

Transcriptome Survey of the Contribution of Alternative Splicing to Proteome Diversity in *Arabidopsis thaliana*

Dear Editor,

Alternative splicing (AS) functions as a key regulatory mechanism and increases transcriptome and proteome diversity. Recent genome-wide studies have substantially expanded our estimation of the frequency of AS in plants (Reddy et al., 2013; Staiger and Brown, 2013). However, the proportion of AS events that lead to increased proteome diversity in plants, rather than imperfect pre-mRNA processing, remains unsolved. Here, we report an analysis of AS events in presumably translated mRNAs, i.e., those associated with polysomes, in *Arabidopsis thaliana*. We found that 35% of AS events identified in total mRNAs occur in polysome-associated mRNAs, suggesting translation into proteins. Furthermore, most (81%) of these presumably translated alternative isoforms lead to diversified protein-coding capacity. By contrast, weakly translated, but not non-translated, alternative isoforms likely undergo nonsense-mediated mRNA decay (NMD). Sequence analyses identified structural features of transcripts and *cis*-elements that were associated with AS. The results suggest that AS in plants increases proteome complexity but shows clear differences from AS in animals.

Genome-wide studies in humans indicate that AS plays important roles and affects nearly 95% of intron-containing genes (Wang et al., 2008). In plants, genome-wide studies have substantially expanded our estimation of the frequency of AS events to about 60% in all intron-containing transcripts in *Arabidopsis* (Filichkin et al., 2010; Marquez et al., 2012). However, the proportion of AS events that lead to increased proteome diversity in plants, rather than imperfect pre-mRNA processing, remains a subject of debate. A recent analysis identified different characteristics of polysome-associated mRNAs than those of total mRNAs (Zhang et al., 2015). However, a comprehensive estimation of the contribution of AS to proteome diversity is still lacking. It's also important to know what determines whether a splicing isoform translates or not. In addition, what kinds of biological processes AS-derived translated mRNAs are involved remain unknown.

To achieve comprehensive and comparable profiles of AS events in *Arabidopsis*, we isolated and deeply sequenced total poly(A)⁺ RNA (transcriptome) and polysome-associated poly(A)⁺ RNA (translatome) in parallel from 10-day-old seedlings and inflorescences (Figure 1A and Supplemental Figure 1). After filtering low-quality reads, about 93 million uniquely aligned read pairs were evenly mapped to the *Arabidopsis* TAIR10 reference genome (Supplemental Figure 2A–2C). Bioinformatic analysis indicates that our sequence depth was sufficient to comprehensively estimate AS events (Supplemental Figure 2D–2F).

We combined the uniquely aligned reads from transcriptome and translatome to ensure comprehensive and reliable detection of AS events. Using a relatively stringent standard for intron identification and transcript assembly (Supplemental Methods), we successfully assembled a total of 31060 transcripts, corresponding to 22896 genes, including 17670 intron-containing genes. Among the genes with introns, we identified 5824 (33.0%) intron-containing genes that have at least two overlapping transcript isoforms (Supplemental Table 1). Of these protein-coding genes, we found 6292 AS events (Figure 1B), including 4058 transcriptome-specific events that were considered to be non-translated (see Supplemental Methods for more details). The remaining 2234 AS events were considered translated (Figure 1C), indicating that about 34.9% of AS events may contribute to proteome diversity. We also applied our analysis pipeline to a recent similar dataset in humans (Sterne-Weiler et al., 2013), and observed that a higher proportion (68.4%) of AS events may contribute to proteome diversity (Supplemental Figure 3).

To further reveal potential protein sequence alterations caused by AS events, we first analyzed the distribution of AS events along transcripts. According to gene structure annotation of the main isoforms, we found that most of them were located in the coding DNA sequence (CDS) (Figure 1D). After normalizing by length, the frequency of AS in the 5' UTR was about three-fold higher (Figure 1D) than in the CDS or the 3' UTR. Among major types of AS events, a lower portion of intron retention (IR) was found in translated transcripts than in the non-translated transcripts in both the CDS and UTR, and this decrease is much more dramatic in the 5' UTR and the CDS than in the 3' UTR (Pearson's chi-squared test, $P < 1.0E-5$). By contrast, all other major types of AS events, including exon skipping (ES), alternative 5' donor (AD), and alternative 3' acceptor (AA), have higher proportions in translated isoforms than in non-translated isoforms (Supplemental Figure 4A).

We then analyzed potential CDS alterations caused by AS events and identified a total of 1751 AS events that could lead to protein sequence alterations, accounting for 80.7% of all AS events detected in the translatome (Figure 1E). On the other hand, we detected 3775 (93.0%) AS events that could lead to protein sequence alterations in non-translated transcripts, which is significantly higher than translated transcripts (Pearson's chi-squared test, $P = 6.0E-49$). The IR type was the major type in transcriptome. However, there was only a small fraction (23.5%) of transcriptome-detected IR contributing to the

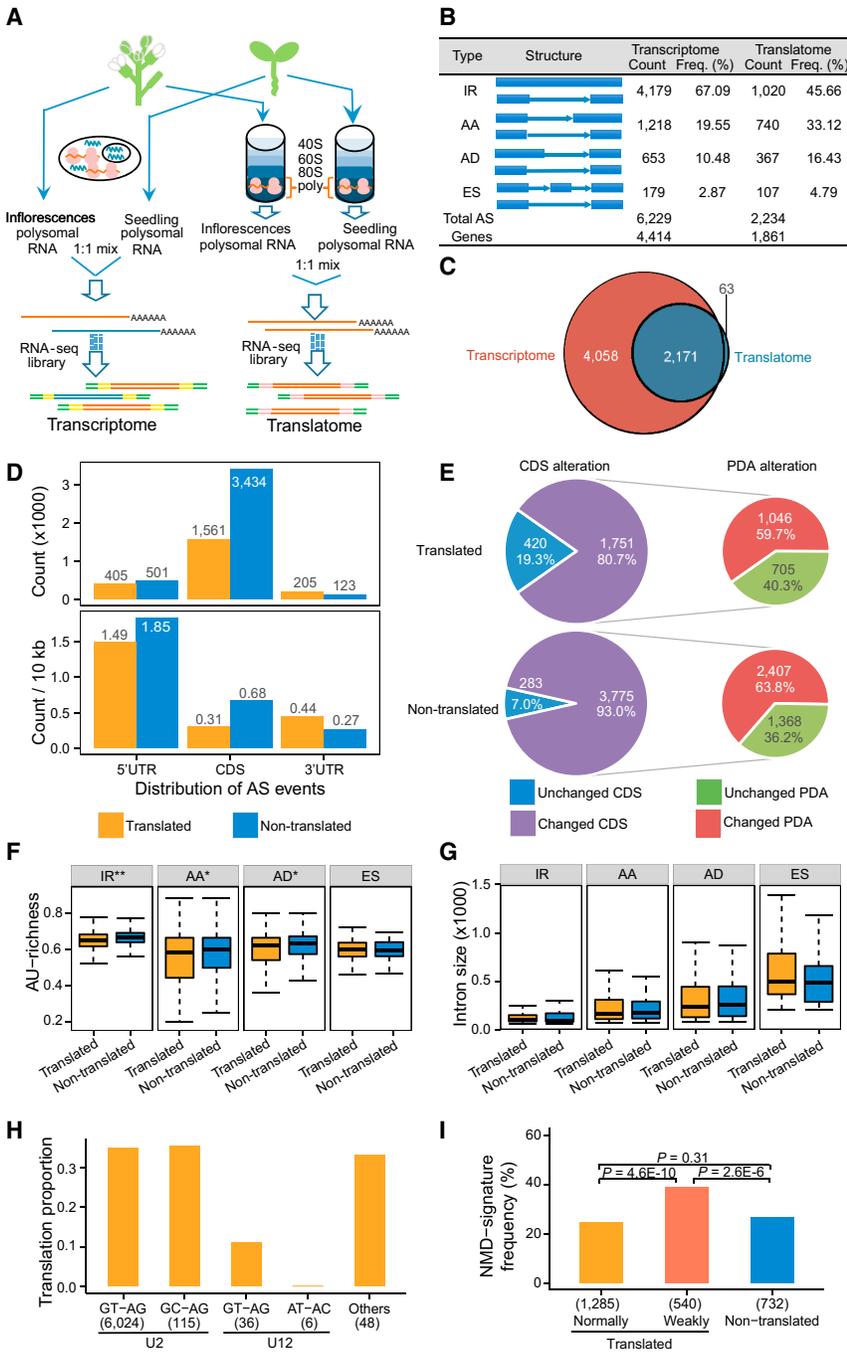


Figure 1. AS Contribution to Proteome Diversity.

(A) Pipelines for construction of RNA-seq libraries for the transcriptome and translatoome. Inflorescences and seedlings were ground to extract total RNA or polysomal RNA separately, and mixed equally before mRNA purification. Then RNA-seq libraries were constructed as described.

(B) Statistics of the four major types of AS events for protein-coding genes.

(C) Location of AS events in the transcriptome and translatoome for protein-coding transcripts.

(D) Distribution of AS events along gene features without (upper) and with (lower) normalization to sequence length.

(E) Effects of AS events on protein sequence (left panel) and protein domain architecture (right panel). AS events in translated transcripts (upper) and in non-translated transcripts (lower) were calculated separately.

(F) AU richness of alternative intron regions shows differences within IR, AA, and AD types between translated and non-translated AS events. Stars indicate significance by Mann-Whitney-Wilcoxon test; * $P < 0.05$, ** $P < 0.01$.

(G) No statistically significant difference in intron size between translated and non-translated AS events. For each AS event, the longest alternative intron was used for calculation.

(H) Distinct translation proportions of associated AS isoforms among intron subtypes. The number of AS events for each subtype is included below the subtype name.

(I) Enrichment of NMD signatures in weakly translated AS isoforms. The Y axis represents the proportion of the NMD signature containing transcripts. The statistical significance (P value) was assessed by Pearson's chi-squared test.

translatome (Supplemental Figure 4A). By contrast, the other three major types of AS have more than half AS events identified in the translatoome data (Supplemental Figure 4A). This observation is consistent with the recent report that although IR is the most abundant AS type, it is enriched in weakly expressed transcripts (Marquez et al., 2012), which are more likely excluded from the translatoome (Supplemental Figure 4B). Exitrons are a class of newly identified exon-like introns, which can be spliced and found on polysomes (Marquez et al., 2015). We could detect a total of 539 exitrons in the transcriptome, among which 305 (56.6%) exitrons could also be detected in the translatoome, suggesting that exitrons resulting from unique AS contribute substantially to proteome diversity.

diversification in *Arabidopsis*. On the other hand, 63.8% of AS events in non-translated isoforms resulted in alteration of PDA.

We also analyzed the relative expression levels of AS isoforms and found that the minor isoforms are expressed at lower levels in non-translated AS isoform pairs than in translated pairs (Mann-Whitney-Wilcoxon test, $P = 4.3E-111$), suggesting that the low-abundance AS isoforms had a lower chance of being translated (Supplemental Figure 4B). This difference is most dramatic in the IR type ($P = 3.4E-60$) but is not significant in the ES type (Supplemental Figure 4B). To test if the observed low rate of minor isoforms in the translatoome could be a detection artifact, we compared low abundant transcripts with

and without AS. For the 2000 transcripts right above our detection cutoff, those with AS were less likely identified in the translome than those without (80.4% versus 93.3%, Pearson's chi-squared test, $P = 4.9E-18$), suggesting that low-abundance AS isoforms had a lower chance of being translated. We also observed that 65.8% of AS-containing genes have minor isoform expression levels above 15% (Supplemental Figure 4B), which is lower than the proportion of 85% in humans (Wang et al., 2008).

To characterize the biological functions of genes that undergo regulation by AS, we deployed a gene ontology (GO) analysis. We found that genes with translated AS isoforms participated in many different processes (Supplemental Figure 5A), but only a few terms related to mRNA processing, including splicing (Supplemental Figure 5B), were enriched in translated isoforms (FDR-adjusted $P < 0.05$). On the other hand, more diverse GO terms are enriched in genes containing non-translated AS isoforms (Supplemental Figure 5C).

To test whether the sequence features related to splicing efficiency also had some influence on translation, we examined the AU richness, intron size, and the type of spliceosomes utilized based on sequence. At the whole-genome scale, we found that the AU richness was slightly, but significantly (Mann-Whitney-Wilcoxon test, $P = 6.5E-101$), lower in introns in translated AS isoforms (Figure 1F). By contrast, no significant difference ($P > 0.05$) in intron size was identified (Figure 1G). Based on the terminal dinucleotide recognized by the spliceosome or the type of spliceosome, the GT-AG group and the U2 group have higher proportions translated isoform than the GC-AG and AT-AC groups, and the U12 group, respectively (Figure 1H).

Because the majority of AS-containing transcripts are non-translated or low in translation (Figure 1B and Supplemental Figure 4B), we speculated that many of these transcripts represent imperfect pre-mRNA processing and would be subject to NMD. We found that known NMD signatures (see Supplemental Discussion for details) were significantly enriched in weakly translated ($\text{RPKM}_{\text{translatome}}/\text{RPKM}_{\text{transcriptome}} < 0.5$) isoforms compared with normally translated ($\text{RPKM}_{\text{translatome}}/\text{RPKM}_{\text{transcriptome}} \geq 0.5$) and non-translated (not detected in the translome) isoforms (Supplemental Methods, Figure 1I, and Supplemental Figure 6A and 6B). Furthermore, we found an enrichment of NMD signatures in the IR and AA types, but not in the AD and ES types (Supplemental Figure 6C). In addition, all known types of NMD signatures are depleted from normally translated isoforms (Supplemental Figure 6E) and show enrichment in either the weakly translated isoforms or the non-translated isoforms (Supplemental Figure 6F and 6G). Because triggering NMD requires the assembly of polysome complexes (Maquat, 2004), non-translated transcripts, which are not associated with ribosomes, may not trigger NMD. These observations are based on genome-scale statistics and may not apply to each individual gene, as suggested recently (Kalyna et al., 2012; Marquez et al., 2012; Gohring et al., 2014). Some transcripts containing retained introns are not NMD sensitive (Kalyna et al., 2012; Marquez et al., 2012) and are probably not exported from the nucleus (Gohring et al., 2014). Many of the non-translated transcripts may be subject to other mRNA surveillance pathways.

ACCESSION NUMBERS

The raw reads data of this study have been submitted to the NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) under accession number PRJNA266359. The aligned reads, assembled transcripts, identified AS events were implemented in a web-based genome browser available at <http://jiaolab.genetics.ac.cn/as>.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by a National Basic Research Program of China (973 Program) grant 2012CB910902, National Natural Science Foundation of China grants 31222033, 31300298, and 31171159, National Program for Support of Top-Notch Young Professionals, and by the State Key Laboratory of Plant Genomics (through grant SKLPG2011A0103).

AUTHOR CONTRIBUTIONS

Y.J. designed the research. C.T. and Y.Y. performed the experiments. H.Y., C.T., and Y.J. analyzed the data. Y.J., C.T., and H.Y. wrote the article.

ACKNOWLEDGMENTS

We thank Wenfeng Qian for critical reading of the manuscript, Taolan Zhao and Xiaofeng Cao for advice on polysome isolation, and Ligeng Ma for his support. No conflict of interest declared.

Received: July 29, 2015

Revised: December 3, 2015

Accepted: December 22, 2015

Published: December 29, 2015

Haopeng Yu^{1,2,4}, Caihuan Tian^{1,4}, Yang Yu^{1,3}
and Yuling Jiao^{1,*}

¹State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, National Center for Plant Gene Research, Beijing 100101, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³The Middle School Attached to Beijing College of Petroleum, Beijing 100083, China

⁴These authors contributed equally to this article.

*Correspondence: Yuling Jiao (ylijiao@genetics.ac.cn)
<http://dx.doi.org/10.1016/j.molp.2015.12.018>

REFERENCES

- Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K., and Mockler, T.C. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* **20**:45–58.
- Gohring, J., Jacak, J., and Barta, A. (2014). Imaging of endogenous messenger RNA splice variants in living cells reveals nuclear retention of transcripts inaccessible to nonsense-mediated decay in *Arabidopsis*. *Plant Cell* **26**:754–764.
- Kalyna, M., Simpson, C.G., Syed, N.H., Lewandowska, D., Marquez, Y., Kusenda, B., Marshall, J., Fuller, J., Cardle, L., McNicol, J., et al. (2012). Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res.* **40**:2454–2469.
- Maquat, L.E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* **5**:89–99.
- Marquez, Y., Brown, J.W., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* **22**:1184–1195.
- Marquez, Y., Hopfler, M., Ayatollahi, Z., Barta, A., and Kalyna, M. (2015). Unmasking alternative splicing inside protein-coding exons

defines exons and their role in proteome plasticity. *Genome Res.* **25**:995–1007.

Reddy, A.S., Marquez, Y., Kalyna, M., and Barta, A. (2013). Complexity of the alternative splicing landscape in plants. *Plant Cell* **25**:3657–3683.

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

Sterne-Weiler, T., Martinez-Nunez, R.T., Howard, J.M., Cvitovik, I., Katzman, S., Tariq, M.A., Pourmand, N., and Sanford, J.R. (2013).

Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.* **23**:1615–1623.

Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**:470–476.

Zhang, X., Rosen, B.D., Tang, H., Krishnakumar, V., and Town, C.D. (2015). Polyribosomal RNA-Seq reveals the decreased complexity and diversity of the *Arabidopsis* translome. *PLoS One* **10**: e0117699.