Plant Biotechnology 23, 199–202 (2006)

Gene Note

Non-biased distribution of tomato genes with no counterparts in *Arabidopsis thaliana* in expression patterns during fruit maturation

Kentaro Yano[1], Taneaki Tsugane[2], Manabu Watanabe[2], Fumi Maeda[2], Koh Aoki[1], Daisuke Shibata[1]*

[1] Kazusa DNA Research Institute, Kazusa-Kamatari 2-6-7, Kisarazu, Chiba 292-0818, Japan; [2] Chiba Prefectural Agriculture Research Center, Daizenno-Cho 808, Midori-ku, Chiba, Chiba 266-0006, Japan
* E-mail: shibata@kazusa.or.jp Tel: +81-438-52-3947 Fax: +81-438-52-3948

Received October 16, 2005; accepted October 27, 2005 (Edited by T. Hashimoto)

Abstract We prepared DNA array nylon filters using 10,905 cDNA clones selected from fruit and leaf cDNA libraries of the miniature tomato (*Solanum lycopersicum*) cultivar, Micro-Tom, as being representative of non-redundant sequences of 37,972 Micro-Tom expressed sequence tags (ESTs). Tomato gene expression during fruit maturation was analyzed using the array filters. Graphs of gene expression patterns were arranged into a 4×4 array using a self-organizing map algorithm. Using non-redundant sequences generated from 188,024 tomato ESTs, we assigned the cDNA clones on the array filters to 1151 genes that had no counterparts in the *Arabidopsis thaliana* genome using stringent conditions for e-values of BLAST searching. We found that the expression patterns of these non-*Arabidopsis* genes were evenly distributed in the self-organizing map, with no statistically significant difference with the distributions of whole genes. These findings suggested that the non-*Arabidopsis* genes participate in a wide variety of diverse functions during fruit maturation.

Key words: *Arabidopsis*, custom DNA array, gene expression analysis, Micro-Tom, *Solanum lycopersicum.*

Accumulation of genomic information from a number of plant species has facilitated searches for genes that have no sequence similarity with any *Arabidopsis thaliana* genes. Comparative sequence analyses between different species including rice, legume, spruce, pine, and Solanaceae species have demonstrated that 25 to 38% of the genes in each species exhibit no sequence similarity to *Arabidopsis* genes (Van der Hoeven et al. 2002; The Rice Full-Length cDNA Consortium 2003; Rensink et al. 2005; Pavy et al. 2005a, 2005b). The functions of these genes, hereafter referred to as non-*Arabidopsis* genes, remain to be elucidated. While analyses of gene expression have facilitated functional comparisons of genes across species (Ogihara et al. 2003; Fei et al. 2004), comprehensive expression analysis of the non-*Arabidopsis* gene set, to our knowledge, has not yet been undertaken.

In this study, we prepared a set of DNA array filters spotted with tomato cDNA clones and examined gene expression during fruit maturation, focusing specifically on the non-*Arabidopsis* gene set. DNA macroarray filters were prepared using cDNA clones selected from fruit and leaf cDNA libraries of the miniature tomato cultivar, Micro-Tom. Of the 37,972 Micro-Tom EST sequences, 10,905 non-redundant sequence groups composed of 5307 singlets and 5598 contigs were generated using the PHRAP program (http://www.phrap.org). The clones that had the 5′-ends of the consensus sequences in each group were selected for array spotting. Inserts of 10,905 cDNA clones were amplified by PCR in 20 μl of the standard reaction mixture using γ Taq polymerase (Takara Bio Inc., Ohtsu, Japan) and T7 (5′- GTAATACGACTCACTATAGGG-3′) and T3 (5′- AATTAACCCTCACTAAAGGG-3′) primers. The reaction solutions were then dried under vacuum before being dissolved in 10 μl of 80% (v/v) formamide, 0.05% (w/v) xylene cyanol and 20 mM EDTA. Spotting of the DNA solutions on 8×12 cm nylon filters (Biodyne Plus, PALL Inc., Ann Arbor, MI, U.S.A.) at a density of 100 spots cm^{-2} was done using a MicroGrid II (Genomic Solutions Inc., Ann Arbor, MI, U.S.A.). Approximately 10 nl of the solutions was placed on single spots. The filters were treated with 0.2 M NaOH and 1.5 M NaCl for 2 min for denaturing DNA before being neutralized with 0.2 M Tris-HCl buffer (pH 7.0). The DNA molecules were fixed on the filter by UV cross-linking (120,000 Joules cm^{-2}). As a negative control, 96 spots of lambda-DNA were distributed evenly on the filters.

Abbreviations: CDS, coding sequence; EST, expressed sequence tag; SOM, self-organizing map.
This article can be found at http://www.jspcmb.jp/

We prepared total RNA from four developmental stages of Micro-Tom fruit. Micro-Toms were grown at 25 °C with a 12 h light/12 h dark photocycle under fluorescent light. The pericarps of fruit were collected at the immature green stage, the mature green stage, the light red stage, and the red ripe stage, which were frozen immediately in liquid nitrogen. The tissues were ground to powder in liquid nitrogen using a mortar and pestle. The ground tissue was mixed with RNA extraction buffer (4.2 M guanidine thiocyanate, 17 mM Sarcosyl, 25 mM trisodium citrate, and 0.1% (v/v) Antifoam), and incubated with acid phenol at 80°C for 10 min. After phenol/chloroform extraction, total RNA was precipitated by 2 M LiCl.

Hybridization of the DNA array filters with labeled cDNA targets prepared from total RNA was carried out as described previously (Ishihara et al. 2004) with minor modifications. Total RNA (5 μg) was suspended in 8.8 μl of deionized distilled water and mixed with 1 μl of 0.5 μg μl^{-1} oligo(dT)$_{12-18}$. Following a heat denaturing step at 65°C for 5 min, 2.5 μl of 10×cDNA synthesis buffer (Invitrogen, Carlsbad, CA, U.S.A.), 2.5 μl of 0.1 M DTT, 0.7 μl of dNTP mixture (20 mM dGTP, 20 mM dATP, 20 mM dTTP and 0.125 mM dCTP each), 2.5 μl of [α^{33}P]dCTP (Amersham Biosciences, Buckinghamshire, U.K.), 1 μl of 40 U μl^{-1} RNaseOUT (Invitrogen) and 1 μl of 50 U μl^{-1} SuperScript II (Invitrogen) were added. The mixture was incubated at 42°C for 50 min. After terminating the reaction by heating at 70°C for 15 min, 2 U of RNase H was added, and the solution was incubated at 37°C for 20 min. The synthesized cDNA was purified using QIAquick PCR purification kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions. One set of DNA array filters (two filters) was prehybridized with 10 ml of 0.5 M Church phosphate buffer (Church and Gilbert 1984) containing 1 mM EDTA, 7% (w/v) SDS and 10 μg of oligo(dA)$_{18}$ in a hybridization bag at 65°C for 3 h. Heat-denatured ^{33}P-labeled cDNA was mixed with 1 ml of the Church phosphate buffer, and then added to the hybridization bag. After incubation at 65°C for 16 h, filters were washed once with 1×SSC containing 0.1% (w/v) SDS at 65°C for 1 min, once with 1×SSC containing 0.1% (w/v) SDS at 65°C for 15 min, and twice with 0.1×SSC containing 0.1% (w/v) SDS at 65°C for 15 min. The filters were wrapped in plastic film and exposed to an IP image plate (Fuji Photo Film, Tokyo, Japan) for 24 h. Signals on the IP image plates were scanned using STORM 830 analyzer (Molecular Dynamics, Sunnyvale, CA, U.S.A.) and quantified using Array Vision 5.1 software (Imaging Research Inc., Ontario, Canada). The median value of signal intensities in each nylon filter was calculated. Normalized signal intensity in nylon filters was calculated by dividing the signal intensity by the median value. For each gene, the average of the normalized signal intensities across three replications was taken as the expression value. For the 96 lambda probes on each filter, the sum of the mean and two-fold standard deviation of the expression values were used to define the threshold value. The expression values of the genes greater than the highest threshold value of the threshold values obtained from these three replicates were employed for further analysis since such genes were regarded as being genes that are actually expressed.

To analyze the expression of the non-*Arabidopsis* gene set during fruit maturation by DNA array hybridization, we identified the non-*Arabidopsis* gene set in the list of the genes spotted on the DNA array filters. The 10,905 cDNA inserts spotted on the macroarray were assigned to 10,176 non-redundant sequences assembled from tomato 186,405 ESTs and 1619 Micro-Tom ESTs that had not been deposited in public databases. Using these non-redundant sequences, the non-*Arabidopsis* gene set was identified as follows: (1) To eliminate short and uncertain sequences, the 10,176 non-redundant sequences were searched against the TIGR Tomato Gene Index database (Lee et al. 2005). Using parameters for the BLASTN program (Altschul et al. 1990) of e-value< 1e-30, identity≥98%, score≥300 and alignment length≥300 bp, we obtained 6432 hit sequences. (2) Of these 6432 hits, we searched for the sequences that shared no sequence similarity with *Arabidopsis* genes. This search was performed using the BLASTN and BLASTX programs, with an e-value threshold<1e-30 against the TAIR *Arabidopsis* DNA and protein databases that include coding sequences (CDS), cDNA and genomic sequences (Rhee et al. 2003). As a consequence, we obtained 1186 sequences that did not correspond with any *Arabidopsis* sequences. (3) We then obtained a set of tomato protein sequences from NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein) and UniProt (Bairoch et al. 2005) which consisted of 3362 tomato protein sequences. We then produced another set of tomato protein sequences that exhibited sequence similarity to *Arabidopsis* protein sequences by searching for sequence similarity between the 3362 tomato protein sequences and *Arabidopsis* proteins (BLASTP with e-value<1e-30). This search yielded a set of 2672 tomato protein sequences that exhibited sequence similarity with *Arabidopsis* proteins. (4) Finally, the 1186 resulting non-redundant sequences from the macroarray were compared with the 2672 tomato proteins having sequence similarity to *Arabidopsis* sequences (BLASTX with e-value<1e-30). We obtained 1151 tomato genes that had no sequence similarity with *Arabidopsis* genes. The 1151 consensus sequences of non-*Arabidopsis* genes corresponded to 1176 probes on the DNA array filters due to clone redundancy on the filters.

To test the stringency of the sequence similarity search on the selection of the non-*Arabidopsis* gene set, we used a different cut-off setting for the e-value threshold. Under more stringent conditions using cut-off e-values<1e-10, 431 non-redundant sequences were identified as being non-*Arabidopsis* genes. These 431 non-redundant sequences corresponded to 440 probes on the DNA array filters.

In the 431 non-*Arabidopsis* gene set, 237 genes did not have any functional annotations. Of 194 genes with annotations, 18 were assigned to the category of transcription factors and 14 were assigned to cell wall-related proteins. However, more than 40% of the 194 genes had annotations such as "expressed protein" or "unknown protein" in databases.

The gene expression during fruit maturation was analyzed using the macroarray. Using the self-organizing map (SOM) program shipped with the GeneSpring DNA array analysis package (Agilent Inc., Foster City, CA, U.S.A.), the patterns of gene expression were pursed and grouped into classes arranged in a 4×4 array (Figure 1A). This arrangement allowed us to have an image of gene expression for the whole genes. For example, the genes responsible for carotenoid biosynthesis, phytoene synthase 1, phytoene desaturase, zeta-carotene desaturase and carotenoid isomerase, could be presented in a few classes, indicating the coordinated expression of the carotenoid biosynthesis genes during fruit maturation (data not shown).

We examined the distributions of the non-redundant *Arabidopsis* gene sets on the SOM (Figure 1B). The distribution patterns of the genes in both the 1151 and 431 non-*Arabidopsis* gene sets appeared to be comparable to that obtained using all of the probes, except for slight differences that appeared in classes (1,1), (3,1), (2,4), and (4,4) (Figure 1C). To determine whether these differences were statistically significant, we performed a χ^2 test to assess goodness of fit at the 1% level. The results showed that there were no statistically significant differences between the patterns obtained for all probes and non-*Arabidopsis* gene sets. We therefore concluded that there was no statistically biased distribution of the non-*Arabidopsis* gene set with respect to the gene expression patterns during fruit ripening.

Transcriptome analyses of tomato fruit during maturation have been reported recently (Alba et al. 2004; Alba et al. 2005). These studies have provided detailed descriptions of the expression profiles of the genes related to physiological processes such as photosynthesis (Alba et al. 2004), and ethylene and carotenoid biosynthesis (Alba et al. 2005). The expression profiles of carotenoid biosynthetic genes were relatively similar to those observed in Micro-Tom fruit. In addition to the enzymes, the dynamic characteristics associated with the expression of numerous transcription and signal

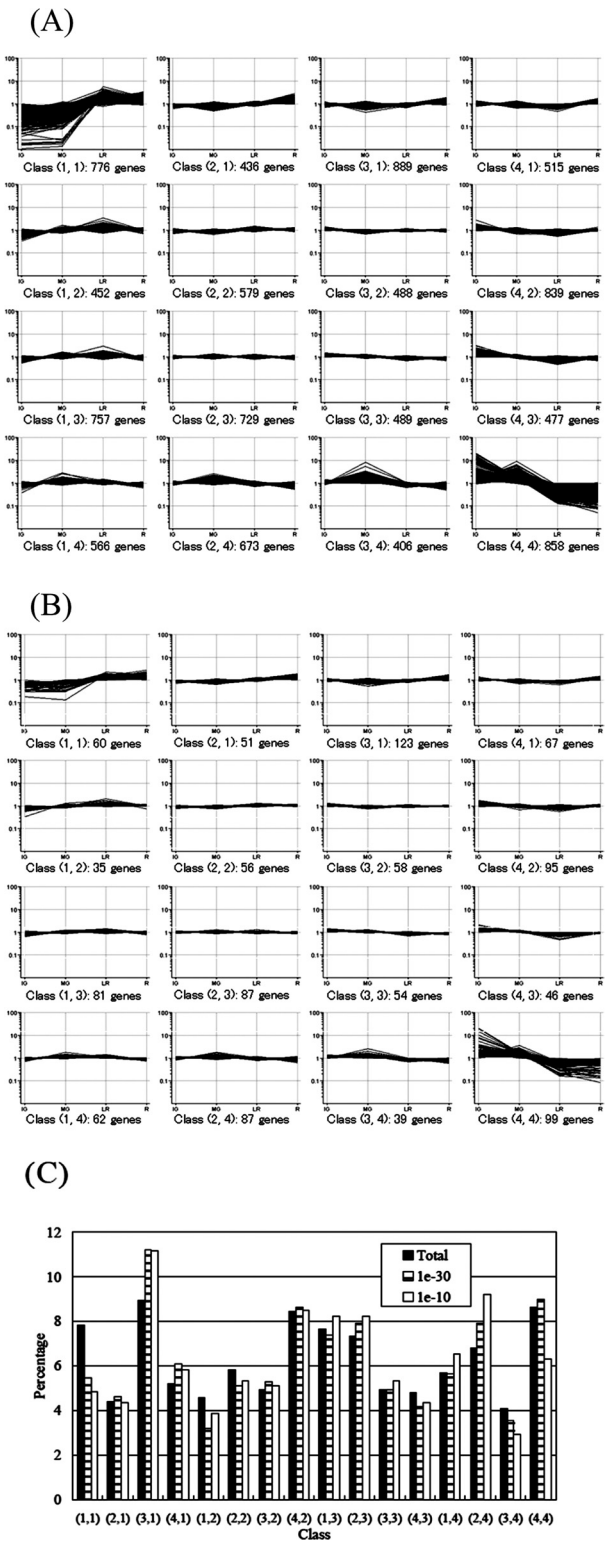

Figure 1. Self-organizing map analysis for gene expression during fruit maturation. Only gene patterns with significant expression levels were represented. (A) Expression patterns of 9929 non-redundant sequences at four stages of fruit maturation arranged in a 4×4 matrix. (B) Expression patterns of the 1100 non-*Arabidopsis* genes that were selected with cut-off value of e-values<1e-30 were extracted from (A). (C) Comparison of the percentage of genes in each class in the matrix. Total: 9929 non-redundant sequences, 1e-30: 1100 non-*Arabidopsis* genes, 1e-10: 431 non-*Arabidopsis* genes.

transduction factors during fruit ripening implied the presence of multiple regulatory points during fruit development and ripening, clearly demonstrating the utility of expression analyses in elucidating gene functions.

A similar approach was adopted to gain new insights into the functions of the non-*Arabidopsis* gene set in the Micro-Tom cultivar. Interestingly, results of expression analysis have suggested that the functions of the non-*Arabidopsis* gene set are distributed among various processes associated with fruit maturation. Detailed analysis of expression patterns for each individual gene may therefore be required for assigning functions to respective non-*Arabidopsis* genes. Graham et al. (2004) have searched for conserved motifs of legume-specific genes to predict their functions. They have successfully identified novel gene families including F-box proteins, proline-rich proteins, and cysteine-cluster proteins. This alternate approach might facilitate the assignment of functions to the non-*Arabidopsis* gene set of the tomato. Analyses of these non-*Arabidopsis* genes could therefore serve as a basis for understanding the genetic diversity of the tomato.

Acknowledgements

We wish to thank Dr. Nozomu Sakurai (Kazusa DNA Research Institute) for critical discussions. This work was supported by a grant from the Kazusa DNA Research Institute Foundation and a grant from the Japanese Ministry of Agriculture, Forestry and Fisheries for research projects utilizing advanced technologies in agriculture, forestry and fisheries.

References

Alba R, Fei Z, Payton P, Liu Y, Moore SL Debbie P, Cohn J, D'Ascenzo M, Gordon JS, Rose JK, Martin G, Tanksley SD Bouzayen M, Jahn MM Giovannoni J (2004) ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *Plant J* 39: 697–714.

Alba R, Payton P, Fei Z, McQuinn R, Debbie P, Martin GB, Tanksley SD, Giovannoni JJ (2005) Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell* 17: 2954–2965

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2005) The Universal Protein Resource (UniProt). *Nucl Acids Res* 33: D154–D159

Church GM, Gilbert W (1984) Genomic sequencing. *Proc Natl Acad Sci USA* 81: 1991–1995

Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannoni JJ (2004) Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. *Plant J* 40: 47–59

Graham MA, Silverstein KA, Cannon SB, VandenBosch KA (2004) Computational identification and characterization of novel genes from legumes. *Plant Physiol* 135: 1179–1197

Van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S (2002) Deduction about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14: 1441–1456

Ishihara T, Sakurai N, Sekine K, Hase S, Ikegami M, Shibata D, Takahashi H (2004) Comparative analysis of expressed sequence tags in resistant and susceptible ecotypes of *Arabidopsis thaliana* infected with cucumber mosaic virus. *Plant Cell Physiol* 45: 470–480

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucl Acids Res* 33: D71–D74

Ogihara Y, Mochida K, Nemoto Y, Murai K, Yamazaki Y, Shin-IT, Kohara Y (2003) Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags. *Plant J* 33: 1001–1011

Pavy N, Laroche J, Bousquet J, Mackay J (2005a) Large-scale statistical analysis of secondary xylem ESTs in pine. *Plant Mol Biol* 57: 203–224

Pavy N, Paule C, Parsons L, Crow JA, Morency MJ, Cooke J, Johnson JE, Noumen E, Guillet-Claude C, Butterfield Y, Barber S, Yang G, Liu J, Stott J, Kirkpatrick R, Siddiqui A, Holt R, Marra M, Seguin A, Retzel E, Bousquet J, Mackay J (2005b) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. *BMC Genomics* 6: 144 (Epub ahead)

Rensink WA, Lee Y, Liu J, Iobst S, Ouyang S, Buell CR (2005) Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. *BMC Genomics* 6: 124–137

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucl Acids Res* 31: 224–228

The Rice Full-Length cDNA Consortium (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301: 376–379