In silico prediction of gene expression patterns in Citrus flavedo

Irving J. Berger1, Juliana Freitas-Astúa1,2, Marcelo S. Reis1, Maria Luísa P.N. Targon1 and Marcos A. Machado1

1Centro APTA Citros Sylvio Moreira, Instituto Agronômico de Campinas, Cordeirópolis, SP, Brazil. 2Embrapa Mandioca e Fruticultura Tropical, Cruz das Almas, BA, Brazil.

Abstract

Out of the 18,942 flavedo expressed sequences (clusters plus singletons) in Citrus sinensis from the Citrus EST Project (CitEST), 25 were statistically supported to be differentially expressed in this tissue after a double in silico hybridization strategy against leaf-, flower-, and bark-derived ESTs. Five of them, two terpene synthases and three O-methyltransferases, are absent in the other citrus tissues with concomitant 2x2 statistics, supporting the hypothesis that they are putative flavedo-specific expressed sequences. The pattern of these differentially expressed sequences during fruit development suggests that most of them are developmentally regulated. Some expressed gene products, including a putative germin-like protein highly expressed in flavedo, are shown to be promising candidates for further characterization. In addition to promoter seeking, this kind of analysis can lead to gene discovery, tissue-specific and tissue-enriched expression pattern predictions (as shown herein) and can also be adopted as an in silico first, and probably reliable approach, for detecting expression profiles from EST sequencing efforts before experimental validation is available or for heuristically guiding that validation.

Key words: EST, fruit, gene expression, orange, tissue-specific.

Received: August 14, 2006; Accepted: April 17, 2007.

Introduction

Gene discovery and analysis of gene expression are attractive fields of the molecular biology, especially in the current genomic and post genomic eras associated to several efforts toward sequencing and characterizing many organisms. Sequencing of cDNA molecules and expressed sequence tags (ESTs) have evolved as very efficient tools to identify transcriptional profiles and novel putative genes among different species or within the same species (Ohlrogge and Benning, 2000). While gene expression comparisons can be done directly among normalized libraries, for those that are not normalized, the comparisons and prediction of expression patterns are performed using their relative sequence abundance (Ewing et al., 1999).

A significant amount of information on citrus fruits has been obtained in the last few years. Several genes have been sequenced, their copy number and expression have been studied and their products analyzed regarding function in various tissues. Particular interest has been paid to fruit genes and gene products involved in the flavonoids, vitamins, and organic acid biosynthesis, among others (Moriguchi et al., 1999; 2001; 2002; Matella et al., 2005; Shimada et al., 2006). However, to our knowledge, the only genomewide study involving citrus ESTs that contains flavedo sequences already published included 5,604 ESTs from Citrus clementina fruits (Forment et al., 2005). This number, although significant, is approximately 10-fold lower than the 51,729 valid C. sinensis flavedo EST sequences from the CitEST Project, a Brazilian initiative coordinated by the Centro APTA Citros Sylvio Moreira. Moreover, the C. clementina study does not provide any information regarding comparative expression patterns, i.e., fruit-specificity or enriched expression in the tissue. Here, we present the first study addressing a large-scale search to predict gene expression patterns in citrus fruits, particularly in flavedo. This includes identification of the most probable candidates to flavedo-specific genes and the ones that are of clear preferential expression in this tissue and, finally, considerations regarding the possible regulation of some genes during fruit development. In order to accomplish this, we propose a combination of two statistical methods, the likelihood R-statistics (Stekel et al., 2000) and the P-statistics (Audic and Claverie, 1997), resulting in a double in silico hybridization strategy to be used for multiple library comparison.
Materials and Methods

ESTs were generated by the CitEST Project from diverse Citrus species and different tissues. Six sweet orange (C. sinensis var. Pera) fruit-derived libraries were constructed, containing cDNAs from flavedo of very young developing fruits (1 cm diameter) and five other developmental stages (up to 9 cm). Information concerning the construction of libraries, sequencing, sequence clustering and nomenclature can be found in Targon et al. and Reis et al. (both in this issue).

A double in silico hybridization strategy, combining the likelihood method (R-statistics) proposed by Stekel et al. (2000) to compare multiple libraries all at once and the P-statistics described by Audic and Claverie (1997), was used to identify differentially expressed clusters (or tentative consensi - TCs) among Citrus tissues. All of the statistically significant TCs from the R-statistics \((R > 8.0)\) were submitted to additional 2x2 P-statistics. Significance of flavedo abundance \((p < 0.05)\) within a cluster, in contrast to the leaf, flower and bark simultaneously, was interpreted as the condition to represent a differentially expressed TC in flavedo. Moreover, double significance \((R > 8.0 \text{ and } p < 0.05)\) of flavedo-exclusive clusters was interpreted as a strong suggestion of tissue specificity. The identification of flavedo-exclusive sequences was based upon the rule that the cluster/singleton must contain only reads derived from the cDNA libraries constructed from flavedo.

The putative identity of each Citrus flavedo TC was established by performing automated BLASTX (Altschul et al., 1997) searches against the GenBank database (Benson et al., 2000), the Arabidopsis Genome Initiative dataset and KEGG: Kyoto Encyclopedia of Genes and Genomes, considering e-values lower than \(e^{-10}\). All flavedo expressed sequences were functionally classified according to MIPS Funcat (Mewes et al., 2004).

In order to increase the reliability of the putative identification of each TC, comparisons were made against the Pfam database of protein domains (Finn et al., 2006), using the HMMER implementation of Hidden Markov Models Profiles (Durbin et al., 1998), considering e-values lower than \(e^{-10}\).

In addition to a comparison between gene expression from sweet orange flavedo and other tissues (flower, bark and leaf), an analysis of differential gene expression among each of the six flavedo libraries corresponding to various fruit developmental stages was also performed adopting the likelihood R-statistics proposed by Stekel et al. (2000). Significance \((R > 8.0)\) of a given cluster was an indication of differential expression of its corresponding sequence, at least in one of the distinct sampled fruit stages, and interpreted as a suggestion that the transcript is under developmental regulation.

Results

Identification of flavedo expressed sequences

Flavedo libraries of the CitEST Project include cDNA sequences derived from six distinct fruit developmental stages of Pera sweet orange with a total of 51,729 valid reads. They represent 21.3% of the whole CitEST databank and 42.6% of the C. sinensis valid sequences (see Targon et al. in this issue for more information concerning the CitEST libraries).

Assembly of all C. sinensis sequences from CitEST (Reis et al., this issue) resulted in 8,513 clusters and 10,429 singletons containing flavedo ESTs, adding up to 18,942 putative coding sequences hereafter named flavedo expressed sequences. From the assembled clusters, 2,852 presented only reads derived from the flavedo cDNA libraries and together with the singletons they represent flavedo-exclusive sequences.

Automated MIPS Funcat based categorization of the 18,942 flavedo expressed sequences resulted in the overview shown in Figure 1. The data presented in the figure represents a subset of the overall CitEST project categorization. A total of 53.4% of the expressed gene products were classified into defined function categories. Metabolism-related encoding sequences represent the most abundant category of expressed sequences (1,992 ESTs), with 38.9% related to C-compound and carbohydrate metabolism. Figure 1 also shows expressed sequences coding for proteins related to protein fate, subcellular localization and binding function or cofactor requirements as the next most abundant ones, with at least 5% each of the identified flavedo expressed sequences of the CitEST Project. A total of 645 sequences were related to transcription (3.4%), and close to 72% of those were putatively involved with transcriptional control.

Flavedo differentially expressed sequences

A global assembly of flavedo (51,729), leaf (34,957), flower (9,082) and bark (7,952) derived ESTs from healthy tissues of C. sinensis was performed (Reis et al., this issue)
to generate a single pool of clusters and run the in silico hybridization likelihood method proposed by Stekel et al. (2000). This resulted in 232 differentially abundant clusters (R > 8.0). Differentially expressed sequences were defined as being over or underexpressed in at least one of the tissues in comparison to the others. Additional support of 2x2 in silico hybridizations (P-statistics from Audic and Claverie, 1997) for these 232 clusters yielded 25 TCs, composed of 21 to 339 reads each, which were significantly more abundant in flavedo than in other tissues simultaneously (p-value < 0.05 in contrast to leaf, flower and bark).

Results of BLASTX and PFAM searches of the deduced amino acid residues from those double significant 25 clusters, hereafter considered differentially expressed transcripts of *C. sinensis* flavedo, can be found in Tables 1 and 2. It should be noted that, overall, the results from these two methodologies were similar. MIPS based categorization of other nine categories (transcription, protein fate, regulation, localization, cell wall, binding or cofactor, storage protein, transport facilitation and not yet clear cut) were all represented by single clusters.

From the 10 secondary metabolism-related clusters, four are terpene synthases or terpene synthase-like coding sequences and six are O-methyltransferases, the latter probably involved in scent biosynthesis (see Table 1).

Within metabolism-related clusters, a putative germin-like protein encoded by TC231 can be highlighted. It is the most abundant transcript from the whole flavedo CitEST databank with 339 reads sequenced from that tissue and a single cDNA sequence from the bark library (BLAST matches BAB10382). It represents around 0.7% of the total differentially expressed clusters. Together with another four metabolism-related, a total of 14 flavedo expressed sequences were found to be involved in metabolism (56%). Two clusters comprising a transferase (candidate to an elicitor inducible gene, TC174) and a hypothetical protein (TC237) were classified as “unclassified proteins.” The other nine categories (transcription, protein fate, regulation, localization, cell wall, binding or cofactor, storage protein, transport facilitation and not yet clear cut) were all represented by single clusters.

### Table 1 - Flavedo differentially expressed sequences, putative flavedo-specific and ‐enriched genes: BLAST results.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>BLASTX against GenBank</th>
<th>Length (aa)</th>
<th>Accession #</th>
<th>E-value</th>
<th>Identities</th>
<th>Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC242</td>
<td>d-limonene synthase [Citrus unshiu]</td>
<td>608</td>
<td>BAD27257.1</td>
<td>0.0</td>
<td>522/523 (99%)</td>
<td>522/523 (99%)</td>
</tr>
<tr>
<td>TC226</td>
<td>Gamma-terpinene synthase [Citrus limon]</td>
<td>600</td>
<td>AAM53943.1</td>
<td>0.0</td>
<td>362/373 (97%)</td>
<td>368/373 (98%)</td>
</tr>
<tr>
<td>TC214</td>
<td>Caffeic acid O-methyltransferase [Rosa chinensis]</td>
<td>359</td>
<td>BAC78828.1</td>
<td>e-113</td>
<td>198/354 (55%)</td>
<td>257/354 (72%)</td>
</tr>
<tr>
<td>TC256</td>
<td>Orcinol O-methyltransferase 3 [R. hybrid cultivar]</td>
<td>346</td>
<td>CAH05085.1</td>
<td>1e-82</td>
<td>147/300 (49%)</td>
<td>205/300 (68%)</td>
</tr>
<tr>
<td>TC244</td>
<td>Orcinol O-methyltransferase 1 [R. chinensis]</td>
<td>367</td>
<td>CAH05077.1</td>
<td>3e-99</td>
<td>184/349 (52%)</td>
<td>246/349 (70%)</td>
</tr>
<tr>
<td>TC231</td>
<td>Germin-like protein [Arabidopsis thaliana]</td>
<td>222</td>
<td>BAB10832.1</td>
<td>8e-77</td>
<td>145/222 (65%)</td>
<td>170/222 (76%)</td>
</tr>
<tr>
<td>TC204</td>
<td>Patatin-like protein 3 [Nicotiana tabacum]</td>
<td>411</td>
<td>AA98369.1</td>
<td>e-146</td>
<td>254/385 (65%)</td>
<td>316/385 (82%)</td>
</tr>
<tr>
<td>TC223</td>
<td>Anthocyanin acyltransferase-like protein [A. thaliana]</td>
<td>449</td>
<td>BAB01191.1</td>
<td>5e-99</td>
<td>213/474 (44%)</td>
<td>291/474 (61%)</td>
</tr>
<tr>
<td>TC227</td>
<td>Orcinol O-methyltransferase 1 [Rosa chinensis]</td>
<td>367</td>
<td>CAH05077.1</td>
<td>3e-92</td>
<td>162/347 (48%)</td>
<td>237/347 (68%)</td>
</tr>
<tr>
<td>TC174</td>
<td>Elicitor inducible gene product EIG-I24 [N. tabacum]</td>
<td>442</td>
<td>BAB16426.1</td>
<td>e-143</td>
<td>254/440 (57%)</td>
<td>314/440 (71%)</td>
</tr>
<tr>
<td>TC201</td>
<td>Catechol O-methyltransferase [N. tabacum]</td>
<td>364</td>
<td>CAA52461.1</td>
<td>e-110</td>
<td>191/340 (56%)</td>
<td>250/340 (73%)</td>
</tr>
<tr>
<td>TC184</td>
<td>Glutamate decarboxylase [Arabidopsis thaliana]</td>
<td>494</td>
<td>BAB20870.1</td>
<td>0.0</td>
<td>390/498 (78%)</td>
<td>443/498 (88%)</td>
</tr>
<tr>
<td>TC237</td>
<td>Hypothetical protein [Oryza sativa (japonica cultivar-group)]</td>
<td>248</td>
<td>BAD46202.1</td>
<td>2e-34</td>
<td>99/205 (48%)</td>
<td>115/205 (56%)</td>
</tr>
<tr>
<td>TC176</td>
<td>Caffeoyl-CoA O-methyltransferase (Populus tremuloides)</td>
<td>247</td>
<td>AAA80651.1</td>
<td>e-103</td>
<td>180/248 (72%)</td>
<td>215/248 (86%)</td>
</tr>
<tr>
<td>TC041</td>
<td>Putative protein [Arabidopsis thaliana]</td>
<td>408</td>
<td>CAB96668.1</td>
<td>e-137</td>
<td>228/384 (59%)</td>
<td>297/384 (77%)</td>
</tr>
<tr>
<td>TC182</td>
<td>Eugenol O-methyltransferase [Rosa chinensis]</td>
<td>366</td>
<td>BAC78826.1</td>
<td>e-110</td>
<td>195/346 (56%)</td>
<td>257/346 (74%)</td>
</tr>
<tr>
<td>TC261</td>
<td>Terpene synthase [Vitis vinifera]</td>
<td>557</td>
<td>AAS66357.1</td>
<td>0.0</td>
<td>348/562 (97%)</td>
<td>429/562 (98%)</td>
</tr>
<tr>
<td>TC395</td>
<td>Limonoid UDP-glucosyltransferase [Citrus unshiu]</td>
<td>511</td>
<td>Q9MB73</td>
<td>e-108</td>
<td>208/467 (44%)</td>
<td>299/467 (64%)</td>
</tr>
<tr>
<td>TC169</td>
<td>Auxin-repressed protein-like - ARP1 [Manihot esculenta]</td>
<td>117</td>
<td>AAX84677.1</td>
<td>1e-49</td>
<td>93/120 (78%)</td>
<td>106/120 (88%)</td>
</tr>
<tr>
<td>TC114</td>
<td>Cytochrome P450-like protein [A. thaliana]</td>
<td>524</td>
<td>CAB79912.1</td>
<td>e-131</td>
<td>274/516 (53%)</td>
<td>370/516 (71%)</td>
</tr>
<tr>
<td>TC247</td>
<td>NADP-isocitrate dehydrogenase [Citrus limon]</td>
<td>414</td>
<td>AAD51361.1</td>
<td>0.0</td>
<td>408/414 (98%)</td>
<td>413/414 (99%)</td>
</tr>
<tr>
<td>TC272</td>
<td>Thiamin biosynthesis [Citrus sinensis]</td>
<td>356</td>
<td>CAB05370.1</td>
<td>0.0</td>
<td>356/356 (100%)</td>
<td>356/356 (100%)</td>
</tr>
<tr>
<td>TC280</td>
<td>MADS-box protein 4 [Vitis vinifera]</td>
<td>242</td>
<td>AAM21344.1</td>
<td>e-115</td>
<td>211/243 (86%)</td>
<td>225/243 (92%)</td>
</tr>
<tr>
<td>TC251</td>
<td>Monoxygenase family protein [A. thaliana]</td>
<td>408</td>
<td>NP196694.1</td>
<td>e-136</td>
<td>228/386 (59%)</td>
<td>296/386 (76%)</td>
</tr>
<tr>
<td>TC248</td>
<td>Dehydrin COR15 [C. clementina x C. reticulata]</td>
<td>138</td>
<td>AAQ92310.1</td>
<td>9e-79</td>
<td>138/138 (100%)</td>
<td>138/138 (100%)</td>
</tr>
</tbody>
</table>

1Putative flavedo-specific genes.
flavedo expressed sequences and, hence, is a great candidate for experimental validation and promoter isolation. It would represent a promoter for driving strong and almost exclusive gene expression in fruit for basic and/or applied transgene approaches. It should be noted that this high level of expression is impressive. Forment et al. (2005), working with 22,635 C. clementina ESTs, mentioned that the number of ESTs per cluster ranged between 2 (1650 clusters) and 76 (one cluster, corresponding to a polyubiquitin gene), while most clusters (80%) contained four or fewer ESTs.

Another interesting assigned cluster is TC280. It encodes the single putative transcriptional control-related protein differentially expressed in flavedo and it is a probable ortholog of the Vitis vinifera MADS-box protein 4 (Table 1). In grapevine, this gene was shown to be expressed on flowers and developing fruits (high levels in the latter), suggesting that it has a role in regulating both grapevine flower and berry development (Boss et al., 2002). Our data are in accordance with these findings, since ESTs from that expressed transcript were only found in fruits and flowers within the C. sinensis CitEST dataset, with a 3.8-fold higher relative abundance in the former.

The 25 sequences differentially expressed in flavedo can be divided into two groups: those predominantly expressed in flavedo, but still expressed in other tissue(s), and those specifically expressed in flavedo (Figure 2). Twenty clusters could be classified into the first group, characterizing sequences of enriched expression in flavedo (Table 1; Figure 2). Only five clusters met the requirements for being specifically expressed in flavedo libraries; they include two terpene synthases and three O-methyltransferases, hereinafter concluded to be putative flavedo-specific genes.

The two terpene synthases (TC242 and TC226, Table 1) presented similarity with already characterized Citrus sp. sequences through BLASTX. TC242 presented 99% identity at the amino acid level with a d-limonene synthase (accession #BAD27257) from Satsuma mandarin (C. unshiu) cloned and validated by Shimada et al. (2005). TC226 is 97% identical, at the amino acid level, to a gamma-terpinene synthase from C. limon (#AAM53943).

There is no information concerning tissue specificity in C. limon; whereas, it was identified from a cDNA sequenced library from the peel of young developing fruits in the species (Lücker et al., 2002).
The O-methyltransferases pointed out herein as putative flavedo-specific transcripts (TC214, TC256 and TC244; Table 1) are all probably related to scent biosynthesis, according to the BLASTX search, since they matched O-methyltransferases related to the synthesis of volatile compounds involved with scent within the *Rosa* genus. TC214 presented 72% of similarity with a caffeic acid O-methyltransferase (#BAC7828) from *R.* *chinensis* (var. spontanea) characterized by Wu *et al.* (2003). TC256 and TC244 showed 69 and 70% of similarity to orcinol-O-methyltransferase genes from *R.* *gallica* (#CAH05079) and *R.* *chinensis* (#CAH05077), respectively, both identified and characterized in Scalliet *et al.* (2006). They were all described as petal-specific in the *Rosa* genus; however, no corresponding ESTs were identified in flower tissues of *Citrus* sp. up to now.

Alignment of these O-methyltransferase clusters (data not shown) by the CAP tool (nucleotides) into BioEdit (Hall, 1999) and Clustal W (amino acids; Thompson *et al.*, 1994) confirmed the first ESTs assembly performed. Indeed, it supports the existence of three distinct consensi sequences coding for O-methyltransferases solely expressed in flavedo. In addition, it shows that TC256 and TC244 are very conservative despite the fact that they are distinct. Alignment of high quality 807 bp partial sequences from both showed that they share a high identity (98.38%) and present comparable abundances in the CitEST flavedo libraries (25 and 21 sequenced reads, respectively).

**Flavedo gene expression patterns during development**

Additional *in silico* hybridization of all flavedo derived sequence tags from the six distinct CitEST libraries (6x6 hybridization analysis, Stekel *et al.*, 2000) resulted in information hypothetically concerning differential expression of flavedo expressed sequences during different fruit development stages. For that particular analysis, a total of 352 clusters were concluded to be differentially expressed at least in one of the distinct fruit stages (data not shown), suggesting that they may be under developmental regulation.

From the five putative flavedo-specific genes, only TC244 was not differentially expressed among the various developmental stages represented by the flavedo libraries (R < 8.0). It means that any difference in relative abundances among distinct fruit stages is not statistically significant and the expression of the gene is expected to be around the same level during all stages of fruit development. It represents an average of 4.00 transcripts per developmental stage, relative to a 10,000 reads library (data not shown).

The terpene synthase genes showed different levels and general patterns of expression. TC242 (d-limonene synthase) is relatively more expressed than TC226 (gamma-terpinene synthase), averaging 17.14 and 5.79 transcripts per developmental stage, respectively, relative to a 10,000 reads library. While TC242 shows stable expression over time, TC226 clearly decreases its expression level from the second to the last studied developmental stage (as detailed in Figure 3).

The O-methyltransferases encoding TC256 and TC214 presented somewhat similar expression patterns during sweet orange fruit development. Differences between their expression patterns (depicted in Figure 3) seem to be mostly related to a higher expression of TC214 at the latter developmental stages. Indeed, despite sequence divergences, their average expression levels during the fruit development do not deviate considerably. TC256 and TC214 have 4.67 and 5.27 transcripts per developmental stage, respectively, relative to a 10,000 reads library.

From those expressed sequences characterized to be of enriched expression in flavedo, TC231 and TC280 (the putative germin-like and MADS-box 4 sequences, respectively) are highlighted. TC231 has a very high relative abundance as mentioned earlier. Figure 3 shows that its expression in flavedo is concentrated especially at the second fruit developmental stage studied and reduces drastically during fruit development and ripening. Whereas with relative abundance around 5-fold less than TC231, TC280 tends to be expressed in increasing levels from the earlier to the latter fruit developmental stages. These genes would represent good candidates for promoter cloning and conse-

![Figure 2 - Digital expression profile of the 25 differentially expressed clusters in flavedo](image-url)
 quaintly to drive gene expression in transgenic approaches. TC280 promoter would have the advantage of increased gene expression at the end of fruit development, leading to the accumulation of the desired gene product in mature fruits.

**Discussion**

Understanding fruit development and ripening as well as the genetic and biochemical basis for the biosynthesis of several secondary compounds in *Citrus* represents a fundamental and still demanding target for both basic and applied research. Gene identification and characterization are part of the important and numerous related objectives. From that point of view, the CitEST Project is a great source of information concerning gene expression patterns of flavedo as a support to gene expression analysis during fruit development and ripening. Moreover, even though the juice is undoubtedly the most important orange product, growing attention has been paid to citrus flavedo-derived compounds such as flavonoids (Frydman *et al.*, 2005) and essential oils (Sawamura *et al.*, 2004; Souza *et al.*, 2005).

Out of the 8,513 CitEST clusters containing at least one sequence from flavedo, 2,852 were detected to include only ESTs from the fruit libraries. Together with the 10,429 singletons, we conclude that 13,281 expressed sequences are exclusive from flavedo and could be considered as candidates to flavedo-specific genes. Indeed, *in silico* EST sequence analysis has been used to identify tissue-specific genes in diverse plant species (Figueiredo *et al.*, 2001; Dornelas and Rodriguez, 2004; 2005; Hecht *et al.*, 2005; Laitinen *et al.*, 2005; Dornelas and Rodriguez, 2006; Dornelas *et al.*, 2006). When ESTs are generated from non-normalized cDNA libraries, gene expression patterns normally can be inferred from the relative abundance of these sequences among different libraries (Ewing *et al.*, 1999) and the indication of tissue-specific sequences is based on a simplified rule. Any sequence observed in a given tissue is accepted to be specifically expressed in it. A rational strategy is to indicate the most abundant clusters (related to the number of reads a cluster is composed of) as the most probable candidates to tissue-specific. However, this can be somewhat misleading when gene expression is significantly higher in one tissue compared to another, particu-

**Figure 3** - Relative abundance (reads per 10,000 reads library) and expression pattern of the flavedo tentative consensi: TC242 (d-Limonene synthase), TC226 (Gamma-terpinene synthase), TC256 (Orcinol O-methyltransferase 3), TC214 (Caffeic acid O-methyltransferase), TC231 (Germin-like protein) and TC280 (MADS-box protein 4) throughout fruit development (from 1 cm to 9 cm fruit diameter).
larly when small libraries are considered. For example, we found a germin-like protein sequence with 339 reads in flavedo and a single read in citrus bark. The latter derived from a relatively large library containing a total of 9,752 valid reads. Using probability, we could not detect that bark sequence in a smaller library (since it was only one per 9,752 sequenced reads) because most of the studies are carried out on a small scale (fewer than 5,000 reads libraries), and we would consider germin-like protein to be erroneously flavedo-specific.

Here, we adopted a double in silico hybridization strategy to give statistical support to the prediction of putative expression patterns. By that strategy, we could first point out differentially expressed (or abundant) sequences in flavedo with respect to the other tissues (enriched expression) and then select the most probable candidates for flavedo-specific genes within all those 2,852 flavedo-exclusive clusters. In fact, using a congruent view, we determined exactly how abundant a cluster must be to be classified as a tissue-specific sequence with more expected reliability considering the present status of CitEST.

A recent application of an in silico double selection strategy for identification of Arabidopsis seed-specific genes in early stages of development proved that in silico subtraction as exclusive criterion shows little correlation with results from the in vitro experimental validation (Becerra et al., 2006). The authors combined the mentioned in silico subtraction with the available microarray data (the Arabidopsis Affymetrix 22k GeneChip®). From 585 candidate genes that were specifically expressed solely on immature seed libraries from the subtraction analyses, only 49 (8%) fulfilled the combined criteria and may represent genes specifically expressed in immature seeds.

Since we do not have extensive available microarray data as the model-plant Arabidopsis does and we were interested in implementing an in silico procedure to first analyze CitEST independently of any other database and second to use it with ease on any EST multilibrary dataset, we adopted the double in silico hybridization strategy as described. The procedure was certainly very stringent. Only 5 of those 2,852 flavedo-exclusive clusters (0.18%) were pointed out as supported candidates for flavedo-specific genes. This can be explained by the fact that transcripts with low abundance may not reach the statistical threshold (R > 8.0 and p < 0.05) to be considered differentially expressed in one tissue in contrast to the others, especially when library sizes are considerably distinct. The size of the flower and bark libraries - around 5 and 6.5-fold less abundant, respectively, in comparison to the flavedo libraries together (51,729 sequence tags) - was the statistical bottleneck. If only flavedo and leaf libraries were hybridized, a larger number of clusters would still be included into the tissue-specific genes. The decision for that more restrictive strategy, dependent on the double hybridization, is based on the intrinsic limitation of the likelihood R-statistics on distinguishing differentially expressed genes from the distinct studied tissues (Stekel et al., 2000). In addition, we also tried to find an extra statistical support to work with the CitEST non-normalized and very distinguishable libraries with respect to the sizes. Association of the Audic and Claveries’ P-statistics with the likelihood R-statistics gave us that support on estimating the probability of a sequence tag existing in a given tissue but not in another because of its expected tissue specificity rather than artifacts generated by considerably fewer cDNA stretches sequenced from the latter tissue. Nevertheless, the adopted strategy was very efficient in separating and indicating differentially expressed sequences for each distinct tissue (data not shown). Whether or not it was excessively restrictive for the identification of flavedo-specific genes, only an experimental validation can elucidate.

In fact, it would be of great interest to validate and correlate the double in silico hybridization strategy herein adopted with experimentally validated expression patterns. Strictly, correlation would represent it to be an efficient in silico approach for allowing tissue-specific gene identification. It is conclusive that the strategy at least produces a small number of most probable candidates for tissue-specific and overexpressed genes and, therefore, the first candidates for directing an experimental validation. These genes are indeed the most abundant flavedo expressed sequences and could be considered to lead promoter search for flavedo specific and enriched expression goals.

Moreover, our data suggest a temporal regulation of gene expression during fruit development for several flavedo expressed sequences, including the most probable flavedo-specific genes here noted, with the exception of TC244 O-methyltransferase (Figure 3). Similarly, most of the differentially expressed sequences that are here considered of enriched expression in flavedo are also expected to be under developmental regulation (R > 8.0; partial data in Figure 3). That indicates a complex gene expression profile in flavedo during fruit development and represents a vast field to be further explored.

Acknowledgments

The authors thank Carolina M. Rodrigues, Juliana M. de Souza, Kleber M. Borges and Silvia O. Dorta for technical assistance in sequencing the libraries and CNPq/ Millennium Institute (62.0054/01-8) for financially supporting this work.

References


Internet Resources

Associate Editor: Ivan de Godoy Maia