EPiC Series in Computing

Volume 83, 2022, Pages 1–9

Proceedings of 14th International Conference on Bioinformatics and Computational Biology

Identification and Analysis of Alternative Splicing in Soybean Plants

Xiangjia Min[1*], Theoni Kasamias[1], Mykaela Wagner[1], Atinuke Ogunbayi[1], Feng Yu[2]

[1] Department of Chemical and Biological Sciences, Youngstown State University, OH 44555, USA

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

xmin@ysu.edu, tkasamias@student.ysu.edu, mwagner01@student.ysu.edu, aeogungbayi@student.ysu.edu, fyu@ysu.edu

Abstract

Alternative splicing (AS) increases the diversities of transcriptomes and proteomes in plants. The work reports identification and analysis of genes and their transcripts with a focus on AS in soybean plants by integrating mapping information of over 1.5 million of mRNAs and expressed sequence tags (ESTs) with more than 6 billons of mapped reads collected from 90 RNA-seq datasets obtained from multiple experiments. A total of 294,164 AS events were detected and categorized into basic events (151,710, 51.57%) and complex events (142,454, 48.43%). The basic AS events include intron retention (18.52%), alternative acceptor sites (16.33%), alternative donor site (8.99%), and exon skipping (7.73%). The AS rate in intron containing genes was estimated to be ~56.3% in soybean based on the current analysis. In addition, a total of 41,453 new genomic loci, which were not previously annotated in the genome, were detected by mapping transcripts to the genome. The annotated data can be accessed through a public database for searching and downloading. This work provides a resource for further detailed functional analysis of gene products in soybean plants.

1 Introduction

Soybean [Glycine max] is an important food and oil crop for humans as well as one of the major food sources for animal husbandry. The first completely sequenced genome of soybean plant was reported by Schmutz et al. (2010) [1]. Currently there are 13 soybean genomes available in the genome database at the National Center for Biotechnology Information (NCBI;

https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/5/). The genome sequences combining with recent RNA-sequencing (RNA-seq) data provide critical information for increasing our understanding of the gene regulations of growth and development as well as responses to various biotic and abiotic stresses at the molecular levels in soybean [2].

RNA-seq is a technology using next-generation sequencing (NGS) to identify and quantify RNA transcripts of a transcriptome in a biological sample [3]. Because RNA-seq has a resolution of one base-pair, it is widely used to determine exon/intron boundaries and thus to identify alternatively spliced transcripts, and to quantify the gene expression levels over time or in different treatments for all types of transcripts including mRNAs and non-coding RNAs, such as small RNA, miRNA, etc. [4]. The Sequence Read Archive (SRA) database at NCBI (https://www.ncbi.nlm.nih.gov) stores RNA-seq data (https://www.ncbi.nlm.nih.gov/sra). Thus, the available genome sequences and RNA-seq data provide researchers a rich data source for examining the gene expressions across different tissues or biological treatments. Recently we reviewed the applications of RNA-seq in transcriptome analysis in different biological samples with various biotic and abiotic treatments in soybean plants, revealing significant progresses have been made in many aspects of soybean biological research [2].

Preliminary RNA (pre-RNA) transcripts in intron containing genes in eukaryotic organisms often undergo alternative splicing (AS) which produces more than one mature RNA transcript from a single pre-RNA. The rate of AS can be as high as more than 90% in human genes and it is reported ~70% genes undergoing AS in plant species [5-8]. In plants AS plays important biological roles in regulation of growth and development as well as biotic and abiotic stresses [9,10]. According to our recent review with recent publications related RNA-seq applications in soybean research, we noticed most of transcriptome analyses focused on identification of differentially expressed genes (DEGs) and only a few of published articles carried out AS analysis [11-14, 2]. In consideration of the important roles AS played in regulation of transcriptome and proteome diversity and its impact on the functions of the products of transcript isoforms in plants we integrated multiple sources of data generated from several different experiments to identify genes undergoing AS and their transcript isoforms.

2 Materials and Methods

2.1 Assembling and mapping of expressed sequence tags (ESTs) and mRNA sequences

The genome sequence and associated genome annotation files (version 2.1) were downloaded from Ensembl database (ftp://ftp.ensemblgenomes.org/pub/plants/release-44/fasta/glycine_max/). A total of 1,558,669 soybean EST and mRNA transcripts were downloaded from NCBI nucleotide database by setting limit to RNA within “Glycine max” organism. Data cleaning was performed on the transcripts using a well implemented procedure as described in our previous analysis [15, 16]. In short, the procedure incudes using EMBOSS trimmest tool for trimming the polyA or polyT end [17], BLASTN search against UniVec and E. coli database for removal of vector and E. coli contaminants, and BLASTN search against the plant repeat database for removal of the repetitive sequences and transposable elements. A final cleaned transcript dataset was further assembled into a non-redundant set using CAP3 [18]. The reason we used CAP3 is that based on our testing this program was particularly useful in assembling ESTS to generate a non-redundant set of unique transcripts suitable for alternative splicing analysis [19]. This set of assembled EST and mRNA transcripts was referred as putative unique transcripts (PUTs). ASFinder and Sim4 programs were used to map the assembled transcripts to soybean genome sequences using cutoff values of a minimum 95% identity and >75% of a PUT sequence length in mapping alignment, as described in our previous work [20 -22]. Mapped

transcripts having an intron size >100 kb were removed for AS identification in order to avoid chimeric transcripts.

2.2 Soybean genome sequences and RNA-seq data collection and mapping to the genome

Soybean genome sequences with GFF3 annotation files were downloaded from Ensembl Plant database (https://plants.ensembl.org/index.html). Soybean RNA-seq data were downloaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/) using SRA Toolkit. Due to recent advance of RNA-seq technology, there are totally over 5000 datasets in soybean plants available, which are beyond our computational processing capacity [2]. Thus, we only selected 89 datasets, which represented a broad range of tissues and different treatments, from recently published projects in this work for mapping the data to soybean genome for identification of AS [11-14, 23-25]. The RNA-seq reads were mapped to soybean genome sequences using TopHat (v2.2.6) with default parameters [26]. Then the transcript alignment file together with annotation GFF3 (v2.1.44) was used as input for Cufflinks (v2.2.1) (http://cole-trapnell-lab.github.io/cufflinks/) [27]. The GTF (Gene Transfer Format) files generated from each RNA-seq dataset after Cufflinks were merged using Cuffcompare script within the Cufflinks package [27]. The GTF file generated from merged RNA-seq GTF files then was further merged using Cuffcompare script with the GTF file that was generated by the ASFinder for mapping the assembled PUTs to the genome to generate a final GTF file for AS analysis. AStalavista was used for AS event classification [28].

2.3 Transcript functional annotation

The transcript sequences were retrieved using gtf_to_fasta tool in the tophat package [26], based on the GTF file generated by Cuffcompare program after merging PUTs mapping GTF file and RNA-seq mapping GTF file. The transcripts were functionally annotated, including protein coding regions (ORF) prediction, assessment of full–length transcript coverage, protein family, and comparison with predicted gene models. The protein family (Pfam) classification was carried using a locally installed BLASTALL package with rpsBLAST tool.

2.4 Data availability

The assembled transcripts and AS events identified in this study are available through Plant Alternative Splicing Database (http://proteomics.ysu.edu/altsplice). This site also provides data from our previously published work including Brachypodium distachyon, pineapple, sacred lotus (*Nelumbo nucifera*), rice, maize, sorghum, and cotton [16, 22, 29-32]. BLAST is provided for searching the transcripts. The datasets for database construction, RNA-seq mapping GTF files, transcript sequences, and the supplementary data are publicly available at: http://proteomics.ysu.edu/publication/data/Soybean/.

3 Results

3.1 EST/mRNA transcripts assembling and mapping

After pre-cleaning steps started with a total of 1,558,669 soybean EST and mRNA transcripts which were downloaded from NCBI nucleotide database, we obtained a final dataset consisting of 1,461,227 sequences with a minimum length of 100 bp for further assembling using CAP3 [18]. After assembling a set of 208,896 PUTs with 47,720 contigs and 161,194 singlets, having an average length of 429 bp,

were generated for mapping to soybean genome sequences. A total of 155, 188 (74.3%) PUTs were mapped to the genome and 8,978 (5.8% of the mapped) PUTs were mapped to more than one genomic locus, using ASFinder program [20]. There were 25.7% of the PUTs not being mapped to the draft genome sequences, which might be due to the incompleteness of the genome sequences or possible sequence errors in either the PUT assembly or genomic sequences. The reason for some PUTs mapped to the more than one genomic locus was likely due to recent gene duplications occurred in the genome. The draft genome sequence was estimated representing about 85% of the predicted whole genome in soybean with 75% of the genes presented in multiple copies due to whole genome duplications [1].

3.2 Mapping RNA-seq data to the genome and merging with PUT mapping information

We mapped 90 RNA-seq datasets, including 19 datasets with single reads and 71 datasets with paired reads, to soybean reference genome. The accession numbers and mapping information of these RNA-seq data were listed in supplementary Table 1. A total of 6.99 billion reads were collected with 6.17 billion reads (~88.3%) mapped to the genome. Among the mapped reads, ~9.1% reads (564 million) were mapped to more than one genomic locus.

Merging EST/mRNA PUTs mapping data with RNA-seq mapping data generated a non-redundant set of 304,677 transcripts from 93,979 genomic loci, with 33,521 loci having two or more transcripts (Table 1). A total of 41,453 new genomic loci were detected with RNA-seq and PUT mapping data that were not previously annotated in the genome. It should be noted that there are 1170 contig DNA sequences, in addition to the assembled 20 chromosome DNA sequences, in the released soybean genome. Thus, the real gene number is expected to be less than the genomic loci detected in the work as a gene sequence might be fragmented into different short genomic contigs.

Total genomic loci	93979
New genomic loci	41453
Total transcripts	304667
Transcripts match to transcripts of gene models	233472 (76.6%)
Average length of transcripts (bp)	2149
BLASTX match against Swiss-Prot database	228997 (75.5%)
Total predicted ORFs	303448 (99.60%)
Average length of ORFs (aa)	345
Total Pfam matches of ORFs	181993 (60.0%)

Table 1. Basic features of RNA transcripts and functional annotation in soybean plants

3.3 Functional annotation of transcripts

A total of 304,677 RNA transcript sequences were retrieved and annotated, including ORF prediction, coding region completeness assessment, function and Pfam prediction. These basic features of these transcripts were summarized in Table 1. A total of 303,448 ORFs with an average length of

345 amino acids were predicted and among them 228,997 (75.5%) were functionally annotated based on BLASTX matches and 181,993 (60%) were classified to a Pfam. In addition, using BLASTN search with a cutoff of 100% identity 233472 (76.7%) transcripts matched with predicted transcripts of gene models.

3.4 Detection and classification of alternative splicing events

We analyzed AS events in EST/mRNA assembled PUTs data, RNA-seq mapping data, and the merged data (Table 2). Among the basic events, intron retention remained to be the most frequent event and exon skipping was the least occurred events in all datasets. A total of 294,164 AS events were detected and categorized into basic events (151,710, 51.57%) and complex events (142,454, 48.43%). The basic AS events include intron retention (18.52%), alternative acceptor sites (16.33%), alternative donor site (8.99%), and exon skipping (7.73%). The complex events consist of two or more basic events in the compared transcript pairs (Table 2). It was noted that the numbers of total events and some categories of basic events in the merged dataset were lower than the numbers obtained in the RNA-seq data, suggesting some transcripts were merged after combining EST/mRNA PUTs mapping data. These AS events were generated from 30,394 genomic loci with 195,098 unique transcripts. As there are a total of 93,979 genomic loci identified with mapping data, thus, the AS rates of genes generating AS isoforms (AS genes) was estimated to be ~32.3%. However, there are a total of 40,011 loci having a transcript without an intron, i. e., only one exon. Thus, the AS rate in genes having at least one intron was estimated to be ~56.3%. The AS rate has been reported ~70% in Arabidopsis, 55% in maize, and 65.0% in tomato previously [8, 32-34]. As we pointed in earlier section, there are more than 5000 RNA-seq data deposited in the SRA database at NCBI currently and our work only used 89 RNA-seq datasets, thus a higher AS rate in soybean plants is expected when more gene expression data are incorporated.

	EST/mRNA	%	RNA-seq	%	Merged	%
Exon skipping	3236	10.26	19359	5.77	22738	7.73
Alternative donor sites	3885	12.32	27587	8.22	26453	8.99
Alternative acceptor sites	5776	18.32	47004	14.01	48045	16.33
Intron retention	9973	31.63	64303	19.17	54474	18.52
Others (complex events)	8664	27.48	177221	52.83	142454	48.43
Total	31534		335474		294164	

Table 2. Summary of identified alternative splicing events based on data of mapping to the genome of EST/mRNA assembly, RNA-seq data, and the merged data in soybean plants

3.5 AS in different protein families

The proteins predicted from mRNA transcripts identified in the work were mapped to Pfam, that facilitate database search and compare the functional domains in proteins translated from different transcript isoforms. We examined the AS distribution of genes encoding different Pfams, a partial list

of Pfam distribution and associated proportion of AS genes in each family is shown in Table 3 and the complete list can be found in the supplementary Table 2. The AS genes refer the genes having AS transcript isoforms. Only one Pfam annotation was selected from each genomic locus. A total of 57,374 Pfam matches from a total of 3,523 unique protein families were obtained from encoded proteins of a total of 93,979 genomic loci. Since different genes encoding different proteins have different exon structures, thus as expected, the AS proportions of genes encoding different protein families are different (Table 3).

4 Discussion

Plant gene expression is a highly regulated process including both differential gene expressions and alternative splicing. RNA-seq technology has been widely applied in soybean research and most of those studies focused on analysis of DEGs [2]. In this work we identified more genes undergoing AS and their associated isoform transcripts than previously published data as we combined all known mRNA sequences including ESTs and multiple RNA-seq datasets generated from different experiments [11-14; 23-25]. The data generated in our work are publicly available from our server that would facilitate researchers to use the information for their research. We are aware that this analysis unfortunately did not include transcriptome data reported by Shen et al. (2014) [35]. Though more AS events and genes subjected to AS were identified in our work, additional AS events, novel isoforms and genes undergoing AS might be reported in the work of Shen et al. (2014) [35]. Thus, we strongly suggest soybean researchers to make a reference of the work of Shen et al. (2014) [35].

Computational identification of genes undergoing AS and their associated transcript isoform sequences are helpful for experimental researchers to design more specific experiments to examine the functions of genes of interests in details. However, it should be noticed that the isoform sequences are derived from assembling fragments of RNA-seq data which may not be accurate, thus, most of these sequences need to be confirmed using specific primers to obtain full-length mRNAs. As there are more RNA-seq data available in the public database and more RNA-seq data will be generated in future, our current work is expected to serve as a foundation for future incorporation of more data, including data reported by Shen et al. [35], for expanding the repertoire of the genes subject to AS and their associated isoform sequences in soybean plants.

5 Author Contributions

XJM and FY designed the experiments. TK, MW and AO collected and processed RNA-seq data for genome mapping. XJM performed EST assembling, transcripts functional annotation and data analysis. XJM and FY prepared the manuscript.

6 Acknowledgements

The work was supported by the Youngstown State University Research Professorship award and reassigned time for scholarship from College of Science, Technology, Engineering, and Mathematics and Department of Chemical and Biological Sciences at YSU to XJM. The Ohio Super Computer Center was acknowledged for providing resources for data processing. We thank the three anonymous reviewers for constructive comments.

Pfam ID	Total	AS genes	%	Pfam	Pfam description
pfam00069	1747	701	40.1	Pkinase	Protein kinase domain
pfam07714	1327	551	41.5	Pkinase_Tyr	Protein tyrosine kinase
pfam13041	1024	256	25.0	PPR_2	PPR repeat family
pfam00067	617	114	18.5	p450	Cytochrome P450
pfam13639	545	166	30.5	zf-RING_2	Ring finger domain
pfam00249	524	124	23.7	Myb_DNA-binding	Myb-like DNA-binding domain
pfam00931	507	140	27.6	NB-ARC	NB-ARC domain
pfam00076	478	265	55.4	RRM_1	RNA recognition motif
pfam00847	386	67	17.4	AP2	AP2 domain
pfam00201	348	28	8.0	UDPGT	UDP-glucoronosyl and UDP-glucosyl transferase
pfam05970	298	27	9.1	PIF1	PIF1-like helicase
pfam14432	294	48	16.3	DYW_deaminase	DYW family of nucleic acid deaminases
pfam00010	285	121	42.5	HLH	Helix-loop-helix DNA-binding domain
pfam00141	283	46	16.3	peroxidase	Peroxidase
pfam00071	276	79	28.6	Ras	Ras family
pfam03171	265	58	21.9	2OG-FeII_Oxy	2OG-Fe(II) oxygenase superfamily
pfam02519	257	5	1.9	Auxin_inducible	Auxin responsive protein
pfam03106	254	83	32.7	WRKY	WRKY DNA -binding domain
pfam01582	250	103	41.2	TIR	TIR domain
pfam00481	242	124	51.2	PP2C	Protein phosphatase 2C
pfam00083	223	70	31.4	Sugar_tr	Sugar (and other) transporter
pfam02458	223	21	9.4	Transferase	Transferase family
pfam03514	222	36	16.2	GRAS	GRAS domain family
pfam00153	220	95	43.2	Mito_carr	Mitochondrial carrier protein
pfam00657	219	74	33.8	Lipase_GDSL	GDSL-like Lipase/Acylhydrolase
pfam01490	210	84	40.0	Aa_trans	Transmembrane amino acid transporter
Total	57374	21202	37.0		

Table 3. A partial list of top protein families encoded by genes with varying proportion of alternative splicing in soybean plants

References

[1] Schmutz J., Cannon S.B., Schlueter J., et al., 2010, Genome sequence of the palaeopolyploid soybean, Nature, 463(7278): 178-183
[2] Min X.J., Wagner M., Kasamias T., 2021, Advances in transcriptome analyses using RNA sequencing technology in soybean plants [Glycine max]. Computational Molecular Biology. 11(1):1-14
[3] Chu Y., and Corey D.R., 2012, RNA sequencing: platform selection, experimental design, and data interpretation, Nucleic Acid Therapeutics, 22(4): 271-274
[4] Wang Z., Gerstein M., and Snyder M., 2009, RNA-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1): 57-63
[5] Kim D., Pertea G., Trapnell C., et al., 2013, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36. doi: 10.1186/gb-2013-14-4-r36
[6] Pan Q., Shai O., Lee L.J., et al., 2008, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40, 1413–1415. doi: 10.1038/ng.259
[7] Chaudhary S., Khokhar W., Jabre I., et al., 2019. Alternative splicing and protein diversity: plants versus animals, Frontiers in plant science, 10, p.708.
[8] Clark S., Yu F., Gu L., et al., 2019, Expanding alternative splicing identification by integrating multiple sources of transcription data in tomato, Frontiers in Plant Sciences. 10:689. doi:10.3389/fpls.2019.00689.
[9] Staiger D., and Brown J.W., 2013, Alternative splicing at the intersection of biological timing, development, and stress responses, Plant Cell, 25: 3640-3656
[10] Reddy A.S., Marquez Y., Kalyna M., et al., 2013, Complexity of the alternative splicing landscape in plants, Plant Cell, 25: 3657-3683
[11] Aghamirzaie D., Nabiyouni M., Fang Y., et al., 2013, Changes in RNA splicing in developing soybean (Glycine max) embryos, Biology, 2(4): 1311-1337
[12] Wang L., Cao C., Ma Q., et al., 2014, RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications, BMC Plant Biol., 14(1): 169
[13] Wang Z., Zhang H., and Gong W., 2019, Genome-wide identification and comparative analysis of alternative splicing across four legume species, Planta, 249(4):1133-1142. doi: 10.1007/s00425-018-03073-3.
[14] Song L., Pan Z., Chen L., et al., 2020, Analysis of whole transcriptome RNA-seq data reveals many alternative splicing events in soybean roots under drought stress conditions, Genes (Basel), 11(12):1520
[15] Min X.J., Powell B., Braessler J., et al., 2015, Genome-wide cataloging and analysis of alternatively spliced genes in cereal crops. BMC Genomics 16:721. doi: 10.1186/s12864-015-1914-5
[16] Min X.J., 2018, A survey of alternative splicing in allotetraploid cotton (Gossypium hirsutum L.). Comp. Mol. Biol. 8, 1-13. doi: 10.5376/cmb.2018.08.0001
[17] Rice, P., Longden, I., Bleasby, A., 2000, EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276-277.
[18] Huang X., Madan A., 1999, CAP3: A DNA sequence assembly program. Genome Res. 9:868-877.
[19] Min X.J., Butler G., Storms R., et al., 2009, Comparative assessment of DNA assemblers for assembling expressed sequence tags. 2009 Ohio Collaborative Conference on Bioinformatics. Cleveland, Ohio. The IEEE Computer Society. 79-82.
[20] Min X.J., 2013, ASFinder: a tool for genome-wide identification of alternatively spliced transcripts from EST-derived sequences. Int. J. Bioinform. Res. Appl. 9, 221-226. doi: 10.1504/IJBRA.2013.053603

[21] 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.
[22] Walters B, Lum G, Sablok G, et al., 2013, Genome-wide landscape of alternative splicing events in Brachypodium distachyon. DNA Res 20:163-171.
[23] Lv L., Yu K., Lü H., et al., 2020, Transcriptome-wide identification of novel circular RNAs in soybean in response to low-phosphorus stress, PloS one, 15(1): e0227243
[24] Jones S.I., and Vodkin L.O., 2013, Using RNA-seq to profile soybean seed development from fertilization to maturity, PloS one, 8(3):e59270.
[25] Neupane S., Mathew F.M., Varenhorst A.J., et al., 2019, Transcriptome profiling of interaction effects of soybean cyst nematodes and soybean aphids on soybean, Scientific Data. 6(1):133. DOI: 10.1038/s41597-019-0140-4
[26] Kim E., Magen A., and Ast G., 2007, Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35, 125–131. doi: 10.1093/nar/gkl924
[27] Trapnell C., Williams B.A., Pertea G., et al., 2010, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511-515. doi: 10.1038/nbt.1621
[28] Foissac S., Sammeth M., 2007, ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Res. 2007; 35:W297–299. doi: 10.1093/nar/gkm311
[29] Wai C.M., Powell B., Ming R., et al., 2016, Analysis of alternative splicing landscape in pineapple (Ananas comosus). Tropical Plant Biol. 9, 150-160. doi: 10.1007/s12042-016-9168-1
[30] Wai C.M., Powell B., Ming R., et al., 2018, Analysis of transcriptome and alternative splicing landscape in pineapple. In: Ming R. (eds) Genetics and Genomics of Pineapple. Plant Genetics and Genomics: Crops and Models, vol. 22. Springer. doi.:10.1007/978-3-030-00614-3_14
[31] Sablok G., Powell B., Braessler J., et al., Comparative landscape of alternative splicing in fruit plants. Current Plant Biol. 9, 29-36. doi: 10.1016/j.cpb.2017.06.001.
[32] Min X.J., 2017, Comprehensive cataloging and analysis of alternative splicing in maize. Comp. Mol. Biol. 7, 1-11. doi: 10.5376/cmb.2017.07.0001
[33] Marquez Y., Brown J.W., Simpson C. G., et al., 2012, Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res. 22, 1184-1195. doi: 10.1101/gr.134106.111
[34] Mei W., Boatwright L., Feng G., et al., 2017, Evolutionarily conserved alternative splicing across monocots. Genetics 207, 465-480. doi: 10.1534/genetics.117.300189
[35] Shen Y., Zhou Z., Wang Z., et al., 2014, Global dissection of alternative splicing in paleopolyploid soybean. The Plant Cell, 26(3), 996-1008.