Phytochem Rev (2009) 8:15–23
DOI 10.1007/s11101-008-9114-4

A robust omics-based approach for the identification of glucosinolate biosynthetic genes

Masami Yokota Hirai

Received: 25 March 2008 / Accepted: 10 October 2008 / Published online: 29 November 2008

Abstract Transcriptome coexpression analysis, which is based on a vast amount of transcriptome data obtained by using DNA arrays, has become a routine method for functional genomics studies in *Arabidopsis*. This analysis enables us to predict the function of genes on the basis of a simple assumption that a set of genes involved in a particular biological process can be coexpressed under the control of a shared regulatory system. Candidate genes involved in glucosinolate biosynthesis were successfully identified by this approach. In this review, the methodology of coexpression analysis is briefly described. The advantages and disadvantages of this analysis are also discussed in the context of its ability to predict gene functions involved in glucosinolate biosynthesis.

Keywords Coexpression · Correlation · Gene function · Network · Transcriptome · Prediction

M. Y. Hirai (✉)
RIKEN Plant Science Center, 1-7-22 Suehiro-cho,
Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan
e-mail: myhirai@psc.riken.jp

M. Y. Hirai
JST, CREST, 4-1-8 Hon-chou, Kawaguchi, Saitama
332-0012, Japan

Introduction

Over the last few years, transcriptome coexpression analysis has become a routine method for functional genomics studies in *Arabidopsis*. In this analysis we predict the function of genes on the basis of a simple assumption that a set of genes involved in a particular biological process can be coexpressed under the control of a shared regulatory system. In other words, if a gene of an unknown function is coexpressed with a set of genes involved in a particular biological process, it can be assumed to be one of the components of the same biological process. The development of comprehensive methods to measure mRNA accumulation such as DNA array and bioinformatics tools for handling large-scale datasets has enabled transcriptome coexpression analysis based on hundreds of transcriptome data. Individual biologists perform transcriptome analysis using DNA arrays to find answers to specific questions, such as how gene expression patterns change in plants under specific conditions of interest. DNA array data thus obtained are deposited in public databases. On the other hand, DNA array data has been systematically acquired by the AtGenExpress (Goda et al. 2008; Kilian et al. 2007; Schmid et al. 2005) and NASCArrays (Craigon et al. 2004) by using the same analytical platform, i.e., Affymetrix GeneChip microarray. This led to the development of secondary databases equipped with web-based coexpression analysis tools that help in calculating and storing the information regarding the level of similarity of gene expression patterns, and this information is made available to users.

We have analyzed the transcriptome of nutrient-starved *Arabidopsis* since 2001. During the data-mining of the in-house dataset obtained in our lab, we found that coexpression analysis is a powerful technique to identify candidate genes involved in glucosinolate (GSL) biosynthesis. More recently, as mentioned above, the development of web-based analytical tools has enhanced the predicting power of transcriptome coexpression analysis. In this review I describe briefly the methodology of coexpression analysis and discuss its advantages and disadvantages of this analysis in the context of its ability to predict gene functions involved in GSL biosynthesis. For a general review of coexpression analysis and network representation of coexpression relationship, please refer to other reviews (Aoki et al. 2007; Saito et al. 2008). In order to avoid any overlap with other reviews in this special issue, I have not discussed the details of the characterization of gene functions.

A brief overview of coexpression analyses

In coexpression analysis the degree of similarity of gene expression patterns across a variety of experimental conditions is evaluated by calculating the similarity between pairs of genes using statistical measures such as Pearson's correlation coefficient (PCC). Both in-house datasets and publicly available datasets can be utilized for calculation of similarity, although the results obtained would differ. In-house transcriptome data are often obtained under specific condition of interest (e.g. sulfur-starvation condition in my study), and hence higher similarity of expression pattern indicates that the coexpression relationship occurs only under the specific condition of interest (e.g. coexpression under sulfur-starvation), that is, condition-dependent coexpression relationship. In contrast, thousands of transcriptome data available in the public databases have been obtained under a wide range of experimental conditions, and hence the higher similarity coefficient calculated on the basis of publicly available dataset indicates a constitutive or condition-independent coexpression relationship; that is, a set of genes with higher similarity are coexpressed across a variety of experimental conditions. Many coexpression analysis tools have been recently released, such as ATTED-II (the *Arabidopsis thaliana* trans-factor and cis-element prediction database) (Obayashi et al. 2007), CSB.DB (the Comprehensive Systems-Biology Database) (Steinhauser et al. 2004), BAR (the Botany Array Resource) (Toufighi et al. 2005), ACT (the Arabidopsis Co-expression Tool) (Jen et al. 2006; Manfield et al. 2006), Genevestigator (Zimmermann et al. 2005; Zimmermann et al. 2004), PED (Plant Gene Expression Database) (Horan et al. 2008) and Cress-Express (Srinivasasainagendra et al. 2008). Some of these provide users the option of calculating the degree of similarity for a dataset. When a whole dataset (e.g. all data obtained by AtGenExpress) is selected for analysis, the constitutive coexpression relationship is elucidated. However, by selecting a subset of dataset (e.g. developmental-series, stress-series, or hormone-treatment-series data of AtGen-Express), condition-dependent, or context-specific coexpression relationship can be determined, as is the case with the coexpression analysis using in-house datasets.

Coexpression analysis has several advantages in predicting gene functions: (1) Researchers do not need to conduct "wet" experiments in order to predict the function of unknown genes of interest. The coexpression relationship of these genes with genes of known function, as well as the sequence similarity between the 2 sets of genes, will provide clues to predict gene function. (2) Researchers can identify the components involved in a particular biological process. Apparently, however, coexpression analysis does work for this purpose only when a complete biological process is coordinately regulated at the level of mRNA accumulation. (3) Even if the knockout of genes belonging to a gene family, the members of which have unknown biological function, fails to reveal any apparent phenotype, the function of these genes can be predicted on the basis of their coexpression relationship with other genes (Rautengarten et al. 2005). (4) Coexpression analysis can even be conducted using a non-targeted approach without any preexisting hypothesis. In other words, coexpression relationship can often be determined from a set of transcriptome data irrespective of the original purpose of the experiments by which the data were obtained.

Prediction of the genes involved in glucosinolate biosynthesis – a case study of coexpression analysis

In this section I briefly describe the transcriptome analysis of nutrient-starved *Arabidopsis* conducted in our lab. As mentioned above, during the course of our study, we realized that coexpression analysis is considerably useful for identifying candidate genes involved in GSL biosynthesis.

In order to understand the plant's response to sulfur deficiency by omics-based approach, we conducted an integrated analysis of the transcriptome and metabolome of sulfur-starved *Arabidopsis* (Hirai and Saito 2008; Hirai et al. 2004, 2005). Time-series data for the transcriptome and the metabolome of leaves and roots were obtained, and analyzed by batch-learning self-organizing mapping (BL-SOM), a sophisticated form of multivariate analysis (Abe et al. 2003; Kanaya et al. 2001). BL-SOM, along with other clustering algorithms such as k-means and hierarchical clustering, can be used for co-occurrence analysis of genes and metabolites. When BL-SOM is applied to transcriptome and/or metabolome data, the genes and/or metabolites can be classified into the cells on a 2-dimensional lattice called a feature map on the basis of the similarity of expression and/or accumulation patterns. In this analysis, we defined a set of co-occurring genes and/or metabolites as a cluster. We identified many clusters, for example, a set of the genes involved in anthocyanin biosynthesis and a set of those involved in sulfate assimilation (Hirai et al. 2005). Several Met- and Trp-derived GSLs were classified into a single cluster, suggesting that GSL metabolism is coordinately regulated under sulfur deficiency. This idea was supported by the finding that the known GSL biosynthetic genes—the *MAM* (*methylthioalkylmalate synthase*), *CYP79* and *CYP83* families, *SUR1* and *AOP2*—were classified into another single cluster. This indicated that GSL biosynthetic genes are coexpressed under sulfur deficiency probably via a shared regulatory mechanism. On the basis of the coexpression relationship with the previously-characterized genes mentioned above, we identified the following genes as candidates involved in GSL biosynthesis: three putative sulfotransferase genes (*AtSOT16*/At1g74100, *AtSOT17*/At1g18590, and *AtSOT18*/At1g74090), an *S*-glucosyltransferase gene (*UGT74B1*/At1g24100), a putative Tyr aminotransferase gene (At5g36160), and two putative glutathione *S*-transferase (GST) genes (*GSTF11*/At3g03190 and *GSTU20*/At1g78370) (Hirai et al. 2005). To date, some of these candidate genes have been characterized experimentally. The predicted functions of the *AtSOT*s and *UGT74B1* have been confirmed by concurrent studies (Hirai et al. 2005; Piotrowski et al. 2004; Douglas Grubb et al. 2004).

In the same analysis using in-house dataset, we identified several genes encoding transcription factors, including *Myb28* (At5g61420) and *Myb29* (At5g07690), as the candidate positive regulators of GSL biosynthesis. We also analyzed constitutive coexpression relationship by ATTED-II (Obayashi et al. 2007) using a whole dataset of AtGenExpress (1,388 ATH1 arrays), and found that *Myb28* and *Myb29* were coexpressed only with the genes involved in Met-derived GSL biosynthesis. The known Met-derived GSL genes were highly coexpressed with *Myb28*, but to a lesser extent with *Myb29*. This analysis suggested that *Myb28* and *Myb29* may be transcription factors positively regulating Met-derived GSL biosynthesis, but not Trp-derived GSL biosynthesis. Reverse-genetic and molecular biological experiments have proved *Myb28* to be a key transcription factor that positively regulates Met-derived GSL biosynthesis and *Myb29* to be a transcription factor probably involved in methyl jasmonate-mediated induction of GSL biosynthesis (Hirai et al. 2007). Concurrently, several groups have independently found that *Myb28*, *Myb29* and *Myb76* (At5g07700) are the positive regulators of Met-derived GSL biosynthesis and that *Myb51* (At1g18570) and *Myb122* (At1g74080), as well as previously-characterized *Myb34* (At5g60890), are the positive regulators of Trp-derived GSL biosynthesis (Beekwilder et al. 2008; Gigolashvili et al. 2007a–c; Sonderby et al. 2007; Malitsky et al. 2008). These authors have discussed the specific functions of individual *Myb*s, the mutual regulation among these *Myb*s and the mutual regulation between Met- and Trp-derived GSL pathways (see other reviews in this issue).

In our analysis, *AtBCAT-3* (At3g49680) and *AtBCAT-4* (At3g19710) were also coexpressed with the Met-derived GSL biosynthetic genes of known function (Hirai et al. 2007), suggesting the involvement of these genes in Met side-chain elongation. The function of these genes has recently been confirmed, and *AtBCAT-3* was shown to function in both GSL and amino acid biosynthesis (Knill et al. 2008; Schuster et al. 2006). We also identified other candidate genes involved in Met-derived GSL biosynthesis, although these predicted functions remain to be confirmed: *AtGSTU20*, *AtGSTF11*, *PMSR2* (At5g07460), and the homologs of bacterial Leu biosynthetic genes named *AtLeuC1* (At4g13430), *AtLeuD1* (At2g43100), *AtLeuD2* (At3g58990), and *AtIMD1* (At5g14200). With regard to the GST genes, it has been suggested that GST-type enzymes may be components of an enzyme complex formed by CYP83s and C-S lyase (Mikkelsen et al. 2004). The *PMSR2* gene encodes a cytosolic peptide methionine sulfoxide reductase. Because a null mutation in this gene resulted in reduced growth in *Arabidopsis* under short-day conditions, it was hypothesized that the role of PMSR2 is to repair oxidized proteins in a short-day photoperiod (Bechtold et al. 2004). We speculate that the PMSR2 protein can recognize the methylsulfinyl moiety of methylsulfinylalkyl GSL as well as that of peptide methionine sulfoxide, and that hence, this enzyme may have some function in the side-chain conversion of Met-GSLs, although $\mathrm{FMO}_{\mathrm{GS\text{-}OX}}$ has been shown to be responsible for the conversion of methylthioalkyl GSLs to methylsulfinylalkyl GSLs (Hansen et al. 2007). We assumed that the homologs of Leu biosynthetic genes are involved in Met side-chain elongation for the following reason. The reactions involved in Met side-chain elongation are similar to those involved in Leu biosynthesis; moreover, the enzymes involved in Met side-chain elongation and Leu biosynthesis are presumably encoded by homologous genes belonging to the same gene families. In fact, *MAM* genes and *IPMS* (*isopropylmalate synthase*) genes, which are responsible for Met side-chain elongation and Leu synthesis, respectively, share sequence similarity with each other and with bacterial *IPMS* (de Kraker et al. 2007; Field et al. 2004; Kroymann et al. 2001). All of the above-mentioned candidate genes are under the transcriptional regulation involving *Myb28* (Hirai et al. 2007). In addition, *UGT74C1* (At2g31790), which is assumed to be involved in Met-derived GSL biosynthesis on the basis of the coexpression analysis (Gachon et al. 2005), is positively regulated by *Myb28* (Hirai et al. 2007). On the other hand, a putative Tyr aminotransferase gene mentioned above is not regulated by *Myb28* (Hirai et al. 2007), suggesting that it may encode a C-S lyase involved only in Trp-/Phe-derived GSL biosynthesis. The reason for this assumption was that the C-S lyase gene *SUR1* had been originally misannotated as a Tyr aminotransferase. Another possibility is that this gene may encode a Phe aminotransferase. *Arabidopsis* ecotype Columbia contains 2-phenylethyl GSL derived from homoPhe. If homoPhe is formed from Phe via a reaction mechanism similar to that involved in the formation of homoMet from Met, Phe must be transaminated by an aminotransferase prior to condensation with acetyl-CoA for the side chain to extend.

The advantages and limitations of coexpression analysis for glucosinolate biosynthetic genes

As described above, the coexpression analysis could predict many, although not all, of the genes involved in the biosynthesis of GSLs, especially Met-derived GSLs. This implies that the genes responsible for Met-derived GSL biosynthetic pathway (side-chain elongation, core structure formation, and side-chain modification) may be coordinately controlled by a limited number of regulatory components including *Myb28*, *Myb29*, and *Myb76*, at the mRNA accumulation level. Coexpression analysis could also effectively predict the candidate genes involved in the other secondary pathways, such as flavonoid and anthocyanin biosynthesis (Tohge et al. 2005; Vandeauwera et al. 2005; Yonekura-Sakakibara et al. 2007).

Quantitative trait locus (QTL) analysis is a powerful tool for identifying candidate genes involved in GSL biosynthesis as well as those involved in hydrolysis, for example, *ESM1* (*Epithiospecifier modifier 1*, At3g14210; Zhang et al. 2006). ATTED-II analysis using a whole dataset showed weak correlation between *ESM1* and *ESP* (*Epithiospecifier protein*, At1g54040) (data not shown). *MAM* genes that encode one of the Met side-chain elongation enzymes were also identified and characterized on the basis of the QTL analysis (Field et al. 2004; Textor et al. 2004; Kroymann et al. 2001, 2003). To my knowledge, however, some other genes that are involved in Met side-chain elongation process, namely, *MAM-I* (coding for methylthioalkylmalate isomerase) and *MAM-D* (coding for methylthioalkylmalate dehydrogenase) have not been identified by QTL analysis, presumably because natural variation of these genes does not result in metabolic natural variation. However, coexpression analysis could distinguish candidate genes i.e., *MAM-I* and *MAM-D*, from the putative Leu biosynthetic genes among the members of the same gene families (*AtLeuC*s, *AtLeuD*s, and *AtIMD*s) (Hirai et al. 2007). However, coexpression analysis requires previously-characterized genes such as *MAM*s as “guide genes” (Lisso et al. 2005), with which genes of unknown function are associated depending on whether coexpression relationship occurs. A combination of QTL analysis and coexpression analysis led to the identification of a flavin-monooxygenase (FMO) gene, $FMO_{GS\text{-}OX}$, which is responsible for the side-chain modification of Met-derived GSLs (Hansen et al. 2007).

Although several *Myb* transcription factors controlling GSL biosynthesis could be predicted by coexpression analysis, this methodology is not sufficiently versatile to identify all regulatory genes. While the functions of at least three *Myb*s–*Myb28*, *Myb29*, and *Myb34* could be predicted by coexpression analysis (see Fig. 1), *SLIM1*, which codes for a transcriptional regulator involved in down-regulation of GSL biosynthetic genes under sulfur deficiency, could never be identified by coexpression analysis, because *SLIM1* itself is not regulated at mRNA accumulation level under sulfur deficiency (Maruyama-Nakashita et al. 2006). Presumably, SLIM1 may be post-transcriptionally regulated in response to sulfur deficiency. Among *Myb28*, *Myb29*, and *Myb34*, at least *Myb34* was shown to be down-regulated via a SLIM1-dependent mechanism in the roots of sulfur-starved *Arabidopsis* (Maruyama-Nakashita et al. 2006). The other regulators of GSL metabolism, *IQD1* (At3g09710) (Levy et al. 2005), *TFL2* (At5g17690) (Kim et al. 2004), and *OBP2* (At1g07640) (Skirycz et al. 2006) did not show any obvious correlation with the known GSL biosynthetic genes in an ATTED-II analysis performed using a whole dataset (data not shown).

Figure 1 is a graph (so-called network) that indicates the coexpression relationship between the characterized and candidate GSL biosynthetic genes, which has been calculated using a whole AtGenExpress dataset. Among 35 query genes (see figure

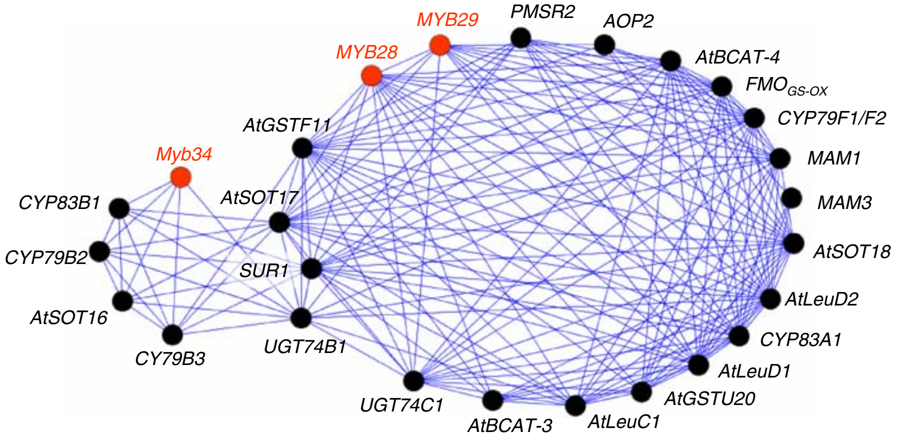

Fig. 1 A correlation network comprising the known and candidate GSL biosynthetic genes. Coexpression relationship was analyzed by using Correlated Gene Search in PRIMe (Platform for RIKEN Metabolomics, http://prime.psc.riken.jp/) (Akiyama et al. 2008) using the following 35 genes as queries: *Myb28*, *Myb29*, *Myb76*, *AtBCAT-4*, *AtBCAT-3*, *MAM1*, *MAM3*, *AtLeuC1*, *AtLeuD1*, *AtLeuD2*, *AtIMD1*, *CYP79F1*, *CYP79F2*, *CYP83A1*, *AtGSTU20*, *AtGSTF11*, *SUR1*, *UGT74B1*, *UGT74C1*, *AtSOT17*, *AtSOT18*, $FMO_{GS\text{-}OX}$, *AOP2*, *PMSR2*, *MYB34*, *MYB51*, *MYB122*, *CYP79B2*, *CYP79B3*, *CYP83B1*, *AtGSTU8*, *AtGSTU3*, *AtSOT16*, putative Tyr aminotransferase, *CYP79A2*. We did not include the genes responsible for Met and Trp biosynthesis into the queries, although some of them are regulated by some *Myb*s described here. *AtGSTU8* and *AtGSTU3* were coexpressed with known GSL biosynthetic genes under sulfur deficiency (Hirai et al. in press). Parameter setting was as follows: Matrix, All data sets v.3 (1,388 data of AtGenExpress); Method, interconnection of sets. The correlation data used in PRIMe have been released by ATTED-II. The gene pairs with PCC greater than 0.50 were selected, and the network was visualized by BioLayout^{Java} (Goldovsky et al. 2005). Transcripts from CYP79F1 and CYP79F2 were cross-hybridized to the same probe sets on a GeneChip microarray and hence are indistinguishable. The lengths of the lines depicted in this type of graph do not have any values

legend), the pairs of coexpressed genes (threshold PCC > 0.5) have been connected by lines. The graph represents 2 partially-overlapping modules. The larger and smaller modules consist mainly of Met- and Trp-derived GSL genes, respectively. The genes specifically involved in Met-derived GSL biosynthesis are not connected directly with those specifically involved in Trp-derived GSL biosynthesis, and vice versa. *SUR1* and *UGT74B1*, the genes involved in both Met- and Trp-derived GSL biosynthesis (Mikkelsen et al. 2004; Douglas Grubb et al. 2004), are in the boundary region of two modules. It has been reported that the preferable substrates of the *AtSOT17* product are Met- and Phe-derived GSLs (Klein et al. 2006; Piotrowski et al. 2004). Although graph structure depends on the dataset and the measure of similarity used, it may possibly suggest the functional relationship of the genes. *Myb51* and *Myb122*, transcriptional regulators of Trp-derived GSL biosynthetic genes (Gigolashvili et al. 2007a), were not connected to any genes in this analysis (Fig. 1). However, *Myb51* and *Myb122* may form a condition-dependent network that can be drawn on the basis of the calculation using a sub dataset such as stress-series data, because at least *Myb51* exhibits an expression pattern different from that of *Myb34* with regards to tissue specificity and response to mechanical stimuli (Gigolashvili et al. 2007a).

Coexpression analysis can be applied to non-model Brassicaceae plants by analyzing their transcript profiles using comprehensive techniques such as cDNA-amplified fragment length polymorphism. In such a study, only a few previously-characterized GSL biosynthetic genes are expected, and hence, parallel analysis of their metabolic profile will help predict candidate genes involved in GSL biosynthesis. Integrated analysis of the transcriptome and the metabolome has led to the elucidation of functions of various other genes in many non-model plants (Saito et al. 2008).

Conclusions and perspectives

As described in this review, coexpression analysis has become an easy-to-use tool for functional genomics studies of *Arabidopsis*. There is certainly a possibility of selecting false positives as candidates, which is the drawback with other genome-wide large-scale analyses. To overcome this problem, novel algorithms for coexpression analysis have been reported in a number of bioinformatics articles and these algorithms have been validated by statistical analysis. However, large-scale analyses only provide clues that help in forming a hypothesis.. Hence, biologists who predict gene function by coexpression analysis should confirm the predicted function by performing wet lab experiments, regardless of the algorithm used.

In our studies, we identified candidate genes on the basis of coexpression relationships, and then selected some genes for further analysis from among the candidate genes on the basis of functional annotation. If a gene that is coexpressed with known GSL biosynthetic genes has a functional annotation, which is not expected on the basis of a priori knowledge of the GSL metabolic pathway, this gene may not be selected for further analysis since there may be a risk of false-positive results due to a coexpression relationship without any functional relationship. However, such a gene might be a novel, unexpected component of GSL metabolic pathway. I believe that new insights into a biological process can be provided by a non-targeted approach that is independent of a priori biological knowledge. An interesting study has recently been reported by Horan et al. (2008), in which 1,541 genes encoding proteins of unknown function were systematically associated with functional annotations of tightly coexpressed genes coding for proteins of known function. This type of genome-wide non-targeted approach will lead to the formation of a novel, data-driven hypothesis. In future, we should utilize large-scale biological methods for understanding a biological process completely, while taking into consideration the drawbacks of the methods (Aoki et al. 2007; Saito et al. 2008).

Acknowledgements I am grateful to the coauthors of my papers cited here. This research was supported in part by CREST of Japan Science and Technology Agency and Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science.

References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. Genome Res 13:693–702

Akiyama K, Chikayama E, Yuasa H, Shimada Y, Tohge T, Shinozaki K, Hirai MY, Sakurai T, Kikuchi J, Saito K (2008) PRIMe: a Web site that assembles tools for metabolomics and transcriptomics. In Silico Biol 8:0027

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol 48:381–390

Bechtold U, Murphy DJ, Mullineaux PM (2004) Arabidopsis peptide methionine sulfoxide reductase2 prevents cellular oxidative damage in long nights. Plant Cell 16:908–919

Beekwilder J, van Leeuwen W, van Dam NM, Bertossi M, Grandi V, Mizzi L, Soloviev M, Szabados L, Molthoff JW, Schipper B, Verbocht H, de Vos RC, Morandini P, Aarts MG, Bovy A (2008) The impact of the absence of aliphatic glucosinolates on insect herbivory in *Arabidopsis*. PLoS ONE 3:e2068

Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res 32:D575–D577

de Kraker JW, Luck K, Textor S, Tokuhisa JG, Gershenzon J (2007) Two Arabidopsis genes (IPMS1 and IPMS2) encode isopropylmalate synthase, the branchpoint step in the biosynthesis of leucine. Plant Physiol 143:970–986

Douglas Grubb C, Zipp BJ, Ludwig-Muller J, Masuno MN, Molinski TF, Abel S (2004) Arabidopsis glucosyltransferase UGT74B1 functions in glucosinolate biosynthesis and auxin homeostasis. Plant J 40:893–908

Field B, Cardon G, Traka M, Botterman J, Vancanneyt G, Mithen R (2004) Glucosinolate and amino acid biosynthesis in *Arabidopsis*. Plant Physiol 135:828–839

Gachon CM, Langlois-Meurinne M, Henry Y, Saindrenan P (2005) Transcriptional co-regulation of secondary metabolism enzymes in *Arabidopsis*: functional and evolutionary implications. Plant Mol Biol 58:229–245

Gigolashvili T, Berger B, Mock HP, Muller C, Weisshaar B, Flugge UI (2007a) The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in *Arabidopsis thaliana*. Plant J 50:886–901

Gigolashvili T, Engqvist M, Yatusevich R, Muller C, Flugge UI (2007b) HAG2/MYB76 and HAG3/MYB29 exert a specific and coordinated control on the regulation of aliphatic glucosinolate biosynthesis in *Arabidopsis thaliana*. New Phytol 177:627–642

Gigolashvili T, Yatusevich R, Berger B, Muller C, Flugge U-I (2007c) The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in *Arabidopsis thaliana*. Plant J 51:247–261

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K, Kiba T, Takatsuto S, Fujioka S, Asami T, Nakano T, Kato H, Mizuno T, Sakakibara H, Yamaguchi S, Nambara E, Kamiya Y, Takahashi H, Hirai MY, Sakurai T, Shinozaki K, Saito K, Yoshida S, Shimada Y (2008) The AtGenExpress hormone- and chemical-treatment data set: experimental design, data evaluation, model data analysis, and data access. Plant J 55:526–542. doi:10.1111/j.0960-7412.2008.03510.x

Goldovsky L, Cases I, Enright AJ, Ouzounis CA (2005) BioLayout(Java): versatile network visualisation of structural and functional relationships. Appl Bioinform 4:71–74

Hansen BG, Kliebenstein DJ, Halkier BA (2007) Identification of a flavin-monooxygenase as the *S*-oxygenating enzyme in aliphatic glucosinolate biosynthesis in *Arabidopsis*. Plant J 50:902–910

Hirai MY, Saito K (2008) Analysis of systemic sulphur metabolism in plants by using integrated "-omics" strategies. Mol Biosyst. doi:10.1039/B802911N

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 101: 10205–10210

Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki H, Sakurai N, Shibata D, Tokuhisa J, Reichelt M, Gershenzon J, Papenbrock J, Saito K (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. J Biol Chem 280:25590–25595

Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, Goda H, Nishizawa OI, Shibata D, Saito K (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci U S A 104:6478–6483

Hirai MY, Sawada Y, Araki R, Saito K (in press) Omics-based identification of the genes involved in glucosinolate biosynthesis. In: Sirko A et al (eds) Sulfur metabolism in higher plants. Backhuys Publishers, Leiden, The Netherland

Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, Zhu J-K, Cushman JC, Gollery M, Girke T (2008) Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol 147: 41–57

Jen CH, Manfield IW, Michalopoulos I, Pinney JW, Willats WG, Gilmartin PM, Westhead DR (2006) The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. Plant J 46:336–348

Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. Gene 276:89–99

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J 50: 347–363

Kim JH, Durrett TP, Last RL, Jander G (2004) Characterization of the Arabidopsis TU8 glucosinolate mutation, an allele of TERMINAL FLOWER2. Plant Mol Biol 54: 671–682

Klein M, Reichelt M, Gershenzon J, Papenbrock J (2006) The three desulfoglucosinolate sulfotransferase proteins in *Arabidopsis* have different substrate specificities and are differentially expressed. FEBS J 273:122–136

Knill T, Schuster J, Reichelt M, Gershenzon J, Binder S (2008) Arabidopsis branched-chain aminotransferase 3 functions in both amino acid and glucosinolate biosynthesis. Plant Physiol 146:1028–1039

Kroymann J, Textor S, Tokuhisa JG, Falk KL, Bartram S, Gershenzon J, Mitchell-Olds T (2001) A gene controlling variation in *Arabidopsis* glucosinolate composition is part of the methionine chain elongation pathway. Plant Physiol 127:1077–1088

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

Levy M, Wang Q, Kaspi R, Parrella MP, Abel S (2005) Arabidopsis IQD1, a novel calmodulin-binding nuclear protein, stimulates glucosinolate accumulation and plant defense. Plant J 43:79–96

Lisso J, Steinhauser D, Altmann T, Kopka J, Mussig C (2005) Identification of brassinosteroid-related genes by means of transcript co-response analyses. Nucleic Acids Res 33: 2685–2696

Malitsky S, Blum E, Less H, Venger I, Elbaz M, Morin S, Eshed Y, Aharoni A (2008). The proximal and distal circles of the transcriptome and metabolome affected by the two clades of Arabidopsis glucosinolate biosynthesis regulators. In: Abstract of 5th international conference on plant metabolomics, Yokohama, Japan, July 2008

Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. Nucleic Acids Res 34:W504–W509

Maruyama-Nakashita A, Nakamura Y, Tohge T, Saito K, Takahashi H (2006) Arabidopsis SLIM1 is a central transcriptional regulator of plant sulfur response and metabolism. Plant Cell 18:3235–3251

Mikkelsen MD, Naur P, Halkier BA (2004) Arabidopsis mutants in the C-S lyase of glucosinolate biosynthesis establish a critical role for indole-3-acetaldoxime in auxin homeostasis. Plant J 37:770–777

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in *Arabidopsis*. Nucleic Acids Res 35:D863–D869

Piotrowski M, Schemenewitz A, Lopukhina A, Muller A, Janowitz T, Weiler EW, Oecking C (2004) Desulfoglucosinolate sulfotransferases from *Arabidopsis thaliana* catalyze the final step in the biosynthesis of the glucosinolate core structure. J Biol Chem 279:50717–50725

Rautengarten C, Steinhauser D, Bussis D, Stintzi A, Schaller A, Kopka J, Altmann T (2005) Inferring hypotheses on functional relationships of genes: analysis of the *Arabidopsis thaliana* subtilase gene family. PLoS Comput Biol 1:e40

Saito K, Hirai MY, Yonekura-Sakakibara K (2008) Decoding genes with coexpression networks and metabolomics—'majority report by precogs'. Trends Plant Sci 13:36–43

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet 37:501–506

Schuster J, Knill T, Reichelt M, Gershenzon J, Binder S (2006) Branched-chain aminotransferase4 is part of the chain elongation pathway in the biosynthesis of methionine-derived glucosinolates in *Arabidopsis*. Plant Cell 18:2664–2679

Skirycz A, Reichelt M, Burow M, Birkemeyer C, Rolcik J, Kopka J, Zanor MI, Gershenzon J, Strnad M, Szopa J, Mueller-Roeber B, Witt I (2006) DOF transcription factor AtDof1.1 (OBP2) is part of a regulatory network controlling glucosinolate biosynthesis in *Arabidopsis*. Plant J 47:10–24

Sonderby IE, Hansen BG, Bjarnholt N, Ticconi C, Halkier BA, Kliebenstein DJ (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. PLoS ONE 2:e1322

Srinivasasainagendra V, Page GP, Mehta T, Coulibaly I, Loraine AE (2008) CressExpress: A tool for large-scale mining of expression data from *Arabidopsis*. Plant Physiol 147:1004–1016

Steinhauser D, Usadel B, Luedemann A, Thimm O, Kopka J (2004) CSB.DB: a comprehensive systems-biology database. Bioinformatics 20:3647–3651

Textor S, Bartram S, Kroymann J, Falk KL, Hick A, Pickett JA, Gershenzon J (2004) Biosynthesis of methionine-derived glucosinolates in *Arabidopsis thaliana*: recombinant expression and characterization of methylthioalkylmalate synthase, the condensing enzyme of the chain-elongation cycle. Planta 218:1026–1035

Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J-I, Awazuhara M, Inoue E, Takahashi H, Goodenowe DB, Kitayama M, Noji M, Yamazaki M, Saito K (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing a MYB transcription factor. Plant J 42:218–235

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J 43:153–163

Vanderauwera S, Zimmermann P, Rombauts S, Vandenabeele S, Langebartels C, Gruissem W, Inze D, Van Breusegem F (2005) Genome-wide analysis of hydrogen peroxide-regulated gene expression in *Arabidopsis* reveals a high light-induced transcriptional cluster involved in anthocyanin biosynthesis. Plant Physiol 139:806–821

Yonekura-Sakakibara K, Tohge T, Niida R, Saito K (2007) Identification of a flavonol 7-*O*-rhamnosyltransferase gene determining flavonoid pattern in *Arabidopsis* by transcriptome coexpression analysis and reverse genetics. J Biol Chem 282:14932–14941

Zhang Z, Ober JA, Kliebenstein DJ (2006) The gene controlling the quantitative trait locus EPITHIOSPECIFIER MODIFIER1 alters glucosinolate hydrolysis and insect resistance in *Arabidopsis*. Plant Cell 18:1524–1536

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol 136:2621–2632

Zimmermann P, Hennig L, Gruissem W (2005) Gene-expression analysis and network discovery using Genevestigator. Trends Plant Sci 10:407–409