the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2009) **57**, 184–194 doi: 10.1111/j.1365-313X.2008.03662.x

TECHNICAL ADVANCE

Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes

Naïra Naouar[1,2], Klaas Vandepoele[1,2], Tim Lammens[1,2], Tineke Casneuf[1,2], Georg Zeller[3,4], Paul van Hummelen[5], Detlef Weigel[3], Gunnar Rätsch[4], Dirk Inzé[1,2], Martin Kuiper[1,2], Lieven De Veylder[1,2,†] and Marnik Vuylsteke[1,2,*,†]

[1]*Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Ghent, Belgium,*

[2]*Department of Molecular Genetics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium,*

[3]*Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany,*

[4]*Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany, and*

[5]*VIB Microarray Facility, UZ-Gasthuisberg, O&N, Leuven 3000, Belgium*

Received 20 March 2008; revised 5 August 2008; accepted 8 August 2008; published online 19 September 2008.

[*]For correspondence (fax +32 9 3313809; e-mail marnik.vuylsteke@psb.ugent.be).

[†]The last two authors should be regarded as joint Senior Authors.

Summary

The Affymetrix ATH1 array provides a robust standard tool for transcriptome analysis, but unfortunately does not represent all of the transcribed genes in *Arabidopsis thaliana*. Recently, Affymetrix has introduced its Arabidopsis Tiling 1.0R array, which offers whole-genome coverage of the sequenced Col-0 reference strain. Here, we present an approach to exploit this platform for quantitative mRNA expression analysis, and compare the results with those obtained using ATH1 arrays. We also propose a method for selecting unique tiling probes for each annotated gene or transcript in the most current genome annotation, TAIR7, generating Chip Definition Files for the Tiling 1.0R array. As a test case, we compared the transcriptome of wild-type plants with that of transgenic plants overproducing the heterodimeric E2Fa-DPa transcription factor. We show that with the appropriate data pre-processing, the estimated changes per gene for those with significantly different expression levels is very similar for the two array types. With the tiling arrays we could identify 368 new E2F-regulated genes, with a large fraction including an E2F motif in the promoter. The latter groups increase the number of excellent candidates for new, direct E2F targets by almost twofold, from 181 to 334.

Keywords: Tiling array, ATH1 array, Arabidopsis, expression analysis, E2F, cell cycle.

Introduction

Recent advances in microarray technologies have resulted in the commercial availability of tiling arrays, making it feasible to interrogate whole genomes in an unbiased way. Probes on tiling arrays either partially overlap one another (true tiling) or are spaced at regular intervals. These arrays are useful for several purposes, and can be used to analyze DNA content, as well as mRNA content. Whereas focused gene expression microarrays seek to measure the relative abundance of transcripts derived from a specifically targeted set of annotated sequences, tiling arrays can also be used, for example, to discover transcribed genomic regions that are independent of previous annotations, to detect non-coding RNA transcripts or to identify alternative RNA isoforms of known genes (Bertone *et al.*, 2004; Kapranov *et al.*, 2002).

The GeneChip® Arabidopsis Tiling 1.0R array, commercially available from Affymetrix, is a single array with over 3.2 million perfect-match and mismatch (PM/MM) probe pairs that are tiled across the complete non-repetitive *Arabidopsis thaliana* genome. The 25-mer probes are spaced (on average) 35 bases apart, as measured from the central position of adjacent 25-mer oligonucleotides, leaving a gap of 10 bases between adjacent probes (http://www.affymetrix.com/products/arrays/specific/arab_tiling.affx). The basis for the design was the NCBI Arabidopsis genome assembly (version 5), and mitochondrial and chloroplast sequences were included as well (NCBI accession numbers NC_001284 and NC_000932). Because of the design, only very limited consideration could be given to probe behavior in hybridiza-

tion. In contrast, the Arabidopsis ATH1 array contains sets of 11 25-mer probes that were designed to uniquely cover 3′ exons of *A. thaliana* genes, based on the now outdated TIGR 3.0 annotation (Redman *et al.*, 2004). In addition to trying to avoid cross hybridization with other genes, probes were chosen such that their hybridization characteristics were as similar as possible. A further potential advantage of the ATH1 array is that probes are all antisense to the annotated transcripts, whereas probes on the Tiling 1.0R are all from one strand of the genome only, such that half of all genes will be represented by antisense probes, and the other half will be represented by sense probes. Thus, to analyze the expression of all genes, the Tiling 1.0R array needs to be hybridized with double-stranded cDNA.

For several years we have been interested in understanding the role of E2F transcription factors in regulating the plant cell cycle and plant growth. E2Fs are conserved regulators of *S* phase-specific genes (Blais and Dynlacht, 2007; Iaquinta and Lees, 2007). The *A. thaliana* genome encodes three E2Fs (E2Fa, E2Fb and E2Fc; De Veylder *et al.*, 2007), which are active in association with the dimerization partners DPa or DPb. A complete understanding of the role of the different E2F isoforms requires the comprehensive identification of their target genes. To this end, we have previously generated plants ectopically overproducing the heterodimer E2Fa-DPa (*E2Fa-DPa*^{OE}; De Veylder *et al.*, 2002; Kosugi and Ohashi, 2003). By interrogating the transcriptome of *E2Fa-DPa*^{OE} plants, which display ectopic cell division and endoreduplication, with ATH1 microarrays, we had identified 181 potentially direct E2F target genes, with many of them encoding proteins involved in cell cycle progression, DNA replication and chromatin dynamics (Vandepoele *et al.*, 2005; Vlieghe *et al.*, 2003).

Here, we describe a procedure to extract and pre-process the expression data from Tiling 1.0R arrays. We use this procedure to benchmark their performance for gene expression studies, taking the results obtained with the ATH1 microarray as a reference. The focus is on quantifying, comparing and evaluating the expression changes between wild-type (WT) and *E2Fa-DPa*^{OE} plants analyzed with both microarray platforms. In addition, we compared a number of different target preparation protocols, in particular with respect to the potential for obtaining quantitative results. Applying our newly developed procedures, we demonstrate that the number of E2F target genes can be significantly enlarged using Affymetrix Tiling 1.0R array analysis.

Results

Annotation

All probes present on the GeneChip® Arabidopsis Tiling 1.0R array were mapped to the genome, and each gene represented in the TAIR7 annotation was characterized by its own set of unique exonic probes (Figure 1). From a total of 31 762 genes/transcripts annotated in TAIR7, the Tiling 1.0R array probes covered either the forward or reverse strand of 29 890 genes. We did not further consider 128 genes that are represented by only one or two probes. Of the remaining 29 767 genes, 20 654 genes are in common with the latest ATH1 chip definition file (CDF), whereas 9113 are only represented on the Tiling 1.0R array.

Effect of different target preparation strategies on the hybridization

We compared three different target preparation protocols, in particular with respect to the extent in which they influence the hybridization signal on the Tiling 1.0R arrays (Figure 2). In Tiling Target Preparation protocol 1 (TTP1), cDNA was prepared in a manner identical as for the ATH1 Target Preparation (ATP), relying on oligo-dT primed cDNA preparation, followed by linear amplification of the cDNA by T7 *in vitro* transcription. TTP2 used total RNA directly reverse transcribed into cDNA. TTP3 used randomly primed cDNA synthesis, starting with total RNA depleted of ribosomal RNA. The distributions of the raw data intensities (Figure 3) are similar for ATP, TTP1 and TTP3, whereas the raw data intensities for TTP2 are lower overall. Because in TTP2 random primers were used for cDNA synthesis, without any prior attempts to remove rRNA, it is very likely that a significant portion of the labeled cDNA will be derived from rRNA, yielding a lower signal intensity overall for non-repetitive probes.

Cross-platform comparison

The set of 20 654 common genes was used to assess the agreement in differential expression measured on both microarray platforms. Initially, we restricted the platform comparison to significantly differentially expressed genes showing large fold changes, which are likely to be of more biological relevance than those showing small, although significant, fold changes. Therefore, all genes showing a significant ($Q < 0.05$), at least twofold change ($FC \geq 2$) in hybridization signal between WT and *E2Fa-DPa*^{OE} plants were considered to be differentially expressed (Table 1). Robust multiarray analysis (RMA) analysis of ATH1 data identified 1562 differentially expressed genes. The three TTP protocols yielded comparable numbers of differentially expressed genes. TTP3, using random primers after depletion of rRNA using locked nucleic acid (LNA) oligonucleotides (see Experimental procedures for further details), produced the highest number of differentially expressed genes, as well as the largest overlap with the ATH1 results. Inclusion of untranslated region (UTR) probes reduced the numbers of differentially expressed genes identified, and hence reduced the overlap with ATP.

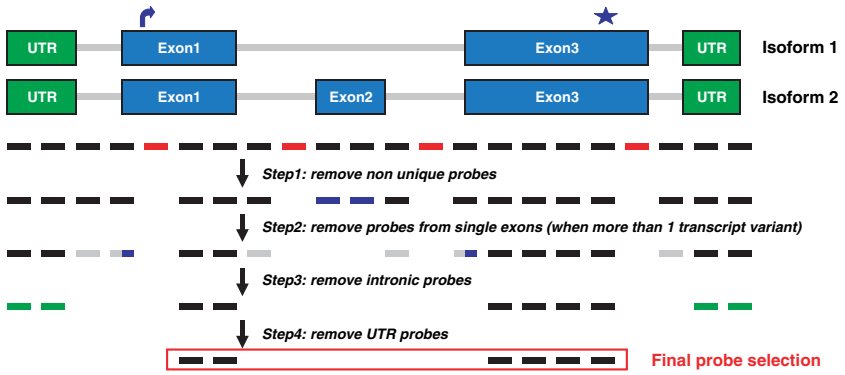

Figure 1. Gene model and probe selection for the chip definition file (CDF).
The 25-mer perfect match (PM) probes with 10-base spacing on average are represented by blocks. Red blocks represent non-unique probes; grey blocks represent intronic probes; green blocks represent untranslated region (UTR) probes; and blue blocks represent probes from unique exons. The final step discriminates between the two CDF variants generated.

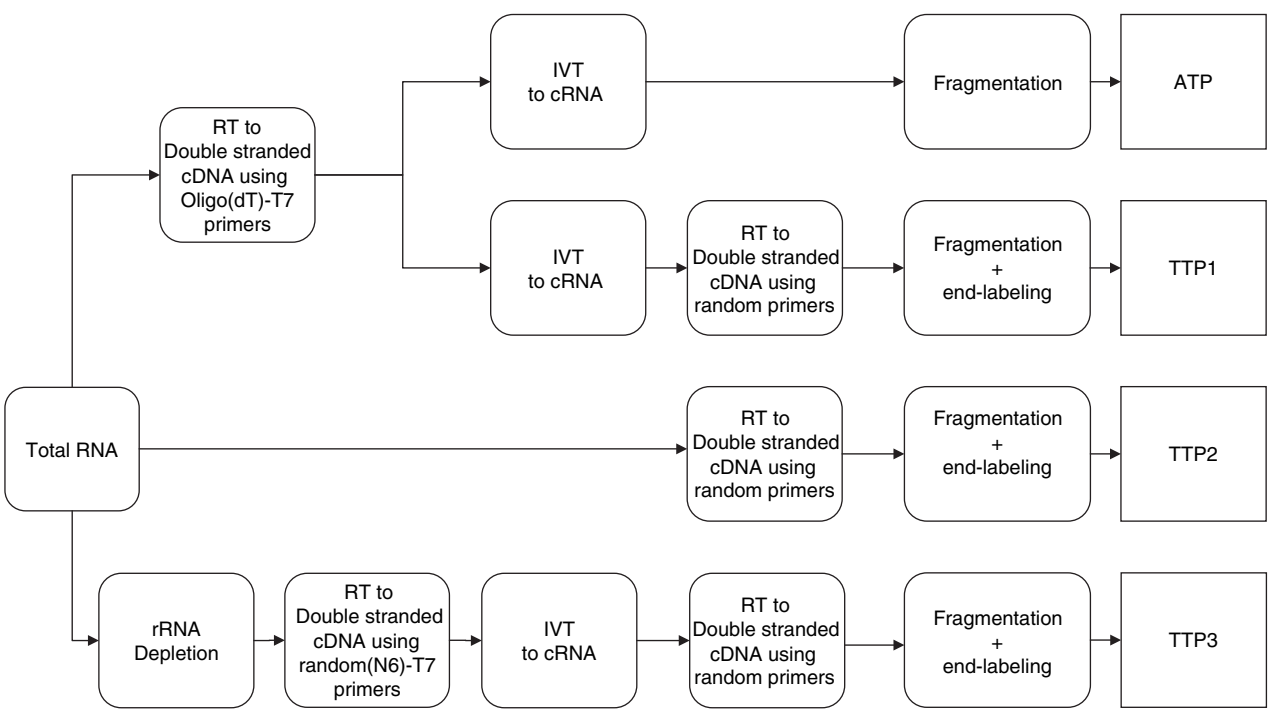

Figure 2. Flow chart showing the Affymetrix ATH1 (ATP) and Tiling 1.0R target preparation protocols (TTP1, TTP2 and TTP3).
Each of the RNA samples was split into separate aliquots that were processed according to the steps shown, leading to the synthesis of four unique labeled targets. ATP and TTP1 derive from amplified cRNA from the same initial oligo(dT) cDNA preparation. These targets therefore primarily represent poly(A) mRNAs. TTP2 and TTP3 were produced with the standard cDNA labeling procedure, as recommended by Affymetrix for tiling arrays, except that in the procedure for TTP3, rRNA was depleted with the Ribominus kit from Invitrogen (see Experimental procedures for further details).

Analyzing the expression data using a mixed model approach increases the proportion of differentially expressed genes in common between ATP and the three TTP strategies, in particular between ATP and TTP2 (Table 1). Again, in all three cases, the highest proportion of differentially expressed genes in common between the two platforms was found when UTR probes were excluded.

Alternatively, we assessed the agreement in the magnitude of fold changes measured on both microarray platforms, irrespective of their statistical significance, by means of a correspondence at the top (CAT) plot. The CAT plots (Figure 4a) show that the ranking of genes based on the fold changes estimated by RMA are highly consistent across the two platforms. The correspondence slightly differs with the

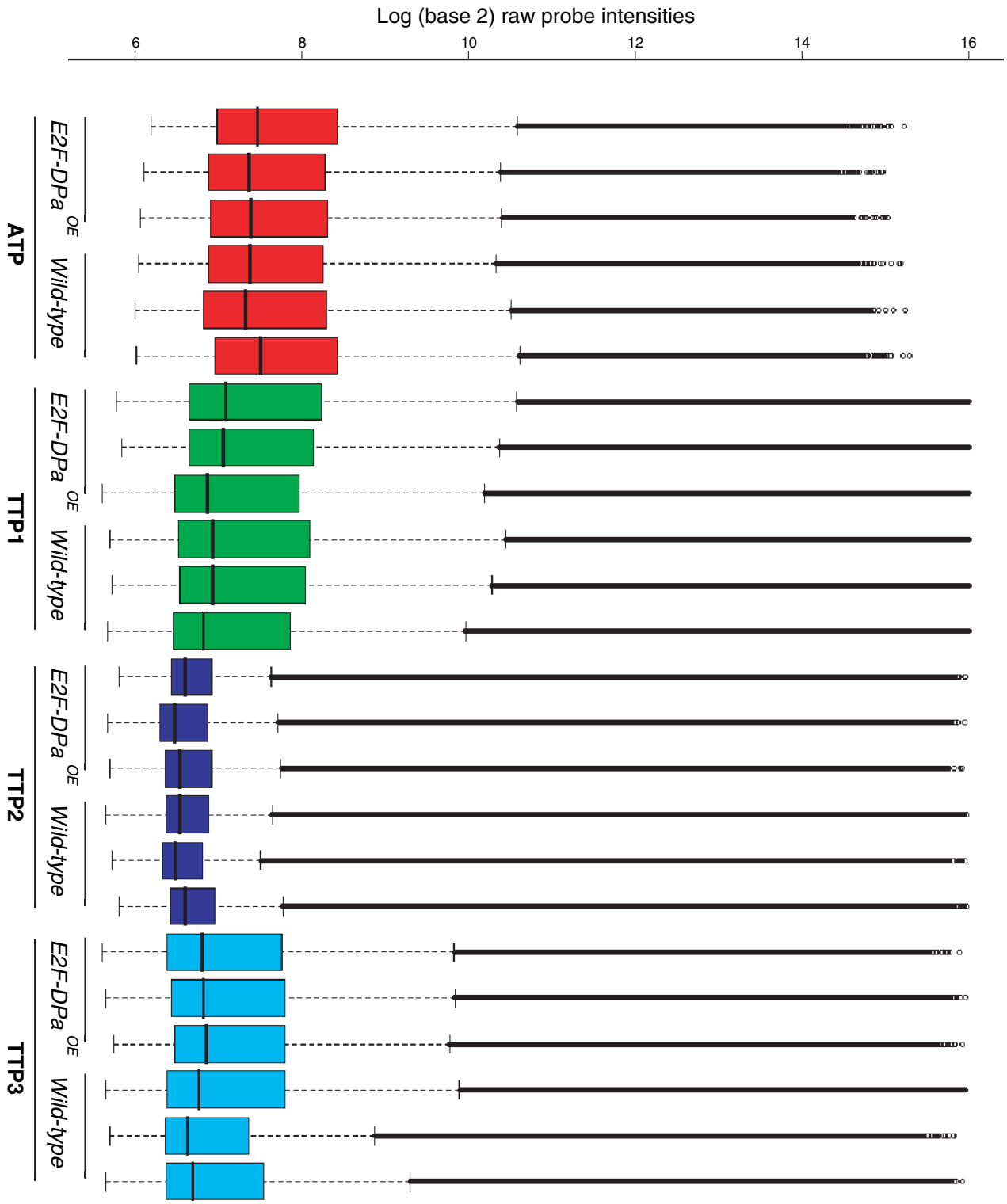

Figure 3. Box plot of log (base 2) raw probe intensities measured for *E2Fa-DPa*^{OE} (first three boxes) and wild-type plants (next three boxes), each in triplicate, on ATH1 arrays (red), and for the three Tiling 1.0R approaches, TTP1 (green), TTP2 (dark blue) and TTP3 (light blue).

Table 1 Number of differentially expressed genes between wild type and *E2Fa-DPa*^{OE} identified using ATH1 and Tiling 1.0R microarrays

Microarray platform	ATH1	Tiling 1.0R microarray					
		TTP1		TTP2		TTP3	
Target amplification protocol	ATP	With UTR	Without UTR	With UTR	Without UTR	With UTR	Without UTR
RMA							
No. differentially expressed genes	1562	1463	1549	1338	1503	1816	1934
No. differentially expressed genes in common with ATP		1079	1111	982	1026	1220	1253
Mixed models							
No. differentially expressed genes	1635	1543	1658	1969	2442	1953	2123
No. differentially expressed genes in common with ATP		1091	1150	1238	1348	1269	1329

For Tiling arrays, three different target preparation (TPP1, TTP2 and TTP3) strategies were applied.
Genes were considered to be differentially expressed when a significant ($Q < 0.05$) two or higher fold change in gene expression was estimated. Expression data were analyzed using the RMA and the mixed model approach. The genes taken into account in this analysis were the 20 654 genes shared in the two platforms.

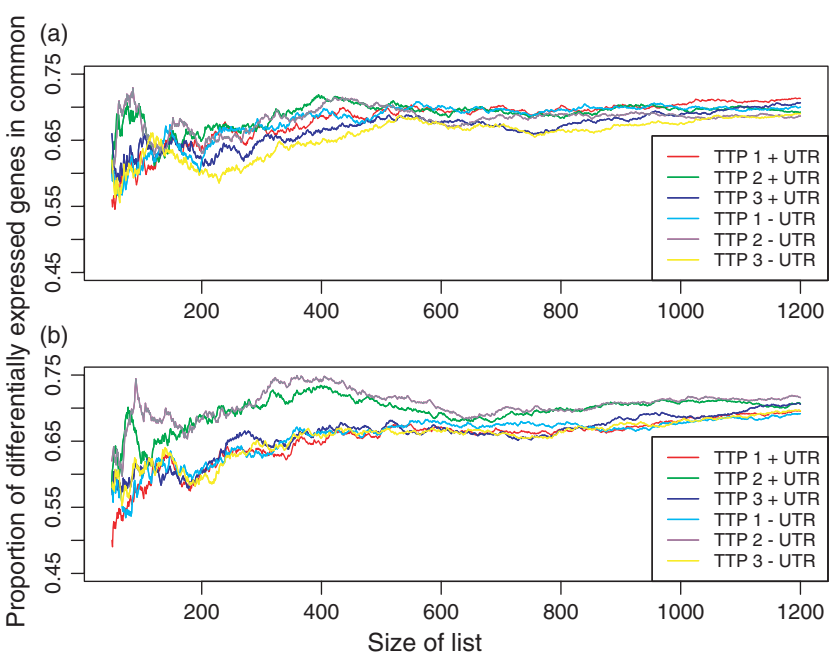

Figure 4. Correspondence at the top (CAT) plots showing agreement in differential expression calls, based on fold changes, between ATH1 and the three Tiling 1.0R experiments.
The fold changes were estimated by either the robust multiarray analysis (RMA) approach (a) or the mixed model approach (b).

nature of the TTP approach used, whereas inclusion of UTR probes has little effect. Again, the overall correspondence in the 'top' genes between the two platforms turns out to be slightly higher when fold changes are estimated by a mixed model approach (Figure 4b).

Cross-TTP comparison

Cross-TTP comparison was performed on the basis of 29 767 genes covered by the Tiling 1.0R array probes. Comparing the different target preparation protocols with respect to the extent to which they affect the number of differentially expressed genes identified shows that TTP2 and TTP3 are superior to TTP1. A possible explanation is that TTP1 selected only for polyadenylated transcripts, whereas TTP2 and TTP3 also amplified the non-polyadenylated transcripts. The number of differentially expressed genes in common between the three TTP approaches is clearly higher when fold changes and statistical significance are assessed by a mixed model approach (Figure 5a,b). The genes that are not in the overlap might be indicative of specific characteristics of the different protocols, such as differences in sensitivity, selectivity or RNA coverage.

To ensure that the differences in the number of differentially expressed genes identified by the three TTP approaches is not caused by error or bias in the sample

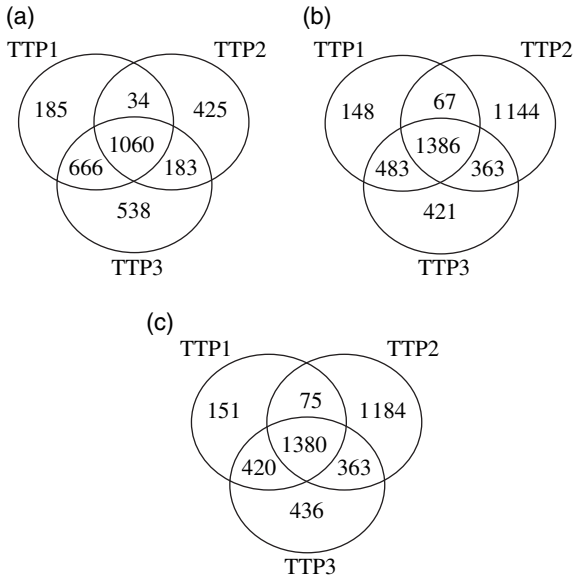

Figure 5. Venn diagram representing the overlapping, differentially expressed genes for the three Tiling 1.0R (TTP) protocols.
The differentially expressed genes were identified by: (a) the robust multiarray analysis (RMA) procedure; (b) the mixed model approach, with significance assessed on the basis of the *F* approximation to the Wald statistic; and (c) the mixed model approach with significance assessed by permutation. Untranslated region (UTR) probes had been removed from the probe-set selection.

preparation, resulting in uneven amplification between samples and increased false-positives passing the biological and statistical significance thresholds, we performed permutation analyses. Permutation based *P* values and the corresponding *Q* values, however, did not differ strongly from those calculated on the basis of the *F* approximation of the Wald statistic, resulting in equal numbers of differentially expressed genes (Figure 5c).

To assess which TTP approach is in best agreement with the changes in gene expression, as measured by RT-PCR, expression levels of 10 differentially expressed genes were measured in triplicate for each of the six samples using RT-PCR, and were compared with those obtained by the three TTP approaches using the mixed model (Figure 6). With the exception of gene AT3G27630, the fold changes measured by the TTP approaches are systematically higher than those measured by RT-PCR. Figure 6 also clearly shows that TTP2 and TTP3 are often in high agreement with the changes in gene expression measured by RT-PCR, confirming the superiority of TTP2 and TTP3 to TTP1.

Identification of E2F target genes

Considering the results obtained with TTP3 using the mixed model procedure, a total of 946 genes were found to have a significant, at least twofold induction in *E2Fa-DPa*^{OE} compared with WT plants, versus the 773 identified with the ATH1 array (Tables 2, S1 and S2). Among the 773 genes identified by ATH1, 195 genes were missing in the Tiling 1.0R data set, in many cases because they just failed to reach the significance threshold or fold-change criterion on the tiling platform. Likewise, out of the 946 genes found to be induced in *E2Fa-DPa*^{OE} plants with the Tiling 1.0R array, 368 were missing in the ATH1 data set (Table S3). Among these 368 genes, 203 genes are among the 9113 genes present only on the Tiling 1.0R array, and thus represent new potential E2Fa-DPa targets. This set of 368 genes showed a gene ontology (GO) enrichment for DNA replication (Table 2), and includes genes involved in cell division (*DEL1*, *CDKB1;1*, *KRP6* and *TONSOKU*), DNA replication (*MCM6*, *ORC1a* and *OCR1b*) and DNA damage response (*MSH3* and *DDB2*). Among the upregulated genes, 153 have a motif in their 1-kb promoter that conforms to known E2F

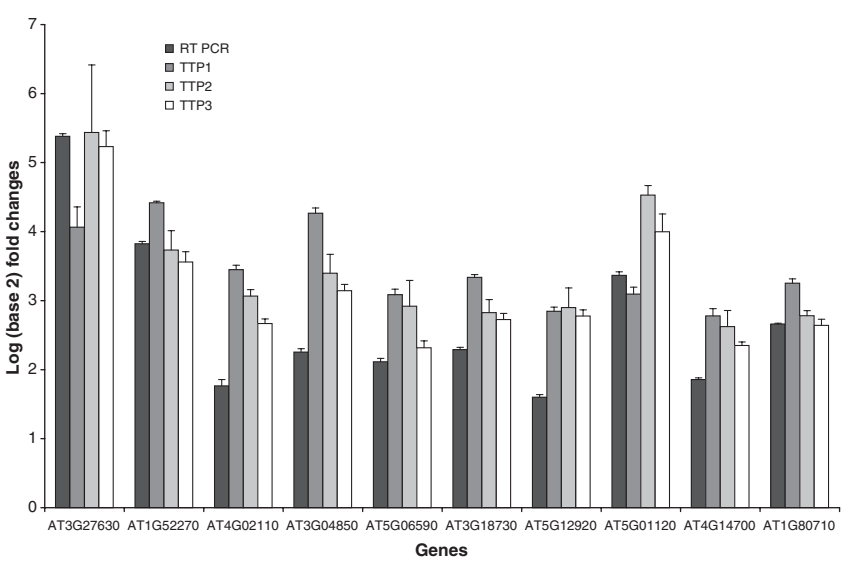

Figure 6. Comparison of log (base 2) expression fold changes as measured by RT-PCR (black) and the three Tiling 1.0R approaches, TTP1 (dark gray), TTP2 (light gray) and TTP3 (white) for 10 genes that were upregulated in *E2Fa-DPa*^{OE} compared with the wild type.
Error bars represent ±SED .

Table 2 Enrichment analysis of genes upregulated in *E2Fa-DPa*^{OE} compared with the wild type, identified using ATH1 and Tiling 1.0R microarrays

Gene set	No. genes	*P*-value 'DNA replication' GO enrichment	Enrichment fold	Number (percentage) of genes in GO category[a]	Number (percentage) of genes with E2F motif[b]	Number (percentage) of genes with conserved E2F[c]
ATH1	773	3.09e–32	13.76	45 (5.8)	389 (50.3)	196 (25.4)
TTP3	946	6.87e–37	13.63	53 (5.6)	468 (49.5)	219 (23.2)
Tiling-specific genes (TTP3–ATH1)	368	2.53e–05	6.24	9 (2.4)	153 (41.6)	50 (13.6)
ATH1-specific genes (ATH1–TTP3)	195	not significant	1.25	1 (0.5)	74 (37.9)	27 (13.8)

For Tiling arrays, data obtained with the target preparation protocol TTP3 with mixed models analysis were considered.
[a]Frequency GO label 'DNA replication' in Arabidopsis genome = 0.4%.
[b]Frequency E2F motif in all Arabidopsis 1-kb promoters = 29.0% (9287/32042).
[c]The presence of an E2F instance in Arabidopsis was scored as conserved if at least one orthologous poplar gene had an E2F motif in its promoter (irrespective of strand or position in the promoter).

binding sites. Of these, the E2F motif is also found in one or more orthologous Populus genes, suggesting that they represent evolutionarily conserved E2F targets (Table 2). To validate some of the new potential E2F targets, we randomly selected 10 genes with an evolutionarily conserved E2F motif. We could confirm their transcriptional upregulation in *E2Fa-DPa*^{OE} plants by RT-PCR analysis (Figure 6).

Moreover, a direct binding of the E2Fa transcription factor to the promoter was confirmed for seven genes by ChIP analysis, supporting the assertion that they are direct E2Fa target genes (Figure 7). The promoters of the three remaining genes were found to bind the related E2Fb transcription factor, indicating that E2Fa indirectly activates these genes through the transcriptional induction of E2Fb.

To assess the biological meaning of the results obtained for the Tiling 1.0R and ATH1 arrays, we first compared the enrichment for the GO class 'DNA replication', as we have demonstrated previously that most genes induced by E2Fa-DPa belong to this GO class (Vandepoele *et al.*, 2005). Using TTP3 in combination with the mixed model procedure, the Tiling 1.0R platform slightly outperforms the ATH1 array in this exercise. Out of 101 annotated DNA replication genes, 53 were found to be induced by more than twofold ($Q < 0.05$) in *E2Fa-DPa*^{OE} transgenic plants, of which eight were detected only on the Tiling array.

In a second step, we compared the number of putative direct E2F target genes identified on both platforms. In *A. thaliana*, there are 9287 genes with an E2F motif in their 1-kb promoter (see Table S4). For 23% (2160) of these, we find that an E2F motif is also present in one or more Populus orthologs. When comparing the presence of E2F binding sites in putative genes identified by both array platforms, again a similar enrichment of genes with a (conserved) E2F element was observed: 49.5% (23.2%) using the Tiling 1.0R array versus 50.3% (25.4%) for the ATH1 array (Table 2), illustrating that the Tiling 1.0R arrays perform well when compared with the ATH1 arrays.

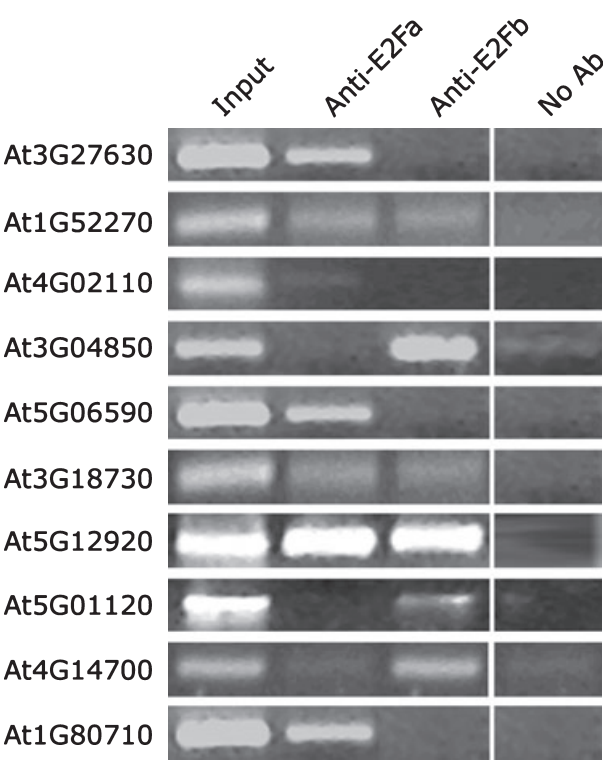

Figure 7. Chromatin immunoprecipitation analysis (ChIP) analysis confirming direct E2Fa/E2Fb binding to the promoter of 10 randomly selected potential direct targets (gene identifiers on the left) of E2Fa-DPa, identified uniquely on the Tiling 1.0R platform.
ChIP was performed on 8-day-old *E2Fa-DPa*^{OE} seedlings using E2Fa- or E2Fb-specific antibodies. An immunoprecipitation without any added antibody served as a control.

Identification of unannotated E2F targets

The identification of genomic regions that are differentially transcriptionally active in WT and *E2Fa-DPa*^{OE} plants that are not part of the overlap between ATH1 and Tiling 1.0R was performed, as explained in detail in the Experimental

procedures. Using this simple algorithm, we found 6974 differentially expressed intervals, of which 6646 matched to annotated genes/transcripts on the TAIR website. For each of the 328 intervals remaining, the corresponding sequence was extracted and blasted against the Arabidopsis expressed sequence tag (EST) bank available on TAIR website. In this way, sequences were checked to see if they were already known as unannotated RNA sequences. Of the 328 intervals, 54 matched to a known EST with identity ≥99%. The remaining 274 differentially expressed regions might represent potential targets for E2F.

Discussion

Although Affymetrix does not explicitly recommend its Arabidopsis Tiling 1.0R array for RNA analyses, this study provides evidence that this platform can be used for quantitative RNA expression analysis. Others have recently described the use of the Tiling 1.0F array, which contains probes that are complementary to those on the Tiling 1.0R array, for expression analysis (Jones-Rhoades Matthew *et al.*, 2007). The increased coverage of tiling arrays compared with the standard ATH1 array, which was developed four years ago (Redman *et al.*, 2004), may significantly aid in the discovery of new transcriptional targets of various processes or transcription factors, including those not represented in the current TAIR7 annotation. Based on this annotation, the Tiling 1.0R array includes over 9000 additional genes, or about 40% more genes than are present on the ATH1 array. In our specific case, this led to the identification of 153 novel putative E2F target genes, where only 184 were known before. This striking increase is proportionally greater than what one might have expected based on the relative numbers of genes on both arrays. It might reflect that many E2F targets are expressed at relatively low levels, and that this group of genes is underrepresented on the ATH1 array. We also analyzed unannotated genomic regions following a published procedure (Jones-Rhoades Matthew *et al.*, 2007), and found 204 differentially active genomic regions, some of which might represent additional potential E2F targets.

With this paper we present two approaches for the analysis of expression data generated with Affymetrix Tiling 1.0R arrays. In addition, we compared a number of different target preparation protocols, in particular with respect to the extent in which they influence the potential to obtain quantitative results, and presented a Tiling 1.0R array CDF containing gene-specific single-copy exonic probe sets for each of the 29 767 genes. Overall, the mixed model approach performed better in terms of numbers of differentially expressed genes identified, as well as differentially expressed genes found in common between the two Affymetrix microarray platforms, than the commonly used RMA approach. A possible explanation is that taking the mean across all of the probes, as implemented in the mixed model approach, is more appropriate than taking the median, especially when the overall level of signal is low and more variable.

Exclusion of UTR probes from the analysis increased the number of differentially expressed genes identified, as well as the proportion of differentially expressed genes in common between ATH1 and Tiling 1.0R platforms, irrespective of the target preparation protocol used. Part of the reason for this may be that, because of the multiple polyadenylation sites (Xiao *et al.*, 2002), mRNA does not hybridize to every single 3′ UTR probe, resulting in heterogeneity of the obtained signal, which in turn lowers the confidence of expression estimates. Our results also provide a useful assessment of three target preparation protocols, and demonstrate the need for capturing ribosomal RNA when random priming of total RNA is performed. We recommend the TPP3 strategy, amplifying all transcripts, irrespective of their polyadenylation status, combined with the Invitrogen Arabidopsis Ribominus kit.

In summary, we hope that this cross-platform study may further assist the Arabidopsis community when choosing a strategy for genome-wide gene expression studies.

Experimental procedures

Experimental design and RNA material

Arabidopsis thaliana WT (Columbia 0) and *E2Fa-DPa*^{OE} plants were grown side-by-side under long-day conditions (with a 16-h light/8-h dark photoperiod) at 22°C on germination medium, as described previously (Vandepoele *et al.*, 2005). Pools of seedlings were harvested in triplicate for each genotype 8 days after sowing, and were quickly frozen in liquid nitrogen and stored at −70°C. Total RNA was extracted using Trizol reagent (Invitrogen; http://www.invitrogen.com). No DNAse treatment was performed. The integrity and purity (A_{260}/A_{280} and A_{260}/A_{230} ratios > 1.8) were determined on an Agilent 2100 Bioanalyzer (http://www.agilent.com) and a Nanodrop ND-1000 UV-Vis spectrophotometer (http://www.nanodrop.com), respectively. Each RNA sample was split into four aliquots for different target preparation and hybridization protocols. In total, six ATH1 arrays and 18 Tiling 1.0R arrays were processed.

Annotation

BIOCONDUCTOR v2.1 contains an updated ATH1 CDF according to the TAIR7 annotation, with 21 021 probe sets uniquely mapping to one gene. We created a Tiling 1.0R array CDF containing gene-specific single-copy exonic probe sets, with the following steps (Figure 1): (i) probes derived from non-unique sequences were masked; (ii) when several transcript variants were annotated, only the probes common to all variants were considered; (iii) probes corresponding to intronic regions, including ones spanning intron/exon junctions, were removed. One CDF variant contained 5′ and 3′ UTR probes; the other variant did not. Mitochondrial and chloroplast genes were not included in the Tiling 1.0R CDFs. The CDF variant that does not contain the probes from UTR regions is available in BIOCONDUCTOR as *athtiling1.0rcdf*; the other variant is available upon request.

Affymetrix ATH1 hybridization

Labeled targets for the Affymetrix ATH1 arrays (catalog no. ATH1121501; Affymetrix, http://www.affymetrix.com) were prepared as recommended by the manufacturer. Briefly, 5 μg of total RNA was reverse transcribed into cDNA using an oligo(dT)-T7 primer, and was then converted into cRNA and linearly amplified by T7 *in vitro* transcription reaction using the Affymetrix IVT Labeling Kit (catalog no. 900449; Affymetrix). The probes were purified with the GeneChip® Sample Cleanup Module (catalog no. P/N 900371; Affymetrix), and were analyzed again for yield, which was 30–120 μg, and purity (A_{260}/A_{280} and A_{260}/A_{230} ratios > 1.8). The resulting cRNA (20 μg) was fragmented by alkaline hydrolysis and resuspended with control spike-in probes in 300 μl of hybridization buffer (Hybridization Control Kit, catalog no. 900457: Affymetrix). The GeneChip® arrays were hybridized in a rotating oven at 45°C for 17 h, and were then washed and stained in the GeneChip® Fluidics Station 400 (Wash and Stain kit, catalog no. 900720; Affymetrix), using the EukGE-WS2v4 washing protocol. Scanning was performed with the GeneChip® Scanner 3000, and image analysis was performed in GCOS (catalog no. 690036; Affymetrix). We refer to this ATH1 target preparation as 'ATP'.

Affymetrix Tiling 1.0R array hybridization

GeneChip® Arabidopsis Tiling 1.0R arrays (catalog no. 900594; Affymetrix) were hybridized with 7.5 μg of double-stranded (ds) cDNA that had been fragmented with uracil DNA glycosylase, and had been terminally labeled with biotin using a terminal deoxynucleotidyl transferase (DLR; Affymetrix WT Double-stranded DNA Terminal Labeling Kits, catalog no. 900812). The cDNA material was prepared following three different target preparation protocols (Figure 2). The first protocol, referred to as TTP1, started with 7 μg of amplified cRNA, prepared as described above for ATP. This was then reversed transcribed to dUTP-containing ds cDNA (WT ds cDNA Synthesis Kit, catalog no. 900813; Affymetrix) using random primers. The second target preparation protocol, referred to as TTP2, started with 7 μg of total RNA, directly reversed transcribed to ds cDNA with T7-(N)6 primers and with the incorporation of dUTP for subsequent fragmentation, also using components of the WT ds cDNA Synthesis Kit. The third target preparation protocol, referred to as 'TTP3', started with 5 μg of total RNA, from which the ribosomal RNA was removed with the Invitrogen Ribominus Kit (catalog no. 45-7013), in which the LNA oligonucleotide for yeast rRNA had been replaced by an LNA oligonucleotide for *A. thaliana* rRNA. This material was reverse transcribed to ds cDNA using T7-(N)6 primers, converted into cRNA and linearly amplified by *in vitro* transcription, and then again reverse transcribed to ds cDNA with random primers, and with the incorporation of dUTP, all using the Affymetrix WT Amplified ds cDNA Synthesis Kit (catalog no. 900811).

The labeled and fragmented cDNA was resuspended with control Oligonucleotide B2 (catalog no. 900301; Affymetrix) in 300 μl of hybridization buffer, hybridized to the arrays in a rotating oven at 45°C for 16 h, and was washed, stained, scanned and analyzed as described above for ATH1 arrays.

Expression analysis

Probe-level data from the ATH1 and Tiling 1.0R arrays were preprocessed using the RMA algorithm (Irizarry *et al.*, 2003), which involves three steps: (i) background correction – where an error component of the intensities is estimated and eliminated; (ii) quantile normalization – where every slide is normalized to have the same cumulative frequency distribution; and (iii) summarization, using the median polish algorithm – where the median values per probe set, adjusted for slide differences, are calculated. RMA ignores the MM probes. On the basis of an empirical Bayes moderated *t*-statistic for the contrasts (Smyth, 2004), as implemented in the BIOCONDUCTOR package LIMMA, *P* values were calculated and then transformed into false-discovery rates, or *Q* values according to the method described by Storey and Tibshirani (2003), as implemented in the R package QVALUE. We compared the results obtained with RMA and those obtained with a mixed-effect ANOVA model (Wolfinger *et al.*, 2001) as a summarization step, after background adjustment (convolution) and quantile normalization had been performed as described for RMA. The gene-specific linear mixed model used here was:

$$Y_{ijk} = \mu + P_i + G_k + A_{j(k)} + \varepsilon_{ijk},$$

in which μ represents the overall mean, P_i is the random probe effect, where *i* ranges from 11 up to 20 for ATH1 (most of the genes present 11 probes, and a few of them present up to 20 probes), and from 3 up to 56 for the Tiling 1.0R array data, respectively, G_k is the fixed genotype effect and $A_{j(k)}$ is the random array (replicate) effect ($j = 1\ldots3$) in genotype *k*. The measurement error is represented by ε_{ijk}. Probe and array effects are considered to be random and statistically independent from each other. This assumption of randomness for the probe effect is reasonable, considering that because of variations in probe hybridization characteristics, different probes interrogating the same transcript can show different signal intensities (Li and Wong, 2001). The mixed model was fitted by restricted maximum likelihood (REML), and differential expression was assessed by means of a Wald test. Significance was assessed either on the basis of the *F* approximation to the Wald statistic for the contrast, or by permutation. The calculated *P* values were subsequently transformed into *Q* values. For each permutation, expression values of each gene were randomly assigned to genotypes and probes, and the gene-specific linear mixed model, as described above, was fitted to the data. The permutation was performed 1000 times, and the sampling distribution of the Wald statistic under the null hypothesis, i.e. no differential expression between WT and *E2Fa-DPa*^{OE} plants, was computed. Finally, two-sided *P* values of the test were calculated as the proportion of sampled permutations where the absolute expression difference was greater than or equal to the observed absolute expression difference.

CAT plots

Correspondence at the top (CAT) plots (Irizarry *et al.*, 2005) were created to compare the Tiling 1.0R procedures with the ATH1 procedure for detecting differentially expressed genes. We made a list of *n* candidate genes found by each procedure, ranked genes according to their measured -fold change and plotted the proportion of genes in common against list sizes with $n = 200–1200$.

Identification of new, unannotated E2F target genes

Identifying genomic regions differentially expressed between WT and *E2Fa-DPa*^{OE} plants not covered by gene models in the TAIR7 annotation was performed according to the approach proposed by Jones-Rhoades Matthew *et al.* (2007). For each probe, a Welch's *t*-test was performed, comparing the $\log_2$ expression values of WT with *E2Fa-DPa*^{OE}. Neighboring probes showing a significant

($P < 0.05$) common change of at least twofold between WT and *E2Fa-DPa*^{OE} plants were taken to indicate a region of the genome giving rise to a differentially expressed transcript. Only intervals with at least three probes were considered further.

E2F motif detection

Detection of E2F binding sites was performed by scanning promoter regions for the presence of an E2F motif [position weight matrix (PWM) representation with consensus motif TTTssCGC derived from 412 upregulated genes; Vandepoele *et al.*, 2005] using MotifLocater (arguments: third-order background model derived from all promoters, –*t* threshold set to 0.95; Thijs *et al.*, 2002). Promoter sequences, extracted from 1 kb upstream of the start codon, were based on the TAIR7 annotation. The upstream sequence was truncated if the intergenic sequence separating a gene from its upstream neighbor was <1 kb.

The assignment of genes to the original Gene Ontology (GO; http://www.geneontology.org) categories was extended to include parental terms (i.e. a gene assigned to a given category was automatically assigned to all the parent categories as well). Enrichment values were calculated as the ratio of the frequency in the selected set relative to the genome-wide frequency. The statistical significance of enrichment was evaluated using the hypergeometric distribution.

Orthologous groups were identified through protein clustering using OrthoMCL (Li *et al.*, 2003). Starting from an all-against-all BLASTP sequence similarity search using the full proteomes of *A. thaliana* (26 541 proteins), and that of the closest sequenced relative, *Populus trichocarpa* (45 554 proteins), 11 707 orthologous clusters were defined, covering 18 088 Arabidopsis and 22 760 Populus genes. These orthologous groups contain inparalogous genes (i.e. genes duplicated after the divergence between Arabidopsis and Populus), and thus offer a more realistic representation of orthology compared with, for example, reciprocal best-hit approaches.

After running the MotifLocater on all Populus promoter sequences, as described above, the presence of an E2F instance in Arabidopsis was scored as conserved if at least one orthologous Populus gene had an E2F motif in its promoter (irrespective of orientation or position in the promoter).

Quantitative real-time PCR analysis

RNA was extracted with the RNeasy Plant Mini Kit (Qiagen, http://www.qiagen.com). First-strand cDNA was prepared from total RNA with the Superscript™ II First-Strand Synthesis System (Invitrogen). Quantitative PCR reactions were performed with the LightCycler® 480 SYBR Green I Master mix (Roche; http://www.roche.com), and were analyzed on the LightCycler® 480 Real-Time PCR System (Roche). All quantifications were normalized to *ACTIN2* cDNA fragments amplified under the same conditions. Quantitative PCR reactions were performed in triplicate and were then averaged. Primer sequences are available upon request.

Chromatin immunoprecipitation analysis

Chromatin immunoprecipitation analysis (ChIP) was performed according to Bowler *et al.* (2004), with some modifications. Briefly, 1 g of 8-day-old seedlings was harvested, rinsed in ddH_2O and cross-linked in 1% formaldehyde for 10 min. Cross-linking was stopped by the addition of glycine to a final concentration of 125 mM. Chromatin was extracted from homogenized tissue obtained by grinding. The chromatin was sheared using a Branson 1200 sonifier (http://www.sonifier.com). After pre-clearing, a 10 μl volume of the appropriate antibodies was added to the chromatin solution and incubated overnight at 4°C. After collection of the immunoprecipitate with protein A agarose beads, beads were washed and antibody–chromatin complexes were eluted. Cross-linking was reversed by overnight incubation at 65°C. Proteinase K digestion was followed by phenol/chloroform extraction and ethanol precipitation. Recovered DNA was subjected to 25 cycles of PCR amplification.

Acknowledgements

We acknowledge Heidi Wouters for her assistance with permutation and her valuable advice, and Sara Maes for her technical assistance. The LNA oligonucleotide probe for *A. thaliana* was kindly provided by Dr. Byung-in Lee (Invitrogen, Carlsbad, CA. This work was supported by FP6 MARIE CURIE ACTIONS in the framework of FP6 IP TRANSISTOR (http://www.transistor-arabidopsis.org, contract MRTN-CT-2004-512285), and by FP6 IP AGRON-OMICS (contract LSHG-CT-2006-037704). LDV and KV are postdoctoral fellows of the Research Foundation – Flanders.

Supporting Information

Additional supporting information may be found in the online version of this article:

Table S1. Upregulated genes in the three tiling array experiments.

Table S2. Upregulated genes in the ATH1 experiment.

Table S3. Tiling array identified putative E2F regulated genes.

Table S4. Arabidopsis genes harbouring and E2F motif in 1-kb promoter region.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Bertone, P., Stolc, V., Royce, T.E. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Blais, A. and Dynlacht, B.D. (2007) E2F-associated chromatin modifiers and cell cycle control. *Curr. Opin. Cell Biol.* **19**, 658–662.

Bowler, C., Benvenuto, G., Laflamme, P., Molino, D., Probst, A.V., Tariq, M. and Paszkowski, J. (2004) Chromatin techniques for plant cells. *Plant J.* **39**, 776–789.

De Veylder, L., Beeckman, T., Beemster, G.T.S. *et al.* (2002) Control of proliferation, endoreduplication and differentiation by the Arabidopsis E2Fa-DPa transcription factor. *EMBO J.* **21**, 1360–1368.

De Veylder, L., Beeckman, T. and Inze, D. (2007) The ins and outs of the plant cell cycle. *Nat. Rev. Mol. Cell Biol.* **8**, 655–665.

Iaquinta, P.J. and Lees, J.A. (2007) Life and death decisions by the E2F transcription factors. *Curr. Opin. Cell Biol.* **19**, 649–657.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Irizarry, R.A., Warren, D., Spencer, F. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–349.

Jones-Rhoades Matthew, W., Borevitz Justin, O. and Preuss, D. (2007) Genome-wide expression profiling of the *Arabidopsis* female gametophyte identifies families of small, secreted proteins. *PLoS Genet.* **3**, 14.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P.A. and Gingeras, T.R. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, **296**, 916–919.

Kosugi, S. and Ohashi, Y. (2003) Constitutive E2F expression in tobacco plants exhibits altered cell cycle control and morphological change in a cell type-specific manner. *Plant Physiol.* **132**, 2012–2022.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Li, L., Stoeckert, C.J. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.

Redman, J.C., Hass, B.J., Tanimoto, G. and Town, C.D. (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array (Vol. 38, p. 545, 2004). *Plant J.* **38**, 1023–1023.

Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. App. Genet. Mol. Biol.* **3**.

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

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. and Moreau, Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.* **9**, 447–464.

Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G.T.S., Gruissem, W., Van De Peer, Y., Inze, D. and De Veylder, L. (2005) Genome-wide identification of potential plant E2F target genes. *Plant Physiol.* **139**, 316–328.

Vlieghe, K., Vuylsteke, M., Florquin, K., Rombauts, S., Maes, S., Ormenese, S., Van Hummelen, P., Van de Peer, Y., Inze, D. and De Veylder, L. (2003) Microarray analysis of E2Fa-DPa-overexpressing plants uncovers a cross-talking genetic network between DNA replication and nitrogen assimilation. *J. Cell Sci.* **116**, 4249–4259.

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.

Xiao, Y.L., Malik, M., Whitelaw, C.A. and Town, C.D. (2002) Cloning and sequencing of cDNAs for hypothetical genes from chromosome 2 of Arabidopsis. *Plant Physiol.* **130**, 2118–2128.

