Parallel analysis of transcript and metabolic profiles: a new approach in systems biology

Ewa Urbanczyk-Wochniak, Alexander Luedemann, Joachim Kopka, Joachim Selbig, Ute Roessner-Tunali, Lothar Willmitzer & Alisdair R. Fernie^{*}

Max-Planck-Institut für Molekulare Pflanzenphysiologie, Golm, Germany

The past few years in the medical and biological sciences have been characterized by the advent of systems biology. However, despite the well-known connectivity between the molecules described by transcriptomic, proteomic and metabolomic approaches, few studies have tried to correlate parameters across the various levels of systemic description. When comparing the discriminatory power of metabolic and RNA profiling to distinguish between different potato tuber systems, using the techniques described here suggests that metabolic profiling has a higher resolution than expression profiling. When applying pairwise transcript-metabolite correlation analyses, 571 of the 26,616 possible pairs showed significant correlation, most of which was novel and included several strong correlations to nutritionally important metabolites. We believe this approach to be of high potential value in the identification of candidate genes for modifying the metabolite content of biological systems.

EMBO reports 4, 989–993 (2003)

doi:10.1038/sj.embor.embor944

INTRODUCTION

Information flow in biological systems follows the sequence DNA to RNA to protein, with gene function generally described by the latter. Most current genomic research is carried out at the DNA and RNA levels; however, it is increasingly clear that a wide range of post-translational factors have functional importance in the cell. The current focus on the nucleic acid level is in part due to the fact that mature technologies are available that allow the sequencing (The Arabidopsis Genome Initiative, 2000; International Human Genome Sequencing Consortium, 2001) and analysis of the expression levels of complete genomes (or large proportions thereof) at the transcript level (Lockhart *et al.*, 1996), whereas strategies to allow similar comprehensive studies of the protein and metabolite complements of the cell are still at relatively early stages in their development. However, rapid progress has been made in proteomic (Shevchenko *et al.*, 1996)

and metabolomic (Fiehn *et al.*, 2000; Roessner *et al.*, 2001) analyses in recent years, and combined studies of the transcriptome and a subset of the proteome have been carried out in *Saccharomyces cerevisiae* (Futcher *et al.*, 1999; Gygi *et al.*, 1999; Ideker *et al.*, 2001a), facilitating the use of systems-biology approaches (Ideker *et al.*, 2001b; Kitano 2002; Oltvai & Barabasi, 2002).

Here, we describe the parallel analysis of potato tuber systems using a recently established platform for metabolic profiling based on gas chromatography–mass spectrometry (GC–MS) analysis in combination with the parallel analysis of gene-expression data using classical array technology. In carrying out these experiments, we set out to answer two questions: first, what is the relative power of the two phenotyping technologies to discriminate biological systems that either differ in developmental state or show well-characterized changes in response to the expression of transgenes? Second, can the combined analysis of transcript and metabolite profiling data present a new and meaningful approach in the identification of candidate genes for changing the metabolic composition of a biological system?

RESULTS AND DISCUSSION

When we designed these experiments, we decided to use a crop-plant system, the potato tuber system, to answer the above questions. The reasons for this were twofold: first, the potato tuber has well-defined but nevertheless highly related developmental stages and allows the assessment of several well-characterized transgenic situations; second, our group is experienced in the analysis of the potato tuber on both the biochemical and molecular levels (Fernie *et al.*, 2002). As a first step, the expression of 1,000 genes (represented by at least two expressed sequence tags (ESTs), representing many transcription factors and a wide variety of biosynthetic genes) was analysed across development and in two transgenic situations using a custom array. After analysing the data sets obtained, we selected 279 ESTs that gave highly reproducible results with respect to the consensus sequence from which they were derived. We next performed a principal-component analysis (PCA) to determine whether various developmental stages could be discriminated from one another and from the transcript profiles of potato tubers that ectopically express a more efficient pathway of sucrose mobilization (Roessner *et al.*, 2001). As shown in Fig. 1A, some of the developmental stages can be readily distinguished from each other on the

^{*}Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

^{*}Corresponding author. Tel: + 49 331 567 8221; Fax: + 49 331 567 8408; E-mail: fernie@mpimp-golm.mpg.de

Received 6 May 2003; revised 9 July 2003; accepted 20 August 2003
Published online 12 September 2003

Fig. 1 | Principal-component analysis of gene-expression and metabolite profiles of genetically and temporally distinct systems. (A) Gene-expression profiles. (B) Metabolite profiles. The distances between these populations were calculated as described in Roessner *et al.* (2001). The percentage of variance explained by each component is shown in parentheses. The transgenic lines INV2-30 (INV) and SP 29 (SP) are represented by black circles, and wild-type samples harvested 8, 9, 10, 13 and 14 weeks after transfer to the greenhouse are represented by coloured circles. Each data point represents an independent sample.

basis of their transcript profiles. Most notable in this respect is the clear differentiation of tubers harvested ten weeks after transfer to the greenhouse from the other developmental stages. The exact reason for this discrimination is unclear. Examination of the upregulated

clones suggests a higher metabolic activity at this stage; however, this finding is not consistent with changes in the metabolite levels documented in identical samples (see below). Despite the fact that these results are surprising, they are consistent for all three replicates harvested at this time point.

We also obtained the surprising result that the transcript levels of the transgenic systems could not be discriminated either from each other or from the corresponding wild-type tuber samples. It is important to note here that a similar situation was seen after PCA of the entire transcript data set (2,200 ESTs). Although it is clear that the tuber samples harvested ten weeks after transfer to the greenhouse were markedly different from those harvested at other developmental stages, the fact remains that transcriptional variation during development is greater than that after a relatively severe genetic perturbation of primary metabolism (Roessner *et al.*, 2001).

Metabolic profiling was then carried out on identical samples to determine the levels of the major metabolites of primary metabolism, including sugars, sugar alcohols, organic acids, amino acids, as well as the nutritionally important ascorbate and tocopherol. When a PCA was carried out on the data set obtained from the metabolic profiling studies (Fig. 1B), a different situation was observed from that seen on analysis of the transcript data. In this case, the transgenic events clustered completely independently (both from the wild-type control and from each other); furthermore, on the basis of their metabolite complement, samples taken after ten weeks of growth were similar to those taken at other time points. As with the results presented for transcript levels, we believe that these data have important ramifications for the potential risks associated with transgenic organisms and theories of substantial equivalence (Kuiper *et al.*, 2001; Trewavas & Leaver, 1999).

As stated above, this study was started with two main objectives. First, we wanted to compare the discriminatory power of both profiling approaches as described above. Second, we were interested in using the combined evaluation of both analyses as a new approach in experimental systems biology (for a definition, see Sweetlove *et al.*, 2003). As a first step, we decided to run all data points through pairwise correlation analysis, determining for each transcript whether it is correlated to any of the metabolites. Of the 26,616 pairs analysed, 363 positive and 208 negative correlations were seen, the total number of 571 correlations being well above the number of correlations that might be expected by chance (266 at $p < 0.01$). Several representative examples of transcript-metabolite correlations are shown in Fig. 2 and are discussed in detail below, and the entire list is available in the supplementary information online.

First, as with any new approach, it is important to determine whether the data obtained are in agreement with observations made using different experimental strategies. This is clearly the case, with a

Fig. 2 | Correlation between metabolite and transcript levels of the analysed systems. All correlations between the relative levels of chemically defined metabolites and those of each selected clone were assessed. All correlations at $p = 0.01$ when assessed by Spearman's rank-order correlation are given in the supplementary information online. Several representative examples are given here (all are plotted using the logarithmic scale). r_s -values are given in parentheses. (A) Sucrose transporter versus sucrose ($r_s = -0.52$). (B) glutamate decarboxylase versus 4-aminobutric acid ($r_s = 0.55$). (C) Tryptophan synthase β -chain 1 versus tryptophan ($r_s = 0.61$). (D) Tryptophan synthase β -chain 1 versus tyrosine ($r_s = 0.65$). (E) Ornithine carbamoyltransferase versus serine ($r_s = 0.53$). (F) Ornithine carbamoyltransferase versus cysteine ($r_s = 0.54$). (G) Aminotransferase-like protein versus fructose-6-phosphate ($r_s = 0.85$). (H) Aminotransferase-like protein versus glucose-6-phosphate ($r_s = 0.85$). (I) Glutamate decarboxylase versus spermidine ($r_s = 0.64$). (J) Glutamate decarboxylase versus tyrosine ($r_s = 0.63$). (K) CONSTANS-like protein versus ascorbate ($r_s = -0.72$). (L) Succinyl-co-enzyme-A synthetase α -subunit versus tocopherol ($r_s = -0.64$). (M) Transcription factor WRKY6 versus lysine ($r_s = 0.51$). (N) S-adenosyl-L-methionine synthetase versus lysine ($r_s = 0.6$). (O) Ornithine carboxyltransferase versus lysine ($r_s = 0.54$). (P) Caffeoyl-co-enzyme-A O-methyltransferase versus lysine ($r_s = -0.52$).

strong negative correlation between sucrose and sucrose transporter expression (Fig. 2A), and a strong positive correlation between 4-aminobutyric acid and glutamate decarboxylase isoform I (Fig. 2B). As both of these relationships have been reported previously (Vaughn *et al.*, 2002; Facchini *et al.*, 2000), these examples provide confirmation of the validity of our approach. Second, many further correlations seem to have a functional basis that can be explained retrospectively. The positive correlations of both tryptophan and tyrosine with the β 2-chain of tryptophan synthase (Fig. 2C,D), and of ornithine carbamoyltransferase with serine and cysteine (Fig. 2E,F), are two such examples. Third, although several of the correlations, such as those described above, were predictable, most of the correlations obtained using this approach were novel and not directly related to the biochemical pathway in which the gene products participate. Although it is clear that many of these correlations are due to chance, the fact that several correlations occurred between transcripts and metabolites from the same or related pathways might strengthen the interpretation of the linkages. An example of this includes aminotransferase, which correlates with both fructose-6-phosphate and glucose-6-phosphate (Fig. 2G,H). Such comparisons may offer hints as to the functions of the genes involved (the elucidation of which is a subject that requires a huge amount of research effort). It is also interesting to note that several transcripts correlate with more than one metabolite, as highlighted above for the aminotransferase. Other examples of this include glutamate decarboxylase isoform I, which correlates with both spermidine (Fig. 2I) and tyrosine (Fig. 2J), and the same is also true for 4-aminobutyric acid and tryptophan, and various transcription factors. Finally, it is exciting to see that nutritionally important metabolites, such as ascorbate, tocopherol and lysine, were closely correlated to the expression levels of various genes or transcription factors: ascorbate was negatively correlated with a homologue of the clock gene *CONSTANS* (Fig. 2K); tocopherol was negatively correlated with succinyl-co-enzyme-A synthetase (Fig. 2L); lysine was positively regulated by the transcription factor WRKY6 (Fig. 2M), S-adenosyl-L-methionine synthetase (Fig. 2N) and ornithine carbamoyltransferase (Fig. 2O), and was negatively regulated by caffeoyl-co-enzyme A *O*-methyltransferase (Fig. 2P). These unexpected results only reveal correlated parameters and do not allow the identification of causality between the transcript and metabolite. Although many possible explanations for such metabolite linkages can be postulated, including unknown metabolic links between pathways (such as, for example, the sharing of a common cofactor; or, in the case of transcription factors, interactions with previously undescribed regulatory genes), further, more direct experimentation is required to elucidate the mechanisms that underlie the correlations presented here. Nevertheless, we believe these essentially unexpected correlations to be of great potential use for biotechnological applications in which the goal is the modification of metabolite compositions through genetic means. Although these data do not, in any way, prove causal or even mechanistic links between molecular entities, the approach of linking transcript and metabolite data through pairwise correlation analysis provides a powerful tool for the rapid identification of candidate genes.

In conclusion, the discriminatory powers of transcript and metabolite profiling approaches are different, with metabolite profiling allowing greater resolution of the different systems studied here. Whether this reflects the fact that changes at the transcript level are less pronounced as compared with changes at the metabolite level or merely highlights limitations of the profiling technologies used,

remains an open question. It is possible that this result merely reflects low sensitivity of ESTs as probes and that profiling using full-length complementary DNAs (Seki *et al.*, 2001), or oligonucleotide-specific probes would allow greater discrimination. However, whatever the reason, these results suggest that the discrimination of biological systems should be performed at more than one level, something that is of particular importance in the establishment of substantial equivalence between transgenic and conventional crops. The second main conclusion of our work is that although the number of strong metabolite-transcript correlations was relatively small, as would be expected from previous experimental work aimed at comparing transcript and protein levels (Futcher *et al.*, 1999; Gygi *et al.*, 1999; Ideker *et al.*, 2001) and mathematical studies (ter Kuile & Westerhoff, 2001), we believe they allow the generation of clear, testable hypotheses. Of particular interest in our data set was the correlation of genes with the essential amino acid lysine and with the vitamins ascorbate and tocopherol, as these linkages define candidate genes for the manipulation of the content of these nutritionally important compounds in plants. To our knowledge, only the combination of the two platforms presented here is able to provide such novel hypotheses; therefore, the pairwise correlation between transcripts and metabolites is a potentially powerful new tool for hypothesis creation in systems biology.

METHODS

Materials. *Solanum tuberosum* L. cv. Desiree was obtained from Saatzeit Lange AG. The generation of the transgenic plants used in this study has been described previously (Sonnewald *et al.*, 1997; Trethewey *et al.*, 2001). Plants were handled as described in the literature (Tauberger *et al.*, 2000; Regierer *et al.*, 2002) and harvested at time points described in the figure legends.

Gene-expression analysis. Microarrays were constructed on nylon filters as described previously (Thimm *et al.*, 2001; Colebatch *et al.*, 2002). More than 2,200 tomato clones were collected from cDNAs constructed in the laboratories of S. Tanksley, G. Martin and J. Giovannoni and provided by The Institute for Genomic Research. These ESTs correspond to approximately 1,000 tomato genes, which are highly homologous to those from potato (Fulton *et al.*, 2002). Microarray analysis was carried out as described previously (Thimm *et al.*, 2001; Colebatch *et al.*, 2002), and full details are given at <http://www.mpimp-golm.mpg.de/willmitzer/zusaetzlich/index-e.html>.

Processing of array data. The 'raw' data sets obtained from microarray analysis were submitted to our in-house database (<http://www.mpimp-golm.mpg.de/haruspex/index-e.html>). In this database, the raw data were annotated, rearranged and normalized, as described previously (Thimm *et al.*, 2001; Colebatch *et al.*, 2002). The normalized data were then subjected to Grubbs' test (Grubbs, 1969; Stefansky, 1972) to detect outliers in the data set.

Metabolite analysis. The preparation and derivation of samples for metabolite analysis, the operation of the GC-MS, the evaluation of the resultant chromatograms and the normalization of the results was carried out exactly as described in Roessner *et al.* (2001).

Pre-selection of relevant genes. As there are no EST/gene libraries publically available for crop-plant systems, including the potato tuber system, the relationships between ESTs and potential genes were checked. Relevant genes for a meaningful co-response analysis were required to show significant changes in gene expression under the experimental conditions analysed, and ESTs representing the

same gene had to show the same hybridization behaviour. We applied non-parametric Spearman's rank-order correlation and parametric Pearson's linear-correlation analyses to those EST pairs representing the same gene. Those genes showing positive correlation at significance threshold $p = 0.001$ in both analyses were used for further analysis.

Discriminant analysis. We tested several data transformations, distance measures and clustering methods on our data set before choosing PCA. PCA was performed independently on the data sets obtained from transcript and metabolite profiling with the software package S-Plus 2000 (Insightful) using the default weighted covariance-estimation function. PCA is a dimensionality-reduction method that attempts to compress the original features into the smallest possible number of components. This is necessary because in analysing transcript and/or metabolite profiles, dimensionality is often a problem: in most cases, the number of features (genes and/or metabolites) exceeds the number of samples considerably. Each principal component is associated with an *eigenvalue*, which corresponds to the amount of variability explained by the corresponding principal component. PCAs are presented as three-dimensional plots that comprise 86.9% and 79.2% of the metabolite and transcript variances, respectively.

The data was $\log_{10}$ -transformed before further analysis, and metabolites or ESTs that were not common to all samples were removed. This transformation reduces the influence of rare high-measurement values, but does not change the order of the data and, therefore, does not influence non-parametric analysis methods such as the statistical tests described below.

Co-response analysis. Correlated expression of the set of pre-selected as well as non-selected genes and correlation of changes in metabolite and messenger RNA levels were detected by non-parametric Spearman's rank-order correlation analysis, with the significance threshold set at $p = 0.01$. Rank-order correlation was chosen because mRNA and metabolite levels can be correlated in a nonlinear manner. Non-parametric methods for correlation analysis are also appropriate in our case, because from relatively few data points it cannot be estimated whether they follow a normal distribution, an assumption required for application parametric methods such as Pearson's linear-correlation analysis. Spearman's rank-order method was preferred to Kendall test statistics because the latter requires more data points.

Supplementary information is available at *EMBO reports* online (<http://www.emboreports.org>).

ACKNOWLEDGEMENTS

We thank K. Koehl for excellent supervision of the plants and T. Altmann, S. Fischer and S. Kloska for consultancy on array preparation and analysis. E.U.-W. thanks the Max Planck Gesellschaft for a fellowship.

REFERENCES

- Colebatch, G. *et al.* (2002) Novel aspects of symbiotic nitrogen fixation uncovered by transcript profiling with cDNA arrays. *Mol. Plant Microbe Interact.*, **15**, 411–420.
- Facchini, P.J., Huber-Allanach, K.L. & Tari, L.W. (2000) Plant aromatic L-amino acid decarboxylases: evolution, biochemistry, regulation, and metabolic engineering applications. *Phytochemistry*, **54**, 121–138.
- Fernie, A.R., Willmitzer, L. & Trethewey, R.N. (2002) Sucrose to starch: a transition in molecular plant physiology. *Trends Plant Sci.*, **7**, 35–42.
- Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. & Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnol.*, **18**, 1157–1161.
- Fulton, T.M., Van der Hoeven, R., Eannetta, N.T. & Tanksley, S.D. (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell*, **14**, 1457–1467.
- Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S. & Garrells, J.I. (1999) A sampling of the yeast proteome. *Mol. Cell Biol.*, **19**, 7357–7368.
- Grubbs, F. (1969) Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1–21.
- Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.*, **19**, 1720–1730.
- Ideker, T. *et al.* (2001a) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.
- Ideker, T., Galitski, T. & Hood, L. (2001b) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, **2**, 343–372.
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Kitano, H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.
- Kuiper, H.A., Kleter, G.A., Noteborn, H.P.J.M. & Kok, E.J. (2001) Assessment of the food safety issues related to genetically modified foods. *Plant J.*, **27**, 503–528.
- Lockhart, D.J. *et al.* (1996) Expression monitoring by hybridisation to high-density oligonucleotide arrays. *Nature Biotechnol.*, **14**, 1675–1680.
- Oltvai, Z.N. & Barabasi, A.L. (2002) Life's complexity pyramid. *Science*, **295**, 1662–1664.
- Regierer, B. *et al.* (2002) Starch content and yield increase as a result of altering adenylate pools in transgenic plants. *Nature Biotechnol.*, **20**, 1256–1260.
- Roessner, U. *et al.* (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11–29.
- Seki, M. *et al.* (2001) *Arabidopsis* encyclopedia using full-length cDNAs and its application. *Plant Physiol. Biochem.*, **39**, 211–220.
- Shevchenko, A. *et al.* (1996) Linking genome and proteome by mass spectrometry: large scale identification of yeast proteins from two-dimensional gels. *Proc. Natl Acad. Sci. USA*, **93**, 14440–14445.
- Sonnenwald, U., Hajirezaei, M.R., Kossmann, J., Heyer, A., Trethewey, R.N. & Willmitzer, L. (1997) Increased potato tuber size resulting from apoplastic expression of a yeast invertase. *Nature Biotechnol.*, **15**, 794–797.
- Stefansky, W. (1972) Rejecting outliers in factorial designs. *Technometrics*, **14**, 469–479.
- Sweetlove, L.J., Last, R.L. & Fernie, A.R. (2003) Predictive metabolic engineering: a goal for systems biology. *Plant Physiol.*, **132**, 420–425.
- Tauberger, E. *et al.* (2000) Antisense inhibition of plastidial phosphoglucosyltransferase provides compelling evidence that potato tuber amyloplasts import carbon from the cytosol in the form of glucose-6-phosphate. *Plant J.*, **23**, 43–53.
- ter Kuile, B.H. & Westerhoff, H.V. (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.*, **500**, 169–171.
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Thimm, O., Essigmann, B., Kloska, S., Altmann, T. & Buckhout, T.J. (2001) Response of *Arabidopsis* to iron deficiency stress as revealed by microarray analysis. *Plant Physiol.*, **127**, 1030–1043.
- Trethewey, R.N. *et al.* (2001) Expression of a bacterial sucrose phosphorylase in potato tubers results in a glucose-independent induction of glycolysis. *Plant Cell Environ.*, **24**, 357–365.
- Trewavas, A. & Leaver, C.J. (1999) Conventional crops are the test of GM prejudice. *Nature*, **401**, 640.
- Vaughn, M.W., Harrington, G.N. & Bush, D.R. (2002) Sucrose-mediated transcriptional regulation of sucrose symporter activity in the phloem. *Proc. Natl Acad. Sci. USA*, **99**, 10876–10880.