Start right to end right: authentic open reading frame selection matters

Mojtaba Bagherian[1], Georgina Harris[1], Pratosh Sathishkumar[1], and James P B Lloyd[1,2,*]

[1]**ARC Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Perth, Australia**

[2]**ARC Centre of Excellence in Plants for Space, School of Molecular Sciences, The University of Western Australia, Perth, Australia**

***e-mail:** james.lloyd@uwa.edu.au

Abstract

Proper identification of start codons is critical for accurate annotation of open reading frames (ORFs) and prediction of nonsense-mediated mRNA decay (NMD) targets. We demonstrate that reference transcriptomes annotations frequently select the longest ORF rather than the authentic start codon, leading to incorrect assessment of NMD features. Our revised ORF annotations significantly improve NMD target identification, and prevents computational hallucinations in protein databases.

Main text

Genome sequencing is becoming cheaper and easier, with telomere to telomere assemblies becoming routine with third generation sequencing technologies[1,2]. However, annotating the transcript models of a genome is still a challenge. With RNA-seq, accurate exon-intron boundaries can be determined and whole transcript models can be generated. Multiple transcript models at a genetic locus can be generated via alternative splicing (AS) to increase the transcriptome diversity. But not all alternative transcript isoforms go on to produce protein isoforms. In the case of AS that introduces a premature termination codon (PTC), then the ribosome will terminate early and this will lead to RNA degradation by nonsense-mediated mRNA decay (NMD)[3], an evolutionarily conserved eukaryotic quality control pathway[4,5]. The commonly reported signals that differentiate a PTC from a normal stop codon are the presence of an exon junction >=50-55 nucleotides (nt) downstream of the stop codon (50-55 nt rule, or EJC model of NMD), or the presence of a long 3' UTR (faux 3' UTR model of NMD)[3,6,7]. However, to accurately predict whether a transcript might be an NMD target, not only do we need accurate structural information for the transcript model (transcriptional start/end sites and exon-intron boundaries), but we also need to know the correct translational start codon and translational stop codon. An accurate 3' UTR annotation allows for accurate prediction of NMD-inducing features[8].

However, in some genome annotation projects, the algorithms used to find the open reading frame (ORF) within the transcript models simply identifies the longest ORF present. For years, researchers have pointed out that simply annotating the longest ORF in a new transcriptome is unwise and leads to biologically incorrect annotations[8–10]. For example, the plant model *Arabidopsis thaliana* transcriptome annotation Araport11 has numerous examples of the longest ORF being selected for alternatively spliced transcripts (Fig. 1)[11], leading to the appearance of a normal stop codon with an unmodified 3' UTR (but a large untranslated 5' UTR). Identifying the longest ORF is computationally simple and logical, but ignores the biological reality: The ribosome cannot predict which ORF is the longest but instead will select the same start codon as if the downstream AS event had not occurred (Fig. 1). By annotating the longest ORF through using an internal methionine codon as the new start codon, the 3' UTR of the AS-coupled NMD target will look normal and the protein will appear as an N-terminal truncation. In reality, the protein would be a C-terminal truncation and the 3' UTR would be much longer and likely contain at least one downstream exon junction (dEJ). Detection of the altered 3' UTR can be used to computationally predict a transcript's NMD-sensitivity.

We took the software TranSuite[10], developed to annotate the authentic start codon and stop codon, and compared the Araport11 transcriptome's base annotations to the TranSuite's improved annotations (Fig. 1). The known NMD targets RS2Z33 and PTB1 in Araport11 - Reference ORF annotations have normal 3' UTRs (Fig. 1a and b), but once annotated with TranSuite, the presence of PTCs as defined by the presence of a dEJ at least 50 nt after the stop codon (PTC_{dEJ}) is clear (Fig. 1c and d).

In order to determine the improvement of the TranSuite ORF annotations of Araport11 at the transcriptome-wide scale, we compared the Revised ORF with the Reference ORF annotations and the frequency of identifying NMD targets in an NMD-deficient mutant[12]. By re-analysis of the data from Drechsel *et al*. (2013)[12] with the Araport11[11] transcriptome and quantification of transcript increase and decrease in the NMD-deficient mutant relative to control, we identified putative NMD targets. However, many transcripts change in NMD mutants not only due to an increase in NMD target expression, but from increased and decreased indirect changes in the transcriptomes. In *A. thaliana*, these include many indirect targets of NMD involved in pathogen response[13–15].

We then compared the fraction of putative NMD targeted transcripts (increased steady state expression) with NMD features with the Araport11 - Reference ORF annotations or the Araport11 - Revised ORF annotations (as reported by us here using a modified TranSuite). We found that only 203 transcripts with increased steady state expression as annotated by Araport11 - Reference ORF annotations had a PTC_{dEJ} (Fig. 2). In contrast, we found that 426 transcripts with increased steady state expression as annotated by Araport11 - Revised ORF annotations had a PTC_{dEJ} (Fig. 2).

Not all transcripts with PTC_{dEJ} are expected to be NMD targets, for example, many intron retained transcripts with PTC_{dEJ} are detained in the nucleus[16,17]. However, correct 3' UTR annotation combined with increased steady state expression in an NMD-deficient mutant is highly indicative of an NMD target.

The ratio of up/down transcripts with PTC_{dEJ} increases when we move from Araport11 Reference to Revised ORF annotations, while the ratio of up/down transcripts with normal termination codons decreases (Fig. 2), suggesting that the TranSuite annotation greatly improves our ability to identify NMD targets. As long-read aided re-annotations of transcripts, such as the *Arabidopsis thaliana* Reference Transcript Dataset 3 (AtRTD3)[18] are produced, improved ORF annotations are vital to correct interpretation of novel datasets.

Taken together, we have shown that simply annotating the longest ORF does not reflect the biological reality of transcript anatomy, and that by annotating the authentic start codon, we can see an improvement in fraction of putative NMD identification (Fig. 2). Not only is this important for studying NMD, but it reflects our understanding of potential proteins in the proteome. If these transcripts are not labelled as putative NMD targets due to the presence of a PTC_{dEJ}, then it might be assumed that they are alternative splicing events leading to a different protein isoform. Computational prediction of protein structure is becoming routine now[19], and thus automated protein sequence and structure databases would be contaminated with incorrect proteins that are likely to never be translated and have no biological impact on the organism. Instead, these represent computational hallucinations and should be purged from such databases.

Methods

Genomes: *Arabidopsis thaliana* TAIR10 genome (fasta) with Araport11 transcriptome (GFF3) downloaded from Phytozome. The GFF3 was converted to GTF by the GFFread utility in Cufflinks[20]. This was also used to create a transcriptome fasta file from the GTF and genome fasta file.

```
gffread -w Athaliana_transcripts.fa -g Athaliana_447_TAIR10.fa Athaliana_447_Araport11.gtf
```

The transcriptome fasta file was converted to a Salmon index (kmer setting at 31), Salmon quant was used to to quant transcript abundances [21]. The following settings were used:

```
Salmon quant -i A_thaliana_Araport11 -l A -r file.fastq.gz -p 16 --validateMappings --fldMean 150 --numBootstraps 100 --seqBias --gcBias -o output
```

Previously published RNA-seq data collected from wild-type and *upf1-1 upf3-1* double mutant plants, which have reduced NMD activity[12], were quantified by Salmon, as above. Differential transcript expression was assessed through the R package Sleuth[22].

To accurately annotate the authentic start and stop codons of ORFs, rather than just the longest ORF, we used a modified version of TranSuite[10]. The original version of TranSuite is available here: https://github.com/anonconda/TranSuite. Our modified version of TranSuite is available here: https://github.com/mojtabagherian/TranSuite. Our modified TransFeat module creates a table (CSV) that lists common features related to NMD per transcript, such as distances to the last exon junction downstream of the stop codon and 3' UTR length. If there was a downstream exon junction and it was at least 50 nucleotides after the stop codon, the transcript isoform is labelled as a Premature Termination Codon downstream Exon-Junction (PTC_{dEJ}).

To get correct annotation of Araport11 ORFs as PTC_{dEJ} or not, we filtered out transcript IDs from the GTF that were missing from the TransFix version to allow for a like-for-like comparison and then used our modified TransFeat to identify whether a transcript was PTC_{dEJ} or not. Because many transcripts in Araport11 are annotated with the longest ORF, and internal methionine codons would generate a longer ORF than using the authentic start codon, the stop codon and 3' UTR does not change in Araport11 - Reference ORF annotations.

A Zenodo archive with genomic data is available here: https://doi.org/10.5281/zenodo.15628154. It contains files used to run TranSuite (Athaliana_447_Araport11.gtf, Athaliana_transcripts.fa), intermediary output from TransFind/TransFix (Transfix.gtf), the output from our custom TransFeat (Athaliana_transfeat_splice_junctions.csv). It also contains the modified Araport11 transcriptome (Athaliana_447_Araport11_Reference_ORF.gtf) used as direct input into TransFeat that was used to generate the Reference ORF annotations used in this study (Athaliana_transfeat_splice_junctions_Reference_ORF.csv).

Acknowledgment

Clifford Bradley Robertson and Gwendoline Florence Robertson fund at The University of Western Australia to JPBL. For supporting the lab in general, JPBL would like to thank the Australian Research Council (ARC) Discovery Project (DP240103385), the UK's Advance Research and Innovation Agency (ARIA) Synthetic Plants programme, and the School of Molecular Sciences at The University of Western Australia. The computational work of the project were made possible by the High Performance Computing platform *Kaya* at The University of Western Australia maintained by Emily Barker, Chris Bonding and David Gray.

References

1. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
2. Koren, S. *et al.* Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res.* **34**, 1919–1930 (2024).
3. Lloyd, J. P. B. The evolution and diversity of the nonsense-mediated mRNA decay pathway. *F1000Res.* **7**, 1299 (2018).
4. Causier, B. *et al.* Conservation of nonsense-mediated mRNA decay complex components throughout eukaryotic evolution. *Sci. Rep.* **7**, 16692 (2017).
5. Lloyd, J. P. B. & Davies, B. SMG1 is an ancient nonsense-mediated mRNA decay effector. *Plant J.* **76**, 800–810 (2013).

6. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
7. Lloyd, J. P. B., French, C. E. & Brenner, S. E. Polysome fractionation analysis reveals features important for human nonsense-mediated mRNA decay. *bioRxiv* 2020.03.08.981811 (2020) doi:10.1101/2020.03.08.981811.
8. Lloyd, J. P. B. *et al.* The loss of SMG1 causes defects in quality control pathways in Physcomitrella patens. *Nucleic Acids Res.* **46**, 5822–5836 (2018).
9. Brown, J. W. S. *et al.* Lost in translation: Pitfalls in deciphering plant alternative splicing transcripts. *Plant Cell* **27**, 2083–2087 (2015).
10. Entizne, J. C. *et al.* TranSuite: a software suite for accurate translation and characterization of transcripts. *bioRxiv* 2020.12.15.422989 (2020) doi:10.1101/2020.12.15.422989.
11. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* **89**, 789–804 (2017).
12. Drechsel, G. *et al.* Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell* **25**, 3726–3742 (2013).
13. Rayson, S. *et al.* A role for nonsense-mediated mRNA decay in plants: pathogen responses are induced in Arabidopsis thaliana NMD mutants. *PLoS One* **7**, e31917 (2012).
14. Riehs-Kearnan, N., Gloggnitzer, J., Dekrout, B., Jonak, C. & Riha, K. Aberrant growth and lethality of Arabidopsis deficient in nonsense-mediated RNA decay factors is caused by autoimmune-like response. *Nucleic Acids Res.* **40**, 5615–5624 (2012).
15. Raxwal, V. K. *et al.* Nonsense-mediated RNA decay factor UPF1 is critical for posttranscriptional and translational gene regulation in Arabidopsis. *Plant Cell* **32**, 2725–2741 (2020).
16. Boutz, P. L., Bhutkar, A. & Sharp, P. A. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev.* **29**, 63–80 (2015).
17. Göhring, J., Jacak, J. & Barta, A. 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 (2014).
18. Zhang, R. *et al.* A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. *Genome Biol.* **23**, 149 (2022).
19. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
20. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
21. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
22. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).

Figures

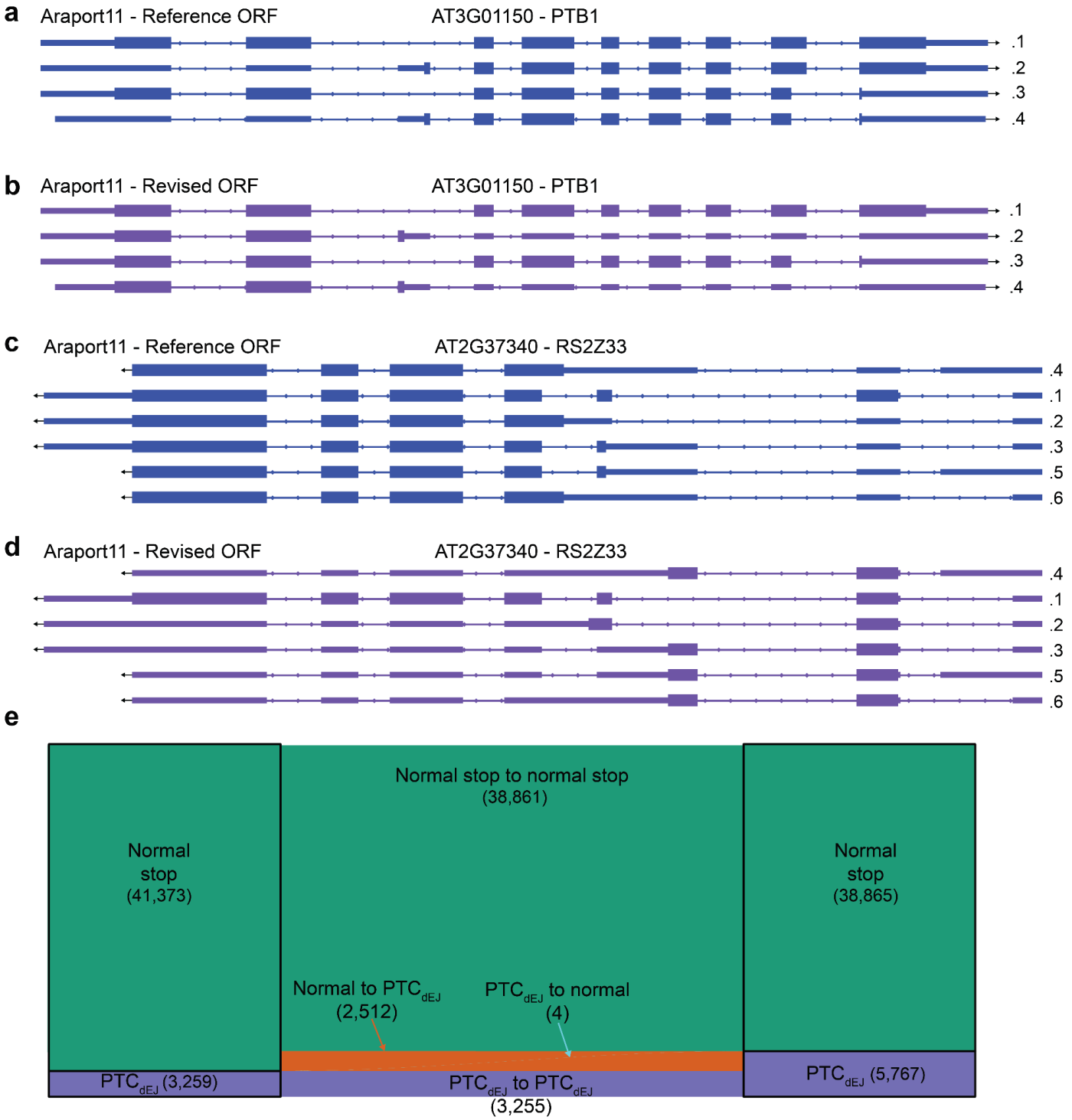


Fig. 1: Araport11 transcripts with ORF annotations

a, Araport11 - Reference ORF annotations of AT3G01150 (PTB1) gene.

b, Araport11 - Revised ORF annotations of AT3G01150 (PTB1) gene.

c, Araport11 - Reference ORF annotations of AT2G37340 (RS2Z33) gene.

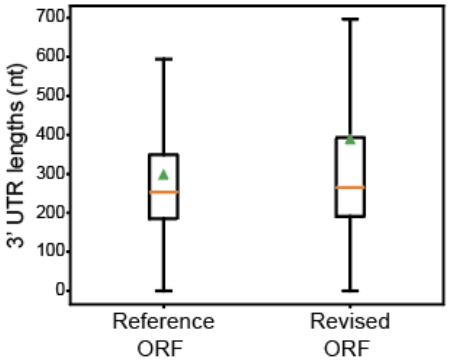
d, Araport11 - Revised ORF annotations of AT2G37340 (RS2Z33) gene.

e, Change in stop codon status from Araport11 Reference to Araport11 Revised. Normal stop to normal stop (38,861), normal stop to PTC_{dEJ} (2,512), PTC_{dEJ} and normal stop (4), and PTC_{dEJ} to PTC_{dEJ} (3,255).

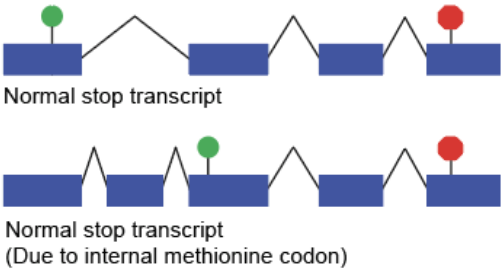
a

	Reference ORFs		Revised ORFs	
	Normal stop	PTC_{dEJ}	Normal stop	PTC_{dEJ}
Up transcripts	3,067	203	2,844	426
No change transcripts	25,623	2,151	23,980	3,794
Down transcripts	4,091	112	4,004	199

b

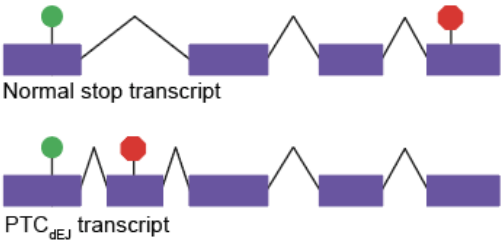


c Araport11 - Reference ORFs



Up/down transcripts with PTC_{dEJ}	Up/down transcripts with normal stop	NMD Enrichment Factor
203/112 = 1.8	3067/4091 = 0.8	= 2.4

d Araport11 - Revised ORFs



Up/down transcripts with PTC_{dEJ}	Up/down transcripts with normal stop	NMD Enrichment Factor
426/199 = 2.1	2844/4004 = 0.7	= 3.0

Fig. 2: Prediction of authentic start-stop codons improves assessment of NMD targets in Araport11

a, Table indicating the number of up (increased steady state expression), no change, and down (decrease steady state expression) with normal stop codons and PTC_{dEJ} in our transcriptomic re-analysis.

b, Boxplot (outliers removed) of 3' UTR lengths (nucleotides; nt) of up (increased steady state expression) transcripts in the NMD-deficient mutant, compared between the Araport11 Reference and Revised ORF annotations. Orange line represents the median. Green triangle represents the mean. Wilcoxon test $p = 1.11 \times 10^{-41}$.

c, Araport11 - Reference ORF annotations and effect on NMD analysis. The number and ratio of up and down transcripts (steady state transcript levels) is shown with either PTC_{dEJ} and normal stop codons. The NMD enrichment factor is the ratio of up/down PTC_{dEJ} transcripts over the ratio of up/down of normal stop transcripts. The higher the number, the higher the enrichment of PTC_{dEJ} changing as predicted based on the EJC model of NMD and not transcripts with normal stop codons.

d, Araport11 - Revised ORF annotations and effect on NMD analysis.