). Technically spoken, we complemented the traditional P-value with a one-sided 'alternative' P_1-value (36), which is the P-value for testing the hypothesis H_1:$|\Delta_{ij}| = \Delta^1$ versus the hypothesis H_0:$|\Delta_{ij}| < \Delta^1$, where Δ_{ij} represents the true underlying difference in GCA effect for a given gene. Hence, we looked for small P and large P_1 as evidence in favor of H_1. From equality (5) it is obvious that in the absence of SCA effects differing from zero, differences in GCA effects measure half the difference in allelic mean $\log_2$ expression. Hence, imposing a 1.5- or 2-fold change in allelic expression

on the original scale implies $\Delta^1 = 0.584/2 = 0.292$ and $\Delta^1 = 1/2 = 0.5$, respectively. Here, we counted genes with $Q < 0.001$ and $P_1 > 0.10$, as carrying evidence in favor of $H_1{:}|\Delta_{ij}| = \Delta^1$.

Verification of ASE differences

The (residual) genetic expression variation observed across the 18 RILs was partitioned into a fixed part associated with the genomic bin where the gene of interest is located (referred to as target genomic bin) and a random part associated with the genetic background consisting of all remaining genomic bins. The GST-specific mixed model used to test for the linkages was as follows:

$$\text{residual} = \mu + \text{dye} + \text{target} + \text{background} + \text{array} + \text{error}, \quad \textbf{12}$$

where target denotes the fixed target genomic bin tagged by a particular marker, background corresponds to the random effect that results from polygenic contributions and background QTLs located in all but the target genomic bins, and array represents the random spot effects. The state of the target genomic bin in a particular RIL, either homozygosity for the L*er* allele, heterozygosity or homozygosity for the Cvi allele, was indicated by the marker genotype. This approach eliminated the need to perform a whole-genome scan for each expression trait, reducing the number of statistical tests to a single-marker test and, therefore, increasing the power to confirm the expected *cis*-regulated ASE. This model was fitted by REML, and the linkage between target genomic bin and expression trait was assessed by a Wald test as implemented in GenStat (29). To test for background effects, likelihoods under a full model, including background markers, and under a reduced model without these markers were compared. On significance for the background variation, the genomic bin with the highest absolute BLUP estimate for the substitution effect was considered to harbor the major *trans*-acting locus.

RESULTS

The GASED analysis relies on the use of a diallel design, in which each line is crossed with several others. Crossing a line to several others allows the assessment of the mean performance of the line across these crosses. This mean performance expressed as a deviation from the mean of all crosses in the diallel, is called the GCA of the line. Any particular cross, then, has an 'expected value' which is the sum of the two parental GCAs. The cross may, however, deviate from this expected value. This deviation, then, is called the SCA.

When a set of inbred lines is used in a diallel crossing scheme, as is the case in this study, the genetic interpretation of GCA and SCA is simplified by the fact that the analysis becomes in reality a 'gamete' combining ability analysis. Therefore, in genetic terms, the GCAs represent the additive effects of the parental gametes, whereas the SCAs represent the non-additive effect of putting gametes together in pairs to make the F_1 genotypes. In statistical terms, the GCAs are the main effects and the SCA is an interaction. GCA is, in fact, equivalent to the breeding value of an individual, and, therefore, in a context of breeding, is of great importance to identify higher yielding combinations of parents. In a context of gene expression, however, where transcript levels are determined by *cis*- and *trans*-elements, the GCAs may be regarded as the summation of expression effects contributed by each gamete (i.e. the set of *cis*- and *trans*-elements in the gamete) and the SCA represents the interaction of the gametes (i.e. the interaction of the *cis*- and *trans*-elements in one gamete with those of the other).

In a purely additive case, i.e. when the dominance deviation SCA equals zero, the allelic expression ratio in the hybrid (H), representing the relative abundance of the allele-specific transcripts in a common hybrid genetic background $i \times j$, can be written as $H = g_i/g_j$, where g_i and g_j represent the GCA of the i-th and j-th parent, respectively. This is equal to the parental expression ratio $P = 2g_i/2g_j$, as parental lines are homozygous for the expressed alleles. On a $\log_2$ scale, the equality of the P and H ratios can be approximated by an equality of differences in means of parental and allelic expression values. According to Wittkopp *et al*. (19), this equality of ratios on the original scale, or equality in differences on the $\log_2$ scale implies that *cis*-regulatory divergence completely explains the expression difference between parents, and that *trans*-regulatory variants are absent. From this it follows that screening for genes with an allelic expression difference caused by a *cis*-regulatory variant in a particular hybrid cross $i \times j$ implies the screening for imbalances in allelic expression, $g_i \neq g_j$, in the absence of gametic interaction, i.e. $s_{ij} = 0$. The strategy is outlined in Figure 2 and aspects of the individual components are discussed below.

To apply GASED, we examined transcript levels in RNA samples collected from a diallel experiment in *Arabidopsis* with 5 parental lines and 10 F_1 hybrids (Figure 1). The hybrid samples consisted of a pooled progeny from reciprocal crosses. To assess the genotype and GCA and SCA variability, and to generate a robust set of data, we analyzed gene expression from two independent samples for each genotype. In total, we carried out 30 hybridizations on CATMA v2.2 arrays (24). This yielded expression data for ~50% of the genes on the array (10 877 out of 22 494 gene features).

Among these 10 877 expressed genes, we first identified those with a significant genetic variance component in their transcript abundance. The expression levels of 6838 genes (62.9%) were significantly different across the 15 genotypes ($P < 0.05$). To correct for multiple testing, we evaluated the FDR at several levels of significance. At $P = 0.05$, the FDR was ~2%, corresponding to 6701 expected true positives. At a more stringent level of $P = 10^{-3}$, the FDR dropped below 0.1%, producing ~3 false discoveries among 4066 features displaying significant genetic variation in their expression. Henceforth, as a rule of thumb, we chose a P-level that produced ⩽10 false discoveries among all significant features as a cut-off for significance.

Next, we fitted a mixed model containing the GCA and SCA variance components to the expression levels of the 4066 genes by REML. According to the GASED procedure, screening for genes with an allelic expression difference caused by a *cis*-regulatory variant implies the screening for transcripts with $s_{ij} = 0$ and $g_i \neq g_j$ in a particular hybrid crossing $i \times j$. In a first step, therefore, t-statistics were calculated for each of the 4066 genes in each of the 10 hybrids to test for the SCA effect differing from zero. Because we were

Table 1. Number of features with a non-significant SCA effect, a significant difference in allelic mean expression, and a significant 1.5- and 2-fold change in allelic expression detected in each of the 10 hybrids

Hybrid	$s_{ij} = 0$ ($Q_a < 0.0005$)	$g_i \neq g_j$ ($Q < 0.001$)	$g_i \neq g_j$ $\lvert\Delta^1\rvert = 0.292$ ($Q < 0.001$; $P_1 > 0.1$)	$g_i \neq g_j$ $\lvert\Delta^1\rvert = 0.5$ ($Q < 0.001$; $P_1 > 0.1$)
Col×L*er*	2051	245	87	31
Col×Cvi	2804	448	218	77
Col×Ws	1196	191	89	41
Col×C24	1222	262	108	39
L*er*×Cvi	1944	368	151	41
L*er*×Ws	2796	257	96	29
L*er*×C24	2292	375	175	51
Cvi×Ws	1718	368	155	46
Cvi×C24	1301	340	150	46
Ws×C24	2346	614	210	60
Total	4047	1574	715	236

interested in selecting a subset of genes for which the null of no interaction, i.e. $s_{ij} = 0$, is likely, we selected genes lacking evidence against the null based on the FNR (33,34) instead of the FDR. Whereas FDR controls the number of false positives, i.e. interactions that are truly null but are falsely declared as significant, the FNR quantity expresses the fraction of false negatives among the genes for which the null is not rejected, i.e. interactions that are truly non-null but are falsely declared as non-significant. Hence, calculated *P*-values were subsequently transformed into point-wise FNRs, Q_a-values, taking into account the total number of 40 660 features tested simultaneously. The number of genes displaying no *trans*-acting effect on their expression are tabulated per hybrid in Table 1.

Next, we aimed to estimate the GCA effects for each gene with a non-significant SCA effect in a particular hybrid and focused on the difference between GCA effects. Rejecting H_0: $g_i = g_j$ resulted in a total of 1574 genes displaying significant ASE differences across the 10 hybrids at a *Q*-value of <0.001 (Table 1 and Supplementary Table 2). However, despite the fact that statistical thresholds provide a better basis to identify a subset of genes for further analysis than does a commonly accepted arbitrary cut-off of fold change in expression, small *P*- or *Q*-values do not automatically imply a sizable difference in effects. Therefore, we also incorporated the requirement of a sizable change, i.e. 1.5- and 2-fold change in allelic expression on the original scale or, as derived before, a difference in GCA effects of 0.292 and 0.5 when data are on the $\log_2$ scale. To avoid missing potentially interesting allelic differences, e.g. differences that fell short of an observed difference in GCA effects of 0.292, we complemented the *Q*-value with a one-sided 'alternative *P*-value', referred to as P_1 (36). Here, P_1 measures how likely it is to observe a difference in GCA effects as small as or smaller than the one observed when the target difference, e.g. 0.292, is true. Of the 1574 genes displaying significant ASE differences ($Q < 0.001$), a total of 715 and 236 genes were likely to carry allelic polymorphisms responsible for at least a 1.5- and 2-fold change in allelic expression, respectively, at a P_1 significance of at least 0.10 (because P_1 values summarize evidence against a target alternative in the direction of the null hypothesis, we looked for large P_1 values) (Table 1 and Supplementary Table 1).

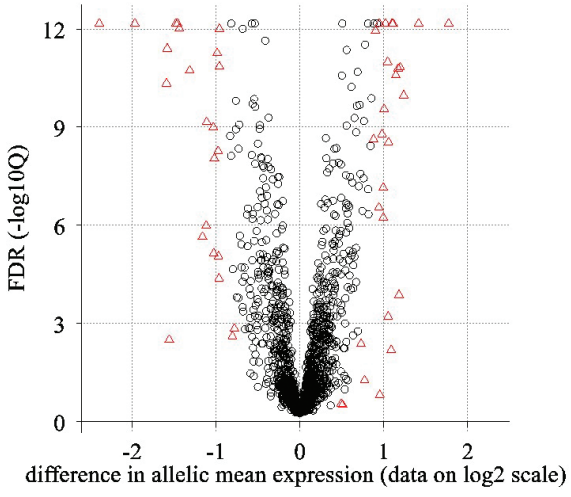

Figure 3. Volcano plot contrasting the significance ($-\log_{10}Q$ on the ordinate) and the magnitude of the difference in allelic mean expression between L*er* and Cvi (data on $\log_2$ scale). The bottom horizontal dashed line corresponds to the FDR acceptance level of $Q = 0.001$ ($-\log_{10} = 3$). The vertical dashed lines demarcate the 0-, 2- and 4-fold change in allelic expression (data on the $\log_2$ scale). Triangles represent the 50 genes with $s_{ij} = 0$ at the FNR acceptance level of 0.0005 and carrying evidence in favor of a 2-fold difference in allelic mean expression, of which 41 genes met the FDR acceptance level of $Q = 0.001$. Circles represent genes of which the allelic expression is not affected by *trans*-regulatory variants ($s_{ij} = 0$; FNR < 0.0005), but do not carry evidence in favor of a 2-fold difference in allelic mean expression.

The significance and the magnitude of the change in allelic expression detected in the L*er*×Cvi hybrid are visualized in Figure 3. This volcano plots contrasts significance on the $-\log_{10}(Q)$ scale against difference in allelic mean expression on the $\log_2$ scale. This plot illustrates that a decision criterion based on the observed difference in allelic mean expression, i.e. >1 in absolute value, combined with the *Q*-value <0.001 was too conservative: not all genes that carried evidence in favor of the alternative, i.e. a 2-fold allelic mean expression, had an observed effect of that magnitude. Therefore, as illustrated by the volcano plot, it is useful to base a decision criterion both on the *Q*-values and the P_1-values.

We then sought to confirm these results by expression QTL (eQTL) mapping. Because an allelic expression difference

caused by *cis*-regulatory variant implies a nearby polymorphism that controls expression of the allele in LD, we expected the expression level of such a gene, when treated as a quantitative trait, to display linkage to its own locus. Furthermore, we also anticipated the allele preferentially expressed in hybrids to be associated with higher expression in the segregants. We examined this for the 41 genes identified as displaying a 2-fold allelic expression difference in the L*er*×Cvi hybrid, a parent pair for which an RIL population is available as well as a detailed AFLP linkage map (22). Expression profiles collected from 18 L*er*/Cvi RILs for the 41 genes were carried through a linkage analysis with 69 markers defining an equal number of genomic bins. Within a mixed model framework, we partitioned the total genetic variation into a fixed and a random part associated with the genomic bin containing the gene in question and the genetic background consisting of possible *trans*-acting eQTL located elsewhere in the genome, respectively. For 31 genes (76%) expression levels displayed the strongest linkage to their own locus at $P < 0.05$, confirming an allelic polymorphism responsible for the allelic expression difference (Figure 4). All of the 31 *cis*-eQTL display higher expression of the allele predicted to be preferentially expressed. For nine genes (22%), expression differences showed the strongest linkage to loci distant to their own locus at $P < 0.05$, and accordingly, failed to confirm their *cis*-regulated ASE as judged by the GASED analysis (Figure 4). Linkage to *trans*-acting loci in three out of these nine cases was associated with elevated expression of the opposite allele. For one gene, a significant linkage could be detected neither to the target *cis*-locus nor to *trans*-loci.

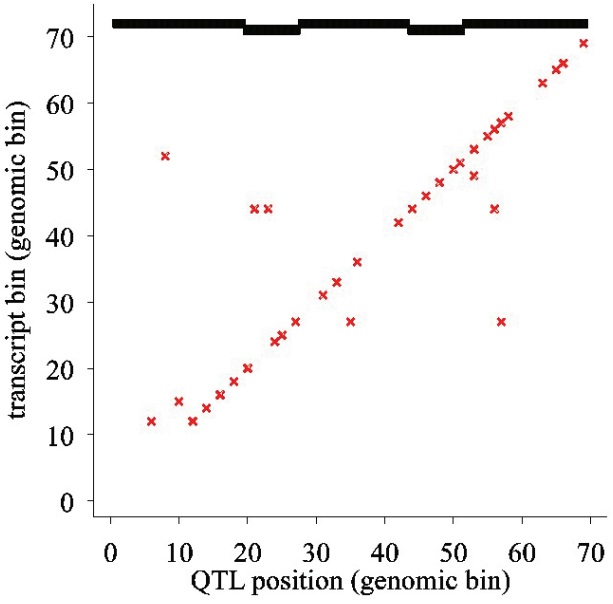

Figure 4. Mapping eQTL that modulate the expression of 41 genes identified in the L*er*×Cvi hybrid as having an allelic polymorphism responsible for a 2-fold difference in allelic mean expression. Variation in the transcript levels across 18 L*er*/Cvi RIL samples was correlated with the presence of the L*er* or Cvi allele at the target genomic bin. Each cross represents a single transcript. The physical position of each transcript is indicated on the ordinate, and the position of the locus most strongly associated with the variation of the transcript levels on the abscissa. Transcripts on the diagonal display the strongest association to their own locus at $P < 0.05$, confirming an allelic polymorphism responsible for the change in allelic mean expression. Transcripts off the diagonal display the strongest association to a locus distant to their own locus at $P < 0.05$, failing to confirm their *cis*-regulated ASE as judged by the GASED procedure. To represent the data graphically, the *Arabidopsis* genome was divided into 69 genomic bins. Actual chromosomal positions are indicated at the top.

DISCUSSION

The science of quantitative genetics has been around for a long time. It has served as the theoretical basis for most plant and animal breeding programs for well over a half century. Quantitative geneticists have become interested recently in applying quantitative genetic methodologies to estimate the genetic variance and heritability of gene expression and to detect eQTL. Wayne *et al*. (37) showed how a quantitative description of variation in mRNA abundance can be presented in terms of GCA, and a number of studies (3–10) have shown that regulatory polymorphisms in *cis*- and *trans*-affect gene expression [for a recent review on the quantitative genetics of transcription see Gibson and Weir (38)]. We have shown that in a context of gene expression, empirical estimates of GCA and SCA generated by a diallel design are valid parameters in large-scale detection of transcripts whose abundance is regulated by strong *cis*-acting variants. Compared to other ASE detection methods, GASED has major advantages. First, allelic variants in multiple genetic backgrounds can be examined at a large number of genes. Second, in contrast with the positional ASE detection methods, such as eQTL mapping, GASED is not affected by local or nearby *trans*-acting variants in LD with the *cis*-acting variants in question. Therefore, GASED provides a complementary strategy to eQTL mapping for identifying *cis*-regulated gene expression and leads to a more accurate identification process of the truly *cis*-acting QTL. Third, GASED detects ASE in a non-mechanistic way and, hence, is not restricted to genes with allelic transcript variants, and this is the major strength of the GASED approach.

Our results suggest that a considerable number of the ~30 000 *Arabidopsis* genes (39) contain functional regulatory variants that affect expression levels by at least 1.5-fold among the five *Arabidopsis* accessions studied. We found indication of such regulatory variants in 715 of the 10 877 genes expressed, corresponding to a frequency of ~7%, which is in the order of frequencies calculated in mouse (13). Such a frequency probably underestimates the true portion of genes that harbor *cis*-regulatory variants because only one developmental stage was examined under a particular environmental condition. A survey of more developmental stages, environmental conditions and, moreover, single tissues rather than whole organs would probably reveal more variation.

Linkage mapping of transcript abundance in a limited set of RILs directed at 41 genes confirmed 31 cases identified to contain functional *cis*-regulatory variants that affect expression levels at least 2-fold among the two accessions L*er* and Cvi. All 31 genes were confirmed in the predicted direction. At least two plausible explanations can be provided for the genes whose expression levels did not provide evidence of linkage. First, when testing the SCA interaction,

the power was relatively low given the low number ($n = 4$) of observations per hybrid in this study. More observations per hybrid would increase the power to decrease the number of false negatives, i.e. genes with allelic expression differences caused by *trans*-regulatory variants, but identified as apparently *cis*-regulated. Second, as SCA variance does not account for additive × additive epistasis coming from a *cis*–*trans* interaction, the GASED procedure may result in wrong detection of *cis*-controlled ASE. It should be noted that, even if the false-positive rate is 9/41, the GASED approach is superior to the large-scale detection of ASE with oligonucleotide arrays (17) with a false-positive rate of 6/11, and to the screening of unselected gene sets (13) that detects ASE in only 4 of the 69 genes studied. Thus, a pre-screening approach, such as GASED, that leads to at least a 1.5-fold enrichment can accelerate systematic surveys of *cis*-regulatory variation. Notably, GASED can also be used to explore naturally occurring *cis*-regulatory variation among inbred lines for laboratory animals, such as mouse, rat and fruitfly. Such surveys should ultimately identify *cis*-regulatory variants to be examined in association studies of complex traits.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was partly supported by the Ministerie van de Vlaamse Gemeenschap-Landbouwkundig Onderzoek (IWT/020716), the European Commission within its FP5 Programme (thematic area 'Quality of Life', CAGE project QLK3-CT-2002-02035) and the Interuniversity Poles of Attraction Programme—Belgian Science Policy (P5/13). R.K. and R.S. are indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship. Funding to pay the Open Access publication charges for this article was provided by the Department of Molecular Genetics of the Ghent University.

Conflict of interest statement. None declared.

REFERENCES

1. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Rev. Genet*., **5**, 276–287.
2. Pastinen,T. and Hudson,T.J. (2004) *Cis*-acting regulatory variation in the human genome. *Science*, **306**, 647–650.
3. Brem,R.B., Yvert,G., Clinton,R. and Kruglyak,L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.
4. Schadt,E.E., Monks,S.A., Drake,T.A., Lusis,A.J., Che,N., Colinayo,V., Ruff,T.G., Milligan,S.B., Lamb,J.R., Cavet,G. *et al*. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.
5. Bystrykh,L., Weersing,E., Dontje,B., Sutton,S., Pletcher,M.T., Wiltshire,T., Su,A.I., Vellenga,E., Wang,H., Manly,K.F. *et al*. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genet*., **37**, 225–232.
6. Chesler,E.J., Lu,L., Shou,S., Qu,Y., Gu,J., Wang,J., Hsu,H.C., Mountz,J.D., Baldwin,N.E., Langston,M.A. *et al*. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genet*., **37**, 233–242.
7. Hubner,N., Wallace,C.A., Zimdahl,H., Petretto,E., Schulz,H., Maciver,F., Mueller,M., Hummel,O., Monti,J., Zidek,V. *et al*. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genet*., **37**, 243–253.
8. Morley,M., Molony,C.M., Weber,T.M., Devlin,J.L., Ewens,K.G., Spielman,R.S. and Cheung,V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
9. DeCook,R., Lall,S., Nettleton,D. and Howell,S.H. (2006) Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics*, **172**, 1155–1164.
10. Vuylsteke,M., Van Den Daele,H., Vercauteren,A., Zabeau,M. and Kuiper,M. (2006) Genetic dissection of transcriptional regulation by cDNA-AFLP. *Plant J*., **45**, 439–446.
11. Ronald,J., Brem,R.B., Whittle,J. and Kruglyak,L. (2005) Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet*., **1**e25 (0213–0222).
12. Doss,S., Schadt,E.E., Drake,T.A. and Lusis,A.J. (2005) *Cis*-acting expression quantitative trait loci in mice. *Genome Res*., **15**, 681–691.
13. Cowles,C.R., Hirschhorn,J.N., Altshuler,D. and Lander,R.S. (2002) Detection of regulatory variation in mouse genes. *Nature Genet*., **32**, 432–437.
14. Guo,M., Rupe,M.A., Danilevskaya,O.N., Yang,X. and Hu,Z. (2003) Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. *Plant J*., **36**, 30–44.
15. Knight,J.C. (2004) Allele-specific gene expression uncovered. *Trends Genet*., **20**, 113–116.
16. Lo,H.S., Wang,Z., Hu,Y., Yang,H.H., Gere,S., Buetow,K.H. and Lee,M.P. (2003) Allelic variation in gene expression is common in the human genome. *Genome Res*., **13**, 1855–1862.
17. Ronald,J., Akey,J.M., Whittle,J., Smith,E.N., Yvert,G. and Kruglyak,L. (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res*., **15**, 284–291.
18. Yan,H., Yuan,W., Velculescu,V.E., Vogelstein,B. and Kinzler,W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
19. Wittkopp,P.J., Haerum,B.K. and Clark,A.G. (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature*, **430**, 85–88.
20. Searle,S.R., Casella,G. and McCulloch,C.E. (1992) *Variance Components, Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*. Wiley, New York, NY.
21. Lynch,M. and Walsh,B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.
22. Alonso-Blanco,C., Peeters,A.J., Koornneef,M., Lister,C., Dean,C., van den Bosch,N., Pot,J. and Kuiper,M.T. (1998) Development of an AFLP based linkage map of L*er*, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a L*er*/Cvi recombinant inbred line population. *Plant J*., **14**, 259–271.
23. Vision,T.J., Brown,D.G., Shmoys,D.B., Durrett,R.T. and Tanksley,S.D. (2000) Selective mapping: a strategy for optimizing the construction of high-density linkage maps. *Genetics*, **155**, 407–420.
24. Hilson,P., Allemeersch,J., Altmann,T., Aubourg,S., Avon,A., Beynon,J., Bhalerao,R.P., Bitton,F., Caboche,M., Cannoot,B. *et al*. (2004) Versatile gene-specific sequence tags for *Arabidopsis* functional genomics: transcript profiling and reverse genetics applications. *Genome Res*., **14**, 2176–2189.
25. Thareau,V., Déhais,P., Serizet,C., Hilson,P., Rouzé,P. and Aubourg,S. (2003) Automatic design of gene-specific sequence tags for genome-wide functional studies. *Bioinformatics*, **19**, 2191–2198.
26. Allemeersch,J., Durinck,S., Vanderhaeghen,R., Alard,P., Maes,R., Seeuws,K., Bogaert,T., Coddens,K., Deschouwer,K., Van Hummelen,P. *et al*. (2005) Benchmarking the CATMA microarray: a novel tool for *Arabidopsis* transcriptome analysis. *Plant Physiol*., **137**, 588–601.
27. Boyes,D.C., Zayed,A.M., Ascenzi,R., McCaskill,A.J., Hoffman,N.E., Davis,K.R. and Görlach,J. (2001) Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell*, **13**, 1499–1510.
28. Baird,D., Johnstone,P. and Wilson,T. (2004) Normalization of microarray data using a spatial mixed model analysis which includes splines. *Bioinformatics*, **20**, 3196–3205.

29. Payne,R.W. and Lane,P.W. (2005) *GenStat® Release Reference Manual, Part 3: Procedure Library PL16*. VSN International, Oxford, UK.
30. Vuylsteke,M., van Eeuwijk,F., Van Hummelen,P., Kuiper,M. and Zabeau,M. (2005) Genetic analysis of variation in gene expression in *Arabidopsis thaliana*. *Genetics*, **171**, 1267–1275.
31. Wolfinger,R.D., Gibson,G., Wolfinger,E.D., Bennett,L., Hamadeh,H., Bushel,P., Afshari,C. and Paules,R.S. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.
32. Allison,D.B., Gadbury,G.L., Heo,M., Fernández,J.R., Lee,C.-K., Prolla,T.A. and Weindruch,R. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.*, **39**, 1–20.
33. Genovese,C. and Wasserman,L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc. B*, **64**, 499–517.
34. Taylor,J., Tibshirani,R. and Efron,B. (2005) The 'miss rate' for the analysis of gene expression data. *Biostatistics*, **6**, 111–117.
35. Moerkerke,B. and Goetghebeur,E. (2006) Selecting significant differentially expressed genes from the combined perspective of the Null and the Alternative. *J. Comput. Biol.*, in press.
36. Moerkerke,B., Goetghebeur,E., De Riek,J. and Roldán-Ruiz,I. (2006) Significance and impotence: towards a balanced view of the null and the alternative hypotheses in marker selection for plant breeding. *J. R Statist. Soc. A*, **169**, 61–79.
37. Wayne,M.L., Pan,Y.-J., Nuzhdin,S.V. and McIntyre,L.M. (2004) Additivity and trans-acting effects on gene expression in male *Drosophila simulans*. *Genetics*, **168**, 1413–1420.
38. Gibson,G. and Weir,B. (2005) The quantitative genetics of transcription. *Trends Genet.*, **21**, 616–623.
39. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.