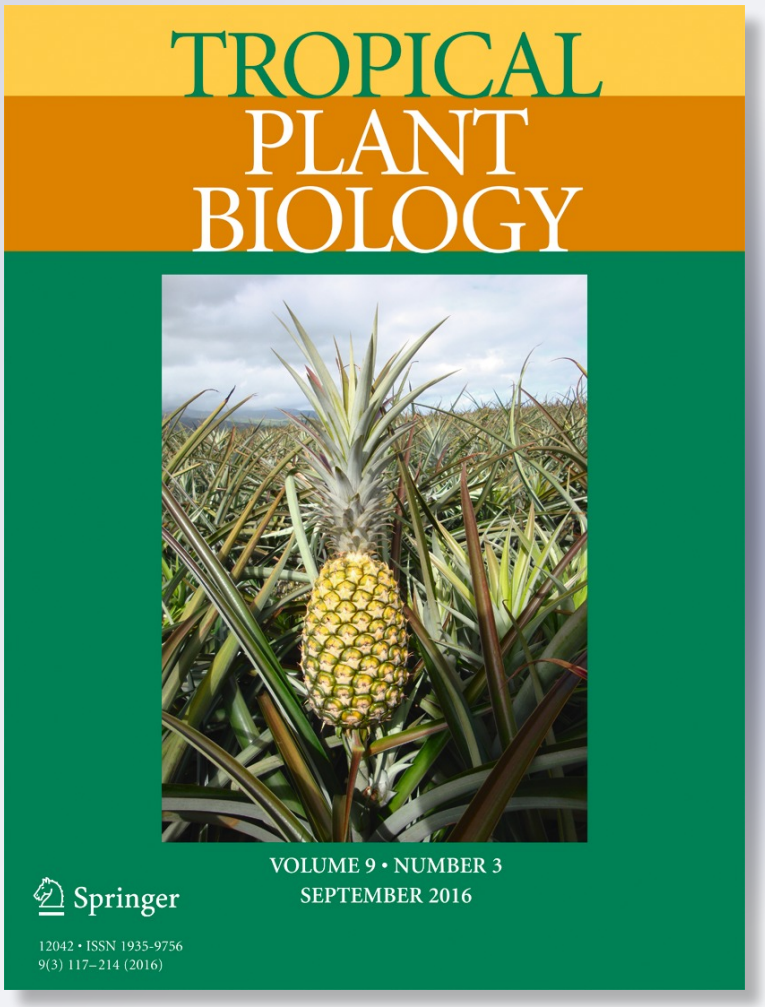
Analysis of Alternative Splicing Landscape in Pineapple (Ananas comosus)

Ching Man Wai, Brian Powell, Ray Ming & Xiang Jia Min

Tropical Plant Biology
An International Journal devoted to original research in tropical plants

ISSN 1935-9756
Volume 9
Number 3

Tropical Plant Biol. (2016) 9:150-160
DOI 10.1007/s12042-016-9168-1



Springer

Tropical Plant Biol. (2016) 9:150–160
DOI 10.1007/s12042-016-9168-1

Analysis of Alternative Splicing Landscape in Pineapple (*Ananas comosus*)

Ching Man Wai[1] · **Brian Powell**[2] · **Ray Ming**[1] · **Xiang Jia Min**[3]

Received: 9 November 2015 / Accepted: 24 April 2016 / Published online: 4 May 2016

Abstract Pineapple (*Ananas comosus* L. Merrill) is an important tropical and subtropical fruit crop and possesses crassulacean acid metabolism (CAM) photosynthesis. Recent release of its genome sequences makes it possible to identify genes transcribed with alternatively spliced isoforms in this plant. Mapping the assembled transcripts generated by next-generation sequencing technology and existing expressed sequence tags as well as mRNA sequences to the published pineapple genome, we identified and analyzed alternative splicing (AS) events. We identified a total of 10,348 AS events involving 13,449 assembled putative unique transcripts, which were mapped to 5146 pineapple gene models that equivalent to 29.7 % of total expressed gene models. Consistent with previous findings in other plant species, intron retention (61.9 %) remains to be the dominant type among the identified AS events. Comparative genomic analysis of genes which generated pre-mRNAs having AS revealed a total of 481 genes conserved among *Oryza sativa* (ssp *japonica*), *Sorghum bicolor*, *Zea mays*, and pineapple, with 51 of them were also conserved with *Brachypodium distachyon*. Gene Ontology classification revealed that the products of these genes which generate AS isoforms are involved in many biological processes with diverse molecular functions. We annotated all assembled transcripts and also associated them with predicted gene models. The annotated information of these data provides a resource for further characterizing these genes and their biological roles. The data can be accessed at Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/).

Keywords Alternative splicing · Expressed sequence tags · mRNA · Pineapple

Communicated by: Paulo Arruda

Electronic supplementary material The online version of this article (doi:10.1007/s12042-016-9168-1) contains supplementary material, which is available to authorized users.

✉ Xiang Jia Min
xmin@ysu.edu

[1] Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2] Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555, USA

[3] Center for Applied Chemical Biology, Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA

Abbreviations

AS	Alternative splicing
ESTs	Expressed sequence tags
GO	Gene ontology
PUT	Putative unique transcript
FPKM	Fragments Per Kilobase of exon model per Million mapped reads
rpsBLAST	Reversed position specific BLAST.

Introduction

Pineapple, *Ananas comosus* L. Merrill, is an important tropical and subtropical fruit plant. It is ranked as the third largest tropical fruit after banana and mango in the world market (Bartholomew et al. 2002). Its fruits can be consumed directly or processed as canned fruit and juice, and the juice can be used for making wines and beers. Its products also include bromelain, a cysteine protease in the family of cysteine proteinases, derived from the stems of pineapples and used as folk medicine and as a meat tenderizer (Taussig and Batkin 1988).

The waste from processed fruits can be used to feed animals. The pineapple fruit peel and the residues including leaves and stems left after harvesting fruits have potential to produce biofuel (Barz and Delivand 2011; Surles et al. 2009).

Pineapple is a perennial monocot belonging to the family Broemeliaceae. It has long been known and studied as a plant possessing crassulacean acid metabolism (CAM) photosynthesis, i. e., carbon dioxide is fixed as malate and compartmentalized during the night and the carbon is remobilized for Calvin cycle using C3-pathway during the day (Bartholomew and Kadzimin 1977; Bartholomew and Malézieux 1994; Nievola et al. 2005). Another character to make pineapple plant an interesting subject for investigation to both farmers and researchers is that its flowering, thus fruiting, can be induced by ethylene (Burg and Burg 1966; Bartholomew 2013 for review). A term called "forcing" has been used in the pineapple industry to describe the practice to use plant growth regulators or other compounds to induce pineapple to flower that results the synchronization of fruit harvesting (Bartholomew 2013 for review). On the contrary to forcing, preventing random precocious flowering before planned forcing is also required in some cultivars in pineapple planting areas. Research efforts have been made to identify growth regulators to inhibit precocious flowering in pineapple (Min and Bartholomew 1996), and eventually delaying natural flowering in pineapple was achieved through foliar application of aviglycine, an inhibitor of ethylene biosynthesis (Wang et al. 2007). Silencing of the ACC synthase gene AcACS2 causes delayed flowering in pineapple, providing further supporting evidence that ethylene plays a central role in initiating flowering in pineapple (Trusov and Botella 2006; Bartholomew 2013).

In the last decade, molecular and genomics approaches have been applied to study genes involved in CAM pathway, fruit ripening, nematode infection, and other biological processes in pineapple (Zhang et al. 2014 for review; Moyle et al. 2005; Ong et al. 2012). Expressed sequence tags (ESTs) and unique gene transcripts generated using next generation sequencing (NGS) technology in pineapple are available (Moyle et al. 2005; Ong et al. 2012). More excitingly the draft, near complete, pineapple genome has been sequenced and annotated recently (Ming et al. 2015). The available genomics resources provide an unprecedented opportunity for studying and understanding this fruit crop at the molecular level using systems biology approaches.

Alternative splicing is a process of generating more than one mRNA transcript from a pre-mRNA, thus this process increases the diversity of mRNA transcripts and the proteins from a single gene in eukaryotes. In humans, it was reported that more than 90 % of intron containing genes have alterative splicing (Pan et al. 2008). Also it was estimated that ~60 % of intron containing genes in Arabidopsis, a model plant, have alternative splicing (Marquez et al. 2012). The biological significances and the complexity of AS in plants have been recently comprehensively reviewed (Staiger and Brown 2013; Reddy et al. 2013). Recently, we have analyzed AS in *Nelumbo nucifera* (sacred lotus), *Brachypodium distachyon*, and several cereal crops including *Oryza sativa* (ssp *japonica* and *indica*) (rice), *Sorghum bicolor* (sorghum), and *Zay mays* (maize) (VanBuren et al. 2013; Walters et al. 2013; Sablok et al. 2011; Sablok et al. 2013; Min et al. 2015). In this study, we identified and analyzed 5146 genes which generated alternatively spliced isoforms in pineapple using the approach of mapping the transcripts data generated with NGS technology and available ESTs and mRNAs in the public database to the newly sequenced pineapple genome.

Results and Discussion

Annotation and Analysis of Putative Unique Transcripts (PUTs)

Assembling the RNA-seq data generated in this project with available EST and mRNA sequences in the public database at NCBI, we generated a total of 63,991 PUT sequences with lengths ranging from 104 bp to 13,919 bp with an average length of 1098 bp (Fig. 1). Using OrfPredictor, a webserver for predicting open reading frames (ORFs) of ESTs (Min et al. 2005a), a total of 63,732 PUTs (99.6 %) were predicted to contain an ORF region for peptide translations, and among them, 33,191 (51.9 %) have a homolog in the UniProtKB/Swiss-Prot database and 22,046 (34.5 %) were assigned with a Pfam domain with a cutoff E-value of 1e-10 searching against the Pfam database. A total of 48,514 PUTs (75.8 %) were mapped to the draft pineapple genomic sequences (version 3) using the parameters defined in the method section. Using BLASTN with a cut off identity of 95 % and a minimum aligned length of 80 bp, 40,834 PUTs (63.8 % of total PUTs, 84.2 % of mapped PUTs) were matched to coding

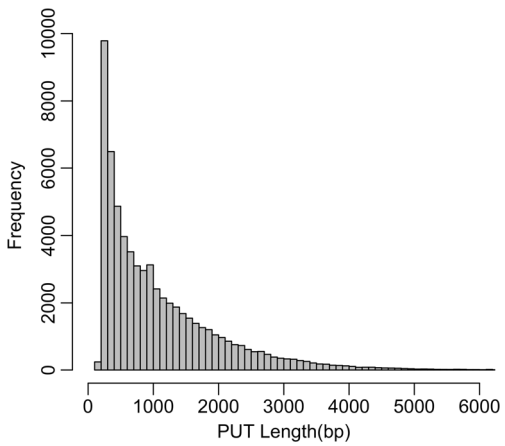


Fig. 1 Length distribution of assembled putative unique transcripts (PUTs) in pineapple

DNA sequences (CDS) of 17,308 predicted genes models. These data can be downloaded from the website mentioned above.

Classification of AS Events

We identified a total of 10,348 AS events (Table 1, Supplementary Table 1), involving 13,449 PUTs that were mapped to 5630 genomics loci. Among the 13,449 PUTs generated by pre-mRNA alternative splicing, 12,118 of them matched with a predicted gene model and a total 5146 unique predicted gene models were found to generate pre-mRNAs undergoing AS. The remaining 1331 PUTs may lie in the untranslated regions of the predicted genes or represent unidentified or transposons genes. We found that a total of 17,308 predicted genes had one or more mapped PUTs, i.e. these genes were expressed and supported with PUT evidence, thus, the percentage of genes generated AS isoforms was estimated to be 29.7 %. However, if we use 27,024, the number of total predicted genes, the percentage of gene with pre-mRNA undergoing AS identified in this study was 19.0 %. Among different AS events, intron retention is the dominant AS type, accounted for 61.9 %, followed by alternative acceptor sites (11.1 %) and alternative donor sites (6.6 %), and exon skipping represents the least AS type (4.6 %). These results are consistent with several previous studies in other plant species (Wang and Brendel 2006; VanBuren et al. 2013; Walters et al. 2013; Min et al. 2015). For the simplicity of description, these genes generated pre-mRNAs undergoing AS were referred as AS genes, and non-AS genes referred those genes not having AS isoforms identified in the study.

Functional Classification of AS Genes and the Impact of AS on Gene Function

The PUTs generated in the present study were annotated for putative protein coding regions by performing a BLASTX search against UniProt/Swiss-Prot database and the ORFs of PUTs were identified using OrfPredictor webserver (Min et al. 2005a). The protein families of the ORF of each PUT were predicted using rpsBLAST searching Pfam database. We further classified AS gene functional products into 1677 protein families using the longest ORF from each AS gene (Table 2, Supplementary Table 2). Among the genes with classified protein family category, 31.6 % of them underwent alternative splicing. However, it is noted that the proportions of AS in different protein families vary greatly (Table 2).

We further performed Gene Ontology (GO) analysis for both AS genes and non-AS genes using only genes having at least one PUT mapped. There are three main categories of GO terms including biological processes, cellular components, and molecular functions (http://www.geneontology.org/). As some proteins may have multiple GO annotations in a category and some of them do not have GO annotation, thus, the GO classification only provides an estimated survey of the sampled genes. AS gene products are involved in nearly all biological processes and have various types of molecular functions. We extracted a total of 26,655 GO entries from non-AS gene dataset and 15,416 GO entries from the AS gene dataset in the category of biological processes of GO, thus, ~36.6 % of expressed genes underwent AS in GO biological processes (Fig. 2a). Similarly, we extracted a total of 12,314 GO entries from non-AS gene dataset and 7176 GO entries from the AS gene dataset in the category of molecular functions, thus ~36.8 % of them underwent AS (Fig. 2b). Among the molecular functions, the top five categories include binding, transferase activity, catalytic activity, nucleotide binding, and hydrolase activity. These proteins are involved in diverse biological processes including metabolic, biosynthesis, cellular component organization, responses to stress, etc. The proteins encoded by AS genes are distributed into several different cellular components. About 34.7 % of GO entries in the cellular component category of GO underwent AS (Fig. 2c).

Isoforms generated by AS can be either functional or non-functional. Non-functional AS isoforms often have a premature stop codon due to non-three nucleotide insertions or deletions within the ORF region. These isoforms often are degraded through the process of "regulated unproductive splicing and translation" (RUST) or nonsense mediated mRNA decay (NMD) surveillance machinery (Morello and Breviario 2008). It was estimated that ~43 % Arabidopsis AS events and ~36 % rice AS events produce NMD candidates (Wang and Brendel 2006). However, functional AS isoforms may have a similar function or a different function depending on if the translation frame is changed and if a premature codon exists and where it is located. We used TargetIdentifer to identify only AS isoform pairs both were predicted to contain a complete ORF and at least one had a Pfam annotation. As TargetIdentifier uses the output of BLASTX for predicting the completeness of ORF region, thus, for PUTs without a hit in BLASTX, there is no prediction of its ORF completeness (Min et al. 2015b). Among a total of 10,348 AS isoform pairs, 2536 pairs were predicted to have a complete ORF in both isoforms and at least one isoform contains a Pfam

Table 1 Alternative splicing events in pineapple

AS event type	Event number	%
exon skipping	474	4.6
Alternative donor sites	684	6.6
Alternative acceptor sites	1145	11.1
intron retention	6404	61.9
others (complex events)	1641	16.9
Total	10,348	

Table 2 Protein family distribution in the proteins encoded by genes not having pre-mRNA alternative splicing (non-AS genes) and genes with pre-mRNA alternative splicing (AS-genes) in pineapple

PFam ID	non-AS Genes	AS Genes	AS (%)	Pfam	Description
pfam00069	331	143	30.2	Pkinase	Protein kinase domain
pfam07714	90	64	41.6	Pkinase_Tyr	Protein tyrosine kinase
pfam00076	69	63	47.7	RRM_1	RNA recognition motif
pfam13041	241	46	16.0	PPR_2	PPR repeat family
pfam13639	95	38	28.6	zf-RING_2	Ring finger domain
pfam00179	24	27	52.9	UQ_con	Ubiquitin-conjugating enzyme
pfam00270	30	26	46.4	DEAD	DEAD/DEAH box helicase
pfam00226	41	21	33.9	DnaJ	DnaJ domain
pfam00153	37	21	36.2	Mito_carr	Mitochondrial carrier protein
pfam00481	31	21	40.4	PP2C	Protein phosphatase 2C
pfam00004	32	19	37.3	AAA	ATPase family associated with various cellular
pfam00067	94	18	16.1	p450	Cytochrome P450
pfam00141	47	18	27.7	peroxidase	Peroxidase
pfam00071	31	17	35.4	Ras	Ras family
pfam00249	78	16	17.0	Myb_DNA-binding	Myb-like DNA-binding domain
pfam12697	38	16	29.6	Abhydrolase_6	Alpha/beta hydrolase family
pfam00010	38	16	29.6	HLH	Helix-loop-helix DNA-binding domain
pfam01501	27	15	35.7	Glyco_transf_8	Glycosyl transferase family 8
pfam00149	24	15	38.5	Metallophos	Calcineurin-like phosphoesterase
pfam00122	15	15	50.0	E1-E2_ATPase	E1-E2 ATPase
Total	8834	4085	31.6		

This is only a partial list. The complete list is showed as Supplementary Table 2

annotation. Within them, 1905 pairs (75.1 %) had identical Pfam identifier, 517 pairs (20.4 %) had one isoform having a Pfam but the other without a Pfam, and 114 pairs (4.5 %) had different Pfam categories. Thus, about 24.9 % of AS event generated isoforms had their protein functionalities changed (Supplementary Table 3). These Pfam loss or changes are most likely caused by the translation frame changes. We further examined the subcellular locations of the predicted protein sequences from the set of 1905 pair isoforms not having Pfam change. The subcellular locations were predicted using six computational tools as described previously (Lum et al. 2014). Within the 1905 pair isoforms which were generated from 1219 genes, we identified 376 genes with 818 unique PUTs, encoded isoform proteins with different subcellular locations (Supplementary Table 4). Thus it was estimated ~30.8 % of AS genes in this subset generated isoforms encoding proteins targeting to different subcellular locations. The biological significance of the change in protein subcellular locations of the isoforms in these genes certainly warrants further evaluation.

Conserved AS Genes in Monocots

Recently, we identified and compared conserved AS genes using assembled EST and mRNA sequences in cereal crops including rice ssp *japonica* and ssp *indica*, sorghum and maize (Min et al. 2015). As the number of AS genes identified in *indica* rice was much less than the number of AS genes identified in *japonica* rice, presumably due to much less number of available ESTs in *indica* rice, we did not include *indica* rice in this analysis. We compared conserved AS homologous genes between each pair of species as well as among each three species and also the four species. A total of 481 genes, which generated AS transcripts, were conserved among rice, sorghum, maize and pineapple (Fig. 3). The detailed species pair-wise conserved gene lists and other intermediate data files can be downloaded at the website provided in the method section. Furthermore, a total of 51 genes, which generated pre-mRNA undergoing AS, were identified to be conserved among pineapple, sorghum, maize, rice, and *Brachypodium* (Table 3). These AS genes conserved within the five monocots, diverged from lineage leading to banana and the palms about 100–120 million years ago (Ming et al. 2015), may warrant further detailed examination of the biological significance of these genes.

In fact, some of these conserved AS genes have been well studied in Arabidopsis and rice (Table 3). For example, genes encoding proteins having a MYB-DNA binding domain play an important role in plant development and defense mechanism, and

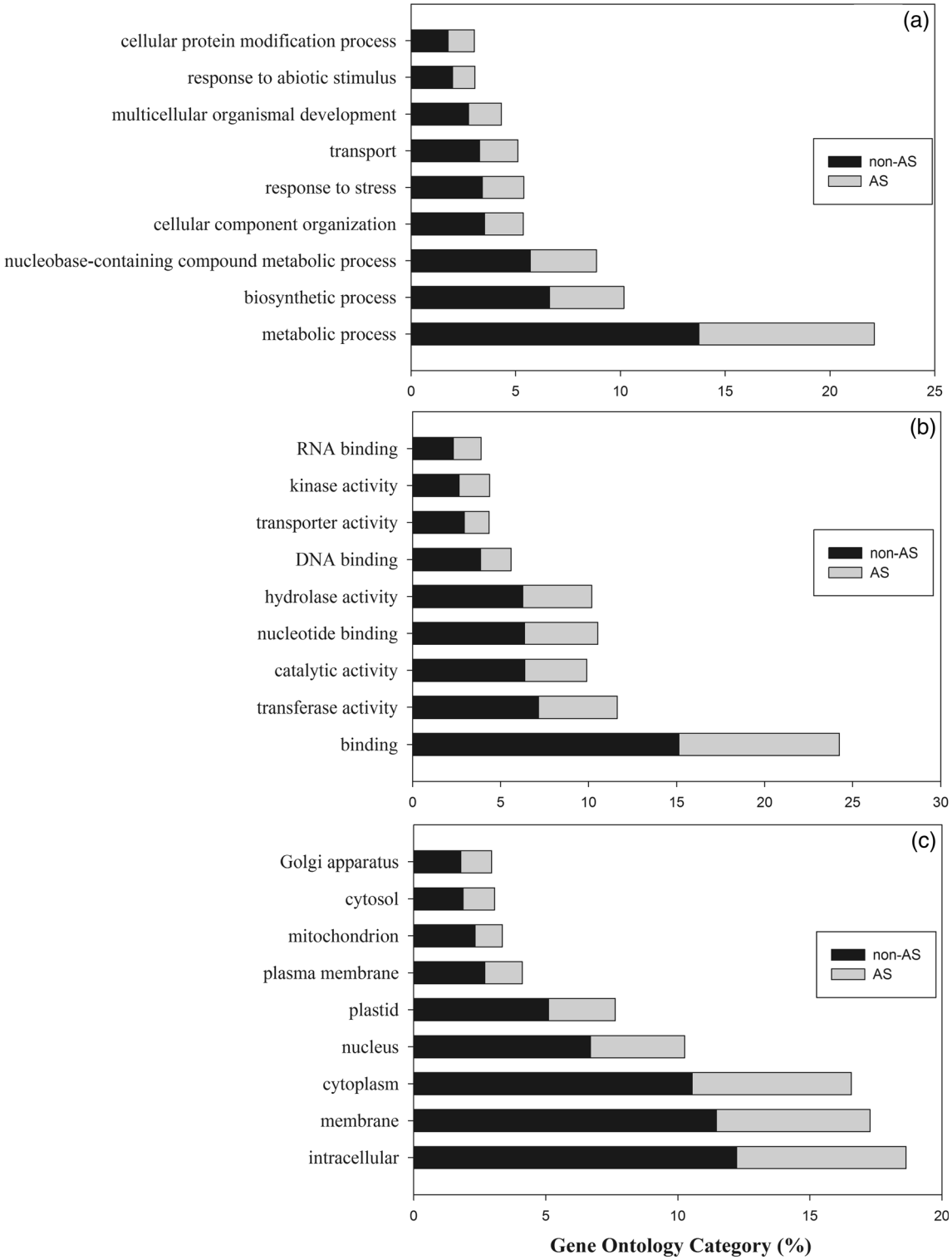


Fig. 2 Gene Ontology (GO) classification of pineapple genes with pre-mRNAs not undergoing alternative splicing (non-AS genes) and genes with pre-mRNA undergoing alternative splicing (AS-genes). **a** Biological process; **b** Molecular function; **c** Cellular component

are transcriptionally regulated by AS in Arabidopsis and rice (Li et al. 2006; Zhao and Beers 2013). Alternative splicing of MYB DNA-binding related genes MYR1 and MYR2 have clearly demonstrated the change in protein dimerization and folding as a consequence of AS, thus contributing to modulation of MYR1 and MYR2 activities as regulators of flowering time (Zhao and Beers 2013). Another example is the reticulon protein family (Table 3). The reticulon family is a large and diverse group of

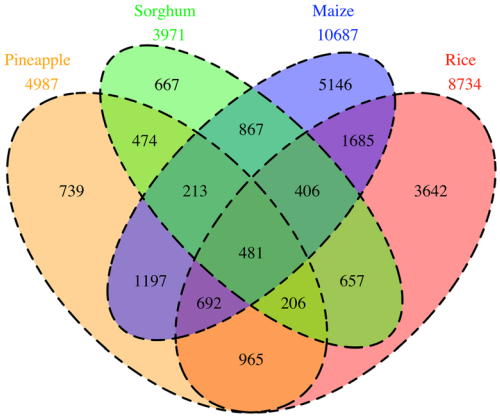


Fig. 3 Conserved alternatively spliced genes among pineapple, rice, sorghum, and maize

membrane-associated proteins found throughout the eukaryotic kingdom. Reticulons principally localize to the endoplasmic reticulum, and there is evidence that they influence endoplasmic reticulum-Golgi trafficking, vesicle formation and membrane morphogenesis (Yang and Strittmatter 2007). The reticulon genes are regulated by AS in both mammals and Arabidopsis (Di Scala et al. 2005; Nziengui et al. 2007). Drosophila ferritin mRNA has been known under AS regulation (Lind et al. 1998), and plant ferritins were found targeting both mitochondria and chloroplast and were proposed that feature resulting from alternative transcription, alternative translation starts, alternative exon splicing (Zancani et al. 2004). Thus biological functional significance of the AS genes encoding ferritin-like domain conserved in these five monocot plants is worth further examination.

Gene Expression

In pineapple genome, a total of 27,024 protein-coding genes were predicted (Ming et al. 2015). Among these predicted gene models, 4967 genes were not expressed in our sampled tissues. A total of 22,057 genes were expressed and detected by RNA-seq, and among them, 9338 genes having a FPKM value of <10 in all tissues were treated as lowly expressed genes; 12,719 genes had ≥10 FPKM in at least one tissue. Further among these 12,719 genes, 9301 showed no significant differential expression among sampled tissues including leaf, root, flower and fruit; and 3418, genes were found differentially expressed between two or more different tissues, i. e. at least two fold changes of the log2 values of FPKM (Supplementary Table 5). As in our experiments we used pooled tissues without biological replicates, thus, no statistical tests were carried out, we compared several genes having gene expression data obtained using quantitative real-time PCR (qRT-PCR) in literature. The FLOWERING LOCUS T-like gene (GenBank accession number: HQ343233, Gene model: Aco010684) was highly expressed in fruit flesh tissue and extremely lowly expressed in root and leaf tissue, our RNA-seq data were consistent with the results obtained by using qRT-PCR (Lv et al. 2012a). However, the pineapple PISTILLATA (PI)-like gene, a regulator of flowering in angiosperms, AcPI (Genbank: HQ717796, Gene model: Aco019365) isolated from pineapple cultivar Comte de Paris, was found to be highly expressed in floral organs but lowly expressed in leaf tissue using qRT-PCR (Lv et al. 2012b) while our data showed the gene were equally highly expressed in both leaf and flower tissues. Thus, RNA-seq data were well suited for screening differentially expressed genes and qRT-PCR might be still needed to validate gene expression levels.

We compared the differentially expressed genes list (3418 genes) with the alternatively spliced gene list (5146 genes) and found that 701 genes were common on both lists, i. e., these genes were differentially expressed among different tissues and were also alternatively spliced (Supplementary Table 6). Thus we suggest that these genes need to be particularly examined at the AS isoform level in future gene expression study.

As the ten key CAM carboxylation and decarboxylation genes has been identified based on its diurnal transcript expression (Ming et al. 2015), we wonder if AS events play a regulatory role in CAM genes among different tissues. Among the ten CAM genes, five of them have AS events detected in at least one of the four tissues (leaf, root, flower, and fruit) examined. These fives genes are beta-carbonic anhydrase (beta-CA, Aco006181), phosphoenolpyruvate carboxylase (PEPC, Aco010025), cytosolic malate dehydrogenase (MDH, Aco013935), chloroplastic malate dehydrogenase (Aco010232) and pyruvate, phosphate dikinase (PPDK, Aco024818). For cytosolic MDH, it contains four isoforms that two contains partial open reading frame and two with complete open reading frame. Isoform c87308_g1_i1 is the dominant transcripts with the highest expression in all four tissues. However, another cytosolic MDH isoform DT336599 could be detected in flower, root and fruit at high abundance but no expression at all in leaf (Table 4), suggesting that isoform DT336599 is not transcribed in leaf. Another interesting AS events was found in the three PPDK isoforms, whereas the two longest ones contain the same Pfam domain while the shortest one has a different Pfam domain. For the two longest PPDK isoforms with same Pfam domain, c81326_g1_i1 and Ac2740, the former has similar expression level in flower and leaf but 2 and 15-fold higher expression in root and fruit, respectively, than Ac2740. The leaf diurnal expression and functional role of the MDH and PPDK isoforms in different tissue types, thus, should be further investigated to dissect the spatio-temporal expression of these CAM gene isoforms.

Table 3 Genes with pre-mRNAs undergoing alternative splicing conserved in five monocot plants

Pineapple	Maize	Rice	Bd	Sorghum	Pfam with description
Aco000284	Zm102071	Osj1997	Bd16220	Sb876	pfam00450, Peptidase_S10, Serine carboxypeptidase
Aco000895	Zm100142	Osj195	Bd2265	Sb1070	No Pfam/Domain predicted
Aco001241	Zm59942	Osj491	Bd27405	Sb3313	pfam00125, Histone, Core histone H2A/H2B/H3/H4
Aco001519	Zm6058	Osj17274	Bd21664	Sb11340	pfam00072, Response_reg, Response regulator receiver domain
Aco001727	Zm66360	Osj36146	Bd14282	Sb6058	pfam06964, Alpha-L-AF_C, Alpha-L-arabinofuranosidase C-terminus
Aco001737	Zm40809	Osj27404	Bd11999	Sb13248	pfam14144, DOG1, Seed dormancy control
Aco003427	Zm104171	Osj9792	Bd7833	Sb19539	pfam01156, IU_nuc_hydro, Inosine-uridine preferring nucleoside
Aco004765	Zm26294	Osj43203	Bd25836	Sb13448	pfam05383, La, La domain
Aco005053	Zm5068	Osj24220	Bd23246	Sb10671	pfam02453, Reticulon, Reticulon
Aco005061	Zm91134	Osj44927	Bd208	Sb15888	pfam06943, zf-LSD1, LSD1 zinc finger
Aco007079	Zm58947	Osj16392	Bd7531597	Sb3303	pfam00210, Ferritin, Ferritin-like domain
Aco007659	Zm46142	Osj16693	Bd7810	Sb13903	pfam05623, DUF789, Protein of unknown function (DUF789)
Aco007682	Zm67002	Osj20481	Bd5031591	Sb7321	pfam07690, MFS_1, Major Facilitator Superfamily
Aco007928	Zm70017	Osj14932	Bd3731595	Sb10575	pfam00676, E1_dh, Dehydrogenase E1 component
Aco008105	Zm35072	Osj15328	Bd29210	Sb10885	pfam00439, Bromodomain, Bromodomain
Aco008819	Zm38497	Osj36865	Bd9583	Sb20674	pfam00447, HSF_DNA-bind, HSF-type DNA-binding
Aco009281	Zm35123	AB091673.1	Bd2755	Sb11681	pfam00083, Sugar_tr, Sugar (and other) transporter
Aco009409	Zm35625	Osj22934	Bd7027	Sb15873	pfam01370, Epimerase, NAD dependent epimerase/dehydratase family
Aco009509	Zm8024	Osj19926	Bd4187	Sb10040	pfam04366, DUF500, Family of unknown function (DUF500)
Aco009671	Zm66716	Osj43789	Bd4539	Sb313	pfam02309, AUX_IAA, AUX/IAA family
Aco010214	Zm92934	Osj954	Bd2565	Sb6267	pfam03171, 2OG-FeII_Oxy, 2OG-Fe(II) oxygenase superfamily
Aco010377	Zm98577	Osj35805	Bd20683	242066955	pfam12076, Wax2_C, WAX2 C-terminal domain
Aco010968	Zm101865	Osj7689	Bd25885	Sb5520	pfam00719, Pyrophosphatase, Inorganic pyrophosphatase
Aco010976	Zm86065	Osj23668	Bd16506	Sb718	pfam00249, Myb_DNA-binding, Myb-like DNA-binding domain
Aco011354	Zm5213	Osj42344	Bd15760	Sb15538	pfam03094, Mlo, Mlo family
Aco012425	Zm24118	Osj47510	Bd8231593	Sb11323	pfam03214, RGP, Reversibly glycosylated polypeptide
Aco012837	Zm34387	Osj26278	Bd17237	Sb1242	pfam00226, DnaJ, DnaJ domain
Aco013154	Zm20714	Osj20205	Bd6374	Sb12056	pfam00762, Ferrochelatase, Ferrochelatase
Aco013886	Zm33939	Osj21144	Bd23758	Sb5787	pfam14299, PP2, Phloem protein 2
Aco014140	Zm37855	Osj20062	Bd24117	Sb6825	No Pfam/Domain predicted
Aco014288	Zm34294	Osj16408	Bd19455	57806619	sd00025, zf-RanBP2, RanBP2-type zinc finger
Aco015388	Zm39479	Osj4519	Bd24331	Sb9831	No Pfam/Domain predicted
Aco015658	Zm55675	Osj48001	Bd25775	18069991	pfam01805, Surp, Surp module
Aco015994	Zm90776	Osj3505	Bd10066	Sb16899	pfam00145, DNA_methylase, C-5 cytosine-specific DNA methylase
Aco016559	Zm2002	Osj43276	Bd7425	Sb5602	pfam00149, Metallophos, Calcineurin-like phosphoesterase
Aco016684	Zm78764	Osj32459	Bd21043	Sb4094	No Pfam/Domain predicted
Aco016862	Zm46954	Osj41915	Bd268	Sb1476	pfam00316, FBPase, Fructose-1-6-bisphosphatase
Aco018139	Zm59883	Osj15126	Bd15932	Sb673	pfam01636, APH, Phosphotransferase enzyme family
Aco018596	Zm94988	Osj22529	Bd13677	Sb9205	pfam02800, Gp_dh_C, Glyceraldehyde 3-phosphate dehydrogenase,
Aco019033	Zm47385	Osj42282	Bd31189	Sb2377	pfam00504, Chloroa_b-bind, Chlorophyll A-B binding protein
Aco019609	Zm29988	Osj42011	Bd19593	30165373	pfam00291, PALP, Pyridoxal-phosphate dependent enzyme
Aco020426	Zm101909	Osj1909	Bd22028	Sb1160	pfam00179, UQ_con, Ubiquitin-conjugating enzyme
Aco021295	Zm104454	Osj19199	Bd16056	Sb12015	pfam00450, Peptidase_S10, Serine carboxypeptidase
Aco022282	Zm39790	Osj42201	Bd8363	Sb2138	pfam00348, polyprenyl_synt, Polyprenyl synthetase
Aco023992	NM_001112569.1	Osj37723	Bd3697	Sb10731	pfam08323, Glyco_transf_5, Starch synthase catalytic domain
Aco024253	FL103380	Osj16673	Bd4031	242042528	pfam07876, Dabb, Stress responsive A/B Barrel Domain
Aco024300	Zm58340	Osj19165	Bd2531589	Sb10339	pfam00650, CRAL_TRIO, CRAL/TRIO domain
Aco024325	EC873417	Osj29038	Bd1692	Sb19219	pfam07415, Herpes_LMP2, Gammaherpesvirus latent membrane protein
Aco026809	Zm88316	Osj22392	Bd7352	45969421	pfam00248, Aldo_ket_red, Aldo/keto reductase family
Aco026855	Zm101869	Osj20257	Bd31434	Sb7733	pfam04777, Evr1_Alr, Erv1/Alr family
Aco029835	Zm34407	NM_001187881.1	Bd25606	Sb11634	pfam03151, TPT, Triose-phosphate Transporter family

Bd *Brachypodium distachyon*

Table 4 CAM carboxylation and decarboxylation gene isoforms expression among four pineapple tissues

Gene	Gene ID	Isoform ID	Pfam domain	Full/partial ORF	Length	FPKM			
						flower	leaf	root	fruit
beta-CA	Aco006181	c80803_g1_i1	Pfam00484	Full	1068	1884	1984	39	66
		c80803_g1_i2	Pfam00484	Full	1146	1932	2565	10	0
		c80803_g1_i3	Pfam00484	Full	744	21	44	2	0
PEPC	Aco010025	JI400975.1	-	Full	262	166	159	70	310
		AJ312629.1	Pfam00311	Partial	715	384	323	204	670
		AB523411.1	Pfam00311	Full	2938	4781	4428	2436	10,233
		Ac6577	Pfam00311	Full	3249	0	0	0	59
		Ac6578	Pfam00311	Full	3349	86	51	118	97
Cytosolic MDH	Aco013935	DT336869	Pfam00056	Partial	929	0	0	0	25
		DT336599	Pfam00056	Partial	895	604	0	591	3162
		c87308_g1_i1	Pfam02866	Full	1490	1053	1118	744	4067
		Ac96	Pfam02866	Full	1260	92	119	61	54
Chloroplastic MDH	Aco010232	JI412072.1	-	Partial	226	44	65	16	11
		c89145_g1_i2	Pfam02866	Full	1239	57	0	11	0
PPDK	Aco024818	c81326_g1_i1	Pfam02896	Full	2627	4398	4498	1083	1371
		c81326_g1_i2	Pfam01326	Full	1631	81	66	2	0
		Ac2740	Pfam02896	Full	3077	4937	3966	471	87

beta-CA beta-carbonic anhydrase; PEPC phosphoenolpyruvate carboxylase; MDH malate dehydrogenase; PPDK pyruvate, phosphate dikinase; ORF open reading frame; FPKM fragments per kilobase of transcript length per million of mapped reads

Concluding Remarks

Pineapple is an important economic fruit species, however, is less well studied at the molecular level. The availability of its genome sequence and a large amount of ESTs and assembled transcripts generated by NGS provides an extraordinary opportunity for studying the growth, development, CAM pathway, regulation of flowering and fruiting, and other biological processes at the molecular level. Our analysis in cereal crops and other intensive studies carried out in Arabidopsis and other plants showed that AS is common in plants (Min et al. 2015; Wang and Brendel 2006; Reddy et al. 2013). In Arabidopsis, phytochrome was found to control alternative splicing to mediate light responses, ~6.9 % (1500–1700 genes) of the annotated genes in its genome had alternative splicing under the control of phytochrome (Shikata et al. 2014). Certainly other environment and developmental factors also regulate plant AS through regulating of splicing factors (reviewed by Staiger and Brown 2013). In this study, we identified 5146 predicted protein coding genes generated pre-mRNAs undergoing AS in pineapple. The detailed annotation of these assembled transcripts along with aligned genomic sequences and graphic visualization presented in our database will facilitate researchers to design experiments to investigate the biological significances of these genes in pineapple growth, development, and responses to environmental stresses.

Materials and Methods

Plant Materials

Leaf, root and flower tissues were collected from pineapple var. F153 plants and 5 different stages of fruits were harvested from pineapple hybrid MD-2 for RNA extraction and transcriptome sequencing. The 'F153' plants were maintained in Hawaii Agriculture Research Center at Kunia, HI and the 'MD-2' plants were maintained in Dole plantation at Wahiawa, HI.

RNA Extraction, Library Construction and Sequencing of Pineapple Tissues

RNA was extracted from leaf, root and flower tissues using Qiagen RNeasy Plant Mini Kit (Qiagen, #74,904), following manufacturer's protocol. For pineapple fruit RNA, it was extracted using salt buffer followed with lithium chloride and isopropanol precipitation. Then, DNA was removed with DNA-free™ DNA Removal Kit (Life Technologies, #AM190 6 M). Single indexed RNA-seq library was constructed using Illumina TruSeq stranded RNA Sample Preparation Kit (Illumina, #RS-122-2001) and then sequenced by Illumina HiSeq2500 in single end 100 nt mode.

Sequence Read Processing and Expression Analysis

For expression profiling at gene level, the trimmed single end reads of each sample were aligned to repeat-masked pineapple assembly version 3 using TopHat v2.0.9 default setting (Trapnell et al. 2012). The normalized FPKM (fragment per kilobase exon per million read mapped) value of each sample were estimated by Cufflinks v2.2.1, followed by Cuffnorm v2.2.1 using default setting with pineapple gene model annotation provided (−g option). For expression profiling at isoform level, the single end reads were aligned to putative unique transcripts (PUTs) set using TopHat2 and Cufflinks package with the same method stated above. We averaged the expression levels of five stages of immature fruit into one fruit sample, and pairwise compared with the other three tissues (leave, root, and flower). We further identified genes down-regulated or up-regulated based on at least two fold changes of the log2 values of FPKM, however, genes having less than 10 FPKM in all tissues were treated as lowly expressed genes and not included in the pairwise expression comparison.

Sequence Datasets

The RNA-seq data generated in this project were assembled using Trinity program, a total of 61,143 tentative transcripts were generated. Combining pineapple 5941 ESTs and 11,907 mRNA sequences downloaded from the National Center for Biotechnology Information (NCBI) dbEST and nucleotide database with the Trinity assembled data, we obtained a total of 78,991 sequences. The ESTs and mRNA sequences in the NCBI nucleotide database were mainly deposited from two previous reports (Moyle et al. 2005; Ong et al. 2012). The data were cleaned using the following procedure: 1) using EMBOSS trimmest tool to trim the polyA or polyT end; 2) using BLASTN to search UniVec and *E. coli* database for removing vector and *E. coli* contaminants, 3) using BLASTN to search a plant repeat database which was built with TIGR gramineae repeat data, (ftp://ftp.plantbiology.msu.edu/pub/data/TIGR_Plant_Repeats/). After cleaning, we obtained 78,950 sequences for assembling using CAP3 with the following parameters: -p 95 –o 50 –g 3 –y 50 –t 1000 (Huang and Madan 1999). After assembling, we obtained a total of 63,991 putative unique transcripts (PUTs) including 7518 contigs (named as Ac#) and 56,473 singletons. The above procedure was previously used for recent work in cereal crop AS analysis (Min et al. 2015).

PUTs Mapping, Identification and Functional Annotation of AS Isoforms

Mapping of putative unique transcripts (PUTs) to their corresponding genomic sequences and identification of AS isoforms was carried out using ASFinder (http://proteomics.ysu.edu/tools/ASFinder.html/) (Min 2013). ASFinder uses SIM4 program (Florea et al. 1998) to map PUTs to the corresponding genome and then subsequently identifies those PUTs that are mapped to the same genomic location but have variable exon-intron boundaries as AS isoforms. For PUT mapping, the following thresholds were used: a 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. The output file of ASFinder was then submitted to AStalavista server (http://genome.crg.es/astalavista/) for AS event analysis (Foissac and Sammeth 2007). The percentage of alternative splicing genes was estimated using the predicted genes models having alternative splicing PUT isoforms among total gene models having at least one PUT.

We predicted the open reading frame (ORF), i.e. protein coding region, of each PUT using the OrfPredictor (Min et al. 2005a) and assessed the full–length transcript coverage using TargetIdentifier (Min et al. 2005b) as previously described. Functional classification was assigned to the PUTs by performing BLASTX searches with an E-value threshold of 1e-5 against UniProtKB/Swiss-Prot dataset. Additionally, predicted protein sequences from OrfPredictor were further annotated using rpsBLAST against the Pfam database using an E-value of 1e-10 (http://pfam.xfam.org/). The classification of Gene Ontologies (GOs) was assigned on the basis of the functional homology using the BLASTX searching against the UniProtKB/Swiss-Prot dataset. The GO IDs were extracted based on the UniProtKB protein identifiers from the ID mapping table downloaded from UniProtKB. The GO categories were further analyzed using GO SlimViewer using plant specific GO terms (McCarthy et al. 2006). To assess the functional coverage of the assembled PUTs, we further compared PUTs against the predicted gene primary transcripts using BLASTN with a cut off E-value of 1e-10, ≥ 95 % identity and minimum aligned length of 80 bp.

Conserved Genes Undergoing AS in Monocot Plants

For the identification of the potentially conserved genes, which generate pre-mRNAs undergoing AS, among pineapple, rice (*japonica*), sorghum, *Brachypodium*, and maize, reciprocal BLASTP (cutoff E-value 1e-10) were carried out using the longest (or longer) ORF of the AS PUT isoforms for classifying the conserved AS pairs between each pair of species. Then the AS pairs were combined for identification of conserved AS genes among the five species.

Data Availability

AS events identified in this study along with annotation of AS genes identified in our previous work in other plants are

available at Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/) (VanBuren et al. 2013; Walters et al. 2013; Min et al. 2015). Additionally, the identified AS events can be visualized and compared with predicted gene models using GBrowse for comparative assessment. We also deployed BLASTN tool to search for the assembled sequences. The sequences and annotation data along with the GO, Pfam, gene expression and other comparative analysis are publicly available for downloading at: http://bioinformatics.ysu.edu/publication/data/Pineapple/.

Acknowledgments The work was supported by the University of Illinois at Urbana-Champaign to RM and Youngstown State University to XJM.

References

Bartholomew DP (2013) History and perspectives on the role of ethylene in pineapple flowering. In: XII international symposium on plant Bioregulators in fruit production. Acta Hortic 1042:269–284

Bartholomew DP, Kadzimin SB (1977) Pineapple. In: Alvin PT, Kozeowski TT (eds) Ecophysiology of tropical crops. Academic Press, New York, NY, pp. 113–156

Bartholomew DP, Malézieux EP (1994) Pineapple. In: Schaffer B, Andersen PC (eds) Handbook of environmental physiology of fruit crops, vol 2. CRC Press, Boca Raton, pp. 243–291

Bartholomew DP, Paull RE, Rohrbach KG (eds) (2002) The pineapple: botany, production, and uses. CABI, Wallingford

Barz M, Delivand MK (2011) Agricultural residues as promising biofuels for biomass power generation in Thailand. J Sustainable Energy Environment Special Issue 2011:21–27

Burg SP, Burg EA (1966) Auxin-induced ethylene formation: its relation to flowering in the pineapple. Science 152:1269–1269

Di Scala F, Dupuis L, Gaiddon C, De Tapia M, Jokic N, Gonzalez de Aguilar JL, Raul JS, Ludes B, Loeffler JP (2005) Tissue specificity and regulation of the N-terminal diversity of reticulon 3. Biochem J 385(Pt 1):125–134

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8:967–974

Foissac S, Sammeth M (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Res 35:W297–W299

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

Lind MI, Ekengren S, Melefors Ö, Söderhäll K (1998) Drosophila ferritin mRNA: alternative RNA splicing regulates the presence of the iron-responsive element. FEBS Lett 436:476–482

Lum G, Meinken J, Orr J, Frazier S, Min XJ (2014) PlantSecKB: the plant secretome and subcellular proteome knowledgebase. Comput Molec Biol 4:1–17

Li J, Li X, Guo L, et al. (2006) A subgroup of MYB transcription factor genes undergoes highly conserved alternative splicing in Arabidopsis and rice. J Exp Bot 57:1263–1273

Lv L, Duan J, Xie J, Wei C, Liu Y, Liu S, Sun G (2012a) Isolation and characterization of a FLOWERING LOCUS T homolog from pineapple (*Ananas comosus* (L.) Merr). Gene 505:368–373

Lv LL, Duan J, Xie JH, Liu YG, Wei CB, Liu SH, Zhang JX, Sun GM (2012b) Cloning and expression analysis of a PISTILLATA homologous gene from pineapple (*Ananas comosus* L. Merr). Int J Mol Sci 13:1039–1053

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res 22:1184–1195

McCarthy FM, Wang N, Magee GB, Williams WP, Luthe DS, Burgess SC (2006) AgBase: a functional genomics resource for agriculture. BMC Genomics 7:229

Min X, Bartholomew DP (1996) Effect of plant growth regulators on ethylene production, 1-aminocyclopropane-1-carboxylic acid oxidase activity, and initiation of inflorescence development of pineapple. J Plant Growth Regul 15:121–128

Min XJ, Powell B, Braessler J, Meinken J, Yu F, Sablok G (2015) Genome-wide cataloging and analysis of alternatively spliced genes in cereal crops. BMC Genomics 16:721

Min XJ, Butler G, Storms R, Tsang A (2005a) OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res 33:W677–W680

Min XJ, Butler G, Storms R, Tsang A (2005b) TargetIdentifier: a web server for identifying full-length cDNAs from EST sequences. Nucleic Acids Res 33:W669–W672

Min XJ (2013) ASFinder: a tool for genome-wide identification of alternatively spliced transcripts from EST-derived sequences. Int J Bioinforma Res Appl 9:221–226

Ming R, VanBuren R, Wai CM, et al. (2015) The pineapple genome and the evolution of CAM photosynthesis. Nat Genet. doi:10.1038/ng.3435

Morello L, Breviario D (2008) Plant spliceosomal introns: not only cut and paste. Curr Genet 9:227–238

Moyle R, Fairbairn DJ, Ripi J, Crowe M, Botella JR (2005) Developing pineapple fruit has a small transcriptome dominated by metallothionein. J Exp Bot 56:101–112

Nievola CC, Kraus JE, Freschi L, Souza BM, Mercier H (2005) Temperature determines the occurrence of CAM or C3 photosynthesis in pineapple plantlets grown in vitro. In Vitro Cellular Dev Biol-Plant 41:832–837

Nziengui H, Bouhidel K, Pillon D, Der C, Marty F, Schoefs B (2007) Reticulon-like proteins in Arabidopsis Thaliana: structural organization and ER localization. FEBS Lett 581:3356–3362

Ong WD, Voo LYC, Kumar VS (2012) De novo assembly, characterization and functional annotation of pineapple fruit transcriptome through massively parallel sequencing. PLoS One 7:e46937

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

Reddy AS, Marquez Y, Kalyna M, Barta A (2013) Complexity of the alternative splicing landscape in plants. Plant Cell 25:3657–3683

Sablok G, Gupta PK, Baek JM, Vazquez F, Min XJ (2011) Genome-wide survey of alternative splicing in the grass Brachypodium Distachyon: an emerging model biosystem for plant functional genomics. Biotechnol Lett 33:629–636

Sablok G, Harikrishna JA, Min XJ (2013) Next generation sequencing for better understanding alternative splicing: way ahead for model and non-model plants. Transcriptomics 1:e103

Shikata H, Hanada K, Ushijima T, Nakashima M, Suzuki Y, Matsushita T (2014) Phytochrome controls alternative splicing to mediate light responses in Arabidopsis. Proc Natl Acad Sci U S A 111:18781–18786

Staiger D, Brown JW (2013) Alternative splicing at the intersection of biological timing, development, and stress responses. Plant Cell 25:3640–3656

Surles T, Foley M, Turn S, Staackmann M (2009) A scenario for accelerated use of renewable resources for transportation fuels in Hawai ‘i. University of Hawaii, Hawaii Natural Energy Institute, School of Ocean and Earth Science and Technology, pp. 1–38

Taussig SJ, Batkin S (1988) Bromelain, the enzyme complex of pineapple (*Ananas comosus*) and its clinical application: an update. J Ethnopharmacol 22:191–203

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc 7:562–578

Trusov Y, Botella JR (2006) Silencing of the ACC synthase gene ACACS2 causes delayed flowering in pineapple [*Ananas comosus* (L.) Merr.]. J Exp Bot 57:3953–3960

VanBuren R, Walters B, Ming R, Min XJ (2013) Analysis of expressed sequence tags and alternative splicing genes in sacred lotus (Nelumbo Nucifera Gaertn.). Plant Omics J 6:311–317

Walters B, Lum G, Sablok G, Min XJ (2013) Genome-wide landscape of alternative splicing events in Brachypodium Distachyon. DNA Res 20:163–171

Wang B, Brendel V (2006) Genome wide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci U S A 103:7175–7180

Wang RH, Hsu YM, Bartholomew DP, Maruthasalam S, Lin CH (2007) Delaying natural flowering in pineapple through foliar application of aviglycine, an inhibitor of ethylene biosynthesis. HortSci 42:1188–1191

Yang YS, Strittmatter SM (2007) The reticulons: a family of proteins with diverse functions. Genome Biol 8:234

Zancani M, Peresson C, Biroccio A, Federici G, Urbani A, Murgia I, et al. (2004) Evidence for the presence of ferritin in plant mitochondria. Eur J Biochem 271:3657–3664

Zhang J, Liu J, Ming R (2014) Genomic analyses of the CAM plant pineapple. J Exp Bot 65:3395–3404

Zhao C, Beers E (2013) Alternative splicing of Myb-related genes MYR1 and MYR2 may modulate activities through changes in dimerization, localization, or protein folding. Plant Signal Behav 11:e27325