Current Plant Biology 9–10 (2017) 29–36

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Comparative landscape of alternative splicing in fruit plants☆

Gaurav Sablok[a,*], Brian Powell[b], Jonathan Braessler[b], Feng Yu[b], Xiang Jia Min[a,*]

[a] Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA
[b] Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555, USA

ARTICLE INFO

Article history:
Received 12 May 2017
Received in revised form 17 June 2017
Accepted 26 June 2017

Keywords:
Alternative splicing
Fruit plant
Apple
Grape
Orange
Strawberry
Expressed sequence tags

ABSTRACT

Alternative splicing (AS) has played a major role in defining the protein diversity, which could be linked to phenotypic alternations. It is imperative to have a comparative resolution of AS to understand the pre-mRNAs splicing diversity. In the present research, we present a comparative assessment of the AS events in four different fruit plants including apple (*Malus domestica*), grape (*Vitis vinifera*), sweet orange (*Citrus sinensis*), and woodland strawberry (*Fragaria vesca*), using spliced mapping of the expressed sequence tags and mRNA sequences. We identified a total of 2039 AS events in apple, 2454 in grape, 1425 in orange, and 631 in strawberry, respectively. In this study grape displayed the maximum number of genes (1588) associated with the splicing, followed by apple (1580), orange (1133) and strawberry (444). Transcripts mapping analysis shows that grape plant has relatively larger intron sizes than introns in other fruit species. The data provide a basis for further functional characterization of the genes undergoing AS and can be accessed at Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/plant/).

1. Introduction

Plant development depends upon a complex interaction of proteins and as such protein diversity contributes to these interactions. Protein diversity has not only played an intricate role in regulating the transcriptional and post-transcriptional responses but also has played a major role in regulating the stress responses [1,2]. The factor that contributes to increasing protein diversity, termed as alternative splicing (AS) is a key mechanism, which leads to spliceosomal alternations resulting in production of more than one splice transcripts [3–6]. It has been widely elucidated that these splicing events not only affect the developmental patterns, but also play an important role in regulating the stress responses in fruit species [2,7], the fate and divergence of the duplicated genes [8], and also contribute to the mechanistic understanding of the miRNA regulation [9]. Patterns of alternatively spliced transcripts have been widely observed with reports suggesting that 90% of human genes containing multiple exons are alternatively spliced [10], thus, demonstrating exon skipping as a major splice event in humans. In plants, due to the presence of long introns, often intron-retention has been seen as a major splice event, with as many as 60% of multi-exon genes undergoing AS in the model plant *Arabidopsis thaliana* [1,11].

mRNA transcript isoforms are generally generated through four basic events in AS: [1] intron retention (IR) in the mature mRNA; [2] exon skipping (ES) resulting from alternative exon usage (AEU); [3] alternative donor site (AltD) and [4] alternative acceptor site (AltA) that are resulted from the use of cryptic splice sites that may elongate or shorten an exon [4,11–13]. Approximately 60–75% of AS events occur within the protein coding regions of mRNAs, resulting changes in binding properties, intracellular localization, protein stability, enzymatic, and signaling activities [14]. In plants, IR has been shown to be the most dominant form with reports suggesting the proportions of intron containing genes undergoing AS in plants ranged from ~30% to >60% depending the depth of available transcriptome data [1,15]. In addition to the above mentioned basic AS events, various complex types can be formed by combination of basic events [1,13,15]. AS isoforms might encode distinct functional proteins or might be nonfunctional, which harbor a premature termination codon. These non-functional isoforms generated through the process called "regulated unproductive splicing and translation" (RUST) are degraded by a process known as nonsense-mediated decay (NMD) [4,12,16]. Recent study on serine/arginine (SR) genes, which are spatiotemporally regulated and also show a varied amount of splicing diversity, has revealed widespread coupling of AS with NMD in SR gene family, suggesting a strong link between unproductive splicing and the abundance of functional transcripts [17]. Nonetheless, association of AS and RNA-binding proteins has been established, specifically RNA-binding protein AtGRP8 up-regulation has been shown to promote the use of cryptic 5′ splice site thus producing a splice transcript, which

☆ This article is part of a special issue entitled "Plant Development", published in the journal Current Plant Biology 9–10, 2017.

* Corresponding authors.
E-mail addresses: sablokg@gmail.com (G. Sablok), xmin@ysu.edu (X.J. Min).

http://dx.doi.org/10.1016/j.cpb.2017.06.001

in mutants revealed a target direct of NMD in *A. thalina* [18]. Identification of AS events have been widely done in several plant species such as *A. thaliana* [12,19–21], *Oryza sativa* (rice) [12], *Zea mays* (maize) [22,23], *Sorghum bicolor* (sorghum) [23,24], *Nelumbo nucifera* (sacred lotus) [25], *Vitis vinifera* (grape) [7,9], *Brachypodium distachyon* [13,15], and *Ananas comosus* (pineapple) [26]. The SR proteins, which belong to the class of RNA-binding proteins, have been shown to play key role in regulating the splicing machinery [27].

With the advent of the next generation sequencing approaches, several classification approaches have been used for the identification of AS types, which includes differential splicing (Diffseq), differential exon-usage (DEXseq) and application of Bayesian (rMATS), splicing graph based detections and count based approaches (Spladder) [28–31]. Although the application of these approaches have revealed the AS landscape variations in different plant species with the splicing information based on the method, application of these approaches are limited in elucidating the AS landscape in polyploid species, mainly due to the heterozygosity and the large genome size of these polyploid species with ancient genome duplication events. Previously, EST/mRNA based approaches have been widely used to understand the AS events in polyploid species and have provided robust estimates of the splice detection in the polyploid species. Taking into account these above mentioned considerations, in this work, we carried out a survey of AS landscape in four fruit plants, which include apple (*Malus domestica*), grape (*Vitis vinifera*), sweet orange (*Citrus sinensis*), and woodland strawberry (*Fragaria vesca*), using mRNA/ESTs spliced alignment. Accurate spliced alignment of the transcripts and identification of these AS events allows for further functional characterization to reveal the role of these identified spliced transcripts played in the regulation process, which can pave the way for understanding the physiological events in fruit plants.

2. Materials and methods

2.1. Genome sequences and transcripts

For the prediction of the comparative AS events, respective genome sequences including gene models for four fruit plant species were downloaded from different data sources respectively. Briefly, the genome sequence for sweet orange data were downloaded from Citrus genome database (http://citrus.hzau.edu.cn/orange/download/data.php) [32], woodland strawberry and apple plant data were downloaded from GDR database (the Genome Database for Rosaceae) (http://www.rosaceae.org/species/fragaria/fragaria_vesca/genome_v1.0; and http://www.rosaceae.org/species/malus/malus_x_domestica/genome_v1.0) [33,34], grape genome was downloaded from Phytozome database (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Vvinifera/annotation/) [36]. The assembled ESTs and mRNA transcripts of orange and grape plants were downloaded from the PlantGDB database (http://www.plantgdb.org/prj/ESTCluster/) [36]. The strawberry and apple plant mRNA and ESTs were downloaded from the ESTs and nucleotide database at National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) and assembled in-house using the CAP3 program with the following parameters: −p 95–o 50–g 3–y 50–t 1000 [37].

2.2. Putative unique transcripts (PUTs) to genome mapping, identification and functional annotation of AS isoforms

The PUTs were mapped to their corresponding genomes using ASFinder (http://proteomics.ysu.edu/tools/ASFinder.html/) [38]. ASFinder uses SIM4 program [39] to align PUTs to the genome and then parse the SIM4 output file to generate a file with gene transfer format (gtf) and also extract those PUTs that are mapped to the same genomic location but have variable exon-intron boundaries. The output file (AS.gtf) of ASFinder was then subsequently submitted to AStalavista server (http://genome.crg.es/astalavista/) for AS event analysis [40]. To avoid the call of the spurious alternative splicing events, we applied a threshold of minimum of 95% identity of aligned PUT with a genomic sequence, a minimum of 80 bp aligned length, and >75% of a PUT sequence aligned to the genome [13]. Application of the above identity percentage and the aligned length minimizes the chance of the false positive AS events calling as a result of gene and genome duplication events. The percentage of alternative spliced genes was estimated using the genome predicted gene models with the spliced genes having at least one PUT spliced alignment.

The assembled PUTs were further annotated for their coding regions using the ORFPredictor [41] and the full–length transcript coverage was assessed using TargetIdentifier [42]. Functional classification was assigned to the PUTs by performing BLASTX searches against UniProtKB/Swiss-Prot with a cutoff E-value of 1E-5. The predicted protein sequences from ORFPredictor were further functionally annotated using rpsBLAST against the PFAM database (http://pfam.xfam.org/). Following the mapping, the exonic and the intronic boundaries were extracted from the AS mapping files and the sequence logo for the intronic and the exonic boundaries were made using the Web Logo 3 available from http://weblogo.threeplusone.com [43]. For the phylogenetic representation, *A. thaliana* SR proteins was used as a query across the sequenced genomes from each clade available from Phytozome [35] and protein alignment was done using MSAProbs [44], followed by RAxML ancestral phylogenetic analysis using RAxML version 8 [45] with PROTCATWAG model. Phylogenetic tree was rooted using *Ambroella trichopoda* as a basal angiosperm.

2.3. Data access and visualization of AS

AS events identified in this study along with the integrated genomic tracks are available from Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/) [13,23]. The user interface allows choosing a species and then searching the database using a PUT ID, gene ID, keywords in functional annotation; PFAM; or AS event types. Additionally; the identified AS events can be visualized and compared with predicted gene models using GBrowse for comparative assessment. BLASTN search for the PUTs and AS isoforms is also supported. The assembled sequence data with annotation information and other related intermediate data files are publicly available for downloading at: http://proteomics.ysu.edu/publication/data/FruitAS/

3. Results and discussion

3.1. Analysis and annotation of PUTs

Genome-wide analyses of alternative splicing have established its nearly ubiquitous role in gene regulation in many organisms [46]. Identification of spliced transcripts plays an important role in understanding the ecotypic responses and also has played fundamental role in understanding the regulome of plant species [1–4]. It is noteworthy to highlight that the previous estimates using ESTs/mRNAs mapping based approaches provided relatively accurate results of the AS events, where high resolution based mRNA-seq is lacking or in sequenced species where the genome fragmentation is largely present [47]. In the present research, we used putative unique transcripts (PUTs) for genome mapping to unravel the splicing diversity in four fruit species and presented a

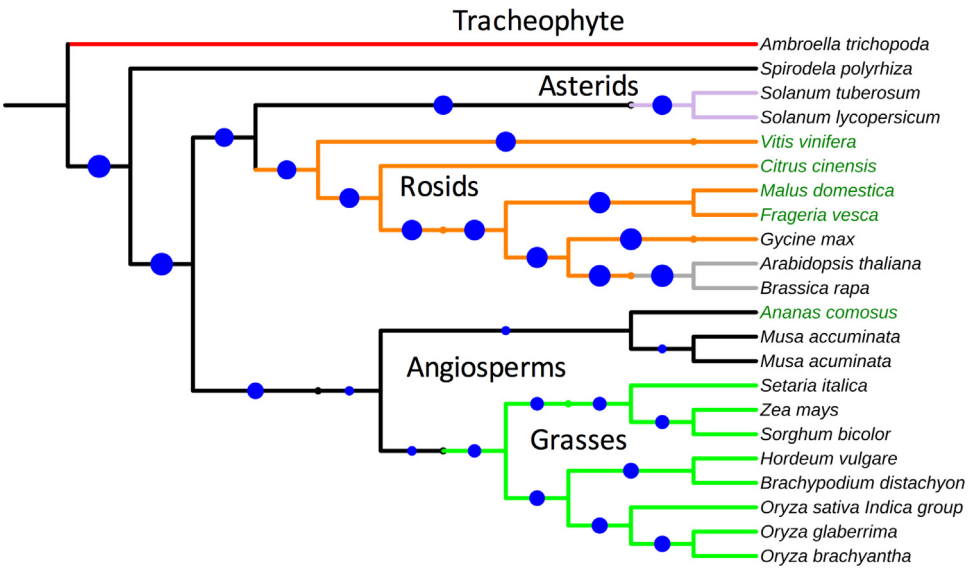

Fig. 1. Phylogenetic classification of the studied species using the serine arginine (SR) rich protein involved in alternative splicing.

Table 1
Summary of the mapping to genome of assembled putative unique transcripts (PUTs) and the percentage of alternative splicing (AS) in fruit plants.

Species	Apple	Grape	Orange	Strawberry
Total PUTs	87,734	64,796	105,294	20,884
Error rate (%)	0.04	0.34	0.37	0.25
Average length (bp)	585	749	858	697
PUTs mapped to genome	42,118	35,470	27,210	12,240
% of PUTs mapped to genome	48.0	54.7	25.8	58.6
Predicted gene models	63,541	26,346	44,275	34,809
PUTs matched to gene models	33,362	20,761	20,347	11,659
Unique gene models with PUTs	21,568	14,326	11,073	11,274
AS genes	1580	1588	1113	444
AS (%)	7.3	11.1	10.1	3.9

comparative assessment of the spliced transcripts with functions conserved across the fruit species. Fig. 1 represents the phylogenetic placement of the studied species using SR (serine arginine) proteins with respect to basal angiosperm *Ambroella trichopoda*. Table 1 presents the summary statistics of the number of the PUTs (20884 in strawberry, 64796 in grape, 87734 in apple, and 105294 in orange) used for the identification of the AS landscape in these four species. As compared to gene models, the numbers of PUTs used were high except in strawberry, revealing more than 50% of the gene models supported with PUTs. Prior to mapping, we estimated the error rate in PUTs to avoid the false mapping of the PUTs and to avoid predicted splice sites supported with stretches of Ns. As compared to the other fruit species, apple datasets revealed low error diversity (0.04%), whereas the error rate was relatively high in other three datasets, from 0.25% in strawberry dataset to 0.37% in grape dataset. The average lengths of PUT datasets varied from 585 bp in apple dataset, 697 bp in strawberry dataset, 749 bp in grape dataset, to 858 bp in orange dataset respectively. Each PUT was functionally annotated including putative ORF prediction, coding region full-length prediction, a putative function and PFAM prediction. The PUTs which were mapped to their corresponding genomes were also visualized and compared with predicted gene models using GBrowse.

Fruit species studied in the present work showed a wide diversity in the genome assembly and several features such as lack of the chromosome based genome assembly. Taking into account the polyploid nature of these fruit species, we first checked whether the predicted gene models were supported by mRNA transcripts by mapping the PUTs to the predicted gene models using BLASTN with an identity of ≥95% to compare PUTs with predicted coding DNA sequences (cds). Our data showed the percentages of gene models supported by at least one expressed transcript were 33.9% in apple, 54.4% in grape, 53.90% in orange, and 32.4% in strawberry plants (Table 1). Spliced mapping of the PUTs to the respective genome revealed the number of genes undergoing AS were 1580 (11.1%) in apple dataset, 1588 (10.1%) in grape dataset, 1113 (10.1%) in orange dataset, and 444 (3.9%) in strawberry dataset (Table 1). The difference of the AS rate in different species may be due to the difference in the number of transcripts used for mapping in respective species and also due to the polyploid nature of these species, lacking the chromosome level scaffolded assembly. Our previous reports on AS in *B. distachyon* clearly illustrated the fact that more ESTs/mRNAs increased the number of identified AS genes [13,15]. This trend was also observed during previous reports of AS analysis in *A. thaliana* [21]. Although as compared to the previous estimates suggesting that 61% of multi-exonic genes in *A. thaliana* are alternatively spliced under normal growth conditions [20], and ~40% of intron containing genes that undergo AS in maize [22], the identified AS events represent a small portion of the AS diversity. However, it is noteworthy to highlight that the fruit species studied in this research are highly heterozygous and identifying the AS patterns in highly heterozygous through short read mapping is still a fundamental challenge. Comparative assessment of the AS events across the grapevine cultivars although revealed a total of 44% multi-exonic genes going under AS, however 70% of the identified AS events were supported with low-expression levels [48].

3.2. Classification of alternative splicing events

Diversity and types of AS events play an important role in deciding the functional aspects of the alternatively spliced isoforms. As compared to humans, where dominant AS events are exon-skipping, plants reveal a large fraction of the identified AS events as intron retention. Classification of the AS events observed in four fruit plants with recently published pineapple data are listed in Table 2 [26], demonstrating the prevalence of the IR as the major splicing type with a frequency varying from 44.7% in grape to 61.9% in pineapple. The high frequency of the IR in fruit plants was consistent with results obtained in other plant species including *A.*

Table 2
Alternative splicing events in fruit plants.

	Apple		Grape		Orange		Strawberry		Pineapple[a]	
	Count	%	Count	%	Count	%	Count	%	Count	%
Exon skipping	75	3.7	117	4.8	47	3.3	18	2.9	474	4.6
Alternative donor sites	166	8.1	189	7.7	88	6.2	36	5.7	684	6.6
Alternative acceptor sites	326	16.0	306	12.5	201	14.1	80	12.7	1145	11.1
Intron retention	1184	58.1	1096	44.7	834	58.5	348	55.2	6404	61.9
others (complex events)	288	14.1	746	30.4	255	17.9	149	23.6	1641	15.9
Total	2039	100	2454	100.0	1425	100.0	631	100.0	10348	100.0

[a] Pineapple data were obtained from Wai et al. (2015).

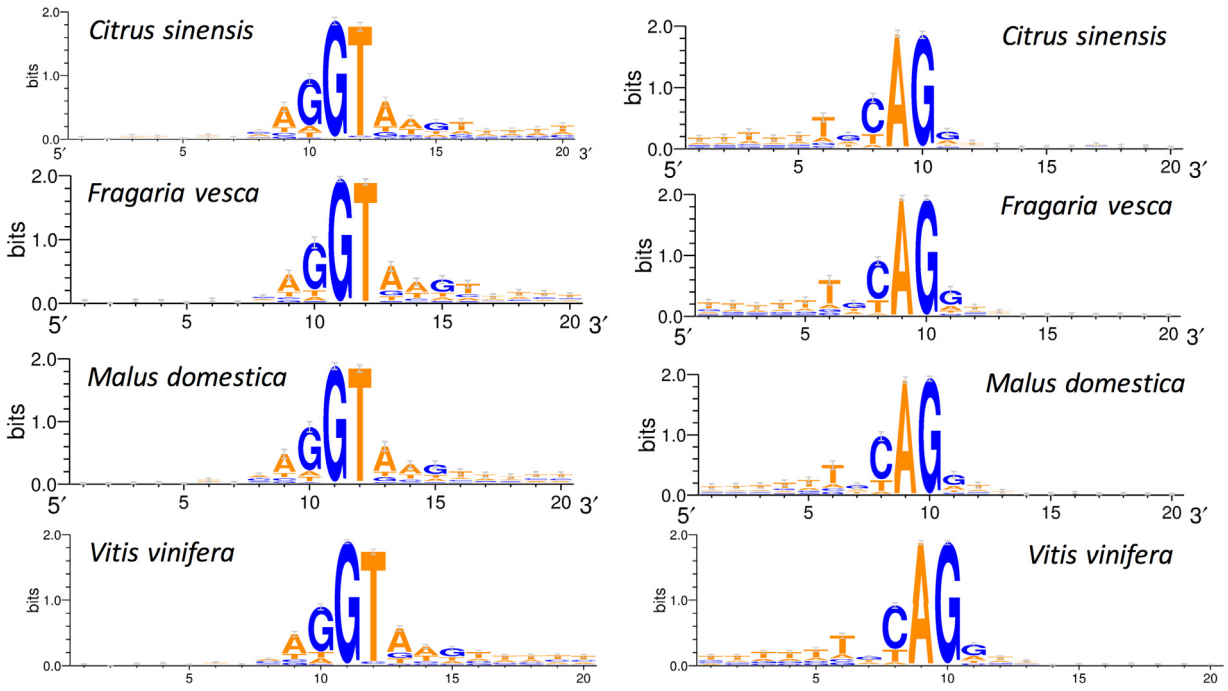

Fig. 2. Sequence logo of the exonic and the intronic boundaries of the splice sites in four fruit plant species.

thaliana, *O. sativa*, and some other grass and cereal species [12,3,23]. The observed frequency of the intron retention events being the most dominant in fruit species is in lines with the RNA-seq based splicing estimates suggesting 77% of the identified events as IR in grapes [9]. Previously, wide spread role of the temperature as a regulatory event in modulating the splice landscape has been widely reviewed [48]. Interestingly, dominance of these splice events have been linked with physiological significance in grape revealing the modulation of the IR at temperature fluctuations [49]. This feature of regulated IR under elevated temperature has been previously shown in *Physcomitrella patens* (moss) [50].

AltA and AltD represent the second and third most abundant type of observed AS events with AltA showing a relatively higher frequency as compared to AltD (Table 2). ES type represents the rarest AS event in plants. The current study was also consistent with numerous previous studies in other plant species [13,23]. The complex events are the AS isoform pairs having more than one AS event and often having more than one basic AS type. The percentage of complex events often varies with the lengths of assembled PUTs, with 61.7% complex AS events were found in sorghum data [23].

To study the post-splicing exonic and intronic sequence features, we extracted the splice site information from the exon based mapping files revealing a total of 37158 sequence features in *Citrus sinensis*, 25266 sequence features in *Fragaria vesca*, 59297 in *Malus domestica* and a total of 54911 sequence features in *Vitis vinifera*, respectively. Weblogo 3 was used to plot the sequence features revealing the conservation of GT-AG splice site rules (Fig. 2), which is consistent and supports the RNA-seq based splice sites identification [9].

3.3. Functional annotation of transcript isoforms generated by AS

Functional impact of the alternative spliced transcripts has been widely elucidated with the recent reports showing the RNA binding proteins SR45, SR30, and SR34, and the nuclear ribonucleic protein U1A revealing the IR based splicing fluctuations regulated by temperature stress [7]. To identify the functional importance of the predicted splice events, we functionally annotated the PUTs including alternatively spliced transcripts with putative protein domains. The ORFs of PUTs were identified using ORFPredictor webserver [41] and the protein families were predicted using rpsBLAST searching PFAM database. Among predicted ORFs of these genes undergoing AS, 1050 in apple, 1137 in grape, 628 in orange, and 256 in strawberry were classified with a putative protein family (Table 3). Among the protein functions encoded by these AS genes, the larger families include proteins with kinase domain, RNA recognition motif, protein tyrosine kinase, protein phosphatase

Table 3
Protein family distribution in the proteins encoded by genes undergoing alternative splicing in four fruit plants.

	Apple	Grape	Orange	Strawberry	Pfam	Description
pfam00069	22	19	17	1	Pkinase	Protein kinase domain
pfam00076	19	17	12	7	RRM_1	RNA recognition motif
pfam07714	17	8	3	1	Pkinase_Tyr	Protein tyrosine kinase
pfam00481	13	7	3	1	PP2C	Protein phosphatase 2C
pfam00179	12	14	6	6	UQ_con	Ubiquitin-conjugating enzyme
pfam00249	11	12	5	2	Myb_DNA-binding	Myb-like DNA-binding domain
pfam00504	10	4	9	7	Chloroa_b-bind	Chlorophyll A-B binding protein
pfam00230	8	7	3	2	MIP	Major intrinsic protein
pfam01370	7	7	2	2	Epimerase	NAD dependent epimerase/dehydratase family
pfam00010	7	4	1	0	HLH	Helix-loop-helix DNA-binding domain
pfam00582	6	8	4	3	Usp	Universal stress protein family
pfam00847	6	6	3	1	AP2	AP2 domain
pfam00226	6	6	1	2	DnaJ	DnaJ domain
pfam03141	6	5	1	0	Methyltransf_29	Putative
pfam03552	6	1	4	1	Cellulose_synt	Cellulose synthase
pfam00083	6	1	2	2	Sugar_tr	Sugar (and other) transporter
pfam07002	6	1	0	0	Copine	Copine
pfam00071	3	12	4	2	Ras	Ras family
pfam00141	4	10	2	3	peroxidase	Peroxidase
pfam09770	3	8	2	3	PAT1	Topoisomerase II-associated protein PAT1
pfam13639	3	7	5	3	zf-RING_2	Ring finger domain
pfam03171	5	7	4	1	2OG-FeII_Oxy	2OG-Fe(II) oxygenase superfamily
pfam01357	4	6	3	1	Pollen_allerg_1	Pollen allergen
pfam02365	4	6	3	1	NAM	No apical meristem (NAM) protein
pfam12796	4	6	1	1	Ank_2	Ankyrin repeats (3 copies)
pfam01490	1	6	1	1	Aa_trans	Transmembrane amino acid transporter protein
pfam00149	2	6	0	0	Metallophos	Calcineurin-like phosphoesterase
pfam00011	0	3	7	3	HSP20	Hsp20/alpha crystallin family
pfam00106	3	3	6	1	adh_short	short chain dehydrogenase
pfam00085	5	3	6	0	Thioredoxin	Thioredoxin
others	841	927	628	256		
Total	1050	1137	748	314		
Unique Pfam	643	690	510	237		

Note: only protein families having at least 5 members in at least one species are listed in the table.

2C, ubiquitin-conjugating enzyme, Myb-like DNA-binding domain, chlorophyll A-B binding domain, etc. (Table 3).

Interestingly, MYB domains proteins have been previously shown to have extensive splice forms regulating the developmental patterns and also in response to pathogens [51]. It is worth to mention that among the top abundant transcripts, we observed the RRM (RNA-recognition motifs), which are also represented in the serine–arginine (SR) rich proteins, which are widely spliced with as much as 90 transcripts in *A. thaliana* and majority of them targeted by NMD [17]. Abundance of the RRM containing motifs has also been linked with the recent class of the NAGNAG motifs, which represents NAGNAG spliced motif and have been shown to be widely regulated under the cold stress [52] and constitutes an important part of the spliceosomal machinery. An interesting hypothesis with abundance of these motifs indicates towards the conservation of the RRM domains containing proteins as spliced transcripts from monocots, dicots to the ploypoid species such as fruits [52]. Another interestingly and the fourth most abundant class of the spliced transcripts encode the PP2C domains, which are represented by as many as 26 genes in *A. thaliana* [53] and were shown to be representing the IR events in wide a variety of stresses and played as key negative regulators of the ABA signaling pathway [54]. It is worth to mention that the ABA signaling pathway plays an important role in fruit ripening and hence the identification of these spliced transcripts presents a resource for the functional characterization of the splicing diversity and their concurrent effect on the ripening process in fruits [55]. It is also worth to mention that the previous report in grape only highlighted the disease resistance motifs as a major class [9] and, thus, the motifs identified in this study present a new source for the identification and functional characterization of splicing in fruits. Among the other categories, we observed widely spliced transcript examples such as Calcineurin-like phosphoesterase [23] and high temperature induced splicing of the HSP20 family [7].

3.4. Conserved alternatively spliced genes in fruit plants

Conservation of the AS events plays a major role in understanding the evolution of the alternatively spliced transcripts and how the splice site motif such as the frequency of the GT-AG or GC-AG pairs evolved over the canonical and non-canonical splicing. Conservation pattern reveals the categories of the AS genes, which have been previously shown in pineapple and cereal plants [23,26], suggesting that these genes undergoing AS to be evolutionarily conserved in different linages of plant species and across the ecotypes in fruits plants [48].

We used the best reciprocal BLASTp search with one representative protein sequence from each gene undergoing AS to identify conserved AS genes. Species pairwise conserved AS genes can be downloaded from the download site mentioned above. A total of 14 AS genes were identified which were conserved among the four fruit species (Fig. 3, Table 4). Interestingly, among them AS genes encoding Ferritin (PFAM00210), UQ_con (ubiquitin-conjugating enzyme, PFAM00179), and TPT (triose-phosphate transporter family, PFAM03151) were also conserved in pineapple, maize, rice, sorghum, and *B. distachyon* [26].

3.5. Features of DNA fragments involved in alternative splicing

We examined the size distributions of DNA fragments involved in basic AS events (Table 5). The retained introns had variable sizes from 2 bp to 1.5 kb with an average size ranged from 140 to 197 bp in different fruit plant species. The range was comparable with previous analysis in *B. distachyon* which had a range of 8–1142 bp and

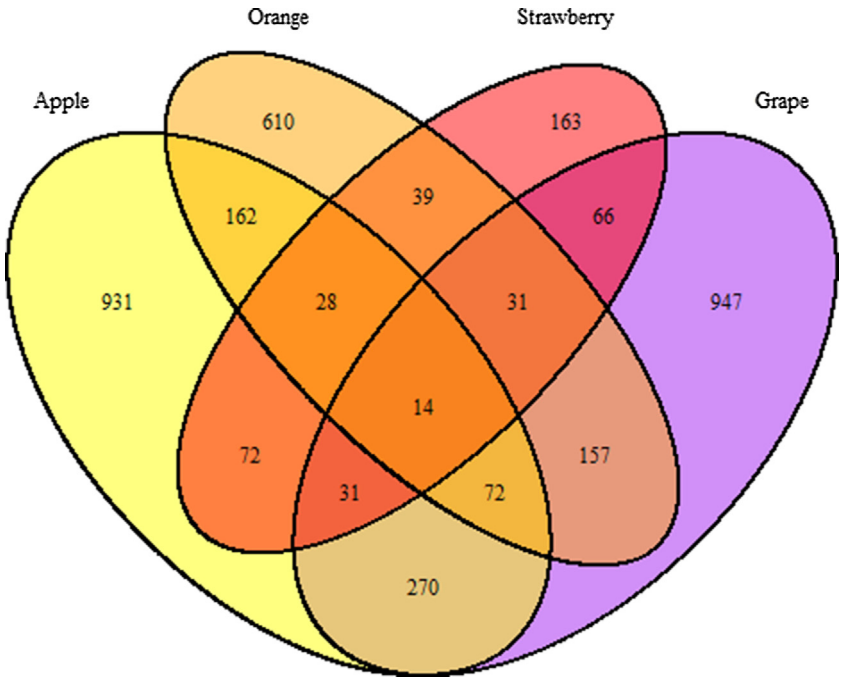

Fig. 3. Conserved alternatively spliced genes among the four fruit plant species including apple, grape, orange, and strawberry.

Table 4
Conserved alternative splicing genes among four fruit plants.

Apple	Grape	Orange	Strawberry	Pfam
Contig9656	Vv3449	Cs60106184	89540794	pfam00076, RRM_1, RNA recognition motif
52024425	Vv63497	Cs194106183	158370739	pfam00101, RuBisCO_small, Ribulose bisphosphate carboxylase, small
Contig20450	Vv41067	Cs46148	Contig1161	pfam00179, UQ_con, Ubiquitin-conjugating enzyme
Contig25482	Vv2430	Cs12987	Contig3834	pfam00210, Ferritin, Ferritin-like domain
Contig21497	Vv8066257	Cs7970	Contig2876	pfam00244, 14-3-3, 14-3-3 protein
Contig10241	Vv43587	Cs48074	Contig4306	pfam00484, Pro_CA, Carbonic anhydrase
Contig7583	Vv57970	Cs27951	Contig1399	pfam01031, Dynamin_M, Dynamin central region
48411817	Vv9179	Cs81333	Contig4113	pfam01070, FMN_dh, FMN-dependent dehydrogenase
Contig26661	Vv51506	Cs65106187	158371956	pfam01554, MatE, MatE domain.
Contig15008	Vv11566260	Cs59461	Contig908	pfam02605, PsaL, Photosystem I reaction centre subunit XI
Contig14799	Vv10975	Cs69106190	Contig2870	pfam03151, TPT, Triose-phosphate Transporter family
Contig27007	Vv19510	Cs80871	158359626	pfam09349, OHCU_decarbox, OHCU decarboxylase
Contig17769	Vv47574	Cs58462	Contig4398	pfam13419, HAD_2, Haloacid dehalogenase-like hydrolase
Contig22262	Vv14651	Cs3951	Contig119	pfam14204, Ribosomal_L18_c, Ribosomal L18 C-terminal region

Table 5
Summary of DNA fragment sizes (bp) involved in alternative splicing events in fruit plants.

		Apple	Grape	Orange	Strawberry
IR	size range	2–1183	2–1533	2–1005	2–1148
	mean (SD)	140 (115)	154 (155)	197 (173)	197 (180)
AltA	size range	4–1192	3–911	3–613	4–238
	mean (SD)	54 (106)	54 (89)	53 (88)	48 (53)
AltD	size range	3–373	3–465	4–604	3–377
	mean (SD)	58 (68)	72 (93)	67 (99)	67 (72)
ES	size range	32–305	32–1236	24–349	42–222
	mean (SD)	105 (69)	126(141)	93 (74)	99 (53)

IR: intron retention; AltA: alternative acceptor site; AltD: alternative donor site; ES: exon skipping; SD: standard deviation.

an average size of 184 bp [13]. Comparing with the average intron sizes in fruit plants (Table 6), the retained introns tend to be smaller. The skipped exons (ES) had a range of 24 bp to 1236 bp with an average size ranged from 93 pb to 126 bp in different fruit plants, which was also similar to the average size (111 bp) obtained in *B. distachyon* [13]. The fragments involved in alternative donor (AltD) or acceptor (AltA) sites ranged from 3 bp to 1192 bp (average size 48–54 bp) in AltA and from 3 bp to 604 bp (average size 58 – 72 bp) in AltD in fruit plants, which were also similar to the data, AltA 49 bp and AltD 67 bp, obtained in *B. distachyon* [13]. Overall the average sizes of the fragments involved in AS were relatively short than the average sizes of exons and introns in these plants (Table 5).

3.6. Features of exons and introns based on PUTs mapping

We extracted and plotted the length distribution of all internal exons and introns from each plant and the results were summarized (Table 6; Figs. 4 and 5). Interestingly, we observed that the average internal exon lengths in four fruit plants were from 129 to 133 bp, i. e. almost similar each other and also similar to the exon lengths obtained in *B. distachyon* (130 bp), *indica* rice (133 bp), and maize (142 bp) [13,23]. Species pairwise T-test showed that the internal exon length of grape was significantly ($p < 0.001$) longer than the exon lengths of apple and citrus, though the difference of the average length seemed small (Fig. 4). However, they were relatively much shorter than the internal exon lengths in *japonica* rice (180 bp) and sorghum (179 bp) [23].

On the other hand, intron lengths varied tremendously (Table 5; Fig. 5). The average intron lengths were 336 bp in strawberry, 365 bp in apple, 402 bp in orange, and 812 bp in grape. The rela-

Table 6
Internal exon and intron size in fruit plants.

	Exon			Intron		
	Sample size	Average length (bp)	SD(bp)	Sample size	Average length (bp)	SD (bp)
Apple	37,582	130	98	59,539	365	569
Grape	36,483	133	101	54,281	812	1357
Orange	23,650	129	91	37,986	402	550
Strawberry	17,209	131	95	25,372	336	437

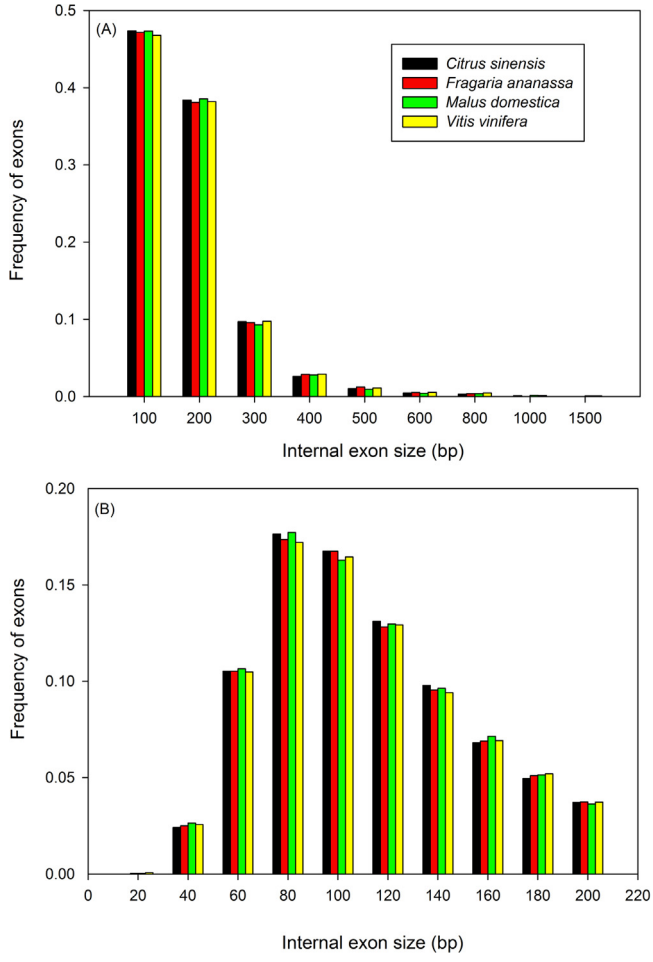

Fig. 4. Distribution of internal exon size: The x-axis indicates the size of internal exons. Bin sizes are right inclusive (e.g., bin 100 comprises sequences of lengths 1–100 bp). The y-axis indicates the frequency of internal exons. (A) Exon length distribution; (B) A detailed distribution of small internal exons.

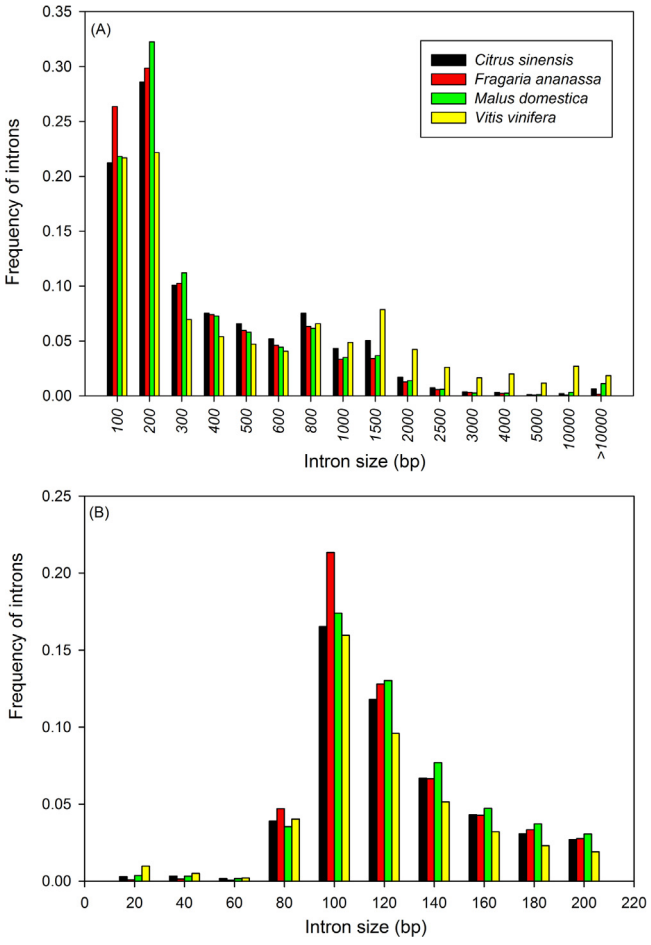

Fig. 5. Distribution of intron size: The x-axis indicates the size of introns. Bin sizes are right inclusive (e.g., bin 100 comprises sequences of lengths 1–100 bp). The y-axis indicates the frequency of introns. (A) Intron length distribution; (B) A detailed distribution of small introns.

tive high frequency of longer introns in grape can be clearly seen in Fig. 5. In our previous analysis maize had a relative longer intron length (554 bp) than introns in other grass and cereal plants. There were some longer introns found in all species, the proportions of introns having a length of >10 kb were 0.1% in strawberry, 0.6% in orange, 1.1% in apple, and 1.8% in grape. However, in considering the possible errors in PUTs and genome assembly the introns with a length of >10 kb were not used for the calculation of the average intron length. Excluding the introns having a length of >10 kb, the remaining data were used for species pairwise T-test. The results showed all the species pairs had significant differences ($P < 0.001$) in intron lengths. It is worth to highlight that larger size of introns and intron expansion in fruits plants especially from the comparative view point of *A. thaliana* and *V. vinifera* has highlighted the role of the transposable elements (TE) expansion in the longer introns [56]. Since alternative splicing has established correlation with the genome duplications [8], the observed large significant variation in introns in grape might be due to the ancient hexa-polyploidization of the genome. However, it is yet to ascertain that whether the TE explanation is affecting the splicing patterns in these species. The biological significance, if there is any, of the variations of the intron sizes in different linages of plant species remains to be further examined.

Author's contribution

XJM and GS conceived the study. BP and JB contributed to the database construction, XJM, GS, and FY contributed to the experiment design, data analysis, and preparation of the manuscript. All authors have read and approved the final version of the manuscript.

Funding

The work was supported by a Research Professorship award to XM by Youngstown State University.

References

[1] A.S. Reddy, Y. Marquez, M. Kalyna, et al., Complexity of the alternative splicing landscape in plants, Plant Cell 25 (2013) 3657–3683.
[2] D. Staiger, J.W. Brown, Alternative splicing at the intersection of biological timing, development, and stress responses, Plant Cell 25 (2013) 3640–3656.
[3] W. Gilbert, Why genes in pieces? Nature 271 (1978) 501.
[4] S.A. Filichkin, H.D. Priest, S.A. Givan, et al., Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*, Genome Res. 20 (2010) 45–58.
[5] B.R. Graveley, Alternative splicing: increasing diversity in the proteomic world, Trends Genet. 17 (2001) 100–107.
[6] G.C. Roberts, C.W. Smith, Alternative splicing: combinatorial output from the genome, Curr. Opin. Chem. Biol. 6 (2002) 375–383.
[7] J. Jiang, X. Liu, G. Liu, et al., Integrating omics and alternative splicing reveals insights into grape response to high temperature, Plant Physiol. 173 (2017) 1502–1508.
[8] L.P. Iñiguez, G. Hernández, The evolutionary relationship between alternative splicing and gene duplication, Front. Genet. 8 (2017) 14.
[9] N. Vitulo, C. Forcato, E.C. Carpinelli, et al., A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype, BMC Plant Boil. 14 (2014) 99.
[10] L. Chen, J.M. Tovar-Corona, A.O. Urrutia, Alternative splicing: a potential source of functional innovation in the eukaryotic genome, Int. J. Evol. Biol. 2012 (2012) 596274.
[11] R.F. Carvalho, C.V. Feijão, P. Duque, On the physiological significance of alternative splicing events in higher plants, Protoplasma 250 (2013) 639–650.
[12] B. Wang, V. Brendel, Genome wide comparative analysis of alternative splicing in plants, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 7175–7180.
[13] B. Walters, G. Lum, G. Sablok, et al., Genome-wide landscape of alternative splicing events in *Brachypodium distachyon*, DNA Res. 20 (2013) 163–171.
[14] S. Stamm, S. Ben-Ari, I. Rafalska, et al., Function of alternative splicing, Gene 344 (2005) 1–20.
[15] G. Sablok, P.K. Gupta, J.M. Baek, et al., Genome-wide survey of alternative splicing in the grass Brachypodium distachyon: an emerging model biosystem for plant functional genomics, Biotechnol. Lett. 33 (2011) 629–636.
[16] B.P. Lewis, R.E. Green, S.E. Brenner, Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 189–192.
[17] S.G. Palusa, A.S. Reddy, Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay, New Phytol. 185 (2010) 83–89.
[18] J.C. Schöning, C. Streitner, I.M. Meyer, et al., Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis, Nucleic Acids Res. 36 (2008) 6977–6987.
[19] P.G. Zhang, S.Z. Huang, A.L. Pin, et al., Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of Arabidopsis, Mol. Biol. Evol. 27 (2010) 1686–1697.
[20] Y. Marquez, J.W. Brown, C. Simpson, et al., Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis, Genome Res. 22 (2012) 1184–1195.
[21] N.H. Syed, M. Kalyna, Y. Marquez, et al., Alternative splicing in plants – coming of age, Trends Plant Sci. 17 (2012) 616–623.
[22] S.R. Thatcher, W. Zhou, A. Leonard, et al., Genome-wide analysis of alternative splicing in Zea mays: landscape and genetic regulation, Plant Cell 26 (2014) 3472–3487.
[23] X.J. Min, B. Powell, J. Braessler, et al., Genome-wide cataloging and analysis of alternatively spliced genes in cereal crops, BMC Genomics 16 (2015) 721.
[24] B. Panahi, B. Abbaszadeh, M. Taghizadeghan, et al., Genome-wide survey of alternative splicing in Sorghum bicolor, Physiol. Mol. Biol. Plants 20 (2014) 323–329.
[25] R. VanBuren, B. Walters, R. Ming, et al., Analysis of expressed sequence tags and alternative splicing genes in sacred lotus (Nelumbo nucifera Gaertn.), Plant Omics 6 (2013) 311–317.
[26] C.M. Wai, B. Powell, R. Ming, et al., Analysis of alternative splicing landscape in pineapple (Ananas comosus), Trop. Plant Biol. 9 (2016) 150–160.
[27] P. Duque, A role for SR proteins in plant stress responses, Plant Signal. Behav. 6 (2011) 49–54.
[28] S. Anders, A. Reyes, W. Huber, Detecting differential usage of exons from RNA-seq data, Genome Res. 22 (2012) 4025.
[29] S. Shen, J.W. Park, Z.X. Lu, et al., rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) (E111: E5593-5601).
[30] A. Kahles, C.S. Ong, Y. Zhong, et al., SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data, Bioinformatics 32 (2016) 1840–1847.
[31] Y. Hu, Y. Huang, Y. Du, et al., DiffSplice: the genome-wide detection of differential splicing events with RNA-seq, Nucleic Acids Res. 41 (2013) e39.
[32] Q. Xu, L.L. Chen, X. Ruan, et al., The draft genome of sweet orange (*Citrus sinensis*), Nat. Genet. 45 (2013) 59–66.
[33] V. Shulaev, D.J. Sargent, R.N. Crowhurst, et al., The genome of woodland strawberry (Fragaria vesca), Nat. Genet. 43 (2011) 109–116.
[34] R. Velasco, A. Zharkikh, J. Affourtit, et al., The genome of the domesticated apple (*Malus x domestica* Borkh.), Nat. Genet. 42 (2010) 833–839.
[35] D.M. Goodstein, S. Shu, R. Howson, et al., Phytozome: a comparative platform for green plant genomics, Nucleic Acids Res. 40 (D1) (2012) D1178–D1186.
[36] J. Duvick, A. Fu, U. Muppirala, M. Sabharwal, et al., PlantGDB: a resource for comparative plant genomics, Nucleic Acids Res. 36 (Suppl. 1) (2008) D959–D965.
[37] X. Huang, A. Madan, CAP3: a DNA sequence assembly program, Genome Res. 9 (1999) 868–877.
[38] X.J. Min, ASFinder: a tool for genome-wide identification of alternatively spliced transcripts from EST-derived sequences, Int. J. Bioinformatics Res. Appl. 9 (2013) 221–226.
[39] L. Florea, G. Hartzell, Z. Zhang, et al., A computer program for aligning a cDNA sequence with a genomic DNA sequence, Genome Res. 8 (1998) 967–974.
[40] S. Foissac, M. Sammeth, ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets, Nucleic Acids Res. 35 (W35) (2007) W297–W299.
[41] X.J. Min, G. Butler, R. Storms, et al., OrfPredictor: predicting protein-coding regions in EST-derived sequences, Nucleic Acids Res. 33 (W33) (2005) W677–680.
[42] X.J. Min, G. Butler, R. Storms, et al., TargetIdentifier: a web server for identifying full-length cDNAs from EST sequences, Nucleic Acids Res. 33 (W33) (2005) W669–W672.
[43] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo:A sequence logo generator, Genome Res. 14 (2004) 1188–1190.
[44] J. González-Domínguez, Y. Liu, J. Touriño, et al., MSAProbs-MPI: parallel multiple sequence aligner for distributed-memory systems, Bioinformatics 32 (2016) 3826–3828.
[45] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.
[46] Y. Xing, C. Lee, Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes, Nat. Rev. Gene 7 (2006) 499–509.
[47] O. Jaillon, J.M. Aury, B. Noel, et al., The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, Nature 449 (2007) 463–467.
[48] G. Capovilla, A. Pajoro, R.G. Immink, et al., Role of alternative pre-mRNA splicing in temperature signaling, Curr. Opin. Plant Biol. 27 (2015) 97–103.
[49] E. Potenza, M.L. Racchi, L. Sterck, et al., Exploration of alternative splicing events in ten different grapevine cultivars, BMC Genomics 16 (2015) 706.
[50] C.Y. Chang, W.D. Lin, S.L. Tu, Genome-wide analysis of heat-sensitive alternative splicing in Physcomitrella patens, Plant Physiol. 165 (2014) 826–840.
[51] J. Li, X. Li, L. Guo, et al., A subgroup of MYB transcription factor genes undergoes highly conserved alternative splicing in Arabidopsis and rice, J. Exp. Bot. 57 (2006) 1263–1273.
[52] S. Schindler, K. Szafranski, M. Hiller, et al., Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes, BMC Genomics 9 (2008) 159.
[53] A. Schweighofer, H. Hirt, I. Meskiene, Plant PP2C phosphatases: emerging functions in stress signaling, Trends Plant Sci. 9 (2004) 236–243.
[54] S. AlShareef, Y. Ling, H. Butt, et al., Herboxidiene triggers splicing repression and abiotic stress responses in plants, BMC Gnomics 18 (2017) 260.
[55] P. Leng, B. Yuan, Y. Guo, The role of abscisic acid in fruit ripening and responses to abiotic stress, J. Exp. Bot. 65 (2014) 4577–4588.
[56] K. Jiang, L.R. Goertzen, Spliceosomal intron size expansion in domesticated grapevine (Vitis vinifera), BMC Res. Notes 4 (2011) 52.