**Direct RNA Sequencing with Nanopores Reveals the Arabidopsis m*6*A Epitranscriptome**

*Or*

**Transcriptome – wide mapping of Arabidopsis m6A in full-length mRNAs**

*Or*

**Context-specific mapping of Arabidopsis m6A**

**Arabidopsis is a crucially important model for plant biology. However, our understanding of what the Arabidopsis genome encodes is subject to revision as our ability to identify RNAs expressed from the genome improves. We investigated the potential of sequencing full-length mRNA molecules directly with protein nanopores to reveal the authentic complexity of Arabidopsis mRNA modifications and processing. N(6)-Methyl-Adenosine (m6A) is a common RNA modification that plays crucial context-specific roles in messenger RNA processing and fate. The Arabidopsis m6A writer complex comprises the methyltransferase and regulatory components that include the conserved protein VIRILIZER (VIR). Here we show that VIR-dependent m6A modifications are detectable in full-length native mRNA using nanopore direct RNA sequencing. In addition to revealing RNA modifications, the same datasets uncover previously unseen combinations of 5’ transcription start sites, splicing events, 3’ site choice and poly(A) tail length. The *vir-1* mutant shows loss of m6A from terminal exons and increases in chimeric RNA formation caused by readthrough into downstream genes. In addition, *vir-1* has an altered circadian period accompanied by poly(A) tail length changes in mRNAs known to be regulated in a circadian-dependent manner. We conclude that nanopore direct RNA sequencing can reveal multiple authentic features of mRNA complexity in full-length single molecule reads. The application of this approach to other species, such as neglected crops, has the potential to transform our understanding of what their genomes encode.**

Patterns of pre-mRNA processing and base modifications determine eukaryotic mRNA coding potential and fate. Alternative transcripts produced from the same gene can differ in the position of the start site, the site of cleavage and polyadenylation and the combination of exons spliced into the mature mRNA. Collectively termed the epitranscriptome, RNA modifications play crucial context-specific roles in gene expression (*1, 2*). The most abundant internal modification of mRNA is methylation of adenosine at the N6 position (m6A) (*1, 2*). Revealing RNA modifications and processing combinations is essential to understand gene expression and what genomes really encode.

RNA sequencing (RNA-Seq) is used to dissect transcriptome complexity. RNA-Seq involves copying RNA into complementary DNA (cDNA) with reverse transcriptases (RT) and it is the subsequent DNA copies that are sequenced. RNA-Seq has transformed our understanding of diverse features of transcriptomes, but limitations can include misidentification of 3’ ends through internal priming (*3*), spurious antisense and splicing events produced by RT template switching (*4*) and the inability to detect some base modifications in the copying process (*5*). The fragmentation of RNA prior to short-read sequencing makes the interpretation of the combination of authentic RNA processing events difficult, and it remains an unsolved problem (*6*).

We asked if long read direct RNA sequencing with nanopores (*7*) could reveal the complexity of *Arabidopsis thaliana* RNA processing and modifications. In nanopore sequencing, intact RNA is fed through a protein pore by a motor protein. The protein pore sits in a membrane through which an electrical current is passed. As RNA moves through the pores, the current is disrupted in a manner by which the raw signal can be interpreted as RNA sequence (*7*). Arabidopsis is a pathfinder model in plant biology, and its genome annotation strongly influences the annotation and understanding of what other plant genomes encode. Viable Arabidopsis mutants defective in RNA processing and modification have helped to establish roles for RNA regulation in controlling fundamental features of plant biology (such as flower development and stress responses) and also facilitate the dissection of features of gene expression that are effectively hidden, for example, if unstable RNAs do not accumulate (*8, 9*).

Here we reveal mRNA m6A and pre-mRNA processing combinations in full-length Arabidopsis mRNAs by applying nanopore direct RNA sequencing and Illumina RNA-Seq to wild-type (Col-0) and mutants defective in m6A (*10*) and exosome mediated RNA decay (*11*).

***Nanopore DRS detects long complex mRNAs and short, structured non-coding RNAs***

We purified poly (A)+ RNA from 4 biological replicates of 14-day old *Arabidopsis thaliana* Col-0 seedlings. We incorporated ERCC synthetic spike-in RNAs into all replicates (*12, 13*) and carried out nanopore direct RNA sequencing. Parallel Illumina RNA-Seq was performed on related material. Using the base-calling and alignment software Guppy (Oxford Nanopore Technologies), we identified around 1 million reads per sample (Table 1). The longest read alignments were 12.7kb for mRNA encoding AT1G48090 (Figure 1A), spanning 64 exons and 12.8kb for mRNA encoding At1G67120, spanning 58 exons (Figure S1A). These represent some of the longest contiguous mRNAs sequenced from Arabidopsis. Among the shortest (but evidently full-length), reads were those that aligned to highly structured non coding RNAs such as UsnRNAs (Figure 1B). Plant UsnRNAs are transcribed from individual genes with well characterised promoters (*14*). In most cases, UsnRNA read alignments were consistent with the corresponding Araport11 annotation. However, nanopore sequencing also revealed that U12 snRNA is processed from a longer, spliced precursor RNA. Consistent with this, U12 precursor RNA accumulates in *hen2-2*, a mutant defective in the activity of the RNA exosome which functions in processing non coding RNAs (*11*) (Figure 1B).

***Base-calling errors in nanopore DRS are non-random***

We used the synthetic ERCC spike-ins as internal controls to monitor properties of the sequencing reads. The spike-ins were detected in a quantitative manner (Figure S1B), consistent with the idea that nanopore sequencing is quantitative (*15*). The error rate in base calling was 8%, as judged against the ERCC spike-in sequences, and were dominated by deletions (Figure S1C). These errors showed evidence of base specificity. For example, guanidine was under-represented and uracil over represented relative to the reference nucleotide (nt) distribution (Figure S1D, E). In some situations, this bias could impact the utility of interpreting nanopore sequence errors. In order to address base calling errors, we used the software tool proovread (*16*) and parallel Illumina RNA-Seq data to error-correct nanopore reads.

***Artefactual splitting of raw signal affects transcript interpretation***

In addition to base-calling errors, we detected artefacts caused by over-splitting of raw signal. As a result of this phenomenon, two distinct transcripts, rather than one, appear to map to the same gene (Figure S1F); the first comprising an apparently novel 3’ end and the second an apparently novel 5’ end (Figure S1F). Over-splitting of reads was recently reported in nanopore DNA sequencing (*17*) and can be detected when two reads sequenced consecutively through the same pore map adjacently to the genome (*17*). Over-splitting generally occurs at low frequency (less than 2% reads) and appears to be associated with MinKNOW software. However, RNAs originating from specific gene loci, such as *RH3* (AT5G26742), appear to be highly susceptible to this effect with over-splitting affecting up to 20% of reads (Fig S1F).

***Spurious antisense reads are rare or absent in nanopore DRS***

Since only 2 reads out of 9,445 (0.02%) mapped antisense to the ERCC spike-in collection (*12*) and 0 of 19665 reads mapped anti-sense to the highly abundant gene RCA (AT2G39730), we concluded that spurious antisense is rare or absent from nanopore data. This simplifies the interpretation of authentic antisense RNAs. For example, we could identify Arabidopsis long non-coding antisense RNAs, such as those at each of a sub-set of the auxin efflux carriers *PIN4*, *5* and *7* (Figure 1C). The existence of these previously unannotated antisense RNAs was supported by Illumina RNA-Seq of wild type Col-0 and the exosome mutant *hen2-2* (Figure 1C), where the abundance of these non-coding antisense RNAs increased 13-fold. Consequently, the low level steady-state accumulation of some antisense RNAs may explain why they are currently unannotated.

***Nanopore DRS confirms sites of RNA 3’ end formation and estimates poly(A) tail length***

The ligation of the motor-protein adapter to RNA 3’ ends results in nanopores sequencing mRNA poly(A) tails first. We used the software tool nanopolish-polyA to estimate poly(A) tail lengths for individual transcripts (Workman et al. 2018). This approach indicated an average length of 76nt for Arabidopsis mRNA poly(A) tails, but with a wide range in estimated lengths for individual mRNAs (95% in the range 13-197nt) (Figure 2A). The generally shorter poly(A) tails of chloroplast and mitochondria-encoded transcripts, that are a feature of RNA decay in these organelles, were also detectable (Figure 2A). We found that poly(A) tail length is negatively correlated (albeit weakly) with gene expression in Arabidopsis (Spearman rho = -0.3, *p*=9.8e-134) (Figure S2A), consistent with other species previously analysed by short read TAILseq (*18*).

We next identified the sites of mRNA cleavage and polyadenylation. We had previously mapped Arabidopsis mRNA 3’ ends transcriptome-wide using Helicos Bio sequencing (*19*). We compared the position of 3’ ends of nanopore read alignments and Helicos Bio data genome-wide. The median genomic distance between nanopore and Helicos Bio 3’ ends is 0+/-13nt (1 standard deviation) demonstrating close agreement between these orthogonal technologies (Figure 2B). Likewise, the overall distribution of the 3’ end of aligned nanopore reads closely resembled the pattern we previously reported with Helicos Bio data (*19*). For example, 97% of nanopore 3’ ends (4,152,800 reads at 639,178 unique sites, 93% of all unique sites) mapped to annotated 3’ untranslated regions (UTRs) or downstream of the current annotation. 3’ ends that mapped to coding sequence or 5’UTRs were rare (2.8%, 119,524 reads at 39,610 unique sites, 5.8% of all unique sites). 3’ ends that mapped to introns were even rarer (0.29%, 12,554 reads at 7,791 unique sites, 1.1% of unique sites), but included examples of alternative polyadenylation with well-established regulatory roles, such as at mRNA encoding the RNA binding protein FPA which controls flowering time (*9*) (Figure 2C), and at mRNA encoding the histone H3K9 demethylase IBM1, which controls levels of genic DNA methylation (*20*), (Table S2B).

Since RT-dependent internal priming can result in the misinterpretation of authentic cleavage and polyadenylation sites (*3*), we asked whether nanopore sequencing was compromised in this way. To address this issue, we examined whether nanopore reads mapped to potential internal priming substrates (comprised of 6 or more consecutive As) predicted to be in transcribed coding sequence according to the Araport11 annotation. Of 10,818 such oligo (A)6 sequences, only 20 mapped within 20 nt of nanopore 3’ ends across all four datasets, 9 of which had Helicos Bio support (at least 3 reads) and 7 of these mapped to the terminal exon of coding sequence annotation, indicating that they may be authentic 3’ ends. Hence, internal priming is rare or absent in nanopore data. Overall, we conclude that nanopore can identify multiple authentic features of RNA 3’ end processing.

***Cap-dependent adapter ligation enables 5’ end detection by nanopore DRS***

In contrast to the clarity of 3’ end detection, nanopore reads frequently truncated prior to annotated transcription start sites, resulting in a 3’ bias of nanopore reads mapped to specific loci (Figure 2A). Consequently, we could not conclude which, if any, of these reads corresponded to full-length transcripts. To address this issue, we used cap-dependent ligation of a biotinylated 5’ adapter RNA to purify capped mRNAs. We re-sequenced two biological replicates of Arabidopsis Col-0 incorporating 5’ adapter ligation (Table 1). We filtered nanopore reads using the sequence alignment tool BLASTN (Table 2) to detect 5’ adapter sequences. We then used high confidence examples of sequences that passed or failed these filters to train a convolutional neural network to detect the 5’ adapter in raw signal (Fig S3A-C). In this way, we improved the identification of 5’ adapter ligated RNAs without requiring genome alignment. As a result, we demonstrated enrichment of full-length, cap-dependent mRNA sequences, filtered for the presence of the 5’ adapter (Figure 3A). This procedure reduced the 3’ bias of nanopore reads (as measured by quartile coefficient of variation of per base coverage) to <0.1 for 56% of genes (Figure 3A). In order to determine whether the 5’ ends we detected reflected full-length mRNAs, we compared them against annotated transcription start sites in datasets derived from full-length Arabidopsis cDNA clones (*21*). We found that 41% of adapter ligated nanopore reads mapped within 5nt of transcription start sites and 60% mapped to within 13nt (Figure 3B). We also detected recently defined examples of alternative 5’ transcription start sites (*22*) at specific Arabidopsis genes (Figure S3D). We therefore conclude that this approach is effective in detecting authentic mRNA 5’ ends.

Reads with adapters had, on average, 11nt more mapped sequence at their 5’ ends compared to reads lacking the 5’ adapter (Figure 3D). This distinction may be explained by loss of processive control by the motor protein when the end of an RNA molecule enters the pore (Workman paper). As a result, the extreme 5’ end of RNA is not correctly sequenced. Consistent with our Arabidopsis transcriptome-wide data, reads mapping to the synthetic ERCC spike-in RNAs and *in vitro* transcribed RNAs also lacked approximately 11 nt of authentic 5’ sequence (Figures S3E, F). However, the precise length of 5’ sequence missing from all these RNAs varied, suggesting that there are sequence or context-specific effects on sequence accuracy associated with the non-processive passage of 5’ RNA through the pore (Figure 3D, S3E, F). Despite the close agreement between nanopore, Illumina RNA-Seq and full-length cDNA start site data at *RCA*, the annotated start site in Araport11 and AtRtD2 is quite different (Figure 3C). This apparent overestimation of 5’UTR length was widespread in Araport11 annotation (Figure S3G). Consequently, nanopore data can be used to revise Arabidopsis transcription start site annotation.

***Detection of novel splice sites and unannotated combinations of established splicing events***

We next investigated the utility of nanopore to reveal patterns of splicing. In single reads, nanopore sequencing revealed some of the most complex splicing combinations known in the Arabidopsis transcriptome. For example, the splicing of the 12.7 kb read, comprised of 66 exons, agreed exactly with the At1g48090.4 isoform annotated in Araport 11. (Figure 1A). Mutually exclusive alternative splicing of *FLM* exons that mediate the thermosensitive response controlling flowering time (*23*) was also clearly detected (Figure 4A). However, considering the non-corrected dataset as a whole, a combination of read and alignment errors probably resulted in the mis-calling of splicing events because 58% (170,702) of unique junctions were absent from Araport11 and AtRTD2 (*24*) annotations, and unsupported by Illumina RNA-Seq (Figure 4B, Table 3). We applied proovread (*16*) error correction with the parallel Illumina RNA-Seq data and re-analysed the corrected and uncorrected nanopore sequences. Most canonical GU/AG splicing events (100,450; 81%) detected in the error corrected nanopore data were found in both annotations and were supported by Illumina RNA-Seq (Figure 4B, Table 3). The four Col-0 biological replicate nanopore datasets capture 75% (102,486) and 69% 104,686) of the Araport11 and AtRTD2 splice site annotations, respectively. 3,234 unique canonical splicing events in the error corrected nanopore data were supported by Illumina RNA-Seq, but were absent from both Araport11 and AtRTD2 annotations, highlighting potential gaps in the complexity of Arabidopsis splicing annotation (Figure 4B, Table 3). Consistent with this, we validated 3 of these splicing events by using RT-PCR followed by cloning and sequencing (Figure S4A). In order to examine the features of the unannotated splices, we applied previously determined splice site position weight matrices of the flanking sequences to categorize U2 or 12 class splice sites (*25*) (Figure S4B). The U2/U12 classifications lend weight to the idea that they constitute authentic splicing events., Unannotated events were generally supported by fewer reads (95% confidence intervals for all the data distributions are 5 counts or fewer) than the annotated splices (median expression of > 10 counts in all categories) (Figure S4B). Consequently, expression level differences may explain the absence of at least some of these events from current annotations.

In addition to previously unannotated splicing events, we identified unannotated combinations of previously established splice sites. For example, we identified 19 *FLM* transcript isoforms which all adhered to known splice junction sites. However, 11 of these transcript isoforms were not previously annotated (Figure 4A). In order to investigate this phenomenon transcriptome-wide, we analysed the 5’ cap-dependent adapter ligated nanopore datasets of full-length mRNAs. Unique sets of co-splicing events were extracted from error-corrected reads to represent isoforms (in order to focus on splicing, we did not consider single exon reads or 5’ and 3’ positions). The read coverage of the 5’ adapter-ligated datasets is lower than the conventional direct RNA sequencing approach (Table 1), but 13,064 transcript isoforms were detected that matched annotations in one or both of Araport11 and AtRTD2 (Figure 4C). 8,659 transcript isoforms were identified in the error corrected nanopore data that were not identified in either annotation (Figure 4C, Table 3). Of these, 50% (4,293) used only splice donor and acceptor pairs that were already annotated in either Araport11 or AtRTD2. Hence, this approach defines novel transcript isoforms produced from alternative combinations of known splice sites.

Overall, we conclude that nanopore can reveal the complexity of splicing in the context of full-length mRNAs. However, accurate splice isoform detection requires error correction of nanopore reads with orthogonal sequence data. Accurate splice detection, even with error-free sequences, can be confounded because equivalent alternative junctions often exist (*26*). Therefore, improved computational tools for not only error correction, but also splicing-aware long-read alignment are required.

***Differential error site analysis reveals the m6A epitranscriptome***

The epitranscriptome has emerged recently as a crucial, but relatively neglected, layer of gene expression (*1, 2*). m6A has been mapped transcriptome-wide using approaches based on antibodies that recognize this mark (*5, 27*). However, in principle, m6A can be detected by nanopore sequencing (*7*). Since m6A is not included in the training data for nanopore base-calling software, we asked whether its incorrect interpretation could be used to identify Arabidopsis m6A transcriptome-wide. We sequenced 4 biological replicates of an Arabidopsis mutant defective in the function of a conserved m6A writer complex component, Virilizer, and 4 biological replicates of a complementing line using nanopore direct RNA sequencing (*10*) (Figure S5). In addition, we sequenced a parallel set of biological replicates with Illumina RNA-Seq. We then used a G-test statistical analysis to determine if there was a differential error rate between the mutant (defective m6A) and the complementing lines. We discovered more than 17,000 sites, with more than two-fold difference in error-rate, specific to the complementing line with restored m6A (Figure 5A). No VIR-dependent error sites mapped to either chloroplast or mitochondrial-encoded RNAs. 99.8% of differential error sites map to mRNA. These sites were enriched in the consensus m6A target sequence DRAYH (D=G or U or A; R=G or A; Y=C or U; H=A or C or U) (Figure 5B) and were asymmetrically distributed to 3’UTRs (Figure 5C). Since approximately 5nt contribute to the observed current at a given time point in nanopore sequencing (*7*), the presence of a methylated adenosine could affect the accuracy of base-calling of surrounding nucleotides. Consistent with this, we identified 4,749 consensus motifs at error sites (FDR < 0.1), with an average of 2.4 error sites per m6A motif. Since the sequence motif and asymmetric distribution of these sites are consistent with established and conserved properties of authentic m6A (*1, 2*), it suggested that nanopore was able to identify thousands of VIR-dependent m6A sites transcriptome-wide.

In order to examine the validity of m6A sites identified by the differential error site analysis, we used an orthogonal technique to map m6A sites. Previous maps of Arabidopsis m6A have used an approach called Me-RIP (*8, 28*). However, a limitation of Me-RIP is the resolution of around 200nt (*29*). Therefore, to examine Arabidopsis m6A sites with a more accurate method, we used miCLIP analysis of 3 biological replicates of wild-type Arabidopsis Col-0 (*27*). In this approach, anti-m6A antibody is cross-linked with UV light to m6A-containing fragments of RNA. Following proteinase digestion of the antibody, residual covalently cross-linked peptide remains in the vicinity of m6A. The peptide adducts, derived from the antibody that we used, predominantly induce truncations in RT copies during library preparation, enabling the more precise mapping of m6A sites (*27*). We found that, like the differential error sites uncovered in nanopore analysis, the Arabidopsis miCLIP reads were enriched in 3’ UTRs, with no enrichment over stop codons (Figure 5D). 77% of called nanopore differential error sites fall within 5nt of an miCLIP peak (Figure 5E,F). We therefore conclude that nanopore can detect authentic Virilizer writer-complex-dependent m6A sites transcriptome-wide.

***The m6A writer complex is required for circadian rhythm and flowering time gene control.***

The combination of transcript processing and modification data obtained using nanopore direct RNA sequencing enabled us to investigate the impacts of m6A on Arabidopsis gene expression. We found a global reduction in protein-coding gene expression in *vir-1* corresponding to transcripts that were methylated in the complemented line (Figure 6A). These findings are consistent with the recent discovery that m6A predominantly protects Arabidopsis mRNAs from endonucleolytic decay (*8*). Therefore, although the writer complex comprises conserved components that target a conserved consensus sequence and asymmetric distribution of m6A to the last exon, it appears that this modification predominantly promotes the decay of mRNAs in human cells (*30*), but promotes the stability of mRNAs in Arabidopsis (*8*).

The changes in gene expression in *vir-1* were wide ranging. For example, we found that the abundance of mRNAs encoding components of the interlocking transcriptional feedback loops that comprise the Arabidopsis circadian oscillator (*31*) such as *CCA1* and *TOC1*, were altered in *vir-1* (Figure S6A). This distinction had a biological consequence because the *vir-1* mutant exhibited a defect in clock period (Figure 6B, C). We also found that previously identified targets of the circadian clock, such as *CAB1,* showed changes in poly(A) tail length in *vir-1* compared to Col-0 and the complemented line (Figure 6D)*,* likely identifying transcripts where poly(A) tail length is controlled in a circadian manner (*32*)*.* One output of the circadian clock is the control of flowering time and we found that, not only was the expression of photoperiod pathway components altered in *vir-1*, but other flowering time genes too (12% of established flowering time genes differentially expressed; Table X). Notably, detectable levels of sense and antisense RNA at the *FLC* locus were reduced compared to wild-type (Figure S6C). Consequently, the proper control of circadian rhythms, flowering time and the regulatory module at *FLC* ultimately requires the m6A writer complex component, VIR.

***Defective m6A writer complex function results in chimeric RNA formation***

In addition to measuring RNA expression, we asked what impact loss of m6A had on pre-mRNA processing. Detectable disruptions to splicing in *vir-1* were modest. For example, using DEX-Seq analysis of annotated splice sites, we found only weak effect size changes to cassette exons, retained introns or alternative donor/acceptor sites compared to the complemented line (Figure S6C). In contrast, a clear defect in RNA 3’ end formation in *vir-1* was apparent. Using a Kolmogorov Smirnov test, we identified 3,579 genes which had a change in nanopore 3’ position profile in the *vir-1* mutant compared to the complemented line (FDR < 0.05, absolute change in position > 13nt). Of these, 3,008 displayed a shift to usage of more proximal poly(A) sites in *vir-1*. 60% of these genes also contain m6A sites detectable by nanopore (70% detectable by miCLIP) and correspond to locations of increased cleavage downstream of m6A sites in the *vir-1* mutant (Figure S6D). A further 571 genes showed increased transcriptional readthrough beyond the 3’ end in *vir-1*. 73% of these loci also contained nanopore-mapped m6A sites (78% by miCLIP). Impacts of altered 3’ processing are complex, but had the potential to change the relative abundance of certain CDS transcripts. For example, we detected increased readthrough of an intronic cleavage site in the Symplekin domain gene *TANG1* (AT1G27595); and increased readthrough and cryptic splicing at *ALG3* (AT2G47760)that alsoresults in chimeric RNA formation with the downstream gene AT2G09955 (Figure 6E). The existence of these chimeric RNAs was supported by Illumina RNAseq (Figure 6E) and confirmed by RT-PCR, cloning and sequencing (Figure S6E). We detected 523 loci with increased levels of chimeric RNAs in *vir-1* resulting from unterminated transcription proceeding into downstream genes on the same strand. However, only 33% of upstream genes that comprised chimeric RNAs had detectable m6A sites in the complementing line restoring VIR activity. Consequently, these findings might be explained either by an m6A-independent role for the writer complex in 3’ end formation or an indirect effect of the writer complex on factors required for 3’ processing. Since m6A independent roles for the human m6A methyltransferases METTL3 (*33*) and METTL16 (*34*) have been found, a role for the writer complex in controlling Arabidopsis RNA processing independent of m6A cannot be overlooked (*34*). In mammals, recognition of the canonical poly(A) signal AAUAAA involves direct binding by CPSF30 (*35, 36*). Notably, an alternatively spliced isoform of Arabidopsis CPSF30 encodes a YTH domain with the potential to bind and read m6A (*37*). Consequently, in plants, m6A may also contribute to the recognition of specific RNA 3’ ends.

***Concluding Remarks***

There is clear evidence for the coupling of transcription and pre-mRNA processing events (*38*). We provide the full-length transcript and modification data required to analyse this phenomenon in Arabidopsis. In addition, we show that defective RNA processing can result in a set of unexpected events (for example, chimeric RNAs, new termination sites, new transcription start sites and cryptic splicing events) and that the interpretation of these *de novo* events is clarified by long read data.

We provide the clearest view of the Arabidopsis epitranscriptome to date. However, since our approach depends upon statistical analysis, we cannot attribute m6A to an individual RNA molecule or sequencing read. In the future, machine learning could refine base calling software to incorporate base modifications and our data provides a validation resource for such approaches.

We show that nanopore direct RNA sequencing has the potential to refine, not only, multiple features of Arabidopsis genome annotation, but also the analysis of gene expression transcriptome-wide. This is even though the sequence of the Arabidopsis genome has been available since 2000 (*39*). Our ability to understand other plant genomes is of pressing importance. Modern agriculture is dominated by a handful of intensely researched crops (*40*). However, it has been estimated that humans have cultivated or collected 5000-7000 different plant species for food. To diversify global food supply, enhance agricultural productivity and tackle malnutrition, there is a need to focus on crops utilized in rural societies that have received little attention for crop improvement (Gigascience reference). The application of genome assisted breeding to evolutionarily diverse neglected crops is hampered by the absence of genome sequences and high quality annotation. Based on our experience with Arabidopsis, we anticipate that the combined application of nanopore direct RNA sequencing and other sequencing approaches will be hugely valuable to orphan crop annotation.

***Acknowledgements***

This work was funded by BBSRC grants BB/J00247X/1, BB/M004155/1, BB/M010066/1 and Dundee University’s Global Challenges Research Fund.

***References***

1. K. D. Meyer, S. R. Jaffrey, Rethinking m(6)A Readers, Writers, and Erasers. *Annu Rev Cell Dev Bi* **33**, 319-342 (2017).

2. I. A. Roundtree, M. E. Evans, T. Pan, C. He, Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187-1200 (2017).

3. C. H. Jan, R. C. Friedman, J. G. Ruby, D. P. Bartel, Formation, regulation and evolution of Caenorhabditis elegans 3 ' UTRs. *Nature* **469**, 97-U114 (2011).

4. J. Houseley, D. Tollervey, Apparent Non-Canonical Trans-Splicing Is Generated by Reverse Transcriptase In Vitro. *Plos One* **5**, (2010).

5. M. Helm, Y. Motorin, Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature Reviews Genetics* **18**, 275-291 (2017).

6. T. Steijger *et al.*, Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**, 1177-+ (2013).

7. D. R. Garalde *et al.*, Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**, 201-206 (2018).

8. S. J. Anderson *et al.*, N-6-Methyladenosine Inhibits Local Ribonucleolytic Cleavage to Stabilize mRNAs in Arabidopsis. *Cell Rep* **25**, 1146-+ (2018).

9. C. Hornyik, L. C. Terzi, G. G. Simpson, The spen family protein FPA controls alternative cleavage and polyadenylation of RNA. *Dev Cell* **18**, 203-213 (2010).

10. K. Ruzicka *et al.*, Identification of factors required for m(6) A mRNA methylation in Arabidopsis reveals a role for the conserved E3 ubiquitin ligase HAKAI. *New Phytol* **215**, 157-172 (2017).

11. H. Lange *et al.*, The RNA Helicases AtMTR4 and HEN2 Target Specific Subsets of Nuclear Transcripts for Degradation by the Nuclear Exosome in Arabidopsis thaliana. *Plos Genetics* **10**, (2014).

12. L. C. Jiang *et al.*, Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **21**, 1543-1551 (2011).

13. L. H. Reid *et al.*, Proposed methods for testing and selecting the ERCC external RNA controls. *Bmc Genomics* **6**, (2005).

14. F. Waibel, W. Filipowicz, Rna-Polymerase Specificity of Transcription of Arabidopsis U-Snrna Genes Determined by Promoter Element Spacing. *Nature* **346**, 199-202 (1990).

15. D. R. Garalde *et al.*, Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods* **15**, 201 (2018).

16. T. Hackl, R. Hedrich, J. Schultz, F. Forster, proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-3011 (2014).

17. A. Payne, N. Holmes, V. Rakyan, M. Loose, BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*, (2018).

18. S. A. Lima *et al.*, Short poly(A) tails are a conserved feature of highly expressed genes. *Nature Structural & Molecular Biology* **24**, 1057-+ (2017).

19. A. Sherstnev *et al.*, Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat Struct Mol Biol* **19**, 845-852 (2012).

20. M. Rigal, Z. Kevei, T. Pelissier, O. Mathieu, DNA methylation in an intron of the IBM1 histone demethylase gene stabilizes chromatin modification patterns. *EMBO J* **31**, 2981-2993 (2012).

21. M. Seki *et al.*, Functional annotation of a full-length Arabidopsis cDNA collection. *Science* **296**, 141-145 (2002).

22. T. Ushijima *et al.*, Light Controls Protein Localization through Phytochrome-Mediated Alternative Promoter Selection. *Cell* **171**, 1316-+ (2017).

23. D. Pose *et al.*, Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature* **503**, 414-417 (2013).

24. R. X. Zhang *et al.*, A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Res* **45**, 5061-5073 (2017).

25. N. Sheth *et al.*, Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34**, 3955-3967 (2006).

26. R. Dehghannasiri, L. Szabo, J. Salzman, Ambiguous splice sites distinguish circRNA and linear splicing in the human genome. *Bioinformatics*, (2018).

27. B. Linder *et al.*, Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* **12**, 767-772 (2015).

28. L. S. Shen *et al.*, N-6-Methyladenosine RNA Modification Regulates Shoot Stem Cell Fate in Arabidopsis. *Dev Cell* **38**, 186-200 (2016).

29. S. D. Ke *et al.*, A majority of m(6)A residues are in the last exons, allowing the potential for 3 ' UTR regulation. *Gene Dev* **29**, 2037-2053 (2015).

30. X. Wang *et al.*, N-6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117-+ (2014).

31. C. R. McClung, The Plant Circadian Oscillator. *Biology (Basel)* **8**, (2019).

32. P. Lidder, R. A. Gutierrez, P. A. Salome, C. R. McClung, P. J. Green, Circadian control of messenger RNA stability. Association with a sequence-specific messenger RNA decay pathway. *Plant Physiology* **138**, 2374-2385 (2005).

33. S. Lin, J. Choe, P. Du, R. Triboulet, R. I. Gregory, The m(6)A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. *Mol Cell* **62**, 335-345 (2016).

34. K. E. Pendleton *et al.*, The U6 snRNA m(6)A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell* **169**, 824-835 e814 (2017).

35. S. L. Chan *et al.*, CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3 ' processing. *Gene Dev* **28**, 2370-2380 (2014).

36. L. Schonemann *et al.*, Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Gene Dev* **28**, 2381-2393 (2014).

37. A. T. Stevens, D. K. Howe, A. G. Hunt, Characterization of mRNA polyadenylation in the apicomplexa. *Plos One* **13**, (2018).

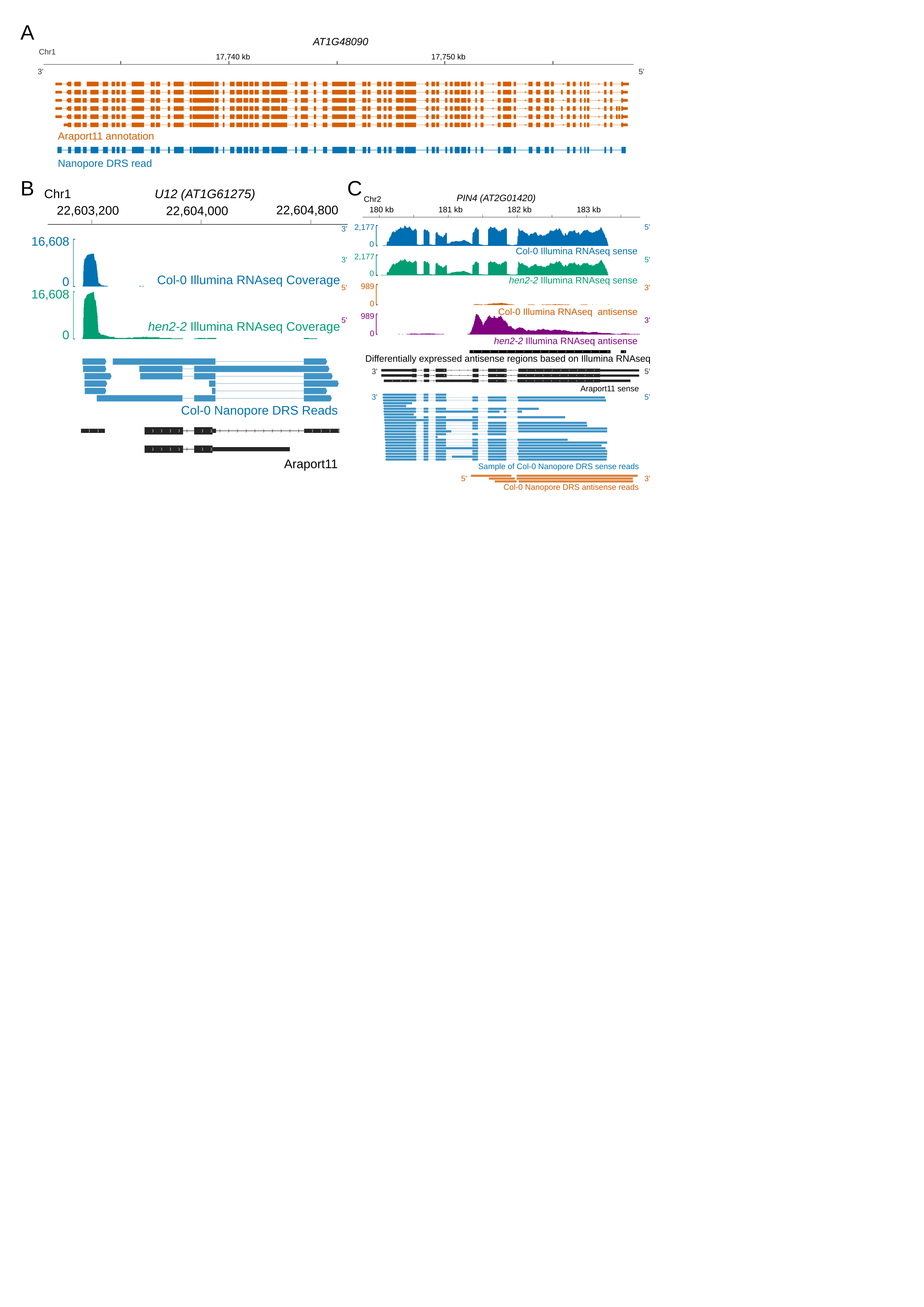
38. S. Y. Anvar *et al.*, Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol* **19**, (2018).

39. S. Kaul *et al.*, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).

40. *World population prospects: the 2017 revision, Key Findings and Advance Tables.* (2017).

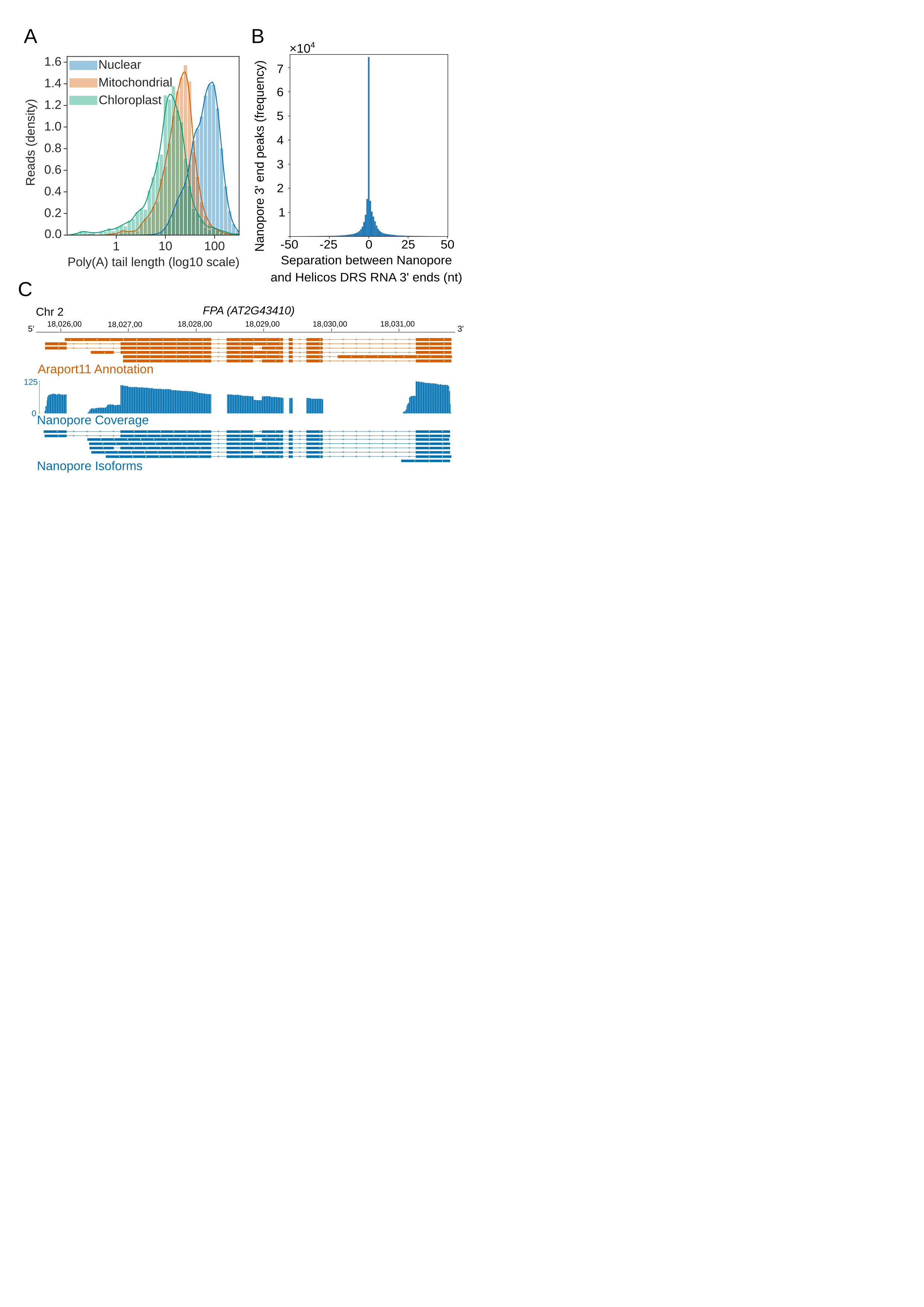
41. K. Jaganathan *et al.*, Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-+ (2019).

42. m. n. g. o. a. a. Genomes Consortium. Electronic address, C. Genomes, 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481-491 (2016).



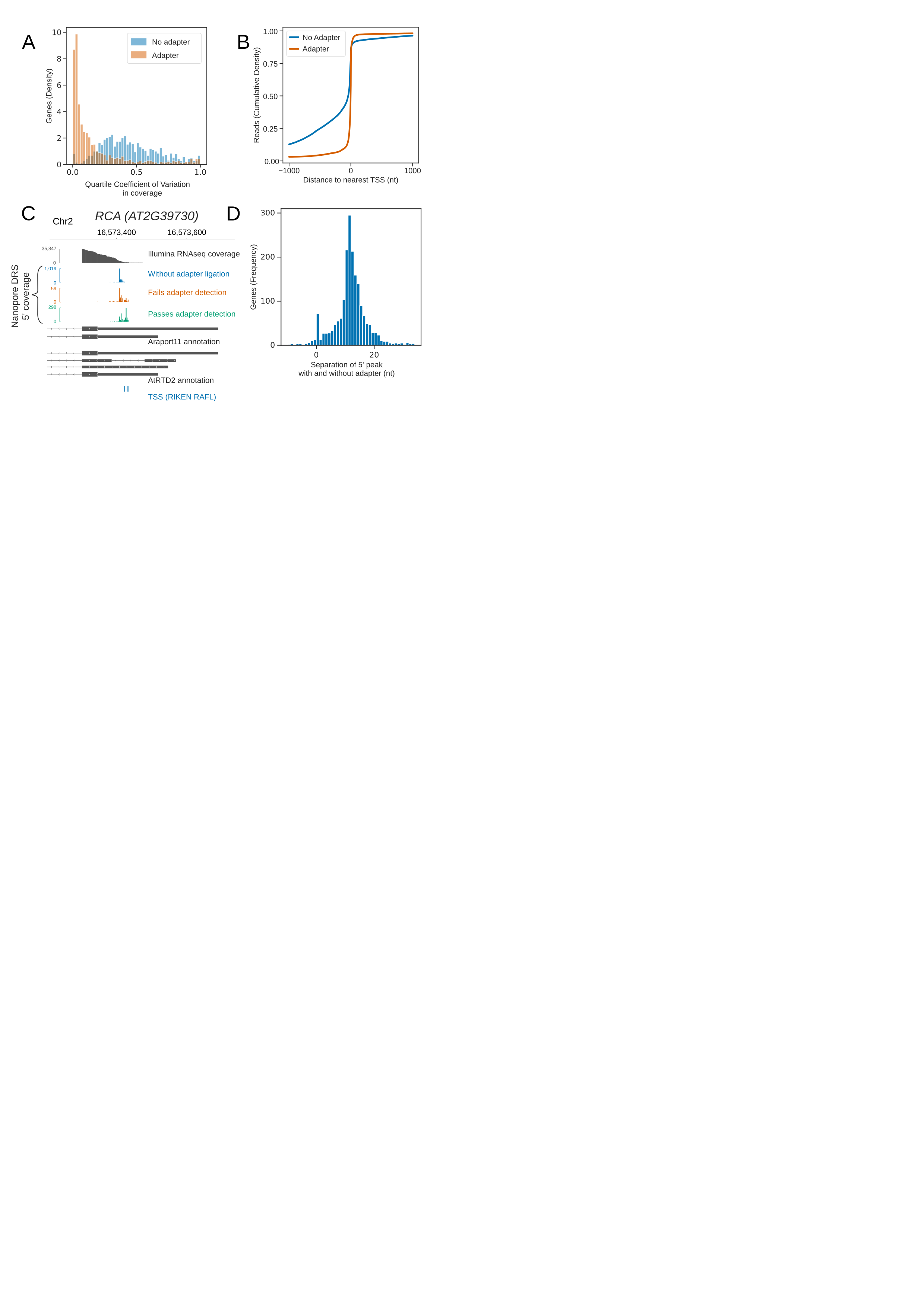
**Figure 1. Nanopore DRS provides new insight into Arabidopsis RNA processing**

1. Nanopore DRS identified 12.66Mb transcript generated from *AT1G48090* gene that includes 66 exons. RNA isoforms present in the *Araport11* annotation are shown in orange and an RNA isoform identified using nanopore DRS – in blue.
2. Nanopore DRS in combination with Illumina RNAseq data of the *hen2-2* exosome mutant identified an unannotated pre-curser RNA of U12 snRNA. Col-0 Illumina RNAseq coverage and Nanopore isoforms are shown in blue; *hen2-2* Illumina RNAseq coverage – in green; Araport11 annotation – in black.
3. *PIN4* long non-coding antisense RNAs were detected using Nanopore DRS. Col-0 sense Illumina RNAseq coverage and Nanopore sense isoforms are shown in blue; Col-0 antisense Illumina RNAseq coverage and Nanopore antisense isoforms – in orange; *hen2-2* mutant sense Illumina RNAseq coverage – in green; *hen2-2* mutant antisense Illumina RNAseq coverage – in purple. Sense RNA isoforms found in Araport11 and antisense differentially expressed regions are shown in black.



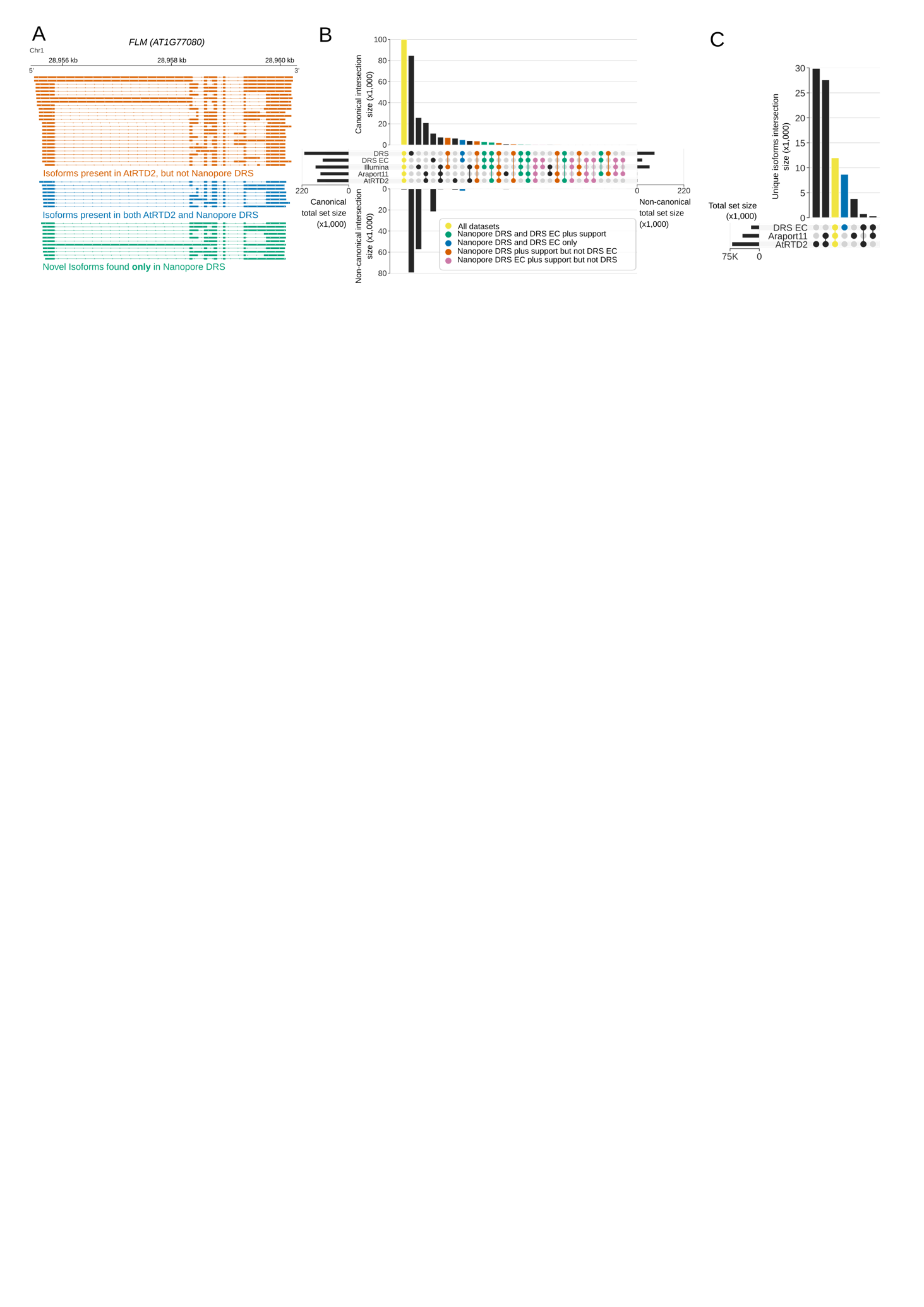
**Figure 2. Nanopore DRS identified length of poly(A) tails and accurately showed locations of 3’ polyadenylation sites.**

1. Normalised histogram of poly(A) tail length of RNAs found in different plant organelles. Nuclear poly(A) tails are shown in blue, mitochondrial – in orange, and chloroplast – in green.
2. Separation between the RNA 3’ end positions in Nanopore DRS read alignments and the nearest polyadenylation sites identified in Helicos Bio DRS data.
3. Nanopore DRS identified new 3’ polyadenylation sites in RNAs transcribed from *FPA (AT2G43410)* gene. Blue track shows coverage of Nanopore DRS reads. Isoforms found in Araport11 annotation are shown in orange and those detected by Nanopore DRS are in blue.



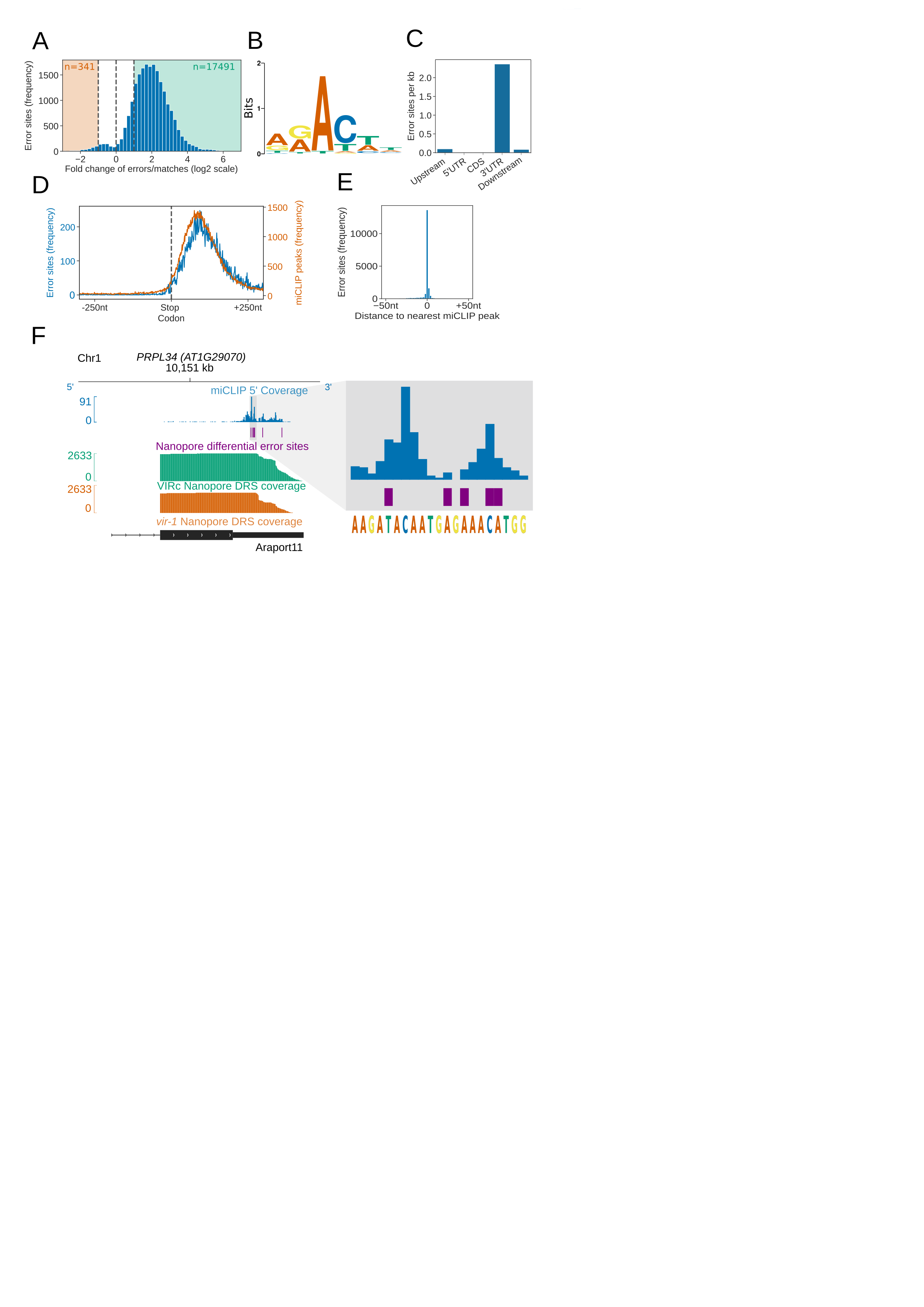
**Figure 3. Cap-dependent ligation of 5’ adapter allowed accurate resolution of RNA 5’ ends.**

1. 5’ adapter ligation reduces 3’ bias in Nanopore DRS data. Histogram shows reduction in quartile coefficient of variation in per base coverage for reads with 5’ adapter (orange), compared to reads without 5’ adapter (blue).
2. Cap-dependent adapter ligation allows identification of authentic 5’ ends using Nanopore DRS. Cumulative distribution function shows the distance to the nearest TSS identified from full length transcripts cloned as part of the RIKEN RAFL project for reads with 5’ adapter (orange), compared to reads without 5’ adapter (blue).
3. For *RCA (AT2G39730),* the 5’ end identified using cap-dependent adapter ligation protocol was consistent with Illumina RNAseq and full-length cDNA start site data, but differed from the 5’ ends found in the Araport11 and AtRTD2 annotations. Illumina RNAseq coverage is shown in grey and Nanopore DRS 5’ end coverage generated without cap-dependent ligation protocol – in blue. Coverage for reads generated using the cap-dependent ligation protocol with and without 5’ adapter are shown in green and orange, respectively. RNA isoforms found in Araport11 and AtRTD2 annotations are shown in grey and TSSs identified from full length transcripts cloned as part of the RIKEN RAFL project - in blue.
4. Cap-dependent adapter ligation allowed resolution of additional 11 nt of sequence at the RNA 5’ end. Histogram shows nt shift in the largest peak of 5’ coverage for each gene in data obtained using protocol with 5’ adapter vs protocol without 5’ adapter.



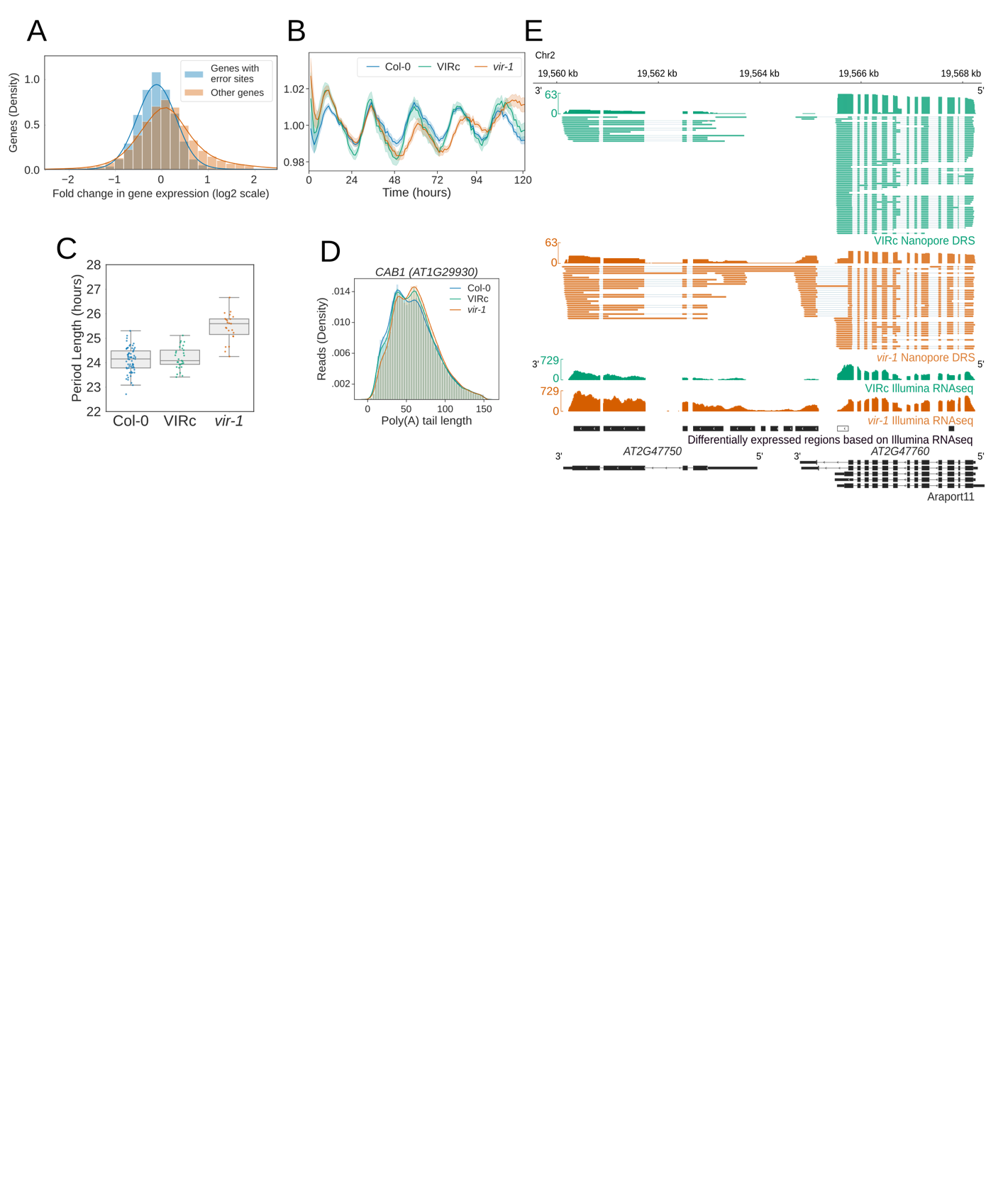
**Figure 4. Nanopore DRS revealed new details of RNA splicing in plants.**

1. Nanopore DRS identified unannotated combinations of established splice sites in alternative splicing pattern of an extensively studied *FLOWERING LOCUS M* (*FLM, AT1G77080)* gene. Isoforms present in AtRTD2 annotation, but not identified using Nanopore DRS are shown in orange, isoforms common to both AtRTD2 and Nanopore DRS are in blue, novel isoforms identified in DRS are shown in green.
2. Comparison of splicing events identified in error-corrected and non-error-corrected Nanopore DRS, Illumina RNA-sequencing, Araport11 and AtRTD2 annotations. Bar size represents number of unique splicing events common to the set intersection highlighted using circles (please refer to Table XX for the exact values). Canonical GT/AG splicing events are shown on the top and non-canonical on the bottom of the plot. Splicing events common to all five datasets are shown in yellow; events common to both error-corrected and non-error-corrected nanopore DRS with support in orthogonal datasets are in green; events common to both Nanopore datasets without orthogonal support are in blue; events found in uncorrected Nanopore DRS (but not error corrected) with orthogonal support are in orange; and events found in error-corrected nanopore DRS (but not uncorrected) with orthogonal support are in pink.
3. Comparison of RNA isoforms (defined as sets of co-spliced introns) identified in error-corrected full-length Nanopore DRS, Araport11 and AtRTD2 annotations. Bar size represents number of splicing events common to a group highlighted using circles below (please refer to Table XX for the exact values). Isoforms identified in Nanopore DRS and both reference annotations are shown in yellow. Novel isoforms are shown in blue.

****

**Figure 5. Differential error rate analysis identified sites of VIR-dependent m6A modifications transcriptome-wide**

1. Loss of VIR function causes a reduction in error rate at a large number of sites in the *Arabidopsis* transcriptome. Histogram shows the log2 fold change in ratio of errors to reference matches at bases with a significant change in error profile in *vir-1* mutant compared to the VIR complemented line. Orange and green shaded regions show sites with increased and reduced errors in *vir-1,* respectively.
2. Motif of the error rate sites matches consensus m6A target sequence. Sequence logo of the motif enriched at sites with reduced error rate in the *vir-1* mutant.
3. Differential error rate sites are primarily found in the 3’ UTRs. Bar plot shows the number of differential error rate sites per kb of genic feature types in the Araport11 reference. Upstream and downstream regions are 200 nt regions before and after the annotated TTS, respectively.
4. Differential error rate sites and miCLIP peaks are similarly distributed within the 3’ UTR without accumulation at the stop codon. Metagene plot centred on stop codons shows the frequency of nanopore error sites in blue and miCLIP peaks in orange.
5. The location of differential error rate sites is in good agreement with the locations of miCLIP sites. Histogram shows distribution of distances to the nearest miCLIP peak for each site of reduced error. 77% of error sites are within 5 nt of an miCLIP peak.
6. Nanopore differential error sites analysis and miCLIP identify m6A sites in the 3’ UTR of *PRPL34* RNA. miCLIP 5’ end coverage is shown in blue; Nanopore differential error sites – in purple; Nanopore DRS coverage of VIR complemented line – in green; Nanopore DRS coverage of *vir-1* mutant – in orange; RNA isoform from Araport11 annotation – in black.

****

**Figure 6. Reduction in m6A RNA modification results in changes to circadian clock and generation of chimeric RNAs.**

1. Genes with differential error rate sites have lower expression. Histogram shows log2 fold change in gene expression based on counts from Nanopore DRS reads in *vir-1* mutant and VIR complemented line. Genes with differential error rate sites are in blue, and those without – in orange.
2. Circadian period is lengthened in *vir-1* mutant. Delayed fluorescence measurements for Col-0 are in blue, VIR complemented line – in green and *vir-1* mutant – in orange.
3. Boxplot shows period lengths of Col-0 (blue), VIR complemented line (green) and *vir-1* mutant (orange) calculated from delayed fluorescence measurements shown in (B).
4. Poly(A) tails of RNA expressed from circadianly regulated gene are longer in *vir-1* mutant. Histogram shows poly(A) tail length distribution of *CAB1 (AT1G29930)* gene in Col-0 (blue), VIR complemented line (green) and *vir-1* mutant (orange).
5. Readthough events and chimeric RNAs are detected in *vir-1* mutant. Nanopore DRS and Illumina RNAseq data for VIR complemented line are shown in green; Nanopore DRS and Illumina RNAseq data for *vir-1* mutant – in orange; RNA isoforms found in Araport11 annotation – in black. Differentially expressed regions detected using Illumina RNAseq data with DERfinder are shown in black (for upregulated regions) or white (for downregulated regions).