Identification of Genes with Potential Roles in Apple Fruit Development and Biochemistry through Large-Scale Statistical Analysis of Expressed Sequence Tags

Sunchung Park, Nobuko Sugimoto, Matthew D. Larson, Randy Beaudry, and Steven van Nocker

Pomology Group (S.P., S.v.N.) and Postharvest Technology and Physiology Laboratory (N.S., R.B.), Horticulture Department; and Bioinformatics Core, Research Technology Support Facility (M.D.L.), Michigan State University, East Lansing, Michigan 48824

Advanced studies of apple (Malus domestica Borkh) development, physiology, and biochemistry have been hampered by the lack of appropriate genomics tools. One exception is the recent acquisition of extensive expressed sequence tag (EST) data. The entire available EST dataset for apple resulted from the efforts of at least 20 contributors and was derived from more than 70 cDNA libraries representing diverse transcriptional profiles from a variety of organs, fruit parts, developmental stages, biotic and abiotic stresses, and from at least nine cultivars. We analyzed apple EST sequences available in public databanks using statistical algorithms to identify those apple genes that are likely to be highly expressed in fruit, expressed uniquely or preferentially in fruit, and/or temporally or spatially regulated during fruit growth and development. We applied these results to the analysis of biochemical pathways involved in biosynthesis of precursors for volatile esters and identified a subset of apple genes that may participate in generating flavor and aroma components found in mature fruit.

Cultivated apple (Malus domestica Borkh) is among the most diverse and ubiquitously cultivated fruit species. Apple is a member of the Rosaceae family, which includes many commercial fruit (e.g. pear [Pyrus communis], strawberry [Fragaria spp.], cherry [Prunus avium], peach/nectarine [Prunus persica], apricot [Prunus armeniaca]), nut (almond [Prunus amygdalus]), forest (black cherry [Prunus serotina Ehrh.]), and ornamental (rose [Rosa hybrida], crab apple [Malus coronaria]) species. In the United States alone, apple production is worth approximately $1.6 billion annually, and Rosaceous fruits, collectively, are the most economically important fruit crops (U.S. Department of Agriculture, National Agricultural Statistics Service, Noncitrus Fruits and Nuts 2004 Summary; http://www.nass.usda.gov/).

In spite of its importance to agriculture and its pervasive role in human health, relatively little is known about apple fruit development, physiology, and biochemistry. This lack of knowledge has contributed to perpetual difficulties in breeding, production, and storage. Unlike tomato (Lycopersicon esculentum Mill.), which has emerged as a model for fruit growth and development, there have been relatively few molecular-oriented studies of apple. However, such studies would be expected to yield novel insights into fruit biology. Like tomato, apple is a climacteric fruit, with a clear respiratory climacteric and ethylene peak associated with ripening. However, unlike tomato, which is a true berry fruit, the majority of apple fruit is derived from proliferated receptacle tissues, with the ovary-derived tissues restricted to the center of the mature fruit (core). The skin (epidermal and subepidermal cell layers) is strikingly different from the cortex, and, in almost all apple varieties studied thus far, biosynthesis of pigments and most volatile esters associated with aroma is concentrated in epidermal and subepidermal tissues. Pigmentation is dominated by anthocyanins (compared with carotenoids in tomato), and tomato fruit is not known to synthesize volatile esters. Compared with most fruits, apple has an extremely long developmental sequence, often exceeding 150 d. Although ripening in apple is accompanied by changes in texture, it is one of the few commercially important fruits that undergo significant softening only after extended storage and deterioration. An additional attraction of apple for studies of fruit biology is the enormous diversity in fruit-related traits among the large number of cultivars and related wild genotypes (>3,000) available for analysis.

Advanced studies of apple development, physiology, and biochemistry have been hampered by the general lack of appropriate genomics tools. An exception is the recent generation and public release of...
extensive expressed sequence tag (EST) data. Public sequence databanks contain in excess of approximately 200,000 apple EST sequences. The majority of the ESTs were contributed through large-scale sequencing efforts in the United States and New Zealand (Korban et al., 2004; Newcomb et al., 2006). The entire available EST dataset resulted from the efforts of at least 20 contributors and was derived from more than 70 cDNA libraries representing diverse transcriptional profiles from a variety of organs, fruit parts, developmental stages, biotic and abiotic stresses, and from at least nine cultivars (Table I; Supplemental Table I).

Analysis of ESTs provides a powerful complement to genome sequencing for model plants and the primary tool for gene discovery in many plant species of agronomic and economic interest. For example, in apple, Newcomb et al. (2006) recently analyzed approximately 150,000 ESTs from 43 libraries for the presence of potential molecular markers useful for genetic mapping and undertook a census of representatives of common protein families and genes potentially important for fruit quality. An additional attractive use of EST data is as an entry point for gene expression analysis. When derived from distinct sources and unbiased (representational) cDNA libraries, EST data can provide a reliable estimate of gene expression. In unbiased cDNA libraries, the frequency of occurrence of ESTs representing a given gene should be proportional to the level of expression (i.e. mRNA abundance) for that gene. This premise allows both estimation of relative expression levels of genes within a specified tissue source, as represented by the derived cDNA library, and comparison of relative gene expression levels between tissue sources. Rigorous statistical tests have been developed to assess the reliability of the information derived from this so-called digital expression approach (Audic and Claverie, 1997). An additional advantage of using EST frequency data to estimate gene expression is the ability to distinguish between closely related sequences, which may not be possible through other expression-profiling techniques, including microarray analysis.

Fruit tissues are well represented in the current apple EST collections (approximately 40% of sequences; Table I). This offers the opportunity to apply EST frequency analysis to a nonmodel fruit and to explore the validity of this approach where data are derived from heterogeneous sources. In this study, we performed an extensive analysis of apple EST sequence data found in public databanks as a first step in identifying genes with important function in apple fruit development, including the biosynthesis of volatile esters during ripening.

RESULTS

We assembled the approximately 200,000 apple sequences found in public sequence databanks into approximately 23,000 contiguous sequences (clusters containing more than one EST) and approximately 21,000 singletons (solitary or nonclustered ESTs; Table II). The number of unique sequences derived from the combined set of clusters and singletons (approximately 44,000) is similar to the number of unique sequences (approximately 43,000) determined by Newcomb et al. (2006) from clustering of approximately 150,000 apple ESTs, but substantially greater than the 6,928 unique sequences identified by the most recent National Center for Biotechnology Information (NCBI) UniGene apple assembly (http://www.ncbi.nlm.nih.gov/UniGene; UniGene Build No. 13). This discrepancy results predominantly from the requirement of UniGene for 3′ anchoring (clusters or singletons must contain an obvious 3′ poly(A) stretch or terminal oligonucleotide tag) and the consequent more limited number of ESTs considered (73,526 ESTs). In addition, UniGene discards clusters likely derived from ribosomal and non-nuclear genomic (mitochondrial and plastid) sequences, whereas we retained and flagged such sequences. We found that our clustering methods, when applied only to those apple sequences selected by UniGene, resulted in comparable numbers of clusters and singletons (data not shown).

The number of unique sequences identified in our study is an overestimate of the number of expressed genes sampled because clusters or singletons representing the same gene may not overlap and distinct clusters or singletons may represent alternatively transcribed or processed mRNAs originating from the same gene. Also, because the ESTs analyzed in this

<table>
<thead>
<tr>
<th>Table I. Apple ESTs by tissue source and variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numbers represent the total number and source of ESTs analyzed based on descriptions provided by contributors. A detailed description of EST sources is given in Supplemental Table I.</td>
</tr>
<tr>
<td>Tissue Source/ Variety</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td><strong>Part/Organ</strong></td>
</tr>
<tr>
<td>Fruit</td>
</tr>
<tr>
<td>Leaf</td>
</tr>
<tr>
<td>Flower</td>
</tr>
<tr>
<td>Shoot</td>
</tr>
<tr>
<td>Stem</td>
</tr>
<tr>
<td>Root</td>
</tr>
<tr>
<td>Cell culture</td>
</tr>
<tr>
<td>Seed</td>
</tr>
<tr>
<td>Other/not described</td>
</tr>
<tr>
<td><strong>Variety</strong></td>
</tr>
<tr>
<td>GoldRush</td>
</tr>
<tr>
<td>Royal Gala</td>
</tr>
<tr>
<td>M9 root stock</td>
</tr>
<tr>
<td>Pinkie</td>
</tr>
<tr>
<td>Braeburn</td>
</tr>
<tr>
<td>Red Delicious</td>
</tr>
<tr>
<td>Pacific Rose</td>
</tr>
<tr>
<td>Aotea</td>
</tr>
<tr>
<td>Fuji</td>
</tr>
<tr>
<td>Northern Spy</td>
</tr>
<tr>
<td>Other/not described</td>
</tr>
</tbody>
</table>


Downloaded from on July 20, 2017 - Published by www.plantphysiol.org
Copyright © 2006 American Society of Plant Biologists. All rights reserved.
study were derived from several cultivars and because cultivated apple is highly heterozygous, there is the potential for the degree of polymorphisms to exceed the sequence match stringency utilized for cluster analysis, leading to the clustering of sequences from distinct alleles of the same gene into separate clusters.

We reasoned that if allelic divergence was a significant factor contributing to cluster number, then this should be evident by frequent occurrence of homologous pairs of clusters each containing ESTs derived from only one cultivar. To explore this, we examined clusters of ESTs originating from young fruit or leaves, two tissue sources that were each represented by non-normalized libraries from both the Royal Gala and GoldRush cultivars (Supplemental Table I). We used a stringent statistical test (see “Materials and Methods”) to identify bias in cultivar-associated EST representation, then eliminated from consideration unbiased clusters and biased clusters containing ESTs derived from both cultivars. The 14,447 ESTs from young fruit were assembled into 7,439 clusters, of which 71 and 49 contained significant numbers of ESTs originating only from Royal Gala or GoldRush, respectively. Of these, 50 and 20, respectively, were found to contain ESTs from the other cultivar when the entire EST collection was considered, indicating that they did not represent a cultivar-specific gene or allele. We used the remaining 21 and 29 clusters as queries in BLAST analyses interrogating the entire clustered EST collection. Only two and four clusters from Royal Gala and GoldRush, respectively, were related to clusters specific to the other cultivar. Similarly, of the 6,354 clusters assembled from 12,677 leaf ESTs, we identified six and 23 clusters containing significant numbers of ESTs originating only from Royal Gala or GoldRush, respectively. Of these, three, 16, and 16, respectively, were found to contain ESTs from the other cultivar when the entire EST collection was considered. None of the remainders was related to clusters specific to the other cultivar. From these results, we concluded that polymorphism was not a significant factor in clustering in this study and that allelic variants were typically clustered together. The cultivar-specific clusters that we identified may correspond to genes that are differentially expressed between cultivars, an idea that we evaluated further below.

Unique sequences in our clustered set were used as queries in homology searches of protein and derived protein databanks. Approximately 40% of unique sequences did not exhibit significant homology with currently cataloged sequences from plant or other genomes (data not shown), suggesting that the respective apple tissues may represent a rich source of novel plant genes.

### Identification of Genes Potentially Highly Expressed in Fruit

As a first step in characterizing the transcriptional profile of fruit, we used EST frequency analysis to identify potential highly expressed genes in fruit tissues. This analysis collectively considered 19 non-normalized cDNA libraries derived from various stages of fruit development (very young to mature) and various fruit parts (core, cortex, and skin/peel; Supplemental Table I).

ESTs generated from these libraries comprised a total of 15,000 unique sequences, including about 10,300 clusters, with about 40% of clusters also including EST representatives from nonfruit tissues (see Table II). Clusters contained up to 284 fruit-derived ESTs, or about 0.7% of the fruit EST pool (Supplemental Table II), with the median cluster containing 39 ESTs. We found that the top second percentile of clusters in terms of number of representative ESTs (282 clusters) contained 24% of all fruit ESTs. A large proportion of these highly represented clusters (27 clusters [about 10%]) did not exhibit significant homology with the sequenced nuclear genome of the reference plant Arabidopsis (Arabidopsis thaliana), but instead were closely related to segments of sequenced plastid or mitochondrial genomes (Supplemental Table II). We presumed that these high-abundance sequences resulted from organellar DNA or RNA contamination of the RNA sources utilized in library construction and, therefore, these were eliminated from further analyses. We also found an additional significant proportion of clusters (9%) that did not exhibit sequence similarity with any currently cataloged sequence, were enriched in A/T nucleotides or contained extended A/T tracts, and/or did not contain an appreciably long open reading frame (ORF). Such clusters may also have originated from high-copy/repetitive apple nuclear genomic or organelar DNA or RNA or from uncharacterized microbial or viral sequences. An alternative explanation is that these could represent authentic products of transcription of nonprotein-coding genes or intergenic sequences (Yamada et al., 2003).

Of the 20 most abundantly represented clusters (Supplemental Table II), none represents genes whose function in relation to fruit biology is unambiguously known. The most highly represented cluster, MD249100, is very closely related to THI4, an Arabidopsis gene that participates in thiamine synthesis (Machado et al., 1997). Another representative of this highly represented group is MD176720, corresponding to a protein included in the glycoside hydrolase family 16 (GH16), which contains members with diverse

<table>
<thead>
<tr>
<th>Table II. Apple clustering data</th>
<th>No. of Sequences</th>
<th>Mean Sequence Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>cDNAs</td>
<td>616</td>
<td>nd *</td>
</tr>
<tr>
<td>ESTs</td>
<td>198,068</td>
<td>nd</td>
</tr>
<tr>
<td>Sequences in clusters</td>
<td>176,174</td>
<td>nd</td>
</tr>
<tr>
<td>Total no. of clusters</td>
<td>23,211</td>
<td>831</td>
</tr>
<tr>
<td>Singleton sequences</td>
<td>20,849</td>
<td>409</td>
</tr>
<tr>
<td>Total no. of unique sequences</td>
<td>44,060</td>
<td>645</td>
</tr>
</tbody>
</table>

*nd, Not determined.*
potential activities, including xyloglucan:xyloglucosyl transferase (xyloglucan endotransglycosylase [XET]) and xyloglucanase (Henrissat and Bairoch, 1993; http://afmb.cnrs-mrs.fr/CAZY). We presume that this gene plays an important role in cell wall modification during fruit growth, softening, or textural changes (see below). Of the remaining 235 highly represented clusters, approximately 20% showed significant homology with genes annotated as expressed, putative, hypothetical, or unknown. These data reveal that even the high-abundance portion of the apple fruit transcriptome remains relatively uncharacterized.

For each highly represented cluster, we also evaluated EST frequency in nonfruit-derived tissues (see below), as well as frequency of ESTs representing closely related sequences (paralogous groups), which may define functionally homologous genes (Supplemental Table II). Resulting data for the clusters were analyzed by k-means grouping (Fig. 1) to identify clusters with similar general frequency profiles across fruit-derived and nonfruit-derived EST sources. We found that there was a high degree of overlap between clusters represented at high frequency by fruit-derived ESTs and those that were represented preferentially by fruit ESTs, such that only a minority of highly represented clusters (Fig. 1, group 1; 87 of 282 clusters [approximately 30%]) were also highly represented by ESTs originating outside of fruit tissues. Genes corresponding to this group likely carry out functions important for both fruit-associated and nonfruit-associated tissues. Examples of group 1 clusters included those representing basic metabolic enzymes (MD249140, corresponding to the small subunit of Rubisco, and several clusters corresponding to chlorophyll a/b-binding proteins) and structural proteins (MD012900, corresponding to histone H2B, and several clusters corresponding to tubulin), among other functional classes.

The majority of clusters showed significantly higher representation by fruit-derived ESTs as compared with nonfruit-derived ESTs (Fig. 1, groups 2–5). Within this subset, clusters in group 2 (approximately 37% of total) either did not exhibit significant sequence homology with other apple sequences or were closely related to clusters only infrequently represented in apple EST libraries, suggesting that they may define genes with essentially nonredundant functions of special relevance for the fruit. A sample cluster, MD170530, corresponds to a gene whose closest homolog in Arabidopsis (E = 3e-59) is BOR1, encoding a protein required for xylem loading of boron in roots (Takano et al., 2002). In apple and other plant species that transport photoassimilates to sink tissues as sorbitol or other sugar alcohols, boron is cotranslocated as an acyl-CoA thioester from free fatty acids. Of 43 ESTs identified for this cluster, a paralogous group was defined, and clusters included in each paralogous group were similarly queried for EST representation from fruit or nonfruit libraries. Resulting data were subjected to k-means analysis and frequency profile groups were defined using the k-means analysis software.

Another example of a group 2 cluster is MD042900, corresponding to an acyl-CoA synthetase, an enzyme that forms CoA thioesters from fatty acids. Of 43 ESTs identified for this cluster, all but three were derived from fruit. The fruit ESTs were exclusively found in ripe or ripening fruit libraries, and it is possible that this gene participates in the biosynthesis of lipid-derived volatile esters (see below).

Clusters included in another profile class (Fig. 1, group 3) showed homology with other clusters that nonfruit libraries (data not shown). Another example of a group 2 cluster is MD042900, corresponding to an acyl-CoA synthetase, an enzyme that forms CoA thioesters from free fatty acids. Of 43 ESTs identified for this cluster, all but three were derived from fruit. The fruit ESTs were exclusively found in ripe or ripening fruit libraries, and it is possible that this gene participates in the biosynthesis of lipid-derived volatile esters (see below).
also exhibited significantly higher representation in libraries from fruit tissues relative to nonfruit tissues. These clusters and their relatives may define paralogous groups of genes with emphasized function in fruit. Examples in this group are MD024600 and MD024610, which show very high sequence homology with a reported progesterone 5β-reductase from Digitalis (Roca-Perez et al., 2004). As opposed to progesterone 5α-reductase, which participates in biosynthesis of brassinolide, progesterone 5β-reductases are relatively uncharacterized. In Digitalis species, this enzyme catalyzes a step in biosynthesis of cardinolide, a steroid best known for antibiotic activity of its derived glycosides (Gartner et al., 1990). Both clusters are strongly represented by ESTs from skin and cortex of ripening fruit (data not shown), and we hypothesize that the genes function in resistance to microbial or insect pathogens. The four additional clusters closely related to these two clusters are represented exclusively by ESTs derived from nonfruit tissues, but with a relatively low collective frequency (data not shown).

Another group 3 cluster, MD033250, represents a likely 1-aminocyclopropane 1-carboxylate oxidase (ACO), an enzyme that catalyzes a key committed step in ethylene biosynthesis and has a well-recognized role in fruit ripening and other ethylene-associated pathways. We identified seven additional ACO-like clusters that collectively showed higher representation in fruit-derived tissues (data not shown).

Clusters in another profile subset (Fig. 1, groups 4 and 5) were not strongly represented in nonfruit-derived libraries but were closely related to clusters represented with high frequency in nonfruit-derived libraries. Thus, group 4 and 5 clusters may define fruit-specific representatives of paralogous groups of genes with biochemical roles that are not limited to fruit. An example in this group is MD187410, encoding a likely lipoxygenase (LOX). Although MD187410 was represented almost exclusively by fruit-derived ESTs, we identified an additional 14 closely related clusters containing ESTs from at least 18 fruit and nonfruit libraries. LOX in vegetative tissues is best characterized in response to wounding through its role in biosynthesis of jasmonic acid (Liechti and Farmer, 2003); in ripening fruit, an additional potential role of LOXs includes generation of fatty acid-derived C6 short-chain volatile compounds important for flavor and aroma (see below; Chen et al., 2004). Consistent with this, MD187410 was represented by ESTs predominantly from cortex and peel of ripening fruit (data not shown), where volatile synthesis predominates.

Identification of Clusters Overrepresented or Underrepresented by Fruit-Derived ESTs

To expand this analysis and identify additional genes with potential fruit-specific roles, we identified all clusters that had statistically higher representation of ESTs from fruit-derived libraries as compared with nonfruit-derived libraries, regardless of their absolute frequency (Supplemental Table I). Using stringent selection criteria (see “Materials and Methods”), we identified 714 clusters overrepresented by fruit ESTs. Conversely, we identified 345 clusters as underrepresented by fruit ESTs (Supplemental Table III). Of these 1,059 clusters, 349 did not exhibit significant similarity (<1e-10) with any sequence cataloged in current databases. In fruit EST-overrepresented clusters, sequences classified as unknown were proportionally much higher than in the fruit EST-underrepresented clusters (Supplemental Table III; data not shown), perhaps reflecting the poor representation of sequences derived from fleshy fruits in current databases.

To help relate these data to gene function, we assigned functional categories to those clusters that were closely related to known genes (see “Materials and Methods”). For the majority of functional classifications, representation by clusters from the fruit-overrepresented and fruit-underrepresented classes was not significantly different. However, fruit-overrepresented clusters showed a marked enrichment for functions related to transcription and signal transduction (Fig. 2A). Analysis of clusters in these categories revealed that several were closely related to genes previously implicated in pathogen response and ethylene signaling. For example, MD094540 and MD116370 were very closely related to genes encoding ethylene-responsive, Ras-related GTP-binding proteins of the RAB8/ARA-3 and RAB11 classes, respectively, from tomato (Zegzouti et al., 1999; Moshkov et al., 2003). RAB11 expression was associated with fruit ripening and was proposed to have a role in mediating exocytosis of degradative enzymes to the cell wall (Lu et al., 2001). Another, MD043140, corresponds to a member of the ethylene response factor (ERF) family of transcriptional regulators most closely related with ERF5 genes from a range of plant species. A third example, MD114390, is substantially similar to the product of an ethylene-responsive gene in tomato classified as a homolog of eukaryotic multiprotein bridging factor 1 transcriptional coactivators (Zegzouti et al., 1999). As expected, genes related to photosynthesis and generation of precursor metabolites and energy were more often seen in the fruit-underrepresented gene category, reflecting the nonphotosynthetic nature of much of the analyzed fruit tissues (Fig. 2A).

To further help characterize function, we subjected fruit-overrepresented clusters to k-means analysis based on EST representation from various fruit libraries. For this analysis, the 19 non-normalized fruit cDNA libraries were assigned to eight classes according to stage and tissues (Supplemental Table I). The clusters were then grouped into eight different frequency profiles (Supplemental Table III; Fig. 2B). We noted that the vast majority of the fruit-overrepresented clusters exhibited an apparent specificity of EST representation according to stage and/or tissue (Fig. 2B). To lend support for this observation, we subjected the clusters to statistical analyses (see “Materials and Methods”) to identify those that were significantly
more or less represented in any one of the eight EST sources. Of the 714 clusters identified as fruit overrepresented, 573 showed statistically significant ($P < 0.01$) library-associated frequency, suggesting that the corresponding genes may act in a temporal and/or tissue-specific pattern during fruit development. Clusters found in group 1 (young fruit libraries) were generally not statistically well supported (median $P$ value = 0.058); this was at least partly as a result of the greater complexity of the source libraries and consequent lower proportional EST representation (Supplemental Table III) rather than to occurrence of large numbers of representative ESTs in other libraries. Of the remaining, statistically supported clusters, those in group 7 (predominantly 126 DAFB cortex and 150 DAFB cortex) generally had the least statistical support (median $P$ value = 0.0013), and this was likely due to broad EST representation among the four related libraries associated with ripening and mature fruit (126 DAFB core, 126 DAFB cortex, 150 DAFB cortex, and 150 DAFB skin).

Identification of Ripening-Related EST Clusters

To extend this analysis to the identification of potential ripening-related genes, we examined all clusters that contained EST representatives from libraries prepared from fruit cortex, regardless of representation in nonfruit-derived libraries, for significant differences in EST frequency among developmental stages. This analysis considered libraries derived from Royal Gala fruit 87 DAFB, 126 DAFB, or 150 DAFB, representing fruit at a late stage of cell expansion and growth, early in the ripening process, and ripe fruit, respectively. This resulted in the identification of 165 clusters, which we subsequently classified into 10 groups according to frequency profiles using $k$-means analysis (Fig. 3; Supplemental Table IV).

Clusters in groups 1, 2, and 3 showed significant frequency increases between 87- and 126-DAFB library sources, suggesting that the corresponding genes are up-regulated associated with the early ripening stage. Group 1 clusters showed a further significant increase between 126- and 150-DAFB library sources, whereas group 3 clusters showed significant decrease in EST frequency between 126- and 150-DAFB libraries.

An example of a group 1 cluster is MD028310, which is very closely related to known adenine

![Figure 2. A, Comparison of frequency of annotations within five functional categories for clusters overrepresented or underrepresented by fruit-derived, relative to nonfruit-derived, ESTs.](image-url)
phosphoribosyltransferases (APTs). A characterized activity for this enzyme is the recycling of adenine into adenylate nucleotides; APTs may also be involved in modification of cytokinins (Schnorr et al., 1996). Representatives of group 2 include the ACO-related cluster MD033250; ripening regulation of ACO genes has been demonstrated in other fruits, including tomato (Nakatsu et al., 1998). Additional representatives of group 2 include MD153210, corresponding to a presumed leucoanthocyanidin reductase (LAR), which catalyzes the first committed step in the biosynthesis of condensed tannins (proanthocyanidins [PAs]). These compounds act as free radical scavengers and also contribute to flavor and astringency of fruit. Group 3 includes the aforementioned progesterone 5-β-reductase-like cluster MD024600 and MD037370, very closely related with a known sarcosine/pipiecololate oxidase from Arabidopsis. Although sarcosine is not known to occur in plants, pipiecolate, a cyclic imino acid, has previously been characterized in seeds as accumulating in response to biotic and abiotic stresses (Romeo, 1998) and, thus, this gene is implicated in response to stresses.

Group 4 clusters showed frequency increases in 150-DAFB libraries compared with 126-DAFB libraries, suggesting a role in a later ripening stage. One of these, MD108530, encodes a likely thiosulfate sulfurtransferase. Sulfurtransferases/rhodaneses are a group of enzymes widely distributed in plants, animals, and bacteria that catalyze the transfer of sulfur from a donor molecule to a thiophilic acceptor substrate. This class of enzyme is not known to be regulated, and the substrate for such an enzyme in ripening fruit has not been proposed. Another group of clusters, group 9, showed statistically significant EST increases in 150-DAFB libraries compared with 87-DAFB libraries, but not with 126-DAFB libraries, and potentially represent genes that are more subtly up-regulated throughout ripening. This group includes cluster MD140930, related to 14-3-3 genes from tomato and other plants.

Finally, we identified a set of clusters, predominantly group 7 (Fig. 3), that showed frequency decreases between 87- and 126-DAFB library sources, suggesting that the corresponding genes are down-regulated associated with ripening. This set includes the XET-related cluster MD176720. Down-regulation of XET gene expression preceding ripening has previously been observed for tomato LeEXT1 (Català et al., 2000). The apparently strong expression of the respective apple gene during the rapid growth phase 87 DAFB, and persistent expression in ripe fruit, implicates this gene in multiple functions of cell wall modification. Another group 7 cluster is MD114970, highly related to acetolactate synthase (ALS; also called acetohydroxyacid synthase). ALS catalyzes the first committed step of branched-chain amino acid biosynthesis, and down-regulation of this gene may reflect a reduced requirement for protein synthesis in the maturing fruit. Branched-chain amino acids are important precursors for a variety of volatile esters in apple and other fruit (Sanz et al., 1997), and regulation of ALS may have an impact on volatile ester profiles.

To evaluate the accuracy of the predicted expression patterns of genes corresponding to these ripening-related clusters, we analyzed the mRNA levels of representative genes of the six major ripening k-means groups during a developmental sequence of fruit growth and ripening in field-grown plants. These plants were...
of a different cultivar (Jonagold) and were maintained independently of plants utilized for EST analysis. The advent of ripening in fruit was apparent by a substantial increase in ethylene concentration from about 0.5 ppm 141 DAFB to about 100 ppm 158 DAFB, and by a transient increase in CO2 production peaking around 155 DAFB (Fig. 4). All six genes analyzed showed an apparent ripening-related expression pattern substantially similar to that predicted by EST frequency analysis (Fig. 4). For example, the APT-related MD028310 (group 1), ACO-related MD033250 (group 2), and progesterone 5-β-reductase-related MD024600 (group 3) showed an increase in expression between 144 and 162 DAFB associated with the climacteric. MD033250 remained at similar levels in ripe fruit at 183 DAFB, whereas MD024600 showed a noticeable decrease. The thiosulfate sulfurtransferase-related MD108530 (group 4) showed similar expression levels at the earliest developmental stages analyzed and was markedly up-regulated late in ripening between 162 and 183 DAFB. Also as predicted, the XET-related MD176720 (group 7) showed a strong decrease in expression associated with ripening, being expressed at high levels at 144 DAFB but nearly undetectable at 162 and 183 DAFB. However, our statistical analysis did not precisely match the observed expression pattern in all cases. For example, MD028310, which was predicted to increase in expression late in ripening, was apparently expressed at similar levels 162 and 183 DAFB, and the observed increase in expression of the 14-3-3-like MD140930 (group 9) apparently occurred earlier and more abruptly than predicted from the statistical analysis.

To determine to what extent presumed ripening-related changes in expression of specific genes might be rendered inconsequential by static or reciprocal changes in expression of functionally similar genes, we analyzed the collective frequency of ESTs representing paralogous groups of clusters in libraries derived from fruit 87, 126, and 150 DAFB. We found that of the 121 clusters that showed significant EST frequency increase associated with fruit ripening (Supplemental Table IV), only seven were members of paralogous groups that collectively showed no significant frequency increase (data not shown). Similarly, of 100 clusters that exhibited EST frequency decrease during ripening, only eight were included in paralogous groups that collectively showed no significant frequency decrease (data not shown). Thus, most of the presumed changes in gene expression that we identified are likely to be biologically meaningful.

To view these results in an evolutionary context, we compared the subsets of clusters representing potential ripening-regulated genes in apple with those in tomato, another plant for which large numbers of ESTs from representational, ripening fruit-derived libraries are available (Fei et al., 2004). We identified 16,268 unique sequences containing ESTs derived from pericarp tissues of mature-green, breaker, and red-ripe tomato fruit cDNA libraries (The Institute for Genomic Research [TIGR]). We then compared the EST frequency for each cluster between these three libraries. Apple cluster sequences were used as queries to identify homologous clusters from tomato. Of the 121 apple clusters that showed statistically supported EST frequency increases associated with ripening, 76 exhibited significant homology with tomato sequences. Of these, 28 were homologous with at least one tomato cluster that similarly showed EST frequency increases during fruit ripening (Supplemental Table V), suggesting that they define a subset of genes that are ripening regulated in both apple and tomato. This class contained several representatives previously shown to be ripening regulated in tomato or other fruits. For example, the LOX cluster MD187410 shares homology with three tomato clusters from this class (Supplemental Table V); two of these represent genes previously shown to be up-regulated in ripening tomato fruit (Ferrie et al., 1994; Chen et al., 2004). Another example is the ACO cluster MD033250; this is homologous to two tomato clusters from this class, with one

Figure 4. Analysis of RNA abundance for selected genes during apple fruit growth and ripening. A, Fruit were harvested from plants maintained under field conditions and analyzed for ethylene and CO2 levels. Vertical gray lines indicate sampling days for which gene expression analysis was carried out. B, RT-PCR analysis of RNA expression for selected clusters/gens representing distinct k-means groups. Amplification primers were designed to be gene specific based on current genomic data. Results reflect the consistent outcome of two independent biological replicates. Number of ESTs for each cluster and library source is indicated, as well as significant differences in EST frequency between libraries (a, significant difference between 87- and 126-DAFB libraries; b, significant difference between 126- and 150-DAFB libraries; c, significant difference between 87- and 150-DAFB libraries).

Copyright © 2006 American Society of Plant Biologists. All rights reserved.
representing the previously characterized, ripening-regulated ACO1 (Nakatsuoka et al., 1998).

The remaining 48 apple clusters that showed sequence homology in tomato were homologous with a total of 243 tomato clusters, none of which exhibited ripening-associated EST frequency increases that were significant even at a less stringent statistical cutoff ($P \leq 0.1$; data not shown), suggesting that these may represent genes up-regulated during ripening only in apple (Supplemental Table V).

**Cultivar Comparison**

The majority of EST sequences analyzed in this study (approximately 85%) were derived from two cultivars, GoldRush and Royal Gala. GoldRush is a very firm variety characterized by a complex spicy flavor and yellowish-orange skin at maturity. This variety produces relatively low levels of ethylene, matures late in the season, and stores well (Janick, 2001). In addition, GoldRush exhibits relatively strong resistance to several diseases, such as apple scab, mildew, and fire blight (Crosby et al., 1994). Royal Gala, a Gala sport, is an early ripening, strongly red-colored fruit. We analyzed all clusters that contained ESTs from non-normalized libraries derived from young fruitlets (10 DAFB or <1 cm diameter), a developmental stage represented by comparable libraries from each cultivar, for cultivar-associated frequency bias (see “Materials and Methods”). This allowed us to define a subset of clusters that are most likely ($P < 0.05$) to represent genes or alleles showing appreciable difference in expression between cultivars.

We identified 169 clusters with EST representation biased toward GoldRush and 292 biased toward Royal Gala (Supplemental Table VI). To determine whether these differences reflected distinct biological processes that could account for cultivar-associated characteristics, we assigned biased clusters into one or more of 25 categories of predicted function (see “Materials and Methods”). Although most of these categories contained similar numbers of clusters from each cultivar (data not shown), some differences were noted. Clusters annotated as related to nucleic acid metabolism were derived predominantly from Royal Gala, and this may reflect an extension in Royal Gala of the period of cell division activity that typically occurs early in apple fruit development. In contrast, clusters annotated as related to response to stress, biotic stimulus, or abiotic stimulus were derived predominantly from GoldRush (Fig. 5), potentially reflecting a relatively higher basal level of activity of the associated response pathways in GoldRush.

We analyzed the results for potential difference in expression of genes with specific roles in fruit quality attributes and found several candidates. One example is the XET-related cluster MD247820, which is represented at high frequency in a GoldRush young fruit library (20 of 7,736 ESTs [0.26%]) but lacks any representation from Royal Gala young fruit libraries. Moreover, this cluster did not contain ESTs derived from any other Royal Gala library, suggesting that the corresponding gene is expressed or exists only in GoldRush. We were also unable to detect any closely related (E < -12) sequence, or any sequences annotated as XET-like or included in GH16, in young fruit libraries from Royal Gala (data not shown). A single XET-related sequence (MD176720) is observed in Royal Gala libraries derived from 87-DAFB fruit, where the ESTs represent nearly 2% of the total EST pool, and EST representation decreased markedly ($P < 0.001$) between 87 and 126 DAFB. MD176720 also contains EST representatives from GoldRush, showing that it is not a Royal Gala-specific gene, but these ESTs originate only from nonfruit libraries (data not shown). XET-like genes have been implicated in both fruit growth and softening, and we speculate that a difference in expression of this gene early in development may contribute to the distinct textural qualities of the mature GoldRush and Royal Gala fruit.

Flavonoids are important secondary metabolites in apple and contribute to antioxidant capacity and pigmentation. We found four clusters tied to flavonoid biosynthesis that were overrepresented in GoldRush: MD245860 and MD244690, related to leucoanthocyanidin dioxygenase (LDOX); MD013330, related to anthocyanidin reductase (ANR); and MD060870, related to anthocyanidin-3-glucoside rhamnosyltransferase. ANR and LAR participate in synthesis of flavan-3-ol monomers required for formation of PA polymers (i.e. tannins), utilizing anthocyanidin and leucoanthocyanidin,

![Figure 5. Comparison of frequency of annotations within four functional categories for clusters showing significant differences in EST representation from Royal Gala- or GoldRush-derived libraries. Functional annotation for each cluster utilized the corresponding GO Slim classification for the closest Arabidopsis homolog. Number of annotations is expressed as a percentage of total annotations for all significantly different clusters.](image-url)
respectively, as substrates (Xie et al., 2003). Anthocyanidin is produced from leucoanthocyanidin by LDOX and is also a precursor of anthocyanin and its derivatives. In grape (Vitis vinifera), tissue and temporal-specific regulation of ANR and LAR genes strongly influences PA accumulation and composition during berry development (Bogs et al., 2005). The ANR cluster MD013330 was strongly represented in GoldRush but absent in Royal Gala in libraries derived from young fruit tissues. Also, we found numerous ESTs for this cluster in nonfruit-derived libraries of Royal Gala, indicating that it does not represent a GoldRush-specific gene. Based on these observations, we speculate that, early in development, GoldRush and Royal Gala may utilize different pathways for the synthesis of flavan-3-ol monomers. Because anthocyanidin serves as a precursor for synthesis of both PAs and anthocyanins, enhanced conversion to PAs by ANR may contribute to the relatively low levels of anthocyanin evident in mature GoldRush fruit. The strong representation in GoldRush of rhamnosyltransferase, which is involved in production of various anthocyanin derivatives (Nakajima et al., 2005), may further contribute to color differences between these varieties.

Analysis of Genes Potentially Involved in Ester Biosynthesis in Ripening Fruit

We applied our EST frequency analyses to the components of biochemical pathways likely to be involved in the generation of precursors for volatile esters in ripening fruit. Ester formation in apple is largely confined to the fruit tissue and, more specifically, the skin (Knee and Hatfield, 1981). It has been demonstrated that the LOX and β-oxidation pathways permit the catabolism of fatty acids during fruit ripening to supply precursors for ester formation (Rowan et al., 1999). Consequently, the expectation is that gene expression for some of the enzymes involved in these pathways will increase during ripening and/or be localized within fruit skin tissues. We identified clusters that contained at least one EST from fruit-derived tissues and that showed strong sequence homology with biochemically characterized enzymes or enzyme classes potentially involved in ester production, including acyl-CoA dehydrogenases, acyl-CoA oxidases, enoyl-CoA hydratases, 3-ketoacyl-CoA thiolases, acyl carrier proteins, malonyl-CoA:ACP transacylases, LOXs, allene oxide synthases, and alcohol dehydrogenases (ADHs; Sanz et al., 1997; Baker et al., 2006). We analyzed these clusters for statistically significant frequency differences between fruit and nonfruit-derived tissue sources and among ripening stages. In addition, we identified those clusters that contained statistically significant enrichment for ESTs derived from libraries from skin of ripe fruit (150 DAFB), where ester biosynthesis predominates, relative to cortex (Table III).

We found a total of 49 clusters collectively representing these genes (data not shown), with 10 showing EST frequency bias among the sources analyzed (Table III). In addition to the acyl-CoA synthetase cluster MD042900

<table>
<thead>
<tr>
<th>Enzyme Class/Cluster</th>
<th>No. of ESTs (%)</th>
<th>Derived Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonfruit</td>
<td>Fruit All Stages</td>
</tr>
<tr>
<td>β-Oxidation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enoyl-CoA hydratase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD011420</td>
<td>0 (0)</td>
<td>16 (0.04)</td>
</tr>
<tr>
<td>MD157240</td>
<td>5 (0.009)</td>
<td>14 (0.035)</td>
</tr>
<tr>
<td>Fatty acid metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acyl-CoA synthetase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD042900</td>
<td>3 (0.005)</td>
<td>29 (0.073)</td>
</tr>
<tr>
<td>MD123520</td>
<td>10 (0.02)</td>
<td>14 (0.035)</td>
</tr>
<tr>
<td>MD216630</td>
<td>0 (0)</td>
<td>5 (0.013)</td>
</tr>
<tr>
<td>Acyl carrier protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD045430</td>
<td>3 (0.005)</td>
<td>5 (0.013)</td>
</tr>
<tr>
<td>Malonyl-CoA:ACP transacylase</td>
<td>2 (0.004)</td>
<td>9 (0.023)</td>
</tr>
<tr>
<td>LOX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD187410</td>
<td>1 (0.002)</td>
<td>22 (0.056)</td>
</tr>
<tr>
<td>Amino acid catabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD075410</td>
<td>3 (0.005)</td>
<td>12 (0.03)</td>
</tr>
<tr>
<td>MD142530</td>
<td>1 (0.002)</td>
<td>20 (0.05)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Increase in 126 DAFB relative to 87 DAFB.  
<sup>b</sup>Increase in 150 DAFB relative to 87 DAFB.  
<sup>c</sup>Decrease in 150 DAFB relative to 126 DAFB.
and LOX cluster MD187410, seven other clusters were overrepresented in fruit tissues relative to vegetative tissues (MD111420 and MD157240, representing potential enoyl-CoA hydratases; MD216630, representing an additional potential acyl-CoA synthetase; MD177270, representing a potential malonyl-CoA:ACP transacylase; and MD075410 and MD142530, representing potential ADHs). In addition to MD187410, the two acyl-CoA synthetase clusters MD042900 and MD123520 showed EST frequency increases associated with ripening. We found that seven clusters from five gene groups showed overrepresentation in skin tissues relative to cortex: the enoyl-CoA hydratase clusters MD111420 and MD157240, the acyl-CoA synthetase cluster MD042900, the acyl carrier protein cluster MD045430, the malonyl-CoA:ACP transacylase cluster MD177270, and the ADH clusters MD075410 and MD142530. We did not observe significant ripening-related or skin-associated overrepresentation of ESTs representing the alcohol acyl transferase (AAT) clusters present in the libraries analyzed (data not shown). This observation is not inconsistent with expression data in studies by Defilippi et al. (2005) and Souleyre et al. (2005), demonstrating significant levels of AAT gene expression even in nonripe apple fruit tissues.

We analyzed mRNA levels of genes corresponding to the LOX cluster MD187410 and acyl-CoA synthetase cluster MD042900 in relation to the ripening-related production of volatile esters in fruit from field-grown plants. In these fruit, substantial amounts of volatiles were first detected 134 DAFB, increased rapidly between 141 and 155 DAFB, peaked at 162 DAFB, and declined thereafter (Fig. 6A). RNA of the LOX gene represented by MD187410 was barely detectable by reverse transcription (RT)-PCR at 123 DAFB, increased in abundance between 123 and 144 DAFB before the onset of significant volatile production, peaked at 162 DAFB, and declined thereafter (Fig. 6A). RNA for the acyl-CoA synthetase gene was easily detectable at 123 DAFB, increased in abundance markedly by 144 DAFB, and declined thereafter; this pattern is also predicted for this cluster, which is included in ripening k-means group 3.

We also analyzed mRNA levels of representative genes predicted to be expressed preferentially in skin relative to cortex tissues of ripe fruit. We confirmed our predictions for the enoyl-CoA hydratase cluster MD111420, the acyl-CoA synthetase cluster MD042900, the acyl carrier protein cluster MD045430, the malonyl-CoA:ACP transacylase cluster MD177270, and the ADH cluster MD142530. Also consistent with predictions based on EST frequency, the LOX cluster MD187410 was found to be expressed to relatively higher levels in cortex compared with skin tissues. The malonyl-CoA:ACP transacylase cluster MD177270 was apparently expressed to similar levels in both samples, although it was predicted to be expressed to higher levels in skin tissues.

**DISCUSSION**

In this study, we applied the technique of EST frequency analysis to apple, a crop for which no additional genomic resources are currently available. Although these predictions of gene expression based on EST frequency serve as an excellent entry point to studies of fruit molecular biology, this type of study is subject to numerous caveats. Identification of the most abundant ESTs, which should represent the most active genes in these tissues, is highly sensitive to artifacts such as differential amplification of cDNAs during library preparation and contamination of libraries with abundant organellar DNAs, highly repetitive
DNA in the nuclear genome, and microbial nucleic acids. We found that a considerable fraction of the most highly represented apple clusters corresponded to known mitochondrial- or plastid-encoded genes. Although the most likely explanation is contamination by organellar DNAs or RNAs, it remains possible that at least a portion of these sequences may be derived from organellar sequences integrated into the genome, as has been observed for Arabidopsis (Arabidopsis Genome Initiative, 2000). Differences in frequency of specific ESTs between library sources, which should represent changes in absolute mRNA abundance, are influenced by factors such as selective mRNA degradation and conversion to cDNA, cloning efficiency, and disproportional changes during library preparation or amplification.

Additional caveats apply to this study where library construction, sequencing, and submission were outside of the investigators’ control. For tree species such as apple, tissue sources will typically be field derived and thus subject to a variety of biotic and abiotic stresses, and that may not be apparent at the time when tissues are collected. In addition, it is not unlikely that absence of ESTs representing some genes reflects selective withholding from submission by investigators, rather than low transcript abundance. However, in spite of the numerous potential caveats, we were able to predict with some accuracy the temporal and/or tissue-associated expression behavior in 13 of the 14 cases attempted, even though the plants subjected to gene expression analysis were a different cultivar than those subjected to EST analysis and were grown in distinct climates and seasons. Rigorous standardization and extensive documentation of tissue sourcing and experimental procedures associated with EST sequencing and submission will likely further expand the utility of this type of analysis.

We also note that the application of EST frequency analysis is limited to proportional (representational) libraries. Many EST sequencing efforts utilize normalization techniques to increase library complexity, precluding this type of analysis. We noted that the 50,721 sequences derived from four normalized GoldRush libraries represented 21,565 unique sequences, whereas the 74,914 sequences derived from 33 non-normalized Royal Gala libraries represented 23,050 unique sequences (data not shown). Therefore, at least in the case studied here, exhaustive sequencing of a few normalized libraries apparently provided some advantage over limited sequencing of diverse and specific cDNA libraries, but at the expense of losing representational data.

We demonstrated the utility of this approach by analyzing a small subset of genes potentially involved in the synthesis of fatty acid precursors for the volatile esters that contribute to aroma and flavor. Although it is well known that levels of fatty acids accumulate during apple fruit ripening (Meigh and Hulme, 1956; Song and Bangerth, 2003), specific genes that participate in this response have not previously been identified. This is due both to the difficulty of targeting specific genes for study within complex gene families and inability of investigators to easily access organized EST data. For example, of the numerous potential LOX-related genes that can be inferred from EST data (Newcomb et al., 2006; this study), we identified the LOX gene represented by MD187410 as potentially important in volatile ester formation in ripening fruit. To complement this investigation, we composed an Internet-based database of apple EST information supplemented with search and analysis tools (http://genomics.msu.edu/fruitdb) that should serve as a useful resource for studies of fruit biology as EST data are continuously updated. This resource allows for anonymous downloading of publicly available ESTs, as well as derived cluster sequences. The database is complemented with additional information linked to each cluster or singleton, including results of BLAST homology searches of comprehensive nonredundant or Arabidopsis sequences. The number, identification, and library source information for ESTs used to build each cluster can be viewed, as well as alignments of ESTs within clusters. Additional functions allow BLAST searches of actual and translated cluster or singleton sequences for homology to nucleotide or amino acid sequence queries and Boolean searches of BLAST homology reports for each cluster or singleton.

Analysis of ESTs is likely to continue to play a dominant role in genome characterization, especially in organisms where other genomics technologies are difficult or intractable, and in organisms of relatively minor collective importance. As methods for high-throughput sequence analysis become more accessible and affordable, EST data for minor crops should become increasingly abundant. Many of these data will likely be in the form of relatively small datasets and heterogeneous with respect to cultivar/genotype analyzed, tissue sources, growth conditions, and methods for library construction. Our results show that this type of analysis can be informative even where based on heterogeneous sources.

**MATERIALS AND METHODS**

**Sequence Sources**

A total of 198,684 apple (*Malus domestica* Borkh) sequences were identified in the EST and nonredundant nucleotide databases at the NCBI (http://www.ncbi.nlm.nih.gov). These sequences consisted of 198,068 ESTs and 616 cDNAs, including 394 full-length cDNAs, and are cataloged at http://genomics.msu.edu/fruitdb (version 3.0). k-means analyses and P-value calculations utilized data in version 2.0 of this database (159,254 sequences). A minimally redundant EST set from tomato (*Lycopersicon esculentum* Mill.) was obtained from the Tomato Gene Index of TIGR (http://www.tigr.org/tigr-scripts/tl/T_index.cgi?species=tomate; Quackenbush et al., 2000).

**Sequence Clustering**

Apple sequences were clustered using StackPACK (version 2.2; http://www.egenetics.com; Miller et al., 1999) using default settings for EST assembly. Vector artifacts and repetitive sequences were masked using CrossMatch (http://www.phrap.org). Masked sequences were clustered based on their similarity or amplification.
relative similarity (≥96% identity) over a window of 150 nucleotides) using a word-based greedy clustering algorithm (Hidée et al., 1994). Loose but related clusters were further aligned and assembled using PHRAP (http://www.phrap.org). We found that this step improved alignment quality by generating particularly distinct sequences as singletons and highly related sequences as subclusters. The subclusters were further analyzed for assembly artifacts and alternative expression forms using the CRAW and CONTIGPROC alignment analysis tools (Chou and Burke, 1999) to produce complete sequence (contigs). The complete catalog of clusters discussed in this study is available at http://genomics.msu.edu/fruitdb (Apple Database version 2.0).

For identification of putative paralogs of presumed highly represented genes in fruit tissue, we used tBLASTx and a single-linkage method to define groups of sequences that exhibited significant similarity (E < 1e–40). The sum of EST counts for all members in the group was considered to be the EST frequency for the paralogous group.

For comparative analysis of ripening with tomato, we utilized the TIGR Tomato Gene Index (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=tomato), version 10.1, containing 162,621 ESTs and 1,587 cDNA sequences.

Sequence Similarity Searches

Similarity searches were carried out with the stand-alone BLAST programs (Altschul et al., 1997) using executable copies obtained from the NCBI (version 2.2.2). Sequence datasets were formatted first as a BLAST-searchable database file using Formatdb (ftp://ftp.ncbi.nih.gov/blast/documents/formatdb.html). Searches were performed through comparisons of database protein sequences (BLASTX) and translation of database nucleotide sequences with translations of nucleotide queries (tBLASTx; Altschul et al., 1997). Protein query sequences, including six-frame translations of ESTs, were filtered with SET (Wootten and Federhen, 1996). Custom Perl scripts and relational databases Microsoft Access and Excel were used to automate searches on large sets of query sequences and to extract summary information. Discussions of the degree of relatedness between sequences in the text followed an arbitrary assignment: “related,” E ≤ 1 × 10e–12; “strongly related,” E ≤ 1 × 10e–25; and “very strongly related,” E ≤ 1 × 10e–40. For functional grouping of genes corresponding to fruit EST-overrepresented or fruit EST-underrepresented clusters, we utilized BLASTx to identify the closest expressed sequence in Arabidopsis (Arabidopsis thaliana) and generated the corresponding GO Slim classification by mapping plant GO Slim to The Arabidopsis Information Resource (TAIR) Arabidopsis GO gene associations, using Perl script map2slim.pl (available at http://www.geneontology.org). Only matches with E values less than 1e-10 were included in the analysis. To avoid inconsistency and error associated with manual or alternative-source annotation, we ignored all clusters without an Arabidopsis relative, even if they exhibited strong similarity with other reported sequences. Of the 1,039 clusters that showed differential representation from fruit and nonfruit-derived libraries, 622 (59%) were selected for functional classification; of the remainder, 78 exhibited significant similarity only with non-Arabidopsis sequences. Where genes were classified into multiple GO Slim categories, each classification was treated separately in the analysis to prevent bias associated with manual selection.

Digital Expression Analysis

Digital analyses of gene expression were performed based on AC statistics (Audic and Claverie, 1997) for pairwise comparisons and general chi-square tests for multiple cDNA library comparisons (Romualdi et al., 2001). P values for the statistical tests were calculated using IDEG6 (Romualdi et al., 2001). To select functional groups of genes potentially enriched in fruit tissues, we applied an AC test statistic of P < 0.005 and a proportional frequency difference of at least 2-fold between the pooled libraries. The raw EST counts were normalized proportionally to the total number of ESTs in the pools. k-means clustering was performed using CLUSTER (Eisen et al., 1998) on log-normalized data and k-means groups were visualized with TREEVIEW (http://rana.lbl.gov/EisenSoftware.htm).

Plant Materials and Analysis

For analysis of mRNA levels during fruit ripening, fruit cv. Jonagold was harvested periodically from trees maintained under field conditions at the Michigan State University Clarksville Horticultural Station. Fruit was allowed to acclimate to ambient laboratory conditions for 24 h before analysis.

Monitoring of fruit ethylene production, respiration, and volatile ester bio-synthesis was carried out essentially as described by Jayanty et al. (2002). Primers used in PCR amplifications are cataloged in Supplemental Table VII. Because of a lack of a gene known to be constitutively expressed during apple fruit ripening, PCR reactions were standardized by using equivalent amounts of a cDNA template.

ACKNOWLEDGMENTS

We thank Curtis Wilkerson for additional bioinformatics support and members of the van Nocker group for helpful critiques.

Received March 27, 2006; revised May 12, 2006; accepted May 16, 2006; published July 6, 2006.

LITERATURE CITED


Chou A, Burke J (1999) CRAWview: for viewing splicing variation, gene families, and polymorphism in clusters of ESTs and full-length sequences. Bioinformatics 15: 376–381


Copyright © 2006 American Society of Plant Biologists. All rights reserved.


Meigh DF, Hulme AC (1956) Fatty acid metabolism in the apple fruit during the respiration climacteric. Phytochemistry 4: 865–871


