Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity

Yamile Marquez,1 Markus Höpfler,1,3 Zahra Ayatollahi,1 Andrea Barta,1 and Maria Kalyna1,2

1Max F. Perutz Laboratories, Medical University of Vienna, Vienna A-1030, Austria; 2Department of Applied Genetics and Cell Biology, BOKU – University of Natural Resources and Life Sciences, Vienna A-1190, Austria

Alternative splicing (AS) diversifies transcriptomes and proteomes and is widely recognized as a key mechanism for regulating gene expression. Previously, in an analysis of intron retention events in Arabidopsis, we found unusual AS events inside annotated protein-coding exons. Here, we also identify such AS events in human and use these two sets to analyse their features, regulation, functional impact, and evolutionary origin. As these events involve introns with features of both introns and protein-coding exons, we name them exitrons (exonic introns). Though exitrons were detected as a subset of retained introns, they are clearly distinguishable, and their splicing results in transcripts with different fates. About half of the 1002 Arabidopsis and 923 human exitrons have sizes of multiples of 3 nucleotides (nt). Splicing of these exitrons results in internally deleted proteins and affects protein domains, disordered regions, and various post-translational modification sites, thus broadly impacting protein function. Exitron splicing is regulated across tissues, in response to stress and in carcinogenesis. Intriguingly, annotated intronless genes can be also alternatively spliced via exitron usage. We demonstrate that at least some exitrons originate from ancestral coding exons. Based on our findings, we propose a “splicing memory” hypothesis whereby upon intron loss imprints of former exon borders defined by vestigial splicing regulatory elements could drive the evolution of exitron splicing. Altogether, our studies show that exitron splicing is a conserved strategy for increasing proteome plasticity in plants and animals, complementing the repertoire of AS events.

[Supplemental material is available for this article.]

In the majority of eukaryotic genes, the protein coding information of exons is interrupted by intervening sequences, introns. Differential inclusion of exons and introns or their parts in mature mRNAs, so-called alternative splicing (AS), results in multiple transcript and protein variants with different fates and functions from a single gene. About 95% of human and 60% of Arabidopsis genes are alternatively spliced (Pan et al. 2008; Wang et al. 2008; Marquez et al. 2012). The repertoire of AS transcripts produced from a single gene is dynamic and changes in different tissues, during development, and in response to environmental cues (Kalsotra and Cooper 2011; Staiger and Brown 2013). Consequently, AS has emerged as a major mechanism to increase the density of information encoded by eukaryotic genomes. Therefore, understanding AS is of paramount importance as further emphasized by linkage of abnormal AS to numerous human diseases, including cancer (Srebrow and Kornblitt 2006; Kelemen et al. 2013). Nevertheless, storing, retrieval, and processing of AS relevant information remain incompletely understood.

Intron removal relies mainly on the core splicing signals present in every intron: 5′ and 3′ splice sites and branch point (Wang and Burge 2008). However, in Arabidopsis and human, these signals represent only part of the information required to define introns (Lim and Burge 2001). Multiple features such as the presence of intronic and exonic splicing regulatory cis-elements, length of introns and exons, their differential guanine-cytosine (GC) content, distinct DNA methylation, histone modifications, and positioning of nucleosomes over introns and exons and at exon/intron boundaries contribute significantly to the recognition of the core splicing signals and can change splice site selection, resulting in AS events (Braunschweig et al. 2013; Reddy et al. 2013). Common types of AS events include intron retention (IR), usage of alternative 5′ and 3′ splice sites (ASSS and A3SS), exon skipping (ES), and mutually exclusive exons, whereby IR is a frequent event both in human and in plants (Ner-Gaon et al. 2004; Marquez et al. 2012; Braunschweig et al. 2013, 2014; Reddy et al. 2013). IR events stall expression of certain genes at particular stages, cell types, or conditions and therefore are thought to control developmental transitions or stress responses (Boothby et al. 2013; Wong et al. 2013; Braunschweig et al. 2014; Shalgi et al. 2014). In spite of their prevalence and functional impact, many questions concerning IRs remain unanswered.

Previously, we conducted a genome-wide survey of the features of retained introns in Arabidopsis thaliana (Marquez et al. 2012). This analysis revealed a subfamily of IRs that constitute internal regions of annotated protein-coding exons, which we referred to as cryptic introns (Marquez et al. 2012). These introns possess all the canonical core splicing signals (5′ and 3′ splice sites and branch point) and, as they are internal parts of the protein-

© 2015 Marquez et al. This article, published in Genome Research, is available under a Creative Commons License (Attribution 4.0 International), as described at http://creativecommons.org/licenses/by/4.0/.
coding exons, they do not contain stop codons. On the basis of their exonic and intronic nature, here we name them exitrons (exonic introns [EIs]) and define them as alternatively spliced internal regions of protein-coding exons. As exitrons are protein-coding sequences directly flanked by protein-coding exonic sequences, they have a great potential to boost protein diversity via AS. Furthermore, these intrinsic features of exitrons raise questions about the origin and evolution of their splicing. Here, we present a comprehensive characterization of this AS event in Arabidopsis and human.

**Results**

Exitron splicing, an alternative splicing event inside protein-coding exons

Overlapping splice junction and exonic reads (Fig. 1) derived from our A. thaliana RNA-seq (flowers and 10-d-old seedlings) (Marquez et al. 2012) mapping to a single annotated protein-coding exon were used to identify EIs and to distinguish them from other IRs (Supplemental Methods). We have defined a set of 1002 exitrons in 892 Arabidopsis genes (Fig. 2A; Supplemental Table 1). As expected from our previous analysis (Marquez et al. 2012), Arabidopsis exitrons have weaker splice site signals than other introns (Supplemental Fig. 1). Intriguingly, 18.9% of exitrons are located in 165 genes annotated as intronless (Supplemental Table 1), suggesting exitron splicing (EIS) to be a novel source of alternative transcripts and protein isoforms for these genes. In total, EIS affects 3.3% of Arabidopsis protein-coding genes (27,206; TAIR10). The exitron subset constitutes 11% of all IRs (9142) and 3.7% of all AS events detected in the same sample (Marquez et al. 2012). We validated EIS events, including those in annotated intronless genes, by various methods (Supplemental Results; Supplemental Figs. 2–4, 6; Supplemental Tables 2–4).

As exitrons constitute intralexonic regions and were identified in the pool of IRs, we next asked how similar are they to exons and introns and, importantly, to IRs. Arabidopsis exitrons are overall longer and have a higher GC content than IRs and constitutive introns (Fig. 2B, C). However, their sizes are closer to the sizes of constitutive exons and other exons of the EI-containing genes (Fig. 2D). Interestingly, EI-containing exons tend to be longer than other exons (Fig. 2D). The GC content of exitrons is lower than in all groups of exons (Fig. 2E), indicating also that this is a specific feature of exitrons and not a general property of EI-containing exons or genes. This lower GC content in exitrons reflects their intronic nature and may be important for EIS. These results show that exitrons possess properties differentiating them from constitutive exons and introns and from IRs.

Furthermore, in contrast to IRs, where a premature termination codon (PTC) is very often generated in the retained intron sequence or in the downstream exon, nonsplicing of exitrons never introduces any stop codons as they are protein-coding sequences. It is only upon exitron splicing that changes in the fate of the resultant transcripts can occur. Splicing of exitrons with lengths of...
Exitron splicing affects protein function in Arabidopsis

As exitrons are protein-coding sequences, we next analyzed the consequences of EIS on protein features. The Arabidopsis ELx3-encoded domains, disorder regions, post-translational modifications (PTMs), and transmembrane domains.
and signal peptides (Fig. 4A; Supplemental Tables 9, 11–13), suggesting that EIS impacts the functional properties of the proteins. Indeed, our analysis of proteogenomic data supports EIS effect on PTMs (buljan et al. 2013). PTMs regulate protein functions by affecting their activity, localization, or affinity to other proteins. Our analysis of published experimental Arabidopsis PTM data sets shows that ELx3-encoded sequences carry sites for various PTMs (Supplemental Table 11). EIS can change sumoylation, ubiquitylation, S-nitrosylation, and lysine acetylation states of the protein isoforms, thus providing the first evidence that AS can influence protein function by affecting other types of PTMs besides phosphorylation (Zhang et al. 2010; Buljan et al. 2012; Merkin et al. 2012). Moreover, phosphopeptides are enriched in the ELx3-encoded sequences when compared to constitutive exons (11.3% vs 4.1%, P-value < 0.001), also when corrected by exon length (Supplemental Table 11).

Figure 3. Regulation of exitron splicing in Arabidopsis. (A) Structures of ELx3-containing genes tested by RT-PCR in B. Dark blue indicates exitron; red car-

ters, exitron splicing. (B) Heatmap for EIS in different tissues, stress conditions, splicing factor mutant and overexpression lines, and Col-0 and Ler-0 ecotypes. The coloring represents only significant changes (P-value < 0.1). a, b, and c refer to cases described in C. (C) SNPs affect EIS in Ler-0 ecotype. SNPs in the splicing signals are indicated in red. 5′SS, 3′SS, and BP indicate 5′ splice site, 3′ splice site, and branch point, respectively. RT-PCR products of the full-length (FL) and EI-spliced isoforms (filled and open triangles, respectively) are shown. Ubiquitin was used as a loading control (*). (D) RBM/SUA motifs and MEME-predicted motifs are enriched in exitrons ([*]) P-value < 0.0001).
Role of exitron splicing in Arabidopsis and human

Methods. Altogether, these results suggest that EIS impacts the dynamics of the Arabidopsis proteome.

The effect of EIS on protein features is illustrated by an EIx3 in the gene encoding the eukaryotic translation initiation factor 4A (EIF4A, AT1G54270), a DEAD-box RNA helicase (Fig. 4C). EIS removes the highly conserved ATP binding motif together with two conserved phosphorylation sites (Fig. 4C; Supplemental Fig. 7). Unwinding of substrates by this RNA helicase is ATP dependent (Cordin et al. 2006), implying that EIS affects this activity of EIF4A.

This EIS event is supported by ESTs in A. lyrata (Syed et al. 2012). To estimate EIS conservation, we used only A. thaliana and A. lyrata EST sets where the coverage is similar. Out of 98 genes with 100 EIS events supported by ESTs in A. thaliana (Supplemental Table 2), 56 genes have orthologs in our A. lyrata set. We found that 40 EIS events in these 56 genes (71.4%) are also supported by ESTs in A. lyrata (Supplemental Table 14). Deeper transcriptome data for plants will likely improve both EIS detection and the estimate of conservation levels. However, EIS can also be species-specific, thus providing a source for adaptation and speciation, as observed for AS exons (Barbosa-Morais et al. 2012; Merkin et al. 2012). In line with this, we found SNPs affecting EIS between two A. thaliana ecotypes (see above, Fig. 3B,C). Our further analysis of 82 ecotypes showed that 2.2% of SNPs either decrease or increase the strength of a splice site signal in at least one ecotype when compared to Col-0 (Supplemental Results; Supplemental Tables S8–S10). As many EI-containing genes are involved in stress responses, the genetic variability affecting EIS could play a role in the adaptation of A. thaliana ecotypes.

EIS in the human EIF4A1 indicates that these events are not restricted to plants. As exitrons could be classified as IRs, we searched for such cases in the literature. Indeed, we found IRs in the mammalian-specific genes human CCKBR, CD55, and FMNL1 and mouse Tgf2 that qualified as exitrons (Supplemental Table 15). Moreover, additional cases of splicing of intraxonic sequences were described in Caenorhabditis elegans, and a hypothesis on their origin was proposed (see below) (Irimia et al. 2008). Altogether, these findings demonstrate that EIS is a common strategy to increase transcriptome diversity in plant and nonplant species.

Exitron splicing is a widespread alternative splicing event in human

To obtain further evidence of EIS in nonplant species, we explored the set of annotated IR transcripts in human Ensembl. In this set, 670 retained introns (in 577 genes) qualified as exitrons (Supplemental Table 16), including the above-described exitrons in FMNL1 and CCKBR, but not in CD55, suggesting that this set is not exhaustive. Thus, we analyzed RNA-seq data sets from six human tissues (brain, heart, liver, lung, ovary, and testis) (Barbosa-Morais et al. 2012) and from a ERBB2-positive breast cancer and the control breast tissue (NBS) (Eswaran et al. 2013). We found 602 exitrons in 488 genes (Supplemental Table 16). Altogether, we detected 923 EIS events in 747 genes (∼3.7% of 20,364 human genes).

Exitron splicing is an evolutionarily conserved strategy to increase versatility of transcriptomes

To find further cases of EIS events in plants, we produced a conserved set of orthologous gene pairs using A. thaliana EI-containing genes (for numbers, see Fig. 7A) and tested it against respective EST collections. We found several conserved examples of EIS: in poplar POPTR_0002s23170, in grape Vv03s0038g03800, and in rice LOC_Os07g08729 and LOC_Os07g05570. The highest number of EIS events (46) was found in Arabidopsis lyrata (Supplemental Table 14). The low level of EIS detection can be explained by different depths of transcriptome coverage in these species (ESTs) in comparison to A. thaliana (RNA-seq), as observed for the discovery of AS events in general (Syed et al. 2012). To estimate EIS conservation, we used only A. thaliana and A. lyrata EST sets where the coverage is similar. Out of 98 genes with 100 EIS events supported by ESTs in A. thaliana (Supplemental Table 2), 56 genes have orthologs in our A. lyrata set. We found that 40 EIS events in these 56 genes (71.4%) are also supported by ESTs in A. lyrata (Supplemental Table 14). Deeper transcriptome data for plants will likely improve both EIS detection and the estimate of conservation levels. However, EIS can also be species-specific, thus providing a source for adaptation and speciation, as observed for AS exons (Barbosa-Morais et al. 2012; Merkin et al. 2012). In line with this, we found SNPs affecting EIS between two A. thaliana ecotypes (see above, Fig. 3B,C). Our further analysis of 82 ecotypes showed that 2.2% of SNPs either decrease or increase the strength of a splice site signal in at least one ecotype when compared to Col-0 (Supplemental Results; Supplemental Tables S8–S10). As many EI-containing genes are involved in stress responses, the genetic variability affecting EIS could play a role in the adaptation of A. thaliana ecotypes.

EIS in the human EIF4A1 indicates that these events are not restricted to plants. As exitrons could be classified as IRs, we searched for such cases in the literature. Indeed, we found IRs in the mammalian-specific genes human CCKBR, CD55, and FMNL1 and mouse Tgf2 that qualified as exitrons (Supplemental Table 15). Moreover, additional cases of splicing of intraxonic sequences were described in Caenorhabditis elegans, and a hypothesis on their origin was proposed (see below) (Irimia et al. 2008). Altogether, these findings demonstrate that EIS is a common strategy to increase transcriptome diversity in plant and nonplant species.

Exitron splicing is a widespread alternative splicing event in human

To obtain further evidence of EIS in nonplant species, we explored the set of annotated IR transcripts in human Ensembl. In this set, 670 retained introns (in 577 genes) qualified as exitrons (Supplemental Table 16), including the above-described exitrons in FMNL1 and CCKBR, but not in CD55, suggesting that this set is not exhaustive. Thus, we analyzed RNA-seq data sets from six human tissues (brain, heart, liver, lung, ovary, and testis) (Barbosa-Morais et al. 2012) and from a ERBB2-positive breast cancer and the control breast tissue (NBS) (Eswaran et al. 2013). We found 602 exitrons in 488 genes (Supplemental Table 16). Altogether, we detected 923 EIS events in 747 genes (∼3.7% of 20,364 human genes).
protein-coding genes; GRCh37) (Fig. 5A), whereby 349 EIS events are shared between the Ensembl and RNA-seq exitron sets (Fig. 5B). Human ELx3-containing genes are enriched in genes with functions in DNA replication, immune response, and the mediator and calcium channel complexes (Supplemental Table 17). Similar to Arabidopsis, the 3n fractions of human exitrons (∼55%) and IRs (32.07%; based on analysis of RNA-seq and Ensembl data sets) are significantly different. Moreover, the 3n exitron fraction is much higher than the fraction of human 3n introns without stop codons (29.8%), previously shown to be counter-selected in the human genome (Jaillon et al. 2008). Human and Arabidopsis exitrons share other features, such as weaker splice sites and higher GC content in comparison to IRs; they have similar size distribution, and their size are closer to exons than to IRs (Fig. 5C–E; Supplemental Fig. 8A–C). In both species, EL-containing exons are considerably longer than other exons (Figs. 2D, 5E). Our analyses show that EIS affects a comparable number of genes in human and in Arabidopsis, and exitrons have similar features in both species.

### EIS affects protein properties in human

To obtain evidence for the translation of exitronic sequences in human, we analyzed published human tissue proteome data sets (Supplemental Methods). We found 382 peptides supporting 81 EL-encoded sequences (Supplemental Table 18), indicating that, as in Arabidopsis, EL-containing isoforms are exported to the cytoplasm and translated in contrast to IR transcripts that are often retained in the nucleus and not translated (Yap et al. 2012; Shalgi et al. 2014).

Analyses of the human ELx3-encoded protein sequences showed that ∼34% overlap with protein domains, whereby splicing boundaries of about one-third of these ELx3s coincide with protein domain borders (Supplemental Tables 19, 20). Interestingly, EIS affects C2H2 type zinc finger (2ZNF) domains in five KRAB-ZNF transcription factors that act as repressors of different endogenous retroviruses. Variation of their DNA binding specificity is achieved by gene duplication and recombination and by duplications and deletions of ZNF repeats that are organized in a single exon (Lukic et al. 2014). EIS affecting ZNF repeats can therefore provide another mechanism for a concerted evolution of combinatorial tools to inactivate retroviruses. As in Arabidopsis, protein domains are underrepresented, while disordered regions and SLiMs are enriched in human ELx3-encoded sequences (Fig. 6A–C). Similar to Arabidopsis, we also detected the first examples of AS affecting various PTM states of protein isoforms in human, whereby the list is expanded to methylation and O-linked glycosylation (Fig. 6D; Supplemental Table 21). Moreover, we detected enrichment not only for phosphorylation sites but also for ubiquitylation sites in the exitron-encoded sequences (see also Supplemental Methods; Fig. 6E,F; Supplemental Fig. 9). In addition, 710 PTM peptides provide further evidence for translation of 190 ELx3s in 161 genes (Supplemental Table 21). These findings indicate that EIS impacts protein features similarly in plants and humans, thus representing an evolutionarily conserved tool for shaping eukaryotic proteomes.

### EIS is differentially regulated across human tissues

We identified a total of 492 EIS events in six human tissue transcriptomes (Fig. 5B; Supplemental Table 16), whereby 217 (44.1%) are found in all samples. Analysis of the latter showed different PSI (percent of spliced in) values for EIS events across human tissues (Fig. 6G). EI-spliced isoforms can be predominant (PSI < 50) (Fig. 6G), however, as in Arabidopsis, EL-containing isoforms are the major transcripts in most cases as revealed by their high PSI values. As splice variants may only appear in a small number of cells types, the analysis of a whole human organ can underestimate the impact of an event. In addition, studies of more tissue samples and conditions would differentiate events with more ample regulation from those that may represent splicing noise.

We found 52 EIS events showing a change of ≥15% (ΔPSI ≥ 15) in at least one tissue, with 10 of them reported previously to have ample evidence for their regulation and physiological relevance (Supplemental Table 22). For example, EIS in the transcription factor CIZ1 changes protein localization, and the EI-spliced isoform is up-regulated in Alzheimer’s disease brains (Dahmcke et al. 2008). In agreement with these studies, we detected the EI-spliced isoform in all but the brain tissues (Supplemental Table 22). It is important to emphasize that previously reported EIS events comprise a wide range of ambiguous definitions such as intron retention, removing an intron from within the exon, intronic splicing, internal splicing event in exon, internal alternative splice sites, or cryptic 5’ and 3’ splice sites located in the exon.
For EIS events with a PSI ≤ 90 in at least one tissue, we observed their differential distribution across human tissues (Supplemental Table 23). The fraction of EIS events in testis (59.4%) is twice as high as in lung (29.2%), heart (27.7%), or liver (26.6%), while being intermediate in ovary and brain (43.5% and 42.4%, respectively). This distribution differs considerably from frequencies of other types of AS events. Usage of A5SS and A3SS is the most prominent in liver, and ES events are the most frequent in brain and testis; this was attributed to tissue-specific combinations and levels of splicing factors regulating particular events (Yeo et al. 2004). It has been suggested that the high number of AS events in testis may be due to splicing noise as many of these events are not conserved between mouse and human (Kan et al. 2005). However, these events can be species-specific, and AS could be one of the mechanisms driving rapid evolution of the reproductive systems (see also Discussion) (Elliott and Grellscheid 2006). A number of EIS events were detected only in a given tissue (25 in brain, 11 in heart, seven in liver, 13 in lung, 17 in ovary, and 44 in testis) (Supplemental Table 23). However, analyses of more samples are required to determine whether they are indeed tissue-specific. Altogether, our results indicate that EIS is tissue regulated and contributes to shaping the human tissue transcriptomes.

(Supplemental Table 22), indicating uncertainty in the interpretation of the type of AS event.

For EIS events with a PSI ≤ 90 in at least one tissue, we observed their differential distribution across human tissues (Supplemental Table 23). The fraction of EIS events in testis (59.4%) is twice as high as in lung (29.2%), heart (27.7%), or liver (26.6%), while being intermediate in ovary and brain (43.5% and 42.4%, respectively). This distribution differs considerably from frequencies of other types of AS events. Usage of A5SS and A3SS is the most prominent in liver, and ES events are the most frequent in brain and testis; this was attributed to tissue-specific combinations and levels of splicing factors regulating particular events (Yeo et al. 2004). It has been suggested that the high number of AS events in testis may be due to splicing noise as many of these events are not conserved between mouse and human (Kan et al. 2005). However, these events can be species-specific, and AS could be one of the mechanisms driving rapid evolution of the reproductive systems (see also Discussion) (Elliott and Grellscheid 2006). A number of EIS events were detected only in a given tissue (25 in brain, 11 in heart, seven in liver, 13 in lung, 17 in ovary, and 44 in testis) (Supplemental Table 23). However, analyses of more samples are required to determine whether they are indeed tissue-specific. Altogether, our results indicate that EIS is tissue regulated and contributes to shaping the human tissue transcriptomes.

Figure 6. Functional implications of EIS in human. (A–C) Enrichment of PFAM domains (A), disordered regions (B), and SLiMs (C) in protein sequences encoded by EIs and other types of exons. (D) Statistics of various post-translational modification (PTM) sites encoded by EIs. (E,F) Enrichment of phosphorylation (E) and ubiquitylation (F) sites in the sequences encoded by EIs and other types of exons. (A–F) Analyses performed for Elx3 subset only. (G,H) Heatmap of EIS (measured by PSI) in different human tissues (G) and in ERBB2-positive breast cancer and normal breast tissue (NBS) samples (H).
Exitron splicing is misregulated in breast cancer

AS is linked to numerous human diseases, including cancer, suggesting its critical role in organism homeostasis (Srebrow and Kornblitt 2006; Kelemen et al. 2013). By inspecting the human exitron list, we found that EIS affects several cancer-related genes: the cancer marker genes BMI1, KRT5, and MUC1 and the genes involved in cell adhesion (CSF1), migration and metastasis (ZEB2 and KLF17) (Supplemental Table 16).

To address a role for EIS in carcinogenesis, we analyzed ERBB2-positive breast cancer and normal breast tissues (NBS). Out of a total of 312 EIS events (Supplemental Table 16), 275 are ERBB2-positive breast cancer and normal breast tissues (NBS). Though EIS facilitates splicing. Indeed, SNPs in the exitron regions in 82
citable splicing. Indeed, SNPs in the exitron regions in 82

citable splicing. Indeed, SNPs in the exitron regions in 82
substitutions favoring the appearance of exitron splice sites in the Arabidopsis genus (Supplemental Fig. 12C). Interestingly, highly conserved short sequences with avoidance of substitutions are present close to the exitron (Supplemental Fig. 12C) that might represent splicing regulatory elements (Fairbrother et al. 2004). Their analysis shows that they are potential binding sites for RBM5/SUA and CELF2 (Supplemental Fig. 12C). Interestingly, the binding specificity of CELF2 homologs is conserved from human to Arabidopsis, and these proteins have multiple functions in RNA processing, including AS (Good et al. 2000; Kim et al. 2013).

In human, intron loss during EIS evolution is illustrated by the annotated intronless gene HSPA1A (HSP70A-1) (Fig. 7C; Supplemental Fig. 13). The EIS event removes the first of two subdomains of the ATPase domain of HSPA1A. This exitron

| Organism                  | # of analyzed orthologues | Coding exon | Intron (coding) | Total coding (%) |
|---------------------------|---------------------------|-------------|-----------------|------------------|
| Arabidopsis lyrata        | 648                       | 631         | 17 (4)          | 635 (98.0)       |
| Brassica rapa             | 442                       | 430         | 12 (1)          | 431 (97.5)       |
| Populus trichocarpa       | 256                       | 252         | 4 (0)           | 252 (98.4)       |
| Vitis vinifera            | 252                       | 246         | 6 (0)           | 246 (97.6)       |
| Oryza sativa              | 163                       | 158         | 5 (0)           | 158 (96.9)       |
| Selaginella moellendorfii | 62                        | 62          | 0               | 62 (100)         |
| Physcomitrella patens     | 73                        | 73          | 0               | 73 (100)         |
| Chlamydomonas reinhardtii | 29                        | 29          | 0               | 29 (100)         |
corresponds exactly to three exons in the *Ciona* sea squirt species. Interestingly, GT and AG dinucleotides are already present at the corresponding exon borders in *Ciona*.

Altogether, our findings suggest that intron loss accompanied the evolution of a subset of EIS events. A full in-depth study of *Arabidopsis* and human EI-containing genes should be performed in the future to estimate the impact of intron loss on exitron splicing evolution.

**Discussion**

We have performed a comprehensive characterization of a subfamily of AS events, the splicing of exitrons, that allows intraxonic protein-coding sequences to be differentially spliced. Previously, we named them cryptic introns (Marquez et al. 2012); however, their hybrid nature combining features of introns and protein-coding exons is reflected better in the new term, exitrons (exonic introns). It will avoid any confusion with cryptic splice sites that are activated by mutations disrupting the usage of the natural splice sites (Roca et al. 2003). Moreover, while some of EIS events were detected previously, their description in published literature has been ambiguous (Supplemental Tables 15, 22, 24; Ner-Gaon et al. 2004). These events were defined as intron retention, as an internal splicing event in the exon, or as usage of cryptic splice sites located in the exon, all reflecting uncertainty in the interpretation of the type of AS event.

The categorization of AS events has promoted studies of mechanistic differences in splicing regulation and their contribution to phenotype. Recent classification of IRs based on their evolutionary origin and conservation has defined a minor group (type B) of IRs located within exons, including noncoding ones (Braunschweig et al. 2014). Though very useful in terms of evolution, this approach fails to detect species-specific AS events. The latter diverge strongly even between closely related species (∼50% of human and chimp AS exons are different), thus contributing to phenotypic differences (Barbosa-Morais et al. 2012; Merkin et al. 2012). As the definition of exitrons is not based on their evolutionary origin, it overcomes this issue and allows detecting this AS event without the need to compare transcriptomes of different species. The only requirement to define exitrons is the protein-coding potential of EI-containing (unspliced) isoform. While our definition does not include events in noncoding regions, it provides a more homogeneous set of AS events for evolutionary, functional, and mechanistic studies.

Though exitrons were found in the IR sets of *Arabidopsis* (Marquez et al. 2012) and human (Ensembl IR-annotated transcripts), their separation is important, as they have clearly distinguishable features and, notably, their splicing results in transcripts with different fates. First, EI-containing transcripts associate with ribosomes and are translated. Splicing of EIs affects essential protein features. In contrast, intron retention is suggested to be a mechanism to forestall translation, when IR transcripts are recognized as incompletely processed and remain in the nucleus until removal of retained introns post-transcriptionally (Boothby et al. 2013; Shalgi et al. 2014). Second, a PTC can be created downstream from an exon (non EI2) only upon splicing, while in the case of IRs, PTCs are generated due to splicing inhibition and retention of intronic sequences. Moreover, while such EIS events can result in NMD-sensitive transcripts, IR transcripts, though possessing PTCs in a NMD-sensible context, avoid the NMD machinery, at least in *Arabidopsis* (Kalyna et al. 2012; Leviatan et al. 2013), probably due to their retention in the nucleus (Göhring et al. 2014). Third, EI-containing transcripts are major isoforms as evidenced by high PSI values, while IR isoforms are usually of low abundance (Marquez et al. 2012; Braunschweig et al. 2014). And fourth, we demonstrate that a subset of exitrons originates from protein-coding exons. Therefore, it is not surprising that they display features characteristic for such sequences (high GC content, absence of stop codons, overrepresentation of EI2S, and the prevalence of synonymous substitutions); these features are totally atypical for IRs or any type of intron. All this clearly separates exitrons from IRs, while the impact of EIS on the proteome is more similar to skipping of protein-coding exons.

It is well documented that chromatin state, nucleosome positioning, RNA polymerase II occupancy and processivity, and binding of splicing factors differ between exons and introns, impacting AS regulation (Braunschweig et al. 2013; Reddy et al. 2013). Therefore, the unique features of exitrons as intraxonic sequences imply distinct mechanisms controlling their splicing; consequently, a clear differentiation of exitrons from IRs is relevant for studies on AS regulation. In addition, finding EIS in annotated intronless genes revises the concept that such genes are devoid of splicing regulatory elements, further impacting AS research.

Our finding of EIS raises the question of how internal, essentially exonic, regions have evolved into exitrons. It has been proposed that mutations in protein-coding sequences, creating a PTC, would promote intronization of the affected region to rescue at least a shortened ORF (Catania and Lynch 2008). EIS evolution must have proceeded differently because exitrons do not contain PTCS. On the contrary, splicing of non-EI3 may actually result in PTCS, albeit downstream from EIS events (Fig. 1). Cases of intronization of exonic sequences in different species have been described previously (Irinia et al. 2008; Zhu et al. 2009; Kang et al. 2012; Braunschweig et al. 2014). It has been suggested that intronization can occur due to single substitutions creating GT/C and AG splicing boundaries (Irinia et al. 2008). In this hypothesis, it is not clear what the driving force for intronization is. Additionally, though these dinucleotides are required at splice sites, they do not create the complete splice site signals. Furthermore, these substitutions are not always needed, as they can be already present in the ancestral sequences (as in *Ciona* *HSP1A* orthologs). Moreover, numerous cryptic splice sites can be present in the pre-mRNA, but they are rarely if ever used (Wang and Burge 2008). Interestingly, in human retrogenes, such cryptic splice sites can be activated due to loss of oppression upon retroposition, thus leading to intronization of exonic sequences (Kang et al. 2012). As such promiscuous splicing is not functionally relevant, it can be detrimental, explaining why these new introns are mainly observed in pseudogenes. This differs from exitrons because they are under positive selection to preserve their coding potential. However, contribution of such cryptic splice sites to the evolution of some exitrons cannot be excluded and needs to be further investigated. Our findings led us to the idea that evolution of at least some exitrons could be a consequence of intron loss, especially at the exon borders. The potential to restore splicing after intron loss is corroborated by recursive splicing of the *Drosophila Ubx* gene, where the splice sites are regenerated at the exon-exon junctions after intron splicing (Hatton et al. 1998). However, it is obvious that exitron splicing does not occur in every region or gene that lost introns. Therefore, we propose a “splicing memory” hypothesis to explain the evolution of exitron splicing. Genes, upon intron loss and retroposition, have footprints of former exon borders and thus “remember” previously defined exons (Fig. 7D). Exonic splicing regulatory sequences at the proximity to the exon borders required for splice site selection
(Reed and Maniatis 1986) can constitute such footprints and can contribute to EIS evolution. If ancestral exons were alternatively spliced, then vestigial exonic splicing regulatory elements could provide the position-dependent information on ancestral AS patterns (“splicing memory”). We found highly conserved short sequences close to the exitron in the T-protein genes that can potentially represent binding sites for RNA processing factors with functions beyond splicing. Binding of such proteins to the motifs still present in the exonic sequences may connect a region that no longer contains introns to the RNA processing and splicing network. Tethering the spliceosomal components to these regions could favor mutations beneficial for retrieval of the splicing-relevant information in response to some cue and support the emergence of splicing signals, thus restoring production of AS transcripts via EIS. This hypothesis would also apply to intronization events in noncoding regions. Interestingly, the highest number of EIS events was detected in the testis. The heritable intron loss mediated by retroposition is limited to germline cells or their embryonic precursor cells and to the genes expressed in these cells (Roy and Gilbert 2006). Therefore, our hypothesis for the evolutionary origin of exitrons that involves intron loss can potentially explain the high number of EIS events detected in the testis.

Nevertheless, in-depth studies of intron loss, ancestral AS events, conserved splicing regulatory elements, and a potential role of RNA processing factors should be performed in the future to test this hypothesis. In addition, since our hypothesis is based on analyses of a limited set of EI-containing genes, other scenarios for exitron splicing evolution cannot be excluded.

Methods

Exitron sets, features of exitrons, and exitron-containing genes

The set of A. thaliana exitrons was obtained using our RNA-seq of normalized cDNA library prepared from flowers and 10-d-old seedlings (Marquez et al. 2012). Human exitrons were identified using Ensembl-annotated intron retention transcripts (GRCh37; Flicek et al. 2013) and RNA-seq data sets for the brain, heart, liver, lung, ovary, and testis (Barbosa-Morais et al. 2012) and for breast organoids (epithelium) samples (NBS) and type ERBB2-positive breast tumor (Eswaran et al. 2013). The new release of the human genome annotation (GRCh38) does not affect the conclusions of our study as (1) the human exitron set has been obtained using CCDS (Consensus CDS) annotation (see Supplemental Methods), and (2) analyses of two very different genomes (A. thaliana and H. sapiens) show that Arabidopsis and human exitrons have very similar features and their splicing results in similar functional consequences at the protein level (see Results). Arabidopsis and human retained introns, constitutive and alternative exons are derived from our RNA-seq data (Marquez et al. 2012) and Ensembl, respectively. Mann-Whitney-Wilcoxon tests were used for test differences in GC content and size distributions. For details, see Supplemental Methods. GO classification for Arabidopsis and human El-containing sets was performed with the classification SuperViewer (http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi) and GOEAST (http://omicslab.genetics.ac.cn/GOEAST/tools.php; Zheng and Wang 2008) tools, respectively.

Validation of Arabidopsis exitron-spliced isoforms

Arabidopsis exitron-spliced isoforms were validated by in vitro transcription, EST data sets (PlantGDB [Duvick et al. 2008] and The Arabidopsis Information Resource [TAIR]), the high-resolution RT-PCR panel, Sanger sequencing, and conventional RT-PCRs (see Supplemental Methods).

Validation of translation of EIS transcripts

Translation of El-containing and El-spliced isoforms was verified in Arabidopsis with two proteogenomic data sets (Baerenfaller et al. 2008; Castellana et al. 2008), and for human, www.proteomicsdb.org (Wilhelm et al. 2014) was used (see Supplemental Methods).

Analyses of exitron splicing regulation

For Arabidopsis, RT-PCRs were done using RNA extracted from different tissues, stress conditions, and genetic backgrounds. In human, PSI values were calculated to determine differential EIS. RBMS experimental binding motifs were obtained from Fushimi et al. (2008) and Song et al. (2012). De novo motif discovery in the exitron sequences was performed with MEME from the MEME suite (http://meme.nbcr.net/meme/; Bailey et al. 2009). Binomial tests were applied for testing motif enrichment. See also Supplemental Methods.

Analysis of the impact of the exitron splicing at the protein level

PFAM protein domains, disordered regions and SLiMs were predicted using HMMER (Finn et al. 2011) (PFAM 25.0), IUPRED (short mode) (Dosztanyi et al. 2005), VSL2B (Obradovic et al. 2005; Peng et al. 2006), and the ANCHOR program (Dosztanyi et al. 2009), respectively. PTM experimental sets for A. thaliana are reported in Supplemental Table 11. Binomial tests were applied for enrichment of the protein features between exitrons and the different groups of exons. See also Supplemental Methods.

Genetic variation in exitron sequences in A. thaliana ecotypes

SNPs of 82 natural A. thaliana ecotypes were mapped against exitron sequences (see Supplemental Methods). The impact of SNPs in the 5’ and 3’ splice site signals was evaluated using PWMs. The position of the SNP in the codon was determined using the phase of the EI-containing exon.

Evolutionary studies of exitrons

Paralogs and orthologs of Arabidopsis El-containing genes were obtained from Bolle et al. (2013) (http://www.gabi-kat.de/db/duplo_genepairs.php) and by using a bidirectional best-hit approach, respectively. Plant EST data sets were obtained from The Joint Genome Institute (http://genome.jgi-psf.org/Aralys1) and from PlantGDB (http://www.plantgdb.org). For human exitron evolution, the presence and positions of introns in the orthologs were analyzed using Ensembl Compara resources and tools. See also Supplemental Methods.

Acknowledgments

We thank John W.S. Brown for valuable discussions, all the laboratory members for comments, and Wim J.J. Soppe for providing seeds of SUA mutants. This work was supported by the Austrian Science Fund (FWF) (M.K. [FWF P26333] and A.B. [DK W1207; ERA-NET Plant Genomics (PASAS) I254; SFB RNAreg F43-P10]).

References

Baerenfaller K, Grossmann J, Grobel MA, Hull R, Hirsch-Hoffmann M, Valovsky S, Zimmermann P, Grossniklaus U, Gruissem W, Baginsky S. Role of exitron splicing in Arabidopsis and human
2008. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320**:938–941.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37:W202–W208.

Bolle C, Hupe G, Kleinforming N, Haberer G, Mayer K, Leister D, Weisshaar B. 2013. GABI-DUPE: a collection of double mutants to overcome genetic redundancy in *Arabidopsis thaliana*. *Plant* **78**:157–171.

Barbosa-Morais NL, Irizarry M, Pan Q, Xiong HY, Gueroussov S, Lee UJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**:1587–1593.

Boothby TC, Zipper RS, van der Weele CM, Wolintak SM. 2013. Removal of retained introns regulates translation in the rapidly developing gametophyte of *Marsilea vestita*. *Dev Cell* **24**:517–529.

Braunschweig U, Gueroussov S, Plocik AM, Gravelle BR, Blencowe BJ. 2013. Dynamic integration of splicing within gene regulatory pathways. *Cell* **152**:1252–1269.

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gotto M, Pouton-Counsell G, Dunker AK, Bateman A, Balaji S, Fuxreiter M, Bateman A. 2012. Tissue-specific splicing of disordered segments and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res* **40**:2454–2469.

Kan Z, Garrett-Engele PJ, Johnson JM, Castle JC. 2005. Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. *Nucleic Acids Res* **33**:5659–5666.

Kang LF, Zhu ZL, Zhao Q, Chen LY, Zhang Z. 2012. Newly evolved introns in human retrogenes provide novel insights into their evolutionary roles. * BMC Evol Biol* **12**:129.

Keseler IM, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. 2013. Function of alternative splicing. *Gene* **514**:1–30.

Kim HS, Abbasi N, Choi SB. 2013. Bruto-like proteins modulate flowering time via 3′-UTR-dependent decay of SOCI mRNA. *New Phytol* **196**:747–756.

Levitan N, Alkan N, Leshkowitz D, Fruh R. 2013. Genome-wide survey of cold stress regulated alternative splicing in *Arabidopsis thaliana* with tilting microarray. *PLoS One* **8**:e66111.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci USA* **98**:11193–11198.

Liu H, Grigoriev A. 2004. Protein domains correlate strongly with exons in multiple eukaryotic genomes: evidence of exon shuffling? *Trends Genet* **20**:399–403.

Lukic S, Nicolas JC, Levine AJ. 2014. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. *Cell Death Differ* **21**:381–387.

Marques V, Loomis PA, Ehmnn S, Beales M, Potashkin JA. 2007. Regulation of retention of FosB intron 4 by PTB. *PLoS One* **2**:e828.

Martin S, Brown JW, Simpson C, McCartney MC, Bart A, Kalyna M. 2012. Transcriptomic survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res* **22**:1184–1195.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**:1593–1599.

Miedlo-Langkosch G, Kappes H, Rietherhof S, Loning T, Bamberger AM. 2003. FosB is highly expressed in normal mammary epithelia, but down-regulated in poorly differentiated breast carcinomas. *Breast Cancer Res Treat* **77**:265–275.

Ner-Gaon H, Halachmi R, Savadlo-Goldstein S, Rubin E, Ophir P, Fruh R. 2004. Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J* **39**:877–885.

O’Connor MK, Clark BJ, McLarnon AF, E’Sylva BJ, O’Donnell L, Wilce JA, Sutherland J, O’Connor AE, Whittle B, Goodnow CC, et al. 2013. Function of alternative splicing landscape in plants. *Science* **339**:1811–1819.

O’Toole R, Hattori M, Andre C, Iida K, Lurin C, Sylva RJ, O’Toole S, Linnéweber C, Sugita M, Small I. 2008. On the expansion of the pentatricopeptide repeat gene family in plants. *Mol Biol Evol* **25**:1120–1128.

Obradovic Z, Peng K, Vucetic S, Radijovic P, Dunker AK. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61**:176–182.

Pan Q, Shai O, Burge CB. 2007. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**:1413–1415.

Peng K, Radijovic P, Vucetic S, Dunker AK, Obradovic Z. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**:208.

Reddy AS, Marquez Y, Kalyna M, Bart A. 2013. Complexity of the alternative splicing landscape in plants. *Plant Cell* **25**:3657–3683.

Redd R, Maniatis T. 1986. A role for exon sequences and splice-site proximity in splice-site selection. *Cell* **46**:681–690.

Roca X, Sachidanandam R, Krainer AR. 2003. Intrinsic differences between authentic and cryptic 5′ splice sites. *Nucleic Acids Res* **31**:6321–6333.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7**:211–221.

Shaliqi R, Hurt JA, Lindquist S, Burge CB. 2014. Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell Rep* **7**:1362–1370.

Song Z, Wu P, Ji P, Zhang J, Gong Q, Wu J, Shi Y. 2012. Solution structure of the second RRM domain of RBMS and its unusual binding characters for different RNA targets. *Biochemistry* **51**:6667–6678.
Srebrow A, Kornblihtt AR. 2006. The connection between splicing and cancer. *J Cell Sci* **119**(Pt 13): 2635–2641.

Staiger D, Brown JW. 2013. Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell* **25**: 3640–3656.

Sugliani M, Brambilla V, Clerckx EJ, Koornneef M, Soppe WJ. 2010. The conserved splicing factor SUA controls alternative splicing of the developmental regulator *ABI3* in *Arabidopsis*. *Plant Cell* **22**: 1936–1946.

Syed NH, Kalyna M, Marquez Y, Barta A, Brown JW. 2012. Alternative splicing in plants: coming of age. *Trends Plant Sci* **17**: 616–623.

Tomita M, Shimizu N, Brutlag DL. 1996. Introns and reading frames: correlation between splicing sites and their codon positions. *Mol Biol Evol* **13**: 1219–1223.

Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**: 802–813.

Wang ET, Sandberg R, Luo S, Khebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**: 582–587.

Wong JJ, Ritchie W, Ebner OA, Selbach M, Wong JW, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, et al. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**: 583–595.

Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. 2012. Coordinated regulation of neuronal mRNA steady-state levels through developmentally-controlled intron retention. *Genes Dev* **26**: 1209–1223.

Yeo G, Holste D, Kreiman G, Burge CB. 2004. Variation in alternative splicing across human tissues. *Genome Biol* **5**: R74.

Zhang C, Frias MA, Mele A, Ruggiu M, Eom T, Marney CB, Wang H, Licatalosi DD, Fak JJ, Darnell RB. 2010. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**: 439–443.

Zheng Q, Wang XJ. 2008. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res* **36**: W358–W363.

Zhu Z, Zhang Y, Long M. 2009. Extensive structural renovation of retrogenes in the evolution of the *Populus* genome. *Plant Physiol* **151**: 1943–1951.

Received October 31, 2014; accepted in revised form April 30, 2015.