Characterization of Transcriptional Complexity during Berry Development in *Vitis vinifera* Using RNA-Seq[1][W]

Sara Zenoni[2], Alberto Ferrarini[2], Enrico Giacomelli, Luciano Xumerle, Marianna Fasoli, Giovanni Malerba, Diana Bellin, Mario Pezzotti, and Massimo Delledonne*

Department of Sciences, Technologies and Grapevine and Wine Markets (S.Z., M.F., M.P.); Department of Biotechnology (A.F., E.G., D.B., M.D.); and Department of Mother and Child and Biology-Genetics, Section of Biology and Genetics (L.X., G.M.), University of Verona, 37134 Verona, Italy

The development of massively parallel sequencing technologies enables the sequencing of total cDNA (RNA-Seq) to derive accurate measure of individual gene expression, differential splicing activity, and to discover novel regions of transcription, dramatically changing the way that the functional complexity of transcriptomes can be studied. Here we report on the first use of RNA-Seq to gain insight into the wide range of transcriptional responses that are associated with berry development in *Vitis vinifera* 'Corvina'. More than 59 million sequence reads, 36 to 44 bp in length, were generated from three developmental stages: post setting, véraison, and ripening. The sequence reads were aligned onto the 8.4-fold draft sequence of the Pinot Noir 40024 genome and then analyzed to measure gene expression levels, to detect alternative splicing events, and expressed single nucleotide polymorphisms. We detected 17,324 genes expressed during berry development, 6,695 of which were expressed in a stage-specific manner, suggesting differences in expression for genes in numerous functional categories and a significant transcriptional complexity. This exhaustive overview of gene expression dynamics demonstrates the utility of RNA-Seq for identifying single nucleotide polymorphisms and splice variants and for describing how plant transcriptomes change during development.

Grapevine (*Vitis* spp.) is a widely cultivated and economically important fruit crop comprising more than 50 species, although almost all the wine produced in the world derives from just one of them, *Vitis vinifera*, which is native to the area south of the Caucasus Mountains and the Caspian Sea (Jay, 1996).

The development and maturation of grape berries has been studied intensely because of the uniqueness of this process in plant biology and a desire to understand the physiological, biochemical, and molecular characteristics that determine fruit and wine quality. Berry development is a dynamic and complex process involving a cascade of biochemical changes (Coombe and McCarthy, 2000). After the fruit has set, the berries undergo a double sigmoidal pattern of growth, which is divided into three distinct stages. The post fruit-set stage involves rapid growth and cell division. Organic acids such as tartrate and malate accumulate in the vacuole, and major precursors of phenolic compounds are synthesized. The véraison stage is characterized by slower growth and the initiation of berry softening. Sugars and pigments begin to accumulate. In the ripening phase, the berries reach their mature size and color, the sugar concentration increases, organic acid production decreases, and volatile secondary metabolites are synthesized that contribute flavor and aroma.

The physiological and biochemical changes described above reflect the transcriptional modulation of many genes, but little is known about these transcriptional changes and their regulation. The analysis of cDNA-AFLP libraries (Zamboni et al., 2008) and microarrays (Pilati et al., 2007; Deluc et al., 2008; Lund et al., 2008) has provided a first picture of transcriptome dynamics during berry development, but these approaches suffer a number of drawbacks and the data are far from complete. For example, cDNA-AFLP analysis covers no more than 60% to 65% of the transcriptome due to the lack of restriction enzyme sites within the remaining cDNAs, and generates a large number of false-positive (comigrating) bands. The sensitivity and accuracy of microarray analysis is limited by background hybridization and the differing performance among probes, and most importantly the Affymetrix *Vitis* GeneChip contains only approximately 14,500 unigenes. The recently published 8.4-fold draft sequence of the grapevine genome indicates there are at least 30,434 protein-coding genes (Jaillon et al., 2007).

Novel, high-throughput, deep-sequencing technologies are making an impact on genomic research by

[1] This work was supported by Fondazione Cariverona (Completamento e attività del Centro di Genomica Funzionale Vegetale). Additional funding was provided by Ministero Politiche Agricole e Forestali (VIGNA project), by Università degli Studi di Verona (Joint Projects), and by Consorzio ORVIT.

[2] These authors contributed equally to the article.

* Corresponding author; e-mail massimo.delledonne@univr.it.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Massimo Delledonne (massimo.delledonne@univr.it).

[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.109.149716

Table I. *Summary of read number*

	Post Fruit Set	Véraison	Ripening
No. of total reads	17,473,026	20,414,476	21,485,042
No. of mapped reads	14,221,092	16,747,270	17,974,725
Unique	10,875,728	13,949,356	14,713,154
Multimatch	3,345,364	2,797,914	3,261,571
No. of reads not mapped	3,249,965	3,526,757	3,423,777

providing new strategies to analyze the functional complexity of transcriptomes. The RNA-Seq approach (Mortazavi et al., 2008) produces millions of short cDNA reads that are mapped to a reference genome to obtain a genome-scale transcriptional map, which consists of the transcriptional structure and the expression level for each gene. The holistic view of the transcriptome and its organization provided by the RNA-Seq method also reveals many novel transcribed regions, splice isoforms, single nucleotide polymorphisms (SNPs), and the precise location of transcription boundaries (Cloonan and Grimmond, 2008; Li et al., 2008; Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Sultan et al., 2008; Wilhelm et al., 2008). Finally, RNA-Seq generates absolute rather than relative gene expression measurements, providing greater insight and accuracy than microarrays (Hoen et al., 2008; Marioni et al., 2008; Mortazavi et al., 2008).

We have carried out the first global analysis of the grapevine (cv Corvina) transcriptome during berry development using the Illumina RNA-Seq method. Although our major effort was to validate the RNA-Seq technology and to set up a pipeline that allows observation of the level of gene expression, new transcripts, splice variants, and expressed SNPs, we report here a comprehensive analysis of transcriptome dynamics that may serve as a gene expression profile blueprint in berry development.

RESULTS

Isolation of RNAs from Berry Tissues and Library Construction

To characterize changes in gene expression during the dynamic process of grape berry development, *V. vinifera* Corvina berries were collected at 5, 10, and 15 weeks post flowering, corresponding to the post fruit-set, véraison, and ripening stages, respectively (Pilati et al., 2007; Deluc et al., 2008). Post fruit-set berries were small, hard, and green, and the total soluble solids content was 4.3 degrees Brix. During véraison, berries started to change color and sugars started to accumulate (11.3 degrees Brix), whereas ripening berries were softened, accumulated anthocyanin pigments, and their sugar content increased to 20 degrees Brix. Three pools of mRNA samples, one representing each stage, were used to build libraries for high-throughput parallel sequencing using an Illumina genome analyzer II.

Illumina Sequencing and Mapping of the Reference Genome

We generated 59,372,544 sequence reads, each 36 to 44 bp in length, encompassing 2.2 Gb of sequence data (Table I). Each stage was represented by at least 17 million reads, a tag density sufficient for quantitative analysis of gene expression (Morin et al., 2008).

The sequence reads were aligned on the Pinot Noir 40024 reference genome (Jaillon et al., 2007) and to a custom splice-junctions database, using the ELAND software (part of GERALD Illumina software) set to allow two base mismatches. Of the total reads, 82.4% matched either to a unique (66.6%) or to multiple (15.8%) genomic locations (Table I).

An in-house Python algorithm was used to assign each unique read to a specific location to determine the genomic distribution (Table II). Most reads (71.5%) were confined to exons and 16.1% mapped within introns. Reads that mapped within a radius of 4 kb from both ends of a gene (8.7%) were ascribed to unannotated exons or untranslated transcribed regions (UTRs), and reads that fell outside this catchment (3.7%) defined intergenic regions that could be ascribed to putatively novel transcripts.

Table II. *Genomic distribution of reads*

Developmental Stage	No. of Total Unique Reads	Exons		Introns		Mapped to 4-kb Gene-Flanking Regions		Intergenic Regions	
		No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Post fruit set	10,875,728	7,558,624	69.5	1,743,988	16.0	998,487	9.2	573,656	5.3
Véraison	13,949,356	10,097,777	72.4	2,321,760	16.6	1,117,381	8.0	411,659	3.0
Ripening	14,713,154	10,616,326	72.2	2,280,757	15.5	1,328,686	9.0	486,550	3.3
Total	39,538,238	28,272,727	71.5	6,346,505	16.1	3,444,554	8.7	1,471,865	3.7

Table III. *Number of genes with alternative and constitutive splicing*

Developmental Stage	No. of Alternative Splicing Events Detected	No. of Constitutive Splicing Events Detected	No. of Genes with Alternative Splicing Detected	No. of Genes with Constitutive Splicing Detected
Post fruit set	185	29,475	173	8,199
Véraison	229	31,126	207	8,049
Ripening	290	30,746	258	7,817
Combined samples	447	41,447	385	9,781

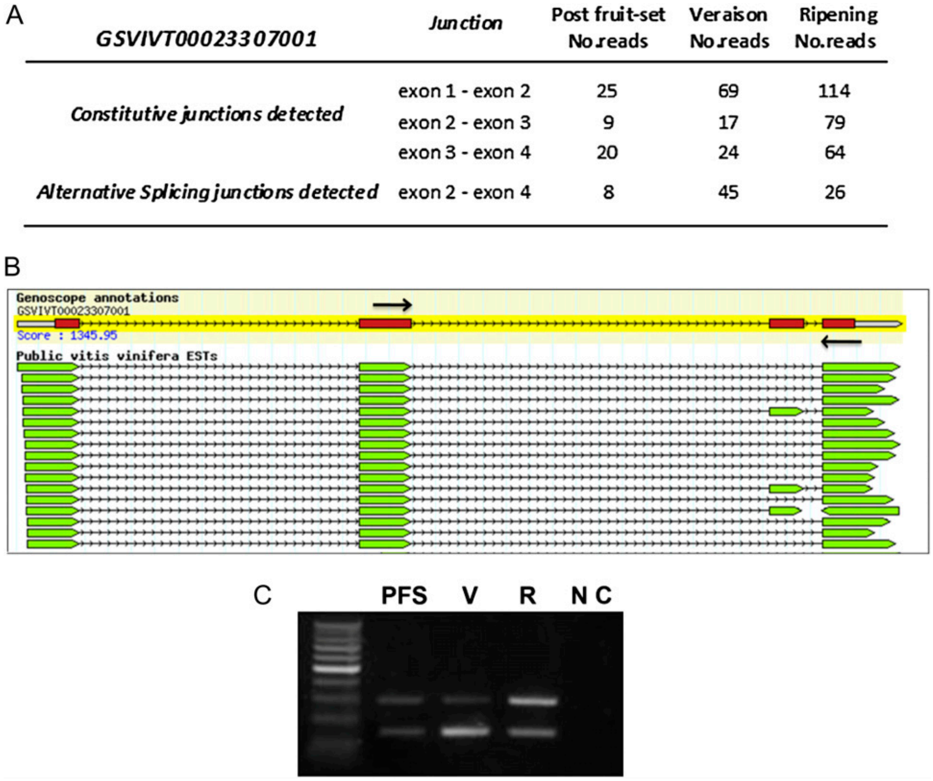
Identification of Splice Variants

As well as evaluating the transcribed portion of the grape genome during berry development, we looked at posttranscriptional processing events such as constitutive and alternative splicing. Because a database of validated alternative splicing for *V. vinifera* is not yet available, we generated an exon-junction database from synthetically computed 54-mer splice junctions enumerating all theoretical constitutive and alternative splice junctions within annotated transcripts using the exon-skipping model (Pan et al., 2008). We then wrote a Python script to count the unmatched reads from the reference genome that mapped onto the exon-junction database. All the sequence reads that mapped to more than one splice junction were discarded. To increase accuracy, we conservatively required that at least five independent tags mapped to the same junction with at least 5 nt across the exon-exon junction (Pan et al., 2008). Among the 92,051 splice junctions we detected (Supplemental Data S1), approximately 0.8% corresponded to alternative splicing events in 385 genes (Table III). Of these, 210 undergo alternative splicing in one stage only, 97 undergo alternative splicing in two stages, and 78 show alternative spliced forms in all the three stages.

Although accurate validation of all the alternative splicing events is beyond the scope of this investigation, we report as an example the splicing of GSVIVT00023307001 (Fig. 1). This gene comprises four exons and, according to our RNA-Seq analysis, is expressed in two forms, both of which are represented in the public *V. vinifera* ESTs database. The presence of GSVIVT00023307001 constitutive and alternative transcript variants during the three stages of berry development, was confirmed by reverse transcription (RT)-PCR and Sanger sequencing.

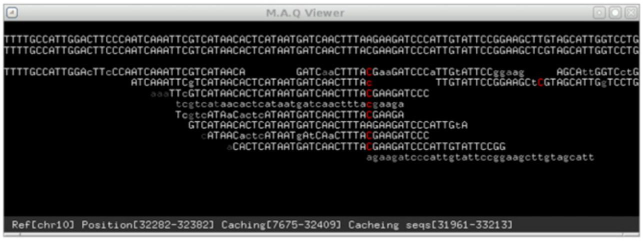
Polymorphism Detection

RNA-Seq data were also used to compare the reference and study genomes (Pinot Noir 40024 and Cor-



GSVIVT00023307001	*Junction*	Post fruit-set No.reads	Veraison No.reads	Ripening No.reads
Constitutive junctions detected	exon 1 - exon 2	25	69	114
	exon 2 - exon 3	9	17	79
	exon 3 - exon 4	20	24	64
Alternative Splicing junctions detected	exon 2 - exon 4	8	45	26

Figure 1. Candidate gene undergoing alternative splicing. A, Summary of reads falling onto constitutive and alternative junctions of the GSVIVT00023307001 gene in berry developmental stages. B, GSVIVT00023307001 gene model is represented by four exons (red tracks) and three introns (black lines). The *V. vinifera* ESTs (green tracks below) confirm the existence of alternative spliced mRNAs lacking the third exon. C, RT-PCR confirmation of GSVIVT00023307001 constitutive and alternative splice events in berry post fruit-set (PFS), véraison (V), and ripening (R) developmental stages. NC is the negative control. The predicted fragment sizes are 295 bp for the constitutive splice and 149 bp for the alternative splice.



```
TTTTGCCATTGGACTTCCCAATCAAATTCGTCATAACACTCATAATGATCAACTTTAAGAAGATCCCATTGTATTCCGGAAGCTTGTAGCATTGGTCCTG
TTTTGCCATTGGACTTCCCAATCAAATTCGTCATAACACTCATAATGATCAACTTTACGAAGATCCCATTGTATTCCGGAAGCTCGTAGCATTGGTCCTG

TTTTGCCATTGGAcTTcCCAATCAAATTCGTCATAACA        GATCaaCTTTACGaaGATCCCaTTGtATTCCggaag     AGCAttGGTCctG
                    ATCAAATTCgTCATAACACTCATAATGATCAACTTTAc          TTGTATTCCGGAAGCtCGTAGCATTGgTCCTG
                       aaaTTcGTCATAACACTCATAATGATCAACTTTACGAAGATCCC
                           tcgtcataacactcataatgatcaactttacgaaga
                           TcgtcATAaCactcATAATGATCAACTTTACGAAGA
                             GTCATAACACTCATAATGATCAACTTTAAGAAGATCCCATTGtA
                               cATAACactcATAATgAtCAaCTTTACGAAGATCCC
                                   aCACTCATAATGATCAACTTTACGAAGATCCCATTGTATTCCGG
                                                         agaagatcccattgtattccggaagcttgtagcatt

Ref[chr10] Position[32282-32382] Caching[7675-32409] Cacheing seqs[31961-33213]
```

Figure 2. Comparative displays from Maqview showing MAQ-aligned reads. The top track shows the reference sequence and the next shows the MAQ-called consensus derived from the aligned individual reads depicted below. Differences from the reference base are shown in red.

vina) and identify expressed SNPs. For initial sequence alignment and candidate SNP identification, we used the publicly available MAQ software (Li et al., 2008). In total, 50,512,088 (85%) reads were successfully aligned, providing an average 30-fold depth of coverage as calculated using an ad hoc Perl script called coverage.pl (50,512,088 reads × 38.02 bp read average length/ 63,722,636 bp of aligned reference sequence). By considering only SNPs detected in unique regions, and using a minimum read depth of 10, we identified 85,870 polymorphisms, 9.14% of which resided in coding regions of 2,973 genes, 28.12% in introns of 4,375 genes, and 62.74% in unpredicted regions (Supplemental Data S2). The MAQ viewer output is shown in Figure 2.

Global Analysis of Gene Expression

One of the primary goals of transcriptome sequencing is to compare gene expression levels in different samples. For robust conclusions about biological differences among samples it is important to utilize biological replication. Due to the high cost of RNA-seq our analysis is limited to a single biological sample of pooled tissue from multiple berry clusters. Although our experiment consisted of a single biological replicate, we were interested in developing approaches to look at the expression differences among these three samples. The following analyses demonstrate methods for using RNA-Seq data to perform global analysis of transcriptome variation during development.

ERANGE measures gene expression in reads per exon kilobase per million mapped sequence reads (RPKM), a normalized measure of exonic read density that allows transcript levels to be compared both within and between samples (Mortazavi et al., 2008). As ERANGE distributes multireads at similar loci in proportion to the number of unique reads recorded, we included in the analysis both unique reads and reads that occur up to 10 times to avoid undercount for genes that have closely related paralogs (Mortazavi et al., 2008). We detected the expression of 17,324 genes during berry development. Their expression in the three developmental stages is summarized in Figure 3.

To obtain statistical confirmation of the differences in gene expression among the developmental stages, we compared the RPKM-derived read count using Fisher's exact test (Bullard et al., 2009). A threshold value of $P = 5.46 \times 10^{-7}$ was used to ensure that differential gene expression was maintained at a significant level (5%) for the individual statistical tests (Supplemental Data S3). To minimize false positives and negatives, we estimated that statistical analysis was reliable when applied to genes showing an RPKM value ≥2 (i.e. six mapped reads on 200 nt of mRNA) in at least one of the three stages. It should be noted that these statistical significances are based on expected sampling distributions. Due to the use of a single biological replicate for each time point these high levels of significance may not reflect biological differences caused by development but may instead reflect other differences among the samples.

To facilitate the global analysis of gene expression, functional categories were assigned to all predicted grapevine genes (Jaillon et al., 2007) by automatic Genoscope annotation (http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/annotation/). This was verified manually and integrated using gene ontology (GO) classification (http://www.geneontology.org). Transcripts were then grouped into 19 GO functional categories (Supplemental Table S1), based on GO biological processes. Some gene families, with members whose functions were clearly linked to the physical and biochemical changes that take place during berry ripening, were characterized in more detail (Supplemental Table S2).

Finally, the expression profiles of the differentially expressed genes were determined by cluster analysis based on the k-means method using Pearson's correlation distance. Genes were divided into eight groups based on their expression modulation, representing the number of profiles indicated by figure of merit analysis (Genesis 1.7.5). Clusters 1 and 2 contained genes positively or negatively modulated along the

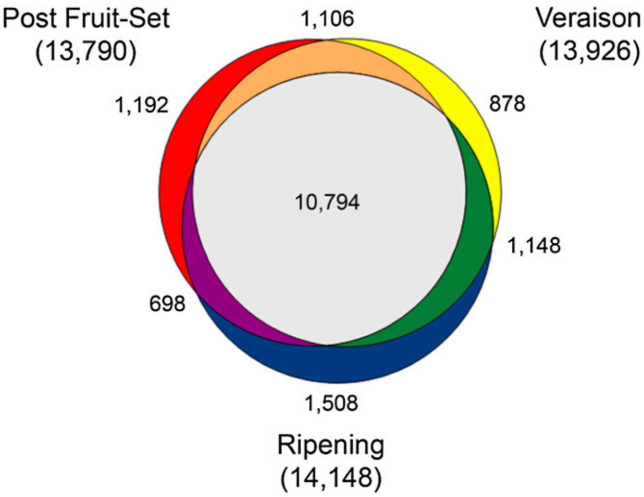


Figure 3. Venn diagram showing the genes expressed in each of the three stages of berry development.

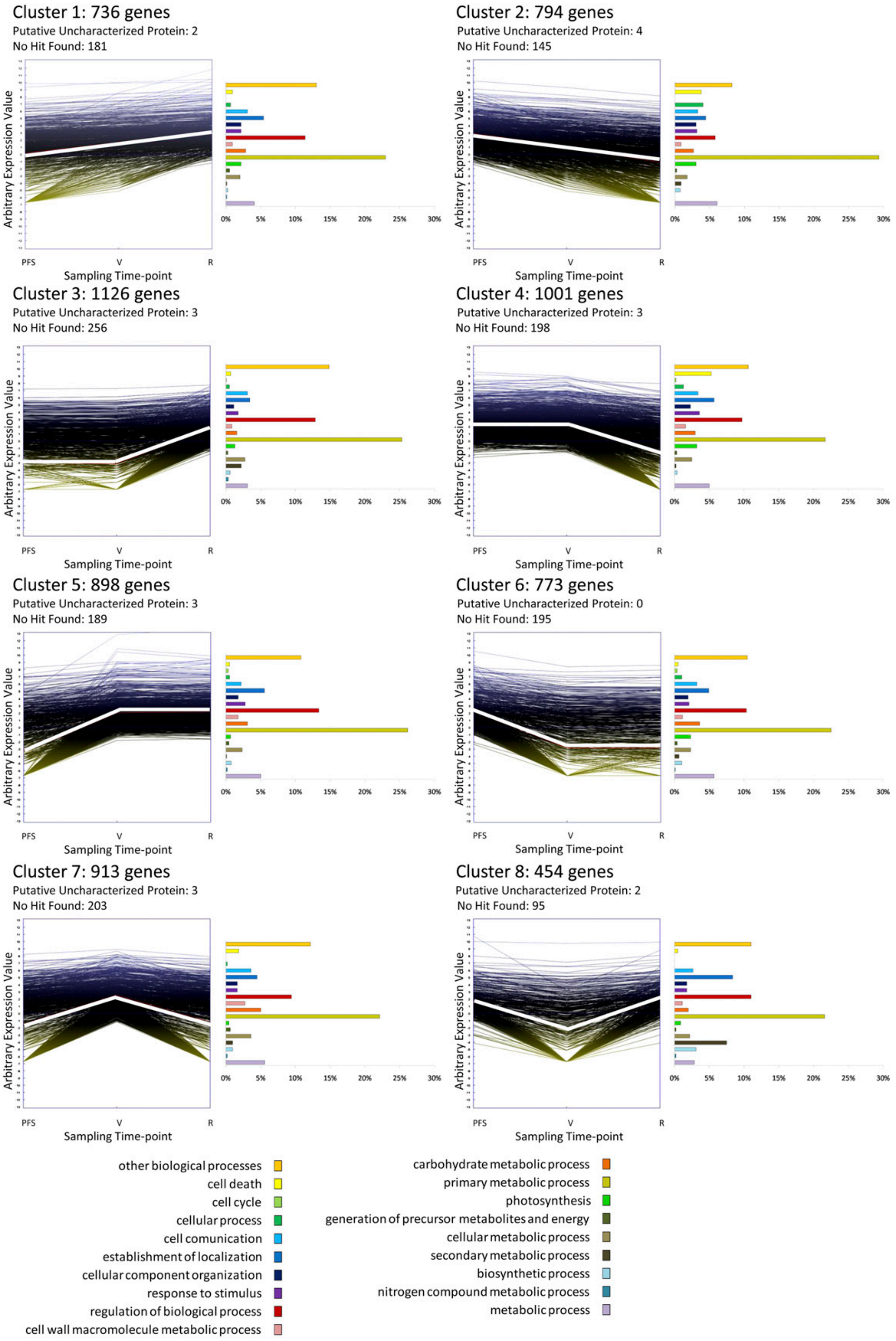


Figure 4. (*Legend appears on following page.*)

Table IV. *Expression pattern of berry ripening marker genes*

Gene ID	*V. vinifera* Gene	Gene Expression Values (RPKM)		
		Post Fruit Set	Véraison	Ripening
GSVIVT00020222001	Grip3	0	0.93	1.45
GSVIVT00020223001	Grip4	0.34	289.95	402.58
GSVIVT00020240001	Grip13	21.92	6,578.93	12,183.49
GSVIVT00020231001	Grip15	13.34	77.73	180
GSVIVT00029882001	Grip28	0	0.06	0.17
GSVIVT00024571001	Grip22	0.04	1.9	0.64
GSVIVT00003839001	Grip32	16.39	66.02	192.18
GSVIVT00001103001	VvTL1 (Grip 51)	0	0.06	0.18
GSVIVT00001105001	VvTL2	4.88	357.3	3,226.72
GSVIVT00030517001	Grip61	0	1,719.39	857.87

whole time course, clusters 3 and 4 contained genes modulated only during the pre-véraison interval, clusters 5 and 6 contained genes positively and negatively modulated after véraison. Genes expressed in early and late stages of berry development fell into cluster 7, and genes specifically induced at véraison were grouped in cluster 8. The functional category distribution frequency was then calculated for each cluster to identify differences in the distribution of genes among the three developmental stages (Fig. 4).

Validation of RNA-Seq-Based Gene Expression

To validate the expression profiles obtained by RNA-Seq, real-time RT-PCR was performed on 15 genes randomly selected for high or low expression levels. The genes encoded members of the MYB transcription factor family, enzymes involved in cell wall metabolism, and proteins synthesized in response to stress. For 12 of the genes, real-time RT-PCR expression profiles were in complete agreement with the RNA-Seq data, the remaining three differing at one or more stages (Supplemental Fig. S1). It is notable that the 3′ UTRs of the three incongruent genes have not been described, so it is possible that the primers may have annealed within a nonspecific or poorly annotated region. Among the 12 congruent genes, GSVIVT00036110001 (*VvMYB4*) was modulated below the level of detection of the array platforms (Cramer et al., 2007; Pilati et al., 2007), confirming that RNA-Seq profiling was suitable for the quantitative assessment of weakly expressed genes. We then selected three MYB genes for real-time RT-PCR on two independent RNA pools (north and south site of vineyard) for each berry developmental stage. The three MYB genes, which were selected among the 12 congruent genes due to their primary importance in controlling several pathways during berry development, showed the same trend of expression on independently collected samples (Supplemental Fig. S2).

Finally we validated the use of the RNA-Seq method by confirming the expression of previously identified markers of berry development. Davies and Robinson (2000) reported that several genes specifically expressed during berry ripening (GRIPs, referring to grape ripening-induced proteins) could be used as markers for the different developmental stages. We analyzed the RNA-Seq data for 10 berry-specific GRIPs (Table IV), confirming the previously reported expression profiles in all cases (Davies and Robinson, 2000; Pilati et al., 2007). In particular *GRIP 3*, *4*, *13*, *15*, and *28*, which are potentially involved in cell wall metabolism, were induced during ripening, whereas *GRIP 22* and *32*, which are potentially involved in the response to abiotic stresses, showed a peak of expression during véraison (*GRIP 22*) and an increase during ripening (*GRIP 32*). Genes involved in disease resistance were induced over the course of berry development (*VvTL1* and *VvTL2*) or modulated with a peak of expression during véraison (*GRIP 61*).

DISCUSSION

The unprecedented level of sensitivity and the high throughput of deep sequencing technologies suggests that RNA-Seq is likely to become the platform of choice for transcriptome, quickly superseding microarray-based methods for comprehensive studies of gene expression, differential splicing activity, and discovery of expressed SNPs. However, the new level of detail offered by RNA-Seq raises novel statistical and computational challenges that are still somehow limiting the large adoption of whole-genome transcriptional profiling to detect genes that are differentially expressed among several experiments.

In this study, we have generated >59 million sequence reads (36–44 bp in length) corresponding to 2.2 Gb of raw sequence data, by Illumina sequencing of

Figure 4. Categories distribution in the eight expression clusters. Clusters were obtained by the k-means method on the gene expression profiles of the 6,695 modulated genes. Histogram representation of the functional categories distribution is expressed as percentage of the amount of genes belonging to the cluster. Gene coding for unknown products were not considered in the analysis. PFS, Post fruit set; R, ripening; V, véraison.

mRNA from the three stages of berry development. Although the study *V. vinifera* genome (Corvina) was not the same as the reference genome (PN40024), >49 million reads were successfully mapped on the 8.4-fold draft sequence of the grapevine genome. By using publicly available softwares adapted by us and integrated with in-house Phyton and Perl scripts, mapped reads were used to identify putative new genes, to detect alternative splicing events and expressed SNPs, and to measure gene expression levels. At the current level of sampling (approximately 17 million tags per sample), the sensitivity of the RNA-Seq approach allowed the accurate identification of differentially expressed genes, as confirmed by the correct profiling of GRIPs (marker genes for berry ripening; Davies and Robinson, 2000), and the precise quantitative expression of genes, even of those weakly transcribed (such as transcription factor genes that are well below a microarray's detection limit), as confirmed by real-time RT-PCR analysis. Furthermore, our analysis led to the identification of alternative splicing events for 385 genes, suggesting that in developing berries there is considerably more transcript complexity than previously appreciated.

This global analysis of gene expression provided a comprehensive dataset (Supplemental Data S3) with each gene represented by its absolute expression level during the three stages of berry development, by a manually verified and integrated GO biological process annotation, and by a new annotation for some gene families whose function is clearly linked to the physical and biochemical changes that take place during berry ripening. Members of the glutathione-*S*-transferase (GST) family, which is involved in anthocyanin accumulation in the vacuole (Kitamura et al., 2004; Conn et al., 2008), were classified using PSI-BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). This led to the identification of 12 genes belonging to the ϕ class and 75 belonging to the τ class, the two most important classes of plant-specific GSTs (Edwards et al., 2000). The RNA-Seq analysis showed that seven ϕ class GSTs and 57 τ class GSTs are expressed during berry development. A previous microarray study identified only four ϕ class and seven τ class GSTs (three of which probably represented the same transcript), and all induced after véraison (Pilati et al., 2007).

Members of the stilbene synthase (STS) family, responsible for the synthesis of resveratrol (Aggarwal et al., 2004), were previously described in the PN40024 sequencing project (Jaillon et al., 2007). We profiled all 43 genes encoding STSs contained in the grapevine genome using the RNA-Seq method. With one exception (GSVIVT00037967001), all the STS genes were found up-regulated during the shift from véraison to ripening. Until this report, only three developmentally regulated STS genes had been identified in berries, all induced after véraison (Pilati et al., 2007).

Members of the MYB transcription factor family, which are directly involved in the regulation of flavonoid biosynthesis (Walker et al., 2007; Deluc et al., 2008; Terrier et al., 2009), were classified as discussed by Matus et al. (2008). A total of 108 candidate *MYB* genes have been identified, few of which have been characterized (Matus et al., 2008). Our RNA-Seq analysis reported the expression of 36 MYB genes during berry development, 28 of which had not been investigated before because of their very low expression level and/or the lack of corresponding probes on microarrays.

Although a global analysis of gene expression based on a single replication does not allow a solid biological interpretation, this RNA-Seq analysis clearly provided a comprehensive view of the participation of several multigene families in berry development and ripening, identifying which members are expressed and characterizing their expression profiles in detail, showing which are likely to participate in the synthesis and accumulation of secondary metabolites. In comparison to previous studies, the RNA-Seq method identified many additional transcripts, paving the way for a more accurate and more detailed description of the molecular processes involved in the development of grape berries and the basis of their organoleptic properties. Therefore, the RNA-Seq method combined with the appropriate bioinformatic tools provides a new approach to study gene expression dynamics on a global scale during a developmental process, allowing specific candidate genes to be highlighted for further functional analysis.

MATERIALS AND METHODS

Sample Preparation and Sequencing

Six berry clusters (*Vitis vinifera* 'Corvina', clone 48) were collected at 5, 10, and 15 weeks after flowering during the 2008 growing season, from a vineyard in the Verona Province (Montorio, Verona, Italy). The clusters were taken from three different plants on the vineyard's north site and three different plants on the south site. Ten berries were randomly selected from each cluster and pooled with berries from the other plants on the same vineyard site, resulting in two independent pools of 30 berries for each developmental stage.

The total soluble solids content (degrees Brix) of berry juice prepared from north and south samples at each developmental stage was assayed using a bench refractometer (PR-32; Atago Co.). Total RNA was also extracted from north and south samples at each time point, using the method described by Zamboni et al. (2008). RNA quality and quantity were determined using a Nanodrop 2000 instrument (Thermo Scientific) and a Bioanalyzer Chip RNA 7500 series II (Agilent). North and south RNA samples were pooled and 10 μg of total RNA from each pool was used to isolate poly(A) mRNA and to prepare a nondirectional Illumina RNA-Seq library with mRNA-Seq 8 sample prep kit (Illumina). We modified the gel extraction step by dissolving excised gel slices at room temperature to avoid underrepresentation of AT-rich sequences (Quail et al., 2008). Library quality control and quantification was performed with a Bioanalyzer Chip DNA 1000 series II (Agilent). Each library had an insert size of 200 bp, and 36 to 44 bp sequences were generated on an Illumina genome analyzer II.

Alignment and Analysis of Illumina Reads against the *V. vinifera* Genome and the Splice Database

Sequence alignments were generated with ELAND (Part of Illumina Pipeline version 1.3.2). ELAND is a very fast ungapped alignment program that supports up to two mismatches, reports reads aligning to multiple locations in the genome, which is very important for an accurate measurement of gene expression, and its output can be directly parsed by the ERANGE package that was used for differential expression analysis. Bowtie (Langmead et al., 2009) also produces an output that can be parsed by ERANGE (Mortazavi et al., 2008) but it's much slower because of its sequence quality processing ability. Nevertheless, tests conducted on our datasets showed a strong correlation among the ERANGE expression data obtained using ELAND or Bowtie mappings (data not shown). The *V. vinifera* RefSeq sequences were based on the 8.4-fold draft of the PN40024 genome (Jaillon et al., 2007). The genome assembly and the *V. vinifera* public ESTs database used for the analysis of alternative splicing are available on the Genoscope site (http://www.genoscope.cns.fr/). The splice junctions database was created with the script getsplicefa.py, included in standard ERANGE distribution, by joining the 3′ end sequence of the 5′ exon to the 5′ end sequence of the 3′ exon, resulting in 54-bp sequence tags (27 bp from each exon). Known gene annotations were loaded into a sqlite database by means of the script createDB.py. Data from the ELAND output were parsed by a Python script (mapreadslocation.py) that compares reads coordinates on the genome to those of known genes contained in the sqlite database. Each sequence read was thus assigned to an exon, intron/UTR, external exon, or intergenic region. Reads mapped onto external exons fell within a 4-kb catchment from both ends of a gene, promoting the investigation of putative undiscovered exons. Intergenic reads represented those sequence reads that fell outside this catchment.

To integrate putative exon splice events in the analysis we developed a Python script (getallsplice.py) that generated an in silico splice database containing junctions (constitutive and alternative) for all exons of every gene by considering the exon skipping model (Pan et al., 2008). The junctions were designed so that the 3′ end of the 5′ exon was connected to the 5′ end of the 3′ exon, resulting in 54-bp sequence tags (27 bp from each exon). By using ELAND, we then mapped 32-bp sequence reads with at least 5 bp mapping over a splice junction, allowing two mismatches (Pan et al., 2008). We considered only unique reads mapped onto the junctions. A Python script (get_alt_splice2.py) parsed all unique reads mapped onto the in silico splice database and extracted constitutive and alternative splicing coverage. All detected junctions covered by at least five reads are shown in Supplemental Data S1.

Differential Gene Expression Analysis

The evaluation of gene expression was performed with ERANGE 3.1 software, available at http://woldlab.caltech.edu/RNA-Seq (Mortazavi et al., 2008). ERANGE requires Cistematic 2.5 to execute RunStandardanalysis.sh. Therefore, a Python script (Vitisvinifera.py) was developed to import *V. vinifera* 8.4× reference sequence and gff annotation into Cistematics Genomes, and a Perl script (gff2knowngene.pl) was developed to convert the gff annotation file to the knowngene.txt file used by RunStandardanalysis.sh. ERANGE reports the number of mapped reads per kilobase of exon per million mapped reads, measuring the transcriptional activity for each gene. To obtain an accurate measure of gene expression not biased by reads mapping to splice junctions in genes with many introns, ERANGE considers both reads mapping to genome or to the custom splice junctions database. ERANGE was preferred over commercial packages such as CASAVA and GenomeStudio platform from Illumina because of its open nature. This allowed us to adapt and reuse code for our own analysis with greater flexibility than a comparable closed source commercial package.

Differential Gene Expression Statistic for mRNA-Seq

ERANGE software computes the normalized gene locus expression level (named RPKM) by assigning reads to their site of origin and counting them. In the case of reads that match equally well to several sites, ERANGE assigns them proportionally to their most likely site(s) of origin (Mortazavi et al., 2008).

The number of reads falling in a given gene locus can be estimated from the RPKM value as follows:

$$N = RPKM \times L \times N_{\mathrm{Tot}} \times 10^{-9}$$

where n = number of mapping reads at a given gene locus, L = estimated length (bp) of the gene locus, N_{tot} = number of total mapping reads, and RPKM = gene locus *RPKM* value.

The null hypothesis of no differential gene expression for each gene was tested using the Fisher's exact test. A stringent threshold value equal to P = 5.46 × 10^{-07} (30,434 genes, three comparisons of gene expression from the three development stages, pair by pair) was used to set the significance level at 0.05 for the individual statistical tests. Each pairwise combination of the three developmental stages was investigated.

Detection of SNPs

We used MAQ version 0.7.1 (Li et al., 2008) to align sequence reads on the Pinot Noir 40024 genome (Jaillon et al., 2007) and then call the putative SNPs. We choose to use MAQ (available at http://maq.sourceforge.net/) as it uses a quality score related to the uniqueness and quality of the alignment. At the multiple alignment reads MAQ assigns a quality score equal to 0. The SNPs were extracted using the cns2snp MAQ command, and the SNPfilter Perl script, part of MAQ package, was used to post process the SNPs using the following parameters: (1) called SNPs must be covered by at least 10 reads; (2) SNPs with base quality lower than 20 must be discarded; (3) SNPs with copy number of the flanking region in the reference genome higher than 1.0 must be discarded; (4) the quality of the 3 bp flanking region around putative SNPs must be higher than 10; (5) the quality of sequence read alignment across SNP sites must be higher than 60. Subsequently, we matched the position of each SNP onto the reference knowngene.txt file, thus obtaining information about the SNPs location (coding sequence, UTR, or nonannotated zone) using a Python script (mapSNPlocation.py), which compares the SNPfilter output file with gene annotations imported into a sqlite database with createDB.py.

RT-PCR Validation

First-strand cDNA synthesis was performed with 800 ng of total RNA using the Improm-II RT system (Promega), according to the manufacturer's instructions. All RNA samples were first treated with DNase I (Promega) and no transcriptase negative controls were performed for each PCR reaction. Forward-specific oligonucleotide primer (5′-GGAGTTGCAAGAATT-TAAGC-3′) and reverse primer (5′-TTCAGCCTTGAACTTCACAT-3′) were designed on second and fourth exon of GSVIV00023307001 gene, respectively. The PCR amplification was carried out in nonsaturating conditions and involved a 98°C hold for 5 min, followed by a 30 cycles at 98°C for 30 s, 52°C for 30 s, and 72°C for 40 s. The PCR products were separated in 2% agarose stained with SYBR Safe DNA gel stain (Invitrogen).

Real-Time PCR Validation

First-strand cDNA synthesis was performed as described above. The transcriptional profiles of 15 genes were analyzed by real-time RT-PCR using the SYBR Green PCR master mix (Applied Biosystems) and a Mx3000P real-time PCR system (Stratagene). Gene-specific primers were designed for 12 genes using the sequence information in the 3′ UTR, whereas for the three genes lacking information covering this region, primers were designed to anneal in the coding sequence. A primer pair was also designed for TC81781 (The Institute for Genomic Research, Release 6.0), encoding an actin protein.

The PCR involved a 50°C hold for 2 min and a 95°C hold for 10 min followed by 40 cycles at 95°C for 30 s, 55°C for 30 s, and 72°C for 20 s. Nonspecific PCR products were identified by the dissociation curves. Amplification efficiency was calculated from raw data using LingRegPCR software (Ramakers et al., 2003). The relative expression ratio value was calculated for development time points relative to the first sampling time point (post fruit set) according to the Pfaffl equation (Pfaffl, 2001). SE values were calculated according to Pfaffl et al. (2002). Each real-time PCR was carried out three times.

Scripts Availability

Scripts source code and help files can be freely downloaded from http://ddlab.sci.univr.it/srtk/.

The short-read sequence data from this article are being submitted to the National Center for Biotechnology Information short read archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession number SRA009962. Data can also be accessed on Genome Browser at http://ddlab.sci.univr.it/cgi-bin/gbrowse/grape/.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Real-time RT-PCR of 15 genes randomly selected for high or low expression levels.

Supplemental Figure S2. Real-time RT-PCR performed on independently collected samples from the vineyard's north site (white histograms) and south site (black histograms).

Supplemental Table S1. List of GO Functional categories.

Supplemental Table S2. Global gene expression of gluthatione-*S*-transferase, STS, and MYB gene families.

Supplemental Data S1. Summary of constitutive and alternative junctions.

Supplemental Data S2. Detection of SNPs with MAQ.

Supplemental Data S3. Global gene expression of grape berry development.

Received October 21, 2009; accepted January 27, 2010; published January 29, 2010.

LITERATURE CITED

Aggarwal BB, Bhardwaj A, Aggarwal RS, Seeram NP, Shishodia S, Takada Y (2004) Role of resveratrol in prevention and therapy of cancer: preclinical and clinical studies. Anticancer Res **24:** 2783–2840

Bullard JH, Purdom EA, Hansen KD, Durinck S, Dudoit S (2009) Statistical inference in mRNA-Seq: exploratory data analysis and differential expression. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 247

Cloonan N, Grimmond SM (2008) Transcriptome content and dynamics at single-nucleotide resolution. Genome Biol **9:** 234

Conn S, Curtin C, Bezier A, Franco C, Zhang W (2008) Purification, molecular cloning, and characterization of glutathione S-transferases (GSTs) from pigmented Vitis vinifera L. cell suspension cultures as putative anthocyanin transport proteins. J Exp Bot **59:** 3621–3634

Coombe BG, McCarthy MG (2000) Dynamics of grape berry growth and physiology of ripening. Aust J Grape Wine Res **6:** 131–135

Cramer GR, Ergul A, Grimplet J, Tillett RL, Tattersall EA, Bohlman MC, Vincent D, Sonderegger J, Evans J, Osborne C, et al (2007) Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. Funct Integr Genomics **7:** 111–134

Davies C, Robinson SP (2000) Differential screening indicates a dramatic change in mRNA profiles during grape berry ripening: cloning and characterization of cDNAs encoding putative cell wall and stress response proteins. Plant Physiol **122:** 803–812

Deluc L, Bogs J, Walker AR, Ferrier T, Decendit A, Merillon JM, Robinson SP, Barrieu F (2008) The transcription factor VvMYB5b contributes to the regulation of anthocyanin and proanthocyanidin biosynthesis in developing grape berries. Plant Physiol **147:** 2041–2053

Edwards R, Dixon DP, Walbot V (2000) Plant glutathione S-transferases: enzymes with multiple functions in sickness and in health. Trends Plant Sci **5:** 193–198

Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res **36:** e141

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467

Jay JM, editor (1996) Modern Food Microbiology, Ed 5. Chapman & Hall, New York, pp 125–147

Kitamura S, Shikazono N, Tanaka A (2004) TRANSPARENT TESTA 19 is involved in the accumulation of both anthocyanins and proanthocyanidins in Arabidopsis. Plant J **37:** 104–114

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10:** R25

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res **18:** 1851–1858

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133:** 523–536

Lund ST, Peng FY, Nayar T, Reid KE, Schlosser J (2008) Gene expression analyses in individual grape (Vitis vinifera L.) berries during ripening initiation reveal that pigmentation intensity is a valid indicator of developmental staging within the cluster. Plant Mol Biol **68:** 301–315

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res **18:** 1509–1517

Matus JT, Aquea F, Arce-Johnson P (2008) Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across Vitis and Arabidopsis genomes. BMC Plant Biol **8:** 83

Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, et al (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res **18:** 610–621

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5:** 621–628

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320:** 1344–1349

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40:** 1413–1415

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res **29:** e45

Pfaffl MW, Horgan GW, Dempfle L (2002) Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res **30:** e36

Pilati S, Perazzolli M, Malossini A, Cestaro A, Dematte L, Fontana P, Dal Ri A, Viola R, Velasco R, Moser C (2007) Genome-wide transcriptional analysis of grapevine berry ripening reveals a set of genes similarly modulated during three seasons and the occurrence of an oxidative burst at veraison. BMC Genomics **8:** 428

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods **5:** 1005–1010

Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett **339:** 62–66

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321:** 956–960

Terrier N, Torregrosa L, Ageorges A, Vialet S, Verries C, Cheynier V, Romieu C (2009) Ectopic expression of VvMybPA2 promotes proanthocyanidin biosynthesis in grapevine and suggests additional targets in the pathway. Plant Physiol **149:** 1028–1041

Walker AR, Lee E, Bogs J, McDavid DA, Thomas MR, Robinson SP (2007) White grapes arose through the mutation of two similar and adjacent regulatory genes. Plant J **49:** 772–785

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature **453:** 1239–1243

Zamboni A, Minoia L, Ferrarini A, Tornielli GB, Zago E, Delledonne M, Pezzotti M (2008) Molecular analysis of post-harvest withering in grape by AFLP transcriptional profiling. J Exp Bot **59:** 4145–4159