Available online at www.sciencedirect.com
ScienceDirect
Gene 402 (2007) 40–50
GENE
www.elsevier.com/locate/gene

Generation of ESTs in *Vitis vinifera* wine grape (Cabernet Sauvignon) and table grape (Muscat Hamburg) and discovery of new candidate genes with potential roles in berry development

Fred Y. Peng [a], Karen E. Reid [a], Nancy Liao [b], James Schlosser [a], Diego Lijavetzky [c], Robert Holt [b], José M. Martínez Zapater [c], Steven Jones [b], Marco Marra [b], Jörg Bohlmann [d], Steven T. Lund [a,*]

[a] *Wine Research Centre, Faculty of Land and Food Systems, University of British Columbia, 2205 East Mall, Vancouver, British Columbia, Canada V6T 1Z4*
[b] *Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 100-570 West 7th Ave, Vancouver, British Columbia, Canada V5Z 4S6*
[c] *Departamento de Genética Molecular de Plantas, Centro Nacional de Biotecnología, CSIC, Campus de la Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain*
[d] *Michael Smith Laboratories, University of British Columbia, 301-2185 East Mall, Vancouver, British Columbia, Canada V6T 1Z4*

Received 8 March 2007; received in revised form 26 June 2007; accepted 17 July 2007
Available online 31 July 2007
Received by I.B. Rogozin

Abstract

We report the generation and analysis of a total of 77,583 expressed sequence tags (ESTs) from two grapevine (*Vitis vinifera* L.) cultivars, Cabernet Sauvignon (wine grape) and Muscat Hamburg (table grape) with a focus on EST sequence quality and assembly optimization. The majority of the ESTs were derived from normalized cDNA libraries representing berry pericarp and seed developmental series, pooled non-berry tissues including root, flower, and leaf in Cabernet Sauvignon, and pooled tissues of berry, seed, and flower in Muscat Hamburg. EST and unigene sequence quality were determined by computational filtering coupled with small-scale contig reassembly, manual review, and BLAST analyses. EST assembly was optimized to better discriminate among closely related paralogs using two independent grape sequence sets, a previously published set of *Vitis* spp. gene families and our EST dataset derived from pooled leaf, flower, and root tissues of Cabernet Sauvignon. Sequence assembly within individual libraries indicated that those prepared from pooled tissues contributed the most to gene discovery. Annotations based upon searches against multiple databases including tomato and strawberry sequences helped to identify putative functions of ESTs and unigenes, particularly with respect to fleshy fruit development. Sequence comparison among the three wine grape libraries identified a number of genes preferentially expressed in the pericarp tissue, including transcription factors, receptor-like protein kinases, and hexose transporters. Gene ontology (GO) classification in the biological process aspect showed that GO categories corresponding to 'transport' and 'cell organization and biogenesis', which are associated with metabolite movement and cell wall structural changes during berry ripening, were higher in pericarp than in other tissues in the wine grape studied. The sequence data were used to characterize potential roles of new genes in berry development and composition.
© 2007 Elsevier B.V. All rights reserved.

Keywords: cDNA library normalization; Gene discovery; Sequence quality; Assembly optimization; Unigene annotation; Gene ontology

1. Introduction

The cultivated grapevine (*Vitis vinifera* L.) is a fruit crop of enormous economic importance with over eight million hectares planted in vineyards worldwide. Grapes are produced for fresh fruit, juice, raisins, and transformed into high value-added products such as wines and spirits. Table grapes and wines represent a considerable share of the economy in many grape and

Abbreviations: bp, base pair(s); EST, expressed sequence tag; GO, gene ontology; kb, kilobase(s) or 1000 bp; Mb, megabase(s) or 1 million bp; ORF, open reading frame.

* Corresponding author. Wine Research Centre, Faculty of Land and Food Systems, University of British Columbia, 230-2205 East Mall, Vancouver, British Columbia, Canada V6T 1Z4. Tel.: +1 604 822 5708; fax: +1 604 822 5143.
E-mail address: stlund@interchange.ubc.ca (S.T. Lund).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.07.016

wine-producing countries. Ripening in grape berries is non-climacteric (Giovannoni, 2001; Adams-Phillips et al., 2004) and the signaling pathways governing ripening onset as well as the metabolism of compounds important for flavor are poorly understood at the molecular level (Lund and Bohlmann, 2006). Thus, grapevine as a plant experimental system that combines substantial economic value along with its physiological characteristics has been gaining attention, despite being complicated by the biology of the species, most notably its perennial growth habit.

The quality of table grape, grape juice, and wine is fundamentally dependent on healthy, high quality fruit. Fruit quality is determined by the genotypic component of the cultivar as well as environmental and cultural management conditions. Contrary to the general trend in annual crops, most improvements in grapevine and wine production rely on technological developments in agronomic (viticultural) and enological practices. Relatively little has been done at the molecular level to exploit the genetic makeup of grapevine in comparison to other economically important crops. Currently, with the development of powerful tools for genetic manipulation and the availability of three nearly fully-sequenced plant reference genomes, Arabidopsis (Arabidopsis Genome Initiative, 2000), rice (*Oryza sativa*; Goff et al., 2002; Yu et al., 2002; International Rice Genome Sequencing Project, 2005), and poplar (*Populus trichocarpa*; Tuskan et al., 2006), genomics-based approaches hold great promise for molecular breeding of grape varieties with novel or improved quality traits. Sequencing of the grapevine genome, which was estimated to be ~500 Mb (Lodhi and Reisch, 1995; Moser et al., 2005), has been hindered by the abundant heterozygosity inherent in *Vitis* genotypes. Alternately, sequencing of cDNAs to generate an EST database offers a cost-effective route to obtain sequence information of transcribed genes representing a significant amount of the gene space in the genome (Adams et al., 1991). By randomly selecting cDNA clones for single-pass sequencing, an EST sequence database can be established in a relatively inexpensive manner, which is particularly valuable for an organism without a sequenced genome. EST sequence data and their corresponding physical clones can serve as a resource for functional genomics studies, including microarrays and real-time PCR (e.g. Aharoni and Vorst, 2002; Pacey-Miller et al., 2003; Terrier et al., 2005; Waters et al., 2005; Ralph et al., 2006; Reid et al., 2006). EST sequences can also be translated into a putative protein database for peptide identification in mass spectrometry-based proteomics studies (e.g. Lippert et al., 2005), as well as used for the development of genetic markers (e.g. Rungis et al., 2005; Lamoureux et al., 2006).

Recently, the importance of grapevine in plant genomics has been reflected by several grape EST projects that were initiated around the world. In 2001, there were only ~400 *V. vinifera* ESTs in GenBank (da Silva et al., 2005; Moser et al., 2005). Since then, this figure has risen drastically; as of July 1, 2006, 195,434 *V. vinifera* ESTs were present in the GenBank dbEST division. When our EST project was launched in 2004, there were three publications based on analyses of grape ESTs, generated from Shiraz, Chardonnay, and Purple Cornichon cultivars on a much smaller scale, ranging from 275 to 4270 ESTs (Ablett et al., 2000; Terrier et al., 2001; Pacey-Miller et al., 2003). Most of these ESTs represent cDNAs sequenced from their 5′ ends, with the exception of the 275 ESTs reported by Terrier et al. (2001) which were sequenced from their 3′ ends. More recently, Moser et al. (2005) reported 8647 5′ ESTs from the cultivars Pinot Noir and Regent. da Silva et al. (2005) performed a comprehensive analysis of the 146,075 grape ESTs and mRNAs deposited in GenBank as of September 30, 2003, from multiple *Vitis* species and revealed a unigene set of 25,746 contig and singleton sequences for *V. vinifera*. The authors estimated that their unigene set might have covered upwards of 67% of the grape transcriptome, assuming that the grapevine genome contains ~38,000 genes (da Silva et al., 2005); however, this number of unique sequences derived via computational clustering and assembly of ESTs likely represents an overestimate of the grape genes identified, since the ESTs analyzed represent cDNAs sequenced from both 5′ and 3′ ends. A higher false positive (and false negative) rate was previously observed with 5′ EST clustering (Pratt et al., 2005; Park et al., 2006). For example, using the information available for the Arabidopsis genome, Wang et al. (2004) demonstrated that the rate of incorrectly separating ESTs from the same gene into two or more clusters is 30% with 5′ ESTs and 3% with 3′ ESTs. As possible evidence that unigenes derived from EST assembly may not accurately represent gene space, there are 110,779 unique sequences in the Arabidopsis Gene Index (AtGI) Release 13.0 (June 16, 2006; http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=arab), yet the latest annotation of Arabidopsis genome only predicted 32,041 genes (April 2007; http://www.arabidopsis.org). Alternate splicing (AS) and alternate transcription initiation (ATI) events, which can account for this discrepancy, are not as common in plants as in mammals (Brett et al., 2002; Iida et al., 2004; Nagasaki et al., 2005); therefore, a large portion of the genes in the grapevine genome is likely not yet identified, supporting additional EST sequencing experiments reported here.

We have generated 77,583 ESTs from cDNA libraries in *V. vinifera* cv. Cabernet Sauvignon (CS; a wine grape) and cv. Muscat Hamburg (MH; a table grape) in order to gain further insight into the grapevine transcriptome and, more specifically, discover new pericarp transcripts. In this study, we report on the generation and analysis of this EST collection with an important focus on EST sequence quality control and assembly optimization. We generated a non-redundant set of sequences for each tissue category and a full assembly among all tissues sampled in CS. Our assembly approach was novel for grapevine in two respects. First, cultivar-specific assemblies were performed in order to attempt to preserve sequence polymorphisms between CS and MH. Second, assemblies by source tissue within cultivars were carried out in order to facilitate bioinformatic delineation of likely paralogous unigenes. Comparative analysis against VvGI Release 5.0 (Quackenbush et al., 2000) and the NCBI UniGene database (Wheeler et al., 2006) for grape indicated that all libraries contributed novel sequences. The EST sequences described

Table 1
Description of cDNA libraries used for EST sequencing

Cultivar	Tissue source[a]	Library name[b]	Construction strategy
Cabernet Sauvignon	Pericarp	CSPCNN	Regular non-normalized
		CSPCNOCot2.5	Normalized to Cot2.5
		CSPCNOCot5	Normalized to Cot5
		CSPCNOCot7	Normalized to Cot7
	Seed	CSSDNOCot5	Normalized to Cot5
		CSSDNOCot7	Normalized to Cot7
		CSSDNOCot5-Alb	Normalized to Cot5 and albumins subtracted
	Non-berry (leaf, flower, root)	CSNBNN	Regular non-normalized
		CSNBNOCot5	Normalized to Cot5
Muscat Hamburg	Pre-veraison berry with post-anthesis flower and seed	MHPRNOCot5	Normalized to Cot5
	Post-veraison berry	MHPONOCot5	Normalized to Cot5

[a] All berry, pericarp, seed, and flower tissues were sampled from different developmental stages; vegetative root and leaf are each one steady-state sample. The CS pericarp libraries included pericarp tissues from eight different developmental stages ranging from fruit set through to full maturity. CS seed libraries represented eight developmental stages ranging from the 1–2 mm stage following fruit set through to seed set. The CS non-berry libraries were constructed by pooling equimolar mRNA samples prepared from root, leaf, and flower. The MH libraries were constructed from berry developmental stages including post-anthesis flowers and berries from fruit set through to the stage just prior to ripening initiation (with seeds) as well as berries from ripening initiation through to full maturity (without seeds).
[b] Libraries were named by abbreviations of cultivar, tissue source and construction strategies.

here have been deposited into the GenBank dbEST under accession numbers EC919418–EC997000.

2. Materials and methods

2.1. Plant material

Berries and seeds of different development stages, pre- and post-anthesis flowers, leaves, and roots produced through air-layering from *V. vinifera* cv. CS clone 15 were collected from a commercial vineyard in Osoyoos, BC, Canada, in 2003. Tissues were frozen immediately in liquid nitrogen, shipped to the laboratory on dry ice, and then stored at −80 °C until RNA extraction. Berries and seeds of different developmental stages, as well as post-anthesis flowers from *V. vinifera* cv. MH were collected from a commercial vineyard in Murcia, Spain, in 2005 and stored in the same manner as the CS samples prior to RNA extraction. For the fruit-derived libraries, harvest time points were carried out at bi-weekly intervals except for every 2 days during the ripening initiation period in order to capture expressed genes associated with the rapid signaling and metabolic changes occurring during this critical period.

2.2. cDNA library construction

High quality RNA was isolated according to the protocol described in Reid et al. (2006). Both non-normalized and normalized cDNA libraries were constructed from the purified mRNA. The non-normalized libraries were created using Zap Express cDNA Gigapack III Gold Cloning Kit (Stratagene, La Jolla, USA) and directionally cloned into pBK-CMV. During the construction of normalized libraries, the primary libraries were created using the cDNA Synthesis Kit (Stratagene) and directionally cloned into pBluescript II SK+. Libraries were then normalized to different Cot levels using a protocol adapted from Bonaldo et al. (1996) and Soares et al. (1994) to create the final normalized libraries after their Cot levels were optimized, based on small-batch sequencing trials and analyses of within-library sequence redundancy rate (data not shown). Clone inserts shorter than 300 bp were discarded. The 11 cDNA libraries used for sequencing of ESTs in this study are described in Table 1.

2.3. EST sequencing and assembly

Sequence data were obtained primarily using universal M13 Forward primer (5′ GTT TTC CCA GTC ACG AC 3′). Poly(T) anchored primer (5′-T21[C/G/A]-3′) was used when poly(A) tails were identified as being excessively long. Some 5′ end sequencing was performed using universal M13 Reverse primer (5′ CAG GAA ACA GCT ATG AC 3′). Sequencing was carried out on a 3730XL DNA Analyzer (Applied Biosystems, Foster City, CA, USA) at Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada. DNA sequence chromatograms were processed using the phred software (Ewing and Green 1998; Ewing et al., 1998). Subsequently, sequences were vector-trimmed using Cross_Match software in the phrap package (http://www.phrap.org/) and quality-trimmed according to the high quality contiguous region determined by phred. The sequencing process was managed by a laboratory information management system (LIMS) and all sequences were stored in a MySQL relational database to facilitate the assemblies and subsequent bioinformatic analyses by an EST analysis pipeline written in Perl (Fig. 1). To identify unigenes, CAP3 was used to assemble ESTs into contigs using the parameters of 60 bp overlap length and 95% overlap identity (Huang and Madan, 1999), following CAP3 parameter optimization.

2.4. Functional annotation of ESTs and unigenes

For annotation of the ESTs and unigenes, sequences were compared using different versions of the BLAST algorithm (Altschul et al., 1997) against 24 sequence databases (Table 2). Gene ontology (GO) annotation (The Gene Ontology Consortium 2004) was generated using BLASTX (*E*-value $\leq 10^{-2}$) against the Arabidopsis proteome (ATH1.pep), and the curated

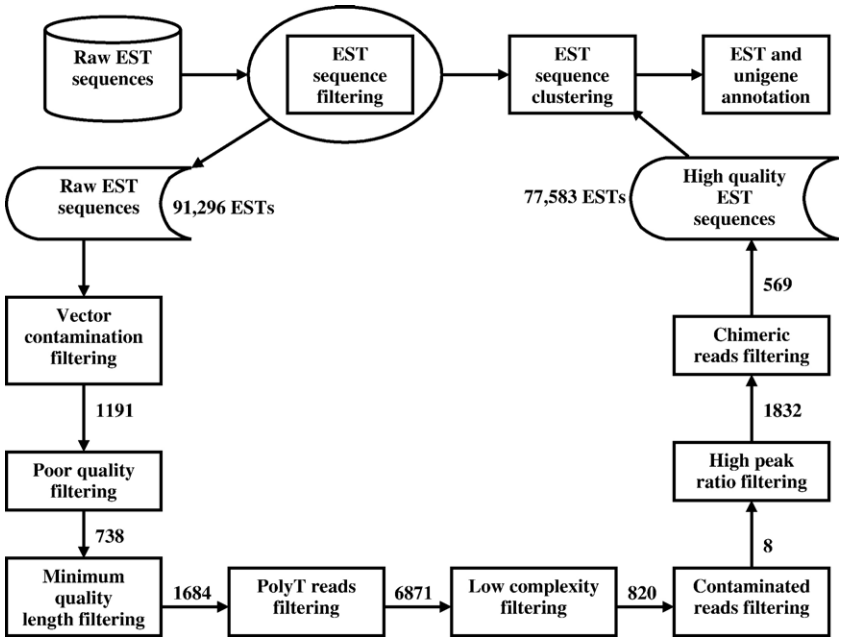

Fig. 1. Flowchart depicting EST clustering and analysis pipeline. Raw EST data generated from the ABI sequencer were entered into a MySQL database designed for EST clustering and analysis. Prior to clustering and batch annotations, filters were implemented to flag and exclude low quality ESTs and chimeric sequences from assembly. The number beside each arrow represents the number of ESTs filtered at each step.

set of Arabidopsis GO terms from The Arabidopsis Information Resource (TAIR July 22, 2006 version; Berardini et al., 2004; The Gene Ontology Consortium, 2004). Custom Perl scripts and the GO-Perl package were used for GO classification according to the GO slim terms in the biological process aspect defined by TAIR.

3. Results and discussion

3.1. Library and EST characteristics

Table 3 summarizes the characteristics of ESTs and unigenes assembled for each cultivar/tissue category. A variety of tissues from the two cultivars, CS and MH, were sampled to construct 11 cDNA libraries (Table 1). A total of 91,296 single-pass sequencing runs were performed, the vast majority of which were primed from the 3′ ends, yielding 77,583 high quality ESTs (Table 3). The sequencing success rate was ~85%. Initial sequencing in a CSPC library generated 3980 5′ ESTs, which were also included in the assembly to extend contigs or in some cases form full-length open reading frames (ORFs). ESTs submitted to the GenBank dbEST division averaged 622 bases in length with an average phred score of nearly 57. The average length of ESTs reported here is substantially longer than those previously reported for grapevine (e.g. 460 bases in Moser et al., 2005, or 527 bases in da Silva et al., 2005). For individual assemblies, CSNB had the largest EST average length of 688 bases, whereas MHBF has the smallest with 558 bases (Table 3). The vast majority of ESTs (>92%) were longer than 300 bases and approximately 88% of ESTs ranged in length from 300 to 899 bases (Fig. 2).

The 77,583 high quality ESTs submitted to the GenBank dbEST division under accession numbers EC919418–EC997000, together with the 47,271 ESTs from Thompson seedless (table grape) and Carmenere (wine grape) cultivars submitted shortly afterwards by a Chilean grape genomics consortium, have brought the total number of grape ESTs in GenBank to 320,503 (as of June 22, 2007). This number currently positions grape as highest among soft fruit species for the number of publicly available ESTs.

3.2. EST filtering and CAP3 assembly optimization

To improve EST and unigene sequence quality, we implemented several filters to identify and remove low quality ESTs, as shown in the flow chart for the EST analysis pipeline (Fig. 1). Empty reads, vector and adaptor sequences, and contaminating bacterial sequences were screened and filtered automatically. The phred program was then used to identify bases with ≥20 quality score (corresponding to ≤1% error probability; Ewing and Green, 1998), and reads that had phred20<100 (i.e. where the length of sequence with ≥20 phred score was <100 bases) were eliminated. In addition to the phred quality score, we used the trace peak area ratio value calculated by the new phred version (0.020425.c) to filter out low quality ESTs. From the trace file, the phred program calculated the ratio of the total uncalled-base peak area over the total called-base peak area in the quality region. A relatively stringent peak ratio of 0.25 (i.e. the total uncalled-base peak area is one quarter of the total called-base peak area) was implemented to filter out ESTs with high-peak ratios. A total of 1832 ESTs were discarded using this parameter.

In addition to the computational detection, we applied some manual inspection measures to further increase the quality of our EST database. Chimeric ESTs, for example, can escape many computational quality check procedures, as their quality

Table 2
The 24 sequence databases included in EST and unigene annotations

Category	Database	Data source (reference)
Comprehensive database	NR proteins	NCBI (Wheeler et al., 2006)
	Swiss-Prot	EBI (Schneider et al., 2004)
Model plant genome database	Arabidopsis (CDS)	TAIR (Rhee et al., 2003)
	Arabidopsis (peptide)	TAIR
	Arab_Swiss	EBI
	Rice	TIGR (Quackenbush et al., 2000)
Plant gene index database	Barley	TIGR
	Cotton	TIGR
	Grape	TIGR
	IcePlant	TIGR
	Lotus	TIGR
	Maize	TIGR
	Medicago	TIGR
	Potato	TIGR
	Rye	TIGR
	Sorghum	TIGR
	Soybean	TIGR
	Strawberry	In-house [a]
	Tomato	TIGR
	Wheat	TIGR
Bacterium and fungus database	Agrobacterium	NCBI
	Aspergillus	NCBI
	Ecoli_k12	NCBI
	Yeast genomic	NCBI

[a] No strawberry gene index is currently available; therefore, an in-house strawberry unigene set was constructed (Supplementary data) by clustering all *Fragaria* EST and cDNA sequences, mostly from the wild strawberry *Fragaria vesca* and the modern garden strawberry *Fragaria ananassa*, downloaded from GenBank as of July 12, 2006.

scores can still be high. For suspicious contigs, we performed the following manual evaluation: Retrieve its constituent ESTs and resubmit them to the CAP3 assembler, download the.ace output file to examine their reassembly using the TIGR assembly viewer (http://www.tigr.org/tdb/tgi/software/), and finally, perform a BLAST analysis for each individual component EST. Through this process, 579 chimeric reads were flagged and excluded from contig assembly.

To assess EST sequence quality improvement gained through the phred scoring, peak area ratio, and chimeric read filters, we used the TIGR sequence validation tool, SeqClean (http://www.tigr.org/tdb/tgi/software/), to compare our EST collection with the 46,900 grape ESTs randomly retrieved from GenBank dbEST. Of the 77,583 ESTs reported in this study, 30 or <0.04% were flagged for discarding by SeqClean, compared to approximately 0.4% of the subset of grape ESTs in GenBank, were these to be used in our assemblies. Through all of these sequence quality control procedures, 13,703 of our reads were eliminated, accounting for nearly 15% of the total number of reads.

The two most critical parameters that can greatly influence the assembly output are overlap length cutoff (CAP3 default is 40 bp overlap length) and overlap percent identity cutoff (CAP3 default is '80', corresponding to 80% minimum sequence identity). Before performing the final assemblies, we tested which combination of these two parameters could best distinguish closely related grape paralogs using two independent data sets: 1) a set of known *Vitis* gene families from GenBank and 2) all ESTs in the CSNB library. Three *Vitis* gene families which have predicted full-length sequences in GenBank, isoflavone reductase-like protein (IFRL; 6 members, ifrl1 to ifrl6), stilbene synthase (STS; 3 members, st1 to st3), and 9-*cis*-epoxycarotenoid dioxygenase (NCED; 2 members, nced1 and nced2) were used for a small-scale assembly using several combinations

Table 3
Summary of assemblies and characteristics of their ESTs and unigenes

Assembly designation [a]	Assembled libraries [b]	ESTs (no.) [c]	Average EST length (nt)	Contigs (no.)	Average contig length (nt)	ESTs in contigs (no.)	Singletons (no.)	Unigenes (no.) [d]	Sequencing depth index [e]	Uniqueness (%) [f]
CSPC	CSPCNN CSPCNOCot2.5 CSPCNOCot5 CSPCNOCot7	12,050	641	1636	895	9229	2821	4457	5.6	37.0
CSSD	CSSDNOCot5 CSSDNOCot7 CSSDNOCot5-Alb	11,888	599	1542	844	8486	3402	4944	5.5	41.6
CSNB	CSNBNN CSNBNOCot5	26,898	688	5060	903	20,194	6704	11,764	4.0	43.7
CSFA	All above libraries	50,834	656	7250	940	42,554	8280	15,530	NA [g]	NA
MHBF	MHPRNOCot5 MHPONOCot5	26,750	558	4462	781	20,998	5752	10,214	4.7	38.2

[a] Assemblies were performed among libraries listed in Table 1 according to cultivar/tissue categories, and were designated by four letters — the first two letters for cultivar abbreviation (CS for Cabernet Sauvignon; and MH for Muscat Hamburg) and the second two letters for tissue source abbreviation.
[b] Library names are as in Table 1.
[c] Number of ESTs is the total number of ESTs sequenced from libraries in the cultivar/tissue category.
[d] Number of unigenes is the total numbers of contigs and singletons.
[e] Sequencing depth index is calculated by dividing the number of ESTs clustered in contigs by the number of contigs.
[f] Percent uniqueness is expressed as a percentage of the number of unigenes divided by the number of ESTs.
[g] Not applicable.

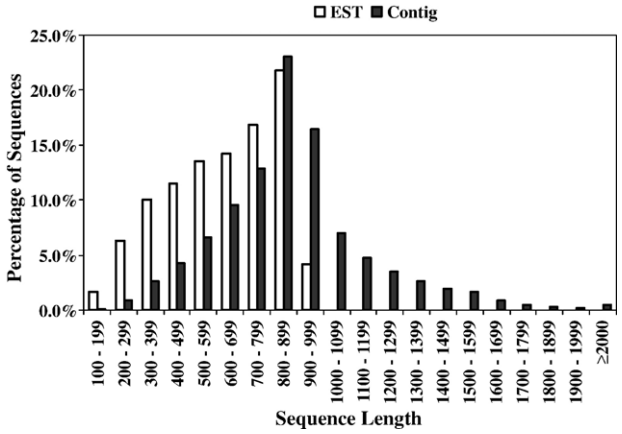

Fig. 2. Distribution of the sequence length of ESTs and contigs after assembly. All ESTs submitted to GenBank dbEST were analyzed without library distinction. Singletons after assembly were excluded for unigene length analysis.

of length and percent identity cutoff settings. We found that setting the length cutoff to either 40 or 60 bp resulted in identical assembly outputs for these gene families (under the same percent identity), but varying the percent identity cutoff affected the assembly output in different ways among the three gene families. For IFRL, all members were separated at 88 percent identity or above, but ifrl5 and ifrl6 were assembled into a contig if percent identity was 87 or lower. For STS, 87 percent identity or above separated all family members, but at 86 percent identity or lower, st1 and st2 were joined. For NCED, all of the valid CAP3 percent identity parameters (>65%) separated the two members. These results suggest that the CAP3 default parameters are not capable of distinguishing some gene family members depending upon their degree of sequence similarity. This finding strongly supported our carrying out CAP3 optimization work prior to executing the final assemblies. To discriminate among paralogous sequences that are more conserved than members in the gene families that we tested, we chose to use a relatively stringent 95 percent overlap identity cutoff value.

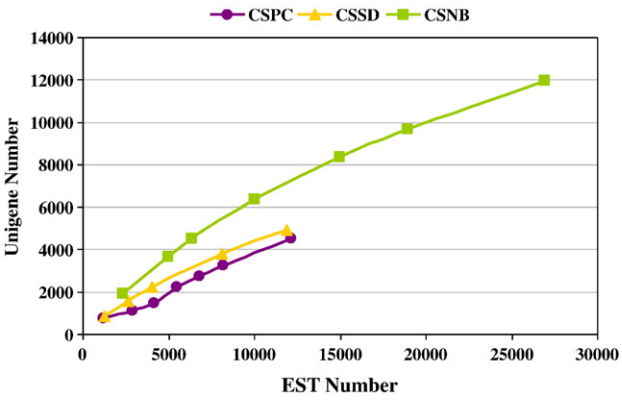

Fig. 3. The number of unigenes as a function of the number of ESTs in the three CS libraries. All data points were obtained from periodic assemblies within individual libraries as sequencing progressed. The two constituent libraries for the MHBF assembly were sequenced in two large batches after library normalization and sequencing/annotation procedures were optimized, so they were not included in this analysis.

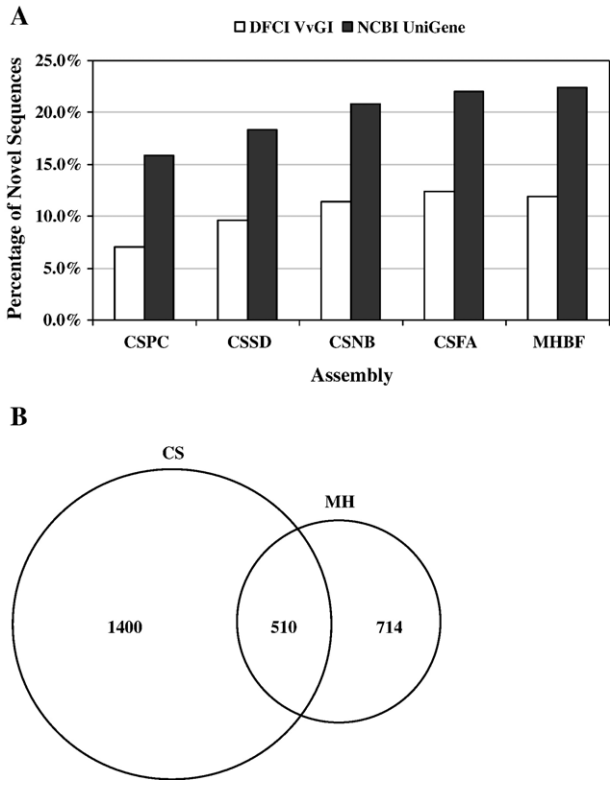

Fig. 4. Gene discovery rate in different libraries as compared to VvGI Release 5.0 and NCBI *Vitis* UniGene Build18 (A) and distribution of new sequences between CS and MH (B). Sequences that have no significant hit in BLASTN analysis (*E*-value $\leq 10^{-2}$) in VvGI or *Vitis* UniGene were considered novel sequences.

For a large-scale EST assembly test, we performed two assemblies in CSNB under two overlap length cutoff settings, 40 or 60 bp. Under the 60 bp window, 11,764 unigenes were produced, 23 unigenes more than with the 40 bp window (11,741 unigenes). This was because increasing the overlap length requirement prohibited some ESTs (e.g. ESTs having 41–59 bp overlap) from forming contigs; however, the difference in unigene numbers was not significant, suggesting that most contigs share substantial overlaps, as all cDNAs in CSNB were sequenced from the 3′ ends. By examining contigs that were assembled using the two window sizes, we found that some closely related gene family members, including different β-tubulins (TUB1, TUB6, TUB8), and vacuolar ATP synthase subunits 2 and 3, were joined into contigs when a 40 bp overlap length cutoff was used but remained separate when 60 was used. This comparison suggested that a 60 bp overlap length cutoff was better able to separate closely related paralogs in this grape EST set. Consequently, we chose to use 60 bp overlap length with a 95% overlap identity for our final assemblies. These findings suggest that careful testing of parameters in assembly software is warranted in order to increase resulting EST database quality.

3.3. Sequence assembly and unigene characteristics

EST data from the 11 cDNA libraries were condensed by cultivar/tissue categories into four EST assemblies to generate a non-redundant gene set (unigenes). Separate assemblies for

each tissue category in CS (CSPC, CSSD, and CSNB), a full assembly of the two MH libraries (MHBF), and a full assembly in the CS cultivar (CSFA) encompassing its three component libraries, were performed. After assembly, there were 4457, 4944, 11,764, and 10,214 unigenes in CSPC, CSSD, CSNB, MHBF, respectively; a full assembly of the three CS libraries yielded 15,530 unigenes (Table 3). Independent assemblies for the two cultivars aimed to preserve potential sequence polymorphisms between these genotypes. The average contig sequence length excluding singleton ESTs was 880 bases, increasing by ~260 bases in length through the assembly. Over 87% of contigs have lengths ranging from 500 to 1399 bases (Fig. 2).

The number of ESTs sequenced from each library was partly determined by the yield of novel sequences in preliminary, small-batch sequencing experiments conducted prior to large-scale sequencing in each library. There were over 11,000 ESTs in each final assembly, with CSNB and MHBF each having nearly 27,000 ESTs. Library uniqueness was assessed by calculating two related parameters — sequencing depth index and percent uniqueness. The sequencing depth index was the number of ESTs clustered into contigs divided by the number of contigs; the percent uniqueness was the number of unigenes divided by the total number of ESTs (Table 3). The sequencing depth index was smallest for CSNB at 4.0, followed by MHFB (4.7), CSSD (5.5) and CSPC (5.6). Roughly conversely, the percent uniqueness was highest in CSNB at 43.7%, followed by CSSD (41.6%), MHFB (38.2%) and CSPC (37.0%).

A periodic sequence assembly approach, performed for seven batches in the case of CSPC and CSNB, as well as for five batches in CSSD, provided an opportunity to explore the relationship between the number of unigenes and ESTs as sequencing progressed. Fig. 3 indicates that in the early stage of sequencing, the number of unigenes increased approximately linearly with the rise in EST number. As sequencing progressed deeper into each library, the rate of increase in unigene number proportional to the number of additional ESTs decreased, as expected. This trend was similar to that reported for Sorghum (*Sorghum bicolor*) cDNA sequencing (Pratt et al., 2005). The steeper slope in the number of unigenes to ESTs in CSNB as compared with CSPC or CSSD suggests greater diversity in CSNB. Fig. 3 also illustrates that more distinct transcripts could potentially be identified in each of our CS libraries if sequencing were continued, albeit at an escalating sequencing cost per unique transcript discovery.

3.4. Gene discovery rate in comparison to public grape ESTs

To assess the contribution of unique grape sequences from our cDNA sequencing, the unigene sequences from each individual assembly were compared to publicly available grape sequences. Two public data sources, the most recent *V. vinifera* Gene Index (VvGI Release 5.0 June 21, 2006) and the NCBI *V. vinifera* UniGene Build 18 were used for BLASTN evaluation under *E*-value 10^{-2} (Fig. 4A). Each of our libraries contributed novel sequences (i.e. sequences with no significant hits to public grape sequences) in comparison to VvGI Release 5.0. The percentages of novel unigenes discovered in each library were 11.4 in CSNB, 7.1 in CSPC, 9.7 in CSSD, and 11.9 in MHBF (Fig. 4A). Overall, the CS full assembly (CSFA) yielded approximately 12.4% novel sequences. MHFB and CSNB provided the highest gene discovery rate, while CSPC had the lowest. We estimate that in comparison to the publicly available sequence data, ~15% novel sequences were discovered in CS and MH, based on the BLAST parameters used.

We also compared our unigene sets with the NCBI *V. vinifera* UniGene Build 18 because UniGene requires 3′ anchoring in order for sequences to be clustered together (Pontius et al., 2003); approximately 95% of our cDNA clones were sequenced from the 3′ ends. Prior to our EST submissions, UniGene Build 18 had 15,194 unique sequences, clustered from 149,691 EST and mRNA sequences. Compared to this set of unique grape sequences, both CSFA and MHBF each had approximately 22% novel sequences (~16%, 18%, and 21% for CSPC, CSSD, and

Table 4
A selected set of grape unigene sequences that have no Arabidopsis hits annotated by BLASTN against tomato and strawberry known genes

Annotation database	Unigene identifier[a]	GenBank accession and annotation[b]	Expect value	Aligned length (bp)/identity (%)
Tomato	CSFA10463	X58885 Ethylene-forming enzyme (EFE)	2e−05	28/96.4
	CSFA12635	AY098732 TDR4 transcription factor	8e−08	219/81.3
	CSFA13220	AY294330 MADS-box protein 5	7e−14	152/84.9
	CSFA2796	AY261512 Mitogen-activated protein kinase 1 (MPK1)	1e−27	293/80.2
	CSFA3095	DQ307488 EIN3-binding F-box protein 1 (EBF1)	9e−08	84/89.3
	CSFA8042	AY940041 Symbiosis receptor-like kinase (SYMRK)	3e−38	223/83.9
	CSFA8631	AY044235 Transcription factor JERF1 (JERF1)	8e−08	87/89.7
	MHBF1727	DQ456876 Gamma-tocopherol methyltransferase (TMT)	4e−22	284/79.6
	MHBF2217	AY840092 Monoterpene synthase 2 (MTS2)	6e−13	80/86.3
	MHBF5787	AJ489278 Carotenoid cleavage oxygenase (CCO)	2e−05	23/100
Strawberry	CSFA10499	AY695817 Anthocyanidin synthase (ANS)	1e−21	172/84.9
	CSFA2613	AJ297513 Ethylene receptor (ein1)	3e−29	578/77.7
	CSFA5337	DQ087253 Leucoanthocyanidin reductase (LAR)	7e−11	127/81.1
	CSFA6613	AF401220 Transcription factor MYB1 (MYB1)	8e−10	73/84.9
	MHBF42	AJ297511 Ethylene receptor (etr1)	4e−14	88/85.2
	MHBF5220	AY679587 Protein phosphatase 2C	3e−06	91/86.8

[a] Unigene identifier is the library designation concatenated with the contig number assigned after assembly.
[b] The GenBank accession and annotation are indicated for the known tomato or strawberry hits.

CSNB alone, respectively) (Fig. 4A). This percentage is higher than when compared to VvGI Release 5.0 because the number of sequences clustered by the NCBI UniGene is lower than that of VvGI due to the requirement of UniGene for 3′ end anchoring.

Because we assembled ESTs from CS and MH separately, there was likely content overlap. In fact, our analyses indicated that 510 highly similar novel sequences in CS and MH would be clustered (Fig. 4B). Taking into account these data in BLASTN comparisons of our unigenes to VvGI Release 5.0, we produced a revised estimate of at least 2725 unique grape sequences discovered in our EST project. This gene discovery estimate was conservative, given that the expectation value cutoff in this analysis was 10^{-2} (corresponding to ~20 base match) in order to take into consideration the partial nature of ESTs and numerous 5′ ESTs in the public domain. By decreasing the *E*-value cutoff to 10^{-25} in separate CS and MH analyses, for example, we determined that the percent novel sequences in comparison to VvGI Release 5.0 increased by 2.5% and 5.4%, respectively, in each assembly.

3.5. Functional annotations of ESTs and unigenes

In addition to the common EST annotations based on the NCBI nr database or Swiss-Prot, we annotated the EST and unigene sequences using other databases, such as the Arabidopsis CDS and peptide databases, the *Agrobacterium*, *Aspergillus*, *E. coli*, and yeast databases, as well as a number of TIGR plant gene index databases including grape, tomato, and strawberry (in-house) (Table 2). EST annotations using the microorganism sequence databases were carried out in order to identify contaminating sequences for removal prior to clustering. Multiple plant gene databases were used for unigene annotations in order to improve our ability to predict potential functions, particularly with respect to berry development and composition traits. We emphasized annotations using unigenes

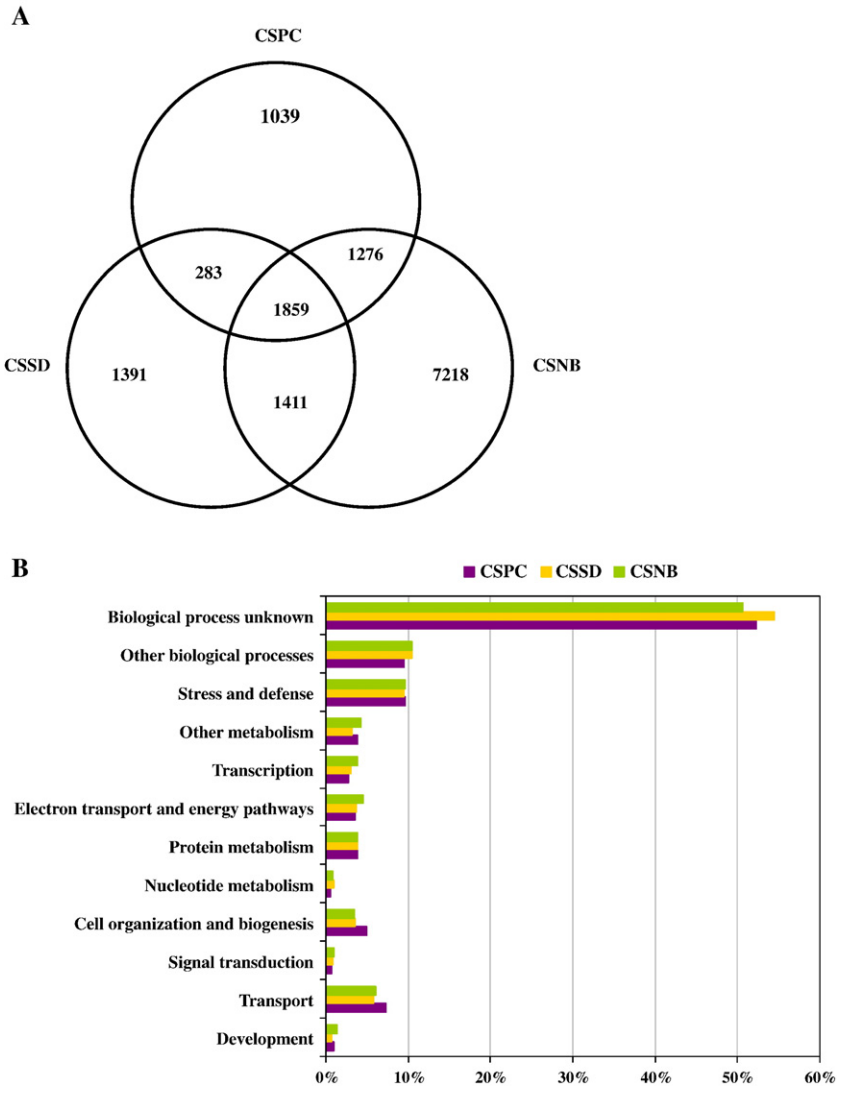

Fig. 5. Venn diagram of similar and distinct sequences (A) and gene ontology classifications in biological aspect (B) among the three CS assemblies. The assembly designations are as in Table 3.

from tomato and strawberry, since the most widely used plant models, Arabidopsis and rice, do not produce fleshy fruits. To demonstrate the utility of this approach, we annotated sequences that have no hits to Arabidopsis sequences using unigene data from tomato, a climacteric ripening species, and strawberry, a non-climacteric species more similar to grape. A total of 9837 unigenes (or 38% of total unigenes in CS and MH) had no hit against Arabidopsis cDNA sequences (*E*-value $\leq 10^{-2}$), compared with 6997 (27%) and 7822 (30%) that have no hits against the unigene sets for tomato (Fei et al., 2004) and strawberry, respectively. Table 4 shows a selection of grape genes of interest annotated using known tomato or strawberry genes, including sequences that are likely involved in signal transduction (kinases, receptors, and transcription factors) and secondary metabolism. The complete lists of sequences without hits in Arabidopsis that were annotated using tomato or strawberry unigenes are included in the Supplementary data.

3.6. Identification of transcripts unique to the CSPC assembly

Comparative sequence analyses among the three CS assemblies identified identical/similar sequences as well as assembly-specific sequences (Fig. 5A). Since the majority of the cDNA libraries were normalized and 95% of the sequences were derived from these libraries, the EST frequency obtained cannot be used for digital gene expression analyses; however, a qualitative analysis at the sequence level to identify those detected in a particular library may shed light on the molecular components underlying biological processes of interest in the tissues from which the ESTs were derived. Since the CS libraries were sequenced to similar depths, this approach seemed reasonable. For example, we determined that the sequences present in all three CS assemblies included a number of classic housekeeping genes, such as actin, cyclophilin, translation elongation factor, glyceraldehyde 3-phosphate dehydrogenase, malate dehydrogenase, polyubiquitin, and ubiquitin-conjugating enzymes. In contrast, sequences found exclusively in CSSD included five genes annotated as late-embryogenesis abundant protein, four genes encoding seed maturation proteins, two seed-specific proteins, and one 11S seed storage globulin. Sequences found specifically in CSPC, potentially indicating preferential expression in this tissue, included putative transcription factors (e.g. putative phloem transcription factor M1, WRKY family transcription factor, MYB-like transcription factor DIVARICATA), other proteins involved in signal transduction (e.g. developmentally-regulated GTP binding protein, receptor-like protein kinase, auxin-responsive family protein), hexose transporters (e.g. sugar transporter family protein, putative sucrose transporter), as well as 9-*cis*-epoxy-carotenoid dioxygenase 1 (Table 5). The MYB transcription factor belongs to a family that is known to regulate anthocyanin biosynthesis during berry development (Xie et al., 2006). Many other genes await functional characterization to clarify their roles in berry development and association with quality traits such as appearance and flavor.

For gene ontology (GO) annotation (The Gene Ontology Consortium, 2004), BLASTX analysis was performed to compare grape unigenes to an Arabidopsis proteome database (Rhee et al., 2003). To gain an overview of the distribution of annotated functions, higher-order GOslim categories were used for GO classifications (Berardini et al., 2004). Over 50% of the sequences in three CS unigene sets could not be assigned informative GO terms ('biological process unknown' or 'other biological process') (Fig. 5B). Those genes either had no significant hit to the Arabidopsis protein sequence database, its putative Arabidopsis ortholog had no GO annotation, or was annotated as 'Unknown' (e.g. the Arabidopsis hits were annotated as expressed, hypothetical, or unknown proteins). Noticeably, the GO analysis of the sequences in the three CS assemblies revealed that the 'transport' and 'cell organization and biogenesis' categories were each >1% higher in CSPC than in CSNB or CSSD (Fig. 5B). 'Cell organization and biogenesis' represents biological processes involved in the assembly and arrangement of cell structures, including the plasma membrane, cell wall, and cell envelope. The overrepresentation of these two categories in the pericarp assembly likely reflects relatively high metabolite transport and cell wall softening events occurring during berry ripening. Our GO classification on the tomato unigene set (Fei et al., 2004), revealed a similar distribution including a high percentage of 'biological process unknown' (data not shown).

Table 5
Selected set of CSPC-specific sequences and their putative functions

Functional class	Unigene identifier[a]	Putative annotation[b]
Transcription factor	CSPC1345	Putative phloem transcription factor M1
	CSPC1833	Myc-like regulatory protein
	CSPC2018	Putative transcriptional coactivator
	CSPC2127	WRKY family transcription factor
	CSPC2048	Pathogenesis-related genes transcriptional activator PTI6
	CSPC3581	MYB-like transcription factor DIVARICATA
	CSPC37	Auxin-responsive family protein
Signaling protein	CSPC114	Signal recognition particle 9 kDa protein SRP9, putative
	CSPC1569	Putative nucellin-like aspartic protease
	CSPC1610	Developmentally regulated GTP binding protein
	CSPC1634	Receptor-like protein kinase
	CSPC1734	Leucine-rich repeat transmembrane protein kinase
	CSPC1741	Receptor-like protein kinase 3
	CSPC2262	Gibberellin 2-oxidase 1
	CSPC244	APETAL2-like protein
	CSPC416	Probable serine/threonine-specific protein kinase
Transport	CSPC191	Sugar transporter family protein
	CSPC229	Putative sucrose transporter
Cell wall enzyme	CSPC124	Putative endoxyloglucan transferase
	CSPC1797	Pectin methylesterase PME1
	CSPC1848	Probable glucosyltransferase
Secondary metabolism	CSPC26	9-*cis*-epoxy-carotenoid dioxygenase 1

[a] Unigene identifier is the library designation concatenated with the contig number assigned after assembly.

[b] The putative annotation was based on GenBank nr (BLASTX *E*-value $\leq 10^{-2}$).

Table 6
Numbers of new genes discovered in whole EST collection and in pericarp-derived libraries

Gene class		Whole ESTs	Pericarp ESTs
Receptor (GPCR, photoreceptor types)		5	1
Kinase	Receptor kinase	24	2
	Serine/Threonine kinase	29	2
	MAPK/K/K	3	0
	Other kinase	22	3
Transcription factor	Auxin response factor	2	0
	bHLH transcription factor	5	0
	bZIP transcription factor	5	0
	MYB transcription factor	8	2
	Zinc finger protein	20	5
	Other transcription factor	19	4
Cytochrome *P450* (unknown functions)		15	2
Terpenoid and flavonoid biosynthesis		7	2
Transporter	Ion transporter	16	2
	Hexose transporter	2	0
	Amino acid/peptide transporter	8	2
	Secondary metabolite transporter (ABC, MRP types)	10	2

3.7. Discovery of new candidate genes in the grapevine unigene set with possible roles in berry development

A total of 2725 novel grape unigenes, clustered from 3747 ESTs, were discovered in this EST study. Among them, 2203 (or ~81%) are singletons, suggesting relatively low expression of the corresponding genes in the sampled tissues, and 1006 (or ~37%) have no similarity (BLASTX *E*-value $\leq 10^{-2}$) to any sequences from other organisms available in the GenBank NR division. To help ascertain that the novelty of these sequences without hits to any publicly available sequences was not likely due to low sequence quality resulting from single-pass sequencing, we predicted their open reading frames (ORFs) using the EMBOSS getorf tool. More than 90% of these sequences contained an ORF encoding a minimum of 30 amino acids.

One of our EST sequencing goals was to discover new genes that may have roles in grape berry development and regulation of berry composition important for wine and table grape quality. Towards this end, we mined this set of 2725 novel sequences from the two separate cultivar assemblies to find unigenes containing at least one EST from the pericarp libraries in CS or MH. We found 357 new grape genes from pericarp libraries, 159 of which have no hits to any public sequences. Several new sequences annotated as signal transduction components including receptors, kinases, and transcription factors were discovered (Table 6). Expression profiling will be required next to further develop hypotheses regarding roles for these genes in e.g., berry ripening initiation, morphology, and/or metabolism underlying berry flavor. We also found many new allozymes and transporters, some of which may be involved in the biosynthesis and storage of those primary and secondary metabolites relevant for the modulation of berry flavor and appearance (Table 6).

Of the 59 transcription factor unigenes that we discovered, 11 were found from pericarp libraries; however, ESTs representing some other functional classes were only discovered in tissues other than pericarp. For interesting unigene sequences that were discovered in tissues other than pericarp, we can use these data for primer design to examine whether they are also expressed in pericarp. For example, a G-protein coupled receptor (GenBank accession numbers EC925845, EC953275) was discovered from the CSNB library, but polymerase chain reaction (PCR) analysis showed that transcripts corresponding to this gene were detectable in the pericarp, as well (data not shown).

4. Conclusion

The gene discovery efforts presented here represent a significant contribution to the identification of the complete grapevine transcriptome. The large number of publicly available grape ESTs will substantially increase the gene coverage in designing next generation microarray platforms to further advance predictions of gene functions in traits of economic importance to the grape and wine industry. This large body of ESTs will also aid in open reading frame annotation when the grapevine genome is fully sequenced and assembled in the next year.

Acknowledgments

We wish to thank Sarah Munro and Robert Kirkpatrick for their help with sequencing and bioinformatics, and Ana Ibañez and Leonor Ruiz-García for their assistance with tissue sampling and RNA extractions. The authors gratefully acknowledge funding from Genome Canada, Genoma España, as well as project management support from Genome British Columbia as part of the Genome Canada-Genoma España collaborative research and development initiative. JB is an NSERC EWR Steacie Fellow.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.07.016.

References

Ablett, E., et al., 2000. Analysis of grape ESTs: global gene expression patterns in leaf and berry. Plant Sci. 159, 87–95.
Adams, M.D., et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656.
Adams-Phillips, L., Barry, C., Giovannoni, J., 2004. Signal transduction systems regulating fruit ripening. Trends Plant Sci. 9, 331–338.
Aharoni, A., Vorst, O., 2002. DNA microarrays for functional plant genomics. Plant Mol. Biol. 48, 99–118.
Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.
Berardini, T.Z., et al., 2004. Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol. 135, 745–755.
Bonaldo, M.D.F., Lennon, G., Soares, M.B., 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6, 791–806.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., Bork, P., 2002. Alternative splicing and genome complexity. Nat. Genet. 30, 29–30.

da Silva, F.G., et al., 2005. Characterizing the grape transcriptome. Analysis of expressed sequence tags from multiple *Vitis* species and development of a compendium of gene expression during berry development. Plant Physiol. 139, 574–597.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–185.

Fei, Z., et al., 2004. Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. Plant J. 40, 47–59.

Giovannoni, J., 2001. Molecular biology of fruit maturation and ripening. Annu. Rev. Plant Physiol. Plant Mol. Biol. 52, 725–749.

Goff, S.A., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296, 92–100.

Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.

Iida, K., et al., 2004. Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. Nucleic Acids Res. 32, 5096–5103.

International Rice Genome Sequencing Project, 2005. The map-based sequence of the rice genome. Nature 436, 793–800.

Lamoureux, D., et al., 2006. Anchoring of a large set of markers onto a BAC library for the development of a draft physical map of the grapevine genome. Theor. Appl. Genet. 113, 344–356.

Lippert, D., et al., 2005. Proteome analysis of early somatic embryogenesis in *Picea glauca*. Proteomics 5, 461–473.

Lodhi, M.A., Reisch, B.I., 1995. Nuclear-DNA content of *Vitis* species, cultivars, and other genera of the Vitaceae. Theor. Appl. Genet. 90, 11–16.

Lund, S.T., Bohlmann, J., 2006. The molecular basis for wine grape quality — a volatile subject. Science 311, 804–805.

Moser, C., et al., 2005. Comparative analysis of expressed sequence tags from different organs of *Vitis vinifera* L. Funct. Integr. Genomics 5, 208–217.

Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., Gotoh, O., 2005. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. Gene 364, 53–62.

Pacey-Miller, T., Scott, K., Ablett, E., Tingey, S., Ching, A., Henry, R., 2003. Genes associated with the end of dormancy in grapes. Funct. Integr. Genomics 3, 144–152.

Park, S.C., Sugimoto, N., Larson, M.D., Beaudry, R., Nocker, S.V., 2006. Identification of genes with potential roles in apple fruit development and biochemistry through large-scale statistical analysis of expressed sequence tags. Plant Physiol. 141, 811–824.

Pontius, J., Wagner, L., Schuler, G., 2003. UniGene: a unified view of the transcriptome. The NCBI Handbook. National Center for Biotechnology Information, Bethesda, Maryland.

Pratt, L.H., et al., 2005. Sorghum expressed sequence tags identify signature genes for drought, pathogenesis, and skotomorphogenesis from a milestone set of 16,801 unique transcripts. Plant Physiol. 139, 869–884.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., Upton, J., 2000. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 28, 141–145.

Ralph, S., et al., 2006. Genomics of hybrid poplar (*Populus trichocarpa x deltoides*) interacting with forest tent caterpillars (*Malacosoma disstria*): normalized and full-length cDNA libraries, expressed sequence tags, and a cDNA microarray for the study of insect-induced defences in poplar. Mol. Ecol. 15, 1275–1297.

Reid, K.E., Olsson, N., Schlosser, J., Peng, F., Lund, S.T., 2006. An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development. BMC Plant Biol. 6, 27.

Rhee, S.Y., et al., 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res. 31, 224–228.

Rungis, D., Hamberger, B., Berube, Y., Wilkin, J., Bohlmann, J., Ritland, K., 2005. Efficient genetic mapping of single nucleotide polymorphisms based upon DNA mismatch digestion. Mol. Breed. 16, 261–270.

Schneider, M., Tognolli, M., Bairoch, A., 2004. The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. Plant Physiol. Biochem. 42, 1013–1021.

Soares, M.B., Bonaldo, M.D., Jelene, P., Su, L., Lawton, L., Efstratiadis, A., 1994. Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. U. S. A. 91, 9228–9232.

Terrier, N., Ageorges, A., Abbal, P., Romieu, C., 2001. Generation of ESTs from grape berry at various developmental stages. J. Plant Physiol. 158, 1575–1583.

Terrier, N., et al., 2005. Isogene specific oligo arrays reveal multifaceted changes in gene expression during grape berry (*Vitis vinifera* L.) development. Planta 222, 832–847.

The Gene Ontology Consortium, 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32, D258–D261.

Tuskan, G.A., et al., 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313, 1596–1604.

Wang, J.P., et al., 2004. EST clustering error evaluation and correction. Bioinformatics 20, 2973–2984.

Waters, D.L., Holton, T.A., Ablett, E.M., Lee, L.S., Henry, R.J., 2005. cDNA microarray analysis of developing grape (*Vitis vinifera* cv. Shiraz) berry skin. Funct. Integr. Genomics 5, 40–58.

Wheeler, D.L., et al., 2006. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 34, D173–D180.

Xie, D.Y., Sharma, S.B., Wright, E., Wang, Z.Y., Dixon, R.A., 2006. Metabolic engineering of proanthocyanidins through co-expression of anthocyanidin reductase and the PAP1 MYB transcription factor. Plant J. 45, 895–907.

Yu, J., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296, 79–92.