Pure Appl. Chem., Vol. 79, No. 4, pp. 811–823, 2007.
doi:10.1351/pac200779040811

Phytochemical genomics in *Arabidopsis thaliana*: A case study for functional identification of flavonoid biosynthesis genes*

Takayuki Tohge[1], Keiko Yonekura-Sakakibara[1], Rie Niida[1], Akiko Watanabe-Takahashi[1], and Kazuki Saito[1,2,‡]

[1]RIKEN Plant Science Center, 1-7-22, Suehiro-cho, Tsurumi-ku, Yokohama-shi, Kanagawa 230-0045, Japan; [2]Graduate School of Pharmaceutical Sciences, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba-shi, Chiba 263-8522, Japan

Abstract: The completion of the whole genome sequence of *Arabidopsis thaliana* has made it possible to explore the phytochemical genomics in this species by determining gene-to-metabolite correlation through the comprehensive analysis of metabolite accumulation and gene expression. In this study, flavonoid profiling of wild-type plants and T-DNA insertion mutants was analyzed using ultra-performance liquid chromatography (UPLC)/photodiode array detection (PDA)/electrospray ionization (ESI)/multiple-stage mass spectrometry (MS^n). Detailed analysis of the metabolite changes in the mutants suggested the functions of genes that have been mutated. In silico coexpression analysis of genes involved in flavonoid metabolism in Arabidopsis was performed using a publicly available transcriptome database of DNA microarrays. We inferred a coexpression framework model of the genes involved in the pathways of flavonol, anthocyanin, and proanthocyanidin synthesis, suggesting specific functions and coregulation of the genes of pathway enzymes and transcription factors. The metabolic profiling of the *omt1* mutant lacking a methyltransferase gene narrowed down by the coexpression analysis showed that *AtOMT1* (At5g54160) is involved not only in the production of lignins and sinapoyl esters but also in the methylation of flavonols forming isorhamnetin. These results suggest that the functional genomics approach by detailed target-metabolite profiling with transcriptome coexpression analysis provides an efficient way of identifying novel gene functions involved in plant metabolism.

Keywords: biosynthesis; flavonoids; glycosylation; enzymes; Arabidopsis.

INTRODUCTION

Plants produce a wide array of compounds that are potentially useful in developing novel medicines, flavors, industrial materials as alternatives for fossil fuel resources, and other specialty chemicals. Cumulatively, plants are thought to produce ~200 000 natural products [1].

Flavonoids, derived from phenylpropanoids and malonyl-CoA, are widely distributed in the plant kingdom, and over 4000 individual compounds have been identified [2]. The widespread distribution of flavonoids and their relatively low toxicity compared to other active plant compounds allows the in-

Pure Appl. Chem.* **79, 467–823 (2007). An issue of reviews and research papers based on lectures presented at the 25th International Symposium on Chemistry of Natural Products (ISCNP-25) and 5th International Conference on Biodiversity (ICOB-5), held jointly in Kyoto, Japan, 23–28 July 2006, on the theme of natural products.
‡Corresponding author: Fax +81-45-503-9489; E-mail: ksaito@faculty.chiba-u.jp

gestion of significant quantities of flavonoids in the diets of many animals, including humans. Many flavonoids share a common phenyl benzopyrane skeleton and differ from one another only in the numbers and positions of the hydroxyl or methylated functional groups. Although the flavonoids share a basic skeleton, a variety of modifications such as glycosylation, acylation, and methylation are responsible for the highly diverse structures of these compounds. Recently, the structures of several anthocyanins and flavonol glycosides [3–5] have been reported in *Arabidopsis thaliana*. To improve the productivity of plants by modifying the genes involved in the synthesis of useful compounds, it is essential to understand the metabolic process in plant and its regulatory mechanisms as a whole. Unfortunately, only a limited number of genes involved in the production of these plant metabolites have been identified by the classical genetic screening of mutants and enzyme purification. Several genes encoding enzymes and regulatory proteins involved in the production of anthocyanins and proanthocyanidins have been isolated mainly by *tt* (transparent testa) or *ttg* (transparent testa glabrous) mutants of seed color of *A. thaliana* [6]. The genes encoding flavonoid 3-*O*-glucosyltransferase (*UGT78D2/F3GT*, At5g17050), flavonol 3-*O*-rhamnosyltransferase (*UGT78D1/F3RT*, At1g30530), flavonol 7-*O*-glucosyltransferase (*UGT73C6/F7RT*, At2g36790), and anthocyanin 5-*O*-glucosyltransferase (*UGT75C1/A5GT*, At4g14090) have been identified by the integrated analysis of the metabolome and transcriptome [7,8]; however, there must be other unidentified genes for flavonoid glycosyltransferases in the *A. thaliana* genome.

The completion of the whole genome sequence of *A. thaliana* has enabled the determination of gene-to-gene and gene-to-metabolite correlation through the comprehensive analysis of metabolite accumulation (metabolomics) and gene expression (transcriptomics) [9–12]. Gene-expression-array technologies designed by commercial enterprises for transcriptome analyses are rapidly emerging for use in various genome-wide studies in biology. To execute metabolome analyses, gas chromatography (GC)/mass spectrometry (MS), high-performance liquid chromatography (HPLC)/MS, capillary electrophoresis (CE)/MS, and Fourier-transform (FT)/MS offer the possibility of the identification of target compounds in a complex mixture by providing a unique combination of resolving power, sensitivity, and selectivity. Multiple-stage mass spectrometry (MS^n) analysis is a tool used more frequently for the identification of metabolites in plant crude plant extracts to provide not only molecular-mass information, but also structural information that is deduced from ion-peak patterns obtained by collision-induced fragmentations. The advantage of HPLC/MS is the ability to determine selected compounds from only a small amount of crude materials with a simple sample preparation. Recent successful studies have indicated the power of the integrated analysis of metabolomics and transcriptomics to identify gene functions involved in secondary metabolism [7,8,13–15]. In order to decipher the Arabidopsis genome, the improvement of high-throughput system for the identification of gene function is required.

Several public Web-based databases for the integrated analysis of the metabolome and transcriptome of Arabidopsis are available, such as KaPPa-View [16], MetaCyc [17], and MapMan [18]. AtGenExpress (RIKEN, <http://pfg.psc.riken.jp/AtGenExpress/>; MaxPlanck, <http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm>)—a database of a multinational consortium for the Affymetrix Arabidopsis Genome ATH1 microarray—has been released to the public. By using these public data, databases of coexpressed Arabidopsis genes are available, for example, on the *A. thaliana trans*-factor and *cis*-element prediction database (ATTED-II) [19], the Platform for RIKEN Metabolomics (PRIMe, <http://prime.psc.riken.jp/>), the Comprehensive Systems-Biology Database (CSB.DB, <http://csbdb.mpimp-golm.mpg.de/index.html>) [20], and the Arabidopsis Coexpression Tool (ACT, <http://www.arabidopsis.leeds.ac.uk/ACT/>) [21]. The coexpressed gene sets can provide clues to gene functions, based upon the assumption that the expression of genes in the same pathway is likely to be coregulated. Very recently, some reports have been published on the successful elucidation of novel gene functions by using these coexpression databases [22–26].

Several Arabidopsis bioresources, including the T-DNA knockout mutant lines (Salk Institute, <http://signal.salk.edu/>) [27] and the full-length cDNA library (RIKEN BioResource Center, <http://www.brc.riken.jp/lab/epd/Eng/>) [28], are available. The practical use of these bioresources and

public Web-based databases allows us to accelerate the identification of genes by the reverse genetics approach and by the biochemical characterization of recombinant proteins. Despite the rich resources of genetic information, chemical information regarding *A. thaliana* metabolites is relatively insufficient. In the present study, we have investigated the flavonoid profiles of different tissues of wild-type plants and T-DNA insertion mutants to obtain the clues to the function of particular genes. Transcriptome co-expression analysis performed using public databases efficiently supports the prediction of gene function. These candidate genes coding for flavonoid biosynthetic enzymes have been confirmed by the metabolite analysis of knockout mutants.

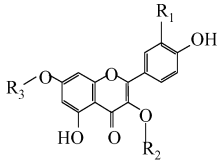
RESULTS AND DISCUSSION

Flavonoid profiles of Arabidopsis leaves

The structures of 8 flavonol glycosides (f1–f8) [3–5,7,8] and 11 anthocyanins (a1–a11) [5,8] in Arabidopsis have been characterized using data from UV spectroscopy, MS, ^{1}H NMR, and ^{13}C NMR spectroscopy, or annotated with HPLC/photodiode array detection (PDA)/electrospray ionization (ESI)/MS analysis (Fig. 1). In our study, the flavonoid accumulation profiles of wild-type Arabidopsis (ecotype Columbia 0, Col-0, used for the Arabidopsis genome sequencing project) leaves were analyzed using ultra-performance liquid chromatography (UPLC)/PDA/ESI/MS. In Col-0 leaves grown under the unstressed condition, three kaempferol glycosides were detected by UPLC/PDA/MS analysis. Figure 2A shows the chromatogram at 280 nm UV absorption and the extracted mass ion chromatograms. In these chromatograms, the peaks of three kaempferol glycosides (f1, kaempferol 3-*O*-rhamnoside 7-*O*-rhamnoside; f2, kaempferol 3-*O*-glucoside 7-*O*-rhamnoside; f3, kaempferol 3-*O*-[rhamnosyl(1 → 2)-glucoside] 7-*O*-rhamnoside) were identified by comparison with the standard compounds. The quasi-molecular ions of *m/z* values of 741 ([kaempferol + rhamnose + rhamnose + glucose – $3H_2O$ + H]$^+$), 595 ([kaempferol + rhamnose + glucose – $2H_2O$ + H]$^+$), and 579 ([kaempferol + rhamnose + rhamnose – $2H_2O$ + H]$^+$) were characteristic and thus valuable for the assignment of the chemical structures.

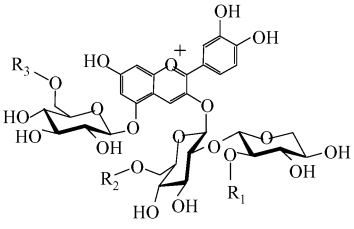
The genes encoding flavonoid 3-*O*-glucosyltransferase (*UGT78D2*/*F3GT*, At5g17050) and flavonol 3-*O*-rhamnosyltransferase (*UGT78D1*/*F3RT*, At1g30530) have previously been identified (Table 1) [7,8]. The *ugt78d2* mutant leaves lacked 2 kaempferol 3-*O*-glycosides (f2 and f3), as indicated by the extracted mass chromatograms of *m/z* 741 and 595 (Fig. 2A-b). On the other hand, kaempferol 3-*O*-rhamnoside (f1) was not detected in the *ugt78d1* mutant leaves (Fig. 2A-c). These results suggested that the changes in the metabolite profiles of specific gene-knockout mutants were well correlated with the loss of gene functions.

Flavonols



f1: R^1= H, R^2= rhamnoside, R^3= rhamnoside
f2: R^1= H, R^2= glucoside, R^3= rhamnoside
f3: R^1= H, R^2= rhamnosyl(1→2)-glucoside, R^3= rhamnoside
f4: R^1= H, R^2= glucosyl(1→6)-glucoside, R^3= rhamnoside
f5: R^1= OH, R^2= rhamnoside, R^3= rhamnoside
f6: R^1= OH, R^2= glucoside, R^3= rhamnoside
f7: R^1= OH, R^2= rhamnoside, R^3= glucoside
f8: R^1= OH, R^2= rhamnosyl(1→2)-glucoside, R^3= rhamnoside
f9: R^1= H, conjugate
f10: R^1= H, -hexose -deoxyhexose (*m/z* 595)
f11: R^1= H, -hexose -deoxyhexose -deoxyhexose (*m/z* 741)
f12: R^1= H, 3-*O*-rhamnoside conjugate (*m/z* 773)
f13: R^1= OH, -hexose -deoxyhexose (*m/z* 611)
f14: R^1= OMe, -hexose -deoxyhexose (*m/z* 625)
f15: R^1= OMe, -hexose -deoxyhexose (*m/z* 625)

Anthocyanins



a1: R^1= H, R^2= H, R^3= H
a2: R^1= H, R^2= H, R^3= malonyl
a3: R^1= H, R^2= *p*-coumaroyl, R^3= H
a4: R^1= sinapoyl, R^2= H, R^3= H
a5: R^1= H, R^2= *p*-coumaroyl, R^3= malonyl
a6: R^1= H, R^2= glucosyl-*p*-coumaroyl, R^3= H
a7: R^1= sinapoyl, R^2= *p*-coumaroyl, R^3= H
a8: R^1= H, R^2= glucosyl-*p*-coumaroyl, R^3= malonyl
a9: R^1= sinapoyl, R^2= *p*-coumaroyl, R^3= malonyl
a10: R^1= sinapoyl, R^2= glucosyl-*p*-coumaroyl, R^3= H
a11: R^1= sinapoyl, R^2= glucosyl-*p*-coumaroyl, R^3= malonyl

Fig. 1 Chemical structures of flavonol glycosides and anthocyanins in Arabidopsis. f1, kaempferol 3-*O*-rhamnoside 7-*O*-rhamnoside; f2, kaempferol 3-*O*-glucoside 7-*O*-rhamnoside; f3, kaempferol 3-*O*-[rhamnosyl(1 → 2)-glucoside] 7-*O*-rhamnoside; f4, kaempferol 3-*O*-[glucosyl(1 → 6)glucoside] 7-*O*-rhamnoside; f5, quercetin 3-*O*-rhamnoside 7-*O*-rhamnoside; f6, quercetin 3-*O*-glucoside 7-*O*-rhamnoside; f7, quercetin 3-*O*-rhamnoside 7-*O*-glucoside; f8, quercetin 3-*O*-[rhamnosyl(1 → 2)-glucoside] 7-*O*-rhamnoside; f9, kaempferol conjugate; f10, kaempferol conjugate (*m/z* 595); f11, kaempferol conjugate (*m/z* 741); f12, kaempferol 3-*O*-rhamnoside conjugate (*m/z* 773); f13, quercetin conjugate (*m/z* 611); f14, isorhamnetin conjugate (*m/z* 625); f15, isorhamnetin conjugate (*m/z* 625); a1, cyanidin 3-*O*-[2-*O*-(xylosyl)-glucoside] 5-*O*-glucoside; a2, cyanidin 3-*O*-[2-*O*-(xylosyl)-glucoside] 5-*O*-[6-(malonyl)glucoside]; a3, cyanidin 3-*O*-[2-*O*-(xylosyl)-6-*O*-(*p*-coumaroyl)glucoside] 5-*O*-glucoside; a4, cyanidin 3-*O*-[2-*O*-(2-(sinapoyl)xylosyl)-glucoside] 5-*O*-glucoside; a5, cyanidin 3-*O*-[2-*O*-(xylosyl)-6-*O*-(*p*-coumaroyl)glucoside] 5-*O*-[6-(malonyl)glucoside]; a6, cyanidin 3-*O*-[2-*O*-(xylosyl)-6-*O*-(4(glucosyl)-*p*-coumaroyl)glucoside] 5-*O*-glucoside; a7, cyanidin 3-*O*-[2-*O*-(2-(sinapoyl)xylosyl)-6-*O*-(*p*-coumaroyl)glucoside] 5-*O*-glucoside; a8, cyanidin 3-*O*-[2-*O*-(xylosyl)-6-*O*-(4(glucosyl)-*p*-coumaroyl)glucoside] 5-*O*-[6-(malonyl)glucoside]; a9, cyanidin 3-*O*-[2-*O*-(2-(sinapoyl)xylosyl)-6-*O*-(*p*-coumaroyl)glucoside] 5-*O*-[6-(malonyl)glucoside]; a10, cyanidin 3-*O*-[2-*O*-(2-(sinapoyl)xylosyl)-6-*O*-(4(glucosyl)-*p*-coumaroyl)glucoside] 5-*O*-glucoside; a11, cyanidin 3-*O*-[2-*O*-(2-(sinapoyl)xylosyl)-6-*O*-(4(glucosyl)-*p*-coumaroyl)glucoside] 5-*O*-[6-(malonyl)glucoside].

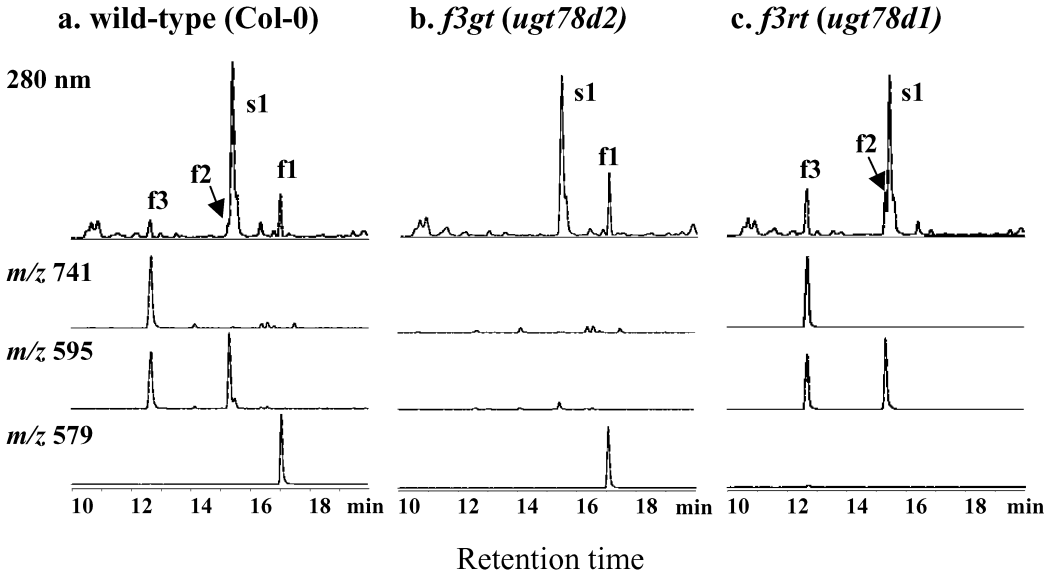


Fig. 2 UPLC/PDA/MS analyses of the extracts of wild-type (Col-0), *ugt78d2*, and *ugt78d1* mutants of leaves. UPLC/PDA and extracted mass chromatograms of aqua-methanol extracts of (a) Arabidopsis wild-type, (b) *f3gt*, and (c) *f3rt* knockout mutants. Absorbance at 280 nm was used for the detection of flavonoids. The extracted mass ion peak chromatograms *m/z* 741, *m/z* 595, and *m/z* 579 correspond to the ions representing [kaempferol + rhamnose + rhamnose + glucose – $3H_2O$ + H]$^+$, [kaempferol + rhamnose + glucose – $2H_2O$ + H]$^+$, and [kaempferol + rhamnose + rhamnose – $2H_2O$ + H]$^+$, respectively. The peak numbers correspond to compounds described in Fig. 1. s1, sinapoyl derivative.

Table 1 The mutants of flavonoid biosynthetic genes used in this study.

AGI code	Mutant name	Enzyme protein		Reference
At5g13930	*tt4*	chalcone synthase	TT4	29,30
At5g07990	*tt7*	flavonoid 3'-hydroxylase	TT7	31
At5g54160	*omt1*	*O*-methyltransferase	AtOMT1	32
At5g17050	*f3gt*	flavonoid 3-*O*-glucosyltransferase	UGT78D2	8
At1g30530	*f3rt*	flavonol 3-*O*-rhamnosyltransferase	UGT78D1	7

Flavonoid profiles of Arabidopsis roots

Flavonoid profiling and peak annotation of Arabidopsis root extracts were carried out by UPLC/PDA/ESI-MS. The flavonoid pattern of roots was more diverse than that of leaves. We detected the known metabolites (f1, f2, f3, f6, and f8) and other unknown ones (f10, f11, f12, f14, and f15) (Fig. 3A-a). Five mutants of flavonoid enzyme genes were analyzed to correlate the accumulation of specific flavonoids and their gene functions.

The *tt4* mutant [29,30] lacking chalcone synthase (TT4/*CHS*, At5g13930)—the first committed enzyme in flavonoid synthesis—was analyzed for its flavonoid profile. No flavonoid glycosides were detected by UV absorbance at 280 nm (Fig. 3A-b). The *tt7* mutant is defective in the gene encoding flavanone 3'-hydroxylase (TT7/*F3'H*, At5g07990) that catalyzes the first step in the synthesis of quercetin and isorhamnetin derivatives. As expected, the *tt7* mutant did not accumulate the quercetin and isorhamnetin glycosides (f6, f8, f14, and f15: Fig. 3A-c) [33]. The *omt1* mutant is defective in the *AtOMT1* gene

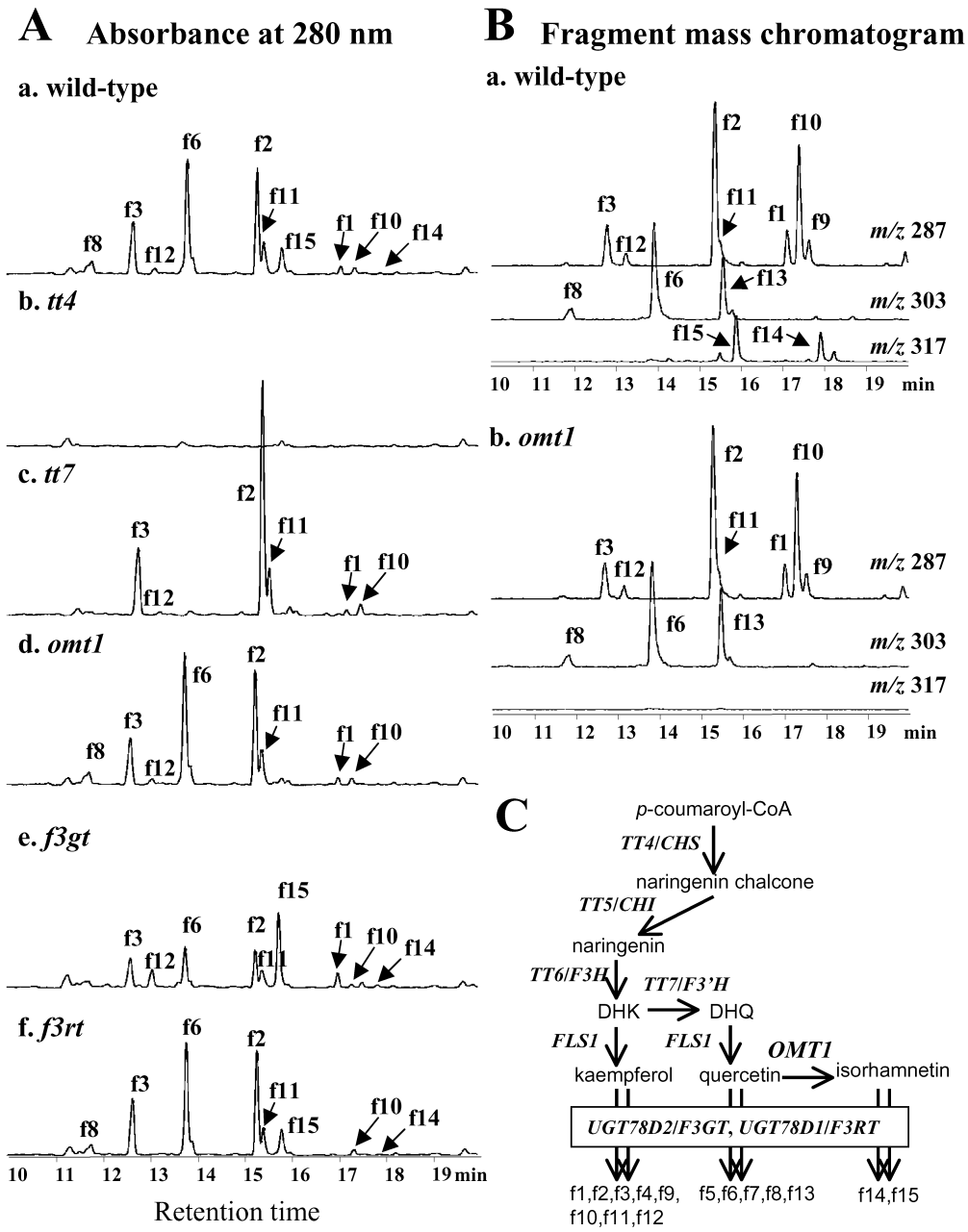


Fig. 3 UPLC/PDA/MS analyses of the extracts of wild-type (Col-0) and knockout mutants of roots. (A) UPLC/PDA profiles of aqua-methanol extracts of (A-a) Arabidopsis wild-type, (A-b) *tt4*, (A-c) *tt7*, (A-d) *omt1*, (A-e) *f3gt*, and (A-f) *f3rt* mutants. Absorbance at 280 nm was used for the detection of flavonoid glycosides. (B) Extracted fragment mass chromatograms of aqua-methanol extracts of (B-a) Arabidopsis wild-type and (B-b) the *omt1* mutant. The extracted fragment mass ions *m/z* 287, *m/z* 303, and *m/z* 317 corresponding to [kaempferol moiety + H]$^+$, [quercetin moiety + H]$^+$, and [isorhamnetin moiety + H]$^+$, respectively, were monitored. (C) Flavonoid pathway, products, and genes in Arabidopsis. *TT4*/*CHS*, chalcone synthase; *TT5*/*CHI*, chalcone isomerase; *TT6*/*F3H*, flavanone 3-hydroxylase, *TT7*/*F3'H*, flavonoid 3'-hydroxylase; *FLS1*, flavonol synthase; *UGT78D2*/*F3GT*, flavonoid 3-*O*-glucosyltransferase; *UGT78D1*/*F3RT*, flavonol 3-*O*-rhamnosyltransferase; DHK, dihydrokaempferol; DHQ, dihydroquercetin.

encoding *O*-methyltransferase (At5g54160). This OMT was reported to be responsible for the biosynthesis of the S units in lignin and of soluble sinapoyl esters [32]. The f14 and f15 flavonoid peaks were absent in the *omt1* mutant as indicated by UV absorbance at 280 nm (Fig. 3A-d). The concentration of the 3-glucosylated flavonoids (f2, f3, f6, f8, and f11) was decreased in the *f3gt* mutant roots but was not completely diminished (Fig. 3A-e). We previously found [8] that, under a high sucrose growth condition, kaempferol 3-glucosides were still detected in the *f3gt* mutant. These results suggested that other 3-*O*-glucosyltransferases besides UGT78D2 should be involved in the flavonoid production in roots and are induced by sucrose stress in Arabidopsis. In *f3rt* mutant roots, two kaempferol 3-*O*-rhamnosides (f1 and f12) were not detected as expected (Fig. 3A-f).

Figure 3B shows the extracted mass chromatograms of the mass fragments due to the flavonoid aglycones in wild-type plants and *omt1* mutants. The chromatograms were monitored for the aglycone ion of *m/z* 287 ([kaempferol moiety + H]$^+$), 303 ([quercetin moiety + H]$^+$), and 317 ([isorhamnetin moiety + H]$^+$), indicating a clearer assignment of the flavonoids accumulated in the mutant. No isorhamnetin glycosides (f14 and f15) were found to be accumulated in the *omt1* mutant despite the same accumulation pattern of kaempferol and quercetin derivatives. The *AtOMT1* gene was previously identified as an OMT involved in the lignin pathway [32]. In the *omt1* mutant, it was found that relative to the wild-type plants, lignins lack syringyl units and the unusual derivatives assigned to 5-OH-feruloyl malate and 5-OH-feruloyl glucose. On the other hand, Muzac et al. (2000) [34] reported that the recombinant AtOMT1 protein could convert quercetin to isorhamnetin. Our present flavonoid profile data of the *omt1* mutant combined with the in vitro results of Muzac et al. (2000) lead to the conclusion that *AtOMT1* (At5g54160) is responsible not only for lignin methylation, but also for isorhamnetin production by methylating quercetin (Fig. 3C) in Arabidopsis.

Coexpression analysis performed using publicly available transcriptome datasets

The coexpressed gene sets with the given correlation coefficients calculated from 771 GeneChip data from the AtGenExpress consortium (RIKEN; <http://pfg.psc.riken.jp/AtGenExpress/>, MaxPlanck; <http://www.weigelworld.org/resources/microarray/AtGenExpress/>) have been released by ATTED-II (<http://www.atted.bio.titech.ac.jp/>) [19]. The coexpression data sets were collected and classified according to the microarray experimental sets from developmental series experiments (237 data), stress treatment experiments (298 data), hormone treatment experiments (236 data), and all experiments (771 data). The Web-based application program at the RIKEN-PRIMe (<http://prime.psc.riken.jp/>) Web site, including ATTED-II ver. 1 data, was used for coexpression analysis in this study. Eighteen genes encoding flavonoid biosynthetic enzymes and transcription factors (Fig. 4) were used for correlation analysis against all Arabidopsis genes. The coexpression combinations of two genes exhibiting a correlation coefficient value of greater than 0.5 ($r > 0.5$) either in all data or in subsets of data were shown as solid or dotted lines, respectively. The coexpression frameworks of flavonoid genes were classified into three clusters. The largest cluster contains *CHS*, *CHI*, *F3H*, *F3'H*, *FLS*, *F3GT*, *F3RT*, and *MYB12*, representing the general flavonoid/flavonol subgroup; the second cluster represents the anthocyanin subgroup containing *DFR*, *A5GT*, *PAP1*, and *PAP2*; and the third cluster containing *BAN*, *LAC*, *TT2*, and *TT8*, represents the proanthocyanidin subgroup. The anthocyanin and proanthocyanidin clusters were composed by the correlations calculated from stress treatment experiments and developmental experiments, respectively. This clustering was supported by the fact that the respective transcription factors regulating each class of compounds were also coexpressed with the structural genes; i.e., MYB12 (At2g47460) for flavonol production [35], PAP1 (MYB75, At1g56650) and PAP2 (MYB90, At1g66390) for anthocyanin production [8,36], and TT2 (MYB123, At5g35550) and TT8 (bHLH042, At4g09820) for proanthocyanidin production [37,38]. The transcription factor PAP1 upregulated several flavonoid pathway genes encoding the general flavonoid/anthocyanin biosynthetic enzymes from CHS to A5GT [8]. However, the *PAP1* gene was coexpressed to a greater extent with the genes encoding anthocyanin-specific enzymes [dihydroflavonol 4-reductase (DFR) and anthocyanin 5-glucosyl-

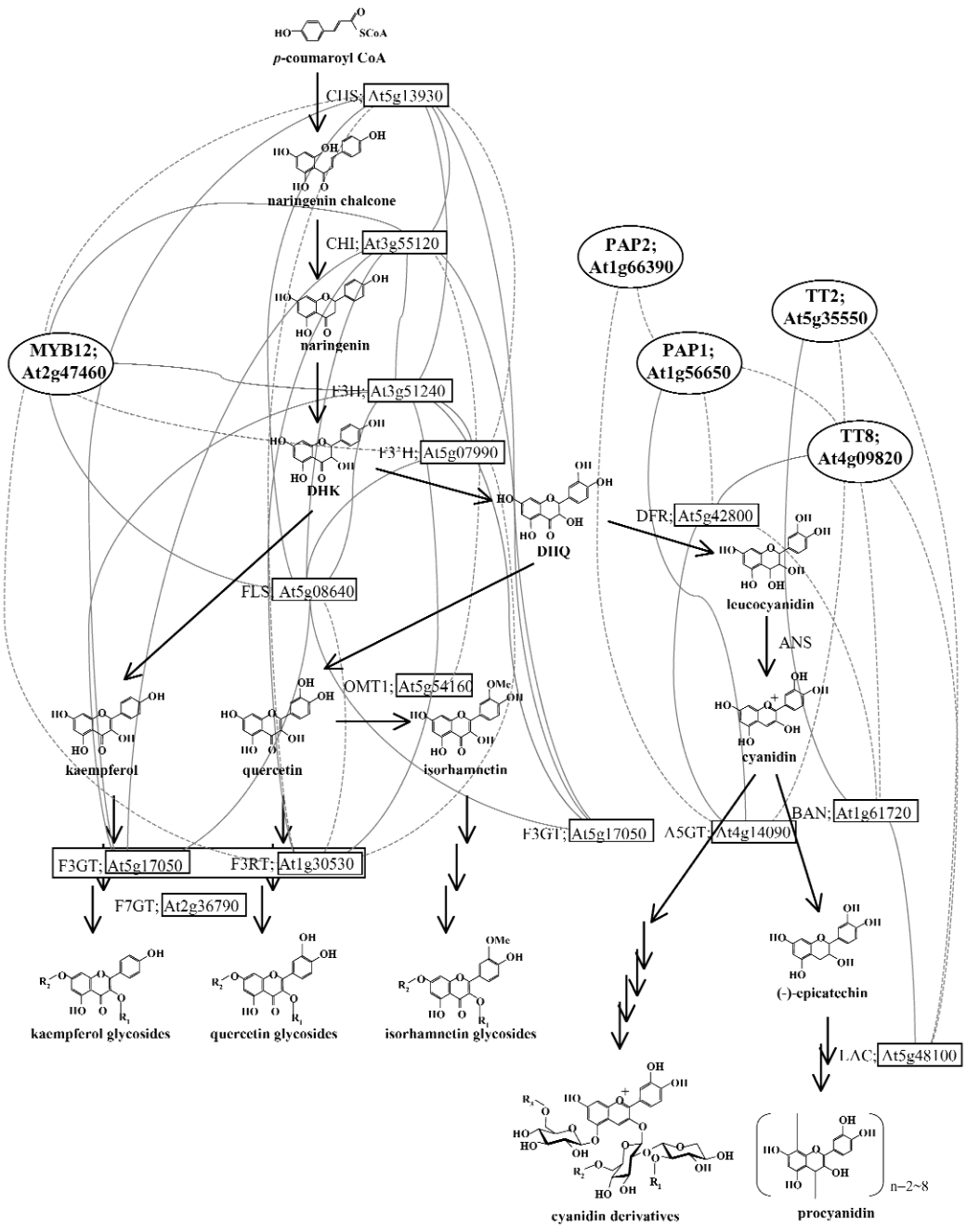


Fig. 4 Coexpression relationships of genes involved in the flavonoid pathway in Arabidopsis. Coexpression relationships among genes encoding flavonoid biosynthetic enzymes and transcription factors were overlaid on known metabolic pathways. Circles represent genes encoding transcription factors. Correlations ($r > 0.5$) are shown by red lines. The solid or dotted lines represent the correlations calculated from all experiments (solid line) or developmental series and/or stress treatment experiments (dotted line). CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavonoid 3-hydroxylase; F3'H, flavonoid 3'-hydroxylase; FLS, flavonol synthase; DFR, dihydroflavonol reductase; ANS, anthocyanidin synthase; OMT, *O*-methyltransferase; BAN, BANYULS anthocyanidin reductase; LAC, laccase; F3GT, flavonoid 3-*O*-glucosyltransferase; F3RT, flavonol 3-*O*-rhamnosyltransferase; F7GT, flavonol 7-*O*-glucosyltransferase; A5GT, anthocyanin 5-*O*-glucosyltransferase.

transferase (A5GT)] than with the common flavonoid genes, suggesting a strong correlation with the anthocyanin-specific genes.

The expression of the *AtOMT1* gene (At5g54160) was highly correlated with the lignin pathway genes, i.e., *CCoAMT1* (At1g34050, $r = 0.823$), *PAL1* (At2g37040, $r = 0.759$), *4CL2* (At3g21240, $r = 0.700$), *4CL1* (At1g51680, $r = 0.653$), *CAD4* (At3g19450, $r = 0.648$), *CAD5* (At4g34230, $r = 0.629$), and *C4H* (At2g30490, $r = 0.618$), based on data from all experiments. In the sub-datasets from developmental experiments, the *AtOMT1* gene exhibits a low but substantial correlation ($r = 0.534$) to *F3'H* (At5g07990). These results indicated that AtOMT1 is primarily responsible for lignin production. However, the results of the flavonoid profiling of the *omt1* mutant indicated that AtOMT1 contributes to isorhamnetin production in the *O*-methylation of quercetin after the F3'H reaction. This result suggests that correlation analysis combined with metabolite profiling should be a powerful technology of the functional prediction of the flavonoid biosynthesis genes.

Proposition of a general strategy for phytochemical genomics in Arabidopsis

We would like to propose a general strategy for functional genomics regarding the production of phytochemicals in *A. thaliana*. This strategy involves the integrated analysis of metabolic profiling, transcriptome coexpression analysis, and the sequence similarity analysis of genes, followed by the reverse genetics and biochemical approaches of narrowed-down candidate genes for a particular reaction or regulation (Fig. 5). A metabolic profiling is performed using high-throughput analysis based on the sensitivity and selectivity of MS with UPLC coupled to UV-PDA. A coexpression analysis of the transcriptome is carried out using Web-based databases calculated from publicly available microarray datasets. A sequence similarity analysis is performed by using genetics and molecular biology databases such as National Center for Biotechnology Information (NCBI, <http://www.ncbi.nlm.nih.gov/>) and The Arabidopsis Information Resource (TAIR, <http://www.arabidopsis.org/index.jsp>). After listing the candidate genes potentially involved in the reactions or regulation of the metabolic pathway, confirmation of gene function can be carried out by the classical reverse genetics approach by using knockout mutants and/or by biochemical characterization of the recombinant proteins. To do this, one can take the advantage of the availability of knockout mutant lines and full-length cDNA libraries, for instance, from SALK Institute (<http://signal.salk.edu/>) [27] and RIKEN BioResource Center (<http://www.brc.riken.jp/lab/epd/Eng/>) [28]. This strategy is actually an expansion of our previously published integrated functional genomics studies [8,14,15]. If the huge metabolome datasets are made available in a public domain in a manner similar to the transcriptome datasets, this strategy will be much better enforced. In this context, we have to collect the metabolome data as comprehensibly as possible and release it in the public domain as a uniform format of the description of data from different laboratories.

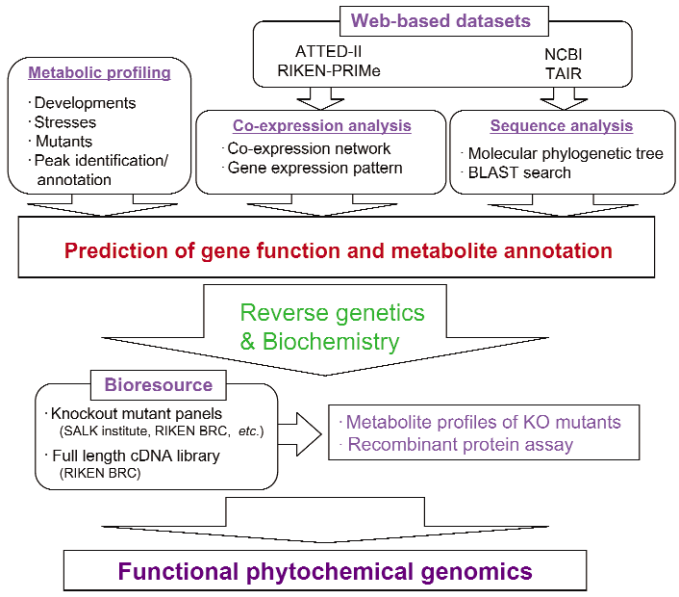


Fig. 5 Proposition of the functional genomics strategy for genes involved in metabolism in *A. thaliana*.

METHODS

Plant materials

A. thaliana ecotype Columbia-0 seed (Lehle Seeds, Round Rock, TX) was used as the wild-type plant in this study. The *f3gt* (At5g17050) mutant and *f3rt* (At1g30530) mutants were described previously [7,8]. The T-DNA-inserted knockout mutant of *omt1* (At5g54160, CS25167) and *tt7* (At5g07990, CS6509) were obtained from the Arabidopsis Biological Resource Center (ABRC). The *tt4* (At1g30530) mutant was obtained from *tt* mutant lines induced by ion beam irradiation of Arabidopsis [30]. The plants were cultured on Murashige and Skoog (MS)-agar medium containing 1 % sucrose in a growth chamber at 22 °C in 16 h/8 h light/dark cycles for 18 days. The leaves and roots of plants were harvested, immediately frozen with liquid nitrogen, and stored at –30 °C until use.

Flavonoid profiling by UPLC/PDA/ESI-Q-TOF/MS

Frozen leaves and roots were homogenized in 5 µl extraction solvent (methanol:H_2O = 4:1) per 1 mg fresh weight of tissues by using a mixer mill (MM300: Retsch GmbH & Co. KG, Haan, Germany) for 3 min at 30 Hz. After centrifugation at 12 000 × g, the cell debris was discarded, and the extracts were centrifuged again. These supernatants were immediately used for the flavonoid analysis.

For flavonoid profiling, a Waters Acquity UPLC^{TM} system (Waters Co., Massachusetts) fitted with Q-Tof Premier mass spectrometer (Micromass MS Technologies, Manchester, UK) was used. UPLC was carried out on a UPLC^{TM} BEH C_{18} column (Φ2.1 mm × 100 mm, Waters) at a flow rate of 0.5 ml/min at 35 °C. An elution gradient with solvent A (0.1% formic acid in H_2O) and solvent B (0.1 % formic acid in acetonitrile) and the elution profile—0 min, 100 % A; 20 min, 20 % B—using linear gradients in between the time points was applied. PDA was used for the detection of UV–vis absorption in the range of 210–600 nm. The time-of-flight (TOF) mass analyzer was used for the detection of flavonoid glycosides $[M + H]^+$ and the peak of fragment ions in a positive ion scanning mode with the following setting: desolvation temperature, 180 °C with a nitrogen gas flow of 500 l/h; capillary spray, 3.2 kV; source temperature, 100 °C; and cone voltage was 35 V.

Kaempferol 3-*O*-rhamnoside 7-*O*-rhamnoside (f1), kaempferol 3-*O*-glucoside 7-*O*-rhamnoside (f2), and kaempferol 3-*O*-[rhamnosyl(1 → 2)glucoside] 7-*O*-rhamnoside (f3) were provided by Dr. H. Takayama, Chiba University, or obtained from AnalytiCon Discovery (Potsdam, Germany). These standard flavonol glycosides were used for the identification of the peaks in the plant extracts based on retention times, UV–vis absorption spectra, and mass fragmentation by tandem MS analysis. Other flavonoids were annotated by comparing their UV–vis absorption spectra, elution time, *m/z* values, and MS^2 fragmentation patterns with the reported data [3–5,7,8].

Coexpression analysis

Coexpression analyses were carried out using the coexpression gene search program at the RIKEN-PRIMe Web site (<http://prime.psc.riken.jp/>). The coexpression gene search program is a Web-based application to search for correlated genes from several gene expression data. The expression data are produced using Affymetrix GeneChip by AtGenExpress (RIKEN, <http://pfg.psc.riken.jp/AtGenExpress/>; MaxPlanck, <http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm>), which is a multinational consortium for Arabidopsis transcriptome analysis. Correlation data have been released in ATTED-II (<http://www.atted.bio.titech.ac.jp/>) [19]. RIKEN-PRIMe (<http://prime.psc.riken.jp/>) includes ATTED-II ver. 1 data on the interconnections of gene sets. Eighteen genes that have been experimentally characterized to be involved in flavonoid production were used as core query genes for coexpression analysis. The genes exhibiting positive correlations ($r > 0.5$) against 18 flavonoid-pathway genes were extracted using "all data sets ver.1 (771 array data)," "tissue and development ver.1 (237 array data)," and "stress treatment ver.1 (298 array data)" in the AtGenExpress datasets. Pajek software (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>) and Microsoft Draw 2003 (Microsoft, USA) were used to draw the coexpression graph.

ACKNOWLEDGMENTS

We thank Dr. Takeshi Obayashi (Chiba University, Japan), Dr. Hideyuki Suzuki, and Dr. Nozomu Sakurai (Kazusa DNA Institute, Japan) for advice on coexpression analysis and Mr. Kenji Akiyama for providing excellent technical support for the RIKEN-PRIMe Web site. We thank Prof. Hiromitsu Takayama (Chiba University, Japan) for providing the flavonoid standards and Mr. Ryo Nakabayashi for confirmation of the chemical structure of a kaempferol derivative. We thank Dr. Satoshi Kitamura (Japan Atomic Energy Research Institute) for providing the *tt4* mutant. This study was supported in part by grants-in-aid from the Ministry of Education, Science, Culture, Sports and Technology, Japan, and the Uehara Memorial Foundation, Japan.

REFFERENCES

1. R. A. Dixon, D. Strack. *Phytochemistry* **62**, 815 (2003).
2. J. B. Harborne (Ed.). *The Flavonoid*: *Advances in Research Since 1986*, Chapman & Hall, London (1994).
3. T. L. Graham. *Plant Physiol*. *Biochem*. **36**, 135 (1998).
4. M. Veit, G. F. Pauli. *J*. *Nat*. *Prod*. **62**, 1301 (1999).
5. S. J. Bloor, S. Abrahams. *Phytochemistry* **59**, 343 (2002).
6. B. Winkel-Shirley. *Plant Physiol*. **126**, 485 (2001).
7. P. Jones, B. Messner, J. Nakajima, A. R. Schaffner, K. Saito. *J*. *Biol*. *Chem*. **278**, 43910 (2003).
8. T. Tohge, Y. Nishiyama, M. Y. Hirai, M. Yano, J. Nakajima, M. Awazuhara, E. Inoue, H. Takahashi, D. B. Goodenowe, M. Kitayama, M. Noji, M. Yamazaki, K. Saito. *Plant J*. **42**, 218 (2005).
9. O. Fiehn. *Plant Mol*. *Biol*. **48**, 155 (2002).

10. R. J. Bino, R. D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. J. Nikolau, P. Mendes, U. Roessner-Tunali, M. H. Beale, R. N. Trethewey, B. M. Lange, E. S. Wurtele, L. W. Sumner. *Trends Plant Sci*. **9**, 418 (2004).
11. J. Kopka, A. Fernie, W. Weckwerth, Y. Gibon, M. Stitt. *Genome Biol.* **5**, 109 (2004).
12. K. M. Oksman-Caldentey, K. Saito. *Curr. Opin. Biotechnol*. **16**, 174 (2005).
13. A. Goossens, S. T. Hakkinen, I. Laakso, T. Seppanen-Laakso, S. Biondi, V. De Sutter, F. Lammertyn, A. M. Nuutila, H. Soderlund, M. Zabeau, D. Inze, K. M. Oksman-Caldentey. *Proc. Natl. Acad. Sci. USA* **100**, 8595 (2003).
14. M. Y. Hirai, M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, K. Saito. *Proc. Natl. Acad. Sci. USA* **101**, 10205 (2004).
15. M. Y. Hirai, K. Marion, Y. Fujikawa, M. Yano, D. B. Goodenowe, Y. Yamazaki, S. Kanaya, Y. Nakamura, M. Kitayama, H. Suzuki, N. Sakurai, D. Shibata, J. Tokuhisa, M. Reichelt, J. Gershenzon, J. Papenbrock, K. Saito. *J. Biol. Chem*. **280**, 25590 (2005).
16. T. Tokimatsu, N. Sakurai, H. Suzuki, H. Ohta, K. Nishitani, T. Koyama, T. Umezawa, N. Misawa, K. Saito, D. Shibata. *Plant Physiol*. **138**, 289 (2005).
17. P. Zhang, H. Foerster, C. P. Tissier, L. Mueller, S. Paley, P. D. Karp, S. Y. Rhee. *Plant Physiol*. **138**, 27 (2005).
18. O. Thimm, O. Blasing, Y. Gibon, A. Nagel, S. Meyer, P. Kruger, J. Selbig, L. A. Muller, S. Y. Rhee, M. Stitt. *Plant J*. **37**, 914 (2004).
19. T. Obayashi, K. Kinoshita, K. Nakai, M. Shibaoka, S. Hayashi, M. Saeki, D. Shibata, K. Saito, H. Ohta. *Nucleic Acids Res*. **35**, D863 (2007).
20. D. Steinhauser, B. Usadel, A. Luedemann, O. Thimm, J. Kopka. *Bioinformatics* **12**, 3647 (2004).
21. C. H. Jen, I. W. Manfield, I. Michalopoulos, J. W. Pinney, W. G. Willats, P. M. Gilmartin, D. R. Westhead. *Plant J*. **46**, 336 (2006).
22. S. Persson, H. Wei, J. Milne, G. P. Page, C. R. Somerville. *Proc. Natl. Acad. Sci. USA* **102**, 8633 (2005).
23. J. Lisso, D. Steinhauser, T. Altmann, J. Kopka, C. Mussig. *Nucleic Acids Res*. **33**, 2685 (2005).
24. C. M. M. Gachon, M. Langlois-Meurinne, Y. Henry, P. Saidrenan. *Plant Mol. Biol.* **58**, 229 (2005).
25. M. Noji, C. G. Kawashima, T. Obayashi, K. Saito. *Amino Acids* **30**, 163 (2006).
26. H. Wei, S. Persson, T. Mehta, V. Srinivasasainagendra, L. Chen, G. P. Page, C. Somerville, A. Loraine. *Plant Physiol*. **142**, 762 (2006).
27. J. M. Alonso, A. N. Stepanova, T. J. Leisse, C. J. Kim, H. Chen, P. Shinn, D. K. Stevenson, J. Zimmerman, P. Barajas, R. Cheuk, C. Gadrinab, C. Heller, A. Jeske, E. Koesema, C. C. Meyers, H. Parker, L. Prednis, Y. Ansari, N. Choy, H. Deen, M. Geralt, N. Hazari, E. Hom, M. Karnes, C. Mulholland, R. Ndubaku, I. Schmidt, P. Guzman, L. Aguilar-Henonin, M. Schmid, D. Weigel, D. E. Carter, T. Marchand, E. Risseeuw, D. Brogden, A. Zeko, W. L. Crosby, C. C. Berry, J. R. Ecker. *Science* **301**, 653 (2003).
28. M. Seki, M. Narusaka, A. Kamiya, J. Ishida, M. Satou, T. Sakurai, M. Nakajima, A. Enju, K. Akiyama, Y. Oono, M. Muramatsu, Y. Hayashizaki, J. Kawai, P. Carninci, M. Itoh, Y. Ishii, T. Arakawa, K. Shibata, A. Shinagawa, K. Shinozaki. *Science* **296**, 141 (2002).
29. B. W. Shirley, W. L. Kubasek, G. Storz, E. Bruggemann, M. Koornneef, F. M. Ausubel, H. M. Goodman. *Plant J*. **8**, 659 (1995).
30. N. Shikazono, Y. Yokota, S. Kitamura, C. Suzuki, H. Watanabe, S. Tano, A. Tanaka. *Genetics* **163**, 1449 (2003).
31. C. Schoenbohm, S. Martens, C. Eder, G. Forkmann, B. Weisshaar. *Biol. Chem*. **381**, 749 (2000).
32. T. Goujon, R. Sibout, B. Pollet, B. Maba, L. Nussaume, N. Bechtold, F. Lu, J. Ralph, I. Mila, Y. Barriere, C. Lapierre, L. Jouanin. *Plant Mol. Biol*. **51**, 973 (2003).
33. L. Lepiniec, I. Debeaujon, J. N. Routaboul, A. Baudry, L. Pourcel, N. Nesi, M. Caboche. *Annu. Rev. Plant Biol*. **57**, 405 (2006).

34. I. Muzac, J. Wang, D. Anzellotti, H. Zhang, R. K. Ibrahim. *Arch. Biochem. Biophys*. **375**, 385 (2000).
35. F. Mehrtens, H. Kranz, P. Bednarek, B. Weisshaar. *Plant Physiol*. **138**, 1083 (2005).
36. J. O. Borevitz, Y. Xia, J. Blount, R. A. Dixon, C. Lamb. *Plant Cell* **12**, 2383 (2000).
37. S. B. Sharma, R. A. Dixon. *Plant J.* **44**, 62 (2005).
38. N. Nesi, I. Debeaujon, C. Jond, G. Pelletier, M. Caboche, L. Lepiniec. *Plant Cell* **12**, 1863 (2000).