Computational Molecular Biology 2017, Vol.7, No.1, 1-11
http://cmb.biopublisher.ca

Research Article **Open Access**

Comprehensive Cataloging and Analysis of Alternative Splicing in Maize

Min X.J.
Department of Biological Sciences, Center for Applied Chemical Biology, Youngstown State University, Youngstown, OH 44555, USA
Corresponding author Email: xmin@ysu.edu

Computational Molecular Biology, 2017, Vol.7, No.1 doi: 10.5376/cmb.2017.07.0001
Received: 20 July, 2017
Accepted: 01 Sep., 2017
Published:,04 Sep., 2017

Preferred citation for this article:
Min X.J., 2017, Comprehensive cataloging and analysis of alternative splicing in maize, Computational Molecular Biology, 7(1): 1-11 (doi: 10.5376/cmb.2017.07.0001)

Abstract Gene expression is a key step in developmental regulation and responses in changing environments in plants. Alternative splicing (AS) is a process generating multiple RNA isoforms from a single gene pre-mRNA transcript that increases the diversity of functional proteins and RNAs. Identification and analysis of alternatively splicing events are critical for crop improvement and understanding regulatory mechanisms. In maize large numbers of transcripts generated by RNA-seq technology are available, we incorporated these data with data assembled with ESTs and mRNAs to comprehensively catalog all genes having pre-mRNAs undergoing AS. A total of 192 624 AS events were detected and classified, including 103 566 (53.8%) basic events and 89 058 (46.2%) complex events which were formed by combination of various types of basic events. Intron retention was the dominant type of basic AS event, accounting for 24.1%. These AS events were identified from 91 128 transcripts which were generated from 26 669 genomic loci, of which consisted of 20 860 gene models. It was estimated that 55.3% maize genes may be subjected to AS. The transcripts mapping information can be used to improve the predicted gene models in maize. The data can be accessed at Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/).

Keywords Alternative splicing; Cereal crops; Gene expression; Maize; mRNA

Introduction

Maize (*Zea mays* subsp. *mays*) is an important food, feed and biofuel crop. It is also an important model organism for fundamental research in genetics, genomics and plant physiology. Its genome consisting of 10 chromosomes and having a size of ~2.3 gigabases has been completed sequenced, with 32 475 protein coding gene models predicted (Schnable et al., 2009; Andorf et al., 2016). Gene expression in plants is a highly regulated process during plant growth and development as well as in response to changing environments. Alternative splicing (AS) is a process generating more than one transcript from one pre-mRNA in gene transcription (Reddy et al., 2013). There are four basic types of AS, including exon skipping (ES), alternative donor site (AltD), alternative acceptor (AltA) site, and intron retention (IR). Various complex types can be formed by combination of basic events (Sablok et al., 2011). In addition to the above mentioned AS types, alternative transcripts may arise as a consequence of the alternative transcription initiation, alternative transcription termination, and alternative polyadenylation (Roberts et al., 2002). An AS transcript isoform may or may not encode a distinct functional protein. However, when harboring a premature termination codon in an AS isoform, the encoded protein may be nonfunctional. The nonfunctional isoforms are degraded by a process known as nonsense-mediated decay (NMD) (Lewis et al., 2003).

AS plays a major role in expanding the transcriptome and proteome diversity in plants, with 60 % of multi-exon genes undergoing alternative splicing in *Arabidopsis thaliana* (Carvalho et al., 2013; Yu et al., 2016). Genome-wide identification and physiological implications of AS have been reported in plant species including *A. thaliana* (Filichkin et al., 2010; Zhang et al., 2010; Marquez et al., 2012; Syed et al., 2012), *Oryza sativa* (Wang and Brendel, 2006), *Nelumbo nucifera* (sacred lotus) (VanBuren et al., 2013), *Vitis vinifera* (Vitulo et al., 2014), *Brachypodium distachyon* (Sablok et al., 2011; Walters et al., 2013), *Zea mays* (maize), and *Sorghum bicolor* (sorghum) (Thatcher et al., 2014; Min et al., 2015). 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 (Stamm et al., 2005). 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 (Reddy et al., 2013; Sablok et al., 2011). On contrast, recent reports suggest the down-regulation of the IR events and up-regulation of the alternative donor/acceptor site (AltDA) and ES under heat stress in model *Physcomitrella patens* (Chang et al., 2014). With the advent of the Next Generation Sequencing (NGS) based approaches, fine scale physiological implications revealed that AS increasing the complexity of the alternative mRNA processing which involved in the microRNA-mediated gene regulation in Arabidopsis (Yang et al., 2012). Complex networks of regulation of gene expression and variation in AS has played a major role in the adaptation of plants to their corresponding environment (Syed et al., 2012) and additionally in coping with environmental stresses.

Rice (ssp *japonica* and *indica*), maize, and sorghum are important cereal crops as major sources of food in many countries. Previously several approaches have widely demonstrated the identification of the quantitative trait loci, genes and proteins linked to the functional grain content in these species (Mao et al., 2010). However, a major portion of the gene functional diversity is controlled by a spliceosomal regulated AS. AS has been shown to be a critical regulator in grass clade, demonstrating several of the genes involved in flowering and abiotic stress depicting alternative splicing (Reddy et al., 2013; Walter et al., 2013; Staiger et al., 2013). Identifying genes with pre-mRNAs undergoing alternative splicing in these cereal plants is critical in understanding the functions and regulations of these genes in plant development and abiotic or biotic stress resistance. Previously, using the homology based mapping approach and expressed sequence tags (ESTs) representing the functional transcripts, we identified a total of 941 AS genes in *Brachypodium distachyon*, a model temperate grass (Sablok et al., 2011; Walters et al., 2013). Previous reports on the identification and prevalence of the alternative splicing events in rice (Campbell et al., 2006; Wang and Brendel, 2006), sorghum (Panahi et al., 2014), and maize (Thatcher et al., 2014) have shown the functional diversity changes through EST/RNA-seq approaches. Recently we also reported our efforts in identification of AS genes in rice (both *japonica* and *indica*), maize, and sorgum (Min et al., 2015). We compared the AS event landscape and the AS gene functional diversity in cereal plants and also comparatively analyzed these AS genes with AS genes identified from *B. distachyon* to reveal conserved patterns of the AS across the grass species. In this work, we incorporated more transcripts data generated using RNA-seq technologies and significantly expanded the list of genes with their mRNAs undergoing AS in maize.

1 Materials and Methods

1.1 Sequence datasets and sequence assembly

In order to comprehensively identify all possible AS events in maize, multiple sources of maize expressed transcripts were integrated including expressed sequence tags (ESTs), mRNA sequences, and transcripts assembled from RNA-seq data. The data sources consisted of a total of pre-assembled 778 172 transcripts obtained from four sources: (1) 488 243 putative unique transcripts (PUTs) assembled with over 2 million of expressed sequence tags and mRNA sequences which were collected from NCBI dbEST and nucleotide database (as of Oct 2013) (Min et al., 2015); (2) 181 779 transcripts assembled from over 200 RNA-seq libraries (Thatcher et al., 2014); (3) 48 432 novel transcript isoforms identified from 147 RNA-seq libraries generated in different developmental stages with and without drought stresses (Thatcher et al., 2016), these sequences were extracted using the version 2 maize genome based on the mapping information provided in the Sup. Table 1 (Thatcher et al., 2015); and (4) recently deposited 59 263 mRNA sequences and 465 ESTs (from Oct 2013 to Dec 2015) with their polyA/T ends trimmed using trimmest tool in the EMBOSS package (Rice et al., 2000). The combined data consisting of a total of 767 717 transcripts were re-assembled using CAP3 with the following parameters: -p95-o40-g3-y50-t1000 (Huang and Madan, 1999). A total of 614 201 putative unique transcripts (named as Mz#) (PUTs) were obtained including 73 089 contigs and 541 112 singlets for downstream mapping to maize genome sequences.

1.2 Mapping PUTs to genome and identification AS events

The maize genome assembly and gene models (B73 RefGen_v3.22) was downloaded from maizeGDB

(http://www.maizegdb.org/assembly/). The assembled PUTs were mapped to their corresponding chromosomes using ASFinder (http://proteomics.ysu.edu/tools/ASFinder.html/) (Min, 2013). ASFinder uses SIM4 program (Florea et al., 1998) to align PUTs to the genome, and then subsequently identifies those PUTs that are mapped to the same genomic location but have variable exon-intron boundaries as AS isoforms. To avoid the spurious mapping, we applied a threshold of minimum of 95% identity for all aligned PUT with a genomic segment (exon), a minimum of 80 bp aligned length, and >75% of a PUT sequence aligned to the genome (Walters et al., 2013). To avoid chimeric assemblies, mapped PUTs having an intron size >100 kb were removed for AS identification. The output file (AS.gtf) of ASFinder was then subsequently submitted to AStalavista server (http://genome.crg.es/astalavista/) for AS event classification (Foissac and Sammeth, 2007). The percentage of alternative splicing genes was estimated using the genome predicted gene models having alternative splicing PUT isoforms among total gene models having at least one mapped PUT. There are a total of 63 241 cDNA sequences generated from 39 475 genes in the recent release of maize gene models (version 3.22, ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/cdna/). Among them 12 627 genes have two or more cDNA sequences, i. e., with isoforms generated by pre-mRNA AS.

1.3 Functional annotation of PUTs

The coding region of each PUT was predicted using the ORFPredictor (Min et al., 2005a) and the full–length transcript coverage was assessed using TargetIdentifier (Min et al., 2005b) as previously described. Functional classification was assigned to the PUTs by performing BLASTX search with an E-value threshold of 1e-5 against UniProtKB/Swiss-Prot. Additionally, predicted protein sequences from ORFPredictor were further annotated using rpsBLAST against the Pfam database (http://pfam.xfam.org/). To assess the 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, the results were summarized in Table 1. Gene Ontology (GO) information was extracted from the UniProt ID mapping table based on the BLASTP of gene model protein sequences against the UniProtKB/Swiss-Prot (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/).The GO categories were further analyzed using GO SlimViewer using plant specific GO terms (McCarthy et al., 2006).

Table 1 Basic features of the assembled putative unique transcripts (PUTs) of maize plant

Total PUTs	614 201
Average length of PUTs (bp)	815
BLASTX matche against UniProt/Swiss-Prot database	247 798
Total ORFs	601 196
Full-length PUTs	128 505
Pfam matches	166 174
PUTs mapped to genome (%)	320 447 (52.2)
PUTs matched to cDNAs of gene models (%)	298 248 (48.6)
PUTs mapped to genome with gene models (%)	206 593 (33.6)
Unique genes supported with matching PUTs (%)	37 751 (95.6)
AS rate of gene models (%)	20 860 (55.3)

1.4 Conserved alternatively spliced genes in cereal plants and visualization of AS

In our previous report, we have identified conserved AS genes among rice, maize, sorghum and *Brachypodium* (Min et al., 2015). In the current work, only maize and rice (ssp *japonica*) conserved AS genes were identified. The reciprocal BLASTP (cutoff E-value 1E-10) was done using the longest ORF of the rice AS isoforms with maize predicted gene model protein sequences for classifying the conserved AS pairs between the species. AS events identified in this study along with the integrated genomic tracks of predicted gene models, as well as data reported previously, are available from Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/) (Walters et al., 2013; VanBuren et al., 2013; Min et al., 2015). BLAST search is also available for searching the

PUTs and AS isoforms. The data analyzed along with the GO and Pfam annotations in the study are publicly available at: http://proteomics.ysu.edu/publication/data/maize2017/. It should be noted that the database also contains AS data from fruit species including pineapple, apple, grape, orange, and strawberry (Wai et al., 2016; Sablok et al., 2017).

2 Results and Discussion

2.1 Transcripts assembling, annotation and mapping to the maize genome

By pooling several sources of maize transcripts assembled from ESTs, mRNAs, and RNA-seq libraries, we obtained a total of 614 201 putative unique transcripts with an average length of 815 bp (Table 1). Compared with our previous assembled maize transcripts (Min et al., 2015), the number of PUTs was increased by ~26%, and the length was also significantly increased from 466 bp to 815 bp. All the assembled PUTs were structurally and functionally annotated including putative open reading frame (ORF) prediction, coding region full-length prediction, a putative function and Pfam prediction. These basic features were summarized in Table 1. A total of 601 196 ORFs were predicted using OrfPredictor with 247 798 of them having a BLASTX hit against the UniProt Swiss-Prot dataset (Min et al., 2015a) and 128 505 PUTs were predicted encoding full-length proteins by TargetIdentifier (Min et al., 2015b). Among the predicted ORFs, 166 174 were annotated with a Pfam match (Table 1).

Using the strict mapping parameters as described in methods, a total of 320 447 PUTs (52.2%) were mapped to maize genome (Table 1). Among the assembled transcripts, a total of 298 248 PUTs matched to cDNA sequences generated from 38 253 unique genes with ≥95% identity of an aligned pair and a minimum of 80 bp of aligned length (Table 1). It should be noted that some PUTs matched to cDNA sequences generated from gene models were not mapped to the genome as strict parameters were applied for mapping using ASFinder (Min, 2013). Among a total of 320 447 PUTs mapped to the genome 206 593 PUTs matched to cDNA sequences generated from a total of 37 751 unique genes. As it was mentioned above a total of 39 475 genes were annotated in the recent release of gene models, thus 95.6% of the gene models were supported by at least one mapped PUT, i.e., transcribed in our assembled data. The mapped PUTs and predicted gene models were also visualized using Generic Genome Browser (GBrowse) (http://gmod.org/wiki/GBrowse).

2.2 Detection and classification of alternative splicing events

ASFinder software was used to identify potential alternatively spliced isoforms based on the SIM4 output of aligning PUTs to the maize genome (Min, 2013; Florea et al., 1998). The AS events were classified using the AStalavista server (Foissac and Sammeth, 2007). A total of 192 624 AS events were detected and classified, including 103 566 (53.8%) basic events and 89 058 (46.2%) complex events which were formed by combination of various types of basic events (Table 2). These AS events were generated from 91 128 PUTs from 26 669 genomic loci. Among 91 128 alternatively spliced PUT isoforms, 81 260 matched to cDNAs of 20 860 gene models. The isoforms not matching a gene models may represent new gene loci or lie in the untranslated regions of known gene models. Similar to our previous studies in maize and other plants (Walter et al., 2013; VanBuren et al., 2013, Min et al., 2015), the IR was the major splicing type among four basic AS types (Table 2). The abundance of IR as a major AS event is consistent with previous reports in maize and other plant species (Min et al., 2015; Wang and Brendel, 2006; Baek et al., 2008; Labadorf et al., 2010; Walters et al., 2013; Thatcher et al., 2014). However, we observed that the proportion of complex events was positively correlated with the average length of assembled transcripts. In this study the average length of the PUTs was 815 bp and the complex AS events was accounted for 46.2%, while in our previous analysis, the average length of the 466 bp and the complex event type was 20.4% (Min et al., 2015). This trend was observed with sorghum AS data (Min et al., 2015). AltA (12.8%) and AltD (9.3%) represent the less abundant observed AS events with AltA showing a slightly higher frequency as compared to AltD (Table 2) (Min et al., 2015). ES (7.5%) was the lowest occurred event in plants, which was in line with the observed results in other studies (Min et al., 2115). Because a large number of transcripts generated using RNA-seq techniques were incorporated in this work, the numbers of AS events in all subtypes were significantly (7-folds) higher than the numbers of AS events previously identified (Table 2).

Table 2 Alternative splicing events in maize

	Previous (%)*	Current (%)
exon skipping	1 568 (5.7)	14 531 (7.5)
alternative donor sites	2 080 (7.6)	17 871 (9.3)
alternative acceptor sites	3 314 (11.4)	24 748 (12.8)
intron retention	11 048 (40.4)	46 416 (24.1)
others (complex events)	5 576 (20.4)	89 058 (46.2)
Total	23 386	192 464

Note: *Previous data from Min et al. (2015).

The percentage of AS genes was estimated based on the proportion of predicted gene models having AS PUT isoforms. As a total of 37 751 gene models have at least on PUT being mapped and among them 2 860 had AS, thus, the rate of AS genes was estimated to be 55.3% in maize. Compared with our previous analysis (Min et al., 2015), the number of genes which were transcribed with alternatively spliced transcripts (AS genes) identified in this study was significantly increased from 10,687 to 20 860, the rate of AS genes was increased from 33.8% to 55.3%. Also the number of AS genes identified in the study is higher than the number reported by Thatcher et al. (2014), which was 15 771, using RNA-seq data. A recent report using the RNA-seq technology revealed that ~61% of multi-exonic genes in *A. thaliana* are alternatively spliced under normal growth conditions (Marquez et al., 2012). The maize AS rate (55.3%) mentioned above was based on all maize gene models with transcript mapping evidence. If we only count gene models having PUTs mapping at least with two exons, there were a total of 31 049 such gene models, thus, the AS rate was 67.2% in maize. We would like to point out that the number of AS events and isoforms in our analysis were higher than the numbers obtained by Mei et al. (2017) as different datasets and assembling approaches were used. However, the AS rate was also reported to be near 60% of expressed multi-exon genes in B73 (Mei et al., 2017).

Recently Yan et al. (2014) developed a database of intron-less genes of Poaceae (PIGD, http://pigd.ahau.edu.cn), which collected 14 623 intron-less maize genes. We compared the list of maize intron-less genes with our mapping data and found 7 152 of them actually had an intron or introns that were directly supported by PUTs mapping (Supplementary Table 1 – file: false_intronless.ids). Thus, the intron-less gene lists collected by Yan et al. (2014) need to be examined thoroughly with gene expression data for other types of analysis. The transcripts mapping to genome information generated in the work can be further used to improve the predicted gene structures in maize.

2.3 Functional classification of AS genes

For simplicity of description below, gene models which have pre-mRNAs generating AS transcript isoforms are referred as AS genes, and gene models having pre-mRNAs with no AS transcripts identified in the current analysis are referred as non-AS genes. To obtain a general picture of AS genes and non-AS genes, Gene Ontology (GO) analyses was performed using the protein sequences of the gene models which had at least one PUT mapped, i. e., they were transcribed and may represent real genes. The predicted protein sequences were used. Thus a total of 37 751 protein sequences were subjected to GO analysis.

Within 37 751 protein sequences 24 061 had GO mapping, and among 20 860 protein sequences of AS genes 15 344 had GO mapping. These mapped GO IDs were further clustered used GOSlimViewer server (http://www.agbase.msstate.edu/cgi-bin/tools/goslimviewer_select.pl). Based on our experiences in analyzing cellular components and protein subcellular location (Lum et al., 2014), GO cellular component analysis based on BLASTP method is not accurate, thus it was not included. We compared the GO classification of biological process and molecular function in AS gene set with the whole set of expressed genes supported with transcript evidence (Table 3; Table 4). AS gene products were involved in all the biological processes with various molecular functions. In average 78.6% and 78.9% of expressed genes had AS with protein products involved in known GO biological processes and molecular functions, respectively (Table 3; Table 4). As the data were collected from pooled ESTs, mRNAs, as well as assembled transcripts from RNA-seq data, it is difficult to make

inferences on the biological significance of the variations of each subcategories of GO. However, numerous detailed experiments have demonstrated the significant biological roles of AS in plant stress responses, growth and development (Reddy et al., 2013; Staiger and Brown, 2013). Identification of these AS genes in maize is the first step in elucidating their biological roles in this plant species.

Table 3 Classification of maize gene products based on Gene Ontology biological processes

GO ID	Total	AS	%	Biological Process
GO:0009987	6 707	5 291	78.9	cellular process
GO:0008152	5 663	4 430	78.2	metabolic process
GO:0009058	3 353	2 575	76.8	biosynthetic process
GO:0006139	2 901	2 290	78.9	nucleobase-containing compound metabolic process
GO:0016043	1 747	1 382	79.1	cellular component organization
GO:0006950	1 608	1 283	79.8	response to stress
GO:0006810	1 461	1 193	81.7	transport
GO:0007275	1 184	910	76.9	multicellular organism development
GO:0006464	1 092	876	80.2	cellular protein modification process
GO:0009056	965	759	78.7	catabolic process
GO:0007154	923	705	76.4	cell communication
GO:0007165	906	698	77.0	signal transduction
GO:0009628	836	675	80.7	response to abiotic stimulus
GO:0019538	787	645	82.0	protein metabolic process
GO:0009719	776	568	73.2	response to endogenous stimulus
GO:0005975	759	584	76.9	carbohydrate metabolic process
GO:0000003	695	535	77.0	reproduction
GO:0006629	659	534	81.0	lipid metabolic process
GO:0009791	598	456	76.3	post-embryonic development
GO:0009605	513	402	78.4	response to external stimulus
GO:0009653	493	398	80.7	anatomical structure morphogenesis
GO:0007049	490	401	81.8	cell cycle
GO:0006259	474	399	84.2	DNA metabolic process
GO:0006412	455	358	78.7	translation
GO:0030154	436	334	76.6	cell differentiation
GO:0009607	373	290	77.7	response to biotic stimulus
GO:0040007	284	222	78.2	growth
GO:0009908	282	217	77.0	flower development
GO:0006091	252	190	75.4	generation of precursor metabolites and energy
GO:0009790	230	183	79.6	embryo development
GO:0015979	188	135	71.8	photosynthesis
GO:0016049	174	138	79.3	cell growth
GO:0008219	172	145	84.3	cell death
GO:0019725	172	132	76.7	cellular homeostasis
GO:0019748	155	106	68.4	secondary metabolic process
GO:0040029	127	111	87.4	regulation of gene expression, epigenetic
GO:0009991	108	85	78.7	response to extracellular stimulus
GO:0009856	72	53	73.6	pollination
GO:0007267	53	45	84.9	cell-cell signaling
GO:0009606	35	30	85.7	tropism
GO:0009875	19	12	63.2	pollen-pistil interaction
GO:0007610	14	13	92.9	behavior
GO:0009835	12	9	75.0	fruit ripening
GO:0009838	12	8	66.7	abscission
Total	39 215	30 805	78.6	

Table 4 Classification of maize gene products based on Gene Ontology molecular functions

	Total	AS	%	Molecular function
GO:0005488	4 541	3 570	78.6	binding
GO:0000166	2 103	1 769	84.1	nucleotide binding
GO:0016740	2 048	1 619	79.1	transferase activity
GO:0016787	1 782	1 443	81.0	hydrolase activity
GO:0003824	1 767	1 364	77.2	catalytic activity
GO:0003677	1 345	1 005	74.7	DNA binding
GO:0003674	1 320	1 006	76.2	molecular_function
GO:0003723	761	625	82.1	RNA binding
GO:0005215	748	581	77.7	transporter activity
GO:0016301	714	588	82.4	kinase activity
GO:0005515	648	532	82.1	protein binding
GO:0003700	627	423	67.5	transcription factor activity, sequence-specific DNA binding
GO:0005198	328	245	74.7	structural molecule activity
GO:0003676	212	176	83.0	nucleic acid binding
GO:0030234	178	133	74.7	enzyme regulator activity
GO:0004518	168	133	79.2	nuclease activity
GO:0008289	157	132	84.1	lipid binding
GO:0030246	134	106	79.1	carbohydrate binding
GO:0004871	129	105	81.4	signal transducer activity
GO:0008135	83	69	83.1	translation factor activity, RNA binding
GO:0004872	77	60	77.9	receptor activity
GO:0003682	64	44	68.8	chromatin binding
GO:0003774	47	39	83.0	motor activity
GO:0005102	42	34	81.0	receptor binding
GO:0019825	4	2	50.0	oxygen binding
GO:0045182	4	3	75.0	translation regulator activity
Total	20 031	15 806	78.9	

2.4 Impact of AS on gene product function

The PUTs were annotated for putative protein coding region by performing a BLASTX search against UniProt/Swiss-Prot database and the ORFs were identified using OrfPredictor webserver (Min et al., 2005a), and the completeness of ORFs were examined using TargetIdentifier (Min et al., 2005b). The protein families of the ORFs of were predicted using rpsBLAST searching Pfam database. 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 events produce NMD candidates (Wang and Brendel, 2006). In our dataset of 192 624 AS isoform pairs, there were 12 146 (6.3%) pairs with one isoform harboring a complete ORF and the other not having an ORF. Lacking an ORF in a transcript could be either due to incompleteness in the PUT sequence or due to loss of a start codon or a premature stop codon. There were also 54 388 (28.2%) pairs having complete ORFs in both isoforms. Thus we further compared if their protein domains were changed or not within the set having complete ORFs.

Within a total of 54 388 AS isoform pairs having complete ORFs, 10 9941 (20.2%) pairs had no Pfam hit, 32 768 (60.2%) pairs had identical Pfam hits, the remaining 10 626 (19.6%) either had one isoform having a Pfam hit and

the other not having a Pfam hit or the pairs had different Pfam categories. Thus, about 19.6% of AS event generated isoforms may have their protein functionalities changed. In pineapple AS analysis it was estimated 24.9% of AS events resulting encoded protein functional changes (Wai et al., 2014). These Pfam loss or changes are most likely caused by the translation frame changes. The biological significance of the change in protein family functional domains in these genes certainly warrants further investigation.

2.5 Conserved alternatively spliced genes

Genes generating AS with biological roles might be conserved during evolution. Previously we have reported conserved AS genes among maize, rice (both *japonica* and *indica*), sorghum, and *B. distachyon* (Min et al., 2015). A total of 8 734 AS genes were identified in *japonica* rice and within them 3 246 were conserved AS genes between rice and maize (Min et al., 2015). As in the current work the number of identified AS genes in maize were increased from 10 687 to 20 860, we re-analyzed the conserved AS gene pairs between these two cereal plants. The conserved AS genes were indeed increased to 4 766 (Figure 1). However, we expect more AS genes in rice as well as more AS genes conserved among cereal or grass plants will be identified if we incorporate more available transcript data generated from RNA-seq experiments in rice and in other plants, as previous transcripts data were assembled using EST and mRNA sequences only (Min et al., 2015).

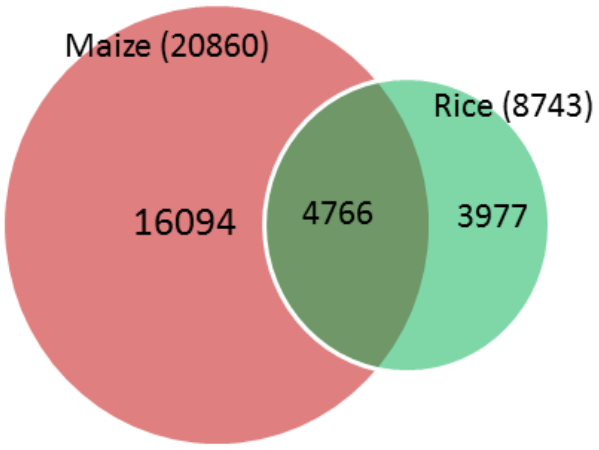

Figure 1 Conserved genes undergoing alternative splicing in maize and rice plants.

3 Conclusion

In this work, we incorporated all available transcripts data including ESTs, mRNAs, and transcripts generated using RNA-seq technology for comprehensively cataloging AS in maize. A total of 192 624 AS events were detected and classified. These AS events were identified from 91 128 transcripts which were generated from 26 669 genomic loci. Of which 20 680 predicted gene models were identified generating mRNAs having AS. Thus about 55.3% maize genes may undergo AS. Based on our work in AS identification in cereal plants as well AS research in other plants by other researchers (Min et al., 2015; Thatcher et al., 2016; Reddy et al., 2013), we believe that AS is common in plant intron-containing genes, thus needs to be considered closely in all research work related to plant gene expression experiments. Systematically identification and cataloging these AS genes in important crop plants and making the AS gene data available to the community would facilitate the crop plant community to better understand the gene regulation in plant growth and development as well as their coping strategies in stress environments.

Acknowledgements

The work was supported by the Youngstown State University Research Professorship award to XJM.

References

Andorf CM, Cannon EK, Portwood JL, et al., 2016, MaizeGDB update: new tools, data and interface for the maize model organism database. Nucleic Acids Res 44(D1):D1195-1201.
https://doi.org/10.1093/nar/gkv1007
PMid:26432828 PMCid:PMC4702771

Campbell MA, Haas BJ, Hamilton JP, et al., 2006 Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 7:327.
https://doi.org/10.1186/1471-2164-7-327
PMid:17194304 PMCid:PMC1769492

Carvalho RF, Feijão CV, Duque P ,2013, On the physiological significance of alternative splicing events in higher plants. Protoplasma 250:639-650.
https://doi.org/10.1007/s00709-012-0448-9
PMid:22961303

Chang CY, Lin WD, Tu SL ,2014, Genome-wide analysis of heat-sensitive alternative splicing in Physcomitrella patens. Plant Physiol. 165:826-840.
https://doi.org/10.1104/pp.113.230540
PMid:24777346 PMCid:PMC4044832

Filichkin SA, Priest HD, Givan SA, et al., 2010, Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. 20:45-58.
https://doi.org/10.1101/gr.093302.109
PMid:19858364 PMCid:PMC2798830

Florea L, Hartzell G, Zhang Z, et al., 1998, A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. 8:967-974.
https://doi.org/10.1101/gr.8.9.967
PMid:9750195 PMCid:PMC310774

Foissac S, Sammeth M, 2007, ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Res.35:W297–299.
https://doi.org/10.1093/nar/gkm311
PMid:17485470 PMCid:PMC1933205

Huang X, Madan A ,1999, CAP3: A DNA sequence assembly program. Genome Res. 9:868-877.
https://doi.org/10.1101/gr.9.9.868
PMid:10508846 PMCid:PMC310812

Lewis BP, Green RE, Brenner SE ,2003, Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA 100:189-192.
https://doi.org/10.1073/pnas.0136770100
PMid:12502788 PMCid:PMC140922

Lum G, Meinken J, Orr J, Frazier S, Min XJ ,2014, PlantSecKB: the plant secretome and subcellular proteome knowledgebase. Computational Molecular Biology. 4(1):1-17 (doi:10.5376/cmb.2014.04.0001)
https://doi.org/10.5376/cmb.2014.04.0001

Mao H, Sun S, Yao J, et al. ,2010, Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. Proc Natl Acad Sci USA. 107:19579-19584.
https://doi.org/10.1073/pnas.1014419107
PMid:20974950 PMCid:PMC2984220

Marquez Y, Brown JW, Simpson C, et al., 2012 Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome research 22:1184-1195.
https://doi.org/10.1101/gr.134106.111
PMid:22391557 PMCid:PMC3371709

McCarthy FM, Wang N, Magee GB, et al., 2006, AgBase: a functional genomics resource for agriculture. BMC Genomics 7:229.
https://doi.org/10.1186/1471-2164-7-229
PMid:16961921 PMCid:PMC1618847

Mei W, Liu S, Schnable JC, et al.,2017, A comprehensive analysis of alternative splicing in paleopolyploid maize. Frontiers Plant Sci. 8:694.
https://doi.org/10.3389/fpls.2017.00694
PMid:28539927 PMCid:PMC5423905

Min XJ, Powell B, Braessler J, et al. ,2015, Genome-wide cataloging and analysis of alternatively spliced genes in cereal crops. BMC Genomics. 16:721.
https://doi.org/10.1186/s12864-015-1914-5
PMid:26391769 PMCid:PMC4578763

Min XJ ,2013, ASFinder: a tool for genome-wide identification of alternatively spliced transcripts from EST-derived sequences. International J Bioinformatics Res Appl 9:221-226.
https://doi.org/10.1504/IJBRA.2013.053603
PMid:23649736

Min XJ, Butler G, Storms R, et al. ,2005a, OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. 33:W677-680. https://doi.org/10.1093/nar/gki394
PMid:15980561 PMCid:PMC1160155

Min XJ, Butler G, Storms R, et al., 2005b, TargetIdentifier: a web server for identifying full-length cDNAs from EST sequences. Nucleic Acids Res. 33:W669-W672.
https://doi.org/10.1093/nar/gki436
PMid:15980559 PMCid:PMC1160197

Morello L, Breviario D ,2008, Plant spliceosomal introns: not only cut and paste. Curr Genet 9:227–238
https://doi.org/10.2174/138920208784533629
PMid:19452040 PMCid:PMC2682935

Panahi B, Abbaszadeh B, Taghizadeghan M, et al., 2014, Genome-wide survey of alternative splicing in Sorghum bicolor. Physiol Mol Biol Plants 20:323-329.
https://doi.org/10.1007/s12298-014-0245-3
PMid:25049459 PMCid:PMC4101146

Reddy AS, Marquez Y, Kalyna M, et al., 2013, Complexity of the alternative splicing landscape in plants. Plant Cell 25:3657-3683.
https://doi.org/10.1105/tpc.113.117523
PMid:24179125 PMCid:PMC3877793

Rice P, Longden I, Bleasby A ,2000, EMBOSS: The European Molecular Biology Open Software Suite. Trends Genetics 16:276-277.
https://doi.org/10.1016/S0168-9525(00)02024-2

Roberts GC, Smith CW ,2002, Alternative splicing: combinatorial output from the genome. Curr Opin Chem Biol. 6:375-83.
https://doi.org/10.1016/S1367-5931(02)00320-4

Sablok G, Gupta PK, Baek JM, et al., 2011, Genome-wide survey of alternative splicing in the grass Brachypodium distachyon: an emerging model biosystem for plant functional genomics. Biotechnology Letters 33:629-636.
https://doi.org/10.1007/s10529-010-0475-6
PMid:21107652

Sablok G, Powell B, Braessler J, Yu F, Min XJ, 2017, Comparative landscape of alternative splicing in fruit plants. Current Plant Biology. DOI: 10.1016/j.cpb.2017.06.001.
https://doi.org/10.1016/j.cpb.2017.06.001

Schnable PS, Ware D, Fulton RS, et al., 2009, The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1125.
https://doi.org/10.1126/science.1178534
PMid:19965430

Staiger D, Brown JW, 2013, Alternative splicing at the intersection of biological timing, development, and stress responses. Plant Cell 25:3640-3656.
https://doi.org/10.1105/tpc.113.113803
PMid:24179132 PMCid:PMC3877812

Stamm S, Ben-Ari S, Rafalska I, et al., 2005, Function of alternative splicing. Gene 344:1-20.
https://doi.org/10.1016/j.gene.2004.10.022
PMid:15656968

Syed NH, Kalyna M, Marquez Y, et al., 2012, Alternative splicing in plants - coming of age. Trends Plant Sci. 17:616-623.
https://doi.org/10.1016/j.tplants.2012.06.001
PMid:22743067 PMCid:PMC3466422

Thatcher SR, Danilevskaya ON, Meng X, et al., 2016, Genome-wide analysis of alternative splicing during development and drought stress in maize. Plant Physiology 170:586-599.
https://doi.org/10.1104/pp.15.01267
PMid:26582726 PMCid:PMC4704579

Thatcher SR, Zhou W, Leonard A, et al., 2014, Genome-wide analysis of alternative splicing in Zea mays: landscape and genetic regulation. Plant Cell 26:3472-3487.
https://doi.org/10.1105/tpc.114.130773
PMid:25248552 PMCid:PMC4213170

VanBuren R, Walters B, Ming R, et al., 2013, Analysis of expressed sequence tags and alternative splicing genes in sacred lotus (Nelumbo nucifera Gaertn.). Plant Omics J. 6:311-317.

Vitulo N, Forcato C, Carpinelli EC, 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 Boil 14:99.
https://doi.org/10.1186/1471-2229-14-99
PMid:24739459 PMCid:PMC4108029

Walters B, Lum G, Sablok G, et al., 2013, Genome-wide landscape of alternative splicing events in Brachypodium distachyon. DNA Res 20:163-171.
https://doi.org/10.1093/dnares/dss041
PMid:23297300 PMCid:PMC3628446

Wai CM, Powell B, Ming R, et al., 2016 Analysis of alternative splicing landscape in pineapple (Ananas comosus). Tropical Plant Biology 9:150-160. https://doi.org/10.1007/s12042-016-9168-1

Wang B, Brendel V, 2006, Genome wide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA 103:7175-7180. https://doi.org/10.1073/pnas.0602039103 PMid:16632598 PMCid:PMC1459036

Yan H, Jiang C, Li X, et al., 2014, PIGD: a database for intronless genes in the Poaceae. BMC Genomics 15:1. https://doi.org/10.1186/1471-2164-15-832 PMid:25270086 PMCid:PMC4195894

Yang X, Zhang H,Li L, 2012, Alternative mRNA processing increases the complexity of microRNA-based gene regulation in Arabidopsis. Plant J. 70:421-431. https://doi.org/10.1111/j.1365-313X.2011.04882.x PMid:22247970

Yu H, Tian C, Yu Y, et al., 2016, Transcriptome survey of the contribution of alternative splicing to proteome diversity in Arabidopsis thaliana. Molecular Plant 9:749-752. https://doi.org/10.1016/j.molp.2015.12.018 PMid:26742955

Zhang PG, Huang SZ, Pin AL, et al., 2010, Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of Arabidopsis. Mol Biol Evol 27:1686-1697. https://doi.org/10.1093/molbev/msq054 PMid:20185454