Isoform-scale annotation and expression profiling of the Cabernet Sauvignon transcriptome using single-molecule sequencing of full-length cDNA

Andrea Minio[1], Melanie Massonnet[1], Amanda M. Vondras[1], Rosa Figueroa-Balderas[1], Barbara Blanco-Ulate[2], and Dario Cantu[1]*

[1] Department of Viticulture and Enology, University of California Davis, Davis, CA, USA
[2] Department of Plant Sciences, University of California Davis, Davis, CA, USA

***Correspondence**:
Dario Cantu
dacantu@ucdavis.edu

Running title: Iso-Seq of the grape transcriptome

Keywords: single molecule real-time sequencing, Iso-Seq, alternative splicing, fruit ripening, genome annotation, transcriptome reconstruction

ABSTRACT

Vitis vinifera cv. Cabernet Sauvignon is one of the world's most widely cultivated red wine grape varieties and often used as a model for studying transcriptional networks governing berry development and metabolism. Here, we applied single-molecule sequencing technology to reconstruct the transcriptome of Cabernet Sauvignon berries during ripening. We added an error-correction step to the standard Iso-Seq pipeline that included using Illumina RNAseq reads to recover lowly-expressed transcripts. From 672,635 full-length non-chimeric reads, we produced 170,860 transcripts capturing 13,402 genes of the Cabernet Sauvignon genome. Full-length transcripts refined approximately one third of the gene models predicted using several *ab initio* and evidence-based methods. The Iso-Seq information also helped identify 563 additional genes, 4,803 new alternative transcripts, and the 5' and 3' UTRs in the majority of predicted genes. Comparisons with the gene content of other grape cultivars identified 549 Cabernet Sauvignon-specific genes, including 65 genes differentially regulated during ripening. Some of these genes were potentially associated with the phenylpropanoid and flavonoid pathways, which may influence unique Cabernet Sauvignon berry attributes. Over 23% of the 36,687 annotated genes in Cabernet Sauvignon had two or more alternative isoforms, predominantly due to intron retention and alternative acceptor and donor sites. We profiled the expression of all isoforms using short read sequencing and identified 252 genes whose alternative transcripts showed different expression patterns during berry development.

INTRODUCTION

The history of *Vitis vinifera* (grape) is deeply intertwined with that of civilization and is closely associated with trade, literature, and culture (Campbell, 2006; McGovern et al., 2003; Unwin, 2005; Westering and Ravenscroft, 2001). Grape was probably domesticated between 6,000 and 22,000 years ago in the Near East (McGovern et al., 2003; Myles et al., 2011; Zhou et al., 2017). Once established, grape-growing (viticulture) and wine-making (enology) often became significant components of countries' economies, with fruit being used for table grapes, raisins, wine, spirits and other products. In terms of gross production value, grape is among the ten most valuable crops globally (69,200.62 million USD; http://www.fao.org/faostat/en/#data). Grape has proven useful for the study of non-climacteric, fleshy fruit (Davies et al., 1997). Though ripening in climacteric fruit like tomato is well-studied and largely governed by ethylene, ripening in non-climacteric fruit like grape, strawberry and citrus is not entirely clear and involves several hormone families (Böttcher et al., 2011; Fortes et al., 2015; Koyama et al., 2010; Symons et al., 2012). Grape has been a useful model for examining the complex crosstalk between these hormones and may give insight into their relationships in other models and contexts (Blanco-Ulate et al., 2017; Chervin et al., 2004; Qian et al., 2016).

Genome-wide expression studies using microarray and, more recently, RNA sequencing (RNAseq) revealed that ripening involves the expression and modulation of ~23,000 genes (Massonnet et al., 2017a) and that the ripening transition is associated with a major transcriptome shift (Fasoli et al., 2012). Transcriptomics has proven invaluable for characterizing a ripening program that is similar across an array of grapevine cultivars (Massonnet et al., 2017a), for assessing differences between them (Da Silva et al., 2013; Jiao et al., 2015; Venturini et al., 2013), identifying key ripening related genes (Massonnet et al., 2017a; Palumbo et al., 2014), and determining the impact of stress and viticultural practices on ripening (Amrine et al., 2015; Blanco-Ulate et al., 2015, 2017; Corso et al., 2015; Deluc et al., 2009; Hopper et al., 2016; Lecourieux et al., 2017; Massonnet et al., 2017b; Pastore et al., 2013; Savoi et al., 2017, 2016; Xi et al., 2014; Zenoni et al., 2017). This knowledge increases the possibility of exerting control over the ripening process, improving fruit composition under suboptimal or adverse conditions, and honing desirable traits in a crop with outstanding cultural and commercial significance.

These genome-wide expression analyses were enabled by the first effort to sequence the grape genome and generate a contiguous assembly for the species (Jaillon et al., 2007); this first effort focused on a highly homozygous line (PN40024) that was created by several rounds of backcrossing to reduce heterozygosity and facilitate genome assembly (Jaillon et al., 2007). Though poor by current standards (contig N50 = 102.7 kb), this pioneering, chromosome-resolved assembly served as the basis for numerous publications. However, the structural diversity of grape genomes makes using a single one-size-fits-all reference genome inappropriate (Golicz et al., 2016a, 2016b). There is substantial unshared gene content between cultivars, with 8 - 10% of the genes missing when two cultivars are compared (Da Silva et al., 2013). Although many of these variable genes are not essential for the plant survival, these genes can account for 80% of the expression within their respective families and expand key gene families possibly associated with cultivar-specific traits (Da Silva et al., 2013). Assembling genome references for all interesting cultivars is impractical in part because the cost of doing so remains prohibitive. In addition, the grape genome has also features that impede the development of high-quality genome assemblies for other cultivars than PN40024. Although the *V. vinifera* genome is relatively small (Jaillon et al., 2007; Lodhi and Reisch, 1995) and as repetitive as other plant genomes of similar size (Jaillon et al., 2007; Michael and Jackson, 2013), it is highly heterozygous (Da Silva et al., 2013; Gambino et al., 2017; Jaillon et al., 2007; Venturini et al., 2013). Most domesticated grape cultivars are crosses between distantly related parents; this may influence the high heterozygosity observed in the species (Bowers and Meredith, 1997; Chin et al., 2016; Cipriani et al., 2010; Di Gaspero et al., 2005; Ibáñez et al., 2009; Lacombe et al., 2013; Lopes et al., 1999; Minio et al., 2017; Myles et al., 2011; Ohmi et al., 1993; Sefc et al., 1998; Strefeler et al., 1992; Tapia et al., 2007). Earlier attempts using short reads struggled to resolve complex, highly heterozygous genomes (Di Genova et al., 2014; Gnerre et al., 2011; Huang et al., 2012;

Kajitani et al., 2014; Safonova et al., 2015). A limited ability to call consensus polymorphic regions yields highly fragmented assemblies where structural ambiguity occurs and alternative alleles at heterozygous sites are excluded altogether (Velasco et al., 2007). Single Molecule Real Time (SMRT) DNA sequencing (Pacific Biosciences, California, USA) has emerged as the leading technology for reconstructing highly contiguous, diploid assemblies of long, highly repetitive genomes that include phased information about heterozygous sites (Chin et al., 2013, 2016; Doi et al., 2014; Gordon et al., 2016; Huddleston et al., 2017; Pryszcz and Gabaldón, 2016; Ricker et al., 2016; Seo et al., 2016; Vij et al., 2016). Recently, we used *Vitis vinifera* cv. Cabernet Sauvignon to test the ability of SMRT reads to resolve both alleles at heterozygous sites in the genome (Chin et al., 2016). The assembly using the FALCON-Unzip assembly pipeline was significantly more contiguous than the original Pinot noir PN40024 assembly (contig N50 = 2.17 Mb) and provided the first phased sequences of the diploid genome of the species (Minio et al., 2017).

Transcriptome sequencing is a useful alternative to whole-genome reconstruction because it captures the functional genome. The ability to reconstruct the transcriptomes of different cultivars gives insight into cultivar-specific gene content that is otherwise unavailable (Da Silva et al., 2013; Jiao et al., 2015; Venturini et al., 2013). SMRT technology has recently enabled the investigation of expressed gene isoforms (Iso-Seq) in a variety of organisms, including a handful of plant species (Filichkin et al., 2018; Liu et al., 2017; Zulkapli et al., 2017); the long reads delivered by this method are full-length transcripts sequenced from their 5'-ends to polyadenylated tails (Dong et al., 2015; Gao et al., 2016; Kuo et al., 2017; Price and Gibas, 2017; Tombácz et al., 2016; Weirather et al., 2015; Workman et al., 2017). More importantly, Iso-Seq is an ideal technology for reconstructing a transcriptome without a reference sequence and for resolving isoforms (Honaas et al., 2016; Ju et al., 2016). Retrieving polyadenylated full-length molecules captures splice variants and some non-coding RNAs that can vary with cell-type (Swarup et al., 2016), developmental stage (Thatcher et al., 2016), or stress (Liu et al., 2016; Yan et al., 2012). Indeed, alternative splicing contributes to the complexity of the genome (Brett et al., 2002) that could not be definitively characterized without transcript information.

This study generated a comprehensive and detailed transcriptome composed of full-length transcripts using Iso-Seq. We show how error-correction with high coverage short-read data recovers an important fraction of the transcriptome otherwise lost by the standard Iso-Seq pipeline. Full-length transcripts were used to annotate the complete gene space of Cabernet Sauvignon, which led to the identification of transcripts associated with berry ripening unique to this cultivar. Full-length isoform information allowed the identification of multiple splice variants for most of the genes in the genome. We show that a transcriptome reference that includes splice variant information allows gene expression profiling at the isoform level and demonstrate the value of our approach by highlighting cases of contrasting expression patterns of isoforms at the same locus, whose differential expression during ripening would have been missed if mapping was carried out without isoform information.

MATERIALS AND METHODS

Plant material and RNA isolation

Grape berries from Cabernet Sauvignon FPS clone 08 were collected in Summer 2016 from vines grown in the Foundation Plant Services (FPS) Classic Foundation Vineyard (Davis, CA, USA). **Supplemental Table S1** provides weather information for the sampling days. Between 10 and 15 berries were sampled at pre-véraison, véraison, post-véraison, and at commercial maturity (harvest). The ripening stages were visually assessed based on color development and confirmed by measurements of soluble solids (**Figure 1; Supplemental Table S2**). On the day of sampling, berries were deseeded, frozen in liquid nitrogen, and ground to powder (skin and pulp). Total RNA was isolated using a Cetyltrimethyl Ammonium Bromide (CTAB)-based extraction protocol as described in Blanco-Ulate *et al*. (2013). RNA purity was evaluated with a Nanodrop 2000 spectrophotometer (Thermo Scientific, Hanover Park, IL, USA). RNA was quantified with a Qubit 2.0 Fluorometer using the RNA broad range kit (Life Technologies, Carlsbad, CA, USA). RNA integrity was assessed using electrophoresis and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Only RNA with an RNA integrity number (RIN) > 8.0 was used for SMRTbell library preparation.

Library preparation and sequencing

Total RNA from 4 biological replicates per developmental stage was pooled in equal amounts and 1 μg of the pooled RNA was used for cDNA synthesis and SMRTbell library construction using the SMARTer PCR cDNA synthesis kit (Clontech Laboratories, Inc. Mountain View, CA, USA). First-strand cDNA synthesis was performed using the SMRTScribe Reverse Transcriptase (Clontech Laboratories, Inc. Mountain View, CA, USA) and each developmental stage was individually barcoded (**Supplemental Table S3**). To minimize artifacts during large-scale amplification, a cycle optimization step was performed by collecting five 5 μl aliquots at 10, 12, 14, 16, and 18 PCR cycles. PCR reaction aliquots were loaded on an E-Gel pre-cast agarose gel 0.8 % (Invitrogen, Life Technologies, Carlsbad, CA, USA) to determine the optimal cycle number. Second-strand cDNA was synthesized and amplified using the Kapa HiFi PCR kit (Kapa Biosystems, Wilmington, MA, USA) with the 5' PCR primer IIA (Clontech Laboratories, Inc. Mountain View, CA, USA) following the manufacturer's instructions. Large-scale PCR was performed using the number of cycles determined during the optimization step (14 cycles). Barcoded double-stranded cDNAs were pooled at equal amounts and used for size selection. Size selection was carried out on a BluePippin (Sage Science, Beverly, MA, USA) and 1-2 kb, 2-3 kb, 3-6 kb, and 5-10 kb fractions were collected. After size selection, each fraction was PCR-enriched prior to SMRTbell template library preparation. cDNA SMRTbell libraries were prepared using 1-3 μg of PCR enriched size-selected samples, followed by DNA damage repair and SMRTbell ligation using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA). A second size selection was performed on the 3-6 Kb and 5-10 Kb fractions to remove short contaminating SMRTbell templates. A total of 8 SMRT cells were sequenced on a PacBio Sequel system (DNA Technologies Core, University of California, Davis, USA) producing 23.6 Gbp of raw reads. Demultiplexing, filtering, quality control, clustering and polishing of the Iso-Seq sequencing data were performed using SMRT Link (ver. 4.0.0) (**Supplemental Table S4**).

RNAseq libraries were prepared using the Illumina TruSeq RNA sample preparation kit v.2 (Illumina, San Diego, CA, USA), following the low-throughput protocol. Each biological replicate was barcoded individually. Final libraries were evaluated for quantity and quality with the High Sensitivity chip on a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). Libraries were sequenced in 100 bp paired-end mode, using an Illumina HiSeq4000 (DNA Technologies Core Facility, University of California, Davis, USA) producing 8,063,142 ± 2,040,693 reads/sample (**Supplemental Table S5**).

Iso-Seq read processing and transcriptome reconstruction

Cabernet Sauvignon primary contigs and haplotigs FALCON-unzip assembly (Chin et al., 2016) were used as genomic reference for *V. vinifera* cv. Cabernet Sauvignon FPS 08. Reads were aligned on the Cabernet Sauvignon genomic contigs using GMAP (ver. 2015-09-29) (Wu and Watanabe, 2005) using the following parameters "-B 4 -f 2 --split-output". Error rates were estimated from the identity and coverage of best alignments. Coding sequences (CDS) were identified using Transdecoder (Haas et al., 2013) as implemented in the PASA (ver. 2.1.0) (Haas et al., 2003) pipeline. Error correction was performed using LSC (ver. 2.0) (Au et al., 2012) using a minimum coverage threshold of 5 read (--short_read_coverage_threshold 5). Genome independent clustering of the isoforms was performed with Evidential Gene (Gilbert). Genome based clustering genome was performed using PASA (ver. 2.1.0) (Haas et al., 2003) with alignments carried out with BLAT (ver. 36x2) (Kent, 2002) and GMAP (Wu and Watanabe, 2005) with parameters reported in **Supplemental File S1** specifying that all the sequences are full-length transcripts.

Cabernet Sauvignon genome annotation

A repeat library was created *ad hoc* for Cabernet Sauvignon following the MAKER-P advanced repeat workflow (Maker-P - Repeat Library Construction -Advanced). MITEs were identified with MITEHunter (Han and Wessler, 2010); LTRs and TRIMs were identified with LTRharvest (Ellinghaus et al., 2008) and LTRdigest (Steinbiss et al., 2009). RepeatModeler (Smit and Hubley) and RepeatMasker (Smit et al.) were then used to combine and classify the information in a custom library of Cabernet Sauvignon repeats models. The custom models were finally combined with plant repeat models database to search for repetitive elements in the genome and in the transcriptome using RepeatMasker (Smit et al.). Iso-Seq reads were considered having a significant match with interspersed repeats when showing a coverage ≥ 75% and an identity ≥ 50%.

To create a high quality training set for *ab initio* gene prediction, PN40024 gene models were aligned on the primary Cabernet Sauvignon assembly with GMAP (Wu and Watanabe, 2005) and uniquely aligning models were kept only if: 1) the alignment length was at least 98% of the original model to ensure no major loss of exons; 2) models contained a full ORF coding for a protein with both identity and coverage ≥90% compared to the protein encoded by the aligned sequence; 3) splice sites were confirmed by Cabernet Sauvignon RNAseq data. In case of redundancy due to multiple different models encoding for the same protein, only one representative was kept.

Ab initio trainings and predictions were carried out with SNAP (ver. 2006-07-28) (Korf, 2004), Augustus (ver. 3.0.3) (Stanke et al., 2006), GeneMark-ES (ver. 4.32) (Lomsadze et al., 2005), GlimmerHMM (ver. 3.0.4) (Majoros et al., 2004), GeneID (ver. 1.4.4) (Parra et al., 2000) and Twinscan (Brent, 2008; Korf et al., 2001) (ver. 4.1.2, using TAIR10 annotation for Arabidopsis as informant species). MAKER-P (ver. 2.31.3) (Campbell et al., 2014a) was used to integrate the *ab initio* predictions with the experimental evidence listed in **Supplemental Table S8**. Only MAKER-P models showing an Annotation Edit Distance (AED) < 0.5 were kept.

Gene structure refinement was carried out with PASA (ver. 2.1.0) (Haas et al., 2003) using as evidence the Iso-Seq data, Clustered isoforms, corrected reads and raw reads, along with all the available grape transcriptomic data. Parameters can be found in **Supplemental Table File 2.** Types of alternative splicing were classified using AStalavista (ver. 3.0) (Foissac and Sammeth, 2007). For structure refinement, all RNAseq data were *de novo* assembled (separately for each sample) using a reference-based approach: HISAT2 alignments were used as input for Stringtie (ver. 1.1.3)(Pertea et al., 2015) without any *a priori* annotation and clustered in a non-redundant dataset using CD-HIT-EST (ver. 4.6)(Li and Godzik, 2006) with an identity threshold of 99%.

Functional annotation was performed with BLAST (Altschul et al., 1990) search using the RefSeq protein database (ftp://ftp.ncbi.nlm.nih.gov/refseq, retrieved January 17th, 2017). Functional domains were identified with InteProScan (ver. 5) (Jones et al., 2014). Enrichment analysis was done the BiNGO (ver. 2.4) (Maere et al., 2005) plugin tool in Cytoscape (ver. 3.0.3) (Shannon et al., 2003) with Biological Process GO categories. Overrepresented Biological Process GO categories were identified using a hypergeometric test with a significance threshold of *P*-value = 0.01. Non-coding RNAs were searched for with Infernal (ver. 1.1.2) (Nawrocki et al., 2009) using the Rfam database (ver. 12.2) (Nawrocki et al., 2015). Secondary overlapping alignments and structures with an *e*-value ≥ 0.01 were rejected. Hits on the minus strand of the Iso-Seq reads were rejected as well as matches that were truncated or covering less than 80% of the entire read.

Short-read alignment and expression profiling

Reads were aligned on transcript sequences using Bowtie2 (ver. 2.26) (Langmead and Salzberg, 2012). Differential gene expression analysis was performed for the 3 pairwise comparisons of consecutive growth stages using DESeq2 (ver. 1.16.1) (Love et al., 2014). Expression of RPKM > 1 was used as minimum threshold to consider a transcript expressed. *K*-means clustering was performed with MeV (ver. 4.9) (Saeed et al., 2003) using the 2,526 gene loci with one or more differentially regulated transcripts (*P*-value < 0.05) at least at one stage of berry development. Before processing, RPKM values were $\log_2$ transformed ($\log_2$ [RPKM average + 1]). *K*-means cluster analysis was performed with 100 iterations and a number of co-expressed clusters equal to three, four and five. The number of clusters was established using figure of merit (FOM) values (1–20 clusters, 100 iterations, **Supplemental Figure S8**). Genomic loci whose alternative transcripts were members of more than one co-expression cluster were considered as genomic loci whose alternative transcripts showed different patterns of gene expression during berry development (**Supplemental File S7 and Figure S9**).

RESULTS AND DISCUSSION

Full-length cDNA sequencing provides comprehensive representation of the Cabernet Sauvignon transcriptome during berry development

To obtain a comprehensive representation of the transcripts expressed during berry development, we isolated RNA from Cabernet Sauvignon berries (**Figure 1**) before the onset of ripening (4.35 ± 0.39 ºBrix), at (10.94 ± 0.26 ºBrix) and after véraison (18.38 ± 0.61 ºBrix), and at commercial ripeness (20.33 ± 0.76 ºBrix). To avoid loading bias, cDNAs were fractionated based on their length to produce four libraries at each developmental stage in size ranges of 1-2 kbp, 2-3 kbp, 3-6 kbp, or 5-10 kbp. Libraries derived from different developmental stages were barcoded and libraries with similar cDNA size were pooled together. Each library pool was sequenced independently on two SMRT cells of a Pacific Biosciences Sequel system generating a total of 23.6 Gbp. In parallel, the same samples were sequenced using Illumina technology to provide high coverage sequence information for error correction and for gene expression quantification (**Supplemental Table S5**). Demultiplexing, filtering and quality control of SMRT sequencing data were performed using SMRT Link as described in the Methods section. A total of 672,635 full-length non-chimeric (FLNC; **Figure 2**) reads with a maximum length of 14.6 kbp and an N50 of 3.5 kbp were generated (**Supplemental Table S4**). FLNC reads were further polished and clustered into 46,675 single representatives of expressed transcripts (henceforth, polished-clustered Iso-Seq reads or PCIRs) ranging from 400 bp to 8.8 kbp with an N50 of 3.6 kbp (**Supplemental Table S4**). The alignment of FLNC and PCIRs to the genomic DNA contigs of the same Cabernet Sauvignon clone (Chin et al., 2016; Minio et al., 2017) confirmed that sequence clustering and polishing successfully increased sequence accuracy, whose median values were 95.4% in FLNC and 99.6% in the PCIRs. The increase in sequence accuracy was also reflected by the significantly longer detectable coding sequences (CDS) in the PCIRs compared to the short and fragmented CDS found in the FLNC reads (**Figure 2**). The residual sequence discrepancy between PCIRs and the genomic contigs could be explained by heterozygosity and/or sequencing errors, but unexpectedly not by coverage (**Supplemental Figure S1**).

Over 18.5% of the FLNC reads did not cluster with any other reads and were discarded by the SMRT Link pipeline. When mapped on the genomic contigs, the uncorrected reads displayed a sequence accuracy that reflected the typical error rate of 10 - 20% of the technology (**Figure 1**) (Giordano et al., 2017; Koren et al., 2016; Zimin et al., 2017). High error rates also resulted in short and fragmented detectable CDS (**Figure 1**). To recover the information carried by these 124,185 uncorrected FLNC, which represented an important fraction of the transcriptome (see below), we error-corrected their sequences with LSC (Au et al., 2012) using the short reads generated using Illumina technology. As for the PCIRs, error correction resulted in greater sequence accuracy and longer CDS (**Figure 2**). This result confirmed the importance of integrating sequencing technologies that provide complementary benefits, long reads covering full-length transcripts of SMRT sequencing together with high coverage and accurate short Illumina reads.

PCIRs and error corrected FLNC (C-FLNC) were finally combined into a single dataset of 170,860 corrected Iso-Seq isoforms (CISIs). As low as 1.7 % (2,826) of the CISIs showed significant homology with interspersed repeats. LTRs and LINEs were the most abundant orders with 778 and 729 representatives, respectively. Chloroplast and mitochondria genes represented a small fraction of the CISIs with only 89 (0.05%) isoforms having a significant match (50% identity and mutual alignment coverage). Excluding these organellar transcribed isoforms, only 164 CISIs (0.1%) failed to align to the Cabernet Sauvignon genomic contigs (**Supplemental Table S6**), confirming the completeness of the genome reference and the negligible biological contamination of the berry samples.

By aligning the CISIs to the Cabernet Sauvignon genomic contigs, we determined the number of genomic loci derived from the different full-length transcripts. The 170,860 isoforms merged into a non-redundant set of 21,680 transcripts that mapped onto 13,402 different loci in the genome with a median number of alternative isoforms per locus of 1.6 ± 1.4. The CISIs were also clustered independently of any genome reference with EvidentialGene (Gilbert). A larger number of non-redundant transcripts (29,482) was retained by clustering, which nonetheless represented a similar number of genomic loci (13,596) when they were aligned to the genomic contigs. In combination, the two methods identified a total of 15,005 expressed loci with over 85% overlap and remarkable agreement in gene structure (~98%). Interestingly, only 25% of the loci were represented by CISIs at all ripening stages, while about one third were detected by Iso-Seq only at specific stages (**Figure 3A**) confirming the importance of collecting different stages of development to capture the complexity of the berry transcriptome (Reddy et al., 2013; Vitulo et al., 2014). As expected, transcripts present in the PCIRs dataset were found associated with higher expression levels than C-FLNC (**Figure 3B**). Importantly, the 15,005 loci identified by Iso-Seq represented about 82% of the total loci detectable by RNAseq using Illumina suggesting that only a minority of lowly expressed genes were not sequenced by Iso-Seq or were lost in the analysis (**Figure 3B**).

Error corrected Iso-Seq isoforms improve gene model prediction

Full-length cDNA sequencing has been recently shown to improve gene annotations in eukaryotic genomes (Chen et al., 2017; Clavijo et al., 2017; Hoang et al., 2017; Korlach et al., 2017; Li et al., 2017; Semler et al., 2017; Wang et al., 2018; Xu et al., 2017; Zhang et al., 2017). We incorporated the Iso-Seq information in the process of protein-coding gene prediction in the Cabernet Sauvignon genome as described in **Figure 4**. We first masked the repetitive regions of the genome using a custom-made library prepared for Cabernet Sauvignon containing MITE, LTR and TRIM information. We identified 412,994 repetitive elements for a total of 313 Mb, which masked ~53% of the genome (**Supplemental Table S7**). LTRs were the most abundant class covering over 240 Mb of the genome, with Gypsy and Copia families accounting for 136 Mb and 64.6 Mb, respectively. MAKER-P (Campbell et al., 2014b) was then used to identify putative protein-coding loci, combining the results of six *ab initio* predictors trained *ad hoc* with publicly available experimental evidences. *Ab initio* predictors were trained using a custom set of 4,000 randomly selected gene models out of the 5,636 high quality, non-redundant, and highly conserved gene models of the PN40024 V1 transcriptome (4,459 multiexonic and 1,177 monoexonic). Prediction processes produced over 296,000 models corresponding to 3.53 ± 4.98 CDSs per transcript with an average CDS length of 810 bp. Experimental evidence from public databases (**Supplemental Table S8**) were incorporated and used to validate the predicted models identifying 41,375 optimal distinct gene loci. Based on similarity to experimental evidence, we finally retained a total of 38,227 high-quality models (AED < 0.5).

To further refine the gene models, introduce alternative splicing events, and update the annotations of UTRs and CISIs, RNAseq Illumina data were introduced sequentially along with all the publicly available grapevine transcriptome assemblies. PCIRs permitted the annotation of 95 loci that were missed by MAKER-P and introduced 953 new alternative transcripts; C-FLNC reads introduced 468 new loci and 1,349 new alternative transcripts; and FLNC reads introduced 2,501 new alternative transcripts. RNAseq data and the other available grapevine transcriptomes allowed the annotation of 662 additional loci and 4,435 new alternative transcripts. At the end of the process, only 15,691 of the original MAKER-P gene models were not updated or modified by the refining procedure. The annotated models were compared to proteins in the RefSeq database and functional domains identified with InterProScan (Jones et al., 2014) in order to assign functional information to each isoform. The 2,477 predicted genes that did not show any similarity to known proteins and did not contain any known functional domain were removed. The final annotation consisted of 55,886 transcripts on 36,687 loci (**Table 1**), up to 29 kb in length with an average of 5.84 exons per transcript. The identified models encoded for proteins comparable in length with known grape proteins, with just 7.3% diverging more than 50% from their most similar and/or co-linear PN40024 protein models (**Supplemental Figure S1**). Gene ontology (GO) terms were assigned to 45,271 transcripts based on homology with protein domains in RefSeq and InterPro databases (**Supplemental Figure S3-S4, Supplemental file S3-S6**).

We scanned both CISIs and the genome assembly for non-coding RNA (ncRNA) using the covariance models of the Rfam database. In the CISI dataset, 182 isoforms were annotated as ncRNAs, all ascribed to ribosomal RNA, 145 of them attributed to the large subunit (clan CL00112) and 37 to the small subunit (clan CL00111). In the genomic contigs, we identified 3,238 non-overlapping putative ncRNA structures belonging to 236 different families covering a total of 638 kb of the assembly (**Supplemental Table S9**).

Overall, these results demonstrate that incorporating complete isoform sequencing information while annotating the gene space not only improved the predicted gene models, but also increased the number of identified coding sequences even when extensive RNAseq data is available. Importantly, because they represent entire molecules and not *de novo* assembled contigs, Iso-Seq reads provided direct experimental evidence supporting the structure and expression of alternative transcripts and UTRs. UTRs are important regulatory elements with strong influence on the post-transcriptional regulation of gene expression; they are hard to predict precisely *ab initio*. Here we show that, by incorporating Iso-Seq and multiple transcriptional evidences, we were able to annotate both 5' and 3' UTRs in the majority of the transcripts.

The Cabernet Sauvignon private transcriptome

Previous analyses of gene content in a limited number of grape cultivars showed that up to 10% of grape genes may not be shared between genotypes. Some of these dispensable genes are associated with cultivar specific characteristics (Da Silva et al., 2013). To identify cultivar-specific genes in Cabernet Sauvignon, all 55,886 annotated transcripts were compared to the predicted CDS of PN40024 (both V1 and V2; (Jaillon et al., 2007; Vitulo et al., 2014)), and the transcriptomes of Corvina (Venturini et al., 2013) and Tannat berries (Da Silva et al., 2013). Only the gene models that did not have a homologous copy in the other cultivars and did not align to PN40024 were considered putative cultivar specific genes. This additional filtering ensured that we did not overestimate the set of cultivar specific genes because of artifacts introduced by gene prediction in Cabernet Sauvignon and PN40024. Our analysis confirmed a mean unshared gene content of 5.25% ± 1.95% between grape cultivars (**Figure 5A**). The set of Cabernet Sauvignon specific isoforms comprised 585 isoforms distributed over 549 gene loci. These genes are involved in various cellular and metabolic processes of grapevine growth and berry ripening (**Figure 5B**). In particular, two GO terms were significantly enriched: "cellular amine metabolic process" and "oxidation reduction process" (adj. P-value ≤ 0.01). Among the genes involved in "cellular amine metabolic process" were two phenylalanine ammonia-lyases (PALs; *P0148F.500780.A*, *P0148F.500740.A*). Both PALs were expressed throughout ripening (RPKM > 1) and significantly up-regulated after véraison. Among the overrepresented Cabernet Sauvignon genes belonging to the "oxidation reduction process" was a putative flavonone 3-hydroxylase (F3H; *P0007F.293800.A*) that was significantly up-regulated between pre-véraison and véraison and between véraison and post-véraison. PAL and F3H are both enzymes involved in the phenylpropanoid and flavonoid biosynthetic pathways that produces polyphenols in berries. During grape berry development, F3H generates intermediate compounds in tannin biosynthesis during the herbaceous phase (pre-véraison), and in flavonol and anthocyanin biosynthesis after véraison (Castellarin et al., 2012). Interestingly, unlike F3H in PN40024 (*VIT_04s0023g03370*; (Castellarin et al., 2007)) and its homolog in Cabernet Sauvignon (*P0009F.302990.A*), this additional F3H paralog does not appear to be expressed before véraison (**Supplemental Figure S5**), suggesting that this particular F3H may contribute to berry coloration rather than astringency or bitterness. Similarly, other Cabernet Sauvignon specific genes were differentially expressed during ripening (65 transcripts) and exhibited different gene expression patterns, suggesting their involvement in berry ripening (**Figure 5C**). We can hypothesize that the expression of additional PALs and F3H as well as of other berry ripening associated genes contribute to Cabernet Sauvignon varietal attributes, such as berry color and organoleptic properties (Heymann and Noble, 1987; Robinson et al., 2014; Roujou de Boubee et al., 2000). For example, Cabernet Sauvignon berries accumulate more anthocyanins than Pinot Noir, Merlot and Cabernet franc berries (Mattivi et al., 2006) leading to wines denser in color (Cliff et al., 2007).

RNAseq data mapping on isoform-aware reference allows genome-wide expression profiling at the isoform resolution

The coding potential and complexity of eukaryotic organisms are known to be increased by the alternative splicing of precursor mRNAs from multiexon genes. Cabernet Sauvignon is no exception: over 23% percent of the 36,687 annotated genes had two or more alternative isoforms, with an average of 1.52 ± 1.27 alternative transcripts per locus, confirming previous reports in PN40024 (Vitulo et al., 2014). The frequency of splicing variant types was similar to those observed in other plant species (Reddy et al., 2013). Intron retention was the most abundant type, counting for over 44% (**Figure 6A**), similarly to what has been observed for rice (45-55%) (Zhang et al., 2015), Arabidopsis (30 - 64%) (Marquez et al., 2012; Reddy et al., 2013; Zhang et al., 2015) and maize (40 - 58%) (Wang et al., 2016; Zhang et al., 2015). Alternative acceptor sites (13%) and donor site (10%), and exon skipping (8%) were the other types of alternative splicing found in the Cabernet Sauvignon genome.

Illumina RNAseq reads were aligned to our new reference transcriptome that included all annotated isoforms to profile the transcriptional levels of all transcripts potentially expressed in the Cabernet Sauvignon genome. Comparison of the four stage transcriptomes showed an obvious distinction of the berry transcriptome before and after véraison (**Supplemental Figure S7**), confirming the well-known transcriptional reprogramming associated with the onset of ripening (Fasoli et al., 2012; Massonnet et al., 2017a). Gene expression analysis showed that 19,717 transcripts belonging to 11,902 loci were differentially expressed (adj. *P*-value < 0.05) at least once during berry development (**Supplemental File S7**). Transcriptional modulation was more intense between pre-véraison and véraison than post-véraison as observed in other studies (**Supplemental Figure 8)** (Massonnet et al., 2017a; Palumbo et al., 2014). Over 76% of the transcripts (82% of the genes) considered expressed following short-read sequencing (RPKM > 1) were detected using Iso-Seq. The transcripts not detected by Iso-Seq were expressed at extremely low levels, with just 1,997 loci (3.6 %) detected over the retention threshold of RPKM > 1. Expression levels measured by mapping on the predicted loci correlated well with the RNAseq results when reads were mapped directly on the CISIs (**Figure 3C**), further supporting the effectiveness of Iso-Seq to generate a reference transcriptome without relying on a genome assembly.

The inclusion of transcript variants in the RNAseq analysis allowed the profiling of each gene at the isoform resolution during berry ripening. We identified 252 loci whose alternative transcripts showed different expression patterns during berry development. **Figure 6** shows two such loci, encoding a *N*-carbamoylputrescine amidase (**Figure 6B**) and a putative hexokinase (**Figure 6C)**, that produce alternative transcripts with different patterns of expression during ripening. *N*-carbamoylputrescine amidase is an enzyme involved in the biosynthesis of polyamines, which are associated with numerous developmental and stress-related processes in plants, including grapevines (Panagiotis et al., 2012). Hexokinases play an important role in sugar sensing and signaling in grape berries (Lecourieux et al., 2014). Two transcripts associated with the same locus encoding a putative *N*-carbamoylputrescine amidase show contrasting patterns of expression; one was significantly up-regulated at véraison and one was significantly down-regulated post-véraison (**Figure 6B**). For the putative hexokinase (**Figure 6C**), one of its three transcripts was significantly up-regulated at véraison and two were significantly down-regulated at and post-véraison. For both genes, considering only a single transcript would have masked the complexity of this locus' usage during ripening.

Conclusions

This study demonstrates that Iso-Seq data can be used to compile a comprehensive reference transcriptome that represents most genes expressed in a tissue undergoing extensive transcriptional reprogramming. The integration of full-length cDNA sequencing with high coverage short read technology allowed to error correct and recover a large number of lowly expressed genes. As established in whole genome reconstruction, our results confirm that the utilization of different technologies with complementary characteristics can have synergistic benefits for the completeness and quality of the final genomic product. Although in this study genomic contigs were available, our results show that Iso-Seq can be used to generate a transcriptome reference without the need of a genome reference. In grapes, this approach can be particularly helpful by giving rapid access to cultivar specific transcripts. Nonetheless, the pipeline described here can be of even greater value for projects aiming to reconstruct the gene space in plant species with complex and large genomes that have not been resolved yet.

Acknowledgments

This work was supported in part by J. Lohr Vineyards and Wines, E. & J. Gallo Winery and was carried out in collaboration with UC Davis Chile with funds from the Chilean Economic Development Agency (CORFO; Project 13CEI2-21852), Viña San Pedro, and Viña Concha y Toro.

Author Contributions

DC and AM designed the experiment. BBU and MM coordinated and executed berry sampling. AM and MM carried out bioinformatics analyses. RFB prepared all sequencing libraries. DC, AM, AV, and MM wrote the manuscript.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability

Sequencing data are accessible through NCBI (SRA SRP132320) and other relevant datasets, such protein coding gene and repeat coordinates, can be retrieved from the Cantulab github repository (http://cantulab.github.io/data.html).

Supplementary material

Supplemental Figure S1: Heatmap representing the distribution of PCIRs in function of base accuracy and maximum measured expression level (RPKM). Accuracy level shows no correlation with isoform expression.

Supplemental Figure S2: Distribution of protein length deviation (percentage) between the annotated transcript and, on the x-axis, the co-linear PN40024 V1 gene model, and, on the y-axis, the most similar PN40024 V1 gene model.

Supplemental Figure S3: Distribution of hits for functional annotation. (A) Venn diagram of transcripts for which InterPro, Refseq Blast hit and GOslim information was available. (B) Number of transcripts for which a GO information was available using InterPro and BLAST against RefSeq databases

Supplemental Figure S4: Distribution of major metabolic process GO annotation available for Cabernet Sauvignon.

Supplemental Figure S5: Flavanone 3-hydroxylase alternative transcripts expression. (A) Schematic representation of flavanone 3-hydroxylase pathway. (B) Expression pattern of flavanone 3-hydroxylase alternative transcripts during berry ripening.

Supplemental Figure S6: Distribution of encoded protein length for expressed transcripts present in PCRIs dataset, C-FLNC isoforms dataset, or missing from any of the corrected Iso-Seq dataset.

Supplemental Figure S7: Heatmap of RNAseq expression distance across the different samples and replicates.

Supplemental Figure S8: Number of differentially expressed genes between consecutive developmental stages. In red are showed the up-regulated genes, in green the down-regulated.

Supplemental Figure S9: Line graph showing Figure of merit value (FOM) values for increasing number of clusters in the k-means clustering algorithm (1-20 clusters, 100 iterations; MeV v.4.9; Saeed et al., 2003).

Supplemental Figure S10: Overlap of gene loci whose alternative transcripts belong to more two or more different clusters when preforming k-means gene expression clustering analysis with 3, 4, or 5 as number of clusters.

Supplemental File S1: Alignment and annotation parameters used in PASA.

Supplemental File S2: Control files with parameters used for MAKER-P annotation.

Supplemental File S3: Results of *K*-means clustering. (A) List of the 2,526 gene with significant difference in expression (*P*-value < 0.05) in at least one comparison of ripening stages. (B) List of the 292 genes whose alternative transcripts showed different patterns of gene expression during berry development. *K*-means gene expression clustering analysis outputs when processing the analysis with three (C), four (D) and five (E) clusters.

Supplemental File S4: Cellular component GO annotation tree for Cabernet Sauvignon.

Supplemental File S5: Molecular function GO annotation tree for Cabernet Sauvignon.

Supplemental File S6: Biological process GO annotation tree for Cabernet Sauvignon.

Supplemental File S7: Expression profiling of Cabernet Sauvignon in berry ripening using RNAseq.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–10. doi:10.1016/S0022-2836(05)80360-2.

Amrine, K. C. H., Blanco-Ulate, B., Riaz, S., Pap, D., Jones, L., Figueroa-Balderas, R., et al. (2015). Comparative transcriptomics of Central Asian Vitis vinifera accessions reveals distinct defense strategies against powdery mildew. *Hortic. Res.* 2. doi:10.1038/hortres.2015.37.

Au, K. F., Underwood, J. G., Lee, L., and Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PLoS One* 7, e46679. doi:10.1371/journal.pone.0046679.

Blanco-Ulate, B., Amrine, K. C., Collins, T. S., Rivero, R. M., Vicente, A. R., Morales-Cruz, A., et al. (2015). Developmental and metabolic plasticity of white-skinned grape berries in response to Botrytis cinerea during noble rot. *Plant Physiol.* 169, pp.00852.2015. doi:10.1104/pp.15.00852.

Blanco-Ulate, B., Hopfer, H., Figueroa-Balderas, R., Ye, Z., Rivero, R. M., Albacete, A., et al. (2017). Red blotch disease alters grape berry development and metabolism by interfering with the transcriptional and hormonal regulation of ripening. *J. Exp. Bot.* 68, 1225–1238. doi:10.1093/jxb/erw506.

Blanco-Ulate, B., Vincenti, E., Powell, A. L. T., and Cantu, D. (2013). Tomato transcriptome and mutant analyses suggest a role for plant stress hormones in the interaction between fruit and Botrytis cinerea. *Front. Plant Sci.* 4, 1–16. doi:10.3389/fpls.2013.00142.

Böttcher, C., Harvey, K., Forde, C. G., Boss, P. K., and Davies, C. (2011). Auxin treatment of pre-veraison grape (Vitis vinifera L.) berries both delays ripening and increases the synchronicity of sugar accumulation. *Aust. J. Grape Wine Res.* 17, 1–8. doi:10.1111/j.1755-0238.2010.00110.x.

Bowers, J. E., and Meredith, C. P. (1997). The parentage of a classic wine grape, Cabernet Sauvignon. *Nat. Genet.* 16, 84–87. doi:10.1038/ng0597-84.

Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. 9, 62–73. doi:10.1038/nrg2220.

Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nat. Genet.* 30, 29–30. doi:10.1038/ng803.

Campbell, C. (2006). *The botanist and the vintner: how wine was saved for the world*. Algonquin Books.

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014a). *Genome annotation and curation using MAKER and MAKER-P*. doi:10.1002/0471250953.bi0411s48.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., et al. (2014b). MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* 164, 513–524. doi:10.1104/pp.113.230144.

Castellarin, S. D., Bavaresco, L., Falginella, L., Gonçalves, M. I. V. Z., and Di Gaspero, G. (2012). Phenolics in grape berry and key antioxidants | BenthamScience. *Biochem. Grape Berry*, 89–110. doi:10.2174/97816080536051120101.

Castellarin, S. D., Matthews, M. A., Di Gaspero, G., and Gambetta, G. A. (2007). Water deficits accelerate ripening and induce changes in gene expression regulating flavonoid biosynthesis in grape berries. *Planta* 227, 101–112. doi:10.1007/s00425-007-0598-8.

Chen, S. Y., Deng, F., Jia, X., Li, C., and Lai, S. J. (2017). A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. *Sci. Rep.* 7, 1–10. doi:10.1038/s41598-017-08138-z.

Chervin, C., El-Kereamy, A., Roustan, J. P., Latché, A., Lamon, J., and Bouzayen, M. (2004). Ethylene seems required for the berry development and ripening in grape, a non-climacteric fruit. *Plant Sci.* 167, 1301–1305. doi:10.1016/j.plantsci.2004.06.026.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–9. doi:10.1038/nmeth.2474.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi:10.1038/nmeth.4035.

Cipriani, G., Spadotto, A., Jurman, I., Gaspero, G. Di, Crespan, M., Meneghetti, S., et al. (2010). The SSR-based molecular profile of 1005 grapevine (Vitis vinifera L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theor. Appl. Genet.* 121, 1569–1585. doi:10.1007/s00122-010-1411-9.

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* 27, 885–896. doi:10.1101/gr.217117.116.

Cliff, M. A., King, M. C., and Schlosser, J. (2007). Anthocyanin, phenolic composition, colour measurement and sensory analysis of BC commercial red wines. *Food Res. Int.* 40, 92–100. doi:10.1016/j.foodres.2006.08.002.

Corso, M., Vannozzi, A., Maza, E., Vitulo, N., Meggio, F., Pitacco, A., et al. (2015). Comprehensive transcript profiling of two grapevine rootstock genotypes contrasting in drought susceptibility links the phenylpropanoid pathway to enhanced tolerance. *J. Exp. Bot.* 66, 5739–5752. doi:10.1093/jxb/erv274.

Da Silva, C., Zamperin, G., Ferrarini, A., Minio, A., Dal Molin, A., Venturini, L., et al. (2013). The high polyphenol content of grapevine cultivar Tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell* 25, 4777–4788. doi:10.1105/tpc.113.118810.

Davies, C., Boss, P. K., and Robinson, S. P. (1997). Treatment of crape berries, a nonclimacteric fruit with a synthetic auxin, retards ripening and alters the expression of developmentally regulated genes. *Plant Physiol* 11, 55–1. doi:10.1104/PP.115.3.1155.

Deluc, L. G., Quilici, D. R., Decendit, A., Grimplet, J., Wheatley, M. D., Schlauch, K. A., et al. (2009). Water deficit alters differentially metabolic pathways affecting important flavor and quality traits in grape berries of Cabernet Sauvignon and Chardonnay. *BMC Genomics* 10, 212. doi:10.1186/1471-2164-10-212.

Di Gaspero, G., Cipriani, G., Marrazzo, M. T., Andreetta, D., Prado Castro, M. J., Peterlunger, E., et al. (2005). Isolation of (AC)n-microsatellites in Vitis vinifera L. and analysis of genetic background in grapevines under marker assisted selection. *Mol. Breed.* 15, 11–20. doi:10.1007/s11032-004-1362-4.

Di Genova, A., Almeida, A. M., Muñoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., et al. (2014). Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.* 14, 7. doi:10.1186/1471-2229-14-7.

Doi, K., Monjo, T., Hoang, P. H., Yoshimura, J., Yurino, H., Mitsui, J., et al. (2014). Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics* 30, 815–822. doi:10.1093/bioinformatics/btt647.

Dong, L., Liu, H., Zhang, J., Yang, S., Kong, G., Chu, J. S. C., et al. (2015). Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. *BMC Genomics* 16, 1039. doi:10.1186/s12864-015-2257-y.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* 9, 18. doi:10.1186/1471-2105-9-18.

Fasoli, M., Dal Santo, S., Zenoni, S., Tornielli, G. B., Farina, L., Zamboni, A., et al. (2012). The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. *Plant Cell* 24, 3489–505. doi:10.1105/tpc.112.100230.

Filichkin, S. A., Hamilton, Mi., Dharmawardhana, P. D., Singh, S. K., Sullivan, C., Ben-Hur, A., et al. (2018). Abiotic stresses modulate landscape of poplar transcriptome via alternative splicing, differential intron retention, and isoform ratio switching. *Front. Plant Sci.* 9. doi:10.3389/fpls.2018.00005.

Foissac, S., and Sammeth, M. (2007). ASTALAVISTA: Dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.* 35, 297–299. doi:10.1093/nar/gkm311.

Fortes, A. M., Teixeira, R. T., and Agudelo-Romero, P. (2015). Complex interplay of hormonal signals during grape berry ripening. *Molecules* 20, 9326–9343. doi:10.3390/molecules20059326.

Gambino, G., Dal Molin, A., Boccacci, P., Minio, A., Chitarra, W., Avanzato, C. G., et al. (2017). Whole-genome sequencing and SNV genotyping of "Nebbiolo" (Vitis vinifera L.) clones. *Sci. Rep.* 7, 17294. doi:10.1038/s41598-017-17405-y.

Gao, S., Ren, Y., Sun, Y., Wu, Z., Ruan, J., He, B., et al. (2016). PacBio full-length transcriptome profiling of insect mitochondrial gene expression. *RNA Biol.* 13, 1–6. doi:10.1080/15476286.2016.1197481.

Gilbert, D. G. EvidentialGene: Evidence Directed gene predictions for eukaryotes. Available at:

http://arthropods.eugenes.org/EvidentialGene/.

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., et al. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci. Rep.* 7, 1–10. doi:10.1038/s41598-017-03996-z.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 108, 1513–8. doi:10.1073/pnas.1017351108.

Golicz, A. A., Batley, J., and Edwards, D. (2016a). Towards plant pangenomics. *Plant Biotechnol. J.* 14, 1099–1105. doi:10.1111/pbi.12499.

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H. R., Martinez, P. A., et al. (2016b). The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* 7, 1–8. doi:10.1038/ncomms13390.

Gordon, D., Huddleston, J., Chaisson, M. J., Hill, C. M., Kronenberg, Z. N., Munson, K. M., et al. (2016). Long-read sequence assembly of the gorilla genome. *Science* 352, aae0344. doi:10.1126/science.aae0344.

Haas, B. J., Delcher, A. L., Mount S.M., S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi:10.1093/nar/gkg770.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi:10.1038/nprot.2013.084.

Han, Y., and Wessler, S. R. (2010). MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 38, 1–8. doi:10.1093/nar/gkq862.

Heymann, H., and Noble, a C. (1987). Descriptive analysis of commercial Cabernet Sauvignon wines from California. *Am. J. Enol. Vitic.* 38, 41–44.

Hoang, N. V., Furtado, A., Mason, P. J., Marquardt, A., Kasirajan, L., Thirugnanasambandam, P. P., et al. (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics* 18, 1–22. doi:10.1186/s12864-017-3757-8.

Honaas, L. A., Wafula, E. K., Wickett, N. J., Der, J. P., Zhang, Y., Edger, P. P., et al. (2016). Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. *PLoS One* 11, 1–42. doi:10.1371/journal.pone.0146062.

Hopper, D. W., Ghan, R., Schlauch, K. A., and Cramer, G. R. (2016). Transcriptomic network analyses of leaf dehydration responses identify highly connected ABA and ethylene signaling hubs in three grapevine species differing in drought tolerance. *BMC Plant Biol.* 16, 118. doi:10.1186/s12870-016-0804-6.

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., et al. (2012). HaploMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* 22, 1581–1588. doi:10.1101/gr.133652.111.

Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27, 677–685. doi:10.1101/gr.214007.116.

Ibáñez, J., Vargas, A. M., Palancar, M., Borrego, J., and De Andrés, M. T. (2009). Genetic relationships among table-grape varieties. *Am. J. Enol. Vitic.* 60, 35–42.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi:10.1038/nature06148.

Jiao, C., Gao, M., Wang, X., and Fei, Z. (2015). Transcriptome characterization of three wild Chinese Vitis uncovers a large number of distinct disease related genes. *BMC Genomics* 16, 223. doi:10.1186/s12864-015-1442-3.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–40.

Ju, C., Zhao, Z., and Wang, W. (2016). Efficient approach to correct read alignment for pseudogene abundance estimates. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* XX, 1–1. doi:10.1109/TCBB.2016.2591533.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–1395. doi:10.1101/gr.170720.113.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res.* 12, 656–64. doi:10.1101/gr.229202. Article published online before March 2002.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2016). Canu : scalable and accurate long- - - read assembly via adaptive k - - - mer weighting and repeat separation. 1–35. doi:10.1101/gr.215087.116.Freely.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5, 59. doi:10.1186/1471-2105-5-59.

Korf, I., Flicek, P., Duan, D., and Brent, M. R. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* 17 Suppl 1, S140-8.

Korlach, J., Gedman, G., Kingan, S. B., Chin, C. S., Howard, J. T., Audet, J. N., et al. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience* 6, 1–16. doi:10.1093/gigascience/gix085.

Koyama, K., Sadamatsu, K., and Goto-Yamamoto, N. (2010). Abscisic acid stimulated ripening and gene expression in berry skins of the Cabernet Sauvignon grape. *Funct. Integr. Genomics* 10, 367–381. doi:10.1007/s10142-009-0145-8.

Kuo, R. I., Tseng, E., Eory, L., Paton, I. R., Archibald, A. L., and Burt, D. W. (2017). Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* 18, 323. doi:10.1186/s12864-017-3691-9.

Lacombe, T., Boursiquot, J. M., Laucou, V., Di Vecchi-Staraz, M., Péros, J. P., and This, P. (2013). Large-scale parentage analysis in an extended set of grapevine cultivars (Vitis vinifera L.). *Theor. Appl. Genet.* 126, 401–414. doi:10.1007/s00122-012-1988-2.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.

Lecourieux, F., Kappel, C., Lecourieux, D., Serrano, A., Torres, E., Arce-Johnson, P., et al. (2014). An update on sugar transport and signalling in grapevine. *J. Exp. Bot.* 65, 821–832. doi:10.1093/jxb/ert394.

Lecourieux, F., Kappel, C., Pieri, P., Charon, J., Pillet, J., Hilbert, G., et al. (2017). Dissecting the biochemical and transcriptomic effects of a locally applied heat treatment on developing Cabernet Sauvignon grape berries. *Front. Plant Sci.* 8. doi:10.3389/fpls.2017.00053.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–9. doi:10.1093/bioinformatics/btl158.

Li, Y., Wei, W., Feng, J., Luo, H., Pi, M., Liu, Z., et al. (2017). Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina- and SMRT-based RNA-seq datasets. *DNA Res.* 0, 1–10. doi:10.1093/dnares/dsx038.

Liu, J., Chen, X., Liang, X., Zhou, X., Yang, F., Liu, J., et al. (2016). Alternative splicing of rice WRKY62 and WRKY76 transcription factor genes in pathogen defense. *Plant Physiol.* 171, pp.01921.2015. doi:10.1104/pp.15.01921.

Liu, X., Mei, W., Soltis, P. S., Soltis, D. E., and Barbazuk, W. B. (2017). Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. *Mol. Ecol. Resour.* 17, 1243–1256. doi:10.1111/1755-0998.12670.

Lodhi, M. A., and Reisch, B. I. (1995). Nuclear DNA content of Vitis species, cultivars, and other genera of the Vitaceae. *Theor. Appl. Genet.* 90, 11–6. doi:10.1007/BF00220990.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* 33, 6494–6506. doi:10.1093/nar/gki937.

Lopes, M. S., Sefc, K. M., Eiras Dias, E., Steinkellner, H., Laimer Câmara Machado, M., and Câmara Machado, A. (1999). The use of microsatellites for germplasm management in a Portuguese grapevine collection. *Theor. Appl. Genet.* 99, 733–739. doi:10.1007/s001220051291.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. 1–21. doi:10.1101/002832.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21, 3448–3449. doi:10.1093/bioinformatics/bti551.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi:10.1093/bioinformatics/bth315.

Maker-P - Repeat Library Construction -Advanced Available at: http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced.

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.* 22, 1184–1195. doi:10.1101/gr.134106.111.

Massonnet, M., Fasoli, M., Tornielli, G. B., Altieri, M., Sandri, M., Zuccolotto, P., et al. (2017a). Ripening transcriptomic program in red and white grapevine varieties correlates with berry skin anthocyanin accumulation. *Plant Physiol.* 174, 2376–2396. doi:10.1104/pp.17.00311.

Massonnet, M., Figueroa Balderas, R., Galarneau, E., Miki, S., Lawrence, D., Sun, Q., et al. (2017b). Neofusicoccum parvum colonization of the grapevine woody stem triggers asynchronous host responses at the site of infection and in the leaves. *Front. Plant Sci.* 8, 1117. doi:10.3389/FPLS.2017.01117.

Mattivi, F., Guzzon, R., Vrhovsek, U., Stefanini, M., and Velasco, R. (2006). Metabolite profiling of grape: Flavonols and anthocyanins. *J. Agric. Food Chem.* 54, 7692–7702. doi:10.1021/jf061538c.

McGovern, P. E., Katz, S. H., and Fleming, S. J. (2003). *The origins and ancient history of wine: Food and nutrition in history and antropology*. Routledge doi:10.4324/9780203392836.

Michael, T. P., and Jackson, S. (2013). The first 50 plant genomes. *Plant Genome* 6, 0. doi:10.3835/plantgenome2013.03.0001in.

Minio, A., Lin, J., Gaut, B. S., and Cantu, D. (2017). How Single Molecule Real-Time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci.* 8, 1–6. doi:10.3389/fpls.2017.00826.

Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., Aradhya, M. K., et al. (2011). Genetic structure and domestication history of the grape. *Proc. Natl. Acad. Sci. U. S. A.* 108, 3530–3535. doi:10.1073/pnas.1009363108.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., et al. (2015). Rfam 12 . 0 : updates to the RNA families database. 43, 130–137. doi:10.1093/nar/gku1063.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25, 1335–7. doi:10.1093/bioinformatics/btp157.

Ohmi, C., Wakana, A., Shiraishi, S., and Alexandria, M. (1993). Study of the parentage of grape cultivars by genetic interpretation of GPI-2 and PGM-2 isozymes. *Euphytica* 65, 195–202.

Palumbo, M. C., Zenoni, S., Fasoli, M., Massonnet, M., Farina, L., Castiglione, F., et al. (2014). Integrated network analysis identifies fight-club nodes as a class of hubs encompassing key putative switch genes that induce major transcriptome reprogramming during grapevine development. *Plant Cell Online* 26, 4617–4635. doi:10.1105/tpc.114.133710.

Panagiotis, M. N., Aziz, A., and Kalliopie, R. A. A. (2012). "Polyamines and grape berry development," in *The Biochemistry of the Grape Berry*, eds. M. N. Panagiotis, A. Aziz, and R.-A. A. Kalliopi (BENTHAM SCIENCE PUBLISHERS), 137–159. doi:10.2174/978160805360511201010137.

Parra, G., Blanco, E., and Guigó, R. (2000). GeneID in Drosophila. *Genome Res.* 10, 511–5.

Pastore, C., Zenoni, S., Fasoli, M., Pezzotti, M., Tornielli, G. B., and Filippetti, I. (2013). Selective defoliation affects plant growth, fruit transcriptional ripening program and flavonoid metabolism in grapevine. *BMC Plant Biol.* 13, 30. doi:10.1186/1471-2229-13-30.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–5. doi:10.1038/nbt.3122.

Price, A., and Gibas, C. (2017). The quantitative impact of read mapping to non-native reference genomes in comparative RNA-Seq studies. *PLoS One*, 1–21. doi:10.1371/ journal.pone.0180904.

Pryszcz, L. P., and Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44, e113. doi:10.1093/nar/gkw294.

Qian, M., Baoju, W., Xiangpeng, L., Xin, S., Lingfei, S., Haifeng, J., et al. (2016). Comparison and verification of the genes involved in ethylene biosynthesis and signaling in apple, grape, peach, pear and strawberry. *Acta Physiol. Plant.* 38, 1–22. doi:10.1007/s11738-016-2067-0.

Reddy, A. S. N., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell* 25, 3657–3683. doi:10.1105/tpc.113.117523.

Ricker, N., Shen, S. Y., Goordial, J., Jin, S., and Fulthorpe, R. R. (2016). PacBio SMRT assembly of a complex multi-replicon genome reveals chlorocatechol degradative operon in a region of genome plasticity. *Gene* 586, 239–247. doi:10.1016/j.gene.2016.04.018.

Robinson, A. L., Boss, P. K., Solomon, P. S., Trengove, R. D., Heymann, H., and Ebeler, S. E. (2014). Origins of grape and wine aroma. Part 1. Chemical components and viticultural impacts. *Am. J. Enol. Vitic.* 65, 1–24. doi:10.5344/ajev.2013.12070.

Roujou de Boubee, D., Van Leeuwen, C., and Dubourdieu, D. (2000). Organoleptic impact of 2-methoxy-3-isobutylpyrazine on red Bordeaux and Loire wines. Effect of environmental conditions on concentrations in grapes during ripening. *J. Agric. Food Chem.* 48, 4830–4834. doi:10.1021/jf000181o.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374–8. doi:12613259.

Safonova, Y., Bankevich, A., and Pevzner, P. A. (2015). dipSPAdes: Assembler for highly polymorphic diploid genomes. *J. Comput. Biol. A J. Comput. Mol. Cell Biol.* 22, 528–545. doi:10.1089/cmb.2014.0153.

Savoi, S., Wong, D. C. J., Arapitsas, P., Miculan, M., Bucchetti, B., Peterlunger, E., et al. (2016). Transcriptome and metabolite profiling reveals that prolonged drought modulates the phenylpropanoid and terpenoid pathway in white grapes (Vitis vinifera L.). *BMC Plant Biol.* 16, 67. doi:10.1186/s12870-016-0760-1.

Savoi, S., Wong, D. C. J., Degu, A., Herrera, J. C., Bucchetti, B., Peterlunger, E., et al. (2017). Multi-Omics and integrated network analyses reveal new insights into the systems relationships between metabolites, structural genes, and transcriptional regulators in developing grape berries (Vitis vinifera L.) exposed to water deficit. *Front. Plant Sci.* 8, 1–19. doi:10.3389/fpls.2017.01124.

Sefc, K. M., Steinkellner, H., Glössl, J., Kampfer, S., and Regner, F. (1998). Reconstruction of a grapevine pedigree by microsatellite analysis. *Theor. Appl. Genet.* 97, 227–231. doi:10.1007/s001220050889.

Semler, M. R., Wiseman, R. W., Karl, J. A., Graham, M. E., Gieger, S. M., and O'Connor, D. H. (2017). Novel full-length major histocompatibility complex class I allele discovery and haplotype definition in pig-tailed macaques. *Immunogenetics*, 1–19. doi:10.1007/s00251-017-1042-2.

Seo, J., Rhie, A., Kim, J., Lee, S., Sohn, M., Kim, C.-U., et al. (2016). De novo assembly and phasing of a Korean human genome. *Nature* 538, 243–247. doi:10.1038/nature20098.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–504. doi:10.1101/gr.1239303.

Smit, A. F. A., and Hubley, R. RepeatModeler Open-1.0. 2008–2015. Available at: http://www.repeatmasker.org.

Smit, A. F. A., Hubley, R., and Green, P. RepeatMasker Open-4.0. 2013–2015. Available at: http://www.repeatmasker.org.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435-9.

Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* 37, 7002–7013. doi:10.1093/nar/gkp759.

Strefeler, M. S., Weeden, N. F., and Reisch, B. I. (1992). Inheritance of chloroplast DNA in two full-sib Vitis populations. *Vitis* 31, 183–187.

Swarup, R., Crespi, M., and Bennett, M. J. (2016). One gene, many proteins: mapping cell-specific alternative splicing in plants. *Dev. Cell* 39, 383–385. doi:10.1016/j.devcel.2016.11.002.

Symons, G. M., Chua, Y.-J., Ross, J. J., Quittenden, L. J., Davies, N. W., and Reid, J. B. (2012).

Hormonal changes during non-climacteric ripening in strawberry. *J. Exp. Bot.* 63, 4741–4750. doi:10.1093/jxb/ers147.

Tapia, A. M., Cabezas, J. A., Cabello, F., Lacombe, T., Martínez-Zapater, J. M., Hinrichsen, P., et al. (2007). Determining the Spanish origin of representative ancient American grapevine varieties. *Am. J. Enol. Vitic.* 58, 242–251.

Thatcher, S. R., Danilevskaya, O. N., Meng, X., Beatty, M., Zastrow-Hayes, G., Harris, C., et al. (2016). Genome-wide analysis of alternative splicing during development and drought stress in maize. *Plant Physiol.* 170, 586–599. doi:10.1104/pp.15.01267.

Tombácz, D., Csabai, Z., Oláh, P., Balázs, Z., Likó, I., Zsigmond, L., et al. (2016). Full-length isoform sequencing reveals novel transcripts and substantial transcriptional overlaps in a herpesvirus. *PLoS One* 11, 1–29. doi:10.1371/journal.pone.0162868.

Unwin, T. (2005). *Wine and the vine: an historical geography of viticulture and the wine trade.* Routledge.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* 2, e1326. doi:10.1371/journal.pone.0001326.

Venturini, L., Ferrarini, A., Zenoni, S., Tornielli, G. B. G. B., Fasoli, M., Santo, S. D. S. D., et al. (2013). De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics* 14, 41. doi:10.1186/1471-2164-14-41.

Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., Van Heusden, P., et al. (2016). Chromosomal-level assembly of the asian seabass genome using long sequence reads and multi-layered scaffolding. *PLOS Genet.* 12, e1005954. doi:10.1371/journal.pgen.1005954.

Vitulo, N., Forcato, C., Carpinelli, E. C., Telatin, A., Campagna, D., D'Angelo, M., et al. (2014). A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biol.* 14, 99. doi:10.1186/1471-2229-14-99.

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7, 11708. doi:10.1038/ncomms11708.

Wang, M., Wang, P., Liang, F., Ye, Z., Li, J., Shen, C., et al. (2018). A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *New Phytol.* 217, 163–178. doi:10.1111/nph.14762.

Weirather, J. L., Afshar, P. T., Clark, T. A., Tseng, E., Powers, L. S., Underwood, J. G., et al. (2015). Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res.* 43, 1–12. doi:10.1093/nar/gkv562.

Westering, J. Van, and Ravenscroft, N. (2001). Wine tourism, culture and the everyday: A theoretical note. *Tour. Hosp. Res.* 3, 149–162. doi:10.1177/146735840100300206.

Workman, R. E., Myrka, A. M., Tseng, E., Wong, G. W., Welch, K. C., and Timp, W. (2017). Single molecule, full-length transcript sequencing provides insight into the extreme metabolism of ruby-throated hummingbird Archilochus colubris. *bioRxiv*, 117218. doi:10.1101/117218.

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–75. doi:10.1093/bioinformatics/bti310.

Xi, H., Ma, L., Liu, G., Wang, N., Wang, J., Wang, L., et al. (2014). Transcriptomic analysis of grape (Vitis vinifera L.) leaves after exposure to ultraviolet C irradiation. *PLoS One* 9. doi:10.1371/journal.pone.0113772.

Xu, Q., Zhu, J., Zhao, S., Hou, Y., Li, F., Tai, Y., et al. (2017). Transcriptome profiling using single-molecule direct RNA sequencing approach for in-depth understanding of genes in secondary metabolism pathways of Camellia sinensis. *Front. Plant Sci.* 8, 1–11. doi:10.3389/fpls.2017.01205.

Yan, K., Liu, P., Wu, C. A., Yang, G. D., Xu, R., Guo, Q. H., et al. (2012). Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in Arabidopsis thaliana. *Mol. Cell* 48, 521–531. doi:10.1016/j.molcel.2012.08.032.

Zenoni, S., Dal Santo, S., Tornielli, G. B., D'Incà, E., Filippetti, I., Pastore, C., et al. (2017). Transcriptional responses to pre-flowering leaf defoliation in grapevine berry from different growing sites, years, and genotypes. *Front. Plant Sci.* 8, 1–21. doi:10.3389/fpls.2017.00630.

Zhang, C., Yang, H., and Yang, H. (2015). Evolutionary character of alternative splicing in plants.

Bioinform. Biol. Insights 9s1, 47–52. doi:10.4137/BBI.S33716.

Zhang, S. J., Wang, C., Yan, S., Fu, A., Luan, X., Li, Y., et al. (2017). Isoform evolution in primates through independent combination of alternative RNA processing events. *Mol. Biol. Evol.* 34, 2453–2468. doi:10.1093/molbev/msx212.

Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D., and Gaut, B. S. (2017). Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. *Proc. Natl. Acad. Sci. U. S. A.* 114, 11715–11720. doi:10.1073/pnas.1709257114.

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marçais, G., et al. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* 27, 787–792. doi:10.1101/gr.213405.116.

Zulkapli, M. M. izzuddin, Rosli, M. A. F., Salleh, F. I. M., Mohd Noor, N., Aizat, W. M., and Goh, H. H. (2017). Iso-Seq analysis of Nepenthes ampullaria, Nepenthes rafflesiana and Nepenthes × hookeriana for hybridisation study in pitcher plants. *Genomics Data* 12, 130–131. doi:10.1016/j.gdata.2017.05.003.

TABLES AND FIGURE LEGENDS

Table 1. Summary statistics of the Cabernet Sauvignon genome annotation after refinement with experimental evidence.

Figure 1. Biological material sampled for transcriptome sequencing. (A) Boxplots showing the concentration of soluble solids in the berries at different stages of development. Representative pictures of Cabernet Sauvignon berry clusters are shown. (B) Size distribution of the Iso-Seq libraries obtained by size fractionation of cDNA.

Figure 2. Diagram depicting the main steps of analysis of the Iso-Seq reads. Raw Iso-Seq reads were processed following the standard SMRT Link pipeline for Iso-Seq data to obtain Full-Length Non-Chimeric (FLNC) reads, and clustered and corrected isoform reads (PCIRs). FLNC reads that did not cluster were error corrected using RNAseq data (C-FLNC). The final dataset described in this study comprised both PCIRs and C-FLNC reads. For each step, sequencing accuracy and CDS length distributions are reported.

Figure 3. Expression profiling of the grape transcriptome using Iso-Seq and RNAseq data. (A) Overlap of loci detected by Iso-Seq in the different stages of berry development. (B) Distribution of the expression level of PCIR, FLNC and C-FLNC datasets measured by RNAseq. (C) Correlation of expression levels between RNAseq conducted by mapping on genomic loci and directly on CIRIs.

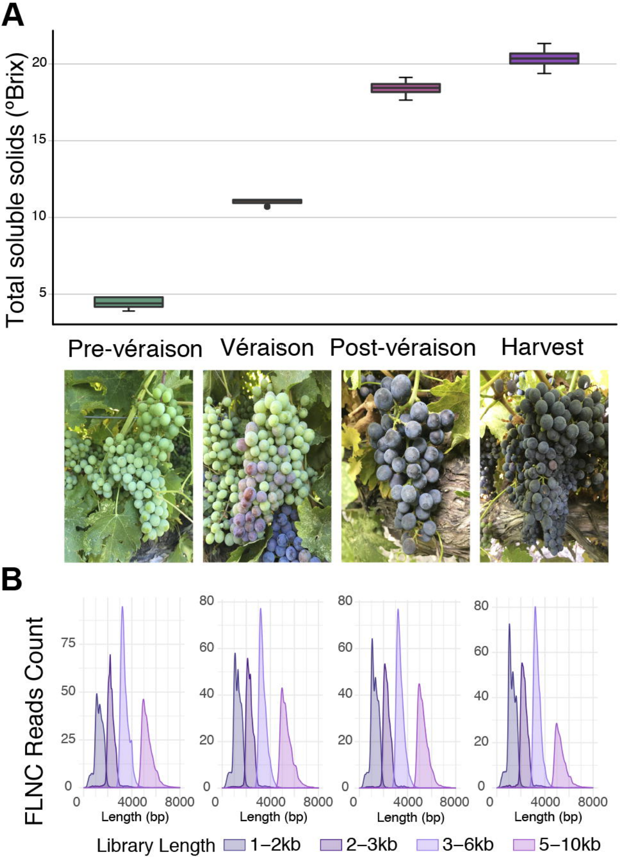
Figure 4. Genome annotation workflow with integration of Iso-Seq data.

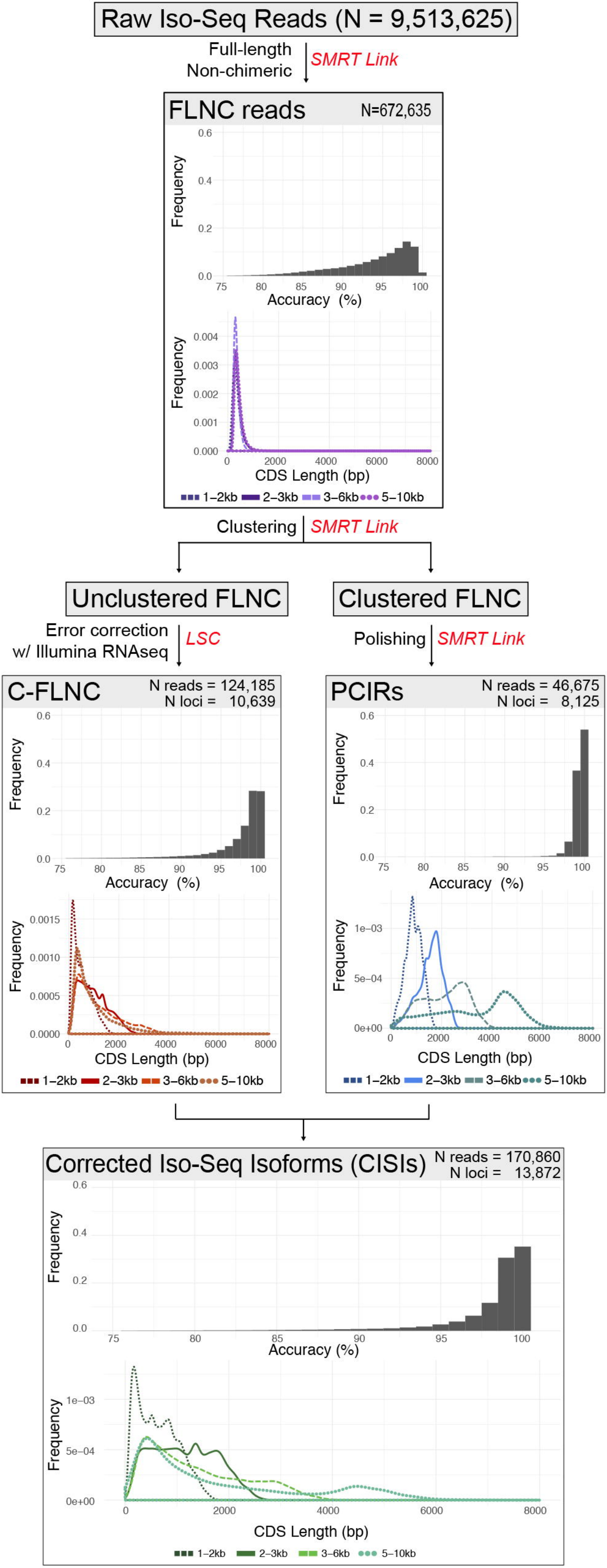
Figure 5. Characterization of unshared gene content with other cultivars. (A) Transcript overlap between Cabernet Sauvignon, PN40024 V1 and V2, Corvina and Tannat. (B) Overrepresented GO terms among the Cabernet Sauvignon cultivar-specific isoforms. Size of the nodes is related to the cardinality of the genes associated with the functional category, while color is proportional to the *P*-value of the enrichment for the category (Benjamini and Hochberg corrected *P*-value < 0.01). (C) Transcriptional modulation of the Cabernet Sauvignon-specific isoforms expressed during berry development. Isoforms were clustered by gene modulation pattern based on a hierarchical cluster analysis using the Ward agglomeration method and Pearson's correlation distance as the metric. Heat maps represent the gene expression level (RPKM) of Cabernet Sauvignon cultivar-specific isoforms at the four growth stages.

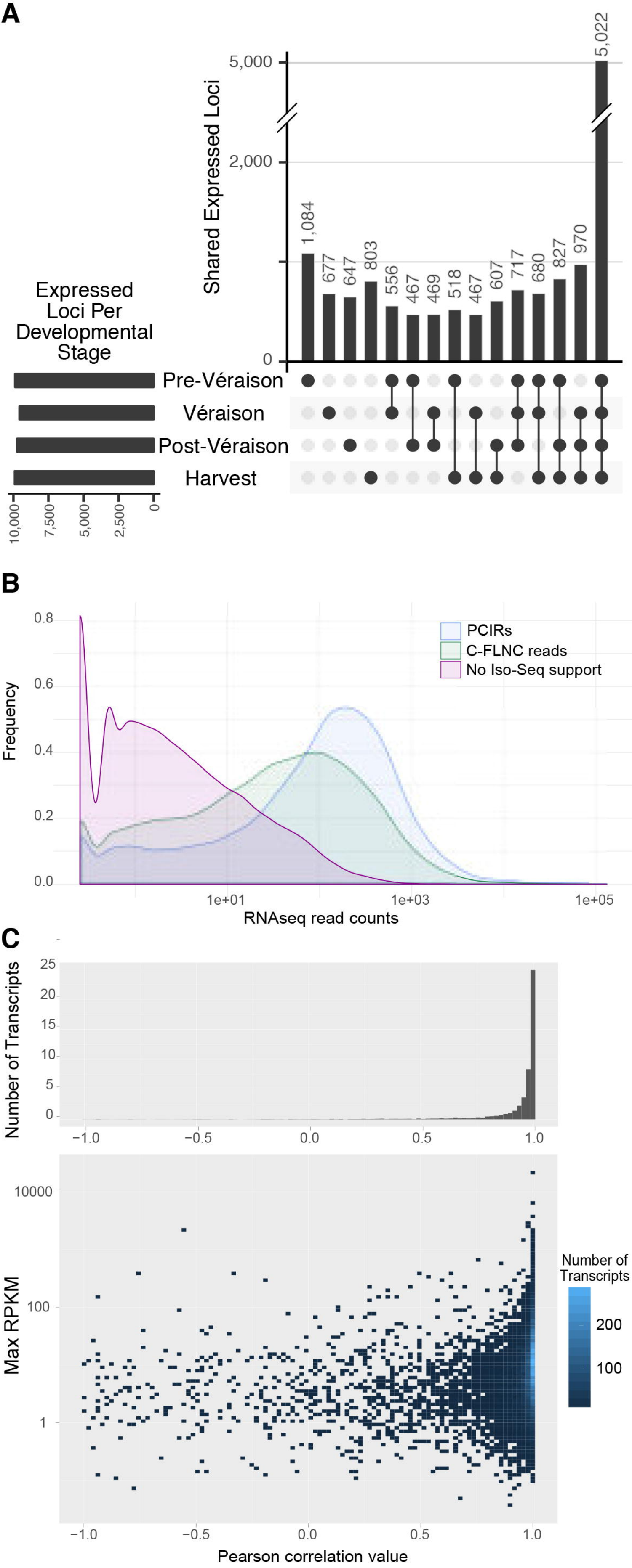
Figure 6. Alternative splicing variants in Cabernet Sauvignon. (A) Relative abundance of the different types of splicing variants. (B, C) Expression profiles of two genes whose annotated alternative ranscripts present a differential transcription modulation during berry development. Expression is calculated over the transcriptome comprising all alternative transcripts per locus and over a reduced representation of the annotation comprising only one transcript per locus. P0029F.365630.A and P0009F.303060.A encode a *N*-carbamoylputrescine amidase and a hexokinase, respectively.

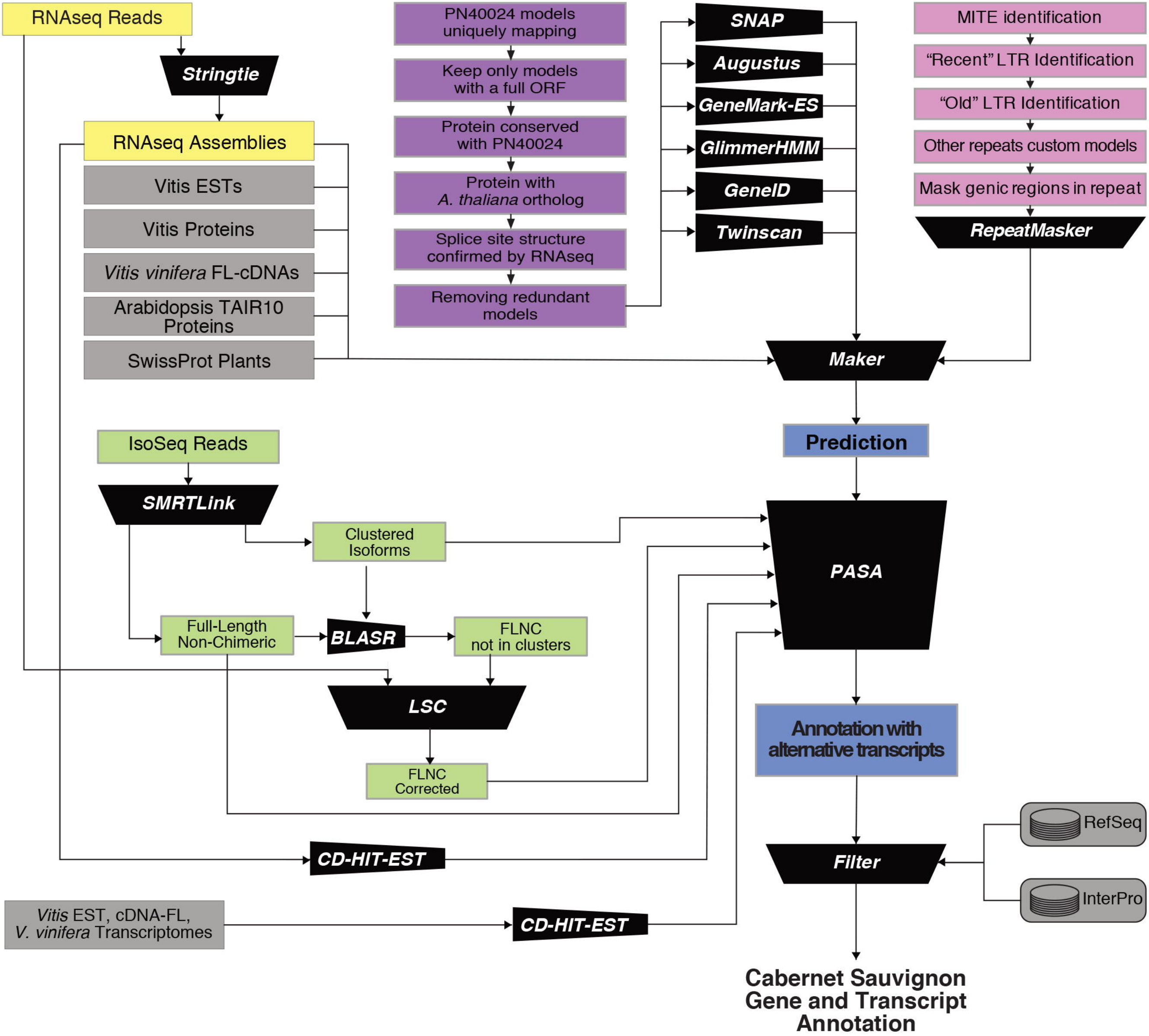
Table 1. Summary statistics of the Cabernet Sauvignon genome annotation after refinement with experimental evidence.

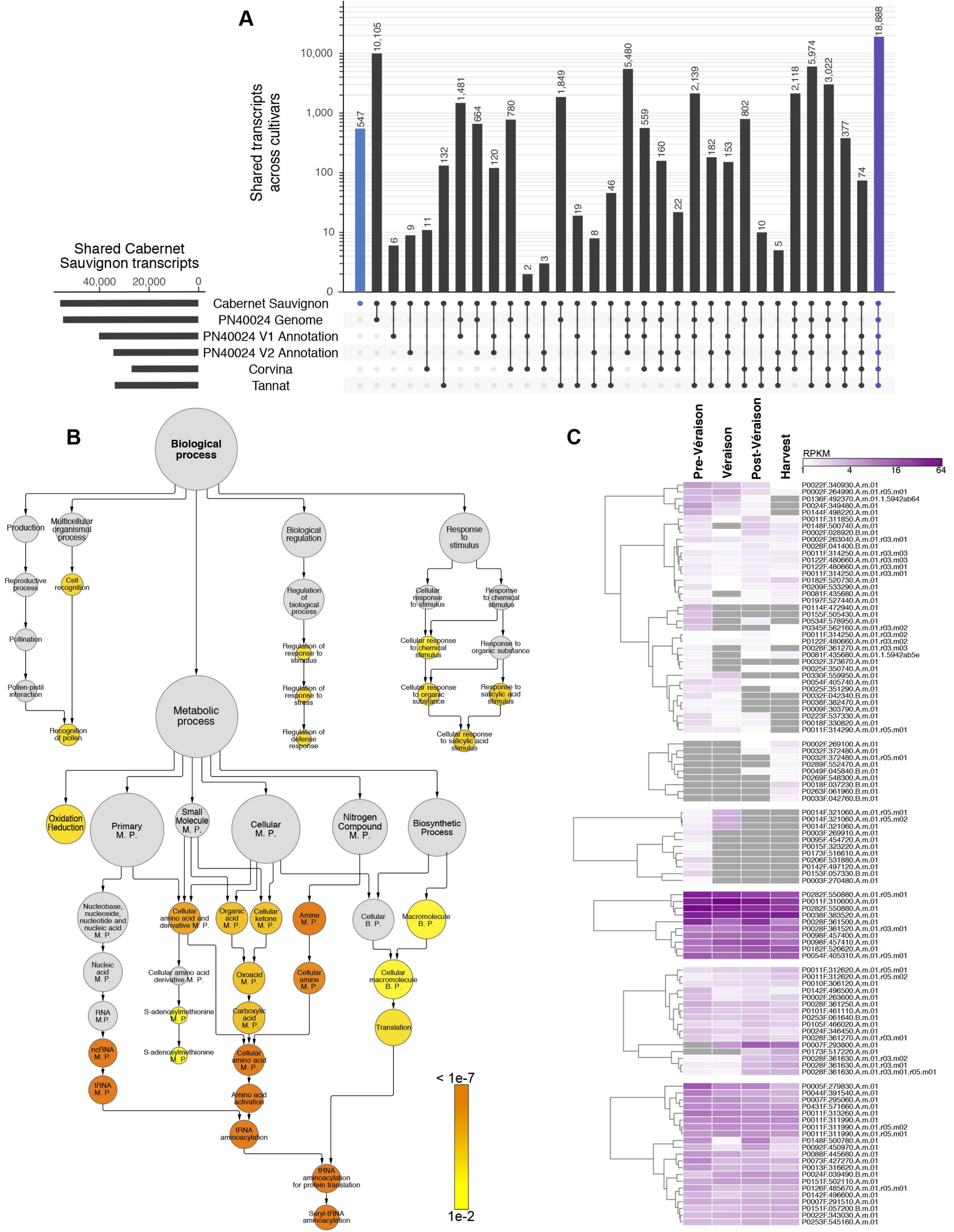
Number of genes	36,687	
Number of monoexonic genes	9,045	
Number of multiexonic genes	27,642	
	Total	**Average per Gene**
Number of Transcripts	55,886	1.52
Number of monoexonic transcripts	9,476	1.05
Number of multiexonic transcripts	46,410	1.68
	Total	**Average per transcript**
Number of exons	326,425	5.84
CDS exon number	296,839	5.31
5'UTR exon number	54,659	0.98
3'UTR exon number	53,433	0.96
	Average Length (bp)	**Max (bp)**
CDS lengths	1,228	29,022
Exon lengths	313	17,750
Intron lengths	809	106,147
5'UTR length	225	13,363
3'UTR length	372	12,798
Intergenic distances	10,349	742,164

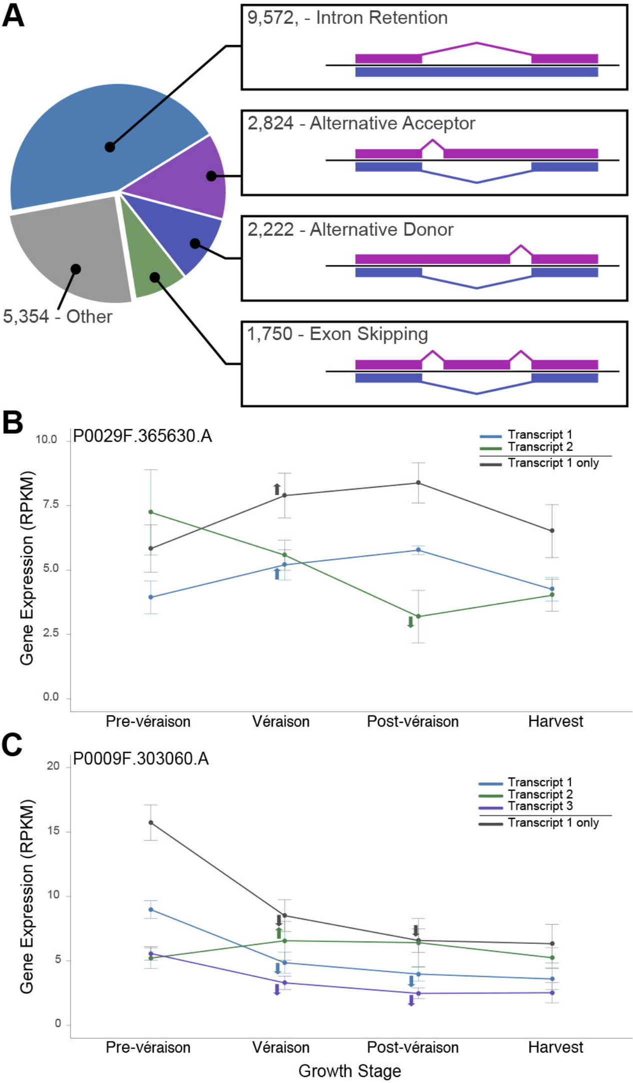

A
Total soluble solids (°Brix)
5
10
15
20
Pre-véraison
Véraison
Post-véraison
Harvest
B
FLNC Reads Count
0
25
50
75
0
20
40
60
80
0
4000
8000
Length (bp)
Library Length
1−2kb
2−3kb
3−6kb
5−10kb


Raw Iso-Seq Reads (N = 9,513,625)
Full-length
Non-chimeric
SMRT Link
FLNC reads
N=672,635
Frequency
0.6
0.4
0.2
0.0
75
80
85
90
95
100
Accuracy (%)
Frequency
0.004
0.003
0.002
0.001
0.000
0
2000
4000
6000
8000
CDS Length (bp)
1−2kb
2−3kb
3−6kb
5−10kb
Clustering
SMRT Link
Unclustered FLNC
Clustered FLNC
Error correction
w/ Illumina RNAseq
LSC
Polishing
SMRT Link
C-FLNC
N reads = 124,185
N loci = 10,639
Frequency
0.6
0.4
0.2
0.0
75
80
85
90
95
100
Accuracy (%)
Frequency
0.0015
0.0010
0.0005
0.0000
0
2000
4000
6000
8000
CDS Length (bp)
1−2kb
2−3kb
3−6kb
5−10kb
PCIRs
N reads = 46,675
N loci = 8,125
Frequency
0.6
0.4
0.2
0.0
75
80
85
90
95
100
Accuracy (%)
Frequency
1e−03
5e−04
0e+00
0
2000
4000
6000
8000
CDS Length (bp)
1−2kb
2−3kb
3−6kb
5−10kb
Corrected Iso-Seq Isoforms (CISIs)
N reads = 170,860
N loci = 13,872
Frequency
0.6
0.4
0.2
0.0
75
80
85
90
95
100
Accuracy (%)
Frequency
1e−03
5e−04
0e+00
0
2000
4000
6000
8000
CDS Length (bp)
1−2kb
2−3kb
3−6kb
5−10kb


A
Shared Expressed Loci
5,000
2,000
0
5,022
1,084
677
647
803
556
467
469
518
467
607
717
680
827
970
Expressed Loci Per Developmental Stage
Pre-Véraison
Véraison
Post-Véraison
Harvest
10,000
7,500
5,000
2,500
0
B
Frequency
0.8
0.6
0.4
0.2
0.0
PCIRs
C-FLNC reads
No Iso-Seq support
1e+01
1e+03
1e+05
RNAseq read counts
C
Number of Transcripts
25
20
15
10
5
0
−1.0
−0.5
0.0
0.5
1.0
Max RPKM
10000
100
1
Number of Transcripts
200
100
−1.0
−0.5
0.0
0.5
1.0
Pearson correlation value


RNAseq Reads
Stringtie
RNAseq Assemblies
Vitis ESTs
Vitis Proteins
Vitis vinifera FL-cDNAs
Arabidopsis TAIR10 Proteins
SwissProt Plants
PN40024 models uniquely mapping
Keep only models with a full ORF
Protein conserved with PN40024
Protein with A. thaliana ortholog
Splice site structure confirmed by RNAseq
Removing redundant models
SNAP
Augustus
GeneMark-ES
GlimmerHMM
GeneID
Twinscan
MITE identification
"Recent" LTR Identification
"Old" LTR Identification
Other repeats custom models
Mask genic regions in repeat
RepeatMasker
Maker
Prediction
IsoSeq Reads
SMRTLink
Clustered Isoforms
Full-Length Non-Chimeric
BLASR
FLNC not in clusters
LSC
FLNC Corrected
PASA
Annotation with alternative transcripts
CD-HIT-EST
Vitis EST, cDNA-FL, V. vinifera Transcriptomes
CD-HIT-EST
Filter
RefSeq
InterPro
Cabernet Sauvignon Gene and Transcript Annotation


A
Shared transcripts across cultivars
10,000
1,000
100
10
0
547
10,105
6
9
11
132
1,481
664
120
780
2
3
1,849
19
8
46
5,480
559
160
22
2,139
182
153
802
10
5
2,118
5,974
3,022
377
74
18,888
Shared Cabernet Sauvignon transcripts
40,000
20,000
0
Cabernet Sauvignon
PN40024 Genome
PN40024 V1 Annotation
PN40024 V2 Annotation
Corvina
Tannat
B
Biological process
Production
Multicellular organismal process
Biological regulation
Response to stimulus
Reproductive process
Cell recognition
Regulation of biological process
Cellular response to stimulus
Response to chemical stimulus
Pollination
Regulation of response to stimulus
Cellular response to chemical stimulus
Response to organic substance
Pollen-pistil interaction
Regulation of response to stress
Cellular response to organic substance
Response to salicylic acid stimulus
Recognition of pollen
Metabolic process
Regulation of defense response
Cellular response to salicylic acid stimulus
Oxidation Reduction
Primary M. P.
Small Molecule M. P.
Cellular M. P.
Nitrogen Compound M. P.
Biosynthetic Process
Nucleobase, nucleoside, nucleotide and nucleic acid M. P.
Cellular amino acid and derivative M. P.
Organic acid M. P.
Cellular ketone M. P.
Amine M. P.
Cellular B. P.
Macromolecule B. P.
Nucleic acid M. P.
Cellular amino acid derivative M. P.
Oxoacid M. P.
Cellular amine M. P.
Cellular macromolecule B. P.
RNA M.P.
S-adenosylmethionine M. P.
Carboxylic acid M. P.
Translation
ncRNA M. P.
S-adenosylmethionine M. P.
Cellular amino acid M. P.
tRNA M. P.
Amino acid activation
tRNA aminoacylation
tRNA aminoacylation for protein translation
Seryl-tRNA aminoacylation
< 1e-7
1e-2
C
Pre-Véraison
Véraison
Post-Véraison
Harvest
RPKM
1
4
16
64
P0022F.340930.A.m.01
P0002F.264990.A.m.01.r05.m01
P0136F.492370.A.m.01.1.5942ab64
P0024F.349480.A.m.01
P0144F.498220.A.m.01
P0011F.311850.A.m.01
P0148F.500740.A.m.01
P0002F.028920.B.m.01
P0002F.263040.A.m.01.r03.m01
P0028F.041400.B.m.01
P0011F.314250.A.m.01.r03.m03
P0122F.480660.A.m.01.r03.m03
P0122F.480660.A.m.01.r03.m01
P0011F.314250.A.m.01.r03.m01
P0182F.520730.A.m.01
P0209F.533290.A.m.01
P0081F.435680.A.m.01
P0197F.527440.A.m.01
P0114F.472940.A.m.01
P0155F.505430.A.m.01
P0534F.578950.A.m.01
P0345F.562160.A.m.01.r03.m02
P0011F.314250.A.m.01.r03.m02
P0122F.480660.A.m.01.r03.m02
P0028F.361270.A.m.01.r03.m03
P0081F.435680.A.m.01.1.5942ab5e
P0032F.373670.A.m.01
P0025F.350740.A.m.01
P0330F.559950.A.m.01
P0054F.405740.A.m.01
P0025F.351290.A.m.01
P0032F.042340.B.m.01
P0038F.382470.A.m.01
P0009F.303790.A.m.01
P0223F.537330.A.m.01
P0018F.330820.A.m.01
P0011F.314290.A.m.01.r05.m01
P0002F.269100.A.m.01
P0032F.372480.A.m.01
P0032F.372480.A.m.01.r05.m01
P0289F.552470.A.m.01
P0049F.045840.B.m.01
P0269F.548300.A.m.01
P0018F.037230.B.m.01
P0263F.061960.B.m.01
P0033F.042760.B.m.01
P0014F.321060.A.m.01.r05.m01
P0014F.321060.A.m.01.r05.m02
P0014F.321060.A.m.01
P0003F.269910.A.m.01
P0095F.454720.A.m.01
P0015F.323220.A.m.01
P0173F.516610.A.m.01
P0206F.531880.A.m.01
P0142F.497120.A.m.01
P0153F.057330.B.m.01
P0003F.270480.A.m.01
P0282F.550880.A.m.01.r05.m01
P0011F.310600.A.m.01
P0282F.550880.A.m.01
P0038F.383520.A.m.01
P0028F.361500.A.m.01
P0028F.361520.A.m.01.r03.m01
P0098F.457400.A.m.01
P0098F.457410.A.m.01
P0182F.520620.A.m.01
P0054F.405310.A.m.01.r05.m01
P0011F.312620.A.m.01.r05.m01
P0011F.312620.A.m.01.r05.m02
P0010F.306120.A.m.01
P0142F.496500.A.m.01
P0002F.263600.A.m.01
P0028F.361250.A.m.01
P0101F.461110.A.m.01
P0253F.061640.B.m.01
P0105F.466020.A.m.01
P0024F.346450.A.m.01
P0028F.361270.A.m.01.r03.m01
P0007F.293800.A.m.01
P0173F.517220.A.m.01
P0028F.361630.A.m.01.r03.m02
P0028F.361630.A.m.01.r03.m01
P0028F.361630.A.m.01.r03.m01.r05.m01
P0005F.279830.A.m.01
P0044F.391540.A.m.01
P0007F.295060.A.m.01
P0431F.571660.A.m.01
P0011F.313260.A.m.01
P0011F.311990.A.m.01
P0011F.311990.A.m.01.r05.m02
P0011F.311990.A.m.01.r05.m01
P0148F.500780.A.m.01
P0092F.450970.A.m.01
P0088F.445680.A.m.01
P0073F.427270.A.m.01
P0013F.316620.A.m.01
P0024F.039490.B.m.01
P0151F.502110.A.m.01
P0126F.485670.A.m.01.r05.m01
P0142F.496600.A.m.01
P0007F.291510.A.m.01
P0151F.057200.B.m.01
P0022F.343030.A.m.01
P0253F.545160.A.m.01


A
9,572, - Intron Retention
2,824 - Alternative Acceptor
2,222 - Alternative Donor
1,750 - Exon Skipping
5,354 - Other
B
P0029F.365630.A
Transcript 1
Transcript 2
Transcript 1 only
Gene Expression (RPKM)
10.0
7.5
5.0
2.5
0.0
Pre-véraison
Véraison
Post-véraison
Harvest
C
P0009F.303060.A
Transcript 1
Transcript 2
Transcript 3
Transcript 1 only
Gene Expression (RPKM)
20
15
10
5
0
Pre-véraison
Véraison
Post-véraison
Harvest
Growth Stage