Computational Molecular Biology, 2018, Vol.8, No.1, 1-13
http://cmb.biopublisher.ca

Research Article **Open Access**

A Survey of Alternative Splicing in Allotetraploid Cotton (*Gossypium hirsutum* L.)

Xiangjia Min
Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA
Corresponding author email: xmin@ysu.edu

Computational Molecular Biology, 2018, Vol.8, No.1 doi: 10.5376/cmb.2018.08.0001
Received: 10 Apr., 2018
Accepted: 23 May, 2018
Published: 27 Jul., 2018

Preferred citation for this article:
Min X.J., 2018, A survey of alternative splicing in allotetraploid cotton (*Gossypium hirsutum* L.), Computational Molecular Biology, 8(1): 1-13 (doi: 10.5376/cmb.2018.08.0001)

Abstract Allotetraploid cotton (*Gossypium hirsutum* L.), accounting for more than 90% of cultivated cotton worldwide, provides textile fibers and seeds. Alternative splicing (AS) is a post-transcriptional process that generates more than one RNA isoforms from a single pre-mRNA transcript, increasing the diversity of functional proteins and RNAs. We surveyed the alternatively spliced genes in cotton using expressed sequence tag (EST) and mRNA sequences available in the public databases. A total of 56,080 AS events, including 41,150 (73.4%) basic events and 14,930 (26.6%) complex events were identified, which were generated from approximately 23,930 genes. Intron retention was the most frequent event, accounting for 34.8%, followed by alternative acceptor site events (18.8%) and alternative donor site events (11.8%), and exon skipping being the least frequent event (8.0%). Complex types, which are formed by more than one basic event, are accounted for 26.6%. The estimated AS rates of genes generating AS isoforms was 27.1% in cotton. Gene Ontology and protein family analysis showed that the products of alternatively spliced genes were involved in many biological processes with diverse molecular functions. The transcripts to cotton genome mapping information can be used to improve the predicted gene models in cotton. The annotation information of AS isoforms of these genes provides a basis for future investigation on the functions of these AS genes in cotton biology. The data can be accessed at Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/).

Keywords Alternative splicing; Cotton; Gene expression; *Gossypium hirsutum*; mRNA; Plant

Background

The most widely cultivated upland cotton (*Gossypium hirsutum* L.) is an allotetraploid species (AtAtDtDt), consisting of both A sub-genomes and D sub-genomes (Lubbers and Chee, 2009; Li et al., 2015; Zhang et al., 2015). *G. hirsutum* accounts for more than 90% of commercial cotton production worldwide and is the main sources of renewable textile fibers and seeds (Wendel and Grover, 2015). The genomes of the two extant progenitor relatives, *G. arboreum* (AA) and *G. rainondii* (DD), and *G. hirsutum* have been sequenced (Wang et al., 2012; Li et al., 2014a; Li et al., 2015; Zhang et al., 2015). Sequencing these genomes provides insights on the genome evolution, gene contents, regulatory elements, genomic signatures of selection and domestication in these species. The genome sequences of *G. hirsutum* (acc. TM-1) have been reported independently by two teams with 66,434 and 76,943 genes annotated from the assembled genomes, respectively (Li et al., 2015; Zhang et al., 2015).

Plant gene expression is a tightly controlled process in regulating growth and development as well as in response to changing environments. In addition to alternative transcription initiation site and polyadenylation site that generate different transcript isoforms, alternative splicing (AS) is a common process in plants that generates two or more transcript isoforms from one pre-mRNA sequence (Reddy et al., 2013). Thus, the diversities of mRNAs and proteins in the organism are significantly increased by AS. There are already well documented experimental data showing AS plays critical roles in many biological processes in plants such as photosynthesis, defense responses, flowering timing, grain quality, and responses to stresses (Reddy et al., 2013; Staiger 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 found in transcript isoforms by combination of basic events (Sablok et al., 2011). AS isoforms may encode a distinct functional protein or become

non-functional due to harboring a premature termination codon in protein coding regions. The nonfunctional isoforms are degraded by a process known as nonsense-mediated decay (NMD) (Lewis et al., 2003).

Arabidopsis thaliana, a model plant species, has been intensively investigated and were reported with ~60-70% of multi-exon genes undergoing AS (Filichkin et al., 2010; Zhang et al., 2010; Marquez et al., 2012; Syed et al., 2012; Carvalho et al., 2013; Yu et al., 2016; Zhang et al., 2017). AS in other plant species also has been examined including *Oryza sativa* (rice) (Wang and Brendel, 2006; Min et al., 2015; Wei et al., 2017; Kater et al., 2018), *Nelumbo nucifera* (sacred lotus) (VanBuren et al., 2013), *Vitis vinifera* (grape) (Vitulo et al., 2014; Sablok et al., 2017), *Brachypodium distachyon* (Sablok et al., 2011; Walters et al., 2013), *Zea mays* (maize) (Thatcher et al., 2014; Min et al., 2015; Thatcher et al., 2016; Mei et al., 2017; Min, 2017), and *Sorghum bicolor* (sorghum) (Panahi et al., 2014; Min et al., 2015; Abdel-Ghany et al., 2016), etc. 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). IR has been shown to be the most frequent AS event in plants with AS rates in the intron containing genes ranged from ~30% to > 60% depending on available transcriptome data (Sablok et al., 2011; Reddy et al., 2013; Sablok et al., 2017). Genome-wide conserved alternatively spliced genes among different plant species have been identified in cereal plants and fruit plants (Min et al., 2015; Sablok et al., 2017). Further, genome-wide conserved AS events across a wide range of plant species such as in flowering plant species as well as in monocot species have also been analyzed (Chamala et al., 2015; Mei et al., 2017). These works lay the foundation for identifying and studying conserved AS genes as well as conserved AS events across evolutionally related plant species (Min et al., 2015; Mei et al., 2017).

There were only three reports related to genome-wide AS analysis in cotton so far. Using RNA-sequencing (RNA-seq) data from *G. raimondii*, 16,437 AS events in 10,197 genes were identified (Li et al., 2014b). Similar RNA-seq analysis identified 14,172 AS events in 6,797 genes *G. davidsonii* growing under salt stress conditions (Zhu et al., 2018). Most recently, Wang et al. (2018) reported that using Pacific Biosciences single molecule long-read isoform sequencing (Iso-Seq) identified 176,849 full-length transcript isoforms, detected a total of 133,229 AS events, from 27,229 gene loci, with 15,102 fiber-specific AS events in *G. barbadense*, an allotetraploid cotton species. In all three reports, the prevalent type of AS events was retained introns. In this work, we report a survey of AS events using currently available expressed sequence tags (ESTs) and mRNA sequences with an aim to generate a preliminary catalog of alternatively spliced genes in the cultivated upland cotton species, *G. hirsutum*.

1 Materials and Methods

1.1 Sequence datasets and sequence assembly

Two draft genome sequences of allotetraploid cotton (*G. hirsutum* L. acc. TM-1) have been generated independently (Li et al., 2015; Zhang et al., 2015). In this work we used the genome sequences (assembly ASM98774v1) generated by Li et al. (2015) as they were available for downloading from the National Center for Biotechnology Information (NCBI) genome database (https://www.ncbi.nlm.nih.gov/genome/?term=cotton). We also downloaded a total of 432,161 nucleotide sequences of *G. hirsutum* including 94,350 mRNA sequences and 337,811 EST sequences. For simplicity of description the term “cotton” only means *G. hirsutum* in the context, otherwise, full species names were specified.

1.2 Transcripts assembly, mapping to genome, and identification AS events

The EST and mRNA sequences were processed to remove contaminants, vector and repetitive sequences using a procedure we implemented previously (Min et al., 2015). The procedure was briefly outlined below: EMBOSS trimmest tool was used to trim the polyA or polyT end (Rice et al., 2000); then trimmed ESTs and mRNA sequences were used to search against UniVec and *E. coli* database using BLASTN for removal of vector and *E. coli* contaminants; finally BLASTN searches against the plant repeat database which was built with TIGR gramineae repeat data, sorghum, maize, and rice repeat data (available from ftp://ftp.plantbiology.msu.edu/pub/data/TIGR_Plant_Repeats/). A total of 430,541 cleaned EST and mRNA sequences were assembled using CAP3 with the following parameters: -p 95 -o 50 -y 20 (Huang and Madan, 1999). A total of 279,050 putative unique

transcripts (PUTs) including 28,316 contigs and 250,734 singlets were obtained for mapping to the genome sequences.

The assembled PUTs were mapped to their corresponding chromosomes using ASFinder (http://proteomics.ysu.edu/tools/ASFinder.html/) (Min, 2013). We applied the threshold values: a minimum of 95% identity, a minimum of 80 bp aligned length, and > 75% of a PUT sequence aligned to the genome (Walters et al., 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 and have variable exon-intron boundaries as AS isoforms. To avoid chimeric PUT assemblies, mapped PUTs having an intron size > 100 kb were removed for AS identification. The output file (AS. gtf) from ASFinder was submitted to AStalavista server (http://genome.crg.es/astalavista/) for AS event classification (Foissac and Sammeth, 2007). The rate of alternative splicing genes was estimated as the ratio of genomic loci having alternative splicing PUT isoforms over total genomic loci having at least one mapped PUT sequence.

1.3 Functional annotation of PUTs and data availability

The PUT sequences were functionally annotated, including prediction protein coding region and domain search. 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). Functional classification was based on the BLASTX search with an E-value threshold of 1e-5 against UniProtKB/Swiss-Prot. In addition, predicted protein sequences from ORFPredictor were further annotated for functional domains using rpsBLAST against the PFam database (http://pfam.xfam.org/). The assembled PUTs were further compared with transcripts of predicted gene models using BLASTN with a cut off E-value of 1e-10, ≥ 95% identity and a minimum aligned length of 80 bp. Gene Ontology (GO) information was extracted from the UniProt ID mapping table based on the BLASTX search of PUTs sequences against the UniProtKB/Swiss-Prot (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/). The GO categories were further analyzed using GO Slim Viewer using plant specific GO terms (http://www.agbase.msstate.edu/cgi-bin/tools/goslimviewer_select.pl) (McCarthy et al., 2006).

1.4 Availability of data

The assembled PUTs and AS events identified in this study along with the predicted gene models, as well as data reported previously in our group, are available from Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice/) (VanBuren et al., 2013; Walters et al., 2013; Min et al., 2015; Wai et al., 2016; Min, 2017; Sablok et al., 2017). BLAST search is also available for searching the PUTs and AS isoforms. The datasets supporting the conclusions of this article including the data used for database construction and the supplementary data are publicly available at: http://bioinformatics.ysu.edu/publication/data/Cotton/.

2 Results and Discussion

2.1 Transcripts assembling and annotation

After removing contaminant and low complexity sequences of the combined ESTs and mRNAs of cotton (*G. hirsutum* L.), we used CAP3 program to assemble the cleaned data. A total of 279,050 putative unique transcripts (PUTs) including 28,316 contigs and 250,734 singlets were obtained for further annotation and mapping to the genome sequences. The PUTs had a length ranged from 100 bp to 20,499 bp and an average length of 975 bp (Table 1). All PUTs were structurally and functionally annotated including ORF prediction, coding region completeness assessment, a putative function and PFam prediction. These basic features were summarized in Table 1. A total of 278,650 ORFs were predicted using OrfPredictor including 201,924 of them were predicted using the frame values obtained from BLASTX search against the UniProt Swiss-Prot dataset and 72,726 ORFs were predicted based the intrinsic sequence signals in the sequences (Min et al., 2005a). Among them 128,505 PUTs were predicted encoding full-length proteins by TargetIdentifier (Min et al., 2005b). Among the predicted ORFs, 166,174 were annotated with a protein family (Pfam) match (Table 1). Further, using BLASTN search with a cutoff of 95% identity 247,871 (88.8%) PUTs matched with predicted mRNA sequences of predicted protein coding gene models (Li et al., 2015).

2.2 Mapping transcripts to cotton genome

We used relatively strict mapping parameters to map PUTs to the genome as described in the method section. The identity threshold of 95% prevented PUTs mismatching to the genome segments with lower similarities due to ancient genome or gene duplications. In other hand, it allowed accurate mapping by tolerating errors in PUT sequences that might be generated in original ESTs or in the assembling process. In addition, there might be sequence errors in the assembled genome sequences or variations in different varieties or ecotypes of the same species. We have used the same procedure of PUTs mapping to the genomes in other plant species including cereal plants and fruit plants (Min et al., 2015; Min, 2017; Sablok et al., 2017). A total of 196,098 (70.3% of the total assembled PUTs) PUTs were mapped to *G. hirsutum* genome, including 113,180 PUTs were mapped to a single genomic locus and 82,918 PUTs were mapped to two or more genomic loci (Table 2). The reason for the relative larger number of PUTs (42.3% of mapped PUTs) having more than one mapping loci was apparently due to *G. hirsutum* AtDt genome consisting of both A subgenome and D subgenome (Li et al., 2015; Zhang et al., 2015). Diploid genomes of *G. arboretum* (AA) and *G. raimondii* (DD), which were diverged about 5-10 million years ago (MYA), have been sequenced (Wang et al., 2012; Li et al., 2014a). Our analysis show that the homologues mRNA sequences of diploid *G. arboretum* (AA) and *G. raimondii* (DD), still share 97-100% identity.

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

Total PUTs	Average length of PUTs (bp)	BLASTX matches against Swiss-Prot data (%)	Total ORFs (%)	Full-length PUTs (%)	Pfam matches (%)	PUTs match with predicted gene models (%)
279050	975	201924 (72.4)	278650 (99.9)	115043 (41.2)	155446 (55.7)	247871 (88.8)

Table 2 Mapping of putative unique transcripts (PUTs) to cotton genome

PUTs mapped to genome (% of total PUTs)	PUTs mapped to single locus (% of mapped PUTs)	PUTs mapped to two or more loci (% of mapped PUTs)	Total genomic loci with mapped PUTs	Genomics loci with alternative splicing (AS)	AS rate (%)
196098 (70.3)	113180 (57.7)	82918 (42.3)	88420	23930	27.1

The PUTs were mapped to a total of 88,420 genomic loci (Table 2). This number was higher than the number of genes reported by the genome sequencing projects, as 76,943 genes were reported by Li et al. (2015) and 66,434 genes were reported by Zhang et al. (2015). The mapped PUTs that were located in the regions outside of the predicted genes may contain genes remained to be annotated.

It should be noted that there were 29.7% of the PUTs not being mapped to the draft genome sequences (Table 2). The reasons for these PUTs not being mapped may include incompleteness of the genome sequences and possible errors in the PUTs or genomic sequences including sequencing errors and misassembling. However, these unmapped PUTs were annotated and available from our database, the information might be useful for identifying new genes from cotton species.

2.3 Detection and classification of alternative splicing events

The PUTs to genome mapping gtf (gene transfer format) file generated by ASFinder was submitted to the AStalavista server for identification and classification of AS events (Foissac and Sammeth, 2007; Min, 2013). A total of 56,080 AS events were detected and classified, including 41,150 (73.4%) basic events and 14,930 (26.6%) complex events which had more than one basic event (Figure 1). These AS events were generated from 23,930 genomic loci (clusters) with 44,239 unique transcripts (Table 2). As a total of 88,150 genomic loci have at least one PUT mapped, thus, the estimated AS rates of genes generating AS isoforms (AS genes) was 27.1% in cotton (Table 2). However, based on the PUTs mapping data, there were 25,427 genomic loci having PUTs not having an intron. Thus, only considering the genomic loci mapped with PUTs having an introns or introns, the AS rate was 40.0% in this dataset in cotton. The AS rate in cotton is lower than the rate in Arabidopsis (~60%) and in maize (55%) reported previously (Marquez et al., 2012; Mei et al., 2017; Min, 2017), this apparently due to relative lower number of available EST and mRNA sequences used in current analysis. Recently, RNA-seq analysis in *G. raimondii* and *G. davidsonii* revealed 31.6% and 32.0% AS rates, respectively, in intron-containing genes (Li et

al., 2014b; Zhu et al., 2018). Thus, more cotton genes undergoing AS are expected to be identified when more gene expression data are available.

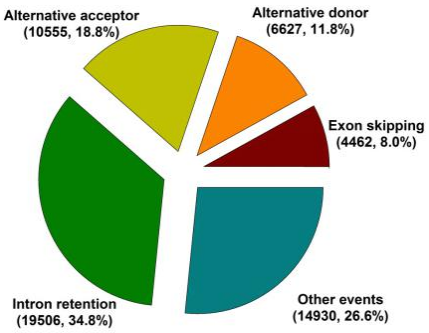

Figure 1 Landscape of alternative splicing events in cotton

Among the AS events, IR (34.8%) was the prevalent type, followed by AltA (18.8%) and AltD (11.8%), with ES (8.0%) as the least type of AS events (Figure 1). Though there were some variations in distributions of AS types, this pattern of AS events was consistent in all cotton species (Li et al., 2014; Zhu et al., 2018) as well as in other plant species so far we have investigated as well as plant species examined by others including Arabidopsis, *Brachypodium distachyon*, cereal plants, and fruit plants (Wang and Brendel, 2006; Baek et al., 2008; Labadorf et al., 2010; Sablok et al., 2012; VanBuren et al., 2013; Walter et al., 2013; Thatcher et al., 2014; Min et al., 2015; Sablok et al., 2017). We also observed that the proportion of complex events varied in different plant species or different analysis of the same species, and the ratio was positively correlated with the average length of assembled transcripts (Min et al., 2015; Min, 2017; Sablok et al., 2017).

One interesting finding of the role of transposons played during AS in plant species was reported recently by Li et al. (2014). Transposable elements (TEs) were found in only 2.9% of all introns, however, 43% of the retained introns were found to have TEs in the AS transcript isoforms. Such an enrichment of TEs in the retained introns in the AS isoforms suggested TE-insertion may play an important role during AS (Li et al., 2014b). In our datasets we retrieved 12,774 retained introns with a length > 30 bp and found only 263 TEs, about 2.1% having TEs in retained introns. Such a discrepancy of TEs in the retained introns of AS isoforms might be resulted by the data processing procedure because in our data cleaning steps, for avoiding misassembling, we purposely removed plant repetitive DNA elements including TEs from the ESTs and mRNA sequences prior to assembling PUTs.

It should be noted that the mapping of PUTs in this work used a cutoff of 95% sequence identity for the aligned regions, this cutoff value could not distinguish homoeologous genes between two subgenomes or homologous genes from recent gene duplications. Full-length mRNA sequences including both 5'-and 3' untranslated regions (UTRs) with strict sequence identity, i.e., 100%, should be able to distinguish transcript isoforms of AS generated from two subgenomes. Recent work using single molecule long-read isoform sequencing (Iso-Seq) identified full-length transcript isoforms and was able to distinguish isoforms from two subgenomes in *G. barbadense*, an allotetraploid cotton species (Wang et al., 2018). It was estimated that ~51.4% of homoeologous genes produced divergent isoforms in each subgenome (Wang et al., 2018).

2.4 Functional classification of PUTs and AS genes

All PUTs including both mapped and unmapped PUTs were annotated functionally as described in section 3.1. To simplify description, predicted gene models having AS transcript isoforms are referred as AS genes, and gene models not having AS transcripts in the current analysis are referred as non-AS genes. To obtain a general picture of protein family distribution in AS genes and non-AS genes, the predicted protein sequences of the PUTs were used to search the Pfam database. For genomic loci having more than one isoform PUT, only one Pfam annotation was selected from each genomic locus. A total of 57,900 Pfam matches from a total of 3,505 protein families were obtained from encoded proteins of a total of 88,420 loci. Among 23,930 genomic loci having AS isoforms, 18,218

genes encoded proteins had Pfam matches to a total of 2,454 protein families. The top protein families in the whole cotton proteome and proteins encoded by genes undergoing AS were listed in Table 3. Among the protein families, many of them were found having AS isoforms in other plant species including cereal plants and fruit plants (Min et al., 2015; Wai et al., 2016; Sablok et al., 2017). These protein families include Pkinase (protein kinase domain), RRM_1 (RNA recoginition motif), Pkinase_Tyr (protein tyrosine kinase), P450 (cytochrome P450), Ras family, UQ_con (uniquitin-conjugating enzyme), etc., suggesting an evolutionally systematic conservation of AS in plant species (Min et al., 2015; Sablok, 2017). We noticed that among 100 genomic loci encoded cellulose synthase (Pfam03552) 39 of them had alternative splicing. In considering the important role played by this enzyme in fiber formation, the functional significance of AS of these genes is warranted for further examination.

Genes undergoing AS during post-transcriptional process produce functional isoforms or non-functional isoforms. We evaluated the impact of AS on the functionalities of the gene products by comparing their Pfam annotation. Among a total of 50,680 isoform pairs generating AS events, 14,214 (25.3%) pairs had no Pfam hit, 30,202 (53.9%) isoform pairs had identical Pfam, 9,046 (16.1%) pairs had one isoform with a Pfam hit and the other not having a Pfam hit, indicating the functional loss of gene products, and 2,708 (4.8%) pairs had different Pfam hits (Supplementary Table 2). Thus, about 20.9% of AS events generated isoforms with functional loss or change. Similar results were obtained in our previous analysis with pineapple and maize data (Wai et al., 2016; Min et al., 2017). The Pfam loss or change in the gene products is most likely caused by the translation frame changes in AS isoforms. The MADS-box genes were alternatively spliced in cotton and some of the alternatively spliced isoforms potentially encoded proteins with altered K-domain and/or C-terminal regions (Lightfoot et al., 2008). The genes were expressed in developing fiber cells suggesting a role in cotton fiber biosynthesis. The biological significance of the change in protein family functional domains in these genes certainly is interesting for further investigation.

2.5 Gene Ontology (GO) analysis of gene products

GO categories provide an overview of the gene products involved in the biological processes, molecular functions, and cellular components. As GO annotation is fairly complex with variable available information for different gene products, thus the analysis is not intended for an accurate quantification but rather providing a broad picture of the functionalities of the gene products. Among the whole set of 27,9031 cotton PUTs sequences a total of 201,924 (72.3%) PUTs had a BLASTX hit (E-value < 1e-5) against the Swiss-Prot database. Then using the Swiss-Prot identifiers we retrieved a total of 1,324,154 GO identifiers. These GO identifiers were further grouped into top categories using GO Slim Viewer (McCarthy et al., 2006). The isoforms from AS genes were also analyzed using the same procedure and a total of 234,362 GO identifiers were obtained. Our previous analysis showed that GO cellular component analysis based on BLASTX method was not accurate, thus we only summarized the GO classification of biological process and molecular function in the whole set of PUTs and isoforms generated by AS genes (Table 4; Table 5). The top categories of molecular functions include binding, catalytic activity, nucleotide binding, transferase activity, hydrolase activity, protein binding etc. (Table 4). These top categories of molecular functions of gene products are more or less similar in all plant species we have examined (Min et al., 2015). The top categories of biological processes include cellular process, metabolic process, biosynthetic process, nucleobase-containing compound metabolic process, response to stress, etc. (Table 5). As expected, the distribution patterns of these processes were also similar in all the plant datasets we analyzed (Min et al., 2015).

GO analysis showed AS gene products were involved in all the biological processes with various molecular functions. In average 46.3% in GO molecular functions and 47.1% in GO biological processes were obtained from the gene products of the AS genes. There are well characterized genes undergoing AS with demonstrated functional significance in regulation of plant growth, development, as well as stress responses (Reddy et al., 2013; Staiger and Brown, 2013). Therefore, the biological roles of AS genes in cotton growth and development need to be examined further.

Table 3 Protein family distributions in the proteins encoded by all genes and by genes with pre-mRNA alternative splicing (AS genes) in cotton

PFam ID	Total	AS	%	Pfam	Pfam description
pfam00078	3419	470	13.7	RVT_1	Reverse transcriptase
pfam00069	1538	590	38.4	Pkinase	Protein kinase domain
pfam07727	1201	142	11.8	RVT_2	Reverse transcriptase
pfam08284	1191	43	3.6	RVP_2	Retroviral aspartyl protease
pfam03732	1049	220	21.0	Retrotrans_gag	Retrotransposon gag protein
pfam07714	1014	326	32.1	Pkinase_Tyr	Protein tyrosine kinase
pfam13041	667	191	28.6	PPR_2	PPR repeat family
pfam00067	655	108	16.5	p450	Cytochrome P450
pfam00931	531	108	20.3	NB-ARC	NB-ARC domain
pfam13639	520	174	33.5	zf-RING_2	Ring finger domain
pfam14223	485	42	8.7	UBN2	gag-polypeptide of LTR copia-type
pfam13456	445	62	13.9	RVT_3	Reverse transcriptase-like
pfam00249	428	127	29.7	Myb_DNA-binding	Myb-like DNA-binding domain
pfam00076	410	233	56.8	RRM_1	RNA recognition motif
pfam13976	398	13	3.3	gag_pre-integrs	GAG-pre-integrase domain
pfam00847	354	40	11.3	AP2	AP2 domain
pfam12776	292	143	49.0	Myb_DNA-bind_3	Myb/SANT-like DNA-binding domain
pfam13359	254	33	13.0	DDE_Tnp_4	DDE superfamily endonuclease
pfam03171	251	70	27.9	2OG-FeII_Oxy	2OG-Fe (II) oxygenase superfamily
pfam00201	240	27	11.3	UDPGT	UDP-glucoronosyl and UDP-glucosyl transferase
pfam00481	226	108	47.8	PP2C	Protein phosphatase 2C
pfam02365	221	86	38.9	NAM	No apical meristem (NAM) protein
pfam00071	216	75	34.7	Ras	Ras family
pfam03106	200	74	37.0	WRKY	WRKY DNA-binding domain
pfam00010	193	77	39.9	HLH	Helix-loop-helix DNA-binding domain
pfam10536	190	17	8.9	PMD	Plant mobile domain
pfam00153	187	66	35.3	Mito_carr	Mitochondrial carrier protein
pfam13966	186	52	28.0	zf-RVT	zinc-binding in reverse transcriptase
pfam13499	186	48	25.8	EF-hand_7	EF-hand domain pair
pfam02519	185	24	13.0	Auxin_inducible	Auxin responsive protein
pfam00141	178	47	26.4	peroxidase	Peroxidase
pfam12796	175	54	30.9	Ank_2	Ankyrin repeats (3 copies)
pfam14432	175	32	18.3	DYW_deaminase	DYW family of nucleic acid deaminases
pfam00106	171	49	28.7	adh_short	short chain dehydrogenase
pfam00004	168	56	33.3	AAA	ATPase family associated with various cellular
pfam00083	164	66	40.2	Sugar_tr	Sugar (and other) transporter
pfam02458	162	15	9.3	Transferase	Transferase family
pfam00665	160	25	15.6	rve	Integrase core domain
pfam00225	159	75	47.2	Kinesin	Kinesin motor domain
pfam03004	158	80	50.6	Transposase_24	Plant transposase (Ptta/En/Spm family)
pfam14244	155	72	46.5	UBN2_3	gag-polypeptide of LTR copia-type
pfam01095	155	20	12.9	Pectinesterase	Pectinesterase
pfam00657	150	30	20.0	Lipase_GDSL	GDSL-like Lipase/Acylhydrolase

Continued Table 3

PFam ID	Total	AS	%	Pfam	Pfam description
pfam00270	149	70	47.0	DEAD	DEAD/DEAH box helicase
pfam03514	147	46	31.3	GRAS	GRAS domain family
pfam00561	147	45	30.6	Abhydrolase_1	alpha/beta hydrolase fold
pfam00226	146	58	39.7	DnaJ	DnaJ domain
pfam00179	143	85	59.4	UQ_con	Ubiquitin-conjugating enzyme
pfam00854	139	44	31.7	PTR2	POT family
pfam13855	137	29	21.2	LRR_8	Leucine rich repeat
pfam04564	136	35	25.7	U-box	U-box domain
pfam01582	135	44	32.6	TIR	TIR domain
pfam01554	129	39	30.2	MatE	MatE
pfam00168	127	42	33.1	C2	C2 domain
pfam00295	127	30	23.6	Glyco_hydro_28	Glycosyl hydrolases family 28
pfam13839	124	29	23.4	PC-Esterase	GDSL/SGNH-like Acyl-Esterase family found
pfam00046	121	37	30.6	Homeobox	Homeobox domain
pfam01490	121	37	30.6	Aa_trans	Transmembrane amino acid transporter protein
pfam00005	120	37	30.8	ABC_tran	ABC transporter
pfam05699	119	53	44.5	Dimer_Tnp_hAT	hAT family C-terminal dimerization
pfam04043	119	7	5.9	PMEI	Plant invertase/pectin methylesterase inhibitor
pfam00501	117	31	26.5	AMP-binding	AMP-binding enzyme
pfam00332	117	27	23.1	Glyco_hydro_17	Glycosyl hydrolases family 17
pfam10250	112	49	43.8	O-FucT	GDP-fucose protein O-fucosyltransferase
pfam00400	111	45	40.5	WD40	WD domain
pfam02362	110	47	42.7	B3	B3 DNA binding domain
pfam03195	110	19	17.3	DUF260	Protein of unknown function DUF260
pfam00085	107	33	30.8	Thioredoxin	Thioredoxin
pfam00230	106	35	33.0	MIP	Major intrinsic protein
pfam01501	106	28	26.4	Glyco_transf_8	Glycosyl transferase family 8
pfam00319	105	28	26.7	SRF-TF	SRF-type transcription factor (DNA-binding and
pfam03018	105	9	8.6	Dirigent	Dirigent-like protein
pfam00248	104	44	42.3	Aldo_ket_red	Aldo/keto reductase family
pfam13414	104	32	30.8	TPR_11	TPR repeat
pfam03151	103	28	27.2	TPT	Triose-phosphate Transporter family
pfam14226	103	25	24.3	DIOX_N	non-haem dioxygenase in morphine synthesis
pfam08263	103	20	19.4	LRRNT_2	Leucine rich repeat N-terminal domain
pfam01370	102	28	27.5	Epimerase	NAD dependent epimerase/dehydratase family
pfam07732	102	23	22.5	Cu-oxidase_3	Multicopper oxidase
pfam00170	101	32	31.7	bZIP_1	bZIP transcription factor
pfam03552	100	39	39.0	Cellulose_synt	Cellulose synthase
Total	57900	18218	31.5		
Unique Pfam	3505	2454			

Note: This is only a partial list; The information for accessing the complete list can be found in the main text

Table 4 Gene Ontology classification of molecular functions of gene products in the whole set of assembled transcripts and in isoforms generated by alternatively spliced genes in cotton

GO ID	GO Description	Total	AS	AS (%)
GO:0005488	binding	9315	4254	45.7
GO:0003824	catalytic activity	5984	2644	44.2
GO:0000166	nucleotide binding	3704	1648	44.5
GO:0016740	transferase activity	3682	1730	47.0
GO:0016787	hydrolase activity	3236	1446	44.7
GO:0005515	protein binding	2614	1318	50.4
GO:0003677	DNA binding	2174	1112	51.1
GO:0003723	RNA binding	1836	792	43.1
GO:0003676	nucleic acid binding	1697	847	49.9
GO:0005215	transporter activity	1254	579	46.2
GO:0016301	kinase activity	1188	652	54.9
GO:0003700	DNA binding transcription factor activity	1150	643	55.9
GO:0005198	structural molecule activity	819	245	29.9
GO:0004518	nuclease activity	674	285	42.3
GO:0030234	enzyme regulator activity	387	159	41.1
GO:0004871	signal transducer activity	359	203	56.5
GO:0030246	carbohydrate binding	309	118	38.2
GO:0008289	lipid binding	291	146	50.2
GO:0008135	translation factor activity, RNA binding	267	84	31.5
GO:0004872	receptor activity	232	138	59.5
GO:0003682	chromatin binding	158	84	53.2
GO:0003774	motor activity	134	77	57.5
GO:0005102	receptor binding	125	51	40.8
GO:0045182	translation regulator activity	8	6	75.0
GO:0019825	oxygen binding	7	3	42.9
Total		41604	19264	46.3

Table 5 Gene Ontology classification of biological processes of gene products in the whole set of assembled transcripts and in isoforms generated by alternatively spliced genes in cotton

GO ID	GO Description	Total	AS	AS (%)
GO:0009987	cellular process	12460	5686	45.6
GO:0008152	metabolic process	10791	4815	44.6
GO:0009058	biosynthetic process	5783	2625	45.4
GO:0006139	nucleobase-containing compound metabolic process	5186	2419	46.6
GO:0006950	response to stress	3007	1457	48.5
GO:0016043	cellular component organization	2995	1454	48.5
GO:0006810	transport	2405	1189	49.4
GO:0007275	multicellular organism development	2145	1134	52.9
GO:0009056	catabolic process	2033	889	43.7
GO:0007154	cell communication	1790	905	50.6
GO:0019538	protein metabolic process	1788	818	45.7
GO:0006464	cellular protein modification process	1769	924	52.2
GO:0007165	signal transduction	1726	868	50.3

Continued Table 5

GO ID	GO Description	Total	AS	AS (%)
GO:0009628	response to abiotic stimulus	1609	842	52.3
GO:0009719	response to endogenous stimulus	1472	759	51.6
GO:0000003	reproduction	1278	707	55.3
GO:0005975	carbohydrate metabolic process	1233	497	40.3
GO:0006629	lipid metabolic process	1147	514	44.8
GO:0009605	response to external stimulus	1096	556	50.7
GO:0009791	post-embryonic development	1061	622	58.6
GO:0006412	translation	947	325	34.3
GO:0009653	anatomical structure morphogenesis	915	455	49.7
GO:0030154	cell differentiation	910	446	49.0
GO:0007049	cell cycle	876	418	47.7
GO:0009607	response to biotic stimulus	837	414	49.5
GO:0006259	DNA metabolic process	778	343	44.1
GO:0040007	growth	574	303	52.8
GO:0006091	generation of precursor metabolites and energy	560	170	30.4
GO:0009908	flower development	435	280	64.4
GO:0009790	embryo development	399	224	56.1
GO:0015979	photosynthesis	371	109	29.4
GO:0019748	secondary metabolic process	361	134	37.1
GO:0019725	cellular homeostasis	355	167	47.0
GO:0016049	cell growth	321	171	53.3
GO:0008219	cell death	309	145	46.9
GO:0009991	response to extracellular stimulus	216	107	49.5
GO:0040029	regulation of gene expression, epigenetic	180	100	55.6
GO:0009856	pollination	141	78	55.3
GO:0007267	cell-cell signaling	82	34	41.5
GO:0009606	tropism	64	36	56.3
GO:0009875	pollen-pistil interaction	42	20	47.6
GO:0009835	fruit ripening	38	15	39.5
GO:0007610	behavior	34	14	41.2
GO:0009838	abscission	18	10	55.6
Total		72537	34198	47.1

Acknowledgments

The work was supported by the Youngstown State University Research Professorship award and a University Research Council (URC) grant to XJM.

References

Abdel-Ghany S.E., Hamilton M., Jacobi J.L., Ngam P., Devitt N., Schilkey F., Ben-Hur A., and Reddy A.S.N., 2016, A survey of the sorghum transcriptome using single-molecule long reads, Nat. Commun, 7: 11706
https://doi.org/10.1038/ncomms11706
PMid:27339290 PMCid:PMC4931028

Carvalho R.F., Feijão C.V., and 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

Chamala S., Feng G., Chavarro C., and Barbazuk W.B., 2015, Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants, Frontiers in Bioengineering & Biotechnology, 3: 33
https://doi.org/10.3389/fbioe.2015.00033
PMid:25859541 PMCid:PMC4374538

Filichkin S.A., Priest H.D., Givan S.A., Shen R., Bryant D.W., Fox S.E., Wong W.K., and Mockler T.C., 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., Rubin G.M., and Miller W., 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., and 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., and 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

Kiegle E.A., Garden A., Lacchini E., and Kater M.M., 2018, A genomic view of alternative splicing of long non-coding RNAs during rice seed development reveals extensive splicing and lncRNA gene families, Front. Plant Sci., 9: 115
https://doi.org/10.3389/fpls.2018.00115
PMid:29467783 PMCid:PMC5808331

Lewis B.P., Green R.E., and Brenner S.E., 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

Li F., Fan G., and Lu C. et al., 2015, Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution, Nat. Biotech., 33: 524
https://doi.org/10.1038/nbt.3208
PMid:25893780

Li F., Fan G., Wang K., Sun F., Yuan Y., Song G., Li Q., Ma Z., Lu C., Zou C., and Chen W., 2014a, Genome sequence of the cultivated cotton *Gossypium arboreum*, Nat. Genet., 46: 567
https://doi.org/10.1038/ng.2987
PMid:24836287

Li Q., Xiao G., and Zhu Y.X., 2014b, Single-nucleotide resolution mapping of the *Gossypium raimondii* transcriptome reveals a new mechanism for alternative splicing of introns, Molecular Plant, 7: 829-840
https://doi.org/10.1093/mp/sst175
PMid:24398628

Lightfoot D.J., Malone K.M., Timmis J.N., and Orford S.J., 2008, Evidence for alternative splicing of MADS-box transcripts in developing cotton fibre cells, Mole. Genet. Genom., 279: 75-85
https://doi.org/10.1007/s00438-007-0297-y
PMid:17943315

Lubbers E.L., and Chee P.W., 2009, The worldwide gene pool of *G. hirsutum* and its improvement, In Genetics and Genomics of Cotton, pp. 23-52, Springer, New York, NY.
https://doi.org/10.1007/978-0-387-70810-2_2

Marquez Y., Brown J.W.S., Simpson C., Barta A., and Kalyna M., 2012, Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis, Genome Res., 22: 1184-1195
https://doi.org/10.1101/gr.134106.111
PMid:22391557 PMCid:PMC3371709

McCarthy F.M., Wang N., and Magee G.B. 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.B., Boatwright J.L., Feng G., Schnable J.C., and Barbazuk W.B., 2017, Evolutionarily conserved alternative splicing across monocots, Genetics, 207: 465-480
https://doi.org/10.1101/120469

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

Min X.J., Butler G., Storms R., and Tsang A., 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 X.J., Butler G., Storms R., and Tsang A., 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

Min X.J., Powell B., Braessler J., Meinken J., Yu F., and Sablok G., 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

Panahi B., Abbaszadeh B., Taghizadeghan M., and Ebrahimie E., 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 A.S.N., Marquez Y., Kalyna M., and Barta A., 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., and 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

Sablok G., Gupta P.K., Baek J.M., Vazquez F., and Min X.J., 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., and Min X.J., 2017, Comparative landscape of alternative splicing in fruit plants, Current Plant Biology

Staiger D., and Brown J.W., 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., Tang Y.H., Zhang Z.Y., Toiber D., Thanaraj T.A., and Soreq H., 2005, Function of alternative splicing, Gene, 344: 1-20
https://doi.org/10.1016/j.gene.2004.10.022
PMid:15656968

Syed N.H., Kalyna M., Marquez Y., Barta A., and Brown J.W.S., 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 S.R., Danilevskaya O.N., Meng X., Beatty M., Zastrow-Hayes G., Harris C., Allen B.V., Habben J., and Li B., 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 S.R., Zhou W., Leonard A., Wang B.B., Beatty M., Zastrow-Hayes G., Zhao X.Y., Baumgarten A., and Li B., 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., and 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., and Carpinelli E.C. 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

Wai C.M., Powell B., Ming R., and Min X.J., 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

Walters B., Lum G., Sablok G., and Min X.J., 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

Wang B., and 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

Wang K., Wang Z., and Li F. et al., 2012, The draft genome of a diploid cotton *Gossypium raimondii*, Nat. Genet., 44: 1098
https://doi.org/10.1038/ng.2371
PMid:22922876

Wang M., Wang P., and Liang F. et al., 2018, A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation, New Phytologist, 217: 163-178
https://doi.org/10.1111/nph.14762
PMid:28892169

Wei H., Lou Q., Xu K., Yan M., Xia H., Ma X., Yu X., and Luo L., 2017, Alternative splicing complexity contributes to genetic improvement of drought resistance in the rice maintainer HuHan2B, Sci. Rep., 7: 11686
https://doi.org/10.1038/s41598-017-12020-3
PMid:28916800 PMCid:PMC5601427

Wendel J.F., and Grover C.E., 2015, Taxonomy and evolution of the cotton genus, *Gossypium*, Cotton, (agronmonogr57): 25-44

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

Zhang P.G., Huang S.Z., Pin A.L., and Adams K.L.,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

Zhang R., Calixto C.P., and Marquez Y. et al., 2017, A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing, Nucleic Acids Res., 45: 5061-73
https://doi.org/10.1093/nar/gkx267
PMid:28402429 PMCid:PMC5435985

Zhang T., Hu Y., and Jiang W. et al., 2015, Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement, Nat. Biotech., 33: 531
https://doi.org/10.1038/nbt.3207
PMid:25893781

Zhu G., Li W., Zhang F., and Guo W., 2018, RNA-seq analysis reveals alternative splicing under salt stress in cotton, *Gossypium davidsonii*, BMC Genomics, 19: 73
https://doi.org/10.1186/s12864-018-4449-8
PMid:29361913 PMCid:PMC5782385