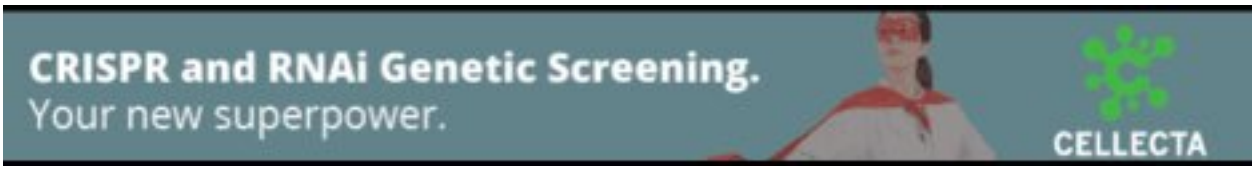
# Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*

Yamile Marquez, John W.S. Brown, Craig Simpson, et al.

*Genome Res.* published online March 5, 2012
Access the most recent version at doi:10.1101/gr.134106.111





Resource

# Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*

Yamile Marquez,[1] John W.S. Brown,[2,3] Craig Simpson,[2] Andrea Barta,[1,4] and Maria Kalyna[1,4]

[1]*Max F. Perutz Laboratories, Medical University of Vienna, A-1030 Vienna, Austria;* [2]*Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom;* [3]*Division of Plant Sciences, University of Dundee at JHI, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom*

Alternative splicing (AS) is a key regulatory mechanism that contributes to transcriptome and proteome diversity. As very few genome-wide studies analyzing AS in plants are available, we have performed high-throughput sequencing of a normalized cDNA library which resulted in a high coverage transcriptome map of *Arabidopsis*. We detect ~150,000 splice junctions derived mostly from typical plant introns, including an eightfold increase in the number of U12 introns (2069). Around 61% of multiexonic genes are alternatively spliced under normal growth conditions. Moreover, we provide experimental validation of 540 AS transcripts (from 256 genes coding for important regulatory factors) using high-resolution RT-PCR and Sanger sequencing. Intron retention (IR) is the most frequent AS event (~40%), but many IRs have relatively low read coverage and are less well-represented in assembled transcripts. Additionally, ~51% of *Arabidopsis* genes produce AS transcripts which do not involve IR. Therefore, the significance of IR in generating transcript diversity was generally overestimated in previous assessments. IR analysis allowed the identification of a large set of cryptic introns inside annotated coding exons. Importantly, a significant fraction of these cryptic introns are spliced out in frame, indicating a role in protein diversity. Furthermore, we show extensive AS coupled to nonsense-mediated decay in *AFC2*, encoding a highly conserved LAMMER kinase which phosphorylates splicing factors, thus establishing a complex loop in AS regulation. We provide the most comprehensive analysis of AS to date which will serve as a valuable resource for the plant community to study transcriptome complexity and gene regulation.

[Supplemental material is available for this article.]

Alternative splicing (AS) is a widespread mechanism which increases transcriptome and proteome complexity and controls developmental programs and responses to the environment in higher eukaryotes. The splicing process, removal of introns and ligation of exons, is performed by a large RNA-protein complex, the spliceosome, consisting of five small nuclear RNAs (snRNAs) and about 180 proteins with different functions (Wahl et al. 2009). Assembly of the spliceosome on introns in a precursor messenger RNA (pre-mRNA) is directed by *cis* elements and *trans*-acting factors (Black 2003; Stamm et al. 2005). The *cis* sequences include the splice sites, branchpoint, and polypyrimidine tract which have degenerate consensus sequences in higher eukaryotes. While many splice sites are selected in all transcripts (constitutive splicing), others are used to various levels, resulting in alternative transcripts. Selection of such alternative splice sites is affected by auxiliary *cis* elements located within exonic and intronic sequences, termed splicing enhancers and silencers. These elements are binding sites for *trans*-acting splicing factors, for example, hnRNP and SR proteins. These proteins, in addition to their functions in constitutive splicing, play a key role in AS by inhibition or promotion of selection of particular splice sites. The presence and abundance of different splicing factors in different cell types, tissues, developmental stages, and environmental conditions determines the AS profiles of expressed genes and ultimately shapes the transcriptome. In addition, alternative transcripts can code for protein isoforms with altered amino acid and domain composition affecting their activity, interaction capacity, localization, and stability, thus affecting the proteome (Stamm et al. 2005).

Alternative splicing was first described in 1977 as peculiar rearrangements in the adenovirus type 2 mRNA (Berget et al. 1977; Chow et al. 1977). Since the discovery of the first example of AS in an endogenous mammalian gene coding for calcitonin (Rosenfeld et al. 1981), the alignment of expressed sequence tag (EST) contigs to genomic DNA allowed the identification of a large number (~35%) of alternatively spliced genes in humans (Mironov et al. 1999). Estimates of AS in many different organisms have been made using EST/cDNA libraries (Okazaki et al. 2002; Zavolan et al. 2003; Iida et al. 2004; Cusack and Wolfe 2005; Wakamatsu et al. 2009). With the advent of tiling arrays and high-throughput sequencing, the number of genes which undergo AS has continued to increase (Jones-Rhoades et al. 2007; Weber et al. 2007; Kwan et al. 2008; Mortazavi et al. 2008; Pan et al. 2008). In particular, the application of high-throughput sequencing to transcriptomes (RNA-seq) has now demonstrated that AS occurs in ~95% of intron-containing genes in human (Pan et al. 2008).

In plants, estimates of the occurrence of AS have been hampered by a low number of ESTs (Brett et al. 2002). However, the levels of AS have continued to increase with greater EST/cDNA coverage: 1.2% (Zhu et al. 2003), 5% (Zhu et al. 2003), 11.6% (Iida et al. 2004), 21.8% (Wang and Brendel 2006), 29% (Xiao et al. 2005), and >30% (Campbell et al. 2006). Many transcriptome studies using high-throughput sequencing have been performed in plants, but few have been used to examine AS (Weber et al. 2007;

**[4]Corresponding authors.**
**E-mail mariya.kalyna@univie.ac.at.**
**E-mail andrea.barta@meduniwien.ac.at.**
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.134106.111. Freely available online through the *Genome Research* Open Access option.

Filichkin et al. 2010; Lu et al. 2010; Zhang et al. 2010). The most recent estimate based on RNA-seq is that ~42% of *Arabidopsis* intron-containing genes undergo AS (Filichkin et al. 2010).

In terms of identifying AS in plants, the expression profile itself influences the representation of many transcripts in databases. For example, an *Arabidopsis* transcriptome study using 454 Life Sciences (Roche) sequencing (Weber et al. 2007) showed that the top 10 most highly expressed genes represent 25% of the total mapped reads, thus tremendously compromising the representation of less abundant transcripts. To improve gene representation and discovery of AS events in *Arabidopsis*, we have used RNA-seq of a normalized cDNA library made from *Arabidopsis* seedlings and flowers. We have shown that normalization significantly increases the coverage of reads across the genes, and we have identified a large number (~47 k) of new splice junctions. Taking advantage of a high-resolution RT-PCR panel (Simpson et al. 2008a,b), we were able to validate many novel AS events. Altogether, our results show that at least 61% of intron-containing genes are alternatively spliced under normal growth conditions, which indicates a high complexity of the *Arabidopsis* transcriptome.

## Results

### RNA-seq of a normalized cDNA library improves gene representation

To facilitate the discovery of alternative splicing events in *Arabidopsis*, we generated a normalized cDNA library (Supplemental Figs. S1, S2) and subjected it to paired-end Illumina sequencing. The library was constructed from total RNA isolated from *Arabidopsis* Col-0 wild-type flowers and 10-d-old seedlings, mixed in equal proportions (Supplemental Fig. S1). Sequencing resulted in nearly 116 million paired-end reads (75-nt read length). To our knowledge, this is the first time that a normalized cDNA library of *Arabidopsis* has been subjected to high-throughput sequencing.

Paired-end reads were aligned to the *Arabidopsis thaliana* reference genome version TAIR9 (www.arabidopsis.org) using Bowtie (Langmead et al. 2009), included in the TopHat software v1.0.14 (Trapnell et al. 2009). We mapped nearly 74% of the starting reads to the reference genome, of which almost 97% aligned uniquely (Fig. 1A). Additionally, the number of aligned reads per chromosome correlated with the chromosome size (Fig. 1B), implying extensive chromosome coverage. The read distribution along each chromosome in windows of 1 kb is shown in Figure 1C (see Methods).

To test whether the uniqueness of the reads was due to the length of the read and/or the cDNA normalization procedure, we compared our results with the 36-nt single-end read Illumina sequences from non-normalized cDNA libraries published previously (Filichkin et al. 2010). In addition, to allow a direct comparison of reads of the same size, we truncated our 75-nt-long reads to 36 nt (Supplemental Table S2). These comparisons suggested that the increased read length and cDNA normalization both significantly influenced read mapping, resulting in an increase in the total number of aligned reads and a higher proportion of uniquely mapped reads (Supplemental Table S2).

Most of the reads mapped to annotated genes (Supplemental Fig. S5A), illustrating that the majority of genes have been predicted in the TAIR9 annotation. Examination of the median coverage along the transcription unit (Supplemental Fig. S5B; see Methods) revealed that the cDNAs are highly covered in almost

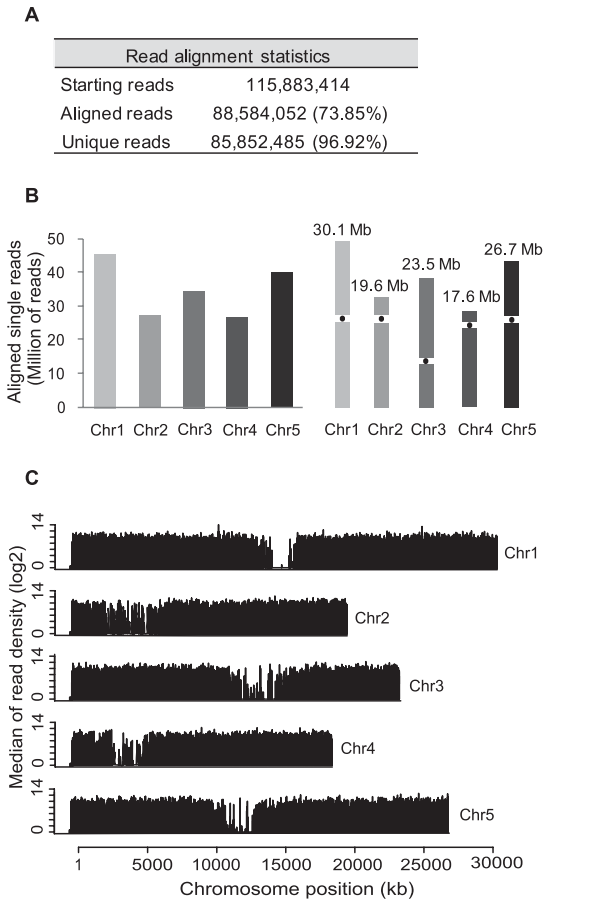



**Figure 1.** Read alignment using Bowtie and TopHat. (*A*) Table showing statistics of the aligned reads to the *Arabidopsis* genome (TAIR9). (*B*) Aligned single reads to *Arabidopsis* chromosomes (*left*) and schematic representation of *Arabidopsis* chromosomes (*right*). (*C*) $Log_2$ scale of median read density in windows of 1 kb by chromosome.

their entire length (median of 600 reads in 80% of the total length of the transcription unit) (Supplemental Fig. S5B).

To test the efficiency of normalization in reducing the abundance of highly expressed transcripts, we analyzed the read coverage of the top 10 most abundantly expressed genes in our normalized library and showed that they only contributed 1.05% of the total of reads (Supplemental Material; Supplemental Table S1). This level of coverage was substantially lower than found in non-normalized libraries sequenced with the Illumina (Filichkin et al. 2010) and 454 (Weber et al. 2007) platforms, where the proportion of reads in the top 10 most abundant genes was 26.76% and 7.76%, and 25%, respectively (Supplemental Table S1). The much lower read coverage of highly expressed genes in our libraries was independent of the read length and paired-end protocol (Supplemental Table S1; Supplemental Fig. S3A,B). A detailed description of the normalization efficiency is given in the Supplemental Material (Supplemental Data; Supplemental Table S1; Supplemental Figs. S3, S4). In general, the comparison between library preparation and sequencing methods suggested that the normalized library achieved much better gene coverage and significantly influenced read mapping, resulting in an increase in the total number of aligned reads and a higher proportion of uniquely mapped reads.

## High level of detection of novel splice junctions in *Arabidopsis*

For splice junction detection, we used TopHat (Trapnell et al. 2009). To reduce the number of false positives, the minimum intron size was set to 60 nt, which is close to that established experimentally (Goodall and Filipowicz 1990) This is in contrast to other studies (Filichkin et al. 2010; Zhang et al. 2010) which used smaller intron sizes of 20 nt and 1 nt, respectively. We filtered the splice junctions originally detected (using alignment with two mismatches) (see Methods) by two criteria: quality of the alignment and coverage of the splice junction (Supplemental Fig. S1). Splice junctions that were obtained from the alignment with two mismatches (see Methods) were removed if the splice junction in question was within 10 nt of another splice junction with better support and if it was not supported by the alignment of perfectly aligned reads (alignment with no mismatches) (Supplemental Fig. S1; see Methods). For the remaining junctions, we have only included those supported by at least three reads (two mismatches). Our results showed that the majority of the splice junctions reside in annotated genes (131,523 of 149,925) (Table 1) and that, in protein-coding genes, they are located mainly in the coding sequence (CDS) (Fig. 2A), highlighting the potential of splice junctions to affect the final protein sequence. From the total of predicted splice junctions, 46,955 (31%) were novel and not annotated in the TAIR9 Database (Table 1). This is most likely due to the use of a normalized library together with longer reads which have considerably improved splice junction detection (Supplemental Table S2).

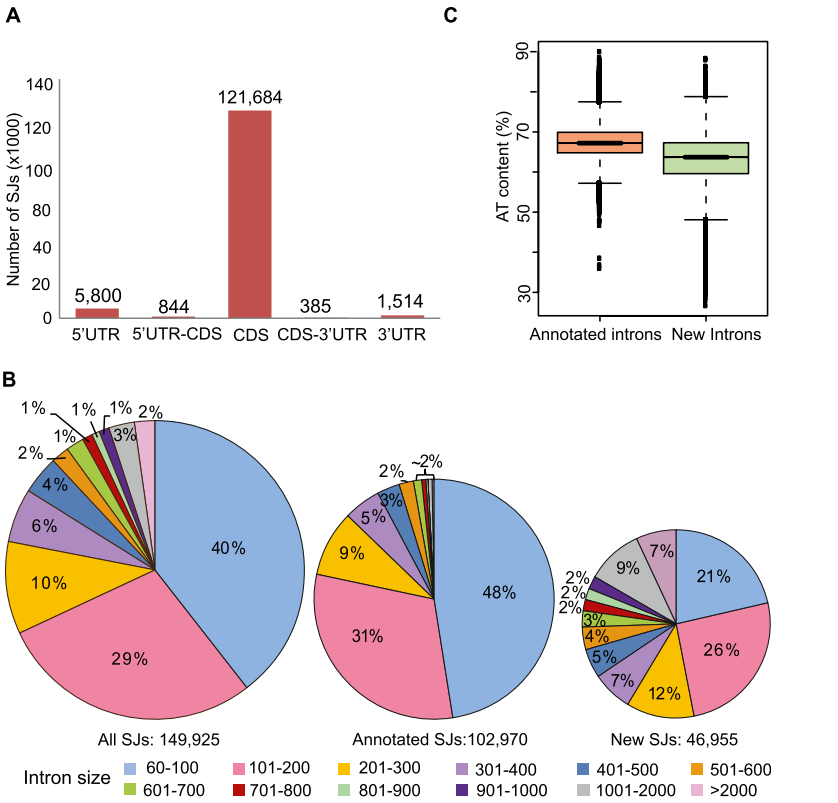



**Figure 2.** Features of splice junctions and predicted introns. (*A*) Distribution of splice junctions along gene features in protein-coding genes, according to TAIR9. (*B*) Distribution of intron sizes defined by splice junctions (SJ) predicted by TopHat (*left* circle). Splice junctions were classified as annotated if they were already in TAIR9 (*middle* circle); otherwise, they were classified as new (*right* circle). (*C*) Distribution of the AT content (%) in introns generated by annotated splice junctions and new splice junctions.

## The majority of splice junctions define typical plant introns

In plants, several features that impact splicing efficiency have been described. These include a minimum intron size, sequences in the 5′ and 3′ splice sites, the branch point, and a minimum percentage of AU-richness (Goodall and Filipowicz 1990; Brown and Simpson 1998; Lorkovic et al. 2000). Although the presence of these signals will define an intron, it is the combination and strength of each signal which determines the splicing efficiency of intron removal (Brown and Simpson 1998).

Inspection of the intron sizes produced by all of the predicted splice junctions had a mean of 298 nt (median = 114 nt) and showed that the majority (~70%) of them were smaller than 200 nt (Fig. 2B). However, the sizes of introns generated by the new splice junctions were bigger than those from the junctions annotated in TAIR (mean = 588.7 nt versus mean = 166.5 nt, respectively, Mann-Whitney-Wilcoxon test, $P$-value < 0.00001). This was also highlighted by only ~47% of them being smaller than 200 nt (Fig. 2B). This is probably due to many of the novel splice junctions defining new AS events and that introns with AS tend to be larger than introns with no evidence of AS (see Fig. 5A, below).

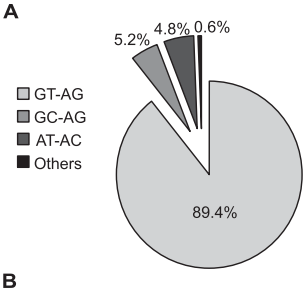
Examination of the dinucleotides at the intron borders exhibited an increase in the number of GC-AG and AT-AC splice sites in comparison to TAIR9 annotations (5.2% and 4.8% vs. 1% and 0.6%, respectively) (Fig. 3A). In order to demonstrate that the introns predicted by these splice junctions were bona fide introns, we looked for sequence signatures of plant introns using position weight matrices (PWMs) (Sheth et al. 2006). Using PWMs of the 5′ and 3′ splice sites of U2 and U12 annotated introns in *Arabidopsis*, together with their respective branch point signatures (Fig. 3B; see Methods), we were able to classify ~93% of the introns. As expected, most of them were classified as U2 introns, but we also identified 2069 potential U12 introns in genic regions (Fig. 3A;

**Table 1.** Predicted splice junctions (SJs)

| | Score ≥ 1 | Score ≥ 2 | Score ≥ 3 |
|---|---|---|---|
| Total of SJs | 172,651 | 159,758 | 149,925 |
| SJs in annotated genes | 144,628 | 137,409 | 131,523 |
| Annotated SJs | 105,237 | 104,070 | 102,970 |
| New SJs | 67,414 | 55,688 | 46,955 |
| Genes with at least one intron[a] | 20,144 | 19,800 | 19,505 |
| Genes with overlapping introns[a] | 11,681 (58.0%) | 10,879 (54.9%) | 10,103 (51.8%) |

[a]Introns are defined by the predicted splice junctions (SJs) (see text).

A




B

| | U2 introns | U12 introns | NC | Total |
|---|---|---|---|---|
| Genic | 124,383 | 2,069 | 5,071 | 131,523 |
| Antisense | 6,071 | 193 | 4,765 | 11,029 |
| Intergenic | 6,086 | 150 | 1,137 | 7,373 |
| Total | 136,540 | 2,412 | 10,973 | 149,925 |

**Figure 3.** Classification of introns defined by predicted splice junctions. (*A*) Splice sites of predicted introns. (*B*) Classification of introns according to Sheth et al. 2006 (see Methods). (NC) Not classified.

Supplemental Fig. S6; see the list of U12 introns in the Supplemental Material), and only 165 and 246 of such introns were identified computationally (Zhu and Brendel 2003; Alioto 2007). Functional classification (GO annotation) of the genes possessing U12 introns shows that they are 2.7-fold enriched in DNA and RNA metabolism class (*P*-value = $1.039 \times 10^{-10}$; data not shown).

We also examined the AU-richness of introns predicted by the splice junctions and showed that ~92% had >59% AU, which is the minimum AU% required for an intron to be efficiently spliced in dicots (Goodall and Filipowicz 1989; Fig. 2C). The AU-richness of introns predicted by the novel splice junctions was slightly lower (~64%) than that predicted by the known (annotated) splice junctions (~67%) (Fig. 2C). This might reflect the fact that the majority of the novel splice junctions are derived from new AS events (see below). Together, the results suggest that a high proportion of the introns generated from the predicted splice junctions in this study have typical characteristics of introns in *Arabidopsis*.

To investigate the potential of the introns (defined by the predicted splice junctions) to produce alternative transcripts, we looked for overlapping introns inside the same gene, as two overlapping introns cannot coexist in the same transcriptional unit. Using this approach, we estimate that 51.8% of the genes in our sample have more than one transcript and, therefore, have evidence of AS (Table 1). It is important to note that this analysis of overlapping introns does not provide data on intron retention events.

## Assembly of putative transcripts predicts that 61.2% of intron-containing genes are alternatively spliced

To determine putative transcripts generated by RNA-seq, we have used the gene coordinates from TAIR9, the aligned reads, and the predicted splice junctions that have passed the quality criteria described above. Putative exons were defined as continuous stretches with at least four reads of coverage. The final transcripts were assembled by joining the putative exons with the predicted junctions. To test whether the predicted transcripts were supported by the read alignment, the putative transcripts were evaluated using Cufflinks (Trapnell et al. 2010). From this analysis, a total of 57,447 transcripts were successfully assembled that corresponded to 23,901 annotated genes (Table 2; see the set of assembled transcripts in the Supplemental Material). In agreement with what was observed in the read alignment (Fig. 1B), the number of genes and transcripts that were assembled by chromosome also correlates with the chromosome size (Table 2). Moreover, in the final set of predicted transcripts, 91% of the genic splice junctions reported in Table 1 (score ≥ 3) are contained in at least one putative transcribed unit.

Next, we evaluated whether the set of putative assembled transcripts corresponded to known isoforms in TAIR9. For this purpose, we calculated how many genes with assembled transcripts have at least one isoform that resembled an annotated transcript in TAIR9, using the definition that the assembled transcript covered contiguously at least 80% of the total length of the annotated isoform. We found that 75% of genes with assembled transcripts had at least one transcript that is annotated in TAIR9 (data not shown).

We estimated the number of *Arabidopsis* intron-containing genes with AS by calculating the ratio of genes with more than one putative transcript divided by the number of genes with at least one splice junction in our sample (intron-containing genes). Therefore, 61.2% of genes with introns in our sample have evidence of AS (Table 2). The difference between this value and the estimate of 51.8% based only on overlapping introns (Table 1) reflects the addition of intron retention transcripts and the filtering criteria which results in loss of some splice junctions (only 91% of splice junctions are included).

To determine the types of AS events in our set of assembled transcripts, we used ASTALAVISTA software (Foissac and Sammeth 2007). Intron retention is the most common event (~40%, Figure 4A; Supplemental Fig. S7) agreeing with previous studies in plants (Iida et al. 2004; Ner-Gaon and Fluhr 2006; Wang and Brendel 2006; Barbazuk et al. 2008; Filichkin et al. 2010) although it is not as high as the estimate of ~56% of the total events (Wang and Brendel 2006). We found that the use of alternative 3′ splice sites is more than twofold more frequent than the use of alternative 5′ splice sites (Fig. 4A; Supplemental Fig. S7), consistent with prior findings (Iida et al. 2004; Wang and Brendel 2006). Moreover, our transcripts show that while skipping of a single exon is a rare event (2.73% of the total events), it is relatively common that multiple exons are skipped together and that exon skipping can utilize alternative 5′ and/or 3′ splice sites (Fig. 4A; Supplemental Fig. S7). In general, ~42% of the events in the assembled transcripts are complex (Fig. 4A; Supplemental Fig. S7).

**Table 2.** Putative assembled transcripts and AS estimates

| Putative assembled transcript statistics | | |
|---|---|---|
| **Chr** | **Gene number** | **Transcript number** |
| Chr1 | 6,253 | 14,888 |
| Chr2 | 3,761 | 9,391 |
| Chr3 | 4,750 | 11,253 |
| Chr4 | 3,610 | 8,935 |
| Chr5 | 5,527 | 12,980 |
| Total | 23,901 | 57,447 |

| Alternative splicing estimates | | | |
|---|---|---|---|
| | RNA-seq (this study) | TAIR9 | TAIR10 |
| Intron-containing genes | 18,719 | 22,403 | 22,523 |
| AS genes | 11,465 | 4,626 | 5,885 |
| % of AS | 61.2% | 20.7% | 26.1% |

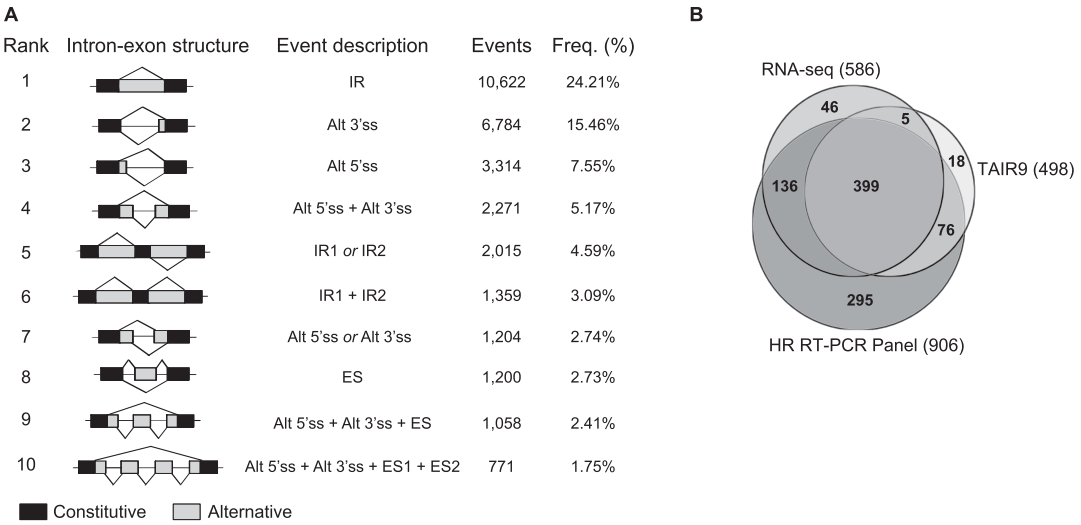



**Figure 4.** Alternative splicing events and validation of assembled transcripts. (*A*) Top 10 most frequent types of AS in the predicted transcripts according to ASTALAVISTA. The first column illustrates the intron-exon structure of the AS event, followed by its description, the raw number of events found in the sample, and their frequency. (IR) Intron retention, (Alt 5′ss) alternative 5′ splice site, (Alt 3′ss) alternative 3′ splice site, (ES) exon skipping. (*B*) Venn diagram of the number of fragments obtained in the HR RT-PCR panel, putative assembled transcripts of RNA-seq, and TAIR9-annotated transcripts according to primer pairs of the HR RT-PCR panel (see text and Methods).

## Validation of assembled transcripts

To validate our set of putative assembled transcripts (PATs), we have used a high-resolution RT-PCR alternative splicing panel (HR RT-PCR panel), which allows monitoring of AS events in multiple genes under different conditions (Simpson et al. 2008a,b). This HR RT-PCR panel is capable of distinguishing AS events involving small size differences in transcripts (as few as 2–3 nt). We analyzed 256 different genes containing AS events which were either published, annotated in TAIR, or found in AS databases (Supplemental Table S3). Applying the HR RT-PCR panel primers to our assembled transcripts, we extracted the sizes of expected RT-PCR products for each of the genes on the panel and compared them to actual products detected on the panel and to predicted products from TAIR9-annotated AS transcripts. Of the 586 products predicted from the RNA-seq assembled transcripts, 92% were supported by at least one of these two resources (Fig. 4B; Supplemental Table S4). Furthermore, 136 (23.2%) of the predicted products are not described in the TAIR9 database but were confirmed by products on the HR RT-PCR panel (Fig. 4B; Supplemental Table S4), again illustrating that many AS events are not annotated. As expected, the HR RT-PCR panel detected the highest number of AS events. This is due, first, to its sensitivity and second, to the different samples and conditions employed. Nevertheless, 33 transcripts of low abundance (relative transcript abundance of <2%) in the HR RT-PCR panel were also indentified by RNA-seq. Moreover, RNA-seq identified a number of AS events which had not been scored on the HR RT-PCR panel previously because they represented very small peaks but were now confirmed by manual inspection. Additionally, some RT-PCR products were subjected to Sanger sequencing, and a total of 124 sequences supported the assemblies in our RNA-seq experiment. Altogether, the results highlight the validity of RNA-seq to detect novel AS events and demonstrate that AS is much more extensive in *Arabidopsis* than previously thought.

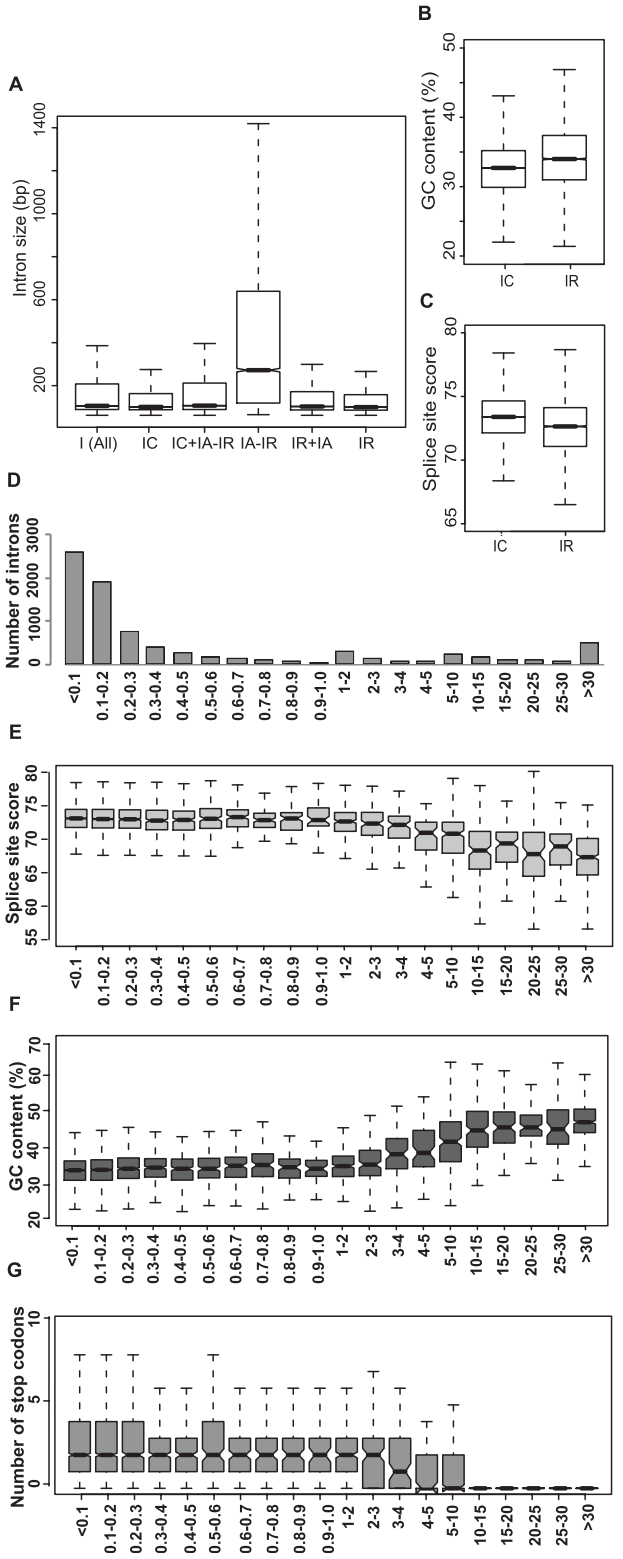
## Low abundance of many intron retention events

Examination of the putative assembled transcripts showed that the predominant type of AS event is intron retention, with ~40% of the total events (Fig. 4A; Supplemental Fig. S7; and see above). We analyzed the characteristics of retained introns identified by RNA-seq and transcript assembly. We compared the size of retained introns (IR) to the sizes of introns with no evidence of retention or any other types of AS (constitutive intron - IC) and to alternatively spliced introns with no evidence of intron retention (IA-IR). We found that retained introns were significantly smaller than non-retained introns which were involved in other types of AS (IR vs. IA-IR, mean 161.39 bp vs. 548.86 bp, $P$-value < 0.00001) (Fig. 5A). On the other hand, there was no significant difference in the size distribution of retained introns compared to constitutive introns (IR vs. IC, $P$-value = 0.5) (Fig. 5A). Therefore, intron size cannot be used as a parameter to distinguish retained introns. Nevertheless, we found that retained introns had a higher GC content ($P$-value < 0.00001) (Fig. 5B) and weaker splice signals ($P$-value < 0.00001) (Fig. 5C) than constitutive introns, which might contribute to lower splicing efficiency and, therefore, a higher probability of intron retention.

We obtained an indication of the degree to which an intron is retained by calculating the median number of reads along the retained intron divided by the number of reads supporting the splice junction in question (see Methods). This value, the intron retention ratio (IRR), reflects to some extent the preference of an intron to be retained. Nearly 73% of the retained introns had an IRR of <1 and around 48% had an IRR of <0.2 (Fig. 5D), suggesting that these introns have a lower read coverage. Deeper analysis of the intron retention events showed that the preference for removing an intron (low IRR) correlated to some extent with stronger splice sites in comparison with introns that have higher IRR ($R^2 = -0.36$, $P$-value = $2.7 \times 10^{-269}$). For example, introns with an IRR < 4 have stronger splicing signals ($P < 0.05$) (Fig. 5E).

We, therefore, analyzed the abundance of intron retention events on the HR RT-PCR panel. Primers were designed to amplify across at least two introns, including the intron(s) which had been described in different plant databases as retained. Of 51 intron retention events from 45 genes, 33 were detected by RT-PCR and were supported by RNA-seq. Nine of the IR products were not de-

tectable on the HR RT-PCR panel, and eight of these were not detected by RNA-seq (Supplemental Table S5). For a further nine IR events, the IR products were barely detectable on the HR RT-PCR panel (<1% of the total transcripts from the gene) and had low coverage of RNA-seq reads (Supplemental Table S5; see also Supplemental Fig. S9). Thus, there is good agreement between the read coverage and detection by RNA-seq and the relative abundance of intron retention transcripts detected experimentally by the HR RT-PCR panel. Based on the lack of detection or very low abundance of some IR events which appear in plant databases, it is most likely that many are derived from ESTs representing partially spliced mRNAs or DNA artifacts.




## Identification of an abundant class of cryptic introns

We examined the distribution of the retained introns along the transcripts to see whether they distribute differently in comparison to all the introns (Fig. 2A). We obtained a significant enrichment of retained introns in the 5′ and 3′ UTRs of genes (twofold and 3.4-fold, respectively, $\chi^2$ test, $\chi^2 = 1176$, $P$-value $= 2.54 \times 10^{-253}$) (Supplemental Table S6), suggesting that intron retention is selected against in the coding regions. This observation appeared to contradict the finding that 1838 introns were preferentially retained in our sample (IRR > 1), and many had very high IRR scores (Fig. 5D). More detailed analysis of these introns with IRR > 1 showed that 66.86% (1227) were inside annotated exons. In addition, introns with an IRR > 3 had higher GC content than introns with lower IRRs ($P$-value < 0.05) (Fig. 5F) and, in general, retained introns are more GC rich than nonretained introns ($P$-value < 0.0001) (Fig. 5B; Supplemental Fig. S8). The position of these retained introns within annotated exons and the higher GC content suggests that they are exonlike. This is also supported by an analysis of the protein coding capacity of these retained intron sequences. First, they contribute significantly to known protein domains compared to nonretained introns or retained introns from UTRs (Supplemental Tables S7, S8), and second, the retained introns have a significantly lower proportion of stop codons than expected from their nucleotide frequency ($\chi^2$ test, $\chi^2 = 1313.7$, $P$-value $= 1.19 \times 10^{-287}$). In particular, retained introns with an IRR > 4 have fewer stop codons ($P$-value < 0.05) (Fig. 5G). Additionally, the codon usage of retained introns with an IRR > 3 is more similar to the codon usage of exons than from nonretained introns (Supplemental Table S9; see Methods).

The exonic features in these retained introns suggest that they represent cryptic introns rather than retained introns. In the following analysis, we define a cryptic intron as an intron with both splice sites inside an annotated coding exon. To validate the removal of such exonic regions identified by RNA-seq, we selected 10 cryptic introns with different IRRs and tested them by RT-PCR. We used cDNA from our initial RNA preparations to exclude any possibility that our results are the consequence of the library preparation or computational analysis. Primer pairs were designed across more than one exon to distinguish any DNA contamination or unprocessed products. From the 10 genes tested, nine showed

**Figure 5.** Features of retained introns. (*A*) Size distribution of different intron classes. [I(All)] All introns in our sample, (IC) the constitutive introns, (IC+IA−IR) and (IA−IR) the categories without retained introns either including or excluding constitutive introns, respectively, (IR+IA) retained introns that are also involved in other AS events, (IR) introns that are only retained. (*B*) GC content distribution of constitutive introns (IC) and retained introns (IR). (*C*) Splice site score distribution of constitutive introns (IC) and retained introns (IR). (*D*) Histogram of number of introns in each intron retention ratio (IRR) category. (*E*) Boxplots of splice site score distribution by IRR category. (*F*) Boxplots of GC content distribution by IRR category. (*G*) Boxplots of stop codons distribution by IRR category. For the splice site score distributions in *C* and *E*, the means of 5′ and 3′ splice sites scores (calculated according to Sheth et al. 2006) were used.

an RT-PCR product that corresponded to cryptic intron splicing (Supplemental Fig. S10; Supplemental Table S10).

In total, we have identified 1289 cryptic introns in annotated coding exons (14.1% of all retained introns). Almost half of these cryptic introns (602) are spliced out in frame, thus having a potential of removing amino acid stretches within a full-length protein and generating a new protein isoform.

### Alternative splicing of *AFC2* Clk/STY (LAMMER-type) kinase

To demonstrate the utility of our RNA-seq data, we have investigated the AS of the *AFC2* gene. AFC2 belongs to Clk/STY (LAMMER-type) protein kinases involved in regulation of constitutive and AS in plants and animals through phosphorylation and interaction with serine/arginine-rich (SR) proteins (Colwill et al. 1996a,b; Duncan et al. 1997; Golovkin and Reddy 1999; Savaldi-Goldstein et al. 2000, 2003). Extensive regulation of Clk/STY kinase genes by usage of alternative promoters and/or AS has been shown in animals (Duncan et al. 1995, 1997; Kpebe and Rabinow 2008), but little is known about regulation of these genes in plants except that two splice variants of *AFC2* have been reported in TAIR.

Our RNA-seq data detected the two splice variants annotated in TAIR and a total of 10 new putative assembled transcripts with an RPKM (reads per kilobase of exon model per million mapped reads) value $\geq 1$ for *AFC2* (Fig. 6A). By comparison to the TAIR reference model (AT4G24740.1), the PATs showed retention of introns 1, 2, 4, 5, 4 + 5, 6, 9, and 10, skipping of exons 2, 3, and 5 + 6, and the use of an alternative 5′ splice site in exon 2 and an alternative 3′ splice site in exon 4. We have validated experimentally the majority of the novel AS events and assembly of the PATs by the HR RT-PCR panel and Sanger sequencing of the RT-PCR products using two primer pairs designed to exons 1 and 4 (primer pair 227) and to exons 4 and 7 (primer pair 226) (Supplemental Table S4). Products corresponding to skipping of exon 3 (PAT3, 278 bp), retention of intron 1 (PAT1, 960 bp, product size is outside the range of HR RT-PCR), and retention of intron 5 (PAT9) were not observed. Retention of intron 4 (PAT10, 460 bp) or intron 6 (PAT8, 418 bp) products are barely detectable by the HR RT-PCR panel (Fig. 6A,B, upper panel), in agreement with RPKM values of the intron retention events (PAT7-10) being much lower than for the two skipped exons 5 and 6 (PAT4-6) or fully spliced PATs (data not shown).

To examine the consequences of the AS events in *AFC2,* we have predicted the putative proteins, taking into consideration alternative transcription start sites (TSSs) and two possible translation initiation sites (TISs) reported in TAIR9 (Fig. 6A). The full-length protein of 427 amino acids containing the N-terminal noncatalytic domain followed by the catalytic kinase domain is encoded by AT424740.1. Translation of other PATs results in either N-terminally or C-terminally truncated proteins which lack either the putative nuclear localization signal encoded by the first exon or the highly conserved LAMMER motif, respectively (Fig. 6A). Interestingly, protein alignment shows that the alternatively spliced exon 2 of *AFC2* corresponds to exon B which undergoes skipping in mammalian orthologs (Duncan et al. 1995; data not shown). It remains to be seen whether *AFC2* putative protein isoforms are, indeed, produced and what are the functional consequences of these variations.

The majority of the AS events in *AFC2* generate premature termination codons (PTCs) which are followed by downstream splice junctions (Fig. 6A) and may, therefore, be turned over by nonsense-mediated decay (NMD). We have analyzed the abundance of these PTC+ transcripts in *upf1-5* and *upf3-1* mutants that are impaired in the NMD pathway using the HR RT-PCR panel. The transcripts with skipped exons 5 and 6 showed a significant increase in the NMD mutants compared to the wild-type control, in contrast to PTC+ PATs with intron retention events (Fig. 6B). These results are in agreement with our recent data showing that intron retention containing transcripts are not targeted to the NMD pathway (Kalyna et al. 2011). Interestingly, intron retention transcripts have also been identified for the orthologous murine *Sty* gene and were shown to be retained in the nucleus (Duncan et al. 1995) and therefore possibly avoid NMD. This example shows that identification of previously unknown AS transcripts of *AFC2* by our RNA-seq experiment can provide insight into the regulation of this important gene and establish the basis for future experimental validation.

## Discussion

Our RNA-seq analysis provides the most comprehensive information on alternatively spliced transcripts in *Arabidopsis* to date and will be of great value in addressing regulation of expression and gene/protein function. We have performed a genome-wide investigation of alternative splicing in *Arabidopsis* by RNA-seq and have demonstrated that over 61% of intron-containing genes are alternatively spliced. The significantly increased occurrence of AS found in this study is due to the use of a normalized cDNA library combined with 75-nt paired-end sequencing reads. The normalization process led to enhanced gene representation due to a significant reduction of read coverage of highly expressed genes, and the longer reads improved splice junction detection (Supplemental Tables S1, S2). The analysis detected almost 150,000 splice junctions, of which 31% were novel. The majority of these splice junctions define introns with features of typical plant introns, providing confidence in the splice junctions identified. The high proportion of splice junctions predicting high confidence introns reflects the filtering parameters based on experimental evidence that we have used in this study. In addition, we identified about eight times more (2069) U12 introns than previously found (165 and 246, Zhu and Brendel 2003, and Alioto 2007, respectively). This comprehensive set of U12 introns will stimulate research of the significance and evolution of minor spliceosomal introns.

Estimates of the number of genes showing AS were derived in two ways. First, the analysis based on overlapping introns which did not take intron retention events into account showed that around 52% of intron-containing genes were alternatively spliced. Second, putative transcripts were assembled from exons (defined by reads) and the splice junctions that were filtered on the basis of read coverage. This approach now including intron retention events indicates that 61.2% of intron-containing genes undergo AS. Although this number is significantly higher than previously reported, we expect that it still represents an underestimate because we used only flower and seedling material grown under normal conditions. Therefore, we have probably not captured AS events specific for other organs, developmental stages, and environmental conditions (abiotic and biotic stresses). Indeed, recent RNA-seq analysis showed that many AS events are stress-associated (Filichkin et al. 2010).

Assembly of transcripts from short sequencing reads remains a computational challenge. It is, therefore, essential to obtain an experimental support for an assembly procedure. We have taken advantage of a high-resolution alternative splicing RT-PCR panel to validate a large set of AS events and transcripts. This method

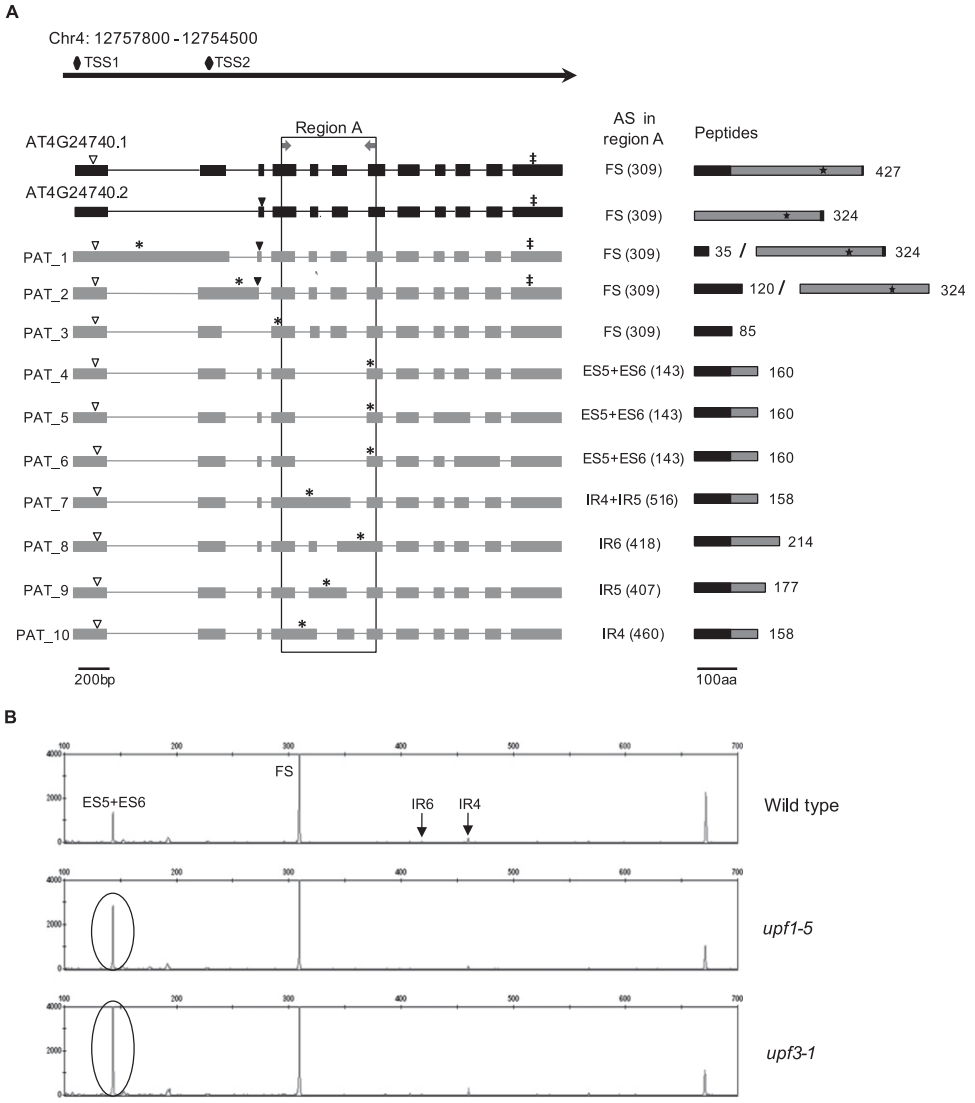



**Figure 6.** Alternative splicing in the *AFC2* gene. (*A*) The black arrow represents the chromosome location of *AFC2* and its orientation indicates the direction of transcription. (Diamonds) Two transcription start sites (TSS) reported by TAIR. (Black) TAIR models for the *AFC2* gene; (gray) new putative assembled transcripts (PATs). The triangles located over the TAIR transcripts and PATs indicate the start codons, while the asterisks or double daggers represent the stop codons. For PAT1 and PAT2, two putative start codons (full and empty triangles) and two putative stop codons (asterisk and double dagger) are depicted. Protein isoforms predicted using the shown start and stop codons are illustrated at the end of each splice variant followed by their sizes in amino acids. (Black) Noncatalytic domain; (gray) kinase catalytic domain. (Star) Location of the LAMMER motif. (Arrows) The primer pair 226 designed for the HR RT-PCR panel. The amplified region is highlighted by a box (region A). The AS events in the region A are shown after each transcript, and the number in parentheses denotes the size of the amplified product in nucleotides. Abbreviations for AS events: (FS) fully spliced, (IRn) intron retention where "n" is the number of the intron, (ESn) exon skipping of exon number "n." (*B*) Electropherograms of RT-PCR products for the primer pair indicated in *A* for wild type and *upf* mutants generated by GeneMapper. Numbers on the *x*-axis represent the size markers in bp; numbers on the *y*-axis represent relative fluorescence, reflecting transcript abundance. The AS event associated with each product is shown in the respective peak. (Ovals) The peaks which increase in abundance in the *upf1-5* and *upf3-1* mutants.

(and Sanger sequencing, when necessary) achieved an excellent level of validation, as we obtained experimental support for 535 of 586 transcripts predicted by RNA-seq for the regions covered on the HR RT-PCR panel. On the other hand, a considerable number of the HR RT-PCR panel transcripts were not identified in this RNA-seq sample (Fig. 4B). This is most probably due to the AS information on the HR RT-PCR panel being derived from a pool of different plant material and growth conditions. In addition, some of the AS transcripts unique to the HR RT-PCR panel were, in fact, explained by RNA-seq splice junctions which were lost due to stringent filtering parameters used in the assembly process. The additional transcripts on the HR RT-PCR panel also suggest that our estimate of the occurrence of AS is quite conservative.

Our transcript assemblies allowed us to identify the frequency of different types of AS events. In particular, we observed a complex alternative splicing landscape where multiple splicing events occurred within the same transcript, raising the possibility that these events are either independent or coordinated. In our analysis, intron retention was the most prominent event, and we conducted a genome-wide analysis of the features of retained introns from our RNA-seq data. In contrast to what was observed in animals (Stamm et al. 2000; Galante et al. 2004; Sugnet et al. 2004;

Zheng et al. 2005; Sakabe and de Souza 2007) and suggested for plant introns (Wang and Brendel 2006), we did not detect a correlation between the retention of an intron and its size. Therefore, in *Arabidopsis*, intron size cannot be used to predict whether an intron will be retained or not. However, introns alternatively spliced by other mechanisms than intron retention were considerably larger than the average intron size.

To further analyze intron retention, we developed an approach to indicate the potential for introns to be retained (the intron retention ratio). The use of IRR allowed us to detect many introns which have a high level of retention and a high GC content. Further examination led to the identification of a set of introns (1289) located within annotated coding exons which we define as cryptic introns. Interestingly, 602 of these cryptic introns are excised in frame and, therefore, would generate protein isoforms by removing a stretch of amino acids within a protein. We also identified a set of introns with low IRRs, many of which had low read coverage. By analyzing all of the intron retention events contained on the HR RT-PCR panel, we showed that ~20% were not detectable by RT-PCR or RNA-seq and a further 20% were barely detectable on the panel but were supported by RNA-seq at low read coverage. These low abundance IR events which are annotated in databases potentially could represent incompletely spliced mRNAs or DNA artifacts in cDNA/EST preparation. The impact of intron retention on alternative splicing in *Arabidopsis* in general is illustrated by the fact that, although IRs represent 40% of AS events, only 23.6% of the assembled transcripts contain one or more retained introns. Indeed, if these transcripts are discounted, 51.3% of intron-containing genes have an alternative transcript which does not include intron retention. This agrees well with our estimate based on overlapping introns from our set of predicted splice junctions (51.8%). Therefore, while intron retention is still the most abundant AS event in plants (~40%), the significance of intron retention for alternative splicing in plants is obviously much less than has been thought previously.

Alternative splicing has recently caught the attention of many plant researchers, as many developmental processes and response to environmental cues have been shown to be controlled by AS (for review, see Ali and Reddy 2008; Terzi and Simpson 2008; Petrillo et al. 2011). An essential level of regulation of pre-mRNA splicing is achieved through phosphorylation of SR proteins (Cao et al. 1997). To illustrate the potential application of our RNA-seq data, we have examined the AS of *AFC2*, coding for a highly conserved Clk/STY (LAMMER-type) kinase, involved in modulation of pre-mRNA splicing through the phosphorylation and interaction with the SR proteins and other splicing factors (Golovkin and Reddy 1999; Savaldi-Goldstein et al. 2000, 2003). In animal systems, alternative transcripts of Clk kinases generated through exon skipping contain PTCs (Duncan et al. 1995, 1997) and are probably degraded by NMD (Hillman et al. 2004). We have demonstrated that AS coupled to NMD plays an important role in the regulation of *AFC2* which, in turn, modulates splicing factors like SR proteins, suggesting a regulatory loop that results in a fine-tuned regulation of AS. Importantly, NMD affects levels of *AFC2* PTC+ transcripts generated through exon skipping; however, no effect was observed on PTC+ splice variants generated by intron retention. This observation is in agreement with our previous findings showing that, although many intron retention transcripts possess NMD features (such as PTCs followed by downstream splice junctions), they are not sensitive to NMD (Kalyna et al. 2011). Therefore, a PTC created by AS is not necessarily a signature to trigger NMD. Furthermore, AS events causing NMD might occur elsewhere in a transcript, making it necessary to study AS linked to NMD at the level of full-length transcripts in order to predict their fates. This will be greatly aided by the advent of high-throughput sequencing technologies providing longer reads.

In summary, we provide a high confidence set of AS events in *Arabidopsis* on a genome-wide level and, therefore, offer a valuable resource in understanding gene regulation. The question remains to what extent the extensive AS described here contributes to regulation of expression and proteome diversity in plants. Our data will be extremely valuable in addressing these aspects of the biological functions of AS. In addition, combining analyses of ecotype genetic variation and AS profiles would add tremendously in dissecting phenotypic features and mechanisms underlying the plasticity of plant development and stress response.

## Methods

### cDNA library preparation and high-throughput sequencing

Total RNA was extracted using RNeasy Mini Kit (Qiagen) from wild-type *Arabidopsis thaliana* (Col-0) flowers of different stages and 10-d-old seedlings grown at 22°C, 60% humidity, light intensity of 150 $\mu$mol $m^{-2}$ $s^{-1}$, and 16-h light/8-h dark cycle. Total RNA samples from seedlings and flowers were mixed in a 1:1 ratio for the synthesis of the cDNA library. A full-length oligo-dT cDNA library was constructed using a Mint-Universal Kit (Evrogen), followed by cDNA library normalization using a Trimmer-direct Kit (Evrogen). The resulting library was fractionated using a Covaris sonicator (cycle 15%, intensity 6.0, cycles burst: 250), and fragments in a size range of 200–800 nt were purified using a QIAquick Gel Extraction Kit (Qiagen). Selected products were amplified by 18 PCR cycles. For high-throughput sequencing, the library was prepared according to the manufacturer's instructions and submitted to five lanes of 75-nt paired-end sequencing using the Illumina GA system.

### Read alignment to the reference *A. thaliana* genome

A total of 115,883,414 paired-end reads were obtained from five lanes of Illumina sequencing. Read mapping to the *Arabidopsis* reference genome TAIR9 (The *Arabidopsis* Genome Initiative 2000), and splice junction detection was done using TopHat (Trapnell et al. 2009) with a maximum of two mismatches. Minimum and maximum intron lengths were fixed at 60 and 6000 nt, respectively. The rest of the parameters were left as default. A Pearson correlation coefficient between the sequencing lanes was calculated using the coverage values obtained by TopHat (Trapnell et al. 2009). As a good correlation between the replicates was obtained ($R^2 > 0.98$), the pool of sequences from the five lanes was used in further analyses.

To examine the influence of read-length and normalization on the read mapping, we used publicly available *Arabidopsis* Illumina data of 36-nt read length (Filichkin et al. 2010). Two libraries were analyzed: an oligo-dT cDNA library from 3-wk-old wild-type plants and one random primed (RP) cDNA library from 12-d-old seedlings (SRA files: SRX006704, SRX006692, SRX006192, SRX006191). To generate 36-nt reads for our data, we extracted the first 36 nt from the 75-nt original reads using homemade Perl scripts. For the single-end reads sets, we used only the left-side reads coming from the original pair-end set. All the sets were aligned with TopHat using the same intron size parameters as described above.

Read coverage along transcription units was obtained using annotated gene models in TAIR 9 (The *Arabidopsis* Genome Initiative 2000). If more than two gene models were annotated for a particular gene, then we chose the longest model as a representative of the gene. The start and the end of each transcription unit were normalized from 1% to 100%, respectively.

### Splice junction detection

We used TopHat software (Trapnell et al. 2009) to predict splice junctions. To diminish the number of potential false positives, we performed a prediction of splice junctions without mismatches in addition to the original alignment of two mismatches (Supplemental Fig. S1). The above approach allowed us to discard splice junctions that are predicted by erroneous alignment, especially in regions where several dinucleotides resembling splice sites are found in tandem. We have filtered out those splice junctions that were not predicted in the alignment of no mismatches and that are in proximity with another better supported junction by <10 nt. The remaining splice junctions were used for further analysis if they were supported by at least three alignments in the original list (alignment with two mismatches).

We have defined as genic splice junctions those junctions that are inside the coordinates of annotated genes in TAIR9 and have the same strand as the gene in question. Those splice junctions that are inside gene coordinates but are in the opposite strand were named antisense. Finally, those junctions that are not inside any gene coordinates were called intergenic.

For the genic splice junctions, we took those coming from protein-coding genes and have examined whether the predicted splice junctions are inside the coding sequence, 5′ UTR or 3′ UTR. If one extreme of the junction was in the coding sequence and the other in the UTR, they were classified as 5′UTR-CDS or CDS-3′UTR.

To determine the signatures of U2 or U12 introns, we used 10 intronic and three exonic bases for the 5′ splice site and 14 intronic bases and three exonic bases for the 3′ splice site. We have evaluated the splice sites using the procedure and PWMs described by Sheth et al. (2006). A strong signal of U2 or U12 intron was considered if the splice site in question (5′ or 3′) score was ≥65 in the respective PWM. For the U12 branch point, a PWM was built using the branch point sequences deposited in the U12 database (Alioto 2007). The procedure of building the PWM was the same as in Sheth et al. (2006). To identify the branch point, we scanned the first 30 bases upstream of the 3′ splice site and evaluated this stretch of sequence with the U12 PWMs. The higher score was retained, and, if this score was ≥65, then it was considered to be a strong signal for a U12 branch point. To define U2 branch points, we have looked for the motif YURAY in the stretch of 30 nt. As U12 introns have more highly conserved 5′ splice site and branch point sequences than U2 introns, they were classified as U12 when they possessed both 5′ splice site and branch point with scores ≥ 65 in the respective U12 PWMs. An intron was considered as U2 if it had a good U2 signal in the 5′ splice site (score ≥ 65 in the respective U2 PWMs) and a good signature in the 3′ splice site (score ≥ 65 in the respective U2 PWMs) or the YURAY U2 branch point.

Gene ontology classification of the genes containing U12 introns was done using Classification SuperViewer Tool (Provart and Zhu 2003) available at: http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi.

### Assembly of putative transcripts

To build the possible set of transcripts that can be explained by our reads, we first started by defining exons inside the longest transcription unit for each annotated gene in TAIR9. A putative exon was defined as a continuous stretch of at least four reads of coverage inside the transcription unit. If the gene in question had predicted junctions in the set of filtered junctions (score ≥ 3) (Table 1), then we used these junctions to connect the putative exons. All the possible combinations of exons that can be connected through the predicted junctions were built for each gene. The final set of transcripts was evaluated using Cufflinks v0.9.3. Those transcripts that passed the expression filter used by Cufflinks default settings (transcripts with <15% of expression level according to the highest expressed transcript in each gene are discarded) were used for the further analyses. To estimate how the above method is able to retrieve known transcripts, we have compared our set of transcripts with those annotated transcripts in TAIR9. If one putative transcript has the same combination of introns as in an annotated transcript and it covers at least 80% of its length contiguously, then it was considered that a known transcript was obtained for that particular gene.

To quantify the types of AS events in our final set of predicted transcripts, we have used ASTALAVISTA software (Foissac and Sammeth 2007).

### Validation of putative assembled transcripts using the HR RT-PCR panel

For corroboration of our putative assembled transcripts, we have used the HR RT-PCR Alternative Splicing Panel (Simpson et al. 2008a,b). The 273 primer pairs were designed to cover AS events previously reported in 256 genes. The primer sequences used in the analyses can be found in Supplemental Table S3. We extracted the sequences defined by each pair of primers using our PATs and transcripts annotated in TAIR9 database. For each sequence, we have calculated its size. Fragments between 100 and 650 nt were used for comparison due to the size marker used in HR-RT PCR panel. A PAT fragment was considered to match between the two resources if their fragment sizes are exactly the same or have a similar size with a deviation of no more than 4 nt. Additionally, Sanger sequences obtained from fragments of the HR RT-PCR panel have been used to corroborate some of the putative transcripts obtained in this study.

### Intron retention analysis

The intron retention events were obtained from our set of PATs. Thus, the retained introns in our analyses are completely covered by RNA-seq reads and have passed the expression filtering by Cufflinks. The intron retention ratio was calculated using the median of reads that aligned inside the intron divided by the number of reads supporting the splice junction. The median was used as the parameter of intron coverage because reads along the introns are not uniformly distributed and this measure is less sensitive to extreme values. Additionally, we have calculated the IRR by counting the reads containing the splice site and adjacent sequence of the retained intron divided by the reads supporting the splice junction. Both methods of IRR calculation gave a Pearson correlation of ~0.9 (data not shown). The further analyses were performed using the first method (median of reads along the intron). The differences between groups of intron features (intron size, GC content, splice site score, and number of stop codons) were evaluated using Mann-Whitney-Wilcoxon tests in groups of introns with the same sample size (Fig. 5).

We analyzed the coding potential of retained introns, which are located inside annotated coding sequences. Translation was performed using the frame of the 5′ neighboring exon annotated in TAIR. For those introns that are inside annotated exons, we have used the frame of the exon in which they are contained. As a negative control, we have used the same number of randomly selected introns inside coding sequences that have no evidence of retention in our sample, and the frame for their translation was obtained as for the retained introns. Furthermore, introns inside UTRs translated in the three frames were also used as a negative control. The resulting translated sequences (exon+intron+exon) were used to search Pfam protein domains employing hidden Markov models (HMMs) with HMMER (Pfam HMM database

release 25.0). Hits were only regarded as significant if they had an E-value ≤ 0.01. An intron was considered to be part of a protein domain if it contributes at least to 20% of the total domain.

The expected frequency of stop codons in introns was calculated by f(T) × f(A) × f(A) + f(T) × f(G) × f(A) + f(T) × f(A) × f(G), where f(A) represents the frequency of the nucleotide A. We obtained the codon usage of the exons and introns as described by Galante et al. (2004). We performed 100 pairwise comparisons between random distributions of 435,283 codons present in the retained introns along the 61 possible codons to obtain the average $\chi^2$ and the standard deviations (Supplemental Table S6).

For testing of the cryptic introns by RT-PCR, we designed primer pairs for 10 randomly selected introns with an IRR > 1 (Supplemental Table S10). The cDNA employed for the experiment was synthesized from a 1:1 mix of total RNA from wild-type flowers and 10-d-old seedlings using the Reverse Transcription System kit (Promega). PCR was performed under the following conditions: denaturation 98°C for 3 min, followed by 30 cycles of denaturation 98°C for 10 sec, annealing 58°C for 10 sec, and extension 72°C for 30 sec, and final extension at 72°C for 5 min. The amplified products were visualized in a 2% agarose gel stained with ethidium bromide (Supplemental Fig. S8).

## Data access

Illumina short read sequences generated in this study have been submitted to the NCBI Sequence Read Archive (SRA) (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession number SRA047499. Splice junctions, U12 introns, and putative assembled transcripts are available in the Supplemental Material.

## Acknowledgments

We thank Andreas Sommer from the Vienna Biocenter CSF-sequencing facility for sequencing, the Center of Integrative Bioinformatics Vienna (CIBIV) for use of cluster facilities, and Gautier Koscielny from the EBI Hinxton for advice to Y.M. in high-throughput sequencing data management. This work was supported by the Austrian Science Fund (FWF) (DK W1207; ERA-NET Plant Genomics [PASAS] I254; SFB RNAreg F43-P10) and the Austria Genomic Program (GENAU III) (ncRNAs), the EU FP6 Programme Network of Excellence on Alternative Splicing (EURASNET) (LSHG-CT-2005-518238), the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/G024979/1, ERA-NET Plant Genomics [PASAS]), and the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD) (WP114).

## References

Ali GS, Reddy AS. 2008. Regulation of alternative splicing of pre-mRNAs by stresses. *Curr Top Microbiol Immunol* **326:** 257–275.

Alioto TS. 2007. U12DB: A database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res* **35:** D110–D115.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Barbazuk WB, Fu Y, McGinnis KM. 2008. Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. *Genome Res* **18:** 1381–1392.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci* **74:** 3171–3175.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72:** 291–336.

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. *Nat Genet* **30:** 29–30.

Brown JW, Simpson CG. 1998. Splice site selection in plant pre-mRNA splicing. *Annu Rev Plant Physiol Plant Mol Biol* **49:** 77–95.

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. 2006. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* **7:** 327. doi: 10.1186/1471-2164-7-327.

Cao W, Jamison SF, Garcia-Blanco MA. 1997. Both phosphorylation and dephosphorylation of ASF/SF2 are required for pre-mRNA splicing in vitro. *RNA* **3:** 1456–1467.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* **12:** 1–8.

Colwill K, Feng LL, Yeakley JM, Gish GD, Caceres JF, Pawson T, Fu XD. 1996a. SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors. *J Biol Chem* **271:** 24569–24575.

Colwill K, Pawson T, Andrews B, Prasad J, Manley JL, Bell JC, Duncan PI. 1996b. The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intracellular distribution. *EMBO J* **15:** 265–275.

Cusack BP, Wolfe KH. 2005. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. *Mol Biol Evol* **22:** 2198–2208.

Duncan PI, Howell BW, Marius RM, Drmanic S, Douville EM, Bell JC. 1995. Alternative splicing of STY, a nuclear dual specificity kinase. *J Biol Chem* **270:** 21524–21531.

Duncan PI, Stojdl DF, Marius RM, Bell JC. 1997. In vivo regulation of alternative pre-mRNA splicing by the Clk1 protein kinase. *Mol Cell Biol* **17:** 5996–6001.

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. 2010. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res* **20:** 45–58.

Foissac S, Sammeth M. 2007. ASTALAVISTA: Dynamic and flexible analysis of alternative splicing events in custom datasets. *Nucleic Acids Res* **35:** W297–W299.

Galante P, Sakabe N, Kirschbaum-Slager N, de Souza S. 2004. Detection and evaluation of intron retention events in the human transcriptome. *RNA* **10:** 757–765.

Golovkin M, Reddy AS. 1999. An SC35-like protein and a novel serine/arginine-rich protein interact with *Arabidopsis* U1-70K protein. *J Biol Chem* **274:** 36428–36438.

Goodall GJ, Filipowicz W. 1989. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* **58:** 473–483.

Goodall GJ, Filipowicz W. 1990. The minimum functional length of pre-mRNA introns in monocots and dicots. *Plant Mol Biol* **14:** 727–733.

Hillman RT, Green RE, Brenner SE. 2004. An unappreciated role for RNA surveillance. *Genome Biol* **5:** R8. doi: 10.1186/gb-2004-5-2-r8.

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K. 2004. Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res* **32:** 5096–5103.

Jones-Rhoades MW, Borevitz JO, Preuss D. 2007. Genome-wide expression profiling of the *Arabidopsis* female gametophyte identifies families of small, secreted proteins. *PLoS Genet* **3:** 1848–1861.

Kalyna M, Simpson CG, Syed NH, Lewandowska D, Marquez Y, Kusenda B, Marshall J, Fuller J, Cardle L, McNicol J, et al. 2011. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res*. doi: 10.1093/nar/gkr932.

Kpebe A, Rabinow L. 2008. Alternative promoter usage generates multiple evolutionarily conserved isoforms of *Drosophila* DOA kinase. *Genesis* **46:** 132–143.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40:** 225–231.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Lorkovic ZJ, Wieczorek Kirk DA, Lambermon MH, Filipowicz W. 2000. Pre-mRNA splicing in higher plants. *Trends Plant Sci* **5:** 160–167.

Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Huang X, et al. 2010. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res* **20:** 1238–1249.

Mironov AA, Fickett JW, Gelfand MS. 1999. Frequent alternative splicing of human genes. *Genome Res* **9:** 1288–1293.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.

Ner-Gaon H, Fluhr R. 2006. Whole-genome microarray in *Arabidopsis* facilitates global analysis of retained introns. *DNA Res* **13:** 111–121.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40:** 1413–1415.

Petrillo E, Sanchez SE, Kornblihtt AR, Yanovsky MJ. 2011. Alternative splicing adds a new loop to the circadian clock. *Commun Integr Biol* **4:** 284–286.
Provart N, Zhu T. 2003. A browser-based Functional Classification SuperViewer for *Arabidopsis* genomics. *Currents in Computational Molecular Biology* **2003:** 271–272.
Rosenfeld MG, Amara SG, Roos BA, Ong ES, Evans RM. 1981. Altered expression of the calcitonin gene associated with RNA polymorphism. *Nature* **290:** 63–65.
Sakabe N, de Souza S. 2007. Sequence features responsible for intron retention in human. *BMC Genomics* **8:** 59. doi: 10.1186/1471-2164-8-59.
Savaldi-Goldstein S, Sessa G, Fluhr R. 2000. The ethylene-inducible PK12 kinase mediates the phosphorylation of SR splicing factors. *Plant J* **21:** 91–96.
Savaldi-Goldstein S, Aviv D, Davydov O, Fluhr R. 2003. Alternative splicing modulation by a LAMMER kinase impinges on developmental and transcriptome expression. *Plant Cell* **15:** 926–938.
Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34:** 3955–3967.
Simpson CG, Fuller J, Maronova M, Kalyna M, Davidson D, McNicol J, Barta A, Brown JW. 2008a. Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant J* **53:** 1035–1048.
Simpson CG, Lewandowska D, Fuller J, Maronova M, Kalyna M, Davidson D, McNicol J, Raczynska D, Jarmolowski A, Barta A, et al. 2008b. Alternative splicing in plants. *Biochem Soc Trans* **36:** 508–510.
Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang M. 2000. An alternative-exon database and its statistical analysis. *DNA Cell Biol* **19:** 739–756.
Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* **344:** 1–20.
Sugnet CW, Kent WJ, Ares M Jr, Haussler D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput* **9:** 66–77.
Terzi LC, Simpson GG. 2008. Regulation of flowering time by RNA processing. *Curr Top Microbiol Immunol* **326:** 201–218.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.
Wahl MC, Will CL, Luhrmann R. 2009. The spliceosome: Design principles of a dynamic RNP machine. *Cell* **136:** 701–718.
Wakamatsu A, Kimura K, Yamamoto J, Nishikawa T, Nomura N, Sugano S, Isogai T. 2009. Identification and functional analyses of 11,769 full-length human cDNAs focused on alternative splicing. *DNA Res* **16:** 371–383.
Wang BB, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci* **103:** 7175–7180.
Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol* **144:** 32–42.
Xiao YL, Smith SR, Ishmael N, Redman JC, Kumar N, Monaghan EL, Ayele M, Haas BJ, Wu HC, Town CD. 2005. Analysis of the cDNAs of hypothetical genes on *Arabidopsis* chromosome 2 reveals numerous transcript variants. *Plant Physiol* **139:** 1323–1337.
Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA, Hayashizaki Y, Gaasterland T. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res* **13:** 1290–1300.
Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, et al. 2010. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res* **20:** 646–654.
Zheng CL, Fu XD, Gribskov M. 2005. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. *RNA* **11:** 1777–1787.
Zhu W, Brendel V. 2003. Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res* **31:** 4561–4572.
Zhu W, Schlueter SD, Brendel V. 2003. Refined annotation of the *Arabidopsis* genome by complete expressed sequence tag mapping. *Plant Physiol* **132:** 469–484.

*Received October 28, 2011; accepted in revised form February 23, 2012.*