Comprehensive EST analysis of tomato and comparative genomics of fruit ripening

Zhangjun Fei1,2, Xuemei Tang1, Rob M. Alba1, Joseph A. White3, Catherine M. Ronning3, Gregory B. Martin1,4, Steven D. Tanksley5 and James J. Giovannoni1,6,*

1Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY 14853, USA, 2Department of Horticultural Sciences, Texas A&M University, College Station, TX 77843, USA, 3The Institute for Genome Research, Rockville, MD 20850, USA, 4Department of Plant Pathology, Cornell University, Ithaca, NY 14853, USA, 5Department of Plant Breeding, Cornell University, Ithaca, NY 14853, USA, and 6United States Department of Agriculture, Plant, Soil, and Nutrition Laboratory, Ithaca, NY 14853, USA

Received 19 March 2004; revised 26 June 2004; accepted 2 July 2004.
*For correspondence (fax +1 607 254 2958; e-mail jjg33@cornell.edu).

Summary

A large tomato expressed sequence tag (EST) dataset (152,635 total) was analyzed to gain insights into differential gene expression among diverse plant tissues representing a range of developmental programs and biological responses. These ESTs were clustered and assembled to a total of 31,012 unique gene sequences. To better understand tomato gene expression at a plant system level and to identify differentially expressed and tissue-specific genes, we developed and implemented a digital expression analysis protocol. By clustering genes according to their relative abundance in the various EST libraries, expression patterns of genes across various tissues were generated and genes with similar patterns were grouped. In addition, tissues themselves were clustered for relatedness based on relative gene expression as a means of validating the integrity of the EST data as representative of relative gene expression. Arabidopsis and grape EST collections were also characterized to facilitate cross-species comparisons where possible. Tomato fruit digital expression data was specifically compared with publicly available grape EST data to gain insight into molecular manifestation of ripening processes across diverse taxa and resulted in identification of common transcription factors not previously associated with ripening.

Keywords: tomato, gene expression, expressed sequence tag, fruit ripening, comparative genomics.

Introduction

Expressed sequence tags (ESTs) are created by partially sequencing randomly isolated gene transcripts that have been converted into cDNA (Adams et al., 1992). ESTs have played significant roles in accelerating gene discovery including gene family expansion (Bourdon et al., 2002; Rogaev et al., 1995), large-scale expression analysis (Ewing et al., 1999; Ogihara et al., 2003; Ronning et al., 2003), and elucidating phylogenetic relationships (Nishiyama et al., 2003). Recent developments in high volume biotechnology combined with advanced DNA sequencing technology have made it feasible to perform large-scale EST sequencing projects. Currently there are near 20 million ESTs in the NCBI public collection-dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/). With many large-scale EST sequencing projects in progress and new projects being initiated, the number of ESTs in the public domain will likely increase substantially providing additional opportunities for intra and inter-specific expression comparisons on a genomics scale.

It has been shown previously that EST databases are a valid and reliable source of gene expression data (Ewing et al., 1999; Ogihara et al., 2003; Ronning et al., 2003). With the rapid expansion of available EST data, opportunities for digital gene expression analysis will continue to expand. As a result of advances in computational molecular biology and biostatistics, it is possible to mine and analyze large-scale EST datasets efficiently and exhaustively (Ewing et al., 1999; Ogihara et al., 2003; Ronning et al., 2003).
Tomato is a member of the family Solanaceae that includes several additional economically important crops such as potato, pepper, and eggplant and as such represents the most valuable plant family in terms of vegetable crops with important contributions to human health and nutrition. Tomato has long served as a model system for plant genetics, development, physiology, pathology and fruit ripening resulting in the accumulation of substantial information regarding the biology of this economically important organism. Recently a large, publicly available tomato EST database has been generated with support from the National Science Foundation Plant Genome Program (http://www.tigr.org/tdb/tgi/lgi; http://www.sgn.cornell.edu) (Quackenbush et al., 2001; Van der Hoeven et al., 2002). Several recently developed bioinformatics and statistical tools have allowed us to perform global expression analysis using EST data generated under this and other projects (Audic and Claverie, 1997; Ewing et al., 1999; Greller and Tobin, 1999; Stekel et al., 2000; Strausberg et al., 2001). Here we present correlated expression profile for 6758 genes across 25 different tomato tissues. We highlight highly expressed genes in various tomato tissues and compare with corresponding Arabidopsis and grape tissues where possible. In addition, we identified sets of differentially expressed genes associated with plant development with an emphasis on fruit ripening. We also performed comparative analysis with available grape ESTs derived from fruit to gain insights into common and distinctive features of these diverse fruit species at the molecular level and identified a set of transcription factors induced during both tomato and grape ripening that had not previously been associated with ripening.

Results

EST foundation for tomato digital expression analysis

A total of 154 054 high quality sequences representing 152 635 distinct ESTs from 27 different cDNA libraries (Table 1) were generated under two NSF-funded tomato functional genomics projects (see Acknowledgements). After clustering and assembly, these ESTs were reduced to 31 012 unique genes, of which 15 925 are Tentative Consensus sequences (TCs) and 15 087 are singletons. The tomato unigene set is available through both TIGR (http://www.tigr.org/tdb/tgi/lgi) and the Solanaceae Genome Network (http://www.sgn.cornell.edu).

<table>
<thead>
<tr>
<th>Category no.</th>
<th>Library description</th>
<th>No. of ESTs</th>
<th>No. of TCs</th>
<th>No. of singletons (%a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1045</td>
<td>Fruit, ovary</td>
<td>9878</td>
<td>3072</td>
<td>623 (16.9)</td>
</tr>
<tr>
<td>T10018</td>
<td>Fruit, developing/immature green</td>
<td>4240</td>
<td>2287</td>
<td>376 (14.1)</td>
</tr>
<tr>
<td>T1356</td>
<td>Fruit, mature green</td>
<td>5317</td>
<td>2540</td>
<td>658 (20.6)</td>
</tr>
<tr>
<td>T1775</td>
<td>Fruit, breaker</td>
<td>15207</td>
<td>4022</td>
<td>815 (16.8)</td>
</tr>
<tr>
<td>T1391</td>
<td>Fruit, red ripe</td>
<td>3895</td>
<td>1605</td>
<td>238 (12.9)</td>
</tr>
<tr>
<td>T1526</td>
<td>Flower, 0-3 mm buds</td>
<td>6259</td>
<td>3018</td>
<td>802 (21.0)</td>
</tr>
<tr>
<td>T1527</td>
<td>Flower, 3-8 mm buds</td>
<td>5524</td>
<td>2597</td>
<td>523 (16.8)</td>
</tr>
<tr>
<td>T1528</td>
<td>Flower, 8 mm preanthesis buds</td>
<td>5759</td>
<td>2665</td>
<td>575 (17.1)</td>
</tr>
<tr>
<td>T1529</td>
<td>Flower, anthesis (open flower)</td>
<td>5643</td>
<td>2630</td>
<td>659 (20.0)</td>
</tr>
<tr>
<td>T10393</td>
<td>Flower, mixed stages</td>
<td>7009</td>
<td>2799</td>
<td>884 (24.0)</td>
</tr>
<tr>
<td>T10227</td>
<td>Flower, wild tomato pollen</td>
<td>5426</td>
<td>796</td>
<td>440 (35.6)</td>
</tr>
<tr>
<td>T1079</td>
<td>Leaf, Pseudomonas susceptible</td>
<td>5243</td>
<td>2469</td>
<td>485 (16.4)</td>
</tr>
<tr>
<td>T1060</td>
<td>Leaf, Pseudomonas resistant</td>
<td>5127</td>
<td>2289</td>
<td>438 (16.1)</td>
</tr>
<tr>
<td>T1297</td>
<td>Leaf, mixed elicitor-induced</td>
<td>9135</td>
<td>3138</td>
<td>599 (16.0)</td>
</tr>
<tr>
<td>T10304</td>
<td>Shoot/meristem</td>
<td>9122</td>
<td>3930</td>
<td>901 (18.7)</td>
</tr>
<tr>
<td>T1005</td>
<td>Shoot</td>
<td>898</td>
<td>627</td>
<td>115 (15.5)</td>
</tr>
<tr>
<td>T1481</td>
<td>Root, plant at pre-anthesis</td>
<td>3259</td>
<td>1775</td>
<td>434 (19.6)</td>
</tr>
<tr>
<td>T1480</td>
<td>Root, plant at fruit set</td>
<td>3142</td>
<td>1782</td>
<td>405 (14.5)</td>
</tr>
<tr>
<td>T1450</td>
<td>Root, nutrient deficient</td>
<td>3175</td>
<td>1788</td>
<td>506 (22.1)</td>
</tr>
<tr>
<td>T1482</td>
<td>Root, etiolated radicle</td>
<td>2373</td>
<td>1436</td>
<td>205 (12.5)</td>
</tr>
<tr>
<td>T1457</td>
<td>Seed, germinating</td>
<td>3927</td>
<td>2004</td>
<td>331 (14.2)</td>
</tr>
<tr>
<td>T1048</td>
<td>Seed, quiescent</td>
<td>547</td>
<td>256</td>
<td>183 (41.7)</td>
</tr>
<tr>
<td>T1451</td>
<td>L. pennellii trichome</td>
<td>2729</td>
<td>1448</td>
<td>579 (28.6)</td>
</tr>
<tr>
<td>T1452</td>
<td>L. hirsutum trichome</td>
<td>2457</td>
<td>1236</td>
<td>463 (27.3)</td>
</tr>
<tr>
<td>T1207</td>
<td>Callus</td>
<td>14114</td>
<td>4440</td>
<td>1288 (22.5)</td>
</tr>
<tr>
<td>T10600</td>
<td>Suspension culture, untreated</td>
<td>8026</td>
<td>2821</td>
<td>579 (17.0)</td>
</tr>
<tr>
<td>T10284</td>
<td>Crown gall</td>
<td>5204</td>
<td>2749</td>
<td>624 (18.5)</td>
</tr>
</tbody>
</table>

*Calculated as the ratio of the number of singletons to the number of total unique sequences in the corresponding library.

Table 1 Tomato cDNA library statistics

In digital expression analysis, the relative abundance of a gene is defined simply as the ratio of homologous ESTs to the total number of ESTs in the corresponding pool (Schmitt et al., 1999). Singletons are likely to represent relatively low abundance transcripts, but are difficult to characterize further via digital expression analysis in a statistically meaningful manner because of their rarity. In each of the 27 tomato cDNA libraries analyzed, singletons represented from 12.5 to 41.7% of the unique genes in the library, while in most libraries, singletons represented less than 20% of the unique genes (Table 1). In contrast, we identified the most abundant transcripts in the combined tomato EST database and in each tissue, in addition to tissue-specific genes [defined as genes with a minimum of four EST members from a single tissue (P < 0.05; Audic and Claverie, 1997)]. The list of all these genes is available at the tomato digital expression database (http://ted.bti.cornell.edu).

Comparative analysis of tomato and Arabidopsis digital expression data

Digital expression analysis allows for genomics-scale cross-species comparisons that cannot be conducted via conventional wet-lab genomics approaches. For example, digital expression analysis can provide gene expression information that can be used to facilitate functional comparison of gene family members across species and on a genomics scale.

To demonstrate the utility of digital expression analysis for cross-species comparisons, we compared tomato EST collections to EST collections from equivalent Arabidopsis tissues. The TIGR Arabidopsis Gene Index contains EST sequence derived from five tissues that we compared with similar tomato EST collections. It should be noted that this comparison can be made because the compared libraries have not been normalized, subtracted or otherwise altered in their ability to represent steady-state gene expression of the tissues examined. The Arabidopsis EST collections that we compared with tomato include leaf (5452 ESTs), root (16 646 ESTs), flower buds (7679 ESTs), green silique (fruit) (13 045 ESTs), and seed (9936 ESTs). During comparative digital expression analysis, the raw EST count was normalized for the proportion of total ESTs. We initiated our analysis by (i) identifying tomato TCs with at least five EST members in the five comparable tissues and (ii) determined their corresponding AtGI top hits (comparing translated amino acids with an e-value of 1e-80 or less). The expression profile of tomato TCs and those of their corresponding AtGI hits across the five tissues were compared using the Pearson correlation coefficient (r) as a measurement of similarity. We identified 74 tomato TCs displaying a statistically similar expression profile in the five tissues analyzed as compared with Arabidopsis homologues (r > 0.9). In this highest quality data subset, the 74 corresponding Arabidopsis best hits also contained at least five EST members. The list of these genes is available at http://ted.bti.cornell.edu. The following examples demonstrate the value of large-scale comparative sequence and expression analysis in assessing functionally conserved homologs across species. In Figure 1(a) tomato TC116371 (likely 14-3-3 protein) has high-sequence similarity to all of its top six AGI hits (e-value less than e-130) making it difficult to identify the corresponding Arabidopsis gene likely to be most functionally similar based on sequence alone. However, TC116371 expression in the five comparable tissues is only similar to that of Arabidopsis TC183911 (r = 0.9689). Combined sequence and digital expression information can thus assist in identification of functionally conserved homologs across species.

Figure 1(b) shows that tomato TC117718 (likely peroxidase) is much more conserved at the sequence level to Arabidopsis TC180382 (e-136) than TC183774 (2e-91), yet its expression profile is more similar to TC183774, suggesting that while the sequence of TC183774 is more diverged, it is likely a better candidate for the homolog of TC117718 with biologically conserved function. Finally, Figure 1(c) shows that tomato TC124329 (likely GDSL-motif lipase/hydrolase) is highly conserved in both sequence and expression profile with both of its top Arabidopsis hits. While this particular result does not further facilitate elucidation of corresponding orthologs, it does suggest that additional homologous gene(s) may remain to be identified in tomato (whose genome remains to be sequenced), a hypothesis that would not have been a logical extension of examination of the tomato data alone. In addition to the 74 genes for which high-quality data is available, lower significance correlation can be viewed for additional genes (2889 from tomato and 2340 from Arabidopsis) at http://ted.bti.cornell.edu.

Coordinated gene expression analysis of tomato

To generate expression patterns of numerous genes across various tissues and to assess the overall similarities and differences between transcriptomes of different tissues/ organs, we performed coordinated gene expression analysis of the tomato EST collection. Only cDNA libraries with more than 1000 independent ESTs and TCs represented by at least five individual ESTs were selected for analysis, as described previously (Ewing et al., 1999). A total of 6758 genes and 25 cDNA libraries, respectively, met these criteria.

Figure 2(a) is the complete cluster of all 6758 genes across 25 tissues. Red regions specific to each tissue represent tissue/treatment-specific gene expression in the context of the 25 tissues analyzed. For example, the red region for pollen in Figure 2(a) represents 226 individual unigenes.
Many have been previously documented to be pollen-specific or pollen-related genes including LePRK1 and LePRK2 (TC118503 and TC125301; Muschietti et al., 1998), LAT52 (TC115959; Tang et al., 2002), and LeFRK4 (TC116279; German et al., 2002).

Correlations of genes with development or biology of specific tissues was observed and two examples are shown in Figure 2. Figure 2(b) depicts a cluster of genes highly expressed in leaf and shoot tissues, many of which have homology to genes involved in photosynthesis. Figure 2(c) shows genes highly expressed in ripening fruit. Most genes in this cluster have previously documented roles in fruit ripening [e.g. E8 (Deikman and Fischer, 1988), β-fructosidase (Elliott et al., 1993), phytoene synthase (Bartley et al., 1992)].

Genes of unknown or minimally defined functions residing in such clusters (i.e. displaying coordinate expression) may represent additional participants in the same or associated biological processes and thus may represent candidates for further analysis.

Gene expression profiles from each tissue/treatment were also clustered (Figure 2a). EST collections derived from the most similar tissues or representing a developmental continuum typically clustered together. All five-fruit, four-root, three-leaf and one shoot and two trichome libraries clustered most tightly together, respectively. Furthermore, most libraries from the two sets of EST libraries representing developmental time courses (i.e. fruit and flower development) clustered in developmental order, further validating the analysis methods employed. The pseudomonas-resistant leaf library and mixed elicitors-treated leaf library also clustered most tightly. The only exception was for the open flower library that clustered most tightly and separately with the pollen EST collection (as opposed to the other floral libraries).
<table>
<thead>
<tr>
<th>Reference</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC115827</td>
<td>16kDa membrane protein</td>
</tr>
<tr>
<td>TC124447</td>
<td>unknown protein</td>
</tr>
<tr>
<td>TC116264</td>
<td>photosystem I reaction center subunit PSI-N</td>
</tr>
<tr>
<td>TC116862</td>
<td>cytochrome b6f complex</td>
</tr>
<tr>
<td>TC117160</td>
<td>glycine cleavage system protein H</td>
</tr>
<tr>
<td>TC124130</td>
<td>photosystem I reaction center subunit X</td>
</tr>
<tr>
<td>TC115970</td>
<td>cutinase homolog (E8)</td>
</tr>
<tr>
<td>TC124252</td>
<td>AP2 protein</td>
</tr>
<tr>
<td>TC116074</td>
<td>flavin-containing monooxygenase</td>
</tr>
<tr>
<td>TC125371</td>
<td>betalain cyclase</td>
</tr>
</tbody>
</table>

**Digital expression analysis of tomato**

---

Identification of tomato fruit ripening-related genes

We initially focused on ripening for a more detailed analysis of differential gene expression, with emphasis on the transition to ripening as demonstrated in the ‘breaker’ stage where the fruit display the first signs of lycopene accumulation (i.e. breaking color), ethylene evolution, and tissue softening (Giovannoni, 2001; Seymour et al., 1993 and references therein).

We compared tomato expression in mature green fruit (2–3 days prior to the onset of ripening) to that in breaker fruit (fruit showing the first visible signs of ripening). Whether or not expression was detected in other tissues was not a factor in this analysis. A total of 333 ripening-induced genes (greater expression in breaker, \( P < 0.05 \)) and 185 ripening-repressed genes (greater expression in mature green fruit, \( P < 0.05 \)) were identified.

The combined 518 ripening-related genes were subdivided into 13 categories according to putative function based on DNA sequence homology (Figure 3 and Tables S1 and S2). As a point of clarification, ethylene response element binding proteins (EREBPs) were placed in the ‘ethylene response’ category although they are also technically transcription factors. Such limitations in annotation, including the possibility of error propagation because of electronic annotation, should be considered in any report of large-scale genomics data.

Among the ripening-induced genes identified through this approach, a number had been previously shown to be associated with fruit ripening, including ACC oxidase 1.

Figure 3. Distribution of tomato ripening-related TCs by different functional categories. A total of 333 ripening-induced TCs (a) and 185 ripening-repressed TCs (b) were annotated by blast against the GenBank nr database and then classified into corresponding categories according to their putative functions. Number in each category represents the number of ripening-induced (a) and ripening-repressed (b) genes classified into the corresponding category.
et al., 1994), polygalacturonase 2A precursor (TC124082; Bird et al., 1988), histidine decarboxylase (TC123915; Picton et al., 1993), and phytoene synthase (TC115970, TC116252; Bartley et al., 1992; Ray et al., 1992). Several additional genes previously associated with ripening via differential display were also identified here (TC116368, TC116059; Cordes et al., 1989; Giovannoni et al., 1999).

Among the ripening-induced genes, 14 share homology with sequences involved in ethylene biosynthesis or signal transduction. Within this group two are previously uncharacterized EREBP genes (TC116548 and TC116831). Twelve ripening-induced genes putatively encode loci involved in cell wall metabolism, 22 are related to defense or stress response sequences, and 11 are homologous to genes involved in fruit pigmentation.

Of the 185 ripening-repressed genes, only three are associated with cell wall metabolism, one with ethylene response (TC116320, an EREBP gene) and none were identified having clear relationships to fruit pigmentation, as would be anticipated. In addition, 14 genes are related to defense/stress response in contrast to 22 members of this classification group being induced during ripening. Particularly prevalent among the ripening-induced stress genes are members of the heat shock protein (HSP) family, suggesting a need for protein stabilization during ripening. Among the three ripening-repressed genes associated with cell wall metabolism, one (TC116388) corresponds to the β subunit of polygalacturonase isoenzyme 1 (PG1), previously reported to be downregulated during fruit ripening and consistent with its role in limiting pectin depolymerization (Zheng et al., 1992).

Few transcription factors other than the RIPENING-INHIBITOR MADS gene LeMADS-RIN (Vrebalov et al., 2002) have been functionally associated with the ripening process. Interestingly, 18 putative transcription factors in addition to three EREBP s were induced during ripening and 14 transcription factors in addition to one EREBP displayed expression negatively correlated with ripening, suggesting a large pool of putative regulatory elements for future functional analysis. Furthermore, 51 genes with no known or putative function are ripening-induced while 34 are ripening-repressed, suggesting an additional pool of genes for further analysis.

Digital expression profiling permits inter-species comparisons on a genomics scale: comparison of tomato and grape ripening-induced genes

Fruit can be physiologically classified as climacteric or non-climacteric depending on the presence or absence of a burst in respiration at the onset of ripening. Climacteric fruit also typically increases ethylene production at the onset of ripening and require this hormone for completion of the ripening process (Biale and Young, 1981). Tomato is a climacteric fruit while grape is considered to be a non-climacteric fruit (Davies and Robinson, 2000). Common ripening regulatory mechanisms operating in both climacteric and non-climacteric fruit remain elusive (Vrebalov et al., 2002). In order to gain insight into common regulatory mechanisms among diverse fruit species and ripening physiologies, we compared EST collections from ripening grape and tomato fruits.

The TIGR Grape Gene Index includes 83 675 sequences representing 66 501 ESTs and 17 091 unique genes. A total of 4126 ESTs in VvGI are from green fruit (stage II) and 5167 are from veraison stage fruit. Veraison represents the onset of grape ripening and thus is approximately comparable with breaker stage of tomato.

In comparing ESTs from green versus veraison fruit, 95 grape ripening-induced genes were identified at $P < 0.05$ (Table S3). Nine of the 95 genes are homologous to genes involved in cell wall metabolism and five represent previously characterized ripening-related genes (Davies and Robinson, 2000; Nunan et al., 2001). Fourteen ethylene synthesis/signaling genes were identified via digital expression analysis of tomato, while none was revealed by analysis of grape ESTs.

The cDNA sequences of all 95 grape and 333 tomato ripening-induced genes were compared at the translated amino acid level using the TBLASTX program (cutoff e-value of 1e-10). Twenty-three tomato ripening-induced genes had homologues in the set of grape-induced genes (Table 2). This number is an overestimate as several pairs of related tomato ripening genes were homologous to a single grape ripening gene. Interestingly, three transcription factors, including members of the MADS box, zinc finger, and bZIP transcription factor families, are among those identified in both species.

During the onset of fruit ripening, both climacteric and non-climacteric fruits typically undergo tissue softening mediated by cell wall modification, changes in the levels of flavor metabolites, and become more susceptible to pathogen attack. A number of genes associated with these ripening traits were identified in both species (Table 2), including xyloglucan endo-1,4-β-d-glucanase, which contributes to cell wall disassembly and is associated with fruit softening (Ishimaru and Kobayashi, 2002), alcohol dehydrogenase, which could impact fruit flavor (Speirs et al., 1998) and a gene encoding a putative PR protein.

A number of additional genes not previously associated with ripening, yet with putative roles in signal transduction, stress response and transport, were also common ripening-induced genes in both tomato and grape (Table 2). We additionally identified 181 grape ripening-repressed genes at $P < 0.05$ (Table S4) and 35 ripening-repressed genes from
tomato with homology to ripening-repressed genes from grape (Table S5).

### Online database – tomato digital expression database

The tomato EST digital expression data can be accessed online through the tomato digital expression database at [http://ted.bti.cornell.edu](http://ted.bti.cornell.edu). A relational database has been implemented, so users can query the data in multiple ways. Several interactive tools for expression data manipulation and analysis have been developed including tools for identifying differentially expressed genes between any two tissues, identifying additional genes with expression patterns similar to any selected gene of interest, and comparative gene expression of tomato and *Arabidopsis* homologues.

### Discussion

#### Utility of digital expression profiling

In this study, we performed large-scale transcriptome analysis and identified differentially expressed genes among diverse plant tissues and across species by using available EST information as a source of expression data for digital expression profiling. Available EST data allowed us to focus on fruit development and ripening. Fruit are major contributors of vitamins, fiber, carbohydrates, and phytonutrient compounds in the diet. The nutritional importance of fruit and vegetables is reflected in current USDA recommendations of five or more servings of fruit or vegetables a day for a healthy diet. The World Health Organization and the United Nations Food and Agriculture Organization (FAO) recently launched an effort to enhance fruit and vegetable consumption worldwide as low consumption is considered one of the top 10 contributing factors to human mortality (http://www.fao.org/english/newsroom/news/2003/24439-en.html).

Digital expression analysis has several advantages over conventional microarray approaches in that the reliability of the later can be reduced by cross-hybridization of closely related sequences (Kuo et al., 2002; Lipshutz et al., 1999) and development of stable probe secondary structures (Southern et al., 1999). The main limitation of digital expression analysis is the availability of large unbiased cDNA libraries from tissues of interest. In this regard, the resolution of digital expression profiling can be low if a given EST collection is small. The overall effort and cost of EST sequencing clearly precludes digital expression analysis as a general tool for expression analysis; nevertheless, as more EST collections are created, this approach becomes a valuable means of extracting expression information from available EST collections and for performing cross-species comparisons. While the data resulting from cross-species comparisons is limited to the number of tissues that can be compared, it can be highly useful, for example, in building the case for functional homology across species.

---

**Table 2** Homologous genes induced by both tomato and grape ripening

<table>
<thead>
<tr>
<th>Tomato TC</th>
<th>Grape TC</th>
<th>E-valuea</th>
<th>Putative function of tomato TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC125305</td>
<td>TC4377</td>
<td>4e-055</td>
<td>MADS box transcription factor</td>
</tr>
<tr>
<td>TC125359</td>
<td>TC4377</td>
<td>3e-072</td>
<td>MADS box transcription factor</td>
</tr>
<tr>
<td>TC124244</td>
<td>TC4730</td>
<td>9e-025</td>
<td>bZIP transcription factor</td>
</tr>
<tr>
<td>TC124112</td>
<td>TC4730</td>
<td>2e-042</td>
<td>bZIP transcription factor</td>
</tr>
<tr>
<td>TC124196</td>
<td>TC9044</td>
<td>1e-025</td>
<td>Putative zinc finger transcription factor</td>
</tr>
<tr>
<td>TC125034</td>
<td>TC9044</td>
<td>2e-025</td>
<td>Similar to zinc finger-like protein</td>
</tr>
<tr>
<td>TC116030</td>
<td>TC4282</td>
<td>5e-033</td>
<td>Xyloglucan endo-1,4-β-D-glucanase</td>
</tr>
<tr>
<td>TC115998</td>
<td>TC4249</td>
<td>2e-080</td>
<td>Pathogenesis-related protein</td>
</tr>
<tr>
<td>TC123883</td>
<td>TC3949</td>
<td>0.0</td>
<td>Alcohol dehydrogenase 2</td>
</tr>
<tr>
<td>TC124274</td>
<td>TC3949</td>
<td>e-160</td>
<td>Alcohol dehydrogenase class III</td>
</tr>
<tr>
<td>TC116962</td>
<td>TC4046</td>
<td>6e-060</td>
<td>Heat shock protein 17.6</td>
</tr>
<tr>
<td>TC116318</td>
<td>TC4181</td>
<td>8e-052</td>
<td>Class I small heat shock protein</td>
</tr>
<tr>
<td>TC116319</td>
<td>TC4193</td>
<td>2e-053</td>
<td>HSP20.0 protein</td>
</tr>
<tr>
<td>TC124903</td>
<td>TC4348</td>
<td>9e-075</td>
<td>HSP70</td>
</tr>
<tr>
<td>TC126297</td>
<td>TC4348</td>
<td>0.0</td>
<td>HSP70</td>
</tr>
<tr>
<td>TC126413</td>
<td>TC4348</td>
<td>1e-017</td>
<td>Heat shock protein MTSHP precursor</td>
</tr>
<tr>
<td>TC122001</td>
<td>TC9134</td>
<td>9e-041</td>
<td>Heat shock protein 1-α</td>
</tr>
<tr>
<td>TC123771</td>
<td>TC4209</td>
<td>0.0</td>
<td>Elongation factor 1-α</td>
</tr>
<tr>
<td>TC115895</td>
<td>TC4236</td>
<td>0.0</td>
<td>Ubiquitin</td>
</tr>
<tr>
<td>TC125239</td>
<td>TC4910</td>
<td>4e-090</td>
<td>Calcinurin B-like protein 1</td>
</tr>
<tr>
<td>TC123892</td>
<td>TC9086</td>
<td>e-102</td>
<td>Calmodulin</td>
</tr>
<tr>
<td>TC124929</td>
<td>TC9284</td>
<td>2e-020</td>
<td>Copper-binding protein family</td>
</tr>
<tr>
<td>TC124731</td>
<td>TC9682</td>
<td>8e-034</td>
<td>Endoplasmatic reticulum retrieval protein</td>
</tr>
</tbody>
</table>

*a*This column represents the e-values obtained by comparing translated amino acid sequences of tomato TCs with the corresponding grape TCs using TBLASTX program.
by providing a second criteria of comparison in addition to sequence homology.

Approach and analysis
Relative expression patterns of 6758 tomato genes across 25 different tissues/treatments were developed (Figure 2a). Unlike other studies which employ raw EST counts (Ewing et al., 1999; Ogihara et al., 2003; Ronning et al., 2003), we normalized the number of ESTs in two ways. First, the raw EST counts for genes represented by sufficient number of ESTs for further analysis were transformed into relative expression values (see Experimental procedures) to correct for the different sampling sizes of the various EST collections. The relative expression value of every gene was then divided by the median expression of said gene across all 25 libraries. This step of normalization puts all genes on the same relative scale for comparison among all tissues. The normalization strategy employed in this study significantly improved the clustering (data not shown). As validation of this approach we found that genes with similar functions or in the same pathways (e.g. photosynthesis and fruit ripening, respectively), often have correlated expression patterns (Figure 2b,c).

Gene expression patterns of 25 different tissues/treatments were clustered according to their digital profiles (Figure 2a). EST collections derived from similar tissues or adjacent developmental stages typically clustered together, providing further validation of the analysis pipeline. Interestingly, open flower and pollen, which clustered together, did not cluster with other flower tissues. Open flower contains a large amount of pollen while other flower tissues (0–3 mm flower buds, 3–8 mm flower buds, and 8 mm preanthesis flower) do not. When re-analyzing the data excluding the subset of pollen-specific genes, open flower clustered with other flower tissues, while pollen remained distant (data not shown). This result combined with the fact that pollen harbored the largest number of tissue-specific genes (Honys and Twell, 2003) provides an explanation for why these two tissues represented an outgroup in this analysis.

Digital gene expression analysis of fruit ripening
Fruit ripening is a complex process influenced by numerous factors including light, hormones, temperature, and genotype. Considerable attention has been directed toward elucidating the molecular basis of fruit ripening in tomato (Giovannoni, 2001, 2004). In this study, we identified genes that are differentially expressed during fruit ripening by comparing mature green and breaker stage fruit EST collections. While a number of ripening-related genes identified in this study were reported previously, more than 300 ripening-induced and 180 ripening-repressed genes that had heretofore not been characterized in terms of ripening were identified.

One novel ripening-induced gene (TC115942) encodes a putative acyltransferase. This gene is highly expressed in ripening fruit (152 ESTs in breaker, 34 in red ripe and none in green fruit or any other tissues) and shows high similarity to acetyl-CoA:benzyl alcohol acetyl transferase from Clarkia breweri responsible for the production of the floral volatile benzyllactate (D’Auria et al., 2002). Alcohol acetyl transferases capable of catalyzing the formation of volatile esters which contribute to fruit flavor and aroma have been identified in strawberry (Aharoni et al., 2000) and this highly regulated gene in ripening tomato fruit is a likely candidate for a role in tomato fruit aroma.

Another interesting gene in terms of ripening expression is TC115905, which encodes pectate lyase and is highly expressed in ripening fruit (105 ESTs in breaker, 105 from red ripe library and only one in green fruit). The relationship between pectate lyase and fruit softening is well characterized in several other species, including banana, strawberry, and grape (Marín-Rodríguez et al., 2002). In tomato, polygalacturonase was thoroughly studied as a key mediator of fruit softening. Nevertheless, transgenic tomato plants with suppressed polygalacturonase expression showed no effects on fruit softening (Smith et al., 1989), implying that this process is more complicated than originally thought. Indeed, additional cell wall hydrolases and expansions have been more recently associated with tomato fruit softening (Rose et al., 1997; Smith and Gross, 2000). The prevalence and tight ripening regulation of pectate lyase during fruit development suggests the possibility of an additional contributor to tomato fruit softening.

Ten ripening-induced genes encoding putative HSPs were identified through digital expression analysis. It has recently been proposed that HSPs play roles in facilitating fruit ripening by protecting cellular machinery against heat stress during the daytime rise in field temperature (Ramakrishna et al., 2003). HSPs, which were originally thought to be protective factors induced specifically by heat stress, were also found to be developmentally regulated in the absence of stress (Vierling, 1991).

Thirty-six ripening-related genes identified by digital expression analysis encode putative transcription factors. Most of these genes were first associated with ripening in this study. Among this group of transcription factors, four belong to the EREBP family, of which three are ripening-induced (TC116548, TC116831, and TC116368) and one is ripening-repressed (TC116320). It has been reported that EREBPs act both as transcriptional activators and repressors in plants (Fujimoto et al., 2000). The fact that we identified both ripening-induced and repressed EREBPs fits such a model in which EREBPs dynamically regulate fruit ripening utilizing antagonistic mechanisms. While EREBP
transcription factors represent a large gene family in tomato, to date their role in fruit ripening remains unknown.

Three putative MADS box genes (TC125359, TC125305, and TC124719) were also among the group of tomato ripening-induced genes. The association of a MADS box gene with fruit ripening was reported recently by cloning the tomato LeMADS-RIN gene (Verhoeyen et al., 2002). MADS box genes are known to act as multimers (Riechmann et al., 1996), thus MADS box genes in addition to LeMADS-RIN could be anticipated to impact ripening. Other putative ripening-related transcription factors identified in this study, such as homeodomain proteins, basic leucine zipper proteins, and zinc finger proteins (Tables S1 and S2), have been found to regulate various processes of plant and animal development, while a relationship between these transcription factors and fruit ripening has not been previously documented.

Included among the ripening-related genes identified in this study are many whose sequences show no similarity to any previously described gene of known function. Association of expression of said genes with fruit development and ripening provides initial insights pertaining to general areas of possible function (proof will, of course, require further study).

Additional expression analyses

We also identified 931 and 250 differentially expressed genes during the five fruit development stages (ovary, immature green, mature green, breaker, and red ripe) and the four flower development stages (0–3, 3–8, 8 mm preanthesis, and open flower), respectively, all at a significance level of P < 0.01. Digital expression analysis was also performed on available tomato EST collections to identify pathogenesis-related genes via analysis of leaves from infected Pseudomonas susceptible plants, infected Pseudomonas-resistant plants (harboring a Pto transgene), and leaves of plants treated with a number of elicitors of plant defense responses (Table 1). A total of 169 differentially expressed genes (P < 0.05) were identified. Information for all these genes and their putative functions based on sequence homology is listed in http://ted.bti.cornell.edu. In addition, we performed digital expression analysis on tomato flavonoid biosynthetic pathway genes (Figure S1 and Table S6). Digital expression data is consistent with that for several previously reported flavonoid pathway genes (Muir et al., 2001; Verhoeyen et al., 2002) and preliminary data for several additional genes that have not been previously characterized in ripening fruit is now available through digital expression analysis. This information may be useful for further characterization of flavonoid pathway genes and eventual manipulation of flavonoid levels in ripening fruit.

Digital expression analysis facilitates cross-species comparisons

Tomato is a climacteric fruit requiring ethylene for ripening, while grape is non-climacteric and the role of ethylene in its maturation remains uncertain. Nevertheless, the ripening processes of both share several common features, including tissue softening, changes in pigmentation, increased vulnerability to pathogen infection, and accumulation of sugars responsible for at least part of the ripening-associated change in flavor quality. Digital expression analysis of the public grape EST collection, we identified 95 grape ripening-induced and 181 ripening-repressed genes and compared them with tomato ripening-related genes at the translated amino acid level. Homologous genes induced by both grape and tomato fruit ripening are shown in Table 2. Of note are three transcription factors, including MADS box, zinc finger, and bZIP transcription factor representatives. The necessary role of a MADS box gene acting upstream of ethylene in climacteric fruit ripening was recently described and proposed to represent a common regulator of ripening in climacteric and non-climacteric fruit (Verhoeyen et al., 2002). Grape MADS box gene TC4377 was previously shown to be induced by ripening (Boss et al., 2002). Although TC4377 is homologous to tomato TC125305 (TDR4) and TC125359 (Table 2), this MADS box gene is most similar (at the amino acid level) to TDR5 of tomato (e-113) which mediates organ differentiation in the three inner whorls of tomato flowers and is not associated with fruit ripening (Pnueli et al., 1994). Digital expression analysis indicates that the expression of TDR5 is not induced by tomato ripening, suggesting the possibility that different MADS box genes may have evolved functions in fruit ripening in different species.

A relationship of either zinc finger proteins or bZIP transcription factors to ripening has not been demonstrated to date, although this analysis suggests their possible roles in a conserved mechanism of ripening control transcending climacteric and non-climacteric distinctions.

Limitations of digital expression analysis

Differentially expressed genes identified via digital expression analysis are biased toward moderate or highly expressed genes to a degree dependent upon the total number of ESTs generated per tissue/treatment. For example, some well-documented differentially expressed genes during tomato fruit ripening (e.g. lycopene 7-cyclase and lycopene epsilon-cyclase; Ronen et al., 1999) may not be revealed by this approach because of the limited availability of corresponding ESTs from tissues analyzed. In addition, genes that are induced by both tomato and grape fruit ripening and expressed at low levels or in specific tissue or cell types not emphasized in EST sequencing efforts may be
missed. It is also important to note that in the analysis described here we focused on only those genes with highly conserved predicted peptide sequences. Most certainly there are additional functionally equivalent genes that were missed because of higher level of sequence divergence. Nevertheless, by setting appropriate statistical criteria, a substantive amount of expression information can be derived from even modest sized EST collections. Here we have clearly identified numerous genes that had not been previously associated with ripening, floral development, or pathogen responses and which may prove valuable in future efforts toward elucidating the genetic mechanisms underlying these processes. The online database developed in conjunction with this effort allows researchers to take advantage of the expression data recovered through digital expression analysis of the tomato EST collection. Continued collection of ESTs and their analysis for expression within and among species holds the promise of becoming a more widely used component of annotation in assessing gene function.

**Experimental procedures**

**cDNA library construction**

Twenty-seven non-normalized nor subtracted tomato cDNA libraries were constructed as described previously (Van der Hoeven et al., 2002). Sequencing of cDNA clones and construction of the tomato unigene build were as described in Van der Hoeven et al. (2002).

**Datasets acquisition and data analyses**

The EST datasets and the Unigene datasets of the TIGR Tomato Gene Index release 9.0 (LeGI; http://www.tigr.org/db/tgi/legi), the TIGR Arabidopsis Gene Index release 10.0 (AtGI; http://www.tigr.org/db/tgi/agi), and the TIGR Grape Gene Index release 2.0 (VvGI; http://www.tigr.org/db/tgi/vvgi) were provided through the TIGR ftp server. ESTs from subtracted and normalized libraries were eliminated for further digital expression analysis. For clones sequenced from both ends, only the 5’ sequence for each EST in LeGI and VvGI was used, while only the 3’ sequence for each EST in AtGI was used, as these are the predominant sequence information for the corresponding EST collections. The unigene dataset from LeGI, AtGI, and VvGI were electronically annotated by performing sequence similarity searches against the NCBI nr database using the BLASTX program with a cutoff e-value at 1e-10. BLASTX analysis was performed on the Cornell Computational Biology Service Unit (CBSU) parallel blast driver (pblast; http://www.ser-loopp.tc.cornell.edu/cbsu/pblast.htm) on a MS Windows PC computer cluster (128 Pentium III 1 GHz CPUs) at the Cornell Theory Center.

**Coordinated gene expression analysis of tomato**

cDNA libraries with more than 1000 ESTs and TCs with at least five members were selected for digital gene expression analysis. A two-way matrix table of raw EST counts was created with rows corresponding to TCs (genes) and columns corresponding to cDNA libraries. Raw data were normalized by: (i) Transforming the raw EST counts into relative expression values as defined by Schmitt et al. (1999) as the ratio of homologous ESTs to the total number of ESTs in the corresponding pool. (ii) The expression values for each gene was then normalized to itself by dividing all gene expression values for a given gene by the median of its expression over all the 25 tissues. If the median of any given gene was less than one then the number one was used for this normalization.

All genes were clustered using the Gene Tree classification and all conditions (cDNA libraries) were clustered using the Experiment Tree classification of GeneSpring software (v6.0; Silicon Genetics, Redwood City, CA, USA). In both classifications, the Pearson correlation was used as a measurement of similarity.

**Identification of differentially expressed genes**

AC statistics (Audic and Claverie, 1997) and general chi-square test were used to identify differentially expressed genes for pair-wise and multiple cDNA library comparisons, respectively, as described in Romualdi et al. (2001). P-values for AC statistics and general chi-square test were calculated using the IDEG6 program (http://telethon.bio.unipd.it/bioinfo/IDEG6/).

**Acknowledgements**

The authors are grateful to Dr Cornelius Barry and Dr Suma Chakravarthy for useful discussion, and to Dr Qi Sun for providing access to the pblast driver. This work was supported by grants from the National Science Foundation (DBI-9872617 and DBI-0116076 to ST, GM, and DBI-9813392 to TIGR).

**Supplementary material**

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ2188/TPJ2188sm.htm

**Figure S1.** Schematic illustration of the flavonoid biosynthesis pathway.

CHI, chalcone isomerase; CHS, chalcone synthase; DFR, dihydroflavonol reductase; F3H, flavanone 3-hydroxylase; F3′H, flavanone 3′-hydroxylase; FLS, flavonol synthase.

**Table S1.** Tomato ripening-induced genes

**Table S2.** Tomato ripening-repressed genes

**Table S3.** Grape ripening-induced genes

**Table S4.** Grape ripening-repressed genes

**Table S5.** Genes repressed by both tomato and grape ripening

**Table S6.** Expression of the flavonoid biosynthetic pathway genes in tomato fruits

**References**


Digital expression analysis of tomato 59


