The *Ds1* Transposon Provides Messages That Yield Unique Profiles of Protein Isoforms and Acts Synergistically With *Ds* to Enrich Proteome Complexity via Exonization

Evolutionary Bioinformatics
1–11
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176934317690410
SAGE

Yuh-Chyang Charng, Lung-Hsin Hsu and Li-yu Daisy Liu

Department of Agronomy, National Taiwan University, Taipei, Taiwan.

ABSTRACT: In exonization events, *Ds1* may provide donor and/or acceptor sites for splicing after inserting into genes and be incorporated into new transcripts with new exon(s). In this study, the protein variants of *Ds1* exonization yielding additional functional profile(s) were studied. Unlike *Ds* exonization, which creates new profiles mostly by incorporating flanking intron sequences with the *Ds* message, *Ds1* exonization additionally creates new profiles through the presence or absence of *Ds1* messages. The number of unique functional profiles harboring *Ds1* messages is 1.3-fold more than that of functional profiles without *Ds1* messages. The highly similar 11 protein isoforms at a single insertion site also contribute to proteome complexity enrichment by exclusively creating new profiles. Particularly, *Ds1* exonization produces 459 unique profiles, of which 129 cannot be built by *Ds*. We thus conclude that *Ds* and *Ds1* are independent but synergistic in their capacity to enrich proteome complexity through exonization.

KEYWORDS: *Ds1* transposon, exonization, alternative splicing, nonsense-mediated decay pathway

RECEIVED: October 4, 2016. **ACCEPTED:** December 7, 2016.

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 621 words, excluding any confidential comments to the academic editor.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by Ministry of Science and Technology in Taiwan (grant no. MOST 105-2321-B-002-037).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Li-yu Daisy Liu, Department of Agronomy, National Taiwan University, No.1, Sec. 4, Roosevelt Rd., Taipei 106, Taiwan. Email: lyliu@ntu.edu.tw

Introduction

Evolution and speciation are believed to be driven in large part by the insertion of transposable elements (TEs) within eukaryotic genes.[1] The TEs can integrate into a gene with either a forward or reverse pattern according to the transcription directionality of the inserted gene and the transposase gene. When inserting into the exons of genes, TEs may disrupt and cause the loss of function of genes into which they are inserted. However, the insertion of TEs within the intronic sequences of a given gene can also have the effect of altering pre–messenger RNA through alternative splicing (AS) and/or exonization.[2] In such instances, the AS that results from TE insertion may cause interference with the normal splicing of the inserted gene's transcribed region, whereas the exonization results in a cryptic splice site of the inserted TE to generate a new exon of the inserted gene. Any additional variant due to AS or exonization may subsequently evolve to form a protein with new functions. Moreover, the operation of natural selection may even serve to enhance the novel splice sites and, in turn, to raise the production level of the new variant if it is advantageous.[3] Alternative splicing is commonly seen in higher eukaryotes, but its role in expanding the functions of the plant proteome is limited.[4] In contrast, exonization can potentially introduce a portion or portions of a TE into the resulting transcripts, thereby altering the reading frames so as to enhance the complexity of proteomes, as was found, for example, in our previous study involving the insertion of a mini *Ds* transposon into the modified tobacco marker gene *epsps*.[5]

The *Ac/Ds* system includes the first TEs recognized in the scientific literature, having been identified by Barbara McClintock 60 years ago. *Ds* transposons of the nonautonomous (transposase defective) variety consist of 11 bp terminal inverted repeats as well as approximately 250 bp of both ends (ie, terminal regions) of their full form transposon, *Activator* (*Ac*).[6] There are 3 different types of *Ds* elements, namely, *Ds*, *Ds1*, and *Ds2*. *Ds1* has 13 bp at the 5′ terminal and 26 bp at the 3′ terminal in common with *Ac*, whereas the internal region of *Ds1* is not homologous to *Ac*.[7] *Ds1* can be mobilized not only by *Ac* but also by another TE, *Uq*, which does not *trans*-activate *Ds* elements of the *Ds* family.[8]

Ds and *Ds1* are identical for the first 19 bp containing 2 discontinuous but in-framed premature termination codons (PTCs; Figure 1). In addition, *Ds* and *Ds1* are both biased toward providing splice donor and/or acceptor sites located close to their terminal regions.[5,9] *Ds* provides only donors (1 forward and 4 reverse insertion patterns), whereas *Ds1* provides 3 donors and 2 acceptors associating with 11 possible exonizing patterns (all in reverse insertion patterns).[10,11] These different features between *Ds* and *Ds1* imply possibly independent evolutionary impacts of *Ds* and *Ds1* induced through exonization. To investigate their roles from an evolutionary perspective, we have simulated all putative *Ds*-exonized as well as *Ds1*-exonized transcripts in the rice genome.[10,11] The exonized transcripts were translated into proteins and characterized as C-terminal variants (ie, those for which the C-terminus of the

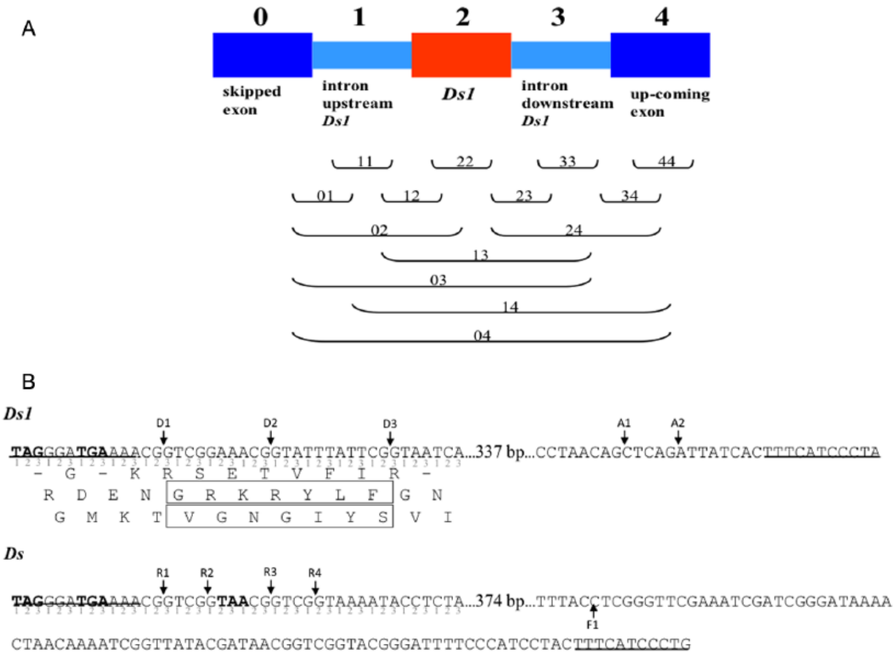
A
0
1
2
3
4
skipped exon
intron upstream Ds1
Ds1
intron downstream Ds1
up-coming exon
11
22
33
44
01
12
23
34
02
24
13
03
14
04
B
Ds1
D1
D2
D3
A1
A2
TAGGGATGAAAACGGTCGGAAACGGTATTTATTCGGTAATCA...337 bp...CCTAACAGCTCAGATTATCACTTTCATCCCTA
- G - K R S E T V F I R -
R D E N G R K R Y L F G N
G M K T V G N G I Y S V I
Ds
R1
R2
R3
R4
TAGGGATGAAAACGGTCGGTAACGGTCGGTAAAATACCTCTA...374 bp...TTTACCTCGGGTTCGAAATCGATCGGGATAAAA
F1
CTAACAAAATCGGTTATACGATAACGGTCGGTACGGGATTTTCCCATCCTACTTTCATCCCTG

Figure 1. (A) Classification of the profiles built by *Ds1* and its flanking sequences. (B) *Ds1* and *Ds* sequences yielding splice acceptor (A) or donor (D) junctions (arrows) as well as premature termination codons in exonized transcripts (bold). The translated products of *Ds1* are also shown, of which gains/losses of 7 amino acids (boxes) yielded using D3/D1 were important to compose functional profiles. The termini repeat sequences of each transposable element are underlined. Note that *Ds* could provide 5 donors, R1, R2, R3, R4, and F1, but no acceptor. The donor, F1, is used for exonization by the opposite insertion pattern.

reference protein was replaced by the output peptides) and interior variants (ie, those for which additional peptides were inserted in the middle of the original transcript even as the same original termination codon was used). We also performed a functional profile analysis based on the PROSITE database[12] and revealed the possibility of proteome enrichment by *Ds* exonization.[13]

In this study, we investigated the behavior of *Ds1* in exonization and compared it with that of *Ds*. A protocol similar to that used by Chien et al,[13] with some modifications (see "Materials and Methods" section), was applied to the *Ds1*-exonized protein variants. We found that the *Ds1*-exonized messages with no more than 59 nucleotides were actively involved in creating diverse functional profiles. In particular, protein variants exonized from a single *Ds1* insertion site with 2-amino-acid differences corresponded to 18 different functional profiles. Although *Ds1* and *Ds* can build functional profiles per intron with similar efficiencies, *Ds1* exonization produces 459 unique profiles, of which 129 are not produced by *Ds*. We thus conclude that *Ds* and *Ds1* are independent but synergistic in their capacity to enrich proteome complexity through exonization.

Materials and Methods

The *Ds1*-exonized transcripts of rice were constructed previously.[11] In this study, open reading frame (ORF) analysis was conducted for all of these exonized transcripts beginning at the original start codon and ending at the first in-frame stop codon. The transcripts were categorized as type I, type II, type III, or type IV transcripts depending on, respectively, whether the in-frame stop codon was located at the conserved region of the original splice junction, at the intron inserted by *Ds1*, at the *Ds1* transposon itself, or at any exon occurring after the *Ds1* insertion. Furthermore, in the event that the ORF analysis revealed no in-frame stop codon, the corresponding transcript was categorized as a type V transcript, and the incomplete transcript (ie, incomplete in that it lacked a stop codon) was output directly. In addition, the transcripts were further categorized as belonging to 1 of 2 subtypes: if the termination codon of the transcript was the same as that of the reference transcript (ie, the transcript without the *Ds1* insertion), it was categorized as an interior transcript; otherwise, it was categorized as a C-terminal transcript.

If a given transcript contained a termination codon that was located more than 55 nucleotides upstream from the last exon/exon junction, it was considered a potential target for the nonsense-mediated decay (NMD) pathway[14,15] and was therefore omitted from isoform prediction. The proteins of these transcripts not targeted by the NMD pathway were called non-NMD protein variants and were further translated to protein sequences. The original protein sequences and protein sequences from type III, IV, and V non-NMD variants were subject to protein profile analysis, in which we scanned the sequences to search for domains (profiles) previously reported in the PROSITE database (version 20.83).[12] The PROSITE database contains a total of 2442 entries that describe the various protein domains, families, and functional sites, in addition to the various amino acid patterns, profiles, and signatures contained within them.

Instead of using all the entries in PROSITE, we only scanned for the 1308 patterned ones to consistently identify the newly developed profiles using the same standard.

The functional profiles of each protein variant were compared with the ones of its reference protein. Only those variants yielding additional functional profile(s) were collected. As described in the text above, the new functional profiles in the protein isoforms were classified into subclasses named with 2 digits from "0" to "4" to indicate the start and the end of the amino acids from which the profiles originated, where "0" to "4" denoted the skipped exon, the flanking intron upstream from *Ds1, Ds1* itself, the flanking intron downstream from *Ds1*, and the upcoming exon, respectively (Figure 1A). Following the same logic, the profile names shown in the text combined the information of interior (I) or C-terminal (C) type as well as the *Ds1* portion used. For example, a class I22-D1 profile indicates a "22" profile observed in an interior functional variant (I) when *Ds1* provided the first donor (D1). Further analyses of the resulting protein variants in different types were conducted using R (version 2.15.1).[16]

Results and Discussion

New functional profiles are introduced by Ds1 alone or together with its flanking exons and introns

Functional profile analyses were performed on the previously simulated *Ds1*-exonized non-NMD protein variants in rice[11] according to the patterned-profile database in PROSITE.[12] From a total of 38 427 898 non-NMD variants, only 14 258 780 (4 303 236 interior and 9 955 544 C-terminal) variants yielding additional functional profile(s) were collected and termed as functional variants (Table 1 and Figure 2). There were 47.33% (6 748 622 of 14 258 780), 27.31% (3 893 991 of 14 258 780), and 25.36% (3 616 167 of 14 258 780) of the functional variants using an acceptor (A) alone, a donor (D) alone, or both a donor and an acceptor (DA) of *Ds1*, respectively (Table 1 and Figure 2). Most of the interior variants (54.59%) originated from *Ds1* providing donors only, whereas most of the C-terminal variants (55.36%) originated from *Ds1* providing acceptors.

The additional functional profiles yielded by the variants were further classified into subclasses named with 2 digits from "0" to "4" to indicate the start and the end of the amino acids from which the profiles originated, where "0" to "4" denoted the skipped exon, the flanking intron upstream from *Ds1, Ds1* itself, the flanking intron downstream from *Ds1*, and the upcoming exon, respectively (Figure 1A). For example, a class "02" profile indicates that the functional profile in question was made from amino acids combining the messages (sequences) of the skipped exon, flanking upstream intron, and *Ds1*; a "22" profile indicates a profile made from the *Ds1* message alone; and a "04" profile indicates a functional profile made from combining the messages all the way from the skipped exon to the upcoming exon. The "04" profiles were expected to be rare because 95% of the patterned profiles presented in PROSITE are composed of less than 30 amino acids. Indeed, "04" profiles accounted for only about 0.4% of the total profiles (Supplementary Table S1A and S1C). Following the same logic, a class I22-D1 profile is a "22" profile observed in an interior functional variant (I) when *Ds1* provided the first donor (D1). Supplementary Table S1C and S1D presents the numbers of unique profiles in all classes of interior and C-terminal variants, respectively. Some classes not existing by definition are labeled "N" in Supplementary Table S1A and S1B. For example, exonization caused using donors alone (D) cannot yield profiles starting/ending with "3" (the intron downstream from *Ds1*).

About 68% (4 975 488 out of 7 368 405) of the interior profiles (ie, additional profiles from interior functional variants) originated from *Ds1* providing donors only (Table 1 and Figure 3A). The classes I11-D1 and I11-D3 yielded the highest number of new profiles (870 463) composed of 59 unique ones (Supplementary Table S1A and Figure 3C). The number of D2 interior profiles was generally lower than the numbers of D1 and D3 interior profiles due to fact that the D2 variants share the same reading frame with the stop codon, TAG, in the upstream *Ds1* (Figure 1B). For C-terminal protein isoforms, 46.3% of the profiles originated from *Ds1* providing acceptors only (Table 1 and Figure 3B). The classes C33-A1 and C33-A2 created the most additional profiles composed of 113 and 120 unique ones, respectively (Supplementary Table S1B and SID and Figure 3D). A C-terminal isoform was constructed using a reading frame different from that of the reference protein, and therefore, a greater diversity of C-terminal profiles than interior profiles were expected. In fact, among the 1308 profiles defined in the PROSITE database, 459 unique profiles were observed in all isoforms in this study, where 334 and 399 unique profiles appeared only in interior variants and only in C-terminal variants, respectively.

On average, 1 variant included 2.52 new profiles (Table 1). The numbers and types of new functional profiles from D-variants (meaning the functional variants using a donor alone), A-variants (meaning the functional variants using an acceptor alone), and DA-variants (meaning the functional variants using both DA) were very different from each other (Figure 3). Most of functional profiles from A-variants resulted from C-terminal variants, for about 1.96 profiles per variant, but DA-variants yielded an average of 4.1 profiles per variant (Table 1). Interestingly, the classes yielding the most additional profiles were not necessarily the classes yielding the most unique profiles. For example, the profile number of class I12-D1 (204 859) was 4.3-fold less than that of I11-D1 (870 463), but more unique profiles were yielded by I12-D1 (111) than by I11-D1 (69). This implies the contribution of incorporated *Ds1* messages, even those consisting of merely 14 bp, for yielding variants with new functions.

Table 1. Number of variants, profiles, and profiles per variant yielded by each donor, acceptor, and donor-acceptor combination.

NUMBER OF VARIANTS				NUMBER OF PROFILES				NUMBER OF PROFILES PER VARIANT			
DS1	INTERIOR	C-TERMINAL	INTERIOR + C-TERMINAL	*DS1*	INTERIOR	C-TERMINAL	INTERIOR + C-TERMINAL	*DS1*	INTERIOR	C-TERMINAL	AVG.
D1	834 649	459 547	1 294 196	D1	1 649 507	1 604 133	3 253 640	D1	1.9763	3.4907	2.5140
D2	538 473	578 306	1 116 779	D2	1 259 036	2 234 207	3 493 243	D2	2.3382	3.8634	3.1280
D3	976 104	506 912	1 483 016	D3	2 066 945	1 821 524	3 888 469	D3	2.1175	3.5934	2.6220
Sub-total	2 349 226	1 544 765	3 893 991	Sub-total	4 975 488	5 659 864	10 635 352	Avg.	2.1440	3.6492	2.7547
D1A1	140 406	500 362	640 768	D1A1	386 510	1 652 797	2 039 307	D1A1	2.7528	3.3032	3.1826
D1A2	88 020	429 108	517 128	D1A2	213 532	1 315 830	1 529 362	D1A2	2.4259	3.0664	2.9574
D2A1	95 605	485 683	581 288	D2A1	288 212	1 729 509	2 017 721	D2A1	3.0146	3.5610	3.4711
D2A2	139 554	443 916	583 470	D2A2	793 056	1 613 625	2 406 681	D2A2	5.6828	3.6350	4.1248
D3A1	156 369	554 037	710 406	D3A1	460 560	1 870 458	2 331 018	D3A1	2.9453	3.3761	3.2812
D3A2	96 485	486 622	583 107	D3A2	250 906	1 503 768	1 754 674	D3A2	2.6005	3.0902	3.0092
Sub-total	716 439	2 899 728	3 616 167	Sub-total	2 392 776	9 685 987	12 078 763	Avg.	3.2370	3.3387	3.3377
A1	653 001	2 911 994	3 564 995	A1	92	6 951 881	6 951 973	A1	0.0001	2.3873	1.9501
A2	584 570	2 599 057	3 183 627	A2	49	6 280 165	6 280 214	A2	0.0001	2.4163	1.9727
Sub-total	1 237 571	5 511 051	6 748 622	Sub-total	141	13 232 046	13 232 187	Avg.	0.0001	2.4018	1.9614
Total	4 303 236	9 955 544	14 258 780	Total	7 368 405	28 577 897	35 946 302	Avg.	1.7123	2.8706	2.5210

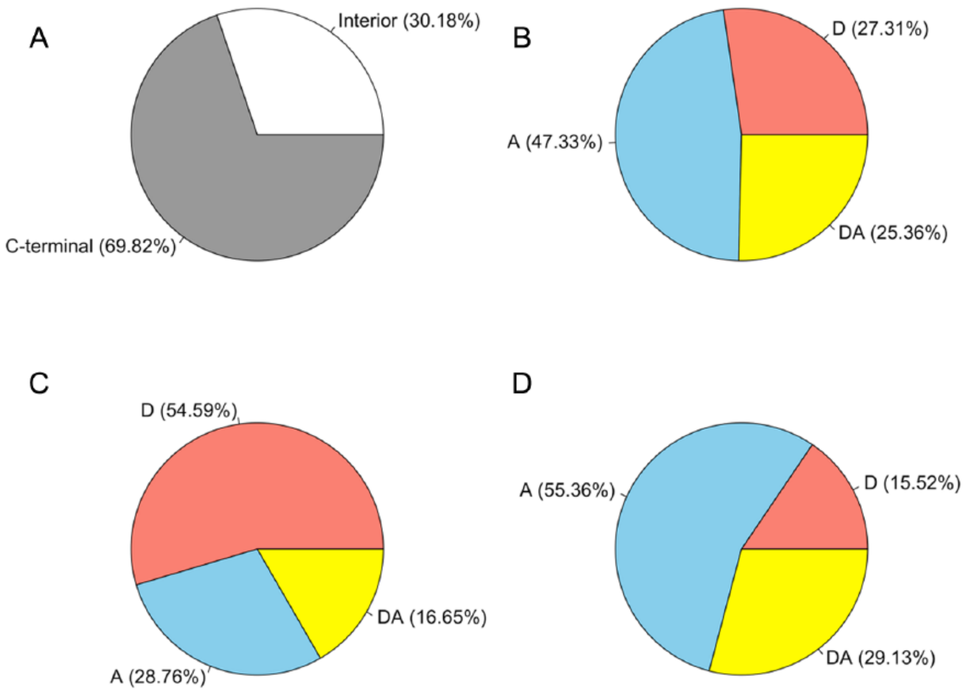

Figure 2. The proportions of functional variants: (A) all variants, (B) all variants, (C) interior variants, and (D) C-terminal variants. A indicates acceptor; D, donor; DA, donor and acceptor.

Exonized messages specifically from Ds1 contributed to new functional profiles

The functional profiles from classes "01," "11," "33," "34," and "44" only using the intron/exon sequences of the affected transcripts are *Ds1* independent, whereas the other classes are *Ds1* dependent. The number of *Ds1*-independent profiles (23 618 314) was about 1.9-fold higher than the number of *Ds1*-dependent ones (12 327 988). Because only a maximum of 59 bp from a *Ds1* message would be incorporated into the resulting protein isoforms, the major contribution of a TE to exonization was expected to be the incorporation of the message of TE-inserted intron of the affected transcripts rather than the message itself.[11] However, the *Ds1*-independent and *Ds1*-dependent profiles were composed of 314 and 398 unique profiles, respectively, with 253 overlapping profiles (Supplementary Table 2). This implies that the exonized *Ds1* message is important for building functional profiles for selective advantage, either via the *Ds1* message alone or together with its flanking intron/exon. There were 1 506 257 (20.44% of all the interior profiles) and 4 302 547 (25.58% of all the C-terminal profiles) class "22" interior and C-terminal profiles, respectively, built using the *Ds1* message alone. These abundant class "22" profiles were composed of only 6 unique profiles, PS00004, PS00005, PS00006, PS00007, PS00008, and PS00009 (the underlined profiles were also yielded by *Ds* alone). The remaining 139 unique *Ds1*-dependent profiles were therefore built using the message of *Ds1* together with its flanking intron/exon, and that number is still 2.3-fold (=139/61) (Supplementary Table 2) more than the number of unique *Ds1*-independent profiles.

Although *Ds1* may yield 11 exonized transcript isoforms at a single insertion site, the translated protein products might be similar to each other. For example, isoforms yielded by D2A1 and D3A2 differ from each other by a mere 2 amino acids (Figure 4). This feature seems to underestimate the contribution of *Ds1* exonization to proteome complexity. However, these small differences in isoforms from a single insertion site surprisingly contributed various new profiles, which was further illustrated using the two particular examples of (1) comparing isoforms using either D3 or D1 and (2) comparing isoforms using one of D1A1, D2A1, and D3A2.

The translated protein isoforms of D3 and D1 at the same insertion site differed by only 7 amino acids (either as "VGNGIYS" or "GRKRYLF" according to the reading frames) because the splice junction of D3 is located 21 bp downstream from that of D1 (Figure 1B). The 7-amino-acid sequences were responsible for 10 more unique profiles in class I12-D3 than in class I12-D1 (Figure 2). However, the lack of these 7-amino-acid sequences meant that class I14-D3 yielded only 21 unique profiles, less than the 78 yielded by class I14-D1. These results indicate that both the presence and absence of a part of the *Ds1* message would act positively for building unique profiles in exonization events. Table 2 shows the IDs of the profiles yielded by gaining or losing 7 amino acids. The presence or absence of the 7 amino acids caused by alternatives of the D1 or D3 sites of *Ds1* provided 133 unique profiles, of which 81 and 27 were exclusively yielded by D1 and D3, respectively. It is notable that the total number of profiles composing a class may not be equal to the number of unique profiles of that class shown in Figures 2 and 3 because messages

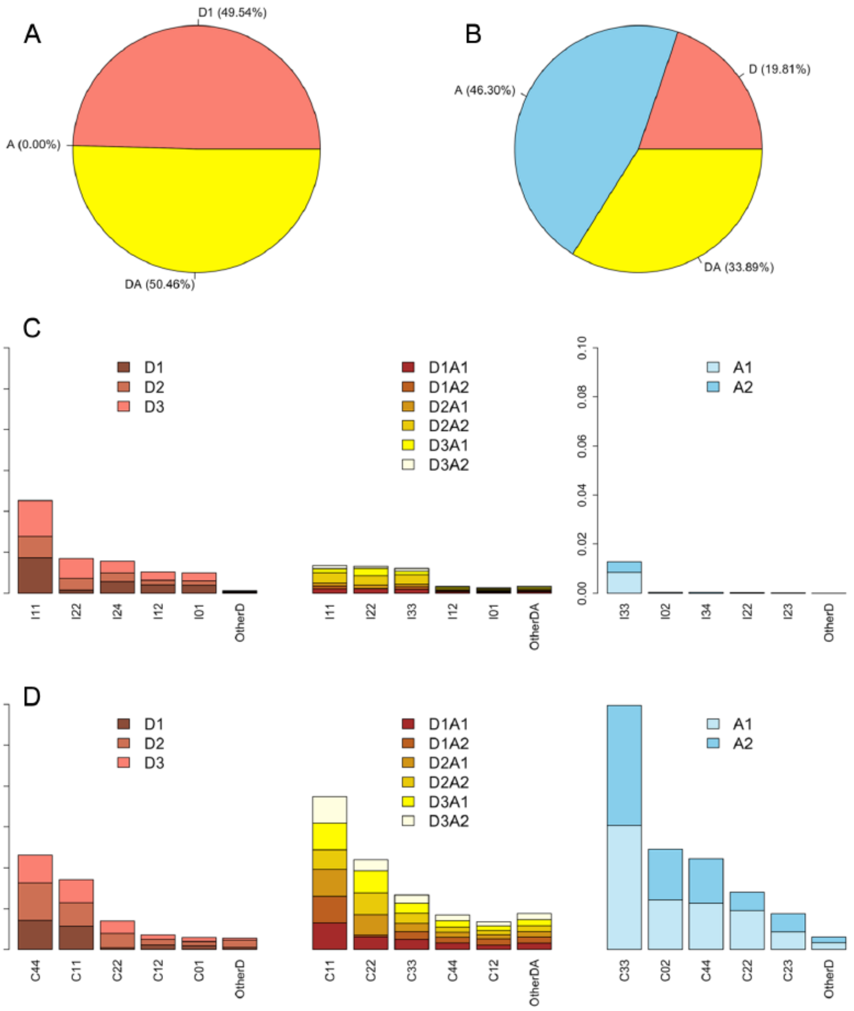

Figure 3. The proportions and numbers of functional profiles: (A) interior profiles, (B) C-terminal profile, (C) numbers of interior profile (in thousands), and (D) numbers of C-terminal profiles (in thousands). A indicates acceptor; D, donor; DA, donor and acceptor.

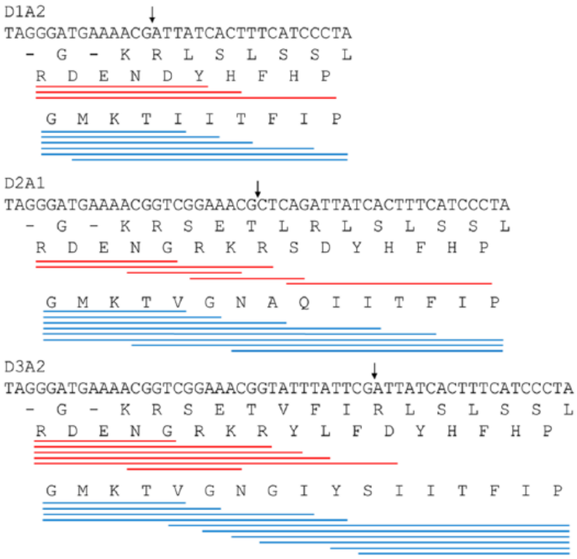
D1A2
TAGGGATGAAAACGATTATCACTTTCATCCCTA
- G - K R L S L S S L
R D E N D Y H F H P
G M K T I I T F I P
D2A1
TAGGGATGAAAACGGTCGGAAACGCTCAGATTATCACTTTCATCCCTA
- G - K R S E T L R L S L S S L
R D E N G R K R S D Y H F H P
G M K T V G N A Q I I T F I P
D3A2
TAGGGATGAAAACGGTCGGAAACGGTATTTATTCGATTATCACTTTCATCCCTA
- G - K R S E T V F I R L S L S S L
R D E N G R K R Y L F D Y H F H P
G M K T V G N G I Y S I I T F I P

Figure 4. Distinct donor/acceptor combinations (ie, D1A2, D2A1, and D3A2) resulting in proteins that differ from other proteins by only a few amino acids.

other than these 7 amino acids can also contribute to creating functional profiles. Similar results were observed when comparing profiles from D1A1 and D3A1 isoforms or those from D1A2 and D3A2 isoforms.

The second example mentioned above consisted of comparing the translated protein isoforms of D1A2, D2A1, and D3A2 differing by only 2 to 7 additional amino acids (Figure 4). Table 3 shows the profiles that were built in protein isoforms by 1 or 2 (but not all) of the D1A2, D2A1, and D3A2 isoforms from a single *Ds1* insertion site. For example, 13, 5, and 6 profiles were yielded only by D1A2, D2A1, and D3A2, respectively (boldfaced profiles in Table 3). D1A2 provided the lowest number of *Ds1* messages, but a specific amino acid sequence, GMKTIITFIP, from D1A2 exonization exclusively contributed 7 profiles, PS00445, PS00622, PS00634, PS00740, PS00838, PS01241, and PS01359, which were not observed in D2A1 and D3A2 isoforms (Figure 4 and Table 3). Although PS00189, PS00371, PS01067, and PS00041 were present in all the D1A2, D2A1, and D3A2 isoforms, they originated from different *Ds1* insertion sites and, consequently, were built by

Table 2. Unique functional profiles yielded by gaining or losing 7 amino acids (either as "VGNGIYS" or "GRKRYLF"), which would be exonized using D1 and D3 donors because the splice junction of D3 is located downstream from D1 by 21 bp.

CLASS	GAIN OR LOSS OF *DS1* MESSAGES FOR TRANSLATED AMINO ACIDS	PROFILE ID
I02-D3	GainVGNGIYS	PS00098; PS00186; PS00371; PS00447; PS01067
I02-D3	GainGRKRYLF	PS00636; PS00743; PS00761
I12-D3	GainVGNGIYS	PS00098; PS00186; PS00371; PS00420; PS00551; PS00595; PS00878; PS01067; PS01143
I12-D3	GainGRKRYLF	PS00027; PS00041; PS00636; PS01143
I22-D3	GainGRKRYLF	PS00009
I04-D1	LossVGNGIYS	PS00012; PS00027; PS00028; PS00029; PS00041; PS00059; PS00079; PS00086; PS00098; PS00189; PS00211; PS00212; PS00216; PS00251; PS00285; PS00310; PS00356; PS00358; PS00371; PS00389; PS00445; PS00527; PS00583; PS00589; PS00592; PS00615; PS00636; PS00652; PS00666; PS00678; PS00770; PS00778; PS00878; PS00909; PS00957; PS01008; PS01067;PS01145; PS01249; PS01353
I04-D1	LossGRKRYLF	PS00007; PS00029; PS00041; PS00079; PS00136; PS00189; PS00299; PS00451; PS00464; PS00595; PS00652; PS00698; PS00743; PS01202
I04-D3	GainVGNGIYS	PS00052; PS00292; PS00634; PS01094; PS01109; PS01171
I04-D3	GainGRKRYLF	PS01117
I14-D1	LossVGNGIYS	PS00012; PS00027; PS00028; PS00029; PS00053; PS00062; PS00079; PS00086; PS00095; PS00098; PS00128; PS00133; PS00146; PS00163; PS00186; PS00189; PS00194; PS00211; PS00212; PS00251; PS00262; PS00285; PS00299; PS00316; PS00338; PS00358; PS00371; PS00389; PS00392; PS00445; PS00551; PS00589; PS00592; PS00615; PS00636; PS00657; PS00672; PS00678; PS00818; PS00889; PS00914; PS01067; PS01103; PS01186; PS01275
I14-D1	LossGRKRYLF	PS00007; PS00022; PS00024; PS00028; PS00029; PS00063; PS00079; PS00107; PS00133; PS00194; PS00216; PS00232; PS00236; PS00280; PS00296; PS00362; PS00410; PS00422; PS00451; PS00464; PS00595; PS00678; PS01159; PS01186; PS60014
I14-D3	GainVGNGIYS	PS00218; PS00559; PS00605; PS01094; PS01109
I14-D3	GainGRKRYLF	PS00605; PS00634; PS01117
I24-D1	LossVGNGIYS	PS00165; PS00187; PS00237; PS00304; PS00671; PS00778; PS00915; PS01319
I24-D1	LossGRKRYLF	PS00018; PS00213
I24-D3	GainVGNGIYS	PS00062; PS00079; PS00107; PS00170; PS00211; PS00259; PS00290; PS00380; PS00588; PS00589; PS00598; PS00606; PS00636; PS01032
I24-D3	GainGRKRYLF	PS00070; PS00214; PS00674; PS01238
C02-D3	GainVGNGIYS	PS00098; PS00186; PS00189; PS00551
C02-D3	GainGRKRYLF	PS00583; PS00636; PS00761
C12-D3	GainVGNGIYS	PS00073; PS00186; PS00189; PS00371; PS00447
C12-D3	GainGRKRYLF	PS00041; PS00098; PS00636; PS00761
C22-D3	GainGRKRYLF	PS00009
C04-D1	LossVGNGIYS	PS00028; PS00029; PS00041; PS00061; PS00079; PS00211; PS00389; PS00551
C04-D1	LossGRKRYLF	PS00007; PS00028; PS00029; PS00041; PS00159; PS00189; PS00464; PS00583; PS01176; PS01249
C04-D3	GainVGNGIYS	PS00107; PS01047; PS01094
C04-D3	GainGRKRYLF	PS00527; PS01117; PS01143
C14-D1	LossVGNGIYS	PS00012; PS00022; PS00028; PS00029; PS00061; PS00073; PS00079; PS00086; PS00132; PS00133; PS00163; PS00189; PS00211; PS00285; PS00371; PS00392; PS00447; PS00527; PS00605; PS00678; PS00878; PS00889

(Continued)

Table 2. (Continued)

CLASS	GAIN OR LOSS OF *DS1* MESSAGES FOR TRANSLATED AMINO ACIDS	PROFILE ID
C14-D1	LossGRKRYLF	PS00007; PS00021; PS00022; PS00028; PS00029; PS00063; PS00079; PS00223; PS00270; PS00296; PS00527; PS00652; PS00678; PS00761; PS01186
C14-D3	GainVGNGIYS	PS00214; PS00217; PS00622; PS01047
C14-D3	GainGRKRYLF	PS00217
C24-D1	LossVGNGIYS	PS00217; PS00615
C24-D1	LossGRKRYLF	PS00018; PS00027; PS00213
C24-D3	GainVGNGIYS	PS00092; PS00107; PS00205
C24-D3	GainGRKRYLF	PS00012; PS00216; PS00276
C44-D1	LossVGNGIYS	PS00435

Note that the total number of profiles composing a class may not be equal to the number of unique profiles of that class shown in Figures 2 and 3 because messages other than these 7 amino acids can also build functional profiles.

Table 3. Unique translated *Ds1* messages for the functional profiles of exonized protein isoforms yielded by joining specific donor and acceptor sites of D1A2, D2A1, and D3A2.

UNIQUE TRANSLATED *DS1* MESSAGES FOR THE FUNCTIONAL PROFILES	INTERIOR	C-TERMINAL
D1A2		
RDENDY	**PS00007**	PS00007; PS00189
RDENDYH	—	**PS01173**
RDENDYHFHP	**PS00028**; PS00223	PS00028
GMKTI	—	PS00371
GMKTII	PS00079; **PS00251**	PS00079; PS00636
GMKTIIT	—	PS01067
GMKTIITFI	**PS00356**	PS00098; PS00356
GMKTIITFIP	PS00041; PS00107; PS00189; **PS00445; PS00622; PS00634; PS00740; PS00838; PS01241; PS01359**	PS00041; PS00189; PS00223; PS00622; PS00634; PS00716; PS00838; PS01047
MKTIITFIP	—	**PS01319**
D2A1		
RDENG	—	PS00761
RDENGRKR	PS00041	—
NGRK	PS00009	PS00009
RKRS	**PS00004**	PS00004
SDYHFHP	**PS00214**	PS00214
GMKTV	—	PS00371
GMKTVG	PS00186; PS00589; PS00878	PS00186; PS00420; PS00589; PS01067
GMKTVGNA	**PS00012**	PS00012
GMKTVGNAQII	—	**PS00373**
GMKTVGNAQIITF	**PS00362**; PS01047	—
GMKTVGNAQIITFIP	PS00041; PS00716; PS01047	PS00716; PS01047

Table 3. (Continued)

UNIQUE TRANSLATED *DS1* MESSAGES FOR THE FUNCTIONAL PROFILES	INTERIOR	C-TERMINAL
TVGNAQIITFIP	PS00223	—
NAQIITFIP	PS00189	—
D3A2		
RDENG	—	PS00761
RDENGRKR	PS00041	—
RDENGRKRY	PS00636	PS00636
RDENGRKRYL	—	**PS00260**
RDENGRKRYLFD	—	**PS00073**
NGRK	PS00009	PS00009
GMKTV	—	PS00371
GMKTVG	PS00186; PS00589; PS00878	PS00186; PS00420; PS00589; PS01067
GMKTVGNGI	—	PS00098
GMKTVGNGIY	—	PS00189
VGNGIYSIITFIP	PS00107	PS00107
GNGIYSIITFIP	PS00189	—
NGIYSIITFIP	**PS00323**	**PS00052**; PS00323
GIYSIITFIP	PS00079	PS00079
YSIITFIP	**PS00027**	PS00027
SIITFIP	**PS00392**	PS00392

The resulting variants differ from others by only a few amino acids, which build unique profiles by 1 (bold) or 2 patterns only. Although PS00189, PS00371, PS01067, and PS00041 present in all three patterns, each profile was built by independent translated *Ds1* message (see text). Profiles yielded by gaining or losing 7 amino acids (either as "VGNGIYS" or "GRKRYLF"), which are exonized using D1A2 and D3A2, were not shown.

independent translated *Ds1* messages. For example, PS00189 was matched by the translated *Ds1* messages "RDENDY," "NAQIITFIP," and "GMKTVGNGIY" from D1A2, D2A1, and D3A2 patterns, respectively, due to its pattern being relatively broadly defined. Taken together, all the functional profiles shown in Tables 2 and 3 contributed about one-third (48 of 145) of the *Ds1*-dependent exclusive profiles (Supplementary Table 2) using merely 59 bp.

Ds1 also differs from *Ds* by providing 2 acceptors for exonization. New interior profiles due to exonization only using *Ds1* acceptors accounted for less than 1% of the interior profiles (Figure 3A), but new C-terminal profiles using *Ds1* acceptors accounted for about 46% of the C-terminal profiles (blue bars in Figure 3). Similarly, either gain or loss of the *Ds1* messages using *Ds1* acceptors enriches proteome complexity through the creation of unique profiles. For example, the number of unique profiles in class C33-A2, 120, yielded by providing an extra 19 bp was a bit higher than that in C33-A1, 118 (Figure 3); however, the number of unique profiles in class C23-A2, 184, was less than that in C23-A1, 193. All these results suggest that *Ds1* messages play a different role than *Ds* messages in building new functional protein isoforms for selective advantage.

Ds1 yields more exclusive profiles than Ds does

As shown in Figure 1B, the *Ds* and *Ds1* transposons share a large degree of similarity in their sequences. However, the behaviors of these 2 TEs for exonization involved in splicing events are different: *Ds* provides only donors (1 forward and 4 reverse insertion patterns), whereas *Ds1* provides both donors and acceptors.[10,11] Only 1 splice donor site, D1 for *Ds1* and R1 for *Ds*, appears in both TEs, and the same set of protein isoforms were yielded in both simulations using this particular loci (Figure 1B). However, the new profiles yielded from functional variants using sites other than R1 in *Ds1* are considered to be *Ds1*-specific profiles.

For interior variants, an intron may yield up to 75 052 (average = 96.92) new functional profiles via *Ds1* insertion and subsequent exonization events (Figure 5A). For C-terminal variants, an intron may yield up to 413 354 (average = 348.84)

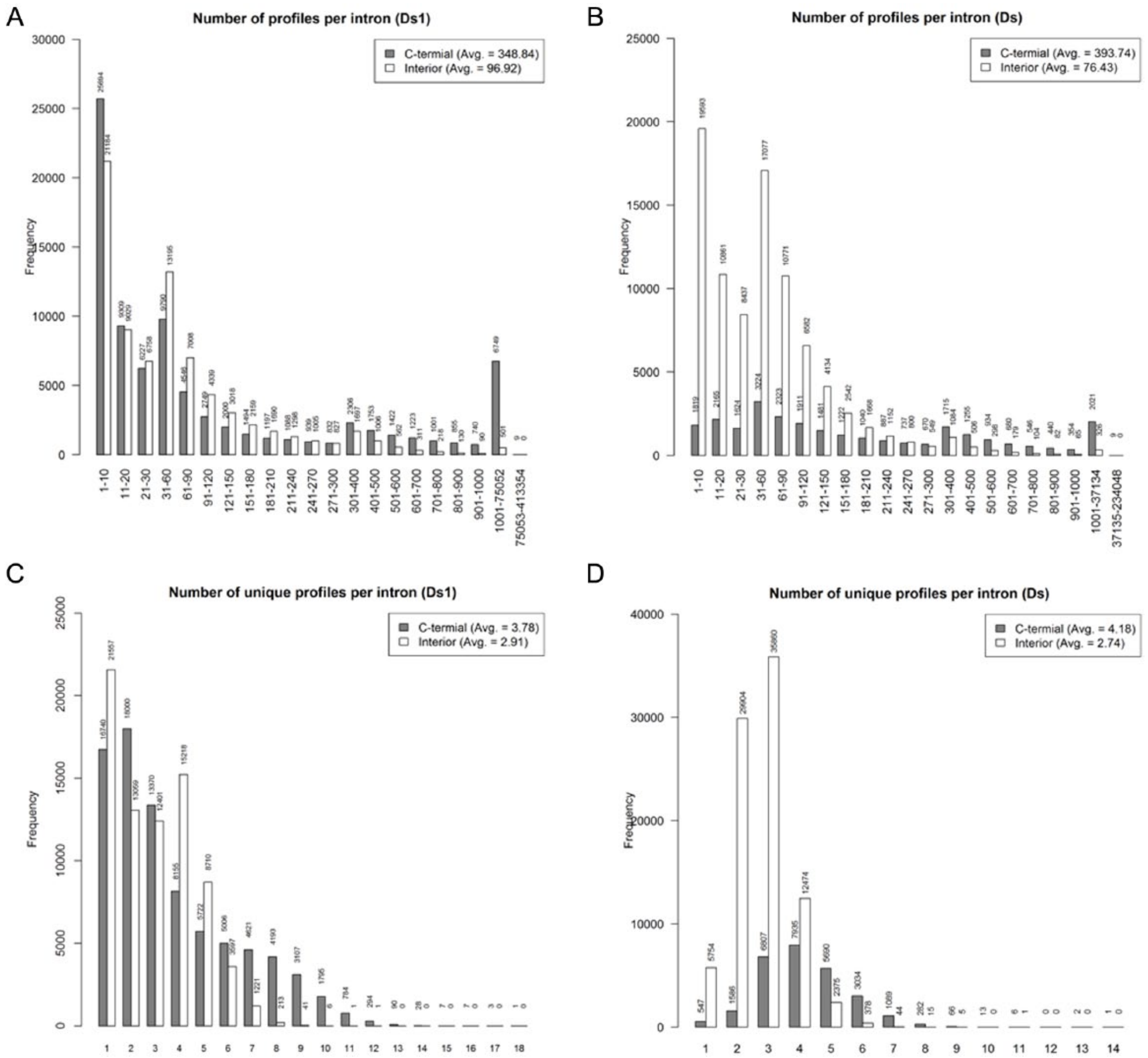

Figure 5. The distributions of (A and B) numbers and (C and D) unique numbers of profiles per intron in rice yielded by *Ds1* and *Ds* exonization, respectively.

new profiles (Figure 5A). The particularly high number of profiles that a given intron could produce in terms of C-terminal variants could result from the peptides generated as a result of a new reading frame being used to replace the reference protein's C-terminus. For the unique profiles, the interior variants can yield up to 12 unique profiles (average = 2.91) via an intron, whereas the C-terminal can yield up to 18 (average = 3.78) unique profiles (Figure 5C).

Through a comparison with those profiles obtained with *Ds* exonization in rice (Figure 5B and 5D), we found that *Ds* insertion yielded more profiles and unique profiles per intron than *Ds1* insertion did via exonization. We reason that the insertion of the specific forward-pattern donor, F1, provided by *Ds* might possibly cause more distinct variants being exonized (Figure 1B); this forward splice donor is, however, absent in *Ds1*. The inflated number of variants by F1 insertion rationally brought on a higher number of (unique) profiles per intron, particularly in C-terminal variants. However, from a total number of 1308 patterned profiles in the PROSITE database, *Ds1* and *Ds* built 459 and 365 unique profiles, respectively, with 330 identical ones (Supplementary Table 3), meaning that there are only 129 and 35 unique profiles exclusively built by *Ds1* and *Ds*, respectively. This implies that although the termini of *Ds* and *Ds1* are highly conserved, they yield independent sets of exonized protein isoforms and would thus be synergistic in contributing to the evolution of proteome complexity.

Conclusions

Ds and *Ds1*, which belong to the same TE family, share an identical 13 bp at 5′-terminal and 26 bp at 3′-terminal sequences. For exonization, the small difference in sequences between *Ds1* and *Ds* may result in different PTCs, donor sites, and incorporated TE messages, which could, consequently, build independent sets of protein variants. As demonstrated in this study, this small difference in sequences makes *Ds1* act unlike *Ds* in exonization. We have previously reported that *Ds* passively enriches proteome complexity in exonization by mainly adopting the messages of the flanking introns after *Ds* insertion sites.[13] However, by offering acceptors that create new A-variants and DA-variants, *Ds1* is more actively involved in exonization, either through its message alone or together with its flanking intron/exon, allowing the building of various new functional protein variants. We also demonstrated that, although a few of the 11 possible exonizing patterns from that *Ds1* inserting into a single site are very similar to each other, even differences of only a few amino acids are enough to result in a wide spectrum of new profiles among these isoforms. All these features suggest that the evolutionary impacts of *Ds* and *Ds1* due to exonization are distinct in various respects. We thus conclude that *Ds* and *Ds1* exonizations are independent and synergistic in their effects on evolutionary proteome complexity enrichment. Incorporating further molecular analysis, for example, determining the changes in the priority of exonization sites under various stresses, would provide more information about the evolutionary impact of TE exonization.

Author Contributions

YC and LDL conceived and designed the experiments. LH and LDL analyzed the data. YC wrote the first draft of the manuscript. YC, LH, and LDL contributed to the writing of the manuscript. YC, LH, and LDL Agree with manuscript results and conclusions. YC, LH, and LDL jointly developed the structure and arguments for the paper. YC and LDL made critical revisions and approved final version. All authors reviewed and approved the final manuscript.

Disclosures and Ethics

As a requirement of publication, author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including, but not limited to, the following: authorship and contributorship, conflicts of interest, privacy and confidentiality, and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

REFERENCES

1. Sela N, Kim E, Ast G. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biol*. 2010;11:R59.
2. Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*. 2008;9:397–405.
3. Schmitz J, Brosius J. Exonization of transposed elements: a challenge and opportunity for evolution. *Biochimie*. 2011;93:1928–1934.
4. Severing E, van Dijk A, Stiekema W, van Ham R. Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics*. 2009;10:154.
5. Huang K-C, Yang H-C, Li K-T, Liu L, Charng Y-C. *Ds* transposon is biased towards providing splice donor sites for exonization in transgenic tobacco. *Plant Mol Biol*. 2012;79:509–519.
6. Haring MA, Rommens CMT, Nijkamp HJJ, Hille J. The use of transgenic plants to understand transposition mechanisms and to develop transposon tagging strategies. *Plant Mol Biol*. 1991;16:449–461.
7. Kunze R, Weil CF. The *hAT* and CACTA superfamilies of plant transposons. In: Craig N, Craigie R, Gellert M, Lambowitz A, eds. *Mobile DNA II*. Washington, DC: ASM Press; 2002:565–610.
8. Pisabarro AG, Martin WF, Peterson PA, Saedler H, Gierl A. Molecular analysis of the Ubiquitous (Uq) transposable element system of *Zea mays*. *Mol Gen Genet*. 1991;230:201–208.
9. Wessler SR. The maize transposable Ds1 element is alternatively spliced from exon sequences. *Mol Cell Biol*. 1992;11:6192–6196.
10. Liu LD, Charng Y-C. Genome-wide survey of *Ds* exonization to enrich transcriptomes and proteomes in plants. *Evol Bioinform Online*. 2012;8:575–587.
11. Charng Y-C, Liu LD. The extent of Ds1 transposon to enrich transcriptomes and proteomes by exonization. *Bot Stud*. 2013;54:14.
12. Hulo N, Bairoch A, Bulliard V, et al. The PROSITE database. *Nucleic Acids Res*. 2006;34:D227–D230.
13. Chien TY, Liu L-YD, Charng Y-C. Analysis of new functional profiles of protein isoforms yielded by *Ds* exonization in rice. *Evol Bioinform Online*. 2013;9:417.
14. Chang Y-F, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem*. 2007;76:51–74.
15. Hori K, Watanabe Y. Context analysis of termination codons in mRNA that are recognized by plant NMD. *Plant Cell Physiol*. 2007;48:1072–1078.
16. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008.