OPEN ACCESS Freely available online

Journal of Horticulture

ISSN: 2376-0354

Research Article

Genome-wide Analysis of mRNA Splicing Variants in Higher Plants

Ying Li, Qianhuan Guo, Chengchao Zheng, Shizhong Zhang*, Kang Yan*

College of Life Sciences, State Key Laboratory of Crop Biology, Shandong Agricultural University, Taian, Shandong 271018, P.R. China

ABSTRACT

Alternative splicing (AS) produces multiple mRNA splicing variants from a single precursor transcript. Recent genome sequencing analyses and increasing experimental evidence in flowering plants have revealed that AS is far more prevalent than previously thought and plays crucial roles in the diversification of gene regulation. Despite numerous studies, the extent and complexity of mRNA variants in plants remain poorly characterized from a global perspective. In present study, 589,034 mRNA variants from 442,541 annotated genes of 12 plant species were investigated. All AS genes were classified into four groups on the basis of the numbers of mRNA variants, namely, 2V (two variants), 3V (three variants), 4V (four variants), and 5V+ (five or more variants). Interestingly, our analysis indicated that more than 50% AS genes generated only two variants in higher plants. A global analysis of gene structure revealed that AS genes contained more but shorter exons and introns as the number of mRNA variants increased. The results also suggested that AS elicited different effects on the improvement of transcriptome and proteome diversity. Taken together, cross-species analysis provided the most comprehensive set of annotated splicing variants in higher plants thus far and extended the current view about mRNA variants.

Keywords: mRNA; Alternative splicing; Deep sequencing; Splicing variants

Abbreviations: AA: Alternative Acceptor Sites; AD: Alternative Donor Sites; AS: Alternative Splicing; ES: Exon Skipping; IR: Intron Retention

INTRODUCTION

Alternative splicing (AS) is a common and fundamental process that contributes to both transcriptome and proteome diversities [1]. AS generates two or more mRNA variants from a single pre-mRNA with multiple introns through different splice sites [2,3]. Most plant genes (80% to 85%) are interrupted by introns [4]. However, the intron and exon organization of higher plant genes is not similar to that of animal genes [5]. Introns in plant genes are generally much shorter in length and fewer in number compared with those in animal genes [6]. In contrast to reports on the differences between plant and animal species, studies that compare the gene structures of different plant species such as food crops and horticultural crops are scarce.

Moreover, AS has been found to be an important regulatory process in different cell types, different developmental stages, and environmental responses. Therefore, individual mRNA variants may play specific spatial or temporal roles [7,8]. The crucial role of AS has been extensively investigated in animals and plants [9,10]. AS is associated with the human genetic diseases, and also involved in a range of functions in plants, such as seed germination, stress response [11,12]. In eukaryotes, AS of pre-mRNAs significantly contributes to the proper expression of the genome and results in new protein products (13). However, not all of the products are functional [13,14]. Several AS variants contain premature termination codons (PTCs) that potentially lead to unproductive transcripts, truncated proteins or mRNA decay [15,16].

Conversely, recent studies indicated AS act as a regulatory mechanism influencing the expression of non-coding RNA [17]. Primary transcripts of miRNAs (pri-miRNAs) contain introns, and their AS events has been detected [18,19]. In *Arabidopsis*, AS event disrupts the secondary structures of pri-miR162a and provides a mechanism for miR400 expression in response to environmental cues [20,21]. Tomato as a kind of very important horticultural crops, research has found that the AS of miR4376 through regulating the expression of an autoinhibited Ca^{2+}-ATPase, *ACA10*, affect the reproductive growth of tomato [22,23]. Thus, growing evidence has indicated that AS in plants is much more prevalent than previously thought and plays crucial roles in the diversification of gene regulation.

Genome-wide studies of AS events in various organisms have

Correspondence to: Kang Yan, College of Life Sciences, Shandong Agricultural University, Taian, handong 271018, P.R. China, Tel: +86-538-8242894, Fax: +86-538-8226399, E-mail: kangyan@sdau.edu.cn

Shizhong Zhang, College of Life Sciences, Shandong Agricultural University, Taian, Shandong 271018, P.R. China, Tel: +86-538-8242894, Fax: +86-538-8226399, E-mail: shizhong@sdau.edu.cn

Received: April 06, 2019; **Accepted:** April 24, 2019; **Published:** May 03, 2019

Citation: Li Y, Guo Q, Zheng C, Zhang S, Yan K (2019) Genome-Wide Analysis of mRNA Splicing vaRiants in Higher Plants. J Hortic 6:255. doi: 10.35248/2376-0354.19.06.255

relied on using the traditional expressed sequence tag and cDNA libraries [24]. However, the occurrence of AS events might be underestimated because data from these libraries do not provide a complete coverage of the transcriptome. High-throughput RNA sequencing (RNA-seq) technologies allow transcriptome analyses at unprecedented levels of sensitivity and precision. Such technologies also provide large data sets for comprehensive analyses [8].

Studies on plant RNA-seq data sets have identified thousands of AS events and confirmed most annotated introns in several vertebrate species. A previous RNA-seq analysis revealed that approximately 95% of human multi-exon genes express multiple splicing variants [25]. Recently, it was shown that 61% of Arabidopsis intron-containing genes and 56% of maize multiexonic genes are subject to AS [26,27]. A large number of variable splicing events of genes have been found in horticultural crops. For example, in strawberries, 66.43% detected multi-exon genes undergo AS [28,29]. And in cucumber, recent studies have found that 58% of the multi-exon genes underwent AS [30]. Also, a large number of genes were alternatively spliced in the soybean genome [31]. To our knowledge, the extent and frequency of plant mRNA variants remain poorly characterized from a global perspective.

According to different AS types, AS events can be mainly classified as exon skipping (ES), intron retention (IR), alternative donor sites (AD), and alternative acceptor sites (AA). IR is the most common splicing form in plants [32]. However, one AS gene variant might simultaneously have multiple different AS types. In this study, to achieve a nonbiased and complete analysis of the plant transcriptome, we analyzed the mRNA splicing variants of 12 plant species (8 dicotyledons and 4 monocotyledons) including food crops and horticultural crops which subjected to high-throughput sequencing database, including Poplar (*Populus trichocarpa*), Medicago (*Medicago truncatula*), soybean (*Glycine max*), Arabidopsis (*Arabidopsis thaliana*), cotton (*Gossypium raimondii*), cacao (*Theobroma cacao*), rose gum (*Eucalyptus grandis*), blue columbine (*Aquilegia coerulea*), sorghum (*Sorghum bicolor*), maize (*Zea mays*), rice (*Oryza sativa*), and Brachypodium (*Brachypodium distachyon*). All mRNA variants from the annotated genes were confirmed by genome alignment and investigated the relationship between the number of AS mRNA variants and their effects on gene features at the genome-wide level.

MATERIAL AND METHODS

Database set-up

A total of 12 recently published RNA-seq data sets from higher plants were selected and downloaded at Phytozome, which is the Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute (https://phytozome.jgi.doe.gov/pz/portal.html). Detail information about mRNA variants of 12 plant species is provided in the Supplemental section (Table 1 and Supplementary Table 1).

mRNA variants alignment to the reference plant genome

A total of 589,034 mRNA variants were aligned to perfectly match gene locations in the reference genome with a maximum of two mismatches. Among the mRNA splicing products, only the annotated genes were selected and classified into groups according to the number of mRNA variants. The correlative gene features (introns, coding sequence, exons, or UTRs) were generated, including their distribution, location, length, and numbers. Detail information about the gene features are given in the Supplemental section (Supplementary Table 1).

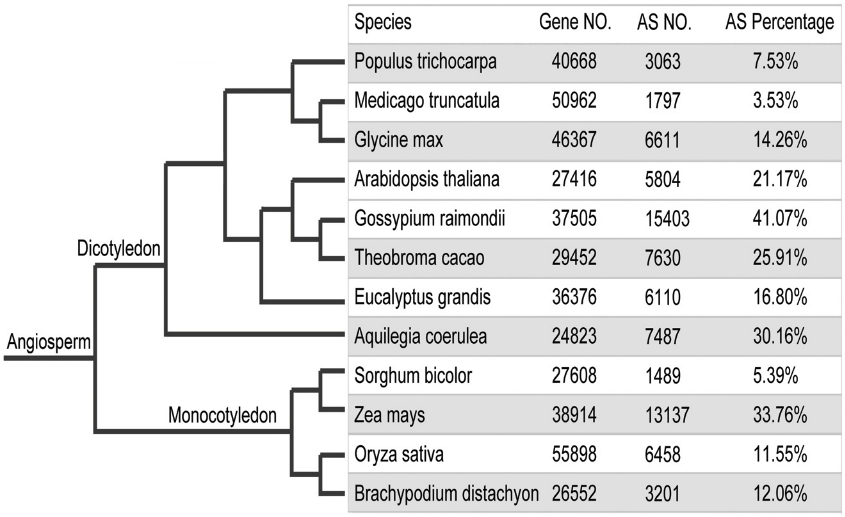
Clustering of AS gene features

The gene features in the different groups were calculated.

RESULTS

Identification and classification of mRNA variants in higher plants

In general, RNA-seq provides broad and deep sequencing of the transcriptome with low false discovery rates. Therefore, to facilitate further investigation of plant mRNA variants from a global perspective, we mined the RNA-seq data sets of 12 plant species to identify thousands of mRNA variants and all the gene sets from 12 annotation genomes downloaded at https://phytozome.jgi.doe.gov/pz/portal.html (Figure 1 and Table 1). All mRNA variants were aligned to perfectly match the corresponding gene locations in the reference genome (Supplementary Table 1). To ensure the



Species	Gene NO.	AS NO.	AS Percentage
Populus trichocarpa	40668	3063	7.53%
Medicago truncatula	50962	1797	3.53%
Glycine max	46367	6611	14.26%
Arabidopsis thaliana	27416	5804	21.17%
Gossypium raimondii	37505	15403	41.07%
Theobroma cacao	29452	7630	25.91%
Eucalyptus grandis	36376	6110	16.80%
Aquilegia coerulea	24823	7487	30.16%
Sorghum bicolor	27608	1489	5.39%
Zea mays	38914	13137	33.76%
Oryza sativa	55898	6458	11.55%
Brachypodium distachyon	26552	3201	12.06%

Figure 1: Characteristics of plant annotated gene features and AS variants.
Statistics of the annotated gene features and AS events in 12 plant species. The line represents the evolutionary relationships among various biological species in the phylogenetic tree.

Table 1: The mRNA variants of 12 plant species used in this study.

Species	Gene number	The number of mRNA variants	AS gene number	Percentage	The number of AS variants	mRNA variants per AS gene	Database
Populus trichocarpa	40668	45033	3063	0.075317203	7428	2.425073457	http://www.plantgdb.org/PtGDB/
Medicago truncatula	50962	53423	1797	0.035261567	4258	2.36950473	http://www.medicago.org/
Glycine max	46367	55788	6611	0.142579852	16031	2.424897897	http://www.plantgdb.org/GmGDB/
Arabidopsis thaliana	27416	35386	5804	0.211701196	13774	2.373190903	http://www.arabidopsis.org/
Gossypium raimondii	37505	77267	15403	0.410691908	55165	3.581445173	http://www.cottondb.org/wwwroot/cdbhome.php
Theobroma cacao	29452	44404	7630	0.259065598	22582	2.959633028	http://www.cacaogenomedb.org/
Eucalyptus grandis	36376	46315	6110	0.167967891	16049	2.626677578	http://www.phytozome.net/eucalyptus.php
Aquilegia coerulea	24823	41063	7487	0.301615437	23727	3.169093095	http://www.phytozome.net/aquilegia.php
Sorghum bicolor	27608	29448	1489	0.053933642	3329	2.235728677	http://www.plantgdb.org/SbGDB/
Zea mays	38914	63540	13137	0.337590584	37763	2.87455279	http://www.maizegdb.org/
Oryza sativa	55898	66338	6458	0.115531862	16809	2.60281821	http://rice.plantbiology.msu.edu/
Brachypodium distachyon	26552	31029	3201	0.12055589	7678	2.39862543	http://www.brachypodium.org/database

authenticity of each mRNA product, only the annotated mRNA were pick up and classified according to the number of mRNA variants into groups (Table 2). The correlative gene features of introns, coding sequence, exons, or untranslated regions (UTRs) were recorded. These features include distribution, location, length and number (Tables 1- 3 and Supplementary Table 1).

Generally, AS can be classified as IR, ES, AD and AA by the type of AS events. However, one AS gene variant might have multiple different AS types, simultaneously. This means that one AS gene can generate two, three, four, or more variants, all AS variants were classified into four groups on the basis of the number of AS variants: 2V (two variants), 3V (three variants), 4V (four variants), and 5V+ (five or more AS variants) (Table 2).

In total, 442,541 annotated genes from 589,034 mRNA products were obtained in the genome annotation of the 12 plant species (Figure 1 and Table 1). Our analysis identified 224,593 annotated mRNA variants. The events were distributed in 78,190 AS genes, which accounted for 17.67% of the total annotated genes (Table 1 and Supplementary Table 1). To eliminate the bias effect of the different species, we calculated the AS gene number and ratio (AS gene number/total gene number) for each species. We generated 3,063 genes in Populus (7.5%), 1,797 in Medicago (3.5%), 6,611 in soybean (14.3%), 5,804 in Arabidopsis (21.2%), 15,403 in cotton (41.1%), 7,630 in Theobroma (25.9%), 6,110 in Eucalyptus (16.8%),7,487 in Aquilegia (30.2%), 1,489 in sorghum (5.4%), 13,137 in maize (33.8%), 6,458 in rice (11.6%), and 3,201 in Brachypodium (12.6%) from the pooled data (Figure 1). A comparison of the AS variants among these species revealed that the AS gene ratios were higher in cotton and maize but lower in Populus, Medicago, and sorghum than in the other species (Figure 1 and Table 1).

Global analysis of gene structure

Gene structure analysis from a global perspective differentiates the architecture of plant and animal genes. In general, the composition of introns differs in plants and animals. Therefore, we evaluated the characteristics of gene structure to determine the putative differences in splicing among these plant species (Table 1). Our analysis revealed that the composition of introns differed in dicots and monocots, especially in terms of the number of introns. The average number of introns in the different dicotyledonous plants ranged from approximately 2.5 to 5.8 (Table 1). This wide range indicated their divergence despite the close phylogenetic relationship. Interestingly, each gene contained an average of four introns in the monocotyledonous plants. These numbers were relatively stable (Table 1). However, the average length of each intron ranged from 0.16 kb to 0.72 kb and greatly differed across the different species. We also found that more than 85% of the plant introns were located in the coding regions (Table 1 and Table 3).

Unlike the composition of introns, the average number and length of exons were highly similar in plants with subtle differences. These differences determined the sizes of the coding regions and proteins (Table 1). The above mentioned exon analysis was consistent with the similar average size of proteins in multiple species (Table 3). This finding explained why the different average gene sizes of various plant species generated similar protein sizes. In addition, the average sizes of UTRs (both 5′UTR and 3′UTR) differed among plant species. The length of the 3′UTR was much longer than that of the 5′UTR, except in Medicago (Table 1 and Table 3).

To gain clues about the features of multiple AS variants in 4 variant groups, we checked the assembled raw numbers and frequencies of AS among the different plant species. Group 2V had the most abundant AS variants, followed by groups 3V, 4V, and 5V+. A high proportion of AS variants in group 2V were generated from each species: Populus, 76.4%; Medicago, 70.8%; soybean, 71.7%; Arabidopsis, 73.2%; cotton, 39.3%; Theobroma, 51.9%; Eucalyptus, 62.5%; Aquilegia, 47.9%; sorghum, 81.8%; maize, 55.0%; rice, 63.8%; and Brachypodium, 72.0% (Figure 2 and Table 2). It is noteworthy that almost more than 50% pre-mRNAs which underwent alternative splicing events produced two variants in plants, whereas the proportions decrease with the increasing numbers of variants in other groups (Figure 2).

Examination of the characteristics of gene structure from different AS groups in plants showed that the average number of exons (ranged from 6.25 to 9.74) and introns (ranged from 4.65 to

Table 2: Classification of annotated mRNA according to the number of mRNA variants into groups.

Populus trichocarpa	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	40668	45033	2790.17	382.78	1151.35	4.98	231.14	3.98	347.09	31860	4358512	30548	8673721
AS gene	3063	7428	4461.6	435.25	1308.76	6.95	188.18	5.95	358.74	9286	1369251	8457	2592763
1V (1 variant)	37605	37605	2655.11	372.42	1120.26	4.59	243.98	3.59	343.28	22574	2989261	22091	6080958
2V (2 variants)	2341	4682	4381.25	438.89	1319.67	6.96	189.39	5.96	367.39	5257	729357	4956	1465027
3V (3 variants)	441	1323	4660.35	414.59	1243.76	6.97	178.52	5.97	357.08	1820	259825	1628	511954
4V (4 variants)	140	560	4693.83	421.74	1265.23	6.34	199.59	5.34	366.9	866	131871	674	202575
5V+ (5 and more variants)	141	863	4943.29	455.97	1367.9	7.22	189.62	6.22	311.62	1343	248198	1199	413207
Medicago truncatula	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	50962	53423	2108.34	243.25	732.74	3.41	214.77	2.41	412.87	28152	9491631	26769	8876325
AS gene	1797	4258	5024.56	328.08	987.23	6.5	151.71	5.5	413.07	6165	1160752	5475	1529566
1V (1 variant)	49165	49165	2002.78	235.9	710.69	3.14	226.05	2.14	412.83	21987	8330879	21294	7346759
2V (2 variants)	1272	2544	4808.59	332.91	1001.74	6.25	160.15	5.25	418.02	3473	658254	3065	866334
3V (3 variants)	416	1248	5483.05	323.74	971.22	6.77	143.66	5.77	415.19	1957	380253	1730	477710
4V (4 variants)	89	356	5565.34	324.37	973.12	7.39	131.72	6.39	364.38	554	84931	527	148381
5V+ (5 and more variants)	20	110	6816.85	277.38	832.15	6.29	132.36	5.29	463.6	181	37314	153	37141
Glycine max	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	46367	55787	3715.72	407.16	1224.49	5.93	206.26	4.93	423.71	43437	5511798	43367	13069327
AS gene	6611	16031	5137.82	381.42	1147.25	6.87	166.73	5.87	446.05	20035	2639875	18579	5918727
1V (1 variant)	39756	39756	3480.4	417.54	1255.63	5.56	225.96	4.56	412.11	23402	2871923	24788	7150600
2V (2 variants)	4738	9476	5156.83	410.17	1233.52	6.9	178.43	5.9	455.25	11083	1465272	10617	3365542
3V (3 variants)	1265	3795	5077.25	356.06	1068.18	6.93	154.37	5.93	432.93	4969	654303	4553	1460803
4V (4 variants)	417	1668	5118.5	329.68	989.04	6.72	147.34	5.72	457.24	2353	323642	2065	639972
5V+ (5 and more variants)	191	1092	5109.7	299.05	897.15	6.61	135.83	5.61	392.68	1630	196658	1344	452410
Arabidopsis thaliana	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	27416	35386	2205.02	409.5	1231.49	5.57	220.87	4.57	156.9	34621	4123096	30634	6636007
AS gene	5804	13774	2967.51	441.42	1327.26	7.24	183.07	6.24	154.82	18585	2308121	15036	3362233
1V (1 variant)	21612	21612	2001.51	389.15	1170.45	4.51	259.57	3.51	159.24	16036	1814975	15598	3273774
2V (2 variants)	4251	8502	2904.55	448.91	1349.72	6.96	193.76	5.96	155.16	10653	1331586	9013	2018121
3V (3 variants)	1133	3399	3151.09	445.73	1337.2	7.78	171.92	6.78	155.08	4883	620999	3827	830656
4V (4 variants)	291	1164	3092.06	408.93	1226.78	7.35	167.08	6.35	153.96	1829	221557	1371	325446
5V+ (5 and more variants)	129	709	3149.03	384.34	1153.02	7.85	146.97	6.85	151.44	1220	133979	825	188010
Gossypium raimondii	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	37505	77267	3242.49	426.55	1282.64	6.29	203.66	5.29	341.58	89649	16338798	86495	27777027
AS gene	15403	55165	4677.3	462.46	1390.38	7.47	185.83	6.47	341.18	73711	13404603	70032	23251420
1V (1 variant)	22102	22102	2244.26	336.91	1013.74	3.35	302.98	2.35	344.31	15938	2934195	16463	4525607
2V (2 variants)	6049	12098	3980.47	436.11	1311.33	5.93	220.81	4.93	332.68	14143	2587001	13674	4400916
3V (3 variants)	3566	10698	4611.7	451.9	1355.71	6.8	199.56	5.8	339.25	13589	2510033	12991	4325895
4V (4 variants)	2213	8852	5116.82	471.31	1413.92	7.57	186.92	6.57	343.62	11710	2144201	11288	3785673

5V+ (5 and more variants)	3575	23517	5649.75	477.49	1432.46	8.53	167.98	7.53	343.92	34269	6163368	32079	10738936
Theobroma cacao	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	29452	44404	6091.33	437.57	1315.71	5.97	220.4	4.97	493.04	60128	11290078	56959	18099236
AS gene	7630	22582	8624.45	535.06	1608.17	8.15	197.07	7.15	449.92	31934	5924456	31173	9963478
1V (1 variant)	21822	21822	5206.98	336.68	1013.05	3.7	273.54	2.7	611.06	28194	5365622	25786	8135758
2V (2 variants)	3963	7926	7785.43	467.22	1404.66	6.76	207.64	5.76	459.79	10430	1960962	9731	3232738
3V (3 variants)	1848	5544	8638.71	530.63	1591.88	8.25	193.17	7.25	455.75	7815	1455309	7354	2420859
4V (4 variants)	918	3672	8456.85	570.26	1710.78	9.24	185.22	8.24	415.2	5258	973926	5200	1695281
5V+ (5 and more variants)	901	5440	12456.36	614.65	1843.96	9.35	197.34	8.35	457.99	8431	1534259	8888	2614600
Eucalyptus grandis	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	36376	46315	3104.6	391.89	1178.68	5.18	227.58	4.18	423.72	37182	5850297	39285	12218689
AS gene	6110	16049	5253.45	428.5	1288.49	7.4	173.86	6.4	438.37	19483	3184895	20353	7133832
1V (1 variant)	30266	30266	2672	372.49	1120.46	4	280.32	3	407.12	17699	2665402	18932	5084857
2V (2 variants)	3818	7636	5019.08	440.13	1323.39	6.9	191.46	5.9	440.79	8640	1406841	9229	3198094
3V (3 variants)	1412	4236	5507.08	431.47	1294.4	7.68	168.73	6.68	432.69	5152	842774	5486	1934006
4V (4 variants)	499	1996	5431.48	392.02	1176.05	7.37	159.57	6.37	427.56	2601	426692	2649	970144
5V+ (5 and more variants)	381	2181	6428.97	415.38	1246.13	8.62	144.56	7.62	449.74	3090	508588	2989	1031588
Aquilegia coerulea	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	24823	41063	3560.47	435.8	1310.41	6.25	209.54	5.25	454.82	40850	6425333	41754	12527440
AS gene	7487	23727	5824.71	490.02	1473.05	8.17	179.96	7.17	460.53	30044	4978721	30126	9726927
1V (1 variant)	17336	17336	2584.03	361.6	1087.81	3.61	301.15	2.61	433.33	10806	1446612	11628	2800513
2V (2 variants)	3590	7180	4991.83	453.29	1362.87	6.51	208.99	5.51	467.93	8123	1203436	8315	2459966
3V (3 variants)	1765	5295	6176.44	501.17	1503.5	8.01	187.88	7.01	472.41	6590	1062144	6615	2114883
4V (4 variants)	969	3876	6683.13	491.79	1475.36	8.48	174.03	7.48	484.4	5100	846089	4983	1631428
5V+ (5 and more variants)	1163	7376	7146.66	516.84	1550.51	9.75	159.11	8.75	438.44	10231	1867052	10213	3520650
Sorghum bicolor	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	27608	29448	3226.51	418.56	1258.69	4.96	253.59	3.96	430.76	15843	2076458	16726	4738354
AS gene	1489	3329	4706.41	411.01	1236.02	7.07	174.55	6.07	393.37	4267	616662	3658	1110048
1V (1 variant)	26119	26119	3143.2	419.53	1261.58	4.69	268.77	3.69	438.59	11576	1459796	13068	3628306
2V (2 variants)	1218	2436	4593.88	414.39	1246.17	6.84	181.87	5.84	397.34	2943	419295	2658	784846
3V (3 variants)	213	639	5097.32	402.48	1207.44	7.85	153.95	6.85	381.36	911	131649	716	229618
4V (4 variants)	43	172	5509.35	378.39	1135.17	7.1	159.88	6.1	412.07	285	43535	185	62900
5V+ (5 and more variants)	15	82	5991.4	445.29	1335.88	7.71	173.4	6.71	350.72	128	22183	99	32684
Zea mays	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	38914	63540	4074	332.02	999.07	4.6	216.96	3.6	634.05	71771	13942080	68219	20414673
AS gene	13137	37763	6029.87	333.26	1002.77	5.37	186.4	4.37	622.83	51975	10331181	48954	14727072
1V (1 variant)	25777	25777	3129.56	330.22	993.65	3.47	286.18	2.47	663.12	19796	3610899	19265	5687601
2V (2 variants)	7220	14440	5503.46	363.23	1092.69	5.11	213.58	4.11	637.92	16299	3227319	15939	4992829
3V (3 variants)	3045	9135	6492.13	338.13	1014.39	5.5	184.65	4.5	640.23	12345	2500623	12000	3636922

4V (4 variants)	1476	5904	6397.92	309.67	929.01	5.54	167.84	4.54	596.55	9065	1798140	8251	2451850
5V+ (5 and more variants)	1396	8284	7354.95	292.45	877.34	5.56	157.94	4.56	598.83	14266	2805099	12764	3645471
Oryza sativa	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	55801	66153	2963.58	382.78	1151.35	4.4	304.6	3.4	381.81	44750	8497020	42193	15794083
AS gene	6457	16809	4851.11	435.25	1308.76	6.26	195.99	5.26	378.58	25122	5139835	22903	8983107
1V (1 variant)	49344	49344	2717.72	372.42	1120.26	3.77	365.94	2.77	383.9	19628	3357185	19290	6810976
2V (2 variants)	4120	8240	4610.08	438.89	1319.67	6.07	205.93	5.07	384.05	11136	2204572	10717	4158976
3V (3 variants)	1450	4350	5106.05	414.59	1243.76	6.42	191.24	5.42	378.1	6498	1314676	6016	2366342
4V (4 variants)	534	2136	5409.58	421.74	1265.23	6.37	183.66	5.37	381.07	3446	743922	3077	1170560
5V+ (5 and more variants)	353	2083	5772.32	455.97	1367.9	6.53	181.54	5.53	357.2	4042	876665	3093	1287229
Brachypodium distachyon	Gene Number	Gene model	The average size of Gene	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number	The average size of Intron	The number of 5'-UTR	The total length of 5'-UTR	The number of 3'-UTR	The total length of 3'-UTR
Total	26552	31029	3580.54	425.96	1280.87	5.39	237.49	4.39	385.02	14174	1763666	24808	7639899
AS gene	3201	7678	4806.08	437.67	1316.02	7.82	168.11	6.82	367.34	6716	866548	9294	3030001
1V (1 variant)	23351	23351	3413.68	422.1	1269.31	4.59	276.31	3.59	396.06	7458	897118	15514	4609898
2V (2 variants)	2304	4608	4745.38	449.03	1350.09	7.68	175.63	6.68	373.04	3583	459424	5433	1778574
3V (3 variants)	644	1932	4960.34	443.49	1330.46	8.29	160.6	7.29	349.46	1795	229394	2382	765742
4V (4 variants)	178	712	4716.65	404.72	1214.15	7.7	157.85	6.7	348.61	781	107905	904	301453
5V+ (5 and more variants)	75	426	5558.56	343.56	1030.68	7.38	139.68	6.38	428.28	557	69825	575	184232

Table 3: The correlative gene features of 12 plant species used in this study.

Species	Gene Number	Gene Models	AS gene Number	Percentage	The average size of Protein	The average size of CDS	Exon Number	The average size of Exon	Intron Number in CDS	Intron number in in 5'UTR	Intron Number in 3'UTR	Intron Number	The average size of Intron	The average size of Gene	The average size of 5'UTR
Populus trichocarpa	40668	45033	3063	0.08	382.72	1151.17	4.98	231.16	3.98 0.14	0.07	4.19	352.53	2916.61	96.78	192.61
Medicago truncatula	50962	53423	1797	0.04	243.21	732.62	3.41	214.77	2.41 0.11	0.06	2.58	430.47	2187.65	177.67	166.15
Glycine max	46367	55788	6611	0.14	421.96	1268.87	5.93	213.82	4.93 0.17	0.08	5.18	427.24	3816.24	98.8	234.27
Arabidopsis thaliana	27416	35386	5804	0.21	409.28	1230.85	5.57	220.91	4.57 0.21	0.07	4.86	164.86	2335.51	116.52	187.53
Gossypium raimondii	37505	77267	15403	0.41	426.04	1281.12	6.29	203.66	5.29 0.31	0.22	5.82	348.92	3882.58	211.46	359.49
Theobroma cacao	29452	44404	7630	0.26	437.26	1314.79	5.97	220.42	4.97 0.38	0.31	5.66	720.68	6058.62	254.26	407.6
Eucalyptus grandis	36376	46315	6110	0.17	391.72	1178.15	5.18	227.6	4.18 0.17	0.13	4.47	433.39	3504.62	126.32	263.82
Aquilegia coerulea	24823	41063	7487	0.3	435.41	1309.24	6.25	209.54	5.25 0.25	0.19	5.69	466.77	4424.45	156.48	305.08
Sorghum bicolor	27608	29448	1489	0.05	418.55	1258.64	4.96	253.62	3.96 0.10	0.05	4.11	440.47	3298.84	70.51	160.91
Zea mays	38914	63540	13137	0.34	331.67	998	4.21	236.86	3.21 0.32	0.26	3.79	710.91	4236.29	219.42	321.29
Oryza sativa	55898	66338	6458	0.12	446.34	1342.01	4.4	304.86	3.40 0.18	0.12	3.7	401.36	3194.44	128.09	238.09
Brachypodium distachyon	26552	31029	3201	0.12	425.82	1280.45	5.39	237.49	4.39 0.09	0.09	4.57	452.57	3652.46	56.84	246.22

9.64) in AS genes were more than the average number of the total annotated genes in each genome, but the average lengths of these exons and introns were slightly shorter (Figure 2 and Table 2). Additionally, the average sizes of UTRs were longer in the genes that generated multiple AS variants. These results implied the presence of a relationship between gene structure and AS. However, the total length of annotated genes did not significantly change (Table 2).

Distribution and frequency of AS variants

The above-mentioned analysis suggested that gene features such as the number and length of introns can influence AS variants. To further explore other genomic features, we performed correlation analysis to investigate the relationship between the AS variants and gene density of higher plants. The distributions of the AS variants in the four groups were pooled in 1 Mb contiguous regions across the chromosome where the genes are located (Figure 3). The analysis revealed extensive transcriptional activity in the chromosomes of the different plant genomes. The AS variants were principally located in the chromosome arms and were associated with gene distribution. As expected, the AS variants in group 2V that matched with multiple locations were predominantly mapped to the most transcribed chromosomal regions. These results were consistent with the high number of group 2V variants in all AS genes (Figure 3).

A comparison of the AS variants to the annotated plant genome showed that most of the AS variants that matched the genome can be mapped to annotated genic regions in plant species. However,

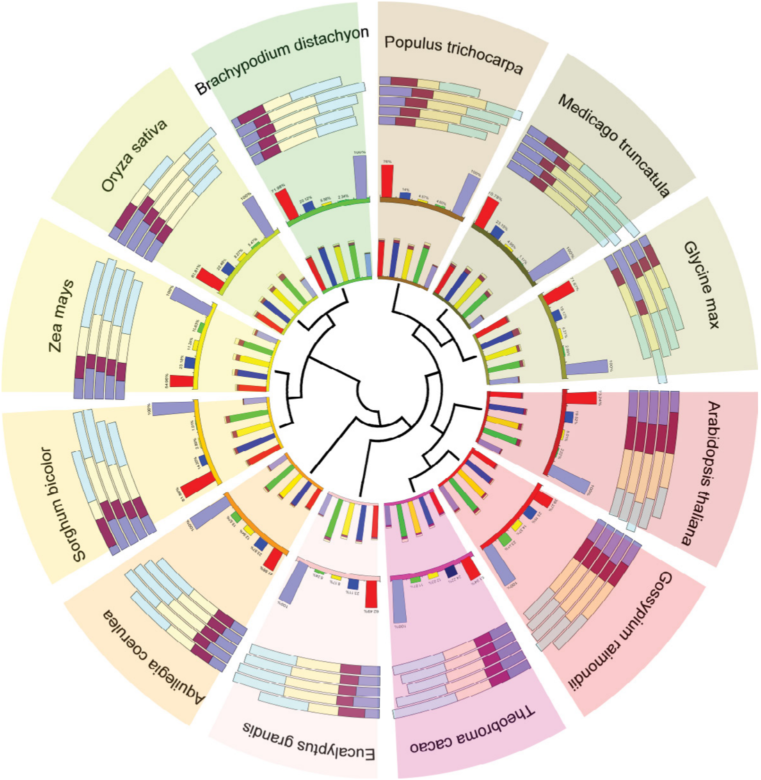


Figure 2: Features of multiple AS variants in plants.
From outside to inside, each circle represents the gene model of multiple AS variants, the distribution of AS variants along annotated genes features, and the proportion of multiple AS variants in plants. See also the Supplemental section (for details including the gene structure, location, number, length, and number of AS events).

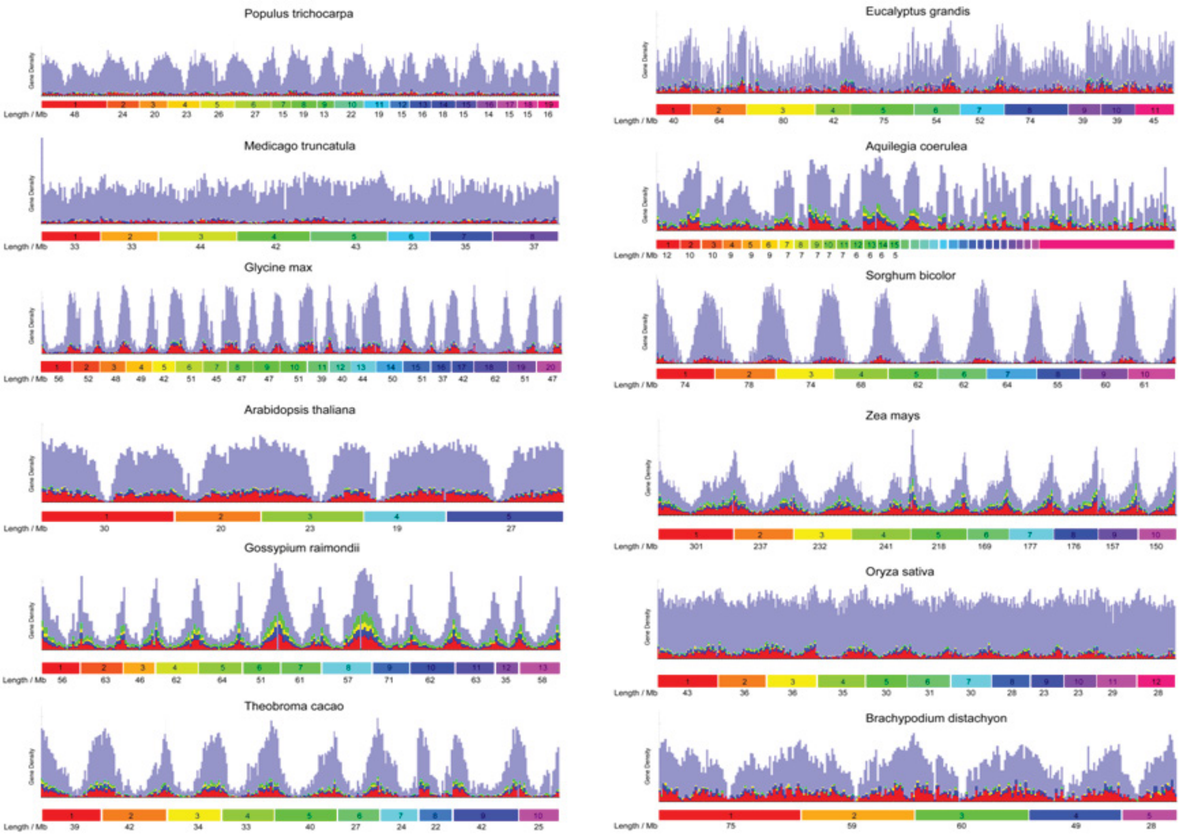


Figure 3: Statistics of multiple AS variants and AS distribution with gene density along chromosome length in plants.
Each vertical bar represents the frequency of an annotated gene in the plant genome. A schematic of each chromosome and its features is shown below. Different colors represent different groups on the basis of the number of AS variants. Red, two variants; deep blue, three variants; yellow, four variants; green, five or more variants.

approximately 50% of the AS variants cannot be mapped to the right chromosome in Aquilegia (Figure 3). Two reasons may explain this discrepancy. First, the depth of coverage of the Aquilegia genome was lower than the gene location coverage. Second, information regarding the annotated genes in each particular chromosome was limited. Overall, the distribution of multiple AS variants significantly positively correlated with gene density. Similar correlation patterns were observed in all plant species.

Multiple AS variants exert different effects toward improving proteome diversity

AS variants improve gene representation at the transcriptional and proteomic levels. The overall proteome from all the pooled data sets contained 211,558 protein isoforms, whereas each AS gene generated 2.87 mRNA variants and 2.71 protein variants. The number of variants per AS gene was not significantly different among the different plant species (Figure 1 and Table 1). These data indicated that AS enhanced transcriptome plasticity and proteome diversity in higher plants.

To explore the effects of multiple AS variants on proteome diversity, we determined the number of protein isoforms that were generated by the AS variants in the four groups. Intriguingly, the proportion of the effective AS events to generate various proteins was the highest in 2V group, but not in 4V or 5V+. Within group 2V, approximately 30% AS genes generated only one protein product, and similar proportions were observed for all plant species, except in Medicago (Figure 4). By contrast, approximately half of the genes from group 3V generated three different proteins, whereas the remaining genes produced one or two proteins (except in Medicago and maize) (Figure 4). The proportion of effective AS events noticeably increased with the number of mRNA variants. The observed proportion was even higher in groups 4V and 5V+ (Figure 4). In brief, a high proportion of the annotated genes which generated the isofroms numbers less than mRNA variants were found, indicating that the effects of AS on the improvement of proteome diversity were lower than previously thought, especially in groups 3V, 4V, and 5V+. The AS variants in the different AS groups exerted distinct influences on proteome diversity (Figure 4).

DISCUSSION

In this study, the total number and ratio of AS were below those of AS in previous studies on some species. Among the multi-exonic genes, approximately 42% to 61% of those in Arabidopsis, approximately 56% of those in maize, and approximately 63% of those in soybean underwent AS. Two possible reasons may explain this result. First, several mRNA variants occurred in specific developmental states might have triggered in response to environmental conditions. Second, the basic information of the annotated gene in some species would be limited. Therefore,

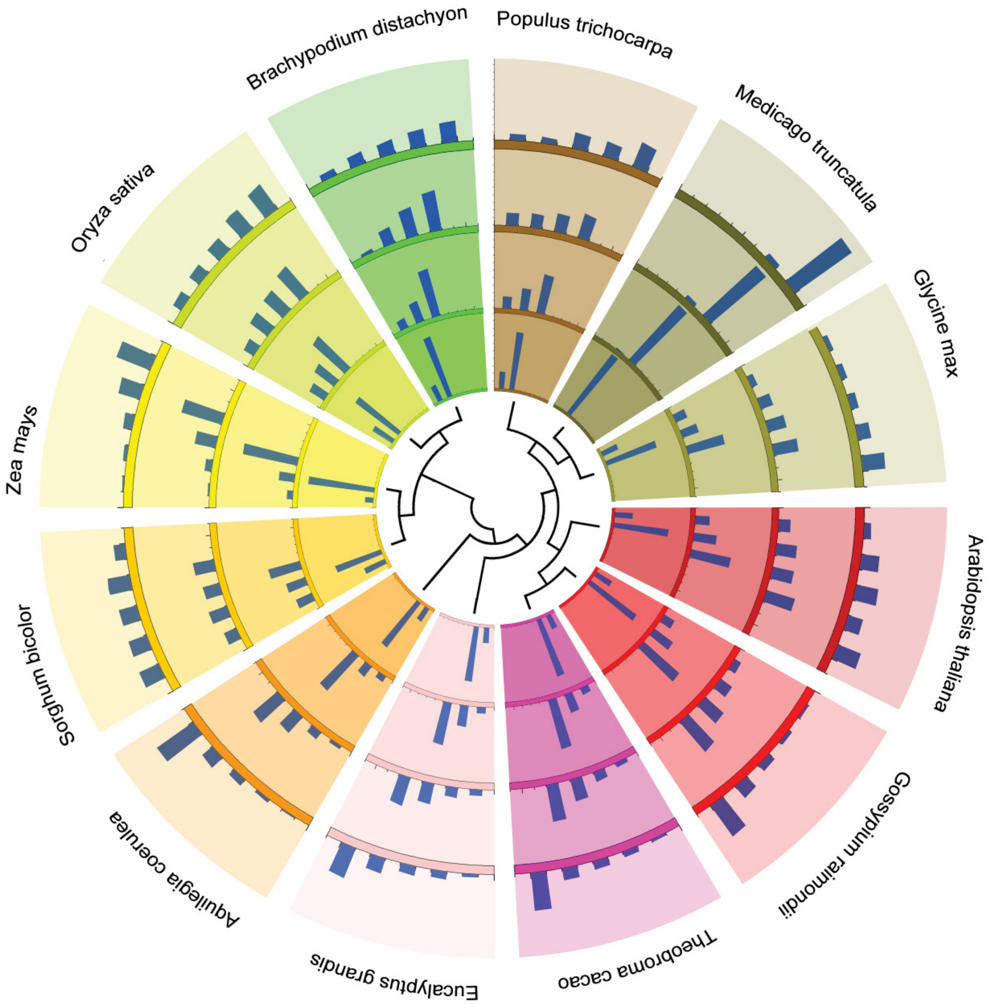


Figure 4: Comparison of studies on proteome diversity with multiple AS variants. From outside to inside, each circle represents the classification of AS as defined by the number of AS variants.

alignment quality, low transcript abundance, and transcriptome coverage directly influenced the number of annotated AS genes.

We noticed that the features of multiple AS variants in 4 variant groups differed in most plants food and horticultural. With the increasing numbers of variants, the AS genes contained more but shorter exons and introns, implying AS might improve gene representation by changing the extent and complexity of gene structure (Figure 2 and Table 2). We speculated that these trends may be related to the evolution of various species from a long-term perspective.

Our analysis suggested that a high proportion of AS events might occur in the UTRs. We hypothesized that a large proportion of the AS events in the UTRs plays crucial roles in gene regulation [33,34]. Growing evidence has indicated that splicing is a regulatory mechanism for noncoding RNA [35]. In a recent study, we found a stress-induced AS event in the 5′UTR that influences intronic miRNA expression in Arabidopsis [21,36]. However, we also found a few introns in the UTRs (both 5′UTR and 3′UTR). Further investigation should be conducted to determine whether or not the AS variants are governed by the same regulators in the UTRs. In summary, we investigated the relationship between the number of AS mRNA variants and their effects on gene features at the genome-wide level through a comprehensive comparison of 12 plant species, as a valuable resource for functional research on food crops and horticultural crops. The present reanalysis and reclassification of previously reported RNA-seq data sets evaluated the global existence of AS variants in plants and suggested the need for future analyses on mRNA variants [37,38].

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation (grant numbers 31570271 and 31870234).

AUTHORS' CONTRIBUTIONS

Ying Li and Kang Yan performed the research and wrote the manuscript. Shizhong Zhang, Qianhuan Guo and Chengchao Zheng conceived of the study, and participated in its design and contributed to revisions of the manuscript. All authors approved the final version of the manuscript.

REFERENCES

1. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. 2010;20(1):45-58.
2. Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. Rna. 2008;14(5):802-813.
3. Lareau LF, Green RE, Bhatnagar RS, Brenner SE. The evolving roles of alternative splicing. Curr Opin Struct Biol. 2004;14(3):273-282.
4. Brown JW, Simpson CG. Splice Site Selection in Plant PRE-mRNA Splicing. Annu Rev Plant Biol. 1998;49:77-95.
5. Lorkovic ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W. Pre-mRNA splicing in higher plants. Trends in plant science. 2000;5(4):160-167.
6. Reddy AS. Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev plant Biol. 2007;58:267-294.
7. Moore MJ, Proudfoot NJ. Pre-mRNA processing reaches back to transcription and ahead to translation. Cell. 2009;136(4):688-700.
8. Ramani AK, Calarco JA, Pan Q, Mavandadi S, Wang Y, Nelson AC, et al. Genome-wide analysis of alternative splicing in Caenorhabditis elegans. Genome Res. 2011;21(2):342-328.
9. Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ. Dynamic integration of splicing within gene regulatory pathways. Cell. 2013;152(6):1252-1269.
10. Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. Epigenetics in alternative pre-mRNA splicing. Cell. 2011;144(1):16-26.
11. Sugliani M, Brambilla V, Clerkx EJ, Koornneef M, Soppe WJ. The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. The Plant cell. 2010;22(6):1936-1946.
12. Mach J. Alternative splicing produces a JAZ protein that is not broken down in response to jasmonic acid. The Plant cell. 2009;21(1):14.
13. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(1):189-92.
14. Modrek B, Lee C. A genomic view of alternative splicing. Nature genetics. 2002;30(1):13-9.
15. Drechsel G, Kahles A, Kesarwani AK, Stauffer E, Behr J, Drewe P, et al. Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. The Plant cell. 2013;25(10):3726-3742.
16. Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, et al. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. Nucleic Acids Res. 2012;40(6):2454-2569.
17. Rogers K, Chen X. Biogenesis, turnover, and mode of action of plant microRNAs. The Plant cell. 2013;25(7):2383-2399.
18. Jia F, Rock CD. MIR846 and MIR842 comprise a cistronic MIRNA pair that is regulated by abscisic acid by alternative splicing in roots of Arabidopsis. Plant Mol Biol. 2013;81(4-5):447-460.
19. Szarzynska B, Sobkowiak L, Pant BD, Balazadeh S, Scheible WR, Mueller-Roeber B, et al. Gene structures and processing of Arabidopsis thaliana HYL1-dependent pri-miRNAs. Nucleic Acids Res. 2009;37(9):3083-93.
20. Hirsch J, Lefort V, Vankersschaver M, Boualem A, Lucas A, Thermes C, et al. Characterization of 43 non-protein-coding mRNA genes in Arabidopsis, including the MIR162a-derived transcripts. Plant Physiol. 2006;140(4):1192-204.
21. Yan K, Liu P, Wu CA, Yang GD, Xu R, Guo QH, et al. Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in Arabidopsis thaliana. Mol Cell. 2012;48(4):521-531.

22. Wang Y, Itaya A, Zhong XH, Wu Y, Zhang JF, van der Knaap E, et al. Function and Evolution of a MicroRNA That Regulates a Ca2+-ATPase and Triggers the Formation of Phased Small Interfering RNAs in Tomato Reproductive Growth. The Plant cell. 2011;23(9):3185-3203.

23. Ben Abdallah RA, Jabnoun-Khiareddine H, Nefzi A, Daami-Remadi M. Evaluation of the Growth-Promoting Potential of Endophytic Bacteria Recovered from Healthy Tomato Plants. J Horticul. 2018;05(02).

24. Chen FC, Wang SS, Chaw SM, Huang YT, Chuang TJ. Plant Gene and Alternatively Spliced Variant Annotator. A plant genome annotation pipeline for rice gene and alternatively spliced variant identification with cross-species expressed sequence tag conservation from seven plant species. Plant Physio. 2007;143(3):1086-1095.

25. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics. 2008;40(12):1413-1415.

26. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome research. 2012;22(6):1184-1195.

27. Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, et al. The developmental dynamics of the maize leaf transcriptome. Nature Gen. 2010;42(12):1060-1067.

28. Li Y, Dai C, Hu C, Liu Z, Kang C. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. Plant J. 2017;90(1):164-176.

29. Verma S, Whitaker V. A New Technology Enabling New Advances in Strawberry Genetics. J Horticul. 2016;3(2).

30. Sun Y, Hou H, Song H, Lin K, Zhang Z, Hu J, et al. The comparison of alternative splicing among the multiple tissues in cucumber. BMC Plant Biol. 2018;18(1):5.

31. Shen Y, Zhou Z, Wang Z, Li W, Fang C, Wu M, et al. Global dissection of alternative splicing in paleopolyploid soybean. The Plant cell. 2014;26(3):996-1008.

32. Chang CY, Lin WD, Tu SL. Genome-wide analysis of heat-sensitive alternative splicing in Physcomitrella patens. Plant Physiol. 2014.

33. Grozeva S. Effect of copper levels in the culture medium on shoot regeneration in pepper. Banats J Biotechnol. 2015;6(12):86-91.

34. Barazesh F, Oloumi H, Nasibi F, Kalantari KM. Effect of spermine, epibrassinolid and their interaction on inflorescence buds and fruits abscission of pistachio tree (Pistacia vera L.), "Ahmad-Aghai" cultivar, Banats J Biotechnol. 2017;8(16):105-115.

35. Idris A. Comparative analysis of 16SrRNA genes of Klebsiella isolated from groundnut and some american type culture collections, Banats J Biotechnol. 2016;7(13):34-40.

36. Hariri Moghadam F, Khalghani J, Moharramipour S, Gharali B, Mostashari Mohasses. Investigation of the induced antibiosis resistance by zinc element in different cultivars of sugar beet to long snout weevil, Lixus incanescens (Col: Curculionidae), Banats J Biotechnol. 2018;9(17):5-12.

37. Ayadi Hassan S, Belbasi Z. Improvemnet of hairy root induction in Artemisia annua by various strains of agrobacterium rhizogenes, Banats J Biotechnol. 2017;8(15):25-33.

38. Kosev VI. Multivariate analysis of spring field pea genotypes. Banats J Biotechnol. 2015;6(11):23-29.