**Nanopore Direct RNA Sequencing Maps the Arabidopsis m6A Epitranscriptome**

Matthew T. Parker1†, Katarzyna Knop1†, Anna V. Sherwood1†, Nicholas J. Schurch1†‡, Katarzyna Mackinnon1, Peter D. Gould2, Anthony Hall3, Geoffrey J. Barton1\* and Gordon G. Simpson1,4\*

1School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK

2Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK

3Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK

4James Hutton Institute, Invergowrie, DD2 5DA, UK

†Contributed equally

\*Correspondence: [g.g.simpson@dundee.ac.uk](mailto:g.g.simpson@dundee.ac.uk) ; [g.j.barton@dundee.ac.uk](mailto:g.j.barton@dundee.ac.uk)

Current Address:

‡Biomathematics and Statistics Scotland, The James Hutton Institute, Aberdeen, AB15 8QH, UK

**Abstract**

**Understanding genome organization and gene regulation requires insight into RNA transcription, processing and modification. We adapted Oxford Nanopore Direct RNA Sequencing to examine RNA from a wild-type accession of the model plant *Arabidopsis thaliana* and a mutant defective in mRNA methylation (m6A). Here we show that m6A can be mapped in full-length mRNAs transcriptome-wide and reveal the complexity of combinations of cap-associated transcription start sites, splicing events, poly(A) site choice and poly(A) tail length. Loss of m6A from terminal exons is associated with decreased transcript abundance and defective RNA 3’ end formation. A functional consequence of disrupted m6A is a lengthening of the circadian period. Hence, nanopore direct RNA sequencing can reveal the complexity of mRNA processing and modification in full-length single molecule reads. These findings refine Arabidopsis genome annotation and, if applied to less well-studied species, this approach has the potential to transform our understanding of what their genomes encode.**

**Introduction**

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 expression1,2. The most abundant internal modification of mRNA is methylation of adenosine at the N6 position (m6A). Revealing RNA modifications and processing combinations is essential to understand gene expression and what genomes really encode.

RNA sequencing (RNAseq) is used to dissect transcriptome complexity. RNAseq involves copying RNA into complementary DNA (cDNA) with reverse transcriptases (RT) and it is the subsequent DNA copies that are sequenced. RNAseq reveals diverse features of transcriptomes, but limitations can include misidentification of 3’ ends through internal priming3, spurious antisense4 and splicing events produced by RT template switching5 and the inability to detect all base modifications in the copying process6. 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 problem7.

We asked if long read direct RNA sequencing (DRS) with nanopores8 could reveal the complexity of Arabidopsis mRNA processing and modifications. In nanopore DRS, intact RNA is fed through a protein pore (nanopore) by a motor protein. The nanopore sits in a membrane through which an electrical current is passed8. Each short RNA sequence (5 bases) within the nanopore can be identified by the magnitude of the signal it produces. Arabidopsis is a pathfinder model in plant biology, and its genome annotation strongly influences the annotation and understanding of what other plant genomes encode. We applied nanopore DRS and Illumina RNAseq to wild-type Arabidopsis (Col-0), and to mutants defective in m6A9 and exosome mediated RNA decay10. We reveal m6A and combinations of RNA processing events (alternative patterns of 5’ capped transcription start sites, splicing, 3’ polyadenylation and poly(A) tail length) in full-length Arabidopsis mRNAs transcriptome-wide.

**Results**

*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 ArabidopsisCol-0 seedlings. We incorporated synthetic ERCC RNA Spike-In mixes into all replicates11,12 and carried out nanopore DRS. Parallel Illumina RNAseq was performed on related material. Using the base-calling and alignment software Guppy (Oxford Nanopore Technologies), we identified around 1 million reads per sample (Supplementary Table S1). The longest read alignments were 12.7kb for mRNA encoding *AT1G48090* (Fig. 1a), spanning 63 exons, and 12.8kb for mRNA encoding *At1G67120*, spanning 58 exons (Supplementary Fig. S1a). These represent some of the longest contiguous mRNAs sequenced from Arabidopsis. Among the shortest read-alignments were those spanning genes encoding highly structured non-coding RNAs such as UsnRNAs and snoRNAs like U3 (Fig. 1b).

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

We used ERCC RNA Spike-Ins11 as internal controls to monitor properties of the sequencing reads. The spike-ins were detected in a quantitative manner (Supplementary Fig. S1b), consistent with the idea that nanopore sequencing is quantitative8. For the portion of the reads that align to the reference, the sequence identity is 92% as measured against the ERCC RNA spike-ins (Supplementary Fig. S1c). The errors showed evidence of base specificity (Supplementary Fig. S1d, e). For example, guanidine was under-represented and uracil over-represented in indels and substitutions relative to the reference nucleotide (nt) distribution. 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 proovread13 and parallel Illumina RNAseq data to error-correct nanopore reads.

*Artefactual splitting of raw signal affects transcript interpretation*

We detected artefacts caused by MinKNOW software splitting raw signal from single molecules into two or more reads. As a result, alignments comprising apparently novel 3’ ends map adjacent to alignments with apparently novel 5’ ends (Supplementary Fig. S1f). A related phenomenon called over-splitting was recently reported in nanopore DNA sequencing14. Over-splitting can be detected when two reads sequenced consecutively through the same pore map adjacently in the genome14. Over-splitting in nanopore DRS generally occurs at low frequency (less than 2% reads). However, RNAs originating from specific gene loci, such as *RH3* (*AT5G26742*), appear to be more susceptible, with 20% of reads affected (Supplementary Fig. S1f).

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

Since only 2 of 9,445 (0.02%) reads mapped antisense to the ERCC RNA Spike-In collection11 and 0 of 19,665 reads mapped antisense to the highly expressed gene *RCA* (*AT2G39730*), we concluded that spurious antisense is rare or absent from nanopore DRS data. This simplifies the interpretation of authentic antisense RNAs. This is important in Arabidopsis because the distinction between RT-dependent template switching and authentic antisense RNAs produced by RNA-dependent RNA polymerases that copy mRNA is not straightforward15. For example, by nanopore DRS, we could identify Arabidopsis long non-coding antisense RNAs, such as those at the auxin efflux carriers *PIN4* and *7* (Fig. 1c). The existence of these previously unannotated antisense RNAs was supported by Illumina RNAseq of wild type Col-0 and the exosome mutant *hen2-2* (Fig. 1c), where the abundance of these 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 transcripts16. 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) (Fig. 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. We found that poly(A) tail length is negatively correlated with gene expression in Arabidopsis (Spearman rho = -0.3, *p*=9.8×10-134, 95% CIs [-0.32, -0.28]) (Supplementary Fig. S2a), consistent with other species previously analysed by short read TAILseq17.

We previously mapped Arabidopsis mRNA 3’ ends transcriptome-wide using Helicos Bio sequencing18. We compared the position of 3’ ends of nanopore DRS read alignments and Helicos Bio data genome-wide. The median genomic distance between nanopore DRS and Helicos Bio 3’ ends was nt (1 standard deviation) demonstrating close agreement between these orthogonal technologies (Fig. 2b). Likewise, the overall distribution of the 3’ end of aligned nanopore DRS reads resembles the pattern we previously reported with Helicos Bio data18. For example, 97% of nanopore DRS 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 time19 (Fig. 2c), and at mRNA encoding the histone H3K9 demethylase IBM1, which controls levels of genic DNA methylation20 (Supplementary Fig. S2b).

Since RT-dependent internal priming can result in the misinterpretation of authentic cleavage and polyadenylation sites3, we asked whether nanopore sequencing was compromised in this way. To address this issue, we examined whether the 3’ ends of nanopore DRS reads mapped to potential internal priming substrates comprised of 6 consecutive adenosines within transcribed coding sequence (according to Araport11 annotation). Of 10,116 such oligo (A)6 sequences, only 4 have read-alignments terminating within 13nt in all four datasets. Of these, 2 were not detectable after error correction with proovread (suggesting that they resulted from alignment errors) and the other 2 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 DRS data. Overall, we conclude that nanopore DRS can identify multiple authentic features of RNA 3’ end processing.

*Cap-dependent 5’ RNA detection by nanopore DRS*

Nanopore DRS reads are frequently truncated prior to annotated transcription start sites resulting in a 3’ bias of genomic alignments (Fig. 3a). Consequently, it is not possible to conclude which, if any, aligned reads correspond 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 (Supplementary Table S1). We filtered the reads for 5’ adapter RNA sequences by using the sequence alignment tool BLASTN and specific criteria (Supplementary Table S2). We then used high confidence examples of sequences that passed or failed these criteria to train a convolutional neural network to detect the 5’ adapter RNA in raw signal (Supplementary Fig. S3a-c). Hence, we improved 5’ adapter-ligated RNA detection without requiring base calling and genome alignment, and demonstrated enrichment of full-length, cap-dependent mRNA sequences (Fig. 3a, b). This procedure reduced the median 3’ bias of nanopore read alignments per gene (as measured by quartile coefficient of variation of per base coverage) from 0.55 to 0.08 (Fig. 3b).

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 clones21. We found that 41% of adapter ligated nanopore DRS reads mapped within 5nt of transcription start sites and 60% mapped to within 13nt (Fig. 3c). We also detected recently defined examples of alternative 5’ transcription start sites22 at specific Arabidopsis genes (Supplementary Fig. S3d). We therefore conclude that this approach is effective in detecting authentic mRNA 5’ ends.

Reads with adapters had, on average, 11nt more sequence at their 5’ ends that could be aligned to the genome compared to the most common 5’ position aligned for reads lacking the 5’ adapter RNA (Fig. 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. As a result, the 5’ end of RNA is not correctly sequenced. Consistent with these Arabidopsis transcriptome-wide nanopore DRS data, reads mapping to the synthetic ERCC RNA Spike-Ins and *in vitro* transcribed RNAs also lacked ~11nt of authentic 5’ sequence (Supplementary Figs. 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 passage of 5’ RNA through the pore (Fig. 3d, Supplementary Figs. S3e, f).

Despite the close agreement between nanopore DRS, Illumina RNAseq and full-length cDNA start site data at *RCA*, the annotated start site in Araport11 and AtRTD223 is quite different (Fig. 3e). This apparent overestimation of 5’UTR length is widespread in Araport11 annotation (Supplementary Fig. S3g). This observation also agrees with the assessment of capped Arabidopsis 5’ ends detected by nanoPARE sequencing24. Consequently, with appropriate modification to the current protocol, such as that which we describe here, nanopore DRS data can be used to revise Arabidopsis transcription start site annotation.

*Nanopore DRS reveals unannotated complexity of splicing events*

In single reads, nanopore sequencing revealed some of the most complex splicing combinations known in the Arabidopsis transcriptome. For example, the splicing pattern of a 12.7kb read alignment, comprised of 63 exons, agreed exactly with the *AT1G48090.4* isoform annotated in Araport11 (Fig. 1a). Mutually exclusive alternative splicing of *FLM* *(AT1G77080*)exons that mediate the thermosensitive response controlling flowering time25 was also detected (Fig. 4a). However, considering the non-corrected dataset as a whole, a combination of sequencing and alignment errors probably resulted in the mis-calling of splicing events because 58% (170,702) of the unique splice junctions detected in the combined set of replicate data were absent from Araport11 and AtRTD2 annotations, and unsupported by Illumina RNAseq (Fig. 4b, Supplementary Table S3). We applied proovread8,13 error correction with the parallel Illumina RNAseq data and re-analysed the corrected and uncorrected nanopore DRS data. The four Col-0 biological replicate nanopore DRS datasets capture 75% (102,486) and 69% 104,686) of the Araport11 and AtRTD2 splice site annotations, respectively. 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 RNAseq (Fig. 4b, Supplementary Table S3). 3,234 unique canonical splicing events in the error corrected nanopore DRS data were supported by Illumina RNAseq, but were absent from both Araport11 and AtRTD2 annotations, highlighting potential gaps in the complexity of Arabidopsis splicing annotation (Fig. 4b, Supplementary Table S3). Consistent with this, we validated 3 of these splicing events by using RT-PCR followed by cloning and sequencing (Supplementary Fig. S4a). In order to examine the features of these unannotated splices, we applied previously determined splice site position weight matrices of the flanking sequences to categorize U2 or 12 class splice sites26. Of the 3,234 novel GU/AG events found in error corrected data, supported by Illumina alignments, 74% were classified as canonical U2 or U12 splice sites, reinforcing the possibility that they are authentic (Supplementary Fig S4b).

In addition to previously unannotated splicing events, we identified unannotated combinations of previously established splice sites. For example, we identified 19 *FLM* splicing patterns which all adhered to known splice junction sites (Fig. 4a). However, 11 of these transcript isoforms were not previously annotated. In order to investigate this phenomenon transcriptome-wide, we analysed the 5’ cap-dependent nanopore DRS datasets of full-length mRNAs (Supplementary Table S1). Unique sets of co-splicing events were extracted from error-corrected reads (in order to focus on splicing, we did not consider single exon reads or 5’ and 3’ positions). 13,064 unique splicing patterns were detected that matched annotations in one or both of Araport11 and AtRTD2 (Fig. 4c). Another 8,659 unique splicing patterns were identified that were not present in either annotation (Fig. 4c, Supplementary Table S3). 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 splicing patterns produced from alternative combinations of known splice sites.

Overall, we conclude that nanopore DRS can reveal the complexity of splicing in the context of full-length mRNAs in a manner that short-read data cannot. However, accurate splice pattern detection requires error correction with, for example, high-accuracy orthogonal short read sequencing data. Accurate splice detection, even with error-free sequences, can be confounded because equivalent alternative junctions often exist27. Therefore, improved computational tools, not only for 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 regulation1,2. m6A has been mapped transcriptome-wide using approaches based on antibodies that recognize this mark6,28. However, in principle, m6A can be detected by nanopore DRS8. 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 applied nanopore DRS to 4 biological replicates of an Arabidopsis mutant defective in the function of a conserved m6A writer complex component, Virilizer (*vir-1*), and 4 biological replicates of a line expressing VIR fused to GFP that restores the activity of VIR in the *vir-1* mutant background9 (Supplementary Fig. S5a). In addition, we sequenced a parallel set of 6 biological replicates with Illumina RNAseq. We then used a G-test statistical analysis to determine if there was a differential error profile in alignments at each reference base between the mutant (defective m6A) and the complementing line. We identified 17,491 sites with greater than two-fold higher error rate (compared to the reference) in the complementing line with restored m6A (Fig. 5A). No VIR-dependent error sites mapped to either chloroplast or mitochondrial-encoded RNAs. 99.8% of differential error sites map within Araport11 annotated protein-coding genes. 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) (Fig. 5b) and preferentially found in 3’UTRs (Fig. 5C). Since approximately 5nt contribute to the observed current at a given time point in nanopore sequencing8, 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, Supplementary Fig. S5b), with a median of 2 error sites per m6A motif (95% CIs [1, 7]). These results agree with the established and conserved properties of authentic m6A sites, suggesting that differential alignment errors can be used to identify thousands of m6A modifications in nanopore DRS datasets.

In order to examine the validity of m6A sites identified by the differential error site analysis, we used an orthogonal technique to map m6A. Previous maps of Arabidopsis m6A are based on Me-RIP29,30 and limited by a resolution of around 200nt31. Therefore, to examine Arabidopsis m6A sites with a more accurate method, we used miCLIP28 analysis of 3 biological replicates of Arabidopsis Col-0. We found that, like the differential error sites uncovered in nanopore DRS analysis, the Arabidopsis miCLIP reads were enriched in 3’ UTRs, with no enrichment over stop codons (Fig. 5D, Supplementary Fig. S5c). 77% of called nanopore DRS differential error sites fall within 5nt of an miCLIP peak (Fig. 5e, f). We therefore conclude that our analysis of nanopore data can detect authentic VIR-dependent m6A sites transcriptome-wide.

*Defective m6A perturbs gene expression patterns and lengthens circadian period*

The combination of transcript processing and modification data obtained using nanopore DRS enabled us to investigate the impact of m6A on Arabidopsis gene expression. We found a global reduction in protein-coding gene expression in *vir-1* (using either nanopore DRS or Illumina RNAseq data) corresponding to transcripts that were methylated in the VIR complemented line (Fig. 6a, Supplementary Fig. S6a). These findings are consistent with the recent discovery that m6A predominantly protects Arabidopsis mRNAs from endonucleolytic cleavage30. Therefore, although the m6A writer complex comprises conserved components that target a conserved consensus sequence and distribution of m6A enriched in the last exon, it appears that this modification predominantly promotes the decay of mRNAs in human cells32, but predominantly promotes the stability of mRNAs in Arabidopsis30.

The changes in gene expression in *vir-1* were wide ranging (Supplementary Fig. S6b-d). For example, we found that the abundance of mRNAs encoding components of the interlocking transcriptional feedback loops that comprise the Arabidopsis circadian oscillator33, such as *CCA1*, was increased in *vir-1* (Supplementary Figs. 6b, c). This distinction was associated with a biological consequence because the *vir-1* mutant exhibited a lengthening in clock period (Fig. 6b, c). We detected m6A at mRNAs encoding the clock components *CCA1*, *PRR7, GI,* and *LNK1/2* in both the nanopore DRS and miCLIP data (Supplementary Fig 6b). We also detected shifts in poly(A) tail length distributions of mRNAs transcribed from genes previously shown to be under circadian control. At *CAB1* mRNAs*,* for example, we detected a periodicity in poly(A) tail length that peaked at approximately 20, 40 and 60nt (Fig. 6d). *vir-1* mutants showed a reduction in abundance of *CAB1* mRNAs with 20 and 40nt poly(A) tails, and an increase in abundance of poly(A) tails 60nt in length (Fig. 6d). Therefore, nanopore DRS may uncover the circadian control of poly(A) tail length previously reported for specific Arabidopsis genes34*.* An output of the circadian clock is the control of flowering time and we found that, not only were photoperiod pathway components differentially expressed, but so too were other flowering time genes (Supplementary Table S4). Notably, *FLC* expression was reduced by more than 40-fold compared to wild-type (Supplementary Fig. S6d). 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 is associated with disrupted RNA 3’ end 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 the software tool DEX-Seq35 for 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 in our Illumina data (Supplementary Fig. S6e). 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 DRS 3’ position profile in the *vir-1* mutant compared to the complemented line (FDR < 0.05, absolute change in position > 13nt; Fig. 6e). 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 DRS (70% detectable by miCLIP) and correspond to locations of increased cleavage downstream of m6A sites in the *vir-1* mutant (Supplementary Fig. S6f). 571 genes showed increased transcriptional readthrough beyond the 3’ end in *vir-1* (Fig. 6e). 73% of these loci also contained nanopore-mapped m6A sites (78% by miCLIP). Impacts of altered 3’ processing can be complex, and have the potential to change the relative abundance of transcripts with different coding potential processed from the same gene. For example, we detected increased readthrough of an intronic cleavage site in the Symplekin-like gene *TANG1* (*AT1G27595*) (Supplementary Fig. S6g); and increased readthrough and cryptic splicing at *ALG3* (*AT2G47760*)that alsoresults in chimeric RNA formation with the downstream gene *GH3.9* (*AT2G4G7750*) (Fig. 6f). The existence of the *ALG3-GH3.9* chimeric RNAs was supported by Illumina RNAseq (Fig. 6f) and confirmed by RT-PCR, cloning and sequencing (Supplementary Fig. S6h). 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. Chimeric RNAs were recently detected in mutants affecting other components of the Arabidopsis m6A writer complex, MTA and FIP3736. 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 VIR or the writer complex in 3’ end formation or an indirect effect of the writer complex on factors required for 3’ processing. m6A independent roles for the human m6A methyltransferases METTL337 and METTL1637 and METTL1637 have been found previously, and a role for the writer complex in controlling Arabidopsis RNA processing independent of m6A cannot be overlooked37. In mammals, recognition of the canonical poly(A) signal AAUAAA involves direct binding by CPSF3038,39. Notably, alternative polyadenylation controls the expression of an Arabidopsis CPSF30 isoform that encodes a YTH domain with the potential to bind and read m6A40. A recent study has indicated that this YTH domain-containing isoform is required to supress chimera formation36. Consequently, in plants, m6A may also contribute to the recognition of specific RNA 3’ ends.

**Concluding Remarks**

We show that nanopore DRS has the potential to refine multiple features of Arabidopsis genome annotation and incorporate the clearest map to date of m6A locations. This is despite the sequence of the genome being available since 200041. Modern agriculture is dominated by a handful of intensely researched crops42, but 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 improvement43. Based on our experience with Arabidopsis, we anticipate that the combination of nanopore DRS and other sequencing approaches will improve genome annotation. Consistent with this, we recently applied nanopore DRS to refine the annotation of the African orphan crop water yam (*Dioscorea alata*). Indeed, we are moving into an era where thousands of genome sequences are available and programmes like the Earth BioGenome Project aim to sequence all eukaryotic life on Earth44. Genome sequences present only part of the puzzle, annotating what they encode will be essential for us to fully utilise this information.

**Materials and Methods**

**Plants**

*Plant material and growth conditions*

The wild-type *Arabidopsis thaliana* accession Col-0 was obtained from Nottingham Arabidopsis Stock Centre. The *vir-1* and VIR complemented (*VIR::GFP-VIR*) lines were provided by K. Ruzicka, Brno45. The *hen2-2* (Gabi\_774HO7) mutant was provided by D. Gagliardi, Strasbourg. The seeds were sown on MS10 media plates, stratified at 4°C for two days, germinated in a controlled environment at 22°C under 16h light/8h dark conditions, and harvested 14 days after transfer to 22°C.

*Clock phenotype analysis*

Clock phenotype experiments were performed as previously described by measuring delay fluorescence as a circadian output46. Briefly, plants were grown in 12‐h light/12‐h dark cycles at 22oC and 80 μmol m−2 sec−1 light for 9 days. Next, delayed fluorescence measurements were recorded every hour for 6 days at constant temperature (22oC) and under constant light (20 µmol m−2 sec−1  red light and 20 µmol m−2 sec−1 blue light mix). FFT non-linear Least Squares fitting to estimate period length was conducted using Biodare47.

**RNA**

*RNA isolation*

Total RNA was isolated using RNeasy® Plant Mini kit (QIAGEN) and treated with TURBO™ DNase (Thermo Fisher Scientific). The total RNA concentration was measured using a Qubit 1.0 Fluorometer and Qubit RNA BR Assay Kit (Thermo Fisher Scientific), and the quality and integrity was assessed using the NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific) and Agilent 2200 TapeStation System (Agilent).

*mGFP in vitro transcription*

mGFP coding sequence was amplified using CloneAmp HiFi PCR Premix (Clontech) and a forward primer containing T7 promoter sequence (Merck) (Table S5). PCR product was purified using GeneJET Gel Extraction (Thermo Fisher Scientific) and DNA Cleanup Micro Kit (Thermo Fisher Scientific) as per manufacturer’s instructions. mGFP was *in vitro* transcribed using mMESSAGE mMACHINE™ T7 ULTRA Transcription Kit (Thermo Fisher Scientific) with and without the addition of a cap analog (Anti-Reverse Cap Analog [ARCA]) according to manufacturer’s instructions. mGFP transcripts were treated with TURBOTM DNase, polyA-tailed using *Escherichia coli* Poly(A) Polymerase (E-PAP) and ATP (Thermo Fisher Scientific) and recovered using MEGAclear™ Kit (Thermo Fisher Scientific) as per manufacturer’s instructions. mGFP mRNAs quantity was assessed using a Qubit 1.0 Fluorometer (as described above), and the quality and integrity was checked using the NanoDrop™ 2000 spectrophotometer and agarose-gel electrophoresis. ARCA-capped and non-capped *in vitro* mGFP mRNAs were used in the library preparation for DRS using nanopores.

*RT-PCR and RT-qPCR*

Total RNA was reverse transcribed using SuperScript™ III polymerase or SuperScript™ IV VILO™ Master Mix (Thermo Fisher Scientific) according to the manufacturer’s protocol. For RT-PCR, the reactions were performed using the Advantage 2 Polymerase Mix (Clontech) using primers (Merck) listed in Table S5. Next, the PCR products were purified from the gel using GeneJET Gel Extraction and DNA Cleanup Micro Kit (Thermo Fisher Scientific), cloned into the pGEM T-Easy vector (Promega) (according to the manufacturer’s instruction) and sequenced. For RT-qPCR, the reactions were carried out using SYBR Green I (Qiagen) mix with primers (Merck) listed in Table S5 following manufacturer’s instructions.

**Illumina RNA sequencing**

*Preparation of libraries for Illumina RNA sequencing*

Illumina RNA sequencing libraries from purified mRNA were prepared and sequenced by the Centre for Genomic Research at University of Liverpool using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England Biolabs® Inc.). Paired-end sequencing with read length of 150bp was carried out on Illumina HiSeq 4000. Illumina RNA libraries from ribosome-depleted RNA were prepared using the Illumina TruSeq Stranded Total RNA with Ribo-Zero Plant kit (Illumina®). Paired-end sequencing with read length of 100bp was carried out on Illumina HiSeq2000 at the Genomic Sequencing Unit of the University of Dundee. ERCC RNA Spike-In mixes (Thermo Fisher Scientific)11,48 were included in each of the libraries using concentrations advised by the manufacturer.

*Mapping of Illumina RNA sequencing data*

Reads were mapped to the TAIR10 reference genome with Araport11 reference annotation using STAR version 2.6.149, a maximum multimapping rate of 5, a minimum splice junction overhang of 8 nt (3 nt for junctions in the Araport11 reference), a maximum of 5 mismatches per read, and intron length boundaries of 60 - 10,000 nt.

*Differential gene expression analysis using Illumina RNA sequencing data*

Transcript level counts for Illumina RNA sequencing reads were estimated by pseudoalignment with salmon version 0.11.250. Counts were aggregated to gene level using tximport51 and differential expression gene expression analyses for *vir-1* mutantvs wild-type and *vir-1* mutant vs VIR complemented line were conducted in R version 3.5 using edgeR version 3.24.352.

*Differentially expressed region analysis using Illumina RNA sequencing data*

Mapped read pairs originating from forward and reverse strands were separated and coverage tracks were generated using samtools version 1.953. Coverage tracks were then used as input for DERfinder version 1.16.154. Mapped read pairs originating from forward and reverse strands were separated and coverage tracks were generated using samtools version 1.953. Coverage tracks were then used as input for DERfinder version 1.16.154. Expressed regions were identified using a minimum coverage of 10 reads, and differential expression between *vir-1* and VIR complemented was conducted using the analyseChr method using 50 permutations.

*Differential exon usage analysis using Illumina RNA sequencing data*

Annotated gene models from Araport11 were divided into transcript chunks (i.e. contiguous regions within which each base is present in the same set of transcript models). Read counts for each chunk were generated using bedtools version 2.27.155 intersect in count mode. Chunk counts were then processed using DEXseq version 1.28.335 to identify differentially expressed chunks between *vir-1* and VIR complemented lines. Chunks were annotated as five prime variation if they included a start site of any transcript, and three prime variation if they contained a termination site. Chunks representing overhangs from alternative donor or acceptor sites were also separately classified. Internal exons were subclassified into cassette exons if they could be wholly contained within any intron.

**Nanopore direct RNA sequencing**

*Preparation of libraries for Direct RNA sequencing using nanopores*

mRNA was isolated from approximately 75 μg of total RNA using the Dynabeads®mRNA purification kit (Thermo Fisher Scientific) following the manufacturer’s instructions. The quality and quantity of mRNA was assessed using the NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific). Nanopore libraries were prepared from 1 μg of poly(A)+ RNA combined with 1 μl of undiluted ERCC RNA Spike-In mix (Thermo Fisher Scientific) using the nanopore DRS Kit (SQK-RNA001 Ltd.) according to manufacturer’s instructions. Quickly poly(T) adapter was ligated to the mRNA using T4 DNA ligase (New England Biolabs® Inc.) in the Quick Ligase reaction buffer (New England Biolabs® Inc.) for 15 min at room temperature. The first strand cDNA was synthesized by SuperScript III Reverse Transcriptase (Thermo Fisher Scientific) using the oligo(dT) adapter. The RNA-cDNA hybrid was purified using Agencourt RNAClean XP magnetic beads (Beckman Coulter). The sequencing adapter was ligated to the mRNA using T4 DNA ligase (New England Biolabs® Inc.) in the Quick Ligase reaction buffer (New England Biolabs® Inc.) for 15 min at room temperature followed by a second purification using Agencourt beads as described above. Libraries were loaded on R9.4 SpotON Flow Cells (Oxford Nanopore Technologies Ltd.) and sequenced on a 48-hour runtime.

To incorporate cap-dependent ligation of a biotinylated 5’ adapter RNA, the following modifications were introduced into the library preparation protocol. 4 µg of mRNA was de-phosphorylated by Calf Intestinal Alkaline Phosphatase (Thermo Fisher Scientific) and 5’ cap was removed by Cap-Clip™ Acid Pyrophosphatase (Cambio) according to the manufacturer’s instructions. Next, 5’ RNA oligo biotinylated at the 5’ end (Integrated DNA Technologies) was ligated to dephosphorylated, de-capped mRNA using T4 RNA ligase I (New England Biolabs® Inc.) and mRNA was purified using Dynabeads™ MyOne™ Streptavidin C1 beads (Thermo Fisher Scientific) according to the manufacturer’s instructions. mRNA quality and quantity were assessed using the NanoDropTM 2000 spectrophotometer and used for nanopore DRS library preparation as described above.

*Processing of nanopore DRS data*

Reads were basecalled with guppy version 2.3.1 (Oxford Nanopore Technologies) using default RNA parameters and converted from RNA to DNA fastq using seqkit version 0.10.056. Reads were aligned to the TAIR10 *Arabidopsis thaliana* genome57 and ERCC RNA spike-in sequences11,48 using minimap2 version 2.858 in spliced mapping mode using a kmer size of 14 and a max intron size of 10,000. SAM and BAM file manipulations were performed using samtools version 1.953. Reads were basecalled with guppy version 2.3.1 (Oxford Nanopore Technologies) using default RNA parameters and converted from RNA to DNA fastq using seqkit version 0.10.056. Reads were aligned to the TAIR10 *Arabidopsis thaliana* genome57 and ERCC RNA spike-in sequences11,48 using minimap2 version 2.858 in spliced mapping mode using a kmer size of 14 and a max intron size of 10,000. SAM and BAM file manipulations were performed using samtools version 1.953.

Proovread version 2.14.113was used to correct errors in the nanopore DRS read. Each nanopore DRS replicate was split into 200 chunks for parallel processing. Each chunk was corrected using 4 samples of Illumina poly(A) RNAseq data, selected randomly from the 36 Illumina files (6 biological replicates sequenced across 6 lanes). Illumina reads 1 and 2 were merged into fragments using FLASh version 1.2.1159. Unjoined pairs were discarded. Error correction with proovread was conducted in sampling free mode using a minimum nanopore read length of 50 nt. Corrected reads were then mapped to the reference using minimap2 as described above. All figures showing gene tracks with Nanopore DRS reads use error corrected reads.

*Error profile analysis using nanopore DRS data*

Error rate analysis of aligned reads was conducted on ERCC RNA Spike-in mix controls using pysam version 0.15.260 for BAM file parsing. Matches, mismatches, insertions and deletions to the reference were extracted from the cs tag (a more informative version of CIGAR string, output by minimap2) and normalised by aligned length of the read. Reference bases and mismatch bases per position were also recorded and used to assess the frequency of each substitution and indel type by reference base.

*Over-splitting analysis of nanopore DRS data*

To identify read pairs which result from over-splitting of signal originating from a single RNA molecule, the sequencing summary files produced by guppy were parsed for sequencing time and channel ID. These were used to identify pairs of consecutively sequenced reads. Genomic locations of reads were parsed from minimap2 mappings and consecutively sequenced reads which aligned adjacently with a genomic distance of between –10 nt and 1000 nt between them were identified. Samples sequenced before or during May 2018 had very low levels of over-splitting (between 0.01 and 0.05% of reads) compared to those sequenced in September 2018 onwards (between 0.8 and 1.5% of reads).

*Analysis of the potential for internal priming in nanopore DRS data*

To determine if internal priming caused by the RT step can occur in nanopore data, the location of oligo(A) hexamers within Arabidopsis CDS regions was determined and reads which terminated within a 20 nt window of each hexamer were counted. Of 10,116 CDS oligo(A) runs, 160 (1.58%) had at least one supporting read in one Col-0 nanopore dataset. Of these, 137 were only supported by one replicate, and only 4 were supported by all four biological replicates. 66 (41%) occurred in terminal exons, suggesting they may be genuine APA sites.

*Poly (A) length estimation using nanopore DRS data*

Poly (A) tail length estimations were produced using Nanopolish version 0.11.016 and added as tags to bam files using pysam version 0.15.260. Per gene length distributions were then produced using the Araport11 annotation, and genes with significant changes in length distribution in the *vir-1* mutant compared to the complemented line were identified using a Kolmogorov-Smirnoff test*.* *p*-values were adjusted for multiple testing using Benjamini-Hochberg correction.

*3’ end analyses of nanopore and Helicos DRS reads*

Helicos DRS data were prepared as described in18 and61. Positions with three or more supporting reads were considered to be peaks of nanopore or Helicos 3’ ends. The distance between each nanopore peak and the nearest Helicos peak was then determined. 37% of nanopore peaks occurred at the same position as a Helicos peak, and the standard deviation in distance was 12.5 nt. To determine the percentage of nanopore DRS 3’ ends mapping within annotated genic features, transcripts were first flattened into a single record per gene. Exonic annotation was given priority over intronic or intergenic annotation and CDS annotation over UTR. Reads were assigned to genes if they overlapped them by >20% of their aligned length and the annotated feature type of the 3’ end position was determined. Counts were generated both for all reads and for unique positions per gene.

*Isoform collapsing of nanopore DRS data*

Error corrected full length alignments were collapsed into clusters of reads with identical sets of introns. These clusters were then subdivided by 3’ end location by using a Gaussian kernel with sigma of 100 to find local minima between read ends, which were used as cut points to separate clusters. The read with the longest aligned length in each cluster was used as the representative in the figure.

*Splicing analysis of nanopore DRS and Illumina RNAseq data*

Splice junction locations, their flanking sequences, and the read counts supporting them were extracted from Illumina RNA sequencing, nanopore DRS, and nanopore error-corrected DRS reads using pysam version 0.1460, and from Araport1162 and AtRTD223 reference annotations. Splice junctions at the same position, but on opposite strands, were counted independently. Junctions were classified by their most likely snRNP machinery using biopython version 1.7163, with position weight matrices calculated by 26. Position weight matrices were scored against the sequence –3 nt to +10 nt of the donor site, and –14 nt to +3 nt of the acceptor site. PWM matrix scores greater than zero indicate a match to the motif, whilst scores around zero, or negative scores, indicate background frequencies or deviation from the motif. Positive scores were normalised into the range 50-100 as in26. Junctions with U12 donor scores >75 and U12 acceptor scores >65 were classified as U12 junctions, whilst junctions with U2 donor and acceptor scores >60 were classified as U2, as in23. Junctions were further categorised as canonical or non-canonical based on the presence or absence of the GT/AG intron border sequences. For isoform analysis, linked splices from the same read were extracted from full length nanopore error corrected reads and counted to create unique sets of splice junctions. Intronless reads were not counted. UpSet plots were generated in Python 3.6 using the package upsetplot.

*Validation of novel splice sites*

To validate novel splice junctions detected in nanopore DRS, 5 splice sites out of 20 most highly expressed splice sites were selected for further validation and 3 out of 5 selected splice sites were amplified in RT-PCR followed by Sanger sequencing (described above).

*5’ adapter detection analyses using nanopore DRS data*

To produce positive and negative examples of 5’ adapter containing sequences, 5’ soft-clipped regions were extracted from aligned reads for the Col-0 replicate 1 datasets (with and without adapter ligation), using pysam60. These soft-clipped sequences were then searched for the presence of the GeneRacer™ adapter sequence using blastn version 2.7.164. Two rules were initially applied to filter blastn results: a match of 10 nt or more to the 44 nt adapter, and an E value of less than 100. Reads from the adapter containing dataset that failed one or both criteria were used as negative training examples. A final rule requiring the match to the adapter sequence to occur directly adjacent to the aligned read was also applied. Reads from the adapter containing dataset that passed all three rules were used as the positive training set. When comparing the ratio of positive to negative examples between datasets containing the adapter and those generated from the same tissue without, we found that these three rules gave a signal to noise ratio of >5000 (Table S2).

72,083 positive and 123,739 negative training examples from Col-0 tissue replicate 1 were collected to train the neural network. We then estimated the amount of raw signal from the 5’ end of the squiggle which was required on average to capture the 5’ adapter. To do this, we used Nanopolish eventalign version 0.11.065 to identify the interval in the raw read which corresponded to the mRNA alignment to the reference in the adapter positive examples. Since the adapter can be identified immediately adjacent to the alignment in sequence space for these reads, the signal after the event alignment should correspond to signal originating from the adapter. The median length of these signals was 1441 points, and 96% of the signals were less than 3000. Therefore, we used a window size of 3,000 to make predictions.

The model was trained in python 3.6 using Keras version 2.2.4 with Tensorflow version 1.10.0 backend66,67. A ResNet style architecture was used68, composed of 8 residual blocks containing two convolutional layers of kernel size 5 and a shortcut convolution with kernel size 1. Down-sampling using maximum pooling layers with a stride of 2 was used between each residual block. A penultimate densely connected layer of size 16 was used, with training dropout of 0.5. Input signals were standardized by median absolute deviation scaling across the whole read before the final 3,000 points were taken, and the negative samples were augmented by addition of random internal signals from reads and pure gaussian, multi-gaussian, and perlin noise signals69. The whole dataset was also augmented on the fly during training by the addition of gaussian noise with standard deviation of 0.1. Models were trained for a maximum of 100 epochs (batch size of 128, 100 batches per epoch, positive and negative examples sampled in a 1:1 ratio) using the RMSprop optimiser with an initial learning rate of 0.001, which was reduced by a factor of 10 after three epochs with no reduction in validation loss. Early stopping was used after five epochs with no reduction in validation loss. Model performance was evaluated using five-fold cross validation and by testing on independently generated datasets from Col-0 replicate 2, produced with and without the adapter ligation protocol (Supplementary Figs S3b, c)66,67.

For evaluation of the reduction in 3’ bias of adapter ligated datasets, we used the Araport11 exon annotations to produce per base coverage for each gene in the Col-0 replicate 2 dataset. Coverage was generated separately for reads predicted to contain adapters and those that did not. Leading zeros at the very 5’ and 3’ of genes were assumed to be caused by mis-annotation of UTRs and so were trimmed. The quartile coefficient of variation (interquartile range / median) was then used as a robust measure of variation in coverage across each gene. To validate the 5’ ends of adapter ligated reads with orthogonal data, full length cDNA clone sequences were downloaded from RIKEN RAFL. These were mapped with minimap258 in spliced mode and distance from each nanopore alignment 5’ end to the nearest RIKEN RAFL alignment21. 5’ end was calculated using bedtools55. The amount of 5’ end sequence which is rescued when 5’ adapters are used was estimated by identifying the largest peak in 5’ end locations per gene in the absence of adapter, and then measured how this peak shifted using reads predicted to contain adapters.

*Differential error site analysis using nanopore DRS data*

To detect sites of Virilizer-dependent m6A RNA modifications, we developed scripts to test changes in per base error profiles of aligned reads. Pileup columns for each position with >10 reads coverage were generated using pysam60 and reads in each column were categorised as either A, C, G, T or indel. The relative proportions of each category were counted. Counts from replicates of the same condition were aggregated and a 2x5 contingency table was produced for each base comparing ­*vir-1* and VIR complemented lines. A G-test was performed to identify bases with significantly altered error profiles. For bases with *p-*value less than 0.05, G-tests for homogeneity between replicates of the same condition were then performed. Bases where the sum of the G statistic for homogeneity tests was greater than the G statistic for the *vir-1* and complemented line comparison were filtered. Multiple testing correction was carried out using the Benjamini-Hochberg method, and an FDR threshold of 0.05 was used. The log2 fold change in mismatch to match ratio (compared to the reference base) between VIR complemented and *vir-1* was calculated using the Haldane correction for zero counts. Bases which had a log fold change > 1 were considered to have a reduced error rate in the *vir-1* mutant.

To identify motifs enriched at sites with reduced error rate, reduced error rate sites were increased in size by 5nt in each direction and overlapping sites were merged using bedtools version 2.27.155. Sequences corresponding to these sites were extracted from the TAIR10 reference and overrepresented motifs were detected in the sequences using MEME version 5.0.270, run in zero or one occurrence mode with a motif size range of 5-7 and a minimum requirement of 100 sequence matches. The presence of these motifs at error sites was then detected using FIMO version 5.0.271. A relaxed FDR threshold of 0.1 was used with FIMO to capture more degenerate motifs matching the m6A consensus.

*Differential gene expression analysis using nanopore DRS data*

Gene level counts were produced for each nanopore DRS replicate using featureCounts version 1.6.372 in long-read mode with strand specific counting. Differential expression analysis between VIR complemented and *vir-1* lines was then performed in R version 3.5 using edgeR version 3.24.352.

*Identification of alternative 3’ end positions and chimeric RNA using nanopore DRS data*

Genes with differential 3’ end usage were identified by producing 3’ profiles of reads which overlapped with each annotated gene locus by >20%. These profiles were then compared between VIR complemented and *vir-1* using a Kolmogorov-Smirmov test to identify changes. Multiple testing correction was performed using the Benjamini-Hochberg method. To approximately identify the direction and distance of the change, the normalised single base level histograms of the complemented line profile was subtracted from that of the mutant profile, and the minimum and maximum points in the difference profile were found. These represent the site of most reduced and increased relative usage, respectively. Results were filtered for FDR < 0.05 and absolute change of site > 13 nt (the measured error range of nanopore DRS 3’ end alignment).

To identify genes with significant increases in chimeras in the *vir-1* mutant, we used the Araport11 annotation62 to identify reads which overlapped with multiple adjacent gene loci (chimeric reads) and those that originated from a single locus (non-chimeric reads). To reduce false positives caused by reads mapping across tandem duplicated loci, reads mapping to genes annotated in the PTGBase73 were filtered out. Chimeric reads were considered to originate from the most upstream gene with which they overlapped. We pooled reads from replicates of a condition and used 50 bootstrapped 75% samples without replacement to estimate the ratio of chimeric to non-chimeric reads at each gene in that condition. The Haldane correction for zero counts was applied. The distributions of complemented line and *vir-1* chimeric to non-chimeric ratios were tested using a Kolmogorov-Smirnoff test to detect loci with altered chimera production. All possible pairwise combinations of VIRc and *vir-1* bootstraps were then compared to produce a distribution of estimates of the change in chimeric to non-chimeric ratio in the *vir-1* mutant. Loci which had more than one chimeric read in *vir-1* and demonstrated at least a two-fold increase in chimeric read ratio in more than 50% of bootstrap comparisons, plus were significantly changed at FDR < 0.05, were considered to be sites of increased chimeric RNA formation in *vir-1* mutant.

**miCLIP**

*Preparation of miCLIP libraries*

Total RNA for miCLIP was isolated from 7.5 mg of 14-day old Arabidopsis Col-0 seedlings as previously described74. mRNA was isolated from ~1 mg of total RNA using oligo(dT) and streptavidin paramagnetic beads (PolyATtract® mRNA Isolation Systems, Promega) as per manufacturer’s instructions. miCLIP was carried out using 15 µg of mRNA as described in75 using antibody against N6-methyladenosine (#202 003 Synaptic Systems) with minor modifications. No-antibody controls were processed throughout the experiment. 4-12% Bis-Tris gel electrophoresis was run at 180 V for 50 minutes, while transfer of the RNA-antibody complexes was carried out at 30 V for 60 minutes. Following the transfer, membranes were exposed to the Medical X-Ray Film Blue (Agfa) at -80°C overnight. Reverse transcription was carried out using barcoded RT primers - RT41, RT48, RT49 and RT50 (Integrated DNA Technologies) (Table S5). After reverse transcription, one cDNA fraction corresponding to 70-200 nt was purified from the 6% TBE-urea gel (Thermo Fisher Scientific). After the final PCR step, all libraries were pooled together and purified using Agencourt Ampure XP magnetic beads (Beckman Coulter) and eluted in nuclease-free water. Paired-end sequencing with read length of 100 bp was carried out on Illumina MiSeq v2 at Edinburgh Genomics of the University of Edinburgh. Input sample libraries were prepared using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England Biolabs® Inc.) and sequenced on Illumina HiSeq2000 at the Tayside Centre for Genomics Analysis at the University of Dundee with pair-end read length of 75bp.

*Processing for miCLIP sequencing data*

miCLIP data were assessed for quality using FastQC version 0.11.876 and MultiQC version 1.777. Only the forward read was used for analysis as the miCLIP site is located at the 5’ position of the forward read. 3’ adapter and poly(A) sequences were trimmed using cutadapt version 1.1878 and Unique Molecular Identifiers were extracted from the 5’ of the reads using UMI-tools version 0.5.579. Immunoprecipitation and no antibody controls were demultiplexed and multiplexing barcodes were trimmed using seqkit version 0.10.056. Reads were mapped to the TAIR10 reference genome with Araport11 reference annotation62, using STAR version 2.6.149, a maximum multimapping rate of 5, a minimum splice junction overhang of 8 nt (3 nt for junctions in the Araport11 reference), a maximum of 5 mismatches per read, and intron length boundaries of 60 - 10,000 nt. SAM and BAM file manipulations were performed using samtools version 1.953. Removal of PCR duplicates was then performed using UMI-tools in directional model79. miCLIP 5’ coverage and matched input 5’ coverage tracks were generated using bedtools version 2.27.155 and these were used to call miCLIP peaks at single nucleotide resolution with Piranha version 1.2.180 with relaxed *p*-value thresholds of 0.5. Reproducible peaks across pairwise combinations of the three replicates were identified by Irreproducible Discovery Rate (IDR) analysis using the python package idr version 2.0.3 using an IDR threshold of 0.0581. The final set of peaks was identified by pooling the three replicates and re-analysing using Piranha, then ranking the peaks by FDR and selecting the top N peaks, where N was the smallest number of reproducible peaks discovered by pairwise comparisons of the three replicates. This yielded 141,198 unique nucleotide level miCLIP peaks.

**m6A LC-MS analysis**

The m6A content analysis using Liquid Chromatography - Mass Spectroscopy (LC-MS) was performed as described in82. Chromatography was carried out by the FingerPrints Proteomics facility at the University of Dundee.

**Code Availability**

All scripts, pipelines, and notebooks used for this study are available on GitHub at https://github.com/bartongroup/NS\_Nanopore\_paper

**Data Availability**

Sequencing datasets described in this study have been deposited at the European Nucleotide Archive (ENA) under the study accession number PRJEB32782.

**Acknowledgements**

This work was funded by BBSRC grants BB/J00247X/1, BB/M004155/1, BB/M010066/1, the University of Dundee Global Challenges Research Fund and the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 799300. The m6A LC-MS analysis was carried out by Abdel Atrih of the FingerPrints Proteomics Facility at the University of Dundee, which is supported by a Wellcome Trust Technology Platform award [097945/B/11/Z]. We thank Kasper Rasmussen and Csaba Hornyik for comments on the manuscript.

**Author contributions**

GGS, NJS and AVS conceived the study; KK, AVS and KM designed and performed the experiments; MTP and NJS undertook data analysis; PDG undertook the clock experiments, supervised by AH; GGS and GJB supervised the study; GGS, MTP and AVS wrote the paper. All authors read and commented on the text.

**Competing interests**

The authors declare no conflicts of interests.

**References**

1 Meyer, K. D. & Jaffrey, S. R. Rethinking m6A readers, writers, and erasers. *Annual Review of Cell and Developmental Biology* **33**, 319-342, doi:10.1146/annurev-cellbio-100616-060758 (2017).

2 Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. *Cell* **169**, 1187-1200, doi:10.1016/j.cell.2017.05.045 (2017).

3 Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of *Caenorhabditis elegans* 3 ' UTRs. *Nature* **469**, 97-101, doi:10.1038/nature09616 (2011).

4 Mourão, K. *et al.* Detection and mitigation of spurious antisense expression with RoSA. *F1000Research* **8**, doi:10.12688/f1000research.18952.1 (2019).

5 Houseley, J. & Tollervey, D. Apparent non-canonical *trans*-splicing is generated by reverse transcriptase *in vitro*. *Plos One* **5**, doi:10.1371/journal.pone.0012271 (2010).

6 Helm, M. & Motorin, Y. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nat Rev Genet* **18**, 275-291, doi:10.1038/nrg.2016.169 (2017).

7 Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**, 1177-1184, doi:10.1038/Nmeth.2714 (2013).

8 Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods* **15**, 201, doi:10.1038/nmeth.4577 (2018).

9 Ruzicka, K. *et al.* Identification of factors required for m6A mRNA methylation in Arabidopsis reveals a role for the conserved E3 ubiquitin ligase HAKAI. *New Phytol* **215**, 157-172, doi:10.1111/nph.14586 (2017).

10 Lange, H. *et al.* The RNA helicases AtMTR4 and HEN2 target specific subsets of nuclear transcripts for degradation by the nuclear exosome in *Arabidopsis thaliana*. *PLoS Genet* **10**, doi:10.1371/journal.pgen.1004564 (2014).

11 Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **21**, 1543-1551, doi:10.1101/gr.121095.111 (2011).

12 Reid, L. H. *et al.* Proposed methods for testing and selecting the ERCC external RNA controls. *Bmc Genomics* **6**, doi:10.1186/1471-2164-6-150 (2005).

13 Hackl, T., Hedrich, R., Schultz, J. & Forster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-3011, doi:10.1093/bioinformatics/btu392 (2014).

14 Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*, doi:10.1093/bioinformatics/bty841 (2018).

15 Matsui, A. *et al.* Novel stress-inducible antisense RNAs of protein-coding loci are synthesized by RNA-dependent RNA polymerase. *Plant Physiology* **175**, 457-472, doi:10.1104/pp.17.00787 (2017).

16 Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*, 459529, doi:10.1101/459529 (2018).

17 Lima, S. A. *et al.* Short poly(A) tails are a conserved feature of highly expressed genes. *Nat Struct Mol Biol* **24**, 1057–1063, doi:10.1038/nsmb.3499 (2017).

18 Duc, C., Sherstnev, A., Cole, C., Barton, G. J. & Simpson, G. G. Transcription termination and chimeric RNA formation controlled by *Arabidopsis thaliana* FPA. *PLOS Genetics* **9**, e1003867, doi:10.1371/journal.pgen.1003867 (2013).

19 Hornyik, C., Terzi, L. C. & Simpson, G. G. The spen family protein FPA controls alternative cleavage and polyadenylation of RNA. *Developmental Cell* **18**, 203-213, doi:10.1016/j.devcel.2009.12.009.

20 Rigal, M., Kevei, Z., Pelissier, T. & Mathieu, O. DNA methylation in an intron of the *IBM1* histone demethylase gene stabilizes chromatin modification patterns. *EMBO J* **31**, 2981-2993, doi:10.1038/emboj.2012.141 (2012).

21 Seki, M. *et al.* Functional annotation of a full-length Arabidopsis cDNA collection. *Science* **296**, 141-145, doi:DOI 10.1126/science.1071006 (2002).

22 Ushijima, T. *et al.* Light controls protein localization through phytochrome-mediated alternative promoter selection. *Cell* **171**, 1316-1325, doi:10.1016/j.cell.2017.10.018 (2017).

23 Zhang, R. *et al.* A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Res* **45**, 5061-5073, doi:10.1093/nar/gkx267 (2017).

24 Schon, M. A., Kellner, M. J., Plotnikova, A., Hofmann, F. & Nodine, M. D. NanoPARE: parallel analysis of RNA 5' ends from low-input RNA. *Genome Res* **28**, 1931-1942, doi:10.1101/gr.239202.118 (2018).

25 Pose, D. *et al.* Temperature-dependent regulation of flowering by antagonistic FLM variants. *Nature* **503**, 414-417, doi:10.1038/nature12633 (2013).

26 Sheth, N. *et al.* Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34**, 3955-3967, doi:10.1093/nar/gkl556 (2006).

27 Dehghannasiri, R., Szabo, L. & Salzman, J. Ambiguous splice sites distinguish circRNA and linear splicing in the human genome. *Bioinformatics*, doi:10.1093/bioinformatics/bty785 (2018).

28 Linder, B. *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* **12**, 767-772, doi:10.1038/nmeth.3453 (2015).

29 Shen, L. S. *et al.* *N*-6-methyladenosine RNA modification regulates shoot stem cell fate in Arabidopsis. *Dev Cell* **38**, 186-200, doi:10.1016/j.devcel.2016.06.008 (2016).

30 Anderson, S. J. *et al.* N6-methyladenosine inhibits local ribonucleolytic cleavage to stabilize mRNAs in Arabidopsis. *Cell Rep* **25**, 1146-1157, doi:10.1016/j.celrep.2018.10.020 (2018).

31 Ke, S. *et al.* A majority of m6A residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev* **29**, 2037-2053, doi:10.1101/gad.269415.115 (2015).

32 Wang, X. *et al.* *N*-6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117-120, doi:10.1038/nature12730 (2014).

33 McClung, C. R. The genetics of plant clocks. *Adv Genet* **74**, doi:10.1016/b978-0-12-387690-4.00004-0 (2011).

34 Lidder, P., Gutierrez, R. A., Salome, P. A., McClung, C. R. & Green, P. J. Circadian control of messenger RNA stability. Association with a sequence-specific messenger RNA decay pathway. *Plant Physiol* **138**, 2374-2385, doi:10.1104/pp.105.060368 (2005).

35 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Research* **22**, 2008-2017, doi:10.1101/gr.133744.111 (2012).

36 Pontier, D. *et al.* The m6A pathway protects the transcriptome integrity by restricting RNA chimera formation in plants. *Life Sci Alliance* **2**, doi:10.26508/lsa.201900393 (2019).

37 Pendleton, K. E. *et al.* The U6 snRNA m6A methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell* **169**, 824-835 e814, doi:10.1016/j.cell.2017.05.003 (2017).

38 Chan, S. L. *et al.* CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3 ' processing. *Genes & Development* **28**, 2370-2380, doi:10.1101/gad.250993.114 (2014).

39 Schonemann, L. *et al.* Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes & Development* **28**, 2381-2393, doi:10.1101/gad.250985.114 (2014).

40 Stevens, A. T., Howe, D. K. & Hunt, A. G. Characterization of mRNA polyadenylation in the apicomplexa. *Plos One* **13**, doi:10.1371/journal.pone.0203317 (2018).

41 Kaul, S. *et al.* Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).

42 United Nations, D. o. E. a. S. A., Population Division. in *Working Paper No. ESA/P/WP/248* (2017).

43 Chang, Y. *et al.* The draft genomes of five agriculturally important African orphan crops. *Gigascience* **8**, doi:10.1093/gigascience/giy152 (2018).

44 Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *P Natl Acad Sci USA* **115**, 4325-4333, doi:10.1073/pnas.1720115115 (2018).

45 Růžička, K. *et al.* Identification of factors required for m6A mRNA methylation in Arabidopsis reveals a role for the conserved E3 ubiquitin ligase HAKAI. *New Phytologist* **215**, 157-172, doi:10.1111/nph.14586 (2017).

46 Gould, P. D. *et al.* Delayed fluorescence as a universal tool for the measurement of circadian rhythms in higher plants. *The Plant Journal* **58**, 893-901, doi:10.1111/j.1365-313X.2009.03819.x (2009).

47 Moore, A., Zielinski, T. & Millar, A. J. in *Plant Circadian Networks: Methods and Protocols* (ed Dorothee Staiger) 13-44 (Springer New York, 2014).

48 Lee, H., Pine, P. S., McDaniel, J., Salit, M. & Oliver, B. External RNA controls consortium beta version update. *Journal of genomics* **4**, 19-22, doi:10.7150/jgen.16082 (2016).

49 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

50 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417, doi:10.1038/nmeth.4197 (2017).

51 Soneson, C., Love, M. & Robinson, M. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]. *F1000Research* **4**, doi:10.12688/f1000research.7563.2 (2016).

52 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

53 Wysoker, A. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

54 Collado-Torres, L. *et al.* Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Res* **45**, e9, doi:10.1093/nar/gkw852 (2017).

55 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

56 Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE* **11**, e0163962, doi:10.1371/journal.pone.0163962 (2016).

57 The Arabidopsis Genome, I. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).

58 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

59 Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

60 Heger, A., Belgrad, T. G., Goodson, M., & Jacobs, K. pysam: Python interface for the SAM/BAM sequence alignment and mapping format. (2014).

61 Schurch, N. J. *et al.* Improved Annotation of 3′ Untranslated Regions and Complex Loci by Combination of Strand-Specific Direct RNA Sequencing, RNA-Seq and ESTs. *PLOS ONE* **9**, e94270, doi:10.1371/journal.pone.0094270 (2014).

62 Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal* **89**, 789-804, doi:10.1111/tpj.13415 (2017).

63 Dalke, A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).

64 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421-421, doi:10.1186/1471-2105-10-421 (2009).

65 Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods* **12**, 733, doi:10.1038/nmeth.3444 (2015).

66 Chollet, F. *Keras*, <<https://github.com/fchollet/keras>> (2018).

67 Abadi, M. *et al.* TensorFlow: large-scale machine learning on heterogeneous distributed systems. (2016).

68 Kaiming He, X. Z., Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. (2015).

69 Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput Biol* **14**, e1006583, doi:10.1371/journal.pcbi.1006583 (2018).

70 Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202-W208, doi:10.1093/nar/gkp335 (2009).

71 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

72 Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids Res* **41**, e108-e108, doi:10.1093/nar/gkt214 (2013).

73 Yu, J. *et al.* PTGBase: an integrated database to study tandem duplicated genes in plants. *Database* **2015**, doi:10.1093/database/bav017 (2015).

74 Quesada, V., Macknight, R., Dean, C. & Simpson, G. G. Autoregulation of *FCA* pre-mRNA processing controls Arabidopsis flowering time. *The EMBO Journal* **22**, 3142-3152, doi:10.1093/emboj/cdg305 (2003).

75 Grozhik, A. V., Linder, B., Olarerin-George, A. O. & Jaffrey, S. R. Mapping m6A at individual-nucleotide resolution using crosslinking and immunoprecipitation (miCLIP). *Methods in molecular biology* **1562**, 55-78, doi:10.1007/978-1-4939-6807-7\_5 (2017).

76 Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).

77 Ewels, P., Magnusson, M., Käller, M. & Lundin, S. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

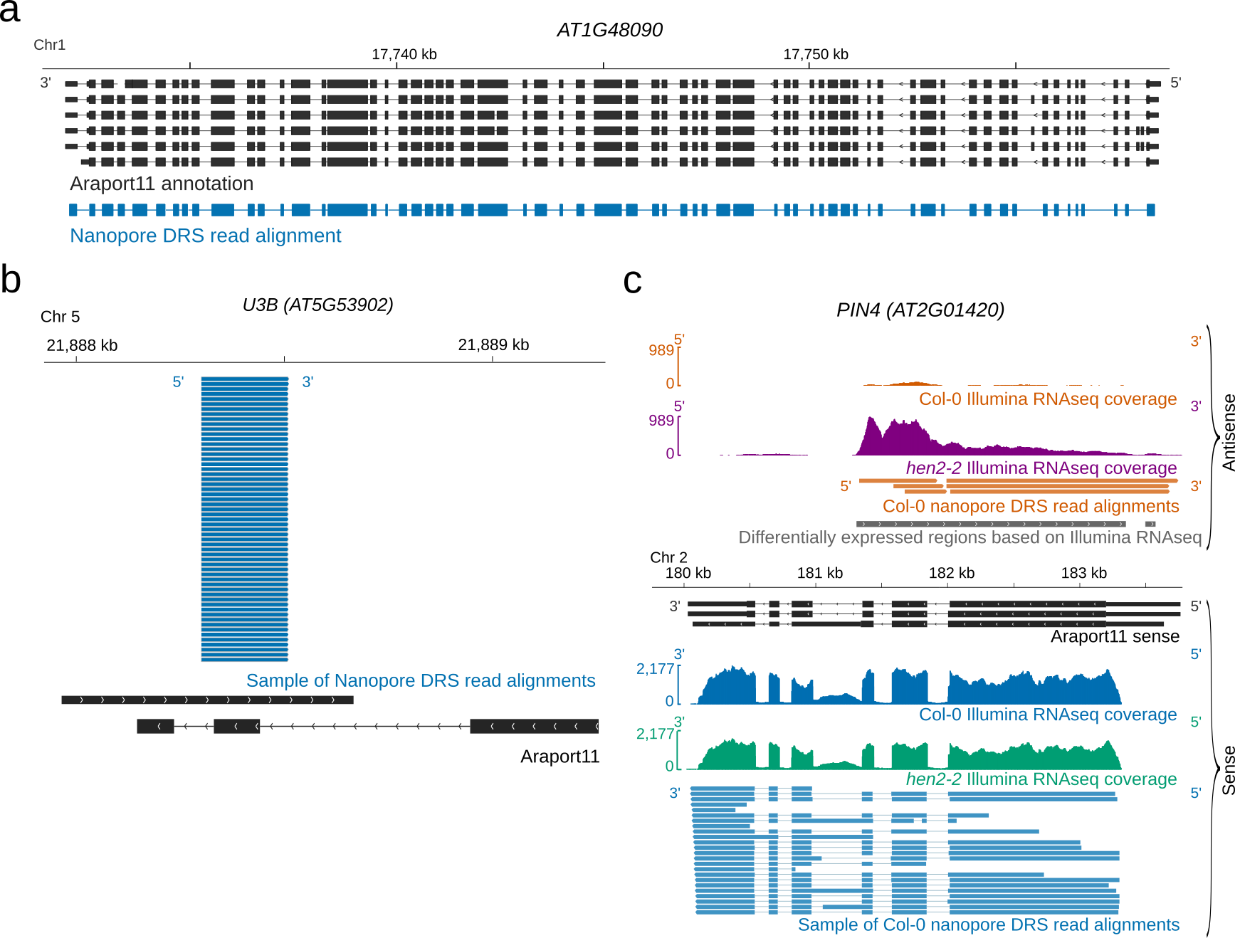
78 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **17**, 3, doi:10.14806/ej.17.1.200 (2011).

79 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).

80 Uren, P. J. *et al.* Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**, 3013-3020, doi:10.1093/bioinformatics/bts569 (2012).

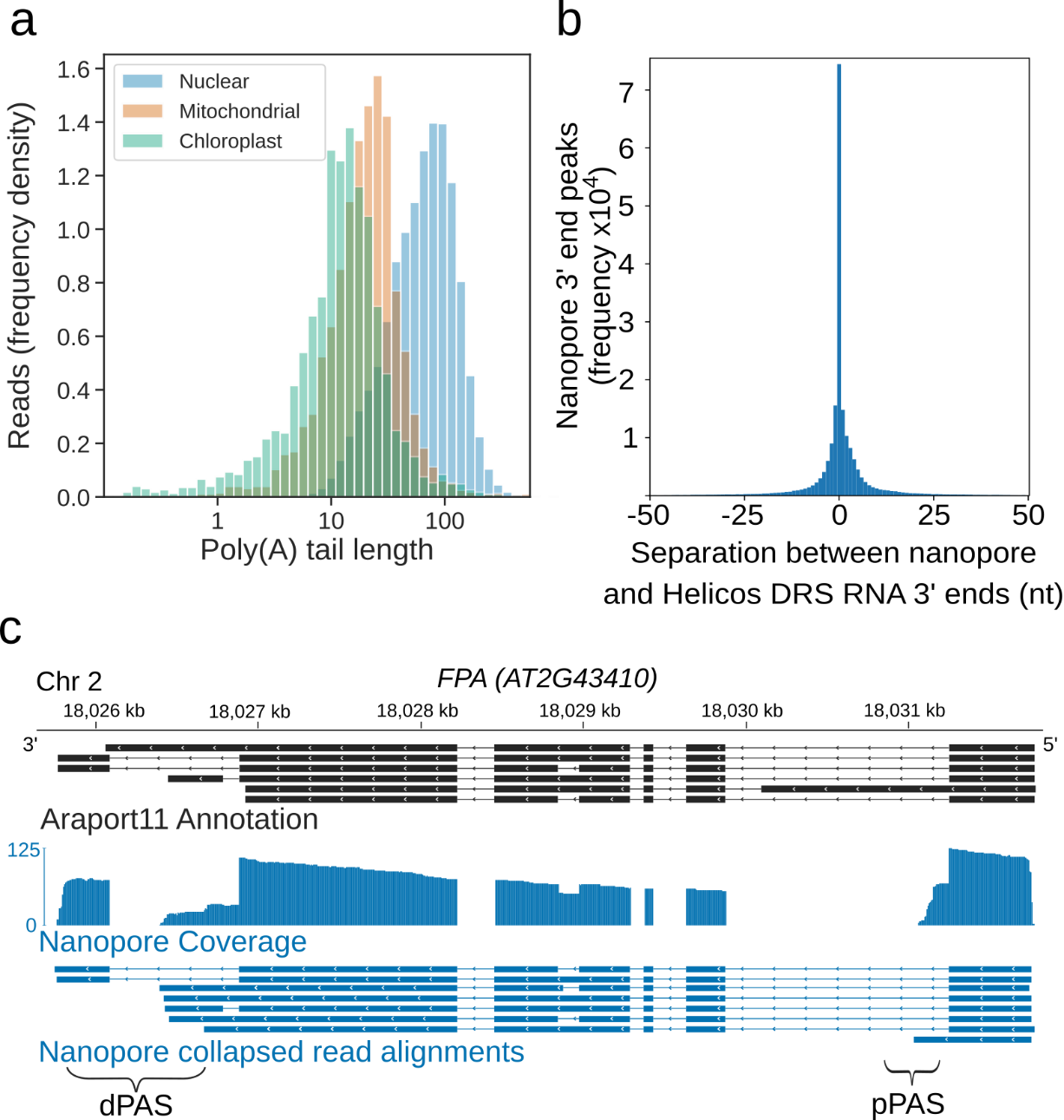
81 Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752-1779, doi:10.1214/11-AOAS466 (2011).

82 Huang, H. *et al.* Recognition of RNA *N*6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nature Cell Biology* **20**, 285-295, doi:10.1038/s41556-018-0045-z (2018).



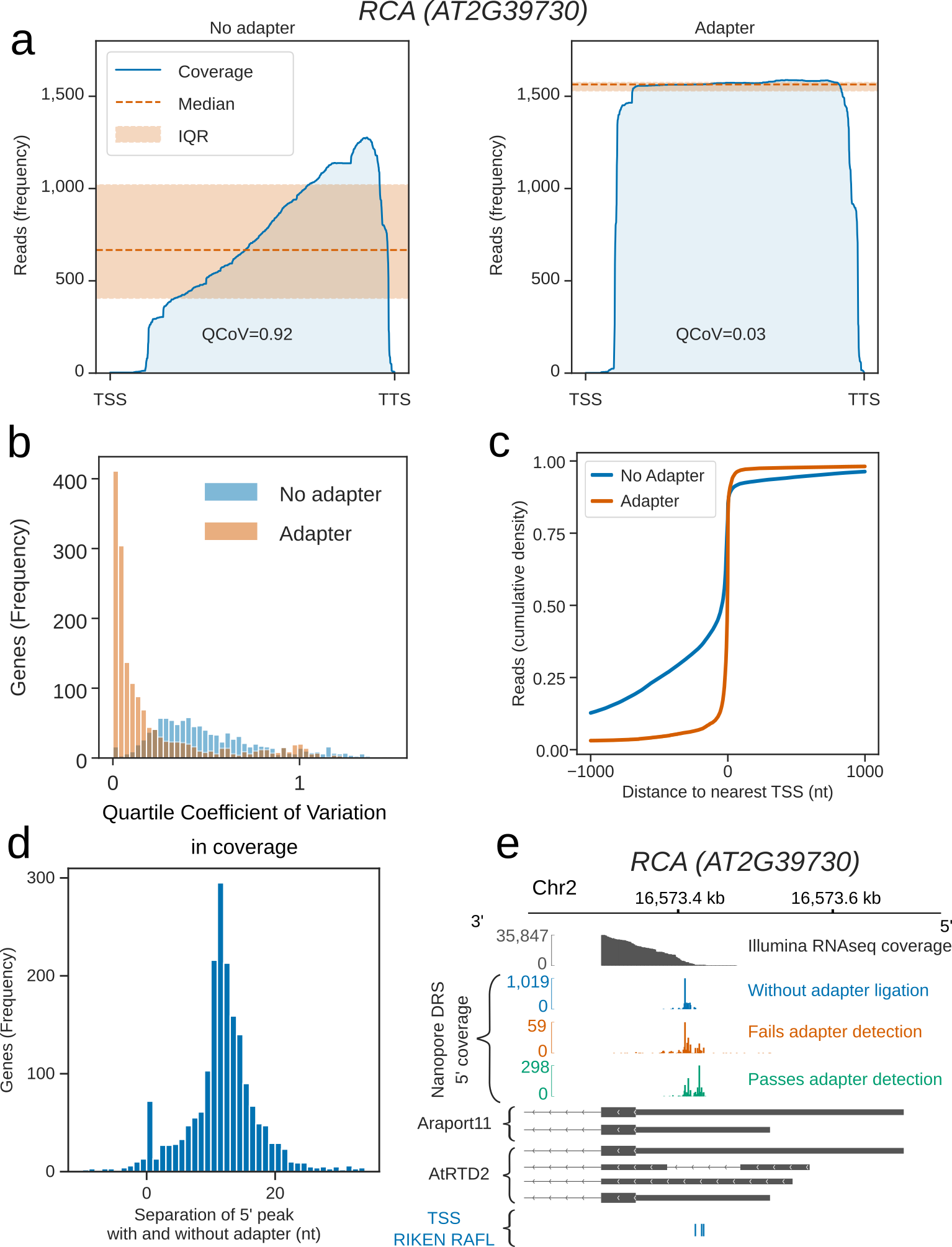
**Figure 1. Diverse Arabidopsis RNAs are detected by nanopore DRS**

1. Nanopore DRS 12.7kb read alignment at *AT1G48090* comprising 63 exons. Araport11 annotation is shown in black, nanopore DRS read alignment – in blue.
2. Nanopore DRS read alignments at the snoRNA gene *U3B*. Nanopore DRS read alignments are shown in blue, Araport11 annotation – in black.
3. *PIN4* long non-coding antisense RNAs detected using nanopore DRS. Col-0 sense Illumina RNAseq coverage and nanopore sense read alignments are shown in blue; Col-0 antisense Illumina RNAseq coverage and nanopore antisense read alignments – 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 are shown in black, and antisense differentially expressed regions detected with DERfinder are shown in grey.



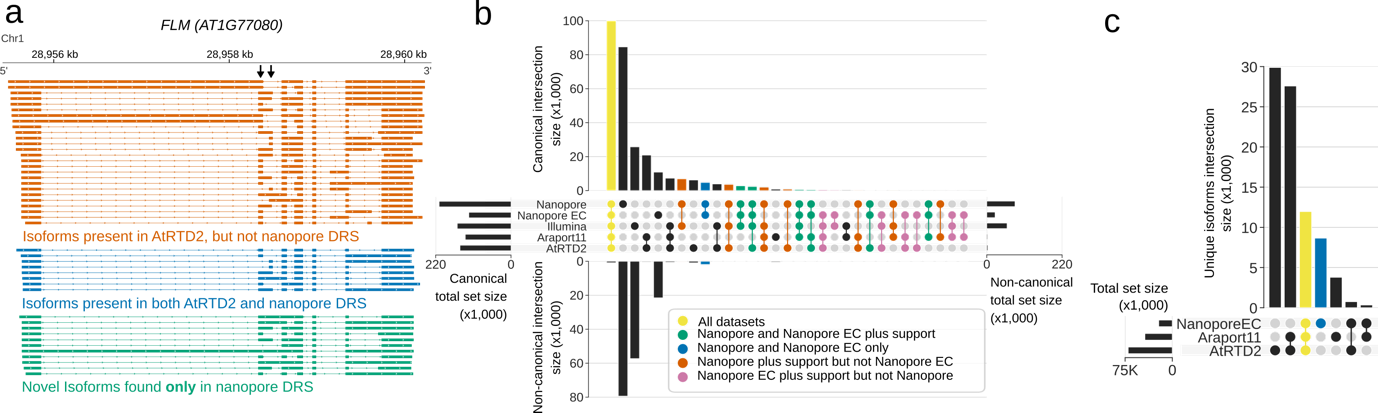
**Fig. 2. Nanopore DRS reveals poly(A) tail length and maps 3’ cleavage and polyadenylation sites.**

1. Normalised histogram of poly(A) tail length of RNAs encoded by different genomes. The poly(A) tail lengths of nuclear (blue), mitochondrial (orange) and chloroplast (in green) genome-derived RNAs are shown.
2. Separation between the RNA 3’ end positions in nanopore DRS read alignments and the nearest polyadenylation sites identified in Helicos Bio data.
3. Nanopore DRS identified 3’ polyadenylation sites in RNAs transcribed from *FPA (AT2G43410)*. Blue track shows coverage of nanopore DRS read alignments, and collapsed read alignments representing putative transcript annotations detected by nanopore DRS. Isoforms found in Araport11 annotation are shown in black and read alignments from nanopore DRS are in blue. pPAS is the proximal polyadenylation site; dPAS – distal polyadenylation sites.



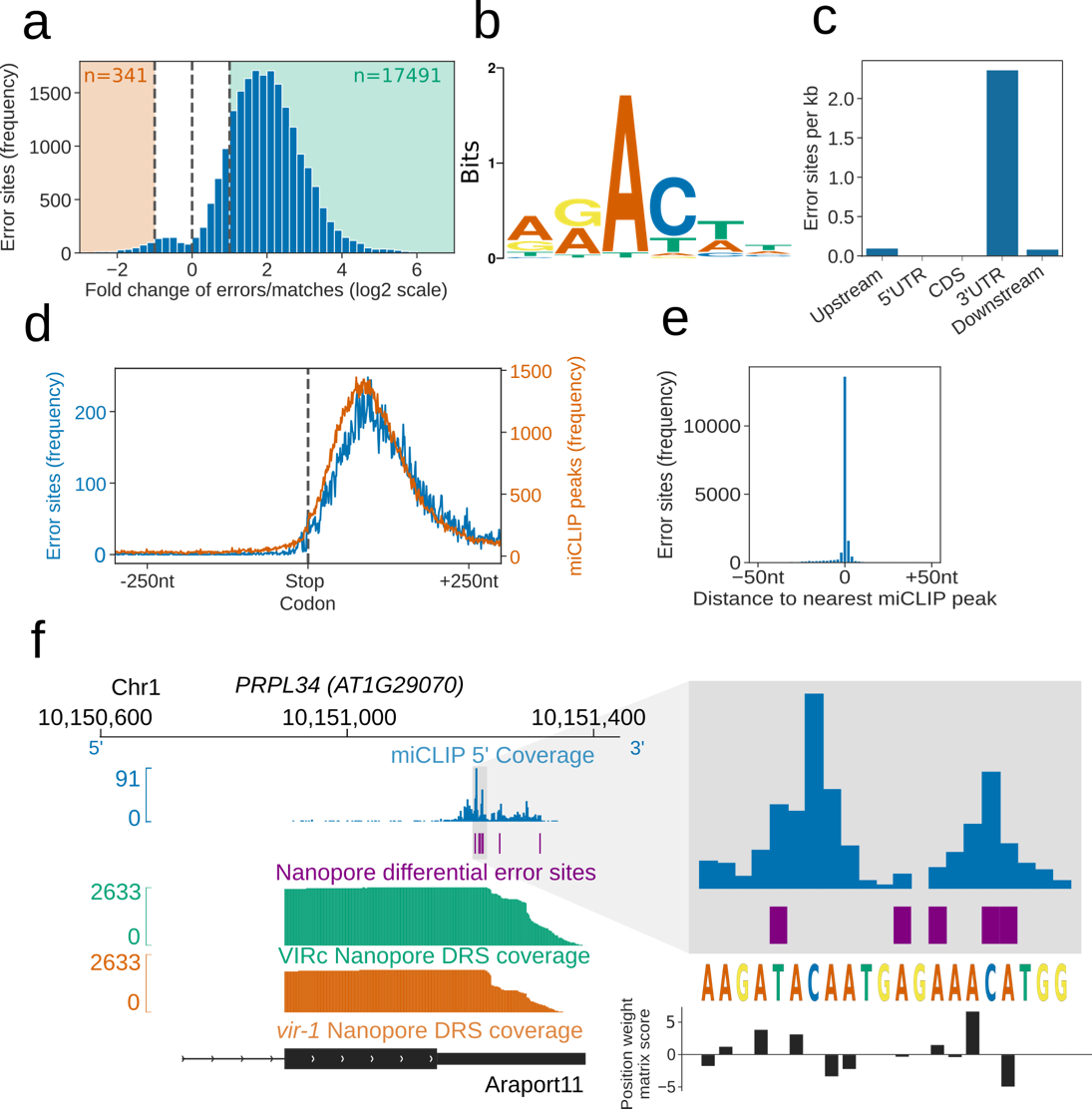
**Fig. 3. Cap-dependent ligation of an adapter enables detection of authentic RNA 5’ ends.**

1. 5’ adapter RNA ligation reduces 3’ bias in nanopore DRS data at *RCA* (*AT2G39730*). Blue line shows exonic read coverage at RCA for reads without adapter – left, and with adapter – right. Orange line shows the median coverage and orange shaded area shows inter-quartile range (IQR). Change in 3’ bias can be measured using the IQR / median (quartile coefficient of variation). 5’ adapter ligation reduces 3’ bias at RCA from 0.92 to 0.03.
2. 5’ adapter RNA ligation reduces 3’ bias in nanopore DRS data. Histogram shows quartile coefficient of variation in per base coverage for reads with 5’ adapter RNA – in orange, compared to reads without 5’ adapter RNA – in blue.
3. 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 RNA (orange), compared to reads without 5’ adapter RNA (blue).
4. Cap-dependent adapter ligation allowed resolution of an 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.
5. For *RCA (AT2G39730),* the 5’ end identified using cap-dependent 5’ adapter RNA ligation protocol was consistent with Illumina RNAseq and full-length cDNA start site data but differed from the 5’ ends 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 read alignments generated using the cap-dependent ligation protocol with and without 5’ adapter RNA is 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.



**Fig. 4. Nanopore DRS reveals complexity of alternative splicing.**

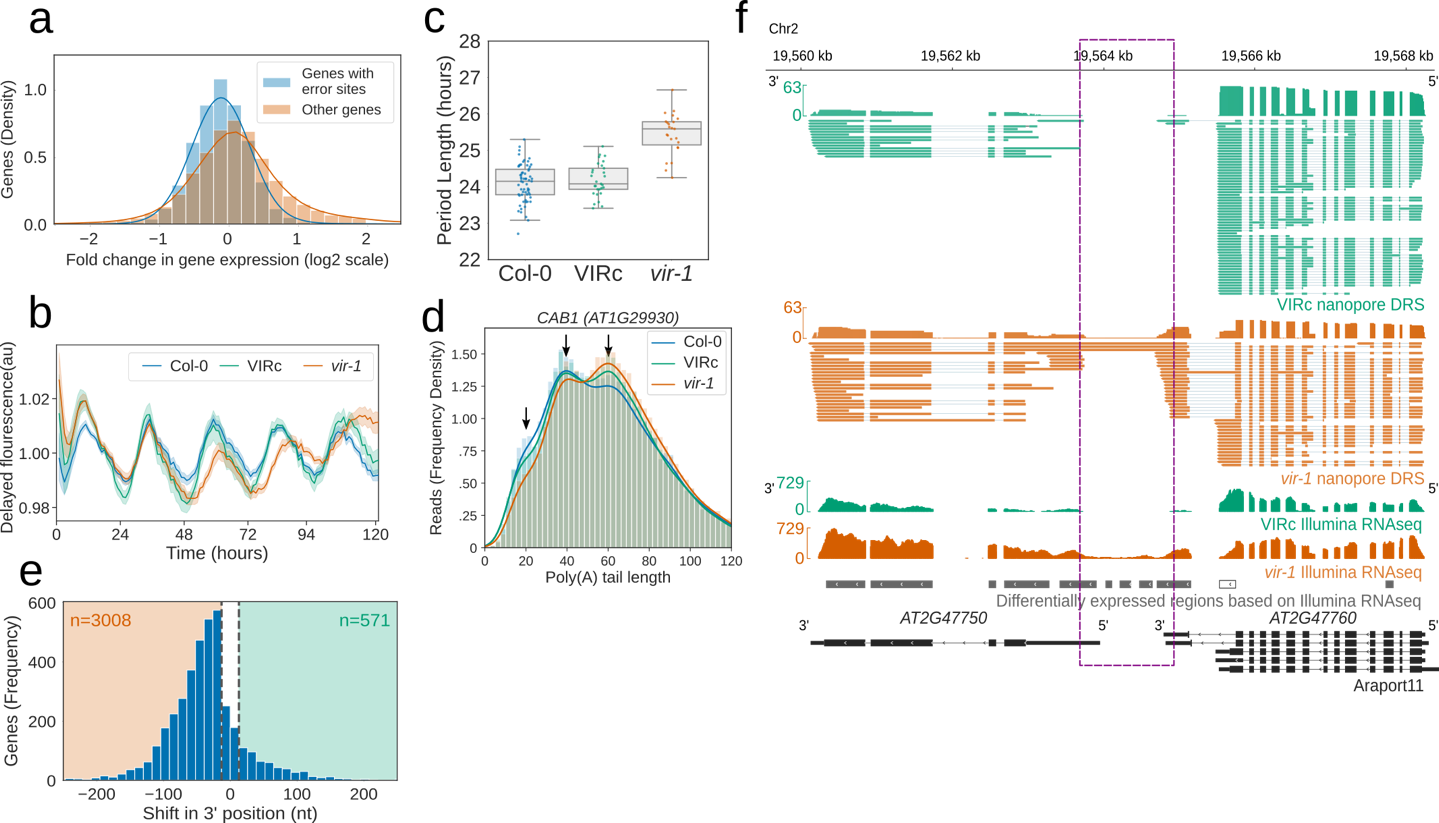
1. Nanopore DRS identified the mutually exclusive alternative splicing of *FLOWERING LOCUS M* (*FLM, AT1G77080)*. Mutually exclusive exons are indicated by black arrows. In addition, novels isoforms were identified: isoforms present in AtRTD2 annotation, but not identified using nanopore DRS are shown in orange, isoforms common to both AtRTD2 and nanopore DRS are in blue, novel isoforms identified in nanopore 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 (see Table S3 for the exact values). GU/AG splicing events are shown on the top and non-GU/AG 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 DRS 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 (see Table S3 for the exact values). Unique splicing patterns nanopore DRS and both reference annotations are shown in yellow. Novel isoforms are shown in blue.



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

1. Loss of VIR function is associated with reduced error rate in nanopore DRS. 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 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 5’ and 3’ of the annotated transcription termination sites (TTSs), 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 DRS 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 5nt of an miCLIP peak.

Nanopore DRS 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 DRS differential error sites – in purple; nanopore DRS coverage of VIR complemented line (VIRc) – in green; nanopore DRS coverage of *vir-1* mutant – in orange; RNA isoform from Araport11 annotation – in black. Zoomed region shows miCLIP coverage in blue and error sites in purple labelled with score using m6A consensus position weight matrix (Figure 5B) in black. More positive scores denote higher likelihood of a match to the consensus sequence.



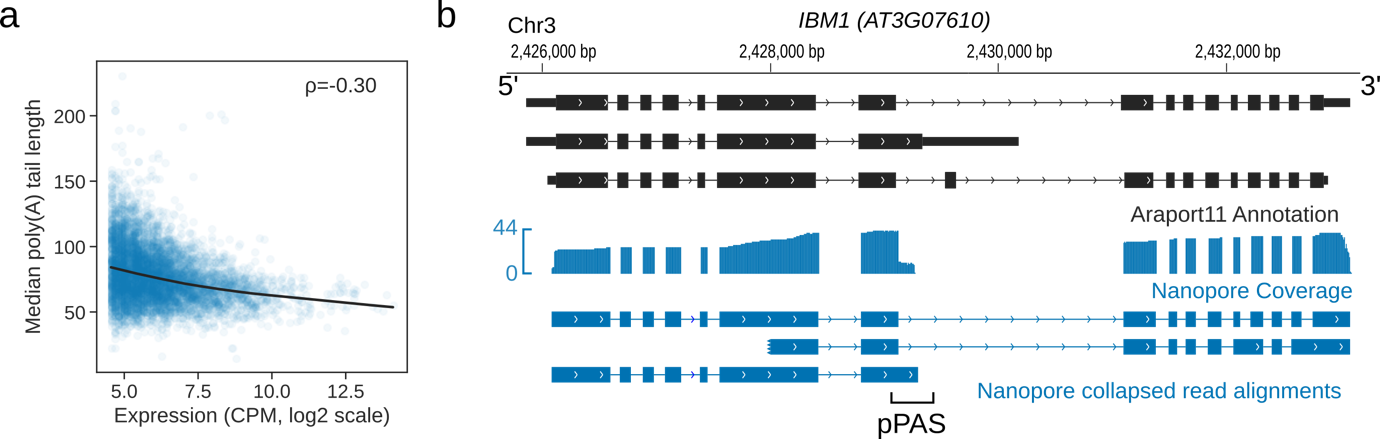
**Fig. 6. Reduction in m6A RNA modification leads to disruption of circadian clock and generation of chimeric RNAs.**

1. Genes with differential error rate sites have lower detectable RNA levels. 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. Mean delayed fluorescence measurements for Col-0 are in blue, VIR complemented line – in green and *vir-1* mutant – in orange. Shaded areas show bootstrapped 95% confidence intervals for mean.
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) tail length alterations in *vir-1* mutant. Histogram shows poly(A) tail length distribution of *CAB1 (AT1G29930)* in Col-0 (blue), VIR complemented line (green) and *vir-1* mutant (orange). Arrows show phased poly(A) length peaks at approximately 20 nt, 40 nt and 60 nt. *vir-1* distribution is significantly different from both Col-0 and VIRc using KS test at 5% level.
5. Global change in 3’ end usage in the *vir-1* mutant compared to the complemented line. Histogram shows the distance in nucleotides between the most reduced and increased 3’ positions, for genes where the 3’ end profile in *vir-1* is altered (detected using KS test, FDR < 0.05). A threshold of 13 nt was used to detect changes in 3’ position.
6. Readthough events and chimeric RNAs are detected in *vir-1*. 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 between *vir-1* and VIRc detected using Illumina RNAseq data with DERfinder are shown in grey (for upregulated regions) or white (for downregulated regions). Intergenic readthrough regions are highlighted using a purple dashed rectangle.



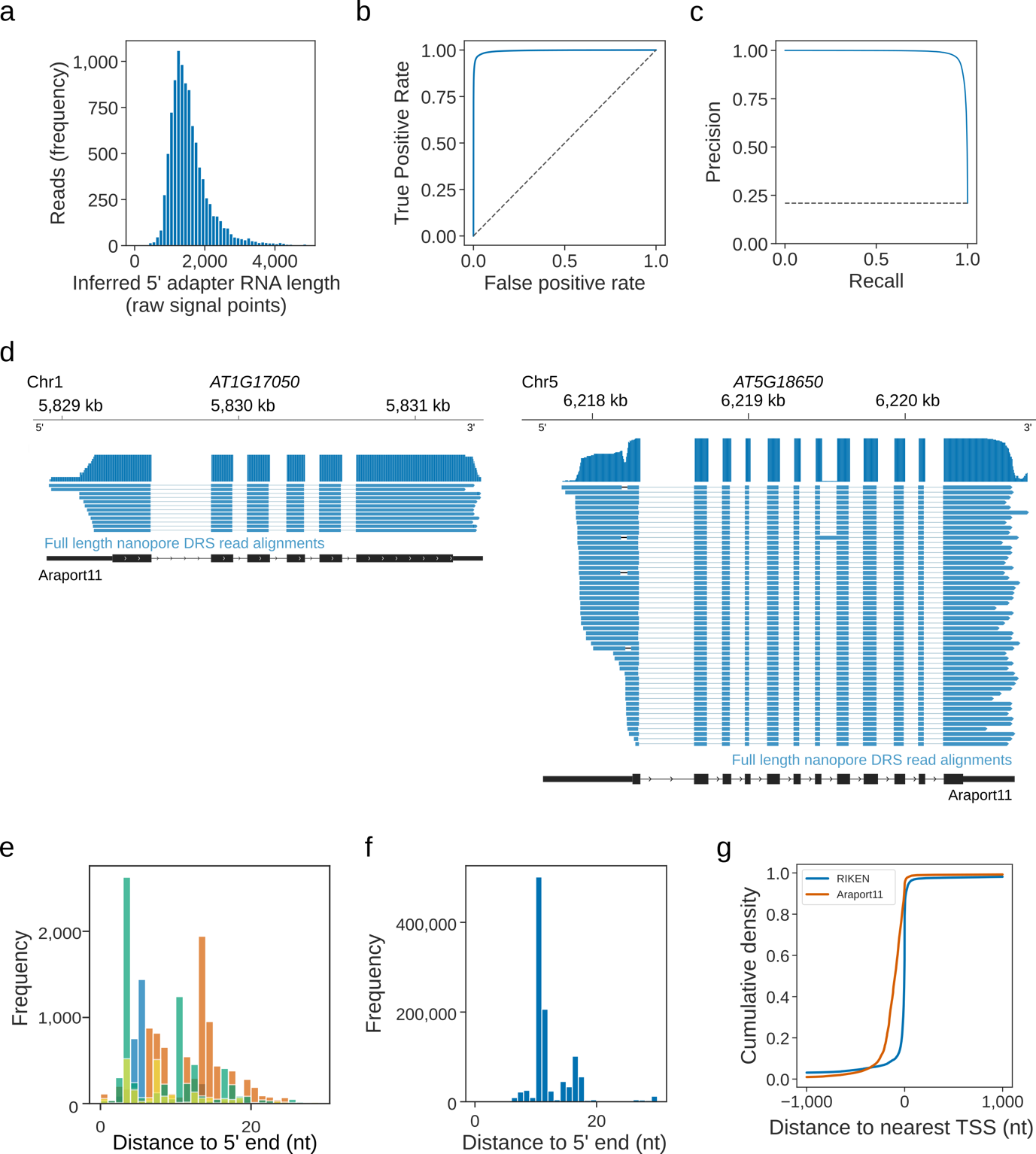
**Figure S1. Properties of nanopore DRS sequencing data.**

1. Nanopore DRS identified 12.8kb transcript generated from *AT1G67120* gene that includes 58 exons. An RNA isoform present in the *Araport11* annotation is shown in black and an RNA isoform identified using nanopore DRS – in blue.
2. Synthetic ERCC RNA Spike-In mixes are detected in a quantitative manner. Absolute concentrations of spike-ins are plotted against counts per million reads in log10 scale. ERCC RNA Spike-In mix 1 is shown in blue and mix 2 – in orange.
3. Overview of the sequencing and alignment characteristics of nanopore DRS data for ERCC RNA Spike-Ins. Distribution of the length fraction of each sequenced read that aligns to the ERCC RNA Spike-In reference is shown on the left; distribution of fraction of identity that matches between the sequence of the read and the ERCC RNA Spike-In reference for the aligned portion of each read – in the centre; distributions of the occurrence of insertions (black), substitutions (orange) and deletions (blue) as a proportion of the number of aligned bases in each read – on the right.
4. Substitution preference for each nucleotide (A, U, G, C, left-to-right). When substituted, guanidine is replaced with adenosine in more than 63% of its substitutions, while cytosine is replaced by uridine 73% of the time. Conversely uridine is rarely replaced with guanidine (12%) and adenine is rarely substituted with cytosine (16%).
5. Nucleotide representation within the ERCC RNA Spike-In reference sequences (black dots) compared with the nucleotide representation within four categories from the nanopore DRS reads. Identity matches between the sequence of the read and the ERCC RNA Spike-In reference (green crosses), insertions (blue pentagons), deletions (yellow stars) and substitutions (purple diamonds). Guanidine (G) is under-represented, and uridine (U) is over-represented, for all three categories of error (insertion, deletion and substitution) relative to the reference nucleotide distribution. Cytosine (C) is over-represented in the set of deletions and substitutions. Adenine (A) is over-represented for insertions and deletions and, under-represented in the set of substitutions.
6. Signals originating from the *RH3* transcripts are susceptible to systematic over-splitting around exons 7-9 (highlighted using a purple dashed box), resulting in reads with apparently novel 5’ or 3’ positions. This appears to only occur at high frequency in datasets collected after May 2018 (Table S1) and may be the result of an update to the MinKNOW software.
7. *PIN7* long non-coding antisense RNAs detected using nanopore DRS. Col-0 sense Illumina RNAseq coverage and nanopore sense read alignments are shown in blue; Col-0 antisense Illumina RNAseq coverage and nanopore antisense read alignments – 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 are shown in black, and antisense differentially expressed regions detected with DERfinder are shown in grey.



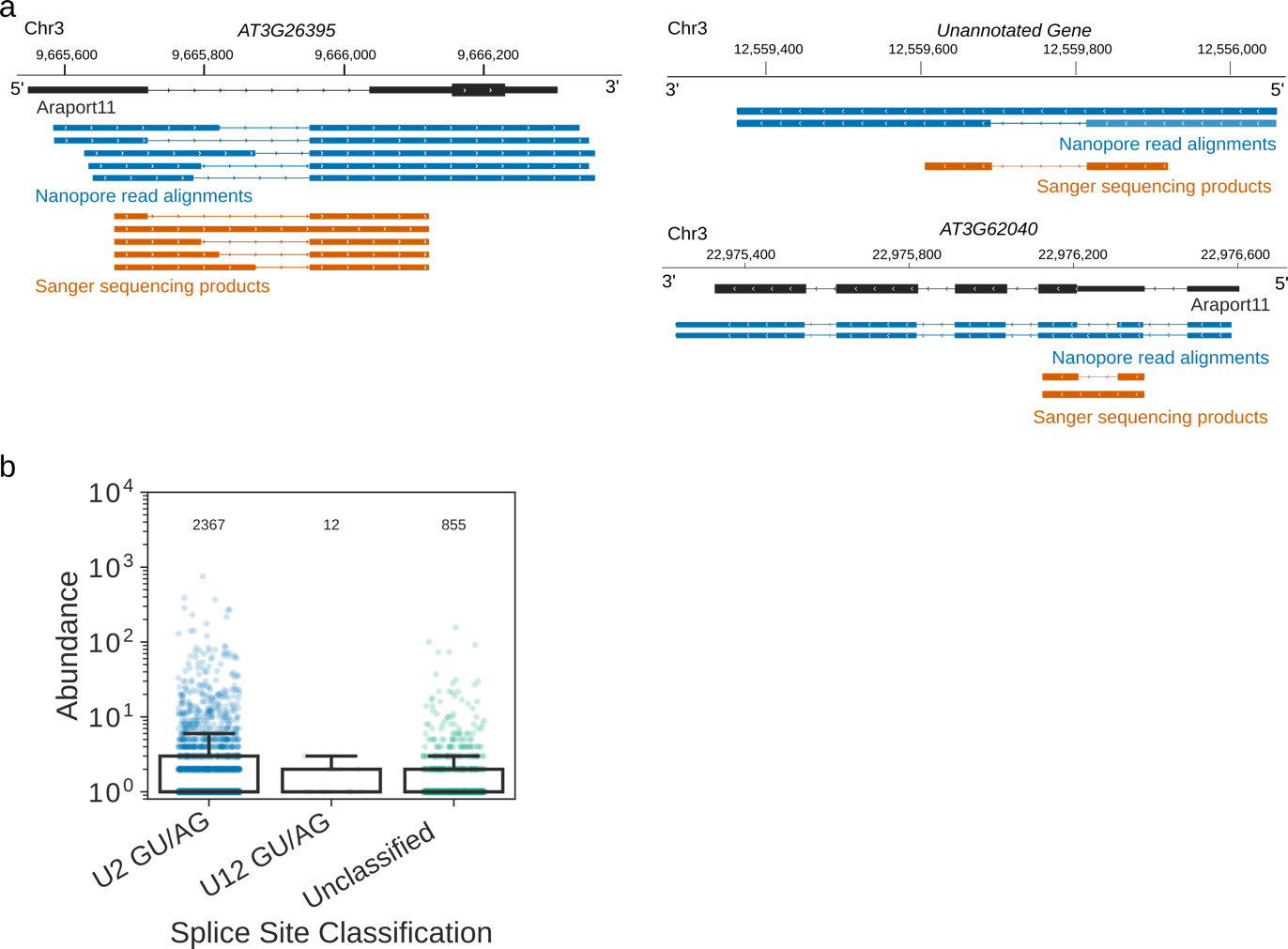
**Figure S2. 3’ end processing revealed by nanopore DRS.**

1. Poly(A) tail length of RNA negatively correlates with the expression of the gene. Expression in log scale of counts per million (CPM) obtained from nanopore DRS data is plotted against median poly(A) tail length. ρ - Spearman’s correlation coefficient.
2. Nanopore DRS identified 3’ polyadenylation sites in RNAs transcribed from *IBM1 (AT3G07610)* gene. Blue track shows coverage of nanopore DRS reads. Isoforms found in Araport11 annotation are shown in black and those detected by nanopore DRS are in blue. pPAS is proximal polyadenylation site.



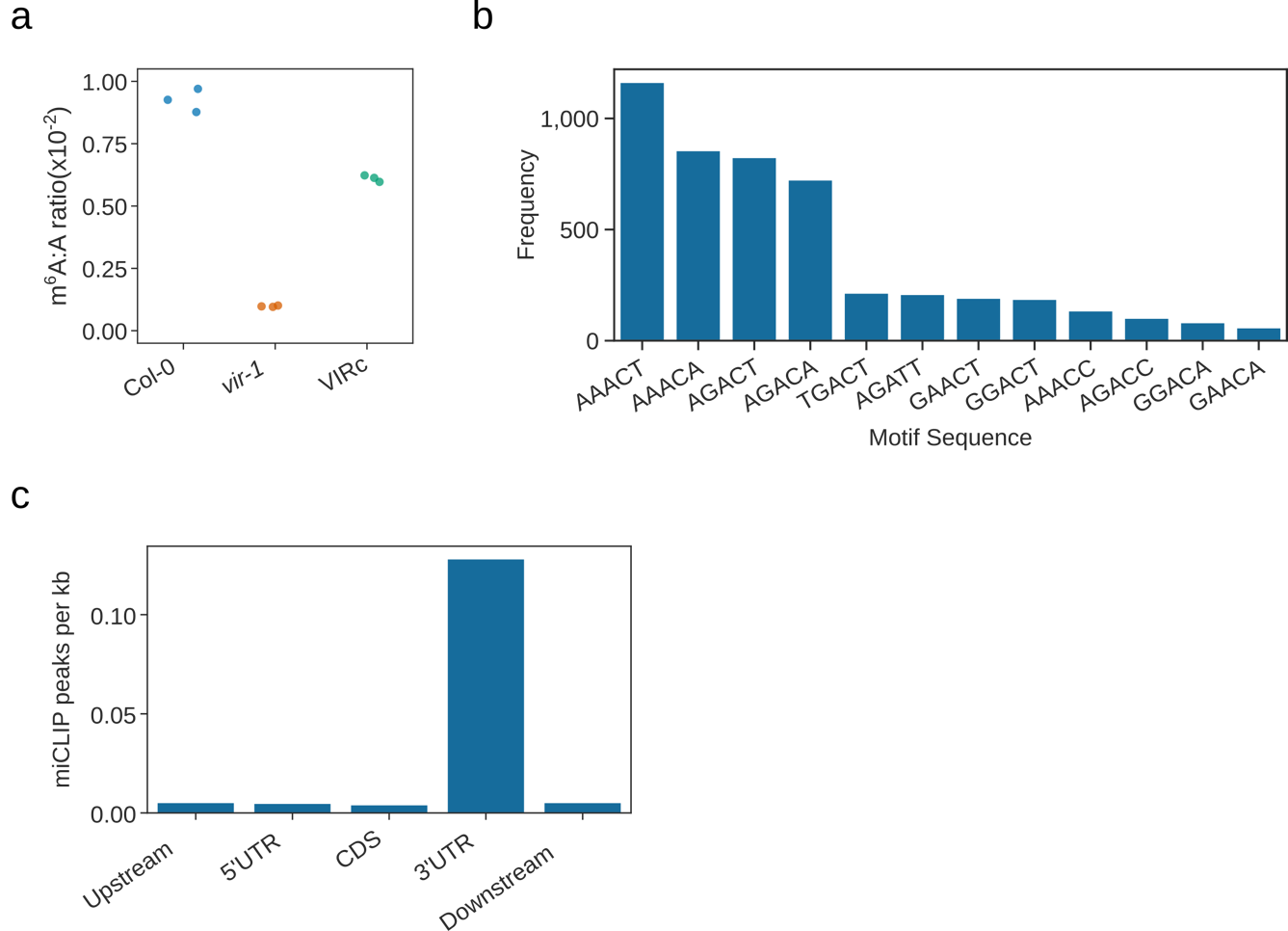
**Figure S3. Nanopore DRS with cap-dependent ligation of 5’ adapter RNA.**

1. Histogram showing the distribution of 5’ adapter RNA length in nanopore raw current signal, inferred from alignment of mRNA sequence to signal using nanopolish eventalign. The median signal length was 1441 points, and 96% of adapter signals were 3000 points or less.
2. Out-of-bag Receiver Operator Characteristic curve showing performance of trained convolutional neural network at detecting 5’ adapter RNA using 3000 points of signal. Curve was generated using 5-fold cross validation.
3. Out-of-bag precision recall curve showing performance of trained neural network, generated using 5-fold cross validation.
4. Alternative transcription start sites were identified using nanopore DRS with cap-dependent ligation of 5’ end adapter at *AT1G17050* and *AT5G18650* genes. Blue track shows coverage of nanopore DRS reads. Isoforms found in Araport11 annotation are shown in black and those detected by nanopore DRS with cap-dependent ligation of 5’ adapter RNA are in blue.
5. Reads mapping to ERCC RNA Spike-Ins lack approximately 11nt of sequence at the 5’ end. Histogram shows distance to the 5’ end for ERCC RNA Spike-In reads (each shown in different colours, only spike ins with more than 1000 supporting reads shown).
6. Reads mapping to *in vitro* transcribed mGFP lack approximately 11 nt of sequence at the 5’ end. Histogram shows distance to the 5’ end for *in vitro* transcribed mGFP.
7. Araport11 annotation overestimates the length of 5’ UTRs. Cumulative distribution function shows the distance to the nearest TSS identified from full length transcripts cloned as part of the RIKEN RAFL project (blue) and Araport11 annotation (orange).



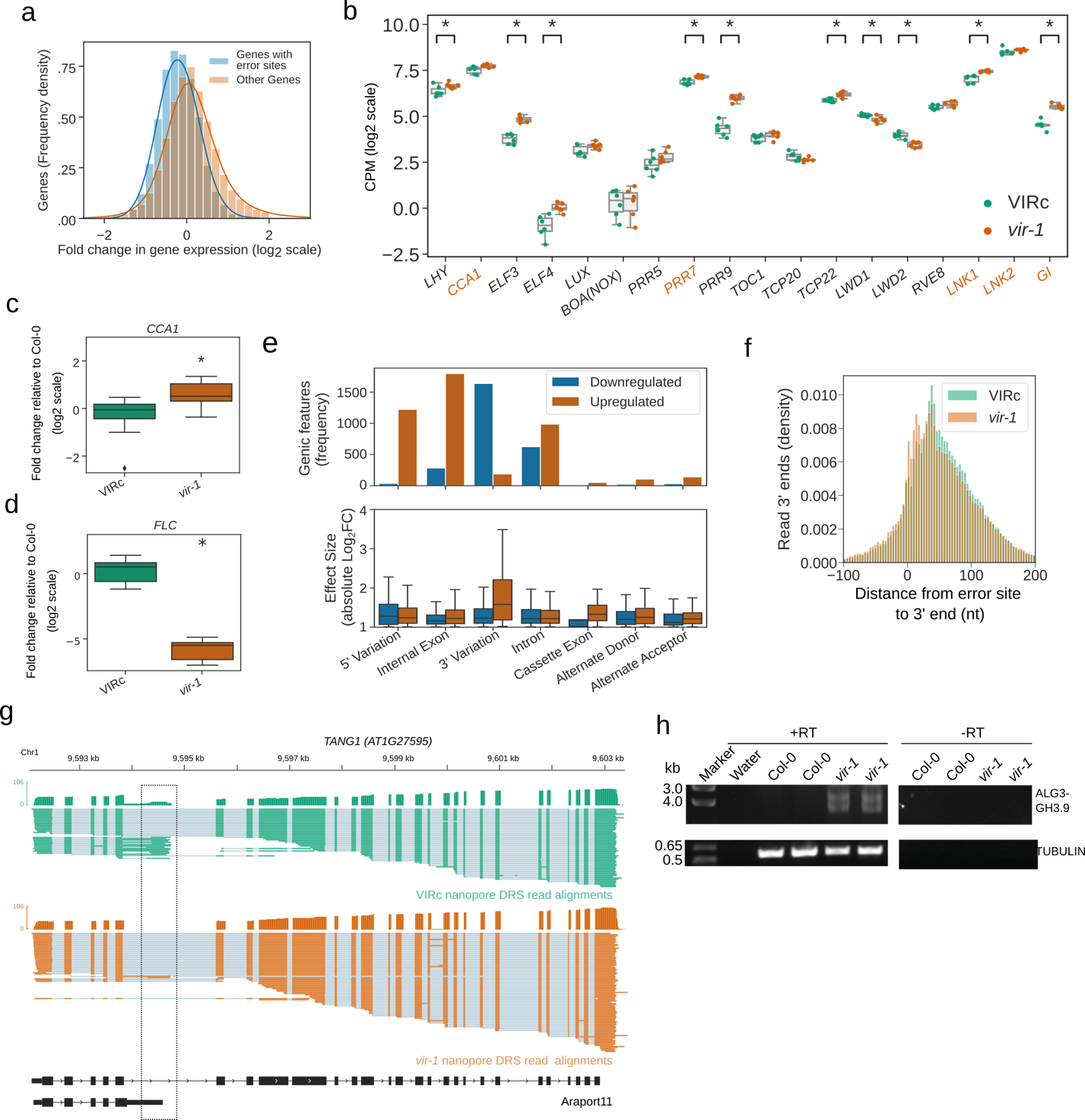
**Figure S4. Patterns of splicing revealed using nanopore DRS.**

1. Nanopore DRS can be validated using RT-PCR. From the top 20 most highly expressed RNAs with novel splice sites, five were selected, and of these, three were validated by RT-PCR followed by Sanger sequencing of the DNA products. RNA isoforms present in Araport11 annotation are shown in black; RNA isoforms found using nanopore DRS – in blue; Sanger sequencing products obtained using RT-PCR – in orange.
2. Splice junction classification of unannotated GU/AG splice sites found in error corrected nanopore DRS data, which also have Illumina support. Counts are plotted in log10 scale and the exact numbers of splice junctions in each set are indicated above each strip/box plot.



**Figure S5. Identification of VIR-dependent m6A transcriptome-wide.**

1. *vir-1* shows reduced levels of m6A compared to Col-0 that are restored in the VIR complemented line**.** Ratio of m6A/A obtained using LC-MS analysis is shown in blue; *vir-1* – in orange; VIR complemented line – in green.
2. Frequency of m6A motifs detected at *vir-1* reduced error sites, as detected by FIMO using the motif detected *de novo* by MEME (Fig 5B) and an FDR threshold of 0.1.
3. Bar plot shows the number of miCLIP peaks per kb of genic feature types in the Araport11 reference. Upstream and downstream regions are 200 nt regions before and after the annotated transcription termination sites (TTSs), respectively.



**Figure S6. Changes in circadian clock components gene expression and RNA 3’ end formation observed *vir-1***

1. Histogram shows log2 fold change in gene expression based on Illumina RNAseq data from *vir-1* mutant and VIR complemented line. Genes with differential error rate sites are in blue, and those without – in orange.
2. Expression of core circadian clock components is perturbed in *vir-1* mutant. Boxplots showing normalised gene expression measured using Illumina RNAseq in log2 counts per million, for VIR complemented line – in green; and *vir-1* mutant*–* in orange. Asterisks denote expression changes which are significant using an FDR threshold of 0.05. Orange labelled genes have 3’ UTR m6A detectable by Nanopore DRS and miCLIP.
3. Expression of *CCA1*, encoding a regulator of circadian rhythm in Arabidopsis, is increased in *vir-1* mutant. Boxplot shows gene expression change from Col-0 by RT-qPCR for VIR complemented line – in green; and *vir-1 –* in orange. Expression change in *vir-1* is significant using a p value threshold of 0.05.
4. Expression of *Flowering Locus C* (*FLC*) gene is decreased in the *vir-1* mutant. Boxplot shows gene expression change from Col-0 by RT-qPCR for VIR complemented line – in green; and *vir-1 –* in orange. Expression change in *vir-1* is significant using a p value threshold of 0.05.
5. Splicing is moderately disrupted in *vir-1* mutant. Results of differential exon usage analysis with DEXseq are shown for contiguous regions (“exon chunks”) which occur in the same sets of transcripts in the Araport11 reference. Regions were labelled as 5’ variation if they were bounded by the TSS of one or more transcripts, and 3’ variation if they were bounded by the TTS of one or more transcripts. Features with increased usage are shown in orange, and reduced usage in blue. Boxplots show the distribution in absolute log fold change for each feature set.
6. A shift in use to more proximal polyadenylation sites is observed in m6A containing transcripts in the *vir-1* mutant. Histogram shows distance from error site to 3’ end in *vir-1* mutant (orange) and VIR complemented line (green).
7. *vir-1* mutants exhibit increased readthrough of an intronic proximation poly(A) site in intron 5 of the Symplekin domain gene TANG1. Nanopore DRS reads from the VIR complemented line are shown in green, and from the *vir-1* mutant – in orange. Araport11 annotation is shown in black. Dashed black box highlights the site of proximal polyadenylation.
8. *ALG3-GH3.9* chimeric RNAs are generated in *vir-1* mutant. RT-PCR gel shows formation of chimeric RNAs in the ­*vir-1* mutant compared to Col-0.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Estimated Over-splitting rate (%)** | 0.047 | 0.045 | 0.111 | 0.035 | 0.047 | 0.040 | 0.019 | 0.019 | 0.069 | 0.809 | 0.959 | 1.600 | 0.878 | 1.467 | 0.887 |
| **Percent Mapped to TAIR10** | 60.89 | 70.25 | 79.04 | 76.94 | 25.32 | 74.89 | 70.12 | 75.98 | 27.30 | 90.52 | 94.62 | 91.85 | 94.74 | 93.78 | 93.53 |
| **Median length of reads mapped to TAIR10** | 865 | 878 | 873 | 878 | 737 | 858 | 840 | 840 | 732 | 871 | 846 | 832 | 884 | 802 | 849 |
| **Longest read mapped to TAIR10** | 12,607 | 12,446 | 11,994 | 11,852 | 5,663 | 11,434 | 12,744 | 11,542 | 5,596 | 12,015 | 15,040 | 10,836 | 15,291 | 11,522 | 11,471 |
| **Median length of reads mapped to ERCC spike-ins** | 481 | 492 | 493 | 492 | 486 | 488 | 486 | 495 | 507 | 497 | 491 | 493 | 480 | 487 | 500 |
| **Longest read mapped to ERCC spike-ins** | 1,905 | 2,304 | 1,108 | 1,099 | 1,928 | 1,094 | 1,772 | 1,292 | 2,308 | 2,236 | 1,880 | 1,947 | 1,134 | 1,128 | 2,127 |
| **Reads mapped to TAIR10** | 1,003,137 | 1,005,664 | 1,101,368 | 743,684 | 345,799 | 930,629 | 757,028 | 765,322 | 557,949 | 1,538,040 | 812,517 | 1,318,857 | 1,746,153 | 1,271,511 | 1,570,040 |
| **Reads mapped to ERCC spike-ins** | 6756 | 1074 | 713 | 556 | 7005 | 701 | 907 | 525 | 11324 | 1824 | 1094 | 2710 | 865 | 1436 | 1548 |
| **Total Reads Base-called** | 1,647,484 | 1,431,457 | 1,393,351 | 966,529 | 1,365,809 | 1,242,616 | 1,079,578 | 1,007,278 | 2,043,751 | 1,699,123 | 858,747 | 1,435,808 | 1,843,192 | 1,355,795 | 1,678,723 |
| **5' Adapter Ligation** | - | *-* | - | - | + | - | - | - | + | - | *-* | - | *-* | *-* | - |
| **Bio Replicate** | 1 | *1* | 1 | 2 | 1 | 2 | 3 | 4 | 2 | 2 | *2* | 4 | *3* | *4* | 3 |
| **Genotype** | Col-0 | *vir-1* | VIRc | Col-0 | Col-0 | Col-0 | Col-0 | Col-0 | Col-0 | VIRc | *vir-1* | VIRc | *vir-1* | *vir-1* | VIRc |
| **Sequencing Date** | 01/02/2018 | 27/02/2018 | 01/03/2018 | 05/04/2018 | 11/04/2018 | 13/04/2018 | 16/04/2018 | 18/04/2018 | 08/05/2018 | 07/09/2018 | 12/09/2018 | 25/09/2018 | 28/09/2018 | 03/10/2018 | 19/10/2018 |

**Table S1. Properties of the nanopore DRS sequencing data.**

Dataset statistics for all nanopore DRS sequencing runs conducted. Datasets are sorted by date of sequencing run. All data was collected using a MinION with R9.4 flow cell and SQK-RNA001 library kit. Increases in mapping and over-splitting rate which occur in samples collected after September 2018 are therefore likely to have resulted from changes in the MinKNOW software.

A

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Replicate 1 | Adapter | | No Adapter | | Signal to Noise |
| Fails | Passes | Fails | Passes |
| Match length > 10 nt | 58,078 | 211,786 | 47,548 | 11,427 | 15.17 |
| Log10 E Value < 2 | 76,926 | 192,938 | 57,874 | 1,101 | 131.8 |
| Adapter matches directly after alignment | 198,296 | 71,568 | 58,936 | 39 | 545.4 |

b

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Replicate 2 | Adapter | | No Adapter | | Signal to Noise |
| Fails | Passes | Fails | Passes |
| Match length > 10 nt | 37,631 | 342,158 | 100,791 | 12,653 | 72.43 |
| Log10 E Value < 2 | 60,605 | 319,184 | 112,781 | 663 | 895.9 |
| Adapter matches directly after alignment | 266,812 | 112,977 | 113,429 | 15 | 3,202 |

**Table S2. Adapter detection using BLASTN rules approach.**

Table shows the number of reads with adapters detected in two biological replicates of Col-0 sequenced (a and b) with and without adapter ligation protocol. Rules are applied cumulatively, *i.e*., row one shows reads that pass first rule row two shows reads that pass first and second rule, etc. Signal to noise ratio shows the number of positive examples detected using rules in adapter ligated dataset divided by the number of false positives from dataset collected without adapters.

a

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **AtRTD2** | **Araport11** | **Illumina** | **Nanopore EC** | **Nanopore** | **Canonical** | **Noncanonical** |
|  |  |  |  |  | 99,966 | 805 |
|  |  |  |  |  | 84,630 | 79,377 |
|  |  |  |  |  | 25,769 | 57,298 |
|  |  |  |  |  | 20,893 | 404 |
|  |  |  |  |  | 10,882 | 21,484 |
|  |  |  |  |  | 7,301 | 545 |
|  |  |  |  |  | 6,941 | 210 |
|  |  |  |  |  | 6,256 | 879 |
|  |  |  |  |  | 4,801 | 1,894 |
|  |  |  |  |  | 3,939 | 160 |
|  |  |  |  |  | 3,665 | 31 |
|  |  |  |  |  | 2,844 | 59 |
|  |  |  |  |  | 2,424 | 28 |
|  |  |  |  |  | 2,006 | 18 |
|  |  |  |  |  | 832 | 453 |
|  |  |  |  |  | 828 | 5 |
|  |  |  |  |  | 585 | 6 |
|  |  |  |  |  | 512 | 6 |
|  |  |  |  |  | 484 | 31 |
|  |  |  |  |  | 390 | 53 |
|  |  |  |  |  | 369 | 33 |
|  |  |  |  |  | 326 | 2 |
|  |  |  |  |  | 193 | 6 |
|  |  |  |  |  | 148 | 8 |
|  |  |  |  |  | 76 | 3 |
|  |  |  |  |  | 47 | 4 |
|  |  |  |  |  | 24 | 0 |
|  |  |  |  |  | 21 | 0 |
|  |  |  |  |  | 20 | 0 |
|  |  |  |  |  | 18 | 0 |
|  |  |  |  |  | 1 | 0 |

b

|  |  |  |  |
| --- | --- | --- | --- |
| **atRTD2** | **Araport11** | **DRS EC** | **Isoforms** |
|  |  |  | 29,894 |
|  |  |  | 27,585 |
|  |  |  | 11,967 |
|  |  |  | 8,659 |
|  |  |  | 3,788 |
|  |  |  | 756 |
|  |  |  | 341 |

**Table S3. Splice junctions supported by nanopore DRS and Illumina RNAseq**

Table shows the numbers for the unique splice junction set intersections upset plot (Figure 4b), and unique linked splicing events upset plot (Figure 4c). Shaded cells denote sets included in the intersection for that row, whilst unshaded cells denote sets excluded from the intersection. Rows are sorted by the size of the intersection for canonical splice junctions.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **AGI ID** | **Gene Name** | **Pathway** | **Log2**  **Fold**  **Change** | **Log2**  **counts**  **per**  **million** | **False**  **Discovery**  **Rate** | **Differentially**  **Expressed**  **(abs logFC > 1,**  **FDR < 0.05)** |
| *AT2G19520* | FVE | Autonomous | -0.40 | 5.49 | 3.95E-08 | - |
| *AT2G21660* | ATGRP7 | Autonomous | 0.03 | 6.11 | 8.57E-01 | - |
| *AT2G43410* | FPA | Autonomous | 0.70 | 5.03 | 1.75E-11 | - |
| *AT3G04610* | FLK | Autonomous | 0.29 | 5.18 | 5.90E-04 | - |
| *AT3G10390* | FLD | Autonomous | 0.31 | 3.80 | 6.68E-03 | - |
| *AT4G02560* | LD | Autonomous | 0.68 | 4.36 | 7.49E-09 | - |
| *AT4G16280* | FCA | Autonomous | 0.71 | 5.02 | 2.90E-09 | - |
| *AT5G13480* | FY | Autonomous | 0.49 | 3.80 | 8.78E-04 | - |
| *AT4G22950* | AGL19 | FLC-independent vernalization pathway | 1.10 | 0.81 | 9.17E-04 | + |
| *AT2G45660* | SOC1 | Floral integrator | -1.64 | 6.72 | 4.55E-84 | + |
| *AT5G60910* | FUL | Floral integrator and flower meristem identity | -1.13 | 2.09 | 1.84E-09 | + |
| *AT1G25560* | TEM1 | Floral repressor | -0.81 | 6.50 | 3.83E-02 | - |
| *AT1G68840* | TEM2 | Floral repressor | -0.19 | 7.08 | 6.30E-01 | - |
| *AT2G22540* | SVP | Floral repressor | -0.29 | 6.96 | 2.20E-03 | - |
| *AT3G57390* | AGL18 | Floral repressor | 0.10 | 3.30 | 5.36E-01 | - |
| *AT5G13790* | AGL15 | Floral repressor | 0.52 | 1.02 | 7.22E-02 | - |
| *AT1G24260* | SEP3 | Flower development | 1.02 | 2.82 | 9.91E-09 | + |
| *AT1G02400* | GA2ox6 | Gibberellin | 0.43 | 2.09 | 3.69E-02 | - |
| *AT1G14920* | GAI | Gibberellin | -0.27 | 6.22 | 5.95E-03 | - |
| *AT1G15550* | GA3ox1 | Gibberellin | -0.68 | 1.36 | 1.72E-02 | - |
| *AT1G18075* | miR159b | Gibberellin | 0.72 | 0.13 | 4.98E-02 | - |
| *AT1G30040* | GA2ox2 | Gibberellin | -0.08 | 1.86 | 8.15E-01 | - |
| *AT1G47990* | GA2ox4 | Gibberellin | -0.12 | 0.76 | 7.18E-01 | - |
| *AT1G78440* | GA2ox1 | Gibberellin | -1.43 | -0.19 | 2.57E-04 | + |
| *AT1G79460* | GA2 | Gibberellin | -0.33 | 2.80 | 4.69E-02 | - |
| *AT2G01570* | RGA | Gibberellin | -0.09 | 7.22 | 4.01E-01 | - |
| *AT3G05120* | GID1A | Gibberellin | -0.33 | 4.93 | 1.74E-03 | - |
| *AT3G11440* | MYB65 | Gibberellin | -0.37 | 1.99 | 4.23E-02 | - |
| *AT3G63010* | GID1B | Gibberellin | 0.76 | 3.51 | 4.26E-12 | - |
| *AT4G02780* | CPS1 | Gibberellin | 0.68 | 0.33 | 2.28E-02 | - |
| *AT4G24210* | SLY1 | Gibberellin | -0.39 | 4.02 | 2.31E-03 | - |
| *AT4G25420* | GA20ox1 | Gibberellin | 0.18 | 1.48 | 5.58E-01 | - |
| *AT5G06100* | MYB33 | Gibberellin | 0.17 | 3.76 | 1.20E-01 | - |
| *AT5G27320* | GID1C | Gibberellin | -0.06 | 4.74 | 6.60E-01 | - |
| *AT5G51810* | GA20ox2 | Gibberellin | -1.75 | 1.94 | 7.84E-14 | + |
| *AT5G03840* | TFL1 | Inflorescence architecture | -1.68 | -0.43 | 1.64E-05 | + |
| *AT2G27990* | PNF | Inflorescence meristem development | 0.52 | 1.78 | 1.60E-03 | - |
| *AT5G02030* | PNY | Inflorescence meristem development | 0.08 | 4.08 | 5.24E-01 | - |
| *AT1G04400* | CRY2 | Light perception | -0.28 | 7.25 | 4.69E-03 | - |
| *AT1G09570* | PHYA | Light perception | 0.01 | 7.56 | 9.35E-01 | - |
| *AT2G18790* | PHYB | Light perception | 0.13 | 6.35 | 2.45E-01 | - |
| *AT1G25540* | PFT1 | Light perception | 0.28 | 4.49 | 9.86E-03 | - |
| *AT4G34530* | CIB1 | Light signalling | -0.99 | 4.14 | 2.13E-04 | - |
| *AT5G46210* | CUL4 | Light signalling | 0.14 | 6.54 | 7.76E-02 | - |
| *AT4G24540* | AGL24 | Meristem response | -1.14 | 2.49 | 1.99E-08 | + |
| *AT4G35900* | FD | Meristem response | -0.31 | 1.78 | 2.46E-01 | - |
| *AT1G53090* | SPA4 | Photoperiod | 0.17 | 4.96 | 6.88E-02 | - |
| *AT1G65480* | FT | Photoperiod | -1.75 | -0.32 | 6.10E-08 | + |
| *AT1G69570* | CDF5 | Photoperiod | -0.34 | 4.05 | 7.97E-02 | - |
| *AT2G28550* | TOE1 | Photoperiod | 0.94 | 6.42 | 1.16E-24 | - |
| *AT2G32950* | COP1 | Photoperiod | 0.53 | 5.43 | 5.84E-09 | - |
| *AT2G34720* | NFYA4 | Photoperiod | -0.31 | 5.45 | 7.93E-03 | - |
| *AT2G38880* | NFYB1 | Photoperiod | -0.44 | 3.93 | 8.79E-06 | - |
| *AT2G39250* | SNZ | Photoperiod | 0.65 | 4.99 | 2.46E-09 | - |
| *AT2G46340* | SPA1 | Photoperiod | 0.58 | 7.45 | 4.89E-11 | - |
| *AT2G47700* | RFI2 | Photoperiod | -0.44 | 5.55 | 3.52E-05 | - |
| *AT3G15354* | SPA3 | Photoperiod | 0.84 | 6.06 | 1.01E-13 | - |
| *AT3G47500* | CDF3 | Photoperiod | 0.62 | 6.44 | 8.79E-11 | - |
| *AT3G48590* | NFYC1 | Photoperiod | -0.67 | 4.61 | 2.85E-09 | - |
| *AT3G54990* | SMZ | Photoperiod | 0.54 | 2.37 | 1.39E-02 | - |
| *AT4G11110* | SPA2 | Photoperiod | 0.28 | 5.26 | 2.56E-03 | - |
| *AT4G14540* | NFYB3 | Photoperiod | 0.23 | 5.21 | 1.93E-02 | - |
| *AT5G12840* | NFYA1 | Photoperiod | -0.30 | 5.15 | 2.97E-04 | - |
| *AT5G39660* | CDF2 | Photoperiod | -0.25 | 5.03 | 1.65E-01 | - |
| *AT5G47640* | NFYB2 | Photoperiod | -0.30 | 6.06 | 2.59E-03 | - |
| *AT5G57660* | COL5 | Photoperiod | -0.67 | 8.85 | 7.86E-07 | - |
| *AT5G60120* | TOE2 | photoperiod | 0.38 | 4.71 | 6.29E-05 | - |
| *AT5G62430* | CDF1 | Photoperiod | -0.01 | 5.50 | 9.25E-01 | - |
| *AT3G22380* | TIC | Photoperiod, circadian clock | 0.68 | 8.16 | 5.64E-15 | - |
| *AT5G57360* | ZTL | Photoperiod, circadian clock | 0.18 | 6.13 | 3.36E-02 | - |
| *AT1G01060* | LHY | Photoperiod, circadian clock | 0.32 | 6.54 | 1.85E-02 | - |
| *AT1G09530* | PIF3 | Photoperiod, circadian clock | -0.31 | 4.76 | 4.49E-02 | - |
| *AT1G12910* | LWD1 | Photoperiod, circadian clock | -0.29 | 4.94 | 3.13E-03 | - |
| *AT1G22770* | GI | Photoperiod, circadian clock | 0.93 | 5.13 | 3.98E-16 | - |
| *AT2G18915* | LKP2 | Photoperiod, circadian clock | -0.38 | 4.47 | 2.23E-03 | - |
| *AT2G21070* | FIO1 | Photoperiod, circadian clock | 0.31 | 2.93 | 1.67E-02 | - |
| *AT2G25930* | ELF3 | Photoperiod, circadian clock | 0.99 | 4.41 | 1.04E-14 | - |
| *AT2G40080* | ELF4 | Photoperiod, circadian clock | 1.01 | -0.36 | 2.20E-03 | + |
| *AT2G46790* | PRR9 | Photoperiod, circadian clock | 1.55 | 5.39 | 3.77E-20 | + |
| *AT2G46830* | CCA1 | Photoperiod, circadian clock | 0.04 | 7.64 | 7.18E-01 | - |
| *AT3G04910* | WNK1 | Photoperiod, circadian clock | -0.41 | 7.16 | 2.29E-04 | - |
| *AT3G26640* | LWD2 | Photoperiod, circadian clock | -0.57 | 3.75 | 1.89E-06 | - |
| *AT3G46640* | LUX | Photoperiod, circadian clock | 0.26 | 3.28 | 1.86E-01 | - |
| *AT3G60250* | CKB3 | Photoperiod, circadian clock | -0.53 | 5.49 | 3.48E-06 | - |
| *AT4G08920* | CRY1 | Photoperiod, circadian clock | -0.36 | 8.36 | 5.54E-05 | - |
| *AT4G16250* | PHYD | Photoperiod, circadian clock | 1.14 | 4.13 | 1.97E-19 | + |
| *AT4G18130* | PHYE | Photoperiod, circadian clock | 0.77 | 5.50 | 1.49E-14 | - |
| *AT5G02810* | PRR7 | Photoperiod, circadian clock | 0.23 | 7.01 | 5.61E-03 | - |
| *AT5G08330* | CHE | Photoperiod, circadian clock | -0.86 | 6.21 | 8.99E-11 | - |
| *AT5G24470* | PRR5 | Photoperiod, circadian clock | 0.31 | 2.67 | 2.29E-01 | - |
| *AT5G35840* | PHYC | Photoperiod, circadian clock | 0.46 | 5.53 | 2.50E-06 | - |
| *AT5G37260* | CIR1 | Photoperiod, circadian clock | 0.25 | 3.12 | 1.60E-01 | - |
| *AT5G59560* | SRR1 | Photoperiod, circadian clock | -0.28 | 3.43 | 4.74E-02 | - |
| *AT5G60100* | PRR3 | Photoperiod, circadian clock | 1.69 | 0.92 | 5.93E-11 | + |
| *AT5G61380* | PRR1 | Photoperiod, circadian clock | 0.12 | 3.91 | 3.92E-01 | - |
| *AT5G64813* | LIP1 | Photoperiod, circadian clock | -0.33 | 4.75 | 5.34E-04 | - |
| *AT1G68050* | FKF1 | Photoperiod, circadian clock | 0.68 | -0.32 | 3.91E-02 | - |
| *AT5G17690* | TFL2 | Photoperiod, vernalization and flower development | 0.00 | 4.46 | 9.97E-01 | - |
| *AT2G42200* | SPL9 | Vegetative and reproductive phase change | 0.42 | 3.03 | 8.60E-03 | - |
| *AT3G57920* | SPL15 | Vegetative and reproductive phase change | 0.89 | 0.49 | 1.34E-02 | - |
| *AT1G53160* | SPL4 | Vegetative to reproductive transition | -1.75 | 2.89 | 1.38E-20 | + |
| *AT2G25095* | miR156a | Vegetative to reproductive transition | 0.49 | 0.32 | 9.78E-02 | - |
| *AT2G33810* | SPL3 | Vegetative to reproductive transition | -1.55 | 4.69 | 4.76E-20 | + |
| *AT3G15270* | SPL5 | Vegetative to reproductive transition | -1.94 | 1.48 | 9.23E-15 | + |
| *AT4G31877* | miR156c | Vegetative to reproductive transition | -0.67 | 0.51 | 2.22E-02 | - |
| *AT1G17760* | Cstf77 | Vernalization | 0.23 | 4.42 | 4.52E-02 | - |
| *AT1G71800* | Cstf64 | Vernalization | 0.52 | 3.29 | 4.32E-05 | - |
| *AT3G18990* | VRN1 | Vernalization | -0.12 | 4.78 | 1.54E-01 | - |
| *AT4G00650* | FRI | Vernalization | 0.26 | 2.78 | 1.51E-01 | - |
| *AT5G10140* | FLC | Vernalization | -6.06 | 1.29 | 2.44E-64 | + |
| *AT1G61040* | VIP5 |  | 0.42 | 4.98 | 3.00E-08 | - |
| *AT2G06210* | VIP6 |  | 0.09 | 5.54 | 3.63E-01 | - |
| *AT2G31650* | ATX1 |  | 0.51 | 2.96 | 7.01E-03 | - |
| *AT3G48430* | REF6 |  | 0.26 | 5.21 | 1.89E-02 | - |
| *AT4G20400* | Jmj4 |  | 0.22 | 5.26 | 1.81E-02 | - |
| *AT5G04240* | ELF6 |  | 0.34 | 4.96 | 4.65E-04 | - |
| *AT1G08970* | NFYC3 |  | -0.64 | 5.32 | 1.57E-15 | - |
| *AT1G14400* | UBC1 |  | -0.59 | 6.65 | 1.10E-10 | - |
| *AT1G54830* | NFYC9 |  | -0.57 | 4.25 | 8.50E-08 | - |
| *AT1G55250* | HUB2 |  | -0.12 | 4.83 | 2.25E-01 | - |
| *AT1G56170* | NFYC2 |  | -0.29 | 3.59 | 2.10E-02 | - |
| *AT1G76710* | SDG26 |  | -0.01 | 3.53 | 9.47E-01 | - |
| *AT1G77080* | MAF1 |  | -0.31 | 5.01 | 2.71E-04 | - |
| *AT1G77300* | EFS |  | 0.87 | 5.69 | 2.31E-11 | - |
| *AT1G79730* | ELF7 |  | 0.58 | 5.26 | 2.31E-11 | - |
| *AT2G02760* | UBC2 |  | -0.35 | 6.12 | 2.08E-05 | - |
| *AT2G23380* | CLF |  | 0.18 | 4.07 | 2.68E-01 | - |
| *AT2G33835* | FES |  | 0.18 | 3.32 | 1.16E-01 | - |
| *AT2G44950* | HUB1 |  | 0.11 | 4.92 | 2.45E-01 | - |
| *AT2G45640* | SAP18 |  | -0.86 | 5.11 | 1.98E-15 | - |
| *AT3G12810* | PIE1 |  | 0.93 | 5.33 | 1.07E-17 | - |
| *AT3G24440* | VIL1 |  | 0.41 | 3.99 | 3.18E-04 | - |
| *AT3G33520* | ESD1 |  | -0.16 | 4.09 | 1.60E-01 | - |
| *AT3G49600* | UBP26 |  | 0.41 | 4.92 | 2.39E-05 | - |
| *AT4G15880* | ESD4 |  | 0.00 | 4.68 | 9.60E-01 | - |
| *AT4G16845* | VRN2 |  | -0.03 | 4.17 | 7.95E-01 | - |
| *AT4G29830* | VIP3 |  | -0.22 | 3.96 | 1.23E-02 | - |
| *AT4G32980* | ATH1 |  | 0.94 | 4.91 | 2.84E-15 | - |
| *AT4G39400* | BRI1 |  | 0.22 | 7.58 | 1.34E-01 | - |
| *AT5G11530* | EMF1 |  | 0.43 | 4.70 | 4.55E-04 | - |
| *AT5G16320* | FRL1 |  | 0.09 | 2.57 | 6.41E-01 | - |
| *AT5G23150* | HUA2 |  | 1.38 | 5.55 | 2.97E-26 | + |
| *AT5G37055* | SEF |  | -0.44 | 2.64 | 2.11E-03 | - |
| *AT5G51230* | EMF2 |  | 0.36 | 5.12 | 8.51E-06 | - |
| *AT5G61150* | VIP4 |  | 0.07 | 5.73 | 4.71E-01 | - |
| *AT5G65050* | MAF2 |  | -0.27 | 5.31 | 2.58E-03 | - |
| *AT5G65060* | MAF3 |  | -0.64 | 3.89 | 9.91E-09 | - |
| *AT5G65070* | MAF4 |  | -0.20 | 0.46 | 4.02E-01 | - |

**Table S4. Flowering time gene expression**

Change in gene expression of curated genes involved in flowering time in Arabidopsis, detected using Illumina RNAseq of *vir-1* compared to the VIR complemented line. 12.2% of flowering time genes change mRNA level expression in the *vir-1* mutant. Source of Flowering Time Genes: George Coupland, Cologne: https://www.mpipz.mpg.de/14637/Arabidopsis\_flowering\_genes

|  |  |  |  |
| --- | --- | --- | --- |
| **Primer name** | **Primer sequence** | **Experiment** | **Gene or gene coordinates** |
| At1g29550 F | ATGGTGGTTACGGATTCTCCA | Chimeric RNA detection RT-PCR | *AT1G29550-AT1G29560* |
| At1g29560 R | TTAAACTCCATCTCCCTCTTT |
| F\_spl2\_A | GCTGTGTTCCTTTTCTCAGCA | Splicing events validation RT-PCR | *AT2G27385* |
| R\_spl2\_A | CTGCCTTGGAGTTCTGCTTC |
| F\_spl5\_A | AGGGATAAGCTCTTCTTTTTCC | Splicing events validation RT-PCR | 3:12559692-12559814 |
| R\_spl5\_A | CGACGGTGACAACGAACC |
| F\_spl7\_AB | GTGGTGTTGAGCCGGAGT | Splicing events validation RT-PCR | *AT3G42806* |
| R\_spl7\_B | CTTTCTCGGTGGCTTTTCTC |
| F\_spl8\_B | ATATGGACGATACTTTATACCCTTTG | Splicing events validation RT-PCR | *AT3G62040* |
| R\_spl8\_B | AGCCATGGTGGTTCCATATT |
| F\_spl9\_AB | AATATTTTCCGGTGAATAGCC | Splicing events validation RT-PCR | *AT3G26395* |
| R\_spl9\_B | AAGCCCAGAATCCGGTGA |
| CCA1\_For | GATGATGTTGAGGCGGATG | RT-qPCR (clock phenotype) | *AT2G46830* |
| CCA1\_Rev | TGGTGTTAACTGAGCTGTGAAG |
| UBC\_For | CTGCGACTCAGGGAATCTTCTAA | RT-qPCR (clock phenotype) | *AT5G25760* |
| UBC\_Rev | TTGTGCCATTGAATTGAACCC |
| F\_GFP\_CDS\_T7prom\_NotI | GCGGCCGC TAATACGACTCACTATAGGGAGA ATGAGTAAAGGAGAAGAACTTTTCACTG | PCR | *mGFP* |
| R\_GFP\_CDS\_AscI | GGCGCGCC TTATTTGTATAGTTCATCCATGCCATG |
| FLC-Fwd | GAGCCAAGAAGACCGAACTC | RT-qPCR (flowering phenotype) | *AT5G10140* |
| FLC-Rev | TTCTGCTCCCACATGATGA |
| RT41clip | NNGTATNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC | miCLIP library preparation |  |
| RT48clip | NNTGTGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC |
| RT49clip | NNTTCTNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC |
| RT50clip | NNTTTCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC |

**Table S5. Primers used in this study**