Native long-read RNA sequencing of the Arabidopsis thaliana transcriptome

### Schurch N.J.1, Knop K.2, Sherwood A.2, Parker M.2, Barton G.J.1 & Simpson G.G.2,3

1 Computational Biology, School of Life Sciences, University of Dundee, DD1 5EH

2 Gene Regulation & Expression, School of Life Sciences, University of Dundee, DD1 5EH

3 Plant Sciences, School of Life Sciences, University of Dundee, DD1 5EH

# Abstract

## Introduction

High-quality gene and transcript annotations are cornerstones of both modern cell biology and evolutionary genetics. Such annotations are currently restricted to model organisms because generating High quality transcriptome annotations currently requires i) a high-quality genome assembly, ii) extensive computational prediction based on protein information from related species, iii) supporting multi-faceted transcriptomic data and, for the highest quality annotations, iv) extensive manual curation [1-3]. As a result, this resource intensive work has been limited to large dedicated organisations (*e.g.* the European Bioinformatics Institute & the National Center for Biotechnology Information) or large, genus-specific consortia (*e.g.* TAIR/Araport, SGD), and is generally restricted to the most important model organisms. Assembling high-quality transcriptome annotations *de novo* from short-fragment cDNA sequencing (RNA-seq) data has proven to be extremely challenging in eukaryotes [4, 5] and requires considerable computational and financial resources. Problems arise with this approach because i) the length of the reads (<~150bp) limits the features that can be reliably tied together, and ii) transcription to cDNA and PCR amplification disconnects the reads from parent transcript resulting in a loss of information about critical RNA modifications and the polyadenylated tail of the transcript. As a result, *de novo* transcriptome assembly is not widely used for annotating model eukaryotes, however in non-model species this is often the only option for annotating the genome. The result is that most eukaryotic life on earth does not benefit from high-quality annotation, even where a genome assembly does exist. This presents a considerable stumbling-block to both the understanding of their unique biology and our broader understanding of their diversity and evolutionary biology.

The power of long-read sequencing to identify novel RNA biology has been highlighted by recent work on deep PacBio cDNA sequencing data [6]. This approach is effective at defining the relationship between splicing, transcription start sites (TSSs) and poly-adenylation sites (PASs), however it is not a scalable or cost-effective solution for making genome annotation accessible to a wide range of species, including those that are primarily the focus for developing countries with limited research resources. Generating PacBio data is expensive, both in terms of initial equipment outlay, operating costs, space requirement, and per transcriptome sequencing consumables. Furthermore, this approach requires separate sequencing of size-selected samples in order to capture the longest mRNAs, amplifying the sequencing cost at least four-fold. These data also suffer from sequence-specific sequencing and amplification biases, as with any amplified cDNA sequencing method.

Long-read direct RNA sequencing with the Oxford Nanopore Technologies MinION (hereafter, DRS [7]) sidesteps many of these issues by sequencing RNA molecules directly. Briefly, a double stranded cDNA adapter including a pre-loaded motor protein is ligated to poly-A tailed RNAs. The motor protein then interacts directly with a nanopore embedded within a membrane, feeding the attached 3’ adapter and the captured RNA molecule through the pore one base at a time. When a voltage is applied across the membrane the specific nucleotide sequence in the pore changes electrical current that flows through the pore and these signals are used to identify the nucleotides present. This technology has several potential strengths that, in principle, make it well-suited for de novo transcriptome annotation. The equipment is cheap, portable, simple to use, and samples the full distribution of mRNA sizes in the sample in a single sequencing run without requiring size-selection steps (Fig 1). Additionally, ONT DRS avoids introducing biases to the dataset by reading the RNA molecule directly, without requiring reverse transcription or amplification, and includes sequencing the polyadenylated tail and RNA modifications. These data have the potential to capture links between the observed splicing patterns and other key transcriptional features such as the transcription start site, poly-adenylation site (PAS), the poly-adenylated tail and any RNA modifications that are present along the molecule. By directly assaying the RNA content of a sample and capturing the relationships between the key features of mRNA transcripts, the relatively recent development of long-read RNA sequencing technology holds the potential to revolutionize our ability to annotate eukaryotic genomes and redefine our understanding of eukaryotic genetics.

Here we illuminate the current strengths and limitations of ONT DRS data to define the transcriptome of a higher eukaryote using a combination of six Arabidopsis thaliana DRS sequencing runs, and two ONT DRS dataset of in vivo transcribed GFP RNA.

## Results

After sequencing, each of the four standard-protocol DRS wild-type col-0 biological replicates, and two modified-protocol DRS wild-type col-0 biological replicates were re-basecalled with albacore (v2.2.7) and aligned to the TAIR10 Arabidopsis thaliana genome with minimap2. The resulting read alignments were processed and analysed to highlight the capabilities of the DRS data.

### ONT DRS captures the full spectrum of mRNA lengths, without requiring size selection

Despite the number of RNA molecules sequenced in a MinION run varies considerably between sequencing runs, the four of the standard protocol DRS datasets from individual biological replicates show that DRS captures sequences spanning the full spectrum of the *Arabidopsis* transcriptome (Figure 1, panel A-D, c.f. Figure 1, panel G & H). The longest read alignments in these datasets are 12,540 & 12,481 kb, align to known genes and span 64 (AT1G48090) & 58 (AT1G67120) exons, respectively (Supp. Figures 1a & 1b). These represent some of the longest contiguous mRNAs sequenced in Arabidopsis and they originate from some of the longest transcript models in the existing annotations (the number of transcripts >12.5kb in Araport is 18 and in atRTD is 31). The median read length for the datasets is considerably less than the median transcript length in the gold-standard annotation (~850bp c.f. 1600 bp), however the read length median confounds the distribution of underlying mRNa lengths with their expression level so a direct comparison of these number is not informative without collapsing the reads to transcript models. To date, there is no published method for doing this with DRS data, in part because of the uses we discuss in later sections of this paper.

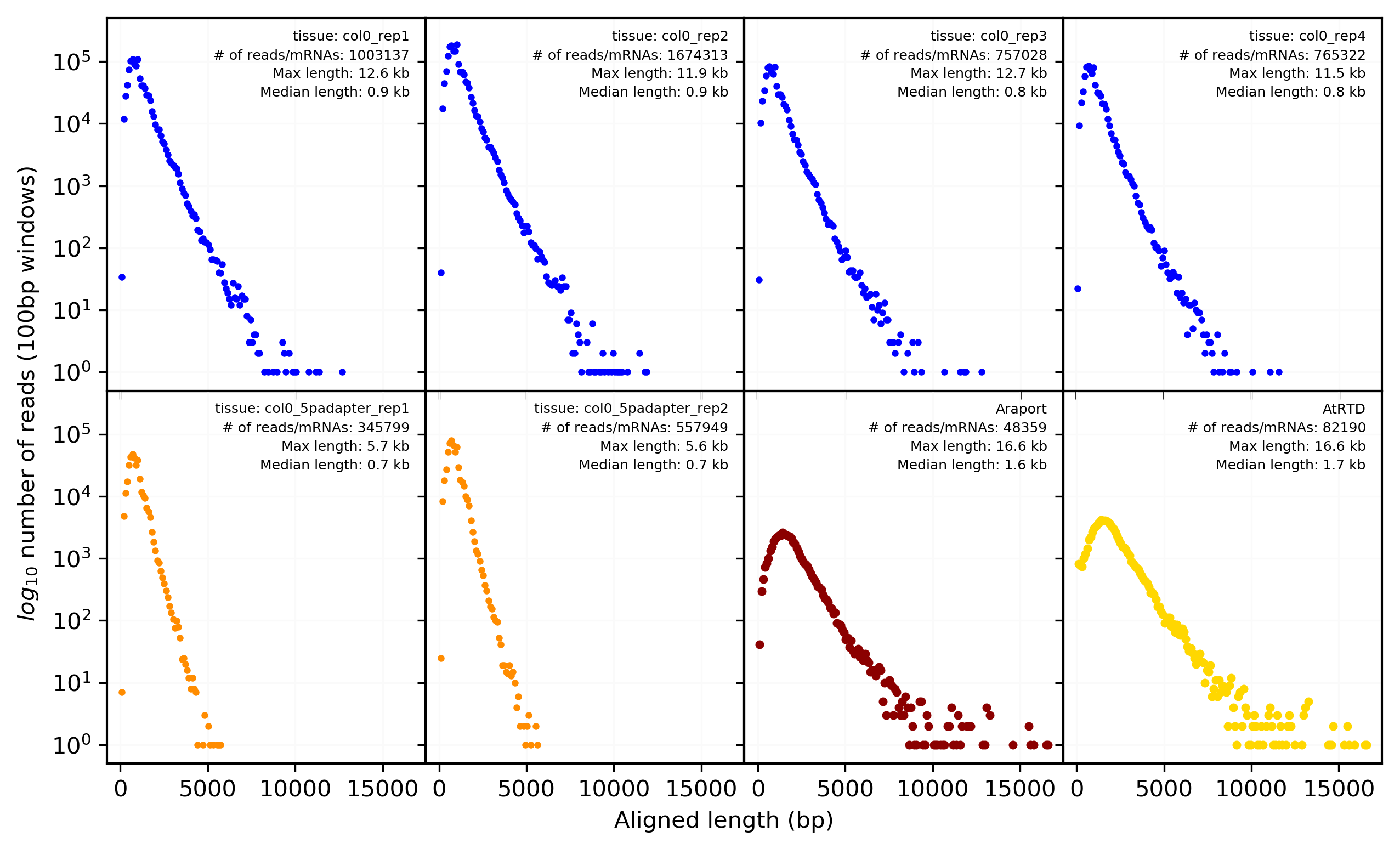


Figure 1:A comparison of ONT DRS read-alignment length distributions for four col-0 biological replicates of RNA prepared using the standard protocol (blue, upper row), two of the col-0 biological replicates prepared with a modified protocol incorporating a 5’ adapter ligation step (orange, bottom row, see Section 2.3.1), and aligned length distributions for the mRNAs models from the two gold-standard Arabidopsis thaliana reference annotations, Araport (dark red, bottom row), and AtRTD (yellow, bottom row). Included for each panel are the total number of aligned reads (or mRNAs) and both the maximum and median alignment lengths. We note that for the ONT datasets, the shape and peak of the distributions convolve both the length distribution of the underlying poly-A mRNA complement with their expression. By contrast, the annotation distributions only include a single copy of each mRNA model, irrespective of poly-adenylation status. Nevertheless, it is clear that the ONT DRS datasets span a large fraction of the mRNA lengths present in the annotations.

### Sequencing and alignment accuracy

Each of our datasets includes the ERCC spike-ins [8], a set of 92 artificial RNA molecules designed by the ERCC consortium to mimic the characteristics of eukaryotic mRNAs, while being unique and distinct from them across a wide range of species. The spike-ins range in length from 273-2022bp, have a GC content ranging from 30-53%, and are included in the kit with a range of concentrations that covers more than six orders of magnitude. We use the known sequence of the spike-ins to precisely quantify the performance of ONT DRS sequencing. We detect 59 of the 92 spike-ins with a total of 27,771 reads, yielding ~18.7Mb of sequence. For these reads, an average of 93% of the read sequence (~17.5Mb) aligns to the ERCC reference sequences (Figure 2, left) with an accuracy (mean sequence identity) of (Figure 2, centre).

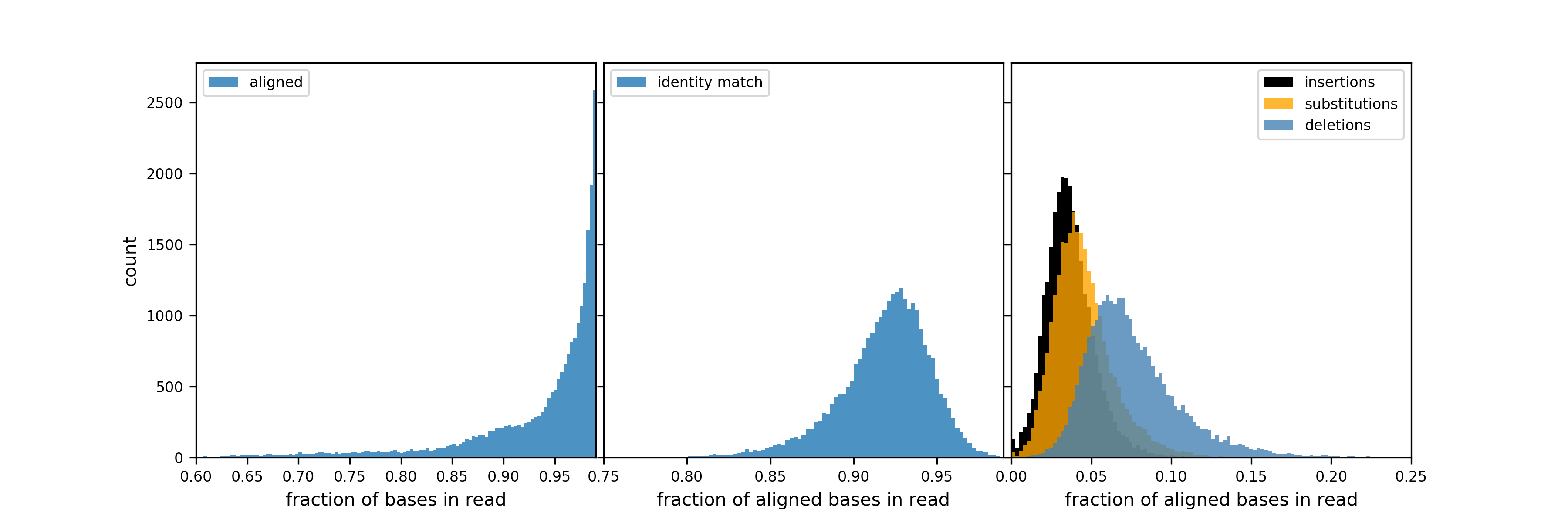
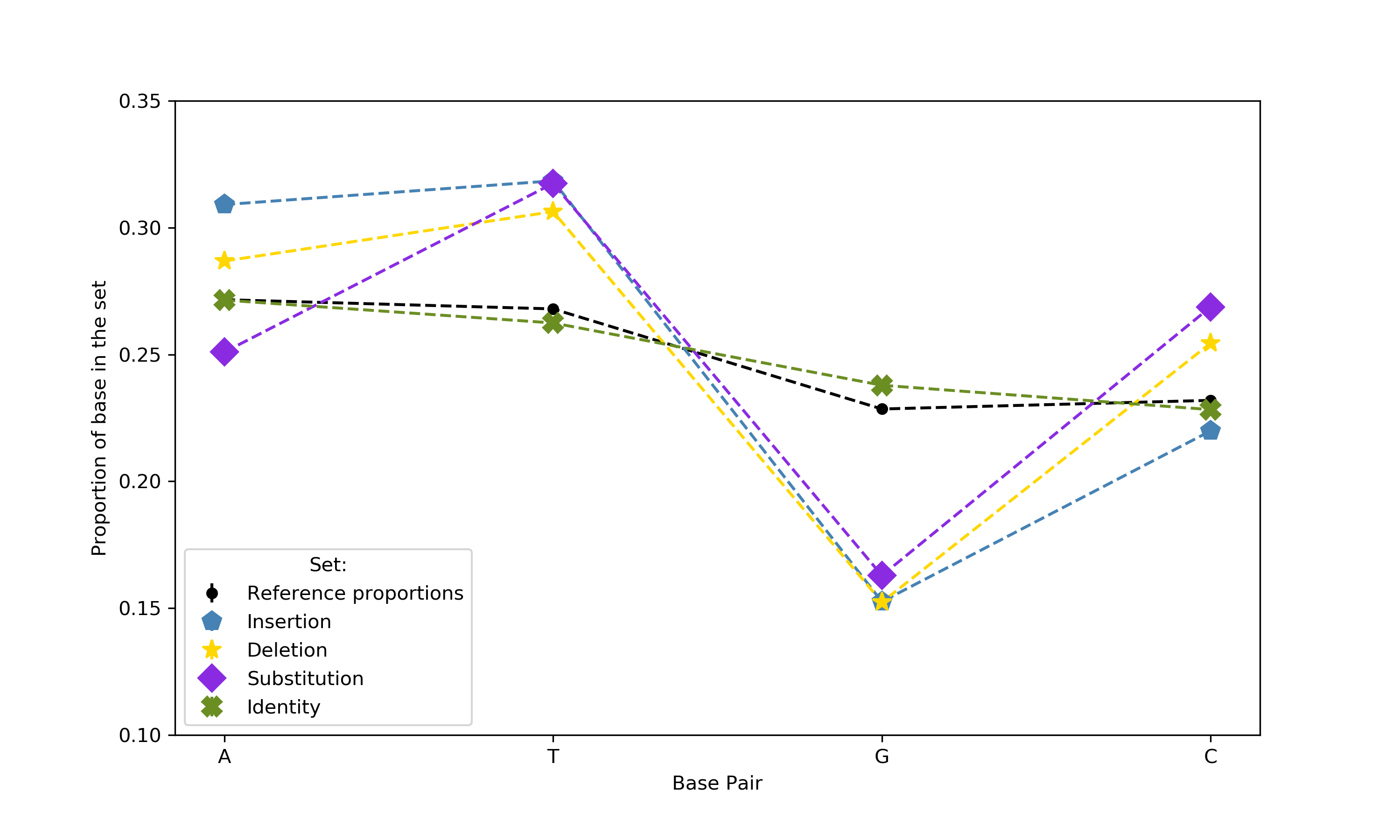


Figure 2: Overview of the sequencing and alignment characteristics of ONT DRS data for ERCC spike-ins. Left: Distribution of the length fraction of each sequenced read that aligns to the ERCC spike-in reference. Center: Distribution of fraction of identity matches between the sequence of the read and the ERCC reference for the aligned portion of each read. Right: Distributions of the occurance of insertions (black), substitutions (orange) and deletions (blue) as a proportion of the number of aligned bases in each read.



*Figure 3a: Nucleotide representation within the ERCC reference sequences (black dots) compared with the nucleotide representation within four categories from the ONT DRS read alignments; identity matches between the sequence of the read and the ERCC reference (green crosses), insertions (blue pentagons), delections (yellow stars) and substitutions (purple diamonds). Guanidine (G) is under-represented, and Thymine (T) 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.*

The remaining 1.5M sequence/alignment errors are broken down into three categories, insertions to the reference, deletions from the reference and base substitutions. Deletions from the reference make up the majority () of the errors for the aligned sequences with both substitutions and insertions accounting for <5%. These errors show evidence for base specificity. Guanidine is under-represented, and Thymine is over-represented, for all three categories of error (insertion, deletion and substitution) relative to the reference nucleotide distributions (Figure 3a). Cytosine is over-represented in the set of deletions and substitutions and, where substituted, is most likely to be replaced with a Thymine (Figure 3b). Adenosine is over-represented for Insertions and deletions and, under-represented in the set of substitutions. The results of the spike-in analysis are reflected for the 4.8Gb of read alignments to the *Arabidopsis* genomic sequences (see Supplementary Figures XX), with small changes in the rates if insertions, deletions and substitutions reflecting the additional sources of biological variation in the data. These results clearly demonstrate that, contry to popular belief, ONT DRS sequencing errors are not unbiased and they suggest caution when interpreting sequencing errors in the raw data as evidence for RNA modifications. In particular the over-representation of substitutions of Thymine for Cytosine has potential implications for the identification of 5mC.

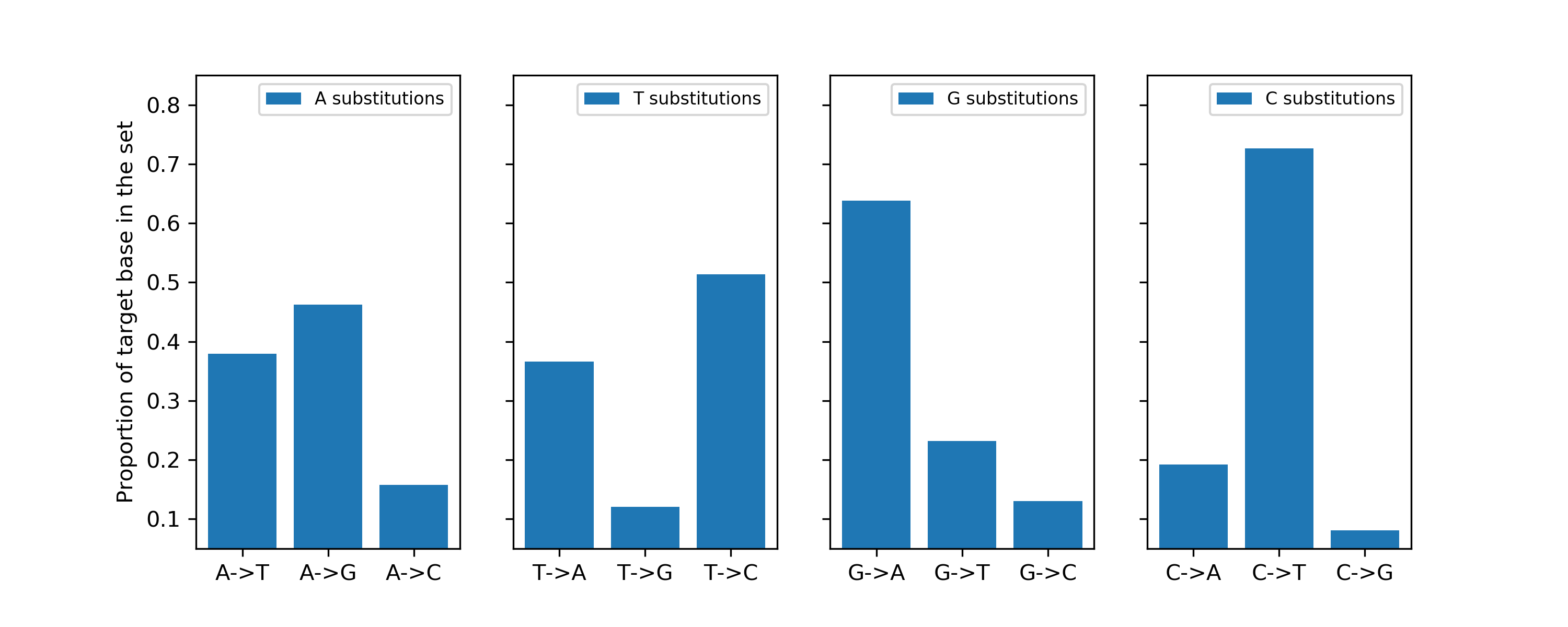


Figure 3b: Substitution preference for each Nucleotide (A, T, G, C, left-to-right). When substituted, Guanidine is replaced with Adenosine in more than 63% of its substitutions, while Cytosine is replaced by Thymine 73% of the time. Conversely Thymine is rarely replaced with Guanidine (12%) and Adenine is rarely substituted with Cytosine (16%).

Although improvements in the MinION sequencing accuracy are expected, an alternative approach in the interim is to attempt to correct the errors in the sequenced DRS reads. Several approaches have been taken previously for dealing with the impact of errors on the downstream analysis of ONT DRS sequencing. Several tools attempt to correct the sequencing errors by employing a consensus-based approach using either other ONT DRS dataset or orthogonal sequencing data (for example, *Mandalorian* https://github.com/christopher-vollmers/Mandalorion-Episode-II] and *proovread* [9]). Others, do not directly correct the raw reads and, instead, attempt to correct the downstream read alignments, replacing features of these read alignments with features from either existing annotations or the alignments of orthogonal sequencing datasets (for example, *FLAIR* [10]). Although the ONT DRS error profile is not unbiased, the base-specific differences are relatively small (<25%) and the overall error rate is low suggesting that consensus based approached may prove effective. Although designed to work with PacBio DNa sequencing, we evaluate the potential for these approaches to improve the utility of ONT DRS sequencing by using *proovread (vXXX)* to correct the ONT DRS data presented here using 100bp paired-end poly-A+ Illumina short-reads from six col-0 biological replicates (Materials and Methods: error correction). Here, *proovread* aligns the short-read data to the ONT DRS and then for each base, corrects the sequence of the DRS read to reflect the consensus of the short-read alignments at this position. As expected, the error correction results in a considerable improvement in the accuracy and length of the read alignments to the *Arabidopsis* genome. For these reads, an average of 97% of the read sequence aligns to the reference sequence with an accuracy (mean sequence identity) of 99.87%. The extremely high accuracy of these alignments is the direct result of *proovread* overwriting the original DRS sequence information with the consensus information from the high-accuracy Illumina data. This improved accuracy comes with several trade-offs, however, including overwriting any genuine biological nucleotide variation (e.g. Single Nucleotide Polymorphisms) from the DRS data that is not present in the Illumina sequencing, removing sequence errors introduced by the passage of RNA modifications through the pore. Furthermore, the error correction process is necessarily non-uniform, correlating with the non-uniform sequencing depth of the Illumina reads aligning to the DRS reads.

### Accurate identification of full-length isoforms from DRS data

### Identifying 5’ transcription start sites

DRS data generated with the standard DRS protocol does not identify the 5’ cap structure of RNA molecules or include a 5’ terminal marker preventing downstream differentiation between full-length reads from polyadenylated and capped mRNAs and reads that originate from truncated or uncapped RNAs, RNA fragments, or incomplete sequencing of a full-length molecule. Here we developed a protocol for ligating a 5’ RNA barcode sequence to 5’ methyl-capped mRNAs and then enriching the sample for these labelled transcripts (see Materials & Methods X). We applied this modified protocol to two sequencing runs using the same biological material as two of our existing biological replicates (tissues 2916 & 2918), which thus act as unlabelled controls for this experiment.

For the two labelled, enriched, datasets and the two standard protocol controls, after sequencing, re-basecalling and alignment, we extract all unaligned (soft-clipped) sequence immediately upstream of the 5’ end of read alignments and use *blastn* to identify which of these soft-clipped sequences have full or significant partial matches to the known adapter sequence. For the initial sequence matching all hits with e-values <104 are reported, the sequences are not masked for low complexity, only one highest scoring pair is reported for each hit, and require a minimum word length of 4. Downstream, we filter the resulting hits by requiring a minimum match length threshold, a minimum e-value threshold, and a positional constraint locating the start of the adapter sequence immediately upstream of the 5’ end of the read alignment. Iterating over a wide range of these threshold parameters, for both the enriched samples and the controls, allowed us to quantify how the signal-to-noise ratio (S/N) of adapter detection varies as a function of these parameters (supp. Figure 2) and then select appropriate values for these thresholds. We select two sets of thresholds; a *relaxed* threshold (match length >10, e-value < 100) and a *stringent* threshold match length >10, e-value < 100, sstart = 1 or send=1). Applying these thresholds to each of the paired tagged and control datasets show that the relaxed threshold has a S/N > 500, while the stringent threshold has a S/N > 5000 (Table 1).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Tissue 2916** | | | **Tissue 2918** | | |
| **+ adapter** | **- adapter** |  | **+ adapter** | **- adapter** |  |
| **Number of aligned reads** | 341779 | 984172 |  | 549429 | 745243 |  |
|  |  |  | **S/N Ratio** |  |  | **S/N Ratio** |
| **Total *blastn* hits** | 227166 | 47797 | 13.7 | 361746 | 83506 | 5.9 |
| **+ Match length >10** | 205440 | 10657 | 55.5 | 336118 | 11752 | 38.8 |
| **+ e-value < 100** | 192778 | 1060 | 523.7 | 318858 | 650 | 665.4 |
| **+ adapter start/end == 1** | 71568 | 39 | 5284.2 | 112977 | 15 | 10216 |

Here, we consider that reads passing the stringent filter accurately identify full-length mRNA molecules and yield base-pair resolution identification of the transcription start site for these molecules. The position of the 5’ ends of these reads are in excellent agreement with existing TSSs annotated in TAIR and peaks identified in publicly available PEAK data [11] (Figure 2A, green track).

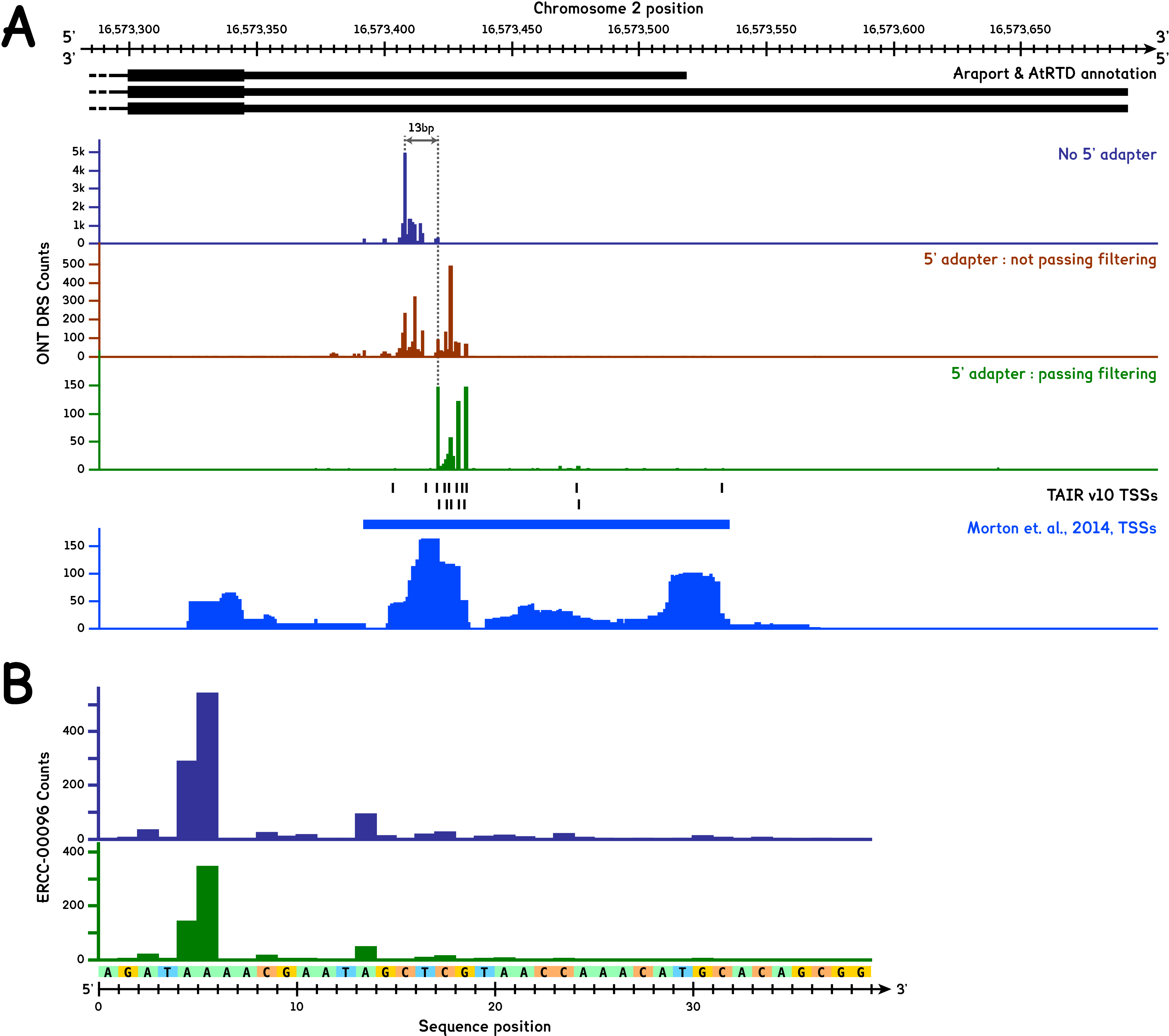


Figure 4: A: ONT DRS sequencing of data that includes a 5’ adapter identifies full-length isoform reads in Rubisco Activase (RCA, AT2G39730). The chromosome locus Chr2:16573280-16573660, reverse strand, showing the current gold-standard reference annotations, Araport, ArRTD & TAIR10 Transcription Start Sites (TSS )identified from EST clones (Black), and i) 5’ end positions of ONT DRS read alignments combined from the two adapter ligated datasets, that pass the stringent filtering for adapter identification (green), ii) 5’ end positions of ONT DRS read alignments combined from the two adapter ligated datasets, that do not pass the stringent filtering for adapter identification (red), and iii) the 5’ end positions of ONT DRS read alignments combined from the two paired sample non-adapter ligated datasets (dark blue). For this locus we also show publically availble PEAK data from Morton et. al. (2014, [11]) and the low-resolution TSS position identified from these data. The stringent filtering identifies adapters in 33% of the reads from this locus and these ends agree well with the TSSs identified from EST clones. A comparison of the end positions that do not pass filtering with those that do, and with the alignment ends from the datasets without adapters, demonstrated that they fall into two regions separated by ~13bp. 51% of those reads that do not pass the stringent filtering and 5’ alignment ends coincident with the ends from reads that do pass filtering. B: ONT DRS sequencing misses 5-15bp from the 5’ ends RNAs. Read alignments for the ERCC spike-ins demonstrating that ONT DRS sequencing commonly fails to sequence to the full-length of RNA molecules, asa result of the motor protein no longer controlling the progression of the molecule through the pore.

Datasets without the adapter sequence are intrinsically unable to identify the 5’ ends of sequenced molecules, even for sequenced reads that appear to represent the full-length of an annotated isoform, because the sequencing requires that the RNA sequence be fed through the nanopore under control of the motor protein attached to the 3’ end of the RNA molecules during library preparation [7]. When the 5’ end of the RNA passes through the motor protein prior to entering the nanopore, the RNA is no longer under the processive control of the motor protein. As a result, it passes through the nanopore in a non-processive fashion and is not correctly sequenced or base-called. The resulting the sequence data is thus truncated at the 5’ end. With the 5’ adapter sequence ligated, this process does not happen until the end of the 44-mer adapter, thus sequencing through the transcription start site of the biological molecule. A comparison of the read alignments of identified full-length transcripts from our adapter tagged datasets with counterpart alignments in the untagged datasets reveals that reads that appear to be full-length isoforms in untagged DRS data are the missing ~13bp of their 5’ sequence (Figure 2). All six of these datasets also include the synthetic ERCC spike-ins; a set of 92 synthetic RNA molecules with ~25mer poly-A tails and no methyl-cap that are added to the total RNA at a range of concentrations. Due to the lack of a 5’ methyl cap these synthetic RNA should not undergo adapter ligation in two adapter ligation datasets, however we leverage the absolutely known sequence (including the 5’ end) of the molecules to validate the length of the missing 5’ ends of ONT DRS reads. An examination of the reads mapping to the Rubisco Activase locus (RCA, AT2G39730, chr2:16570746-16573692 rev) reveals that 33% (599) of the read alignments from the adapter-tagged datasets pass the stringent filtering for adapter identification. Of the remaining reads alignments in the region ~52% (622) have 5’ end positions that are in good agreement with the TSS positions identified by reads that pass the stringent filtering criteria. This suggests that these reads contain the presence of the adapter sequence and are full length mRNAs. The remaining ~48% are consistent with 5’ ends of alignments from our datasets that do not contain the adapter, suggesting that these molecules likewise do not contain the adapter sequence. Comparing the combined counts of reads that contain the adapter sequence (both passing the stringent filtering and those with coincident 5’ end positions that do not pass the filter), with those without evidence for the presence of a 5’ adapter sequence yields an estimate of >60% as the combined result of the adapter tagging efficiency and the pulldown enrichment chemistry for sequencing full-length transcripts using this approach.

For the first time, the 5’ adapter-tagged datasets also enable us to identify novel short unannotated isoforms from DRS sequence fragments that are unidentifiable in datasets processed with the standard DRS protocol and would be overlooked by purely computational isoform identification methods (such as FLAIR [10] and Mandalorian [https://github.com/christopher-vollmers/Mandalorion-Episode-II]) (Figure 3).

### Reliable identification of 3’ poly-A sites

ONT DRS sequenced transcripts from 3’ -> 5’ and the standard protocol relies on the presence of a poly-A tail in order to ligate the adapters and the sequencing motor protein to the biological molecule. As such, we expect all ONT reads to include a poly-A tail sequence that will not align to the genome immediately downstream of the aligned sequence and, thus, that the aligned sequence will reliably identify the poly-A sites of sequenced transcripts. Sequencing long homopolymer runs remains a challenge for ONT sequencing technology (both DNA & RNA), and this impacts the ability of accurately base-call downstream sequences. In order to assess the ability of ONT DRS read alignments to accurately identify 3’ poly-A sites we compared the position of the 3’ ends of ONT DRS read alignments supported by 3 or more reads with polyadenylation sites identified with bp-level resolution from Helicos Bio DRS data [12], genome wide. The median distance between ONT DRS 3’ ends and Helicos Bio polydenylation sites is bp (1 standard deviation), demonstrating that the overall positional agreement between these orthogonal datasets is excellent. The spread of this distribution is reflects the inherent positional accuracy of the ONT DRS data due to the high position accuracy of the Helicos Bio polydenylation position identifications (±1 bp).

As we expect from the chemistry of the ONT DRS library preparation, our DRS datasets show no strong evidence for read alignments resulting from the ONT DRS adapters binding to homopolymer A runs occurring within the main body of the sequenced isoforms. To validate this, we first identified 10,818 homopolymer A runs with a length >6 that occur within the coding sequences of isoforms in Araport 11 (hereafter pA6; the length threshold was chosen to match the threshold used in [12] to assess the incidence of internal priming in Helicos Bio datasets). We compare the 3’ end position of the read alignments for each ONT DRS read in our datasets with these genomic positions allowing a +/- 20bp window to account for the positional accuracy of 3’ end identification in ONT DRS data (see Section 2.2.2). Of the pA6 regions, only 2.3% (248) were coincident with more than three reads across all six datasets, 79% (n=197) of which were detected in just a single replicate and only 8% of which (n=20) were identified in four or more datasets. Of the these, 9 are supported by Helico Bio data (>3 reads), 7 of these are found in the terminal exon of the CDS annotation, suggesting that these may represent new sites of transcription termination.

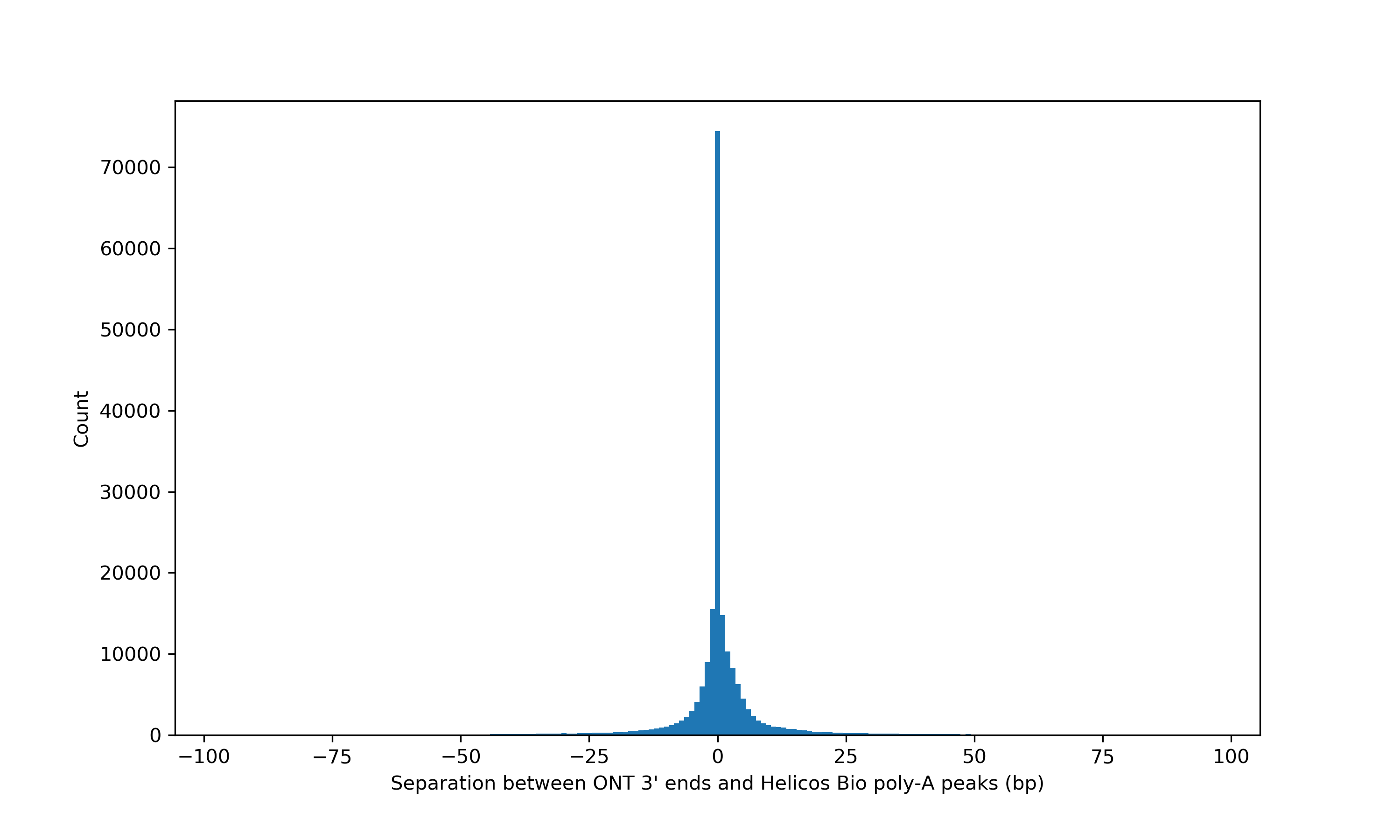


Figure 5: Separation between the polyadenylations sites identified from the 3’ end positions of ONT DRS read alignments and the nearest polyadenylation sites identified from published Helicos Bio DRS data [12].

### 2.4 Capturing alternative splicing with ONT DRS data

Alternative splicing is a characteristic feature of Eukaryotic transcriptomes and is thus an important aspect of biology for RNA sequencing techniques to be able to successfully capture. Uniquely, long read sequencing technology has the capacity to capture splicing patterns across the whole RNA molecule and link these patters to transcription initiation and polyadenylation (see, for example, [6, 13]). This decisive advantage over current short-read sequencing technology can only be fully realized, however, if individual splicing events in long read alignments can be relied upon. Each of our DRS datasets identifies more than 130k unique splicing events (Figure 6, green bars), capturing a similar level of transcriptomic complexity as existing gold-standard Arabidopsis annotations and supporting Illumina short-read sequencing datasets.

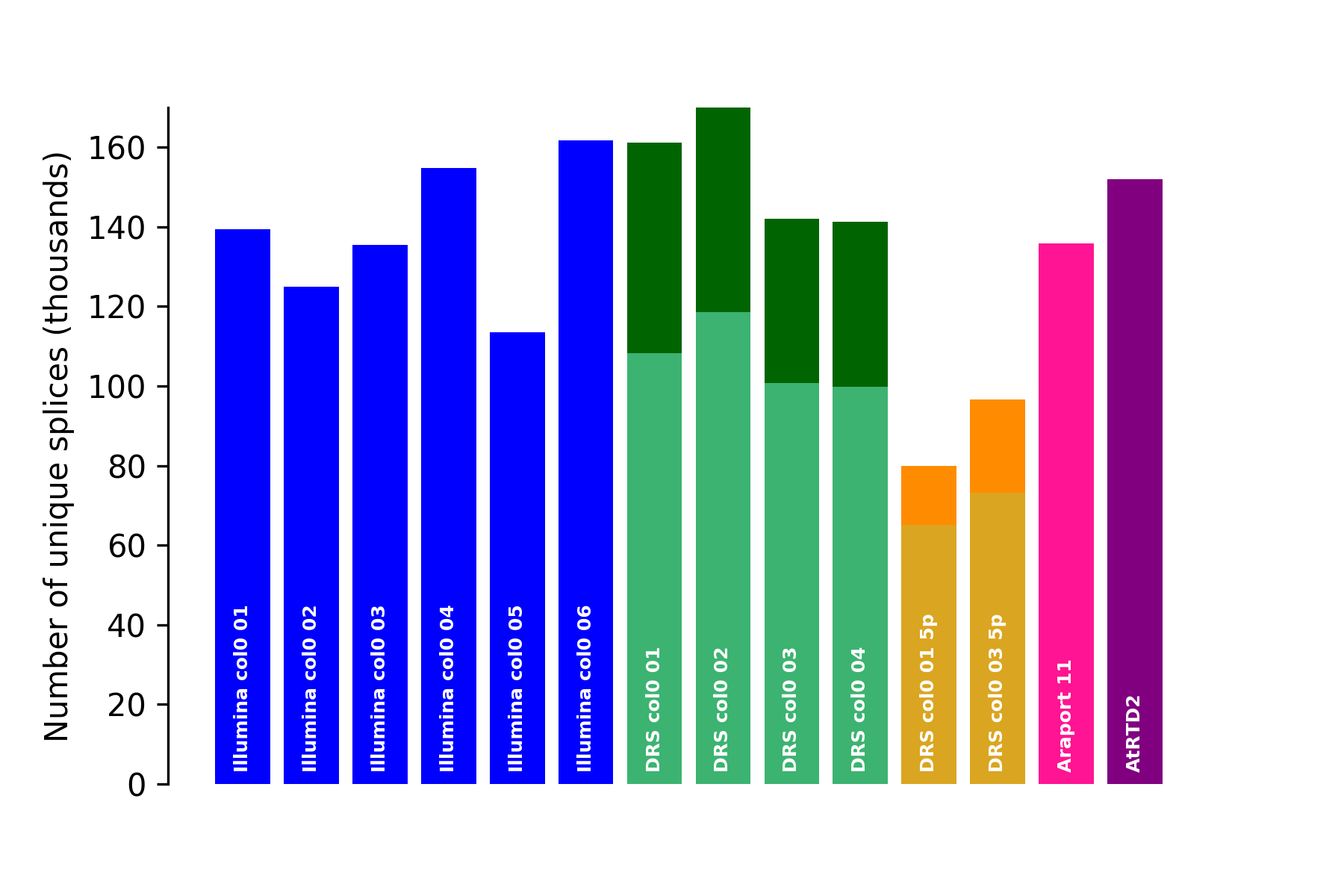


Figure 6: A comparison of the number of unique splicing events detected in various datasets. Blue: Illumina paired-end poly-A col-0 biological replicates. Green: ONT DRS col-0 biological replicates. The number of splices is shown before error correction (dark green), and after error-correction (light green). Orange: ONT DRS col-0 5’ adapter tagged replicates. The number of splices is shown before error correction (dark orange), and after error-correction (light orange). Pink: Araport 11. Purple: AtRTD2.

For unique splicing events detected in the ONT DRS datasets, however, the relatively high (~8%, *c.f.* <<1% for Illumina datasets) error-rate of the read alignments represents a significant potential impediment to reliable splicing identification. Several existing tools address this issue by using a pre-existing annotation and splice junctions identified from short-read Illumina data to ‘correct’ the read alignments to match the reference ([14, 15]). The principle shortcoming of this approach is that it overwrites the original splicing information in the data with nearby splice junctions found in existing reference annotations and/or alternative sequencing datasets. This removes potentially novel splicing from the ONT DRS dataset, preserving any biases present in the pre-existing data that is employed, and relies on the existence of extensive alternative sequencing data. An alternative approach is to correct the sequencing errors in the underlying sequencing data (see Section 2.2). Applying *proovread* to error-correct the DRS reads before their alignment results a reduction in the number of unique splicing events detected in the DRS data.



Figure 8

Here we take a different approach, preferring to classify the splicing events observed in the ONT DRS data and to use these classifications in context with any existing annotations and alternative data, to identify likely novel splice isoforms. We parse the ONT DRS read alignments for splicing events, classifying them as either canonical (GUAG-motif) or non-canonical. We then compare the detected splicing events against existing reference annotations, classifying each splicing event as annotation or unannotated. Finally, we also compare the ONT DRS splices against the unique splice sites detected across seven wild-type Col-0 *Arabidopsis thaliana* Illumina ribominus (biological replicate) datasets from Froussios *et. al.* (2017, [16]).

Combined, the four biological replicate ONT DRS datasets capture ~78% & ~73% of the Araport 11 and AtRTD2 annotations, respectively, including 9,000 unique splicing events that exist in the ONT DRS data and at least one of the annotations, but that are not detected in the deep Illumina data (Figure 7, gold sets). These include 1,237 splices that are detected in only one of the annotations, along with an additional 9,622 unique splices that the detected in both ONT DRS and Illumina datasets (Figure 7, blue set), but not in either reference annotation, highlighting the pitfalls of discarding splicing events based on using combinations of existing reference annotations or pre-existing orthogonal data. The majority of the splicing events detected in the ONT DRS data (56.4%), are not supported by either the Illumina data or the reference annotations. Whilst it is likely that some of these represent real biological molecules, many may be the result of incorrect alignment due to sequencing errors. In particular, the large number of non-canonical splice sites detected in the ONT DRS data are unlikely to represent real undiscovered biology given the well-studies nature of Arabidopsis thaliana. We observe a similar situation in the Illumina data, where 59% of the detected splice sites are unsupported in with the ONT DRS data or in the reference annotations, the majority of which are non-canonical splice sites. Mourao *et. al.* (2018, [17]) show that dUTP-based strand-specific Illumina libraries often show low-levels of incorrect stranded-ness for short read alignments, that is characterised by antisense splicing patterns that match the expected sense strand splicing. The Illumina data used here forms part of the data used in Mourao *et. al.* (2018, [17]) and show levels of spurious antisense expression around 1%. Importantly, any spliced spurious antisense short read alignments will, by definition, result in spurious novel non-canonical splicing events, explaining the large number of unique novel non-canonical splicing events identified here in this Illumina dataset.

In order to identify candidate novel splicing events (CSNEs) from our ONT DRS data we further refine these splicing event classifications in two ways. Firstly, for splice sites classified as unannotated, we examine the immediate sequence neighbourhood (+/-10bp) around the splice acceptor/donor sites for nearby canonical di-nucleotide splicing motifs, classifying these alternatives as annotated or novel according to the reference annotations. The presence of annotated canonical splicing events or, for novel non-canonical splices, a nearby novel canonical splicing motif, suggests that a sequencing and/or alignment error may be driving an incorrect read alignment. Conversely, the absence of any nearby annotated alternative or, for novel non-canonical splices, a nearby novel canonical splice motif, lends weight to the identification of these as CSNEs. We then classify each identified, and nearby alternative, splicing event as either U2 or U12 snRNP-dependent (hereafter U2 & U12, respectively), using the *Arabidopsis thaliana* -specific position weight matrices (PWMs) from Sheth *et. al.* (2006, [18]). In *Arabidopsis thaliana*, U2 snRNP-dependent splicing is the predominant classification, representing splicing events where the intron is excised by splicosomes containing the U1,2,4,5,& 6 snRNPs. These typically have the terminal splice dinucleotide motifs GT-AG, GC-AG & AT-AC. U12 snRNP-dependent splicing is a minor class of splicing events where the intron is excised by splicosomes containing the U11,12,4,5,& 6 snRNPs, and typically have the terminal splice dinucleotide motifs AT-AC & GT-AG. We classify a splice as U2(GT\_AC)/U2(GC-AG)/U12(GT-AG)/U12(ATAC) only if the individual donor and acceptor sites are both classified in the same category by the PWMs. Again, strong U2/U12 classifications lend weight to the identification of potential CSNEs.

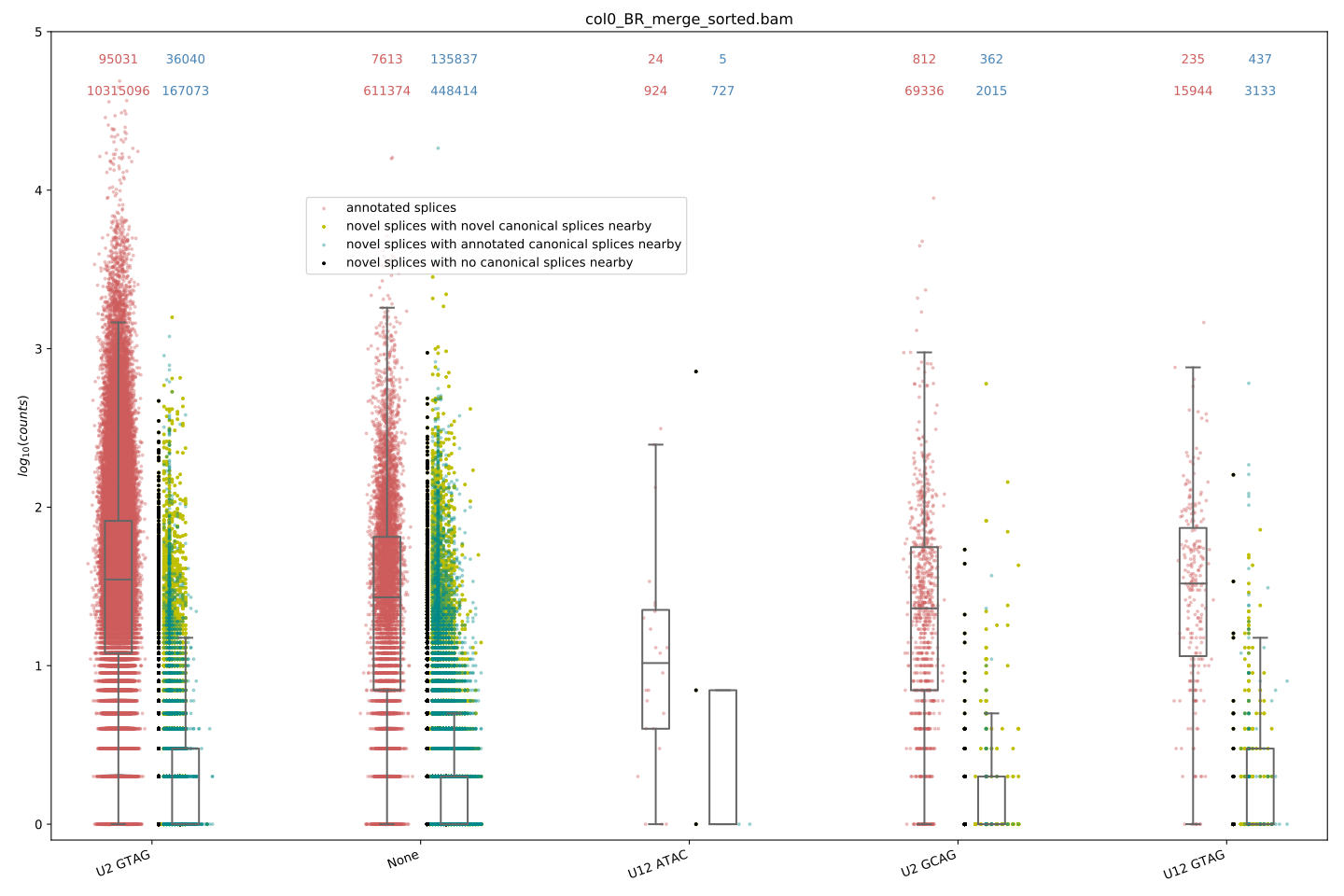


Figure 9

92.6% of the detected Araport splice sites (and 91.3%of the AtRTD2 splice sites) are categorized by the PWMs.

Large majority of unannotated splice sites in each category and low expression (95% confidence intervals for all the data distributions are 5 counts or fewer), whereas the annotated splices have a median expression of >10 counts in all categories.

### DRS datasets show no evidence for internal priming at expressed poly-A sequences

As we expect from the chemistry of the ONT DRS library preparation, our DRS datasets show no strong evidence for read alignments resulting from the ONT DRS adapters binding to homopolymer A runs occurring within the main body of the sequenced isoforms. To validate this, we first identified 10,818 homopolymer A runs with a length >6 that occur within the coding sequences of isoforms in Araport 11 (hereafter pA6; the length threshold was chosen to match the threshold used in [12] to assess the incidence of internal priming in Helicos Bio datasets). We compare the 3’ end position of the read alignments for each ONT DRS read in our datasets with these genomic positions allowing a +/- 20bp window to account for the positional accuracy of 3’ end identification in ONT DRS data (see Section 2.2.2). Of the pA6 regions, only 2.1% (229) were coincident with more than three reads across our datasets, 70% (n=160) of which were detected in just a single replicate and only 7% of which (n=16) were identified in all four biological replicates. Of the pA6 regions detected in all four replicates, 6 are supported by Helico Bio data (>3 reads) and 5 of these are found in the terminal exon of the CDS annotation, suggesting that these are likely to be new sites of transcription termination.

### DRS datasets show no evidence for spurious antisense signal

The strand-specificity of the ONT DRS reads is essentially perfect, as we might expect from the lack of PCR or RT steps in the ONT DRS library preparation.

26,469 reads from across the 16 datasets align to the spike-ins. Only one of which aligns to the incorrect strand.

### Only a small fraction of true full-length reads are sequenced by the ONT DRS library preparation

### ONT DRS data are quantitative

### Detection of novel isoforms in FLM by ONT DRS Data

### DRS data can detect rare antisense lincRNAs

1.5 modification detection by virilizer comparison.

## Discussion

### Sample specific transcriptome annotation with ONT DRS data

## Conclusions

## Bibliography

## Methods

### ONT DRS poly-A Arabidopsis thaliana sample and library preparation

### ONT DRS 5’ adapter ligation

### ONT DRS data processing pipeline

### Base-calling

### Alignment

### Quantifying ONT DRS bp sequencing/alignment errors

### Measuring gene and transcript expression

### Classifying splicing events

### Comparison of 3’ ends to Helicos Bio Data

### Identifying ONT DRS reads with 5’ adapters

### Illumina strand-specific ribo-minus sample and library preparation

### Illumina data processing and analysis

Supplimentary:

1. Guigo, R., et al., *EGASP: the human ENCODE Genome Annotation Assessment Project.* Genome Biol, 2006. **7 Suppl 1**: p. S2 1-31.

2. Zerbino, D.R., et al., *Ensembl 2018.* Nucleic Acids Res, 2018. **46**(D1): p. D754-D761.

3. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome biology, 2006. **7 Suppl 1**: p. S4 1-9.

4. Moreton, J., A. Izquierdo, and R.D. Emes, *Assembly, Assessment, and Availability of De novo Generated Eukaryotic Transcriptomes.* Front Genet, 2015. **6**: p. 361.

5. Smith-Unna, R., et al., *TransRate: reference-free quality assessment of de novo transcriptome assemblies.* Genome Res, 2016. **26**(8): p. 1134-44.

6. Anvar, S.Y., et al., *Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing.* Genome Biol, 2018. **19**(1): p. 46.

7. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores.* Nat Methods, 2018. **15**(3): p. 201-206.

8. Jiang, L., et al., *Synthetic spike-in standards for RNA-seq experiments.* Genome Res, 2011. **21**(9): p. 1543-51.

9. Förster, F., et al., *proovread : large-scale high-accuracy PacBio correction through iterative short read consensus.* Bioinformatics, 2014. **30**(21): p. 3004-3011.

10. Tang, A.D., et al., *Full-length transcript characterization of <em>SF3B1</em> mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.* bioRxiv, 2018: p. 410183.

11. Morton, T., et al., *Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures.* Plant Cell, 2014. **26**(7): p. 2746-60.

12. Sherstnev, A., et al., *Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation.* Nature structural & molecular biology, 2012. **19**(8): p. 845-52.

13. Clark, M., et al., *Long-read sequencing reveals the splicing profile of the calcium channel gene CACNA1C in human brain.* bioRxiv, 2018.

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

15. Wyman, D. and A. Mortazavi, *TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts.* Bioinformatics, 2018. **35**(2): p. 340-342.

16. Froussios, K., et al., *How well do RNA-Seq differential gene expression tools perform in a eukaryote with a complex transcriptome?* bioRxiv, 2017.

17. Mourao, K., et al., *Detection and Mitigation of Spurious Antisense RNA-seq Reads with RoSA.* bioRxiv, 2018.

18. Sheth, N., et al., *Comprehensive splice-site analysis using comparative genomics.* Nucleic Acids Research, 2006. **34**(14): p. 3955-3967.