BaRTv2

Coulter, Max; Entizne, Juan Carlos; Guo, Wenbin; Bayer, Micha; Wonneberger, Ronja; Milne, Linda

Published in:
Plant Journal

DOI:
10.1111/tpj.15871

Publication date:
2022

the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) doi: 10.1111/tpj.15871

RESOURCE

BaRTv2: a highly resolved barley reference transcriptome for accurate transcript-specific RNA-seq quantification

Max Coulter[1], Juan Carlos Entizne[1], Wenbin Guo[2], Micha Bayer[2], Ronja Wonneberger[3], Linda Milne[2], Miriam Schreiber[1], Allison Haaning[4], Gary J. Muehlbauer[4], Nicola McCallum[5], John Fuller[5], Craig Simpson[5], Nils Stein[3,6], John W. S. Brown[1,5], Robbie Waugh[1,5,7] and Runxuan Zhang[2,*]

[1]*Division of Plant Sciences, University of Dundee, James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK,*
[2]*Information and Computational Sciences, James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK,*
[3]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3,D-06466 Stadt Seeland, Germany,*
[4]*Department of Agronomy and Plant Genetics, University of Minnesota, 1991 Upper Buford Circle, 542 Borlaug Hall, St Paul, Minnesota 55108, USA,*
[5]*Cell and Molecular Sciences, James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK,*
[6]*Center for Integrated Breeding Research (CiBreed), Georg-August-University, Göttingen, Germany, and*
[7]*School of Agriculture and Wine & Waite Research Institute, University of Adelaide, Waite Campus, Glen Osmond, South Australia 5064, Australia*

Received 18 October 2021; revised 2 May 2022; accepted 9 June 2022
*For correspondence (e-mail runxuan.zhang@hutton.ac.uk).
The author(s) responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plcell/pages/General-Instructions) is: Runxuan Zhang (runxuan.zhang@hutton.ac.uk)

SUMMARY

Accurate characterisation of splice junctions (SJs) as well as transcription start and end sites in reference transcriptomes allows precise quantification of transcripts from RNA-seq data, and enables detailed investigations of transcriptional and post-transcriptional regulation. Using novel computational methods and a combination of PacBio Iso-seq and Illumina short-read sequences from 20 diverse tissues and conditions, we generated a comprehensive and highly resolved barley reference transcript dataset from the European 2-row spring barley cultivar Barke (BaRTv2.18). Stringent and thorough filtering was carried out to maintain the quality and accuracy of the SJs and transcript start and end sites. BaRTv2.18 shows increased transcript diversity and completeness compared with an earlier version, BaRTv1.0. The accuracy of transcript level quantification, SJs and transcript start and end sites have been validated extensively using parallel technologies and analysis, including high-resolution reverse transcriptase-polymerase chain reaction and 5′-RACE. BaRTv2.18 contains 39 434 genes and 148 260 transcripts, representing the most comprehensive and resolved reference transcriptome in barley to date. It provides an important and high-quality resource for advanced transcriptomic analyses, including both transcriptional and post-transcriptional regulation, with exceptional resolution and precision.

Keywords: barley, RNA-seq, Iso-seq, reference transcript datasets, transcriptomic analysis.

INTRODUCTION

Barley, the fourth most important cereal crop worldwide, is cultivated across a wide range of environments, and is used widely in food and drink for humans and feed for animals (Dawson et al., 2015). Abundant genetic diversity reflects differences in gene expression, whereby developmental, environmental, biotic and abiotic stresses lead to changes in the transcriptome, promoting downstream protein and regulatory responses from the plant (Barakate et al., 2020; Calixto et al., 2016; Kintlová et al., 2017; Ren et al., 2018). Such changes occur at the individual transcript level as well as at the gene level, with alternative promotor usage, alternative splicing (AS) and alternative polyadenylation leading to an increased transcript repertoire (Baralle & Giudice, 2017; Brown et al., 2017; Laloum et al., 2018; Wang et al., 2019).

Detecting changes in gene expression has become routine in plant research using RNA-seq, while identifying transcript variants that are the product of AS is a more recent development (Calixto et al., 2018; Calixto et al., 2019). A variety of methods exist for RNA-seq data analysis, but all require as part of the analysis pipeline either the assembly of transcripts from the dataset (Kim et al., 2013; Maretty et al., 2014; Pertea et al., 2015) or an existing reference transcriptome for gene/transcript quantification. Quantification of transcripts with a reference transcript dataset (RTD) can be achieved with speed and precision using established pseudo-alignment programs such as Kallisto and Salmon (Bray et al., 2016; Patro et al., 2017). However, the precision of transcript expression quantification is only as good as the RTD itself, and such resources are currently not available for most plant species (Brown et al., 2017). The impact of transcript references on quantification accuracy is reflected not only at the transcript-level but also at the gene-level (Soneson et al., 2015; Yi et al., 2018). Thus, comprehensive, high-quality RTDs are key resources for quantitative analysis of gene expression.

Considerable progress has been made in developing plant RTDs using short-read data; first in Arabidopsis (R., Zhang et al., 2015a; Zhang et al., 2017), and then barley (Rapazote-Flores et al., 2019). The approaches adopted for their construction were largely based around the accurate detection of splice junctions (SJs) along with the implementation of assembly pipelines that incorporate stringent quality control filters to remove redundant, poorly assembled or fragmentary transcripts (Zhang et al., 2017). During the construction of Arabidopsis AtRTD2, variation in 5′- and 3′-UTR lengths among different transcripts from the same gene were found to skew subsequent transcript quantifications (Zhang et al., 2017). A method for extending transcript ends to the longest transcript in its gene ('padding') resulted in an overall improvement in quantification accuracy (Zhang et al., 2017), and so was also applied in the creation of an initial barley reference transcriptome BaRTv1.0-QUASI (Rapazote-Flores et al., 2019). However, close inspection revealed that despite these improvements, incorrect transcripts are still present, and these can be generally attributed to the fact that short-read data does not accurately capture transcript 5′- and 3′-ends or phasing of different AS events (Brown et al., 2017; Mao et al., 2020).

Single-molecule sequencing on Pacbio Iso-seq or Oxford Nanopore platforms should in principle overcome issues with 5′- and 3′-end determination and phasing of AS events. Iso-seq has been used to sequence cDNAs of up to 10 kb and has been applied to a number of crop species including maize (Wang et al., 2016), sorghum (Abdel-Ghany et al., 2016), cotton (Wang et al., 2018; Wang et al., 2019), rubber (Chow et al., 2019) and tea (Chen et al., 2020) crops. However, depending on the number of sequencing cycles, some Iso-seq reads suffer from a high sequencing error rate. Other issues include incomplete gene and transcript coverage and high cost per base (Rhoads & Au, 2015), particularly for data generated from the Sequel platform. Sequencing errors have been addressed by self-correction (Salmela et al., 2017) or hybrid-correction methods (Au et al., 2012; Hackl et al., 2014; Salmela & Rivals, 2014). While these reduce overall error, incorrect clustering of transcripts during self-correction and mis-mapping during hybrid correction can cause over-correction resulting in loss of real transcripts with small AS events, incorrect SJ determination and generation of new, false SJs and transcripts (Kuo et al., 2020). Major challenges remain in accurately determining SJs as well as the start and end sites of transcripts.

Defining the 5′- and 3′-ends of transcripts is important in order to identify alternative promotor usage and alternative poly(A) sites. Both of these features are prevalent across biology and add another layer of complexity to regulation of gene expression (Mejía-Guerra et al., 2015; Morton et al., 2014; Wang et al., 2019). In Arabidopsis, alternative transcription start sites (TSS) lead to changes in the subcellular localisation of proteins (Ushijima et al., 2017), while alternative poly(A) selection has a role in the regulation of processes such as flowering (Y., Zhang et al., 2015b). Cap analysis gene expression sequencing (CAGE-Seq) has been used to identify TSS in both Arabidopsis and cotton (Morton et al., 2014; Wang et al., 2019), while polyA-Seq has been used to identify widespread alternative poly(A) sites in the latter (Wang et al., 2019). Nanopore sequencing has also been used to identify alternative poly(A) sites in Arabidopsis (Parker et al., 2020; Zhang et al., 2020).

Here we present a new barley reference transcriptome based on the European 2-row spring barley cultivar Barke, making use of its new and accurate TRITEX assembled genome as a reference (Jayakodi et al., 2020). The new transcriptome combines filtered full-length transcripts from PacBio Iso-seq with stringently filtered Illumina RNA-seq based transcripts. Novel computational methods were used to determine accurate SJs, TSS and transcription end sites (TES), which were utilised to generate an Iso-seq-only dataset (BaRT2.0-Iso). In parallel, transcripts from Illumina sequencing of the same barley samples were assembled and stringently filtered to produce a short-read transcript assembly (BaRT2.0-Illumina). Finally, the Iso-seq dataset was supplemented by high-quality transcripts from the short-read assembly to overcome the incomplete coverage of some genes in the Iso-seq dataset to generate a new barley RTD (BaRTv2.18). Currently BaRTv2.18 comprises 39 434 genes and 148 260 transcripts. High-resolution (HR) reverse transcriptase-polymerase chain reaction (RT-PCR) and a variety of

benchmarks demonstrated a significant improvement in BaRTv2.18 compared with BaRTv1.0. Finally, a BaRTv2.18 transcript annotation is provided that includes coding regions and protein variants as well as unproductive transcripts together with functional annotations.

RESULTS

Construction of the BaRTv2.0 Iso-seq transcriptome (BaRT2.0-Iso)

To maximise the number and diversity of transcripts, 21 samples from a range of tissues, growth stages and treatments were used (Supplementary Table S1). The workflow for the Iso-seq data processing is described in Figure 1. Briefly, Iso-seq reads were pre-processed to full-length non-chimaeric reads (FLNC) and mapped to the Barke genome before being filtered against high-confidence (HC) SJ, TSS and TES datasets to generate the Barke Iso-seq transcriptome, BaRTv2.0-Iso (Figure 1a). This latter step is described in more detail in Figure 2. In parallel, Illumina reads were generated from the 20 samples, pre-processed and transcripts assembled with three different assemblers (Figure 1b). Assemblies were merged and processed to give the short-read assembled transcriptome, BaRTv2.0-Illumina (Figure 1b). Finally, the Iso-seq and Illumina datasets were merged and filtered to remove redundancy and prioritise Iso-seq transcripts over Illumina transcripts. The resultant RTD, BaRTv2.10 underwent a comprehensive series of final filtering steps to, for example, remove low-expressed mono-exonic transcripts to give the final barley (Barke) RTD, BaRTv2.18. The production of BaRTv2.18, generation of HC SJ, TSS and TES datasets and the quality control steps are described below.

For Iso-seq, the first stage processed raw Iso-seq reads with the PacBio Iso-seq 3 pipeline to yield a combined total of 8 113 088 CCS reads with a mean of 405 654 reads per sample. After demultiplexing, removal of polyA tails and concatemers, a total of 7 395 557 FLNC reads were generated of which 93.7% (6930934) mapped to the Barke genome with a mean of 346 546 mapped FLNC reads per sample (Figure 1a; Table S5). For the second stage of the analysis we used the TAMA suite of programs (Kuo et al., 2020). 'TAMA collapse' integrated redundant transcript models in each sample providing 20 initial annotation files. These were then merged with TAMA merge (merges multiple transcriptomes while maintaining source information) to create an unfiltered transcripts dataset containing a total of 33 550 genes and 2 004 544 transcripts with an average of 15 733 genes and 166 048 transcripts per sample (Figure 1a; Table S5). At this stage transcripts were defined at 1 bp resolution (i.e. any small differences in mapping between two reads would lead to two transcripts being annotated).

Determination of HC SJs, TSS and TES

Major challenges remain in the accurate definition of SJs and TSS and TES due to sequencing errors and degradation of mRNAs *in vivo* or during processing. We applied newly developed methods of Iso-seq analysis, which use sets of HC SJs and TSS/TES to define these key features (Kuo et al., 2020; Zhang et al., 2022). Some FLNC reads still have relatively high error rates (about 3%; Kuo et al., 2020). Mapping FLNCs to the genome can be inaccurate due to the phenomenon of edge wander (Holmes & Durbin, 1998) where inaccuracies in defining the boundaries of a gap are the main source of error, for example when aligning a sequence with spliced intron to a genomic reference. Mis-mapping of reads containing spliced introns is exacerbated by local sequence errors ($\pm$ 10 nt) that frequently generate false SJs leading to transcripts with incorrect intron–exon boundaries (Zhang et al., 2022). To address this issue, a HC SJ dataset was created based on both Iso-seq and short-read RNA-seq data (see below). A total of 257 496 SJs were identified in the Iso-seq dataset, of which 164 860 were designated HC (64%; Table S6a). Of the 92 636 SJs that were filtered out, 31 541 contained mis-matches within 10 bp of a SJ, 2687 were template switching events, and the rest (58 408) had non-canonical splice site dinucleotides (Table S6b). Template switching events are an artefact of cDNA synthesis, which can lead to exonic sequence loss and incorrect SJ designation (Cocquet et al., 2006). Iso-seq transcripts that only contains SJs in the HC dataset are kept for downstream processing (Figures 1a and 2a).

Despite applying the Teloprime cap capture system to enrich for full-length Iso-seq reads, numerous 5′ and 3′ truncated reads were observed in our dataset that likely reflect degradation of transcripts either in the cell or during RNA preparation. As for SJs, datasets of HC TSS and TES sites were generated from the Iso-seq data (Figure 2b). An important consideration is that the read abundance of genes in the Iso-seq data covered a large dynamic range with, for example, over 8000 genes having only one or two FLNC reads, while the top 10% of expressed genes contained 79% of FLNC reads in the dataset. The identification of HC TSS/TES sites for high- and low-expressed genes presents very different issues. For example, for highly expressed genes, the major challenge is to reduce false end sites from relatively high numbers of transcript fragments (degradation products) while identifying the dominant, *bona fide* sites. For low-abundance genes, the problem is to obtain enough experimental evidence to define a TSS/TES.

We used two methods to determine TSS and TES depending on the level of expression. Firstly a binomial probability (Loader, 2000) was calculated for all genes in the dataset to identify TSS and TES sites that were

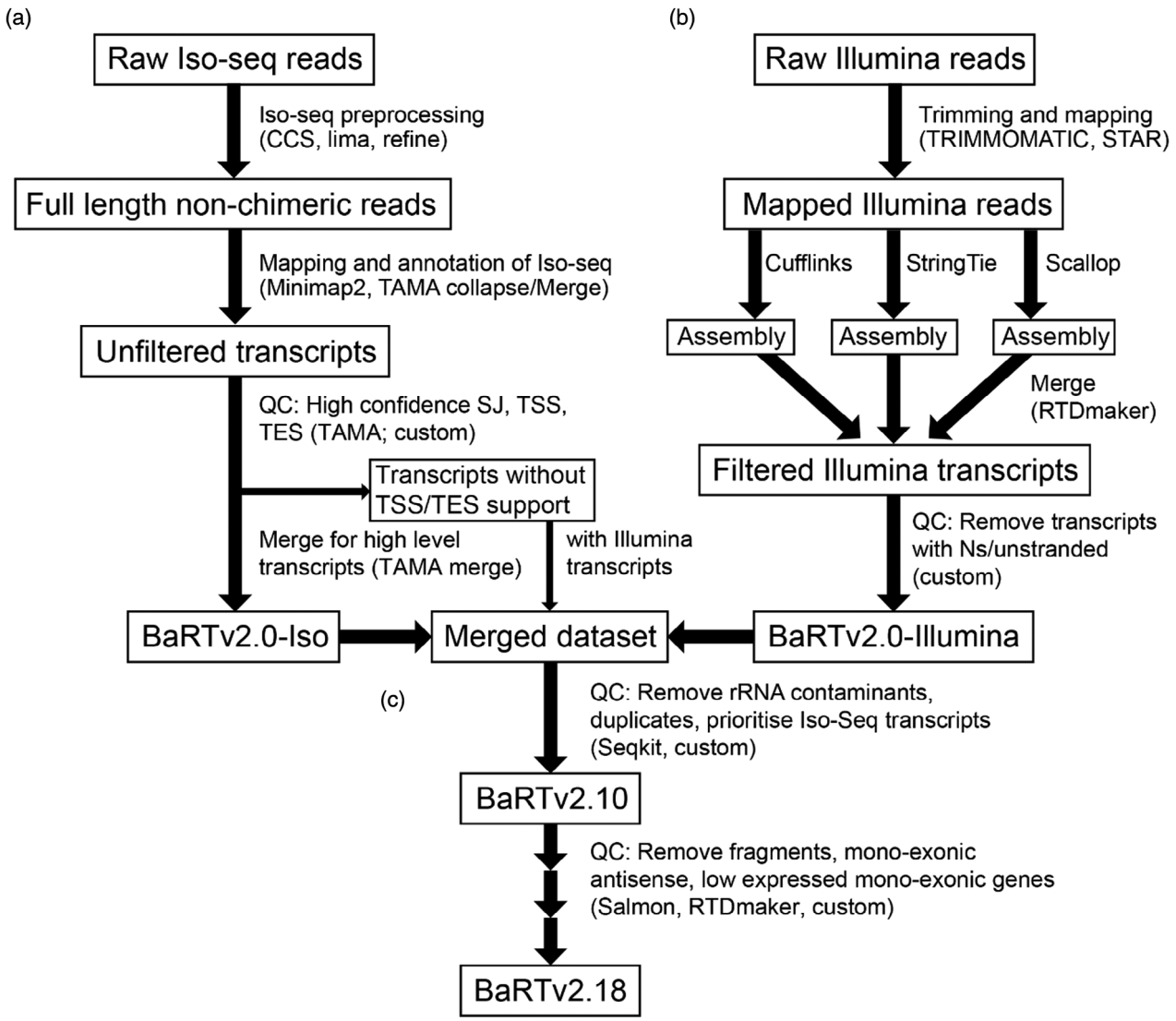


Figure 1. Outline of the pipeline used to generate BaRTv2.18.
Iso-seq and Illumina datasets were processed separately (a,b) and then combined at the final stage (c). Software and scripts are indicated in brackets, with 'custom' referring to custom code.
(a) For the Iso-seq dataset, raw subreads were pre-processed using Isoseq3 software (CCS, lima and refine) to create full-length non-chimeric reads. These reads were mapped to the Barke genome using Minimap2. Mapped reads were used as input for Transcriptome Annotation by Modular Algorithms (TAMA) collapse and TAMA merge to create a set of unfiltered transcripts. Filtering was carried out using custom codes described in Figure 2, to remove fragments, unsupported splice junctions (SJs) and transcription start sites (TSS)/transcription end sites (TES). Transcripts from genes with low read end support that had HC SJs were kept in a separate file for potential inclusion depending on similarity to Illumina transcripts. Redundant transcripts were removed using TAMA merge to create BaRTv2.0-Iso.
(b) Illumina reads were trimmed using Trimmomatic and mapped to the Barke genome using STAR. Transcripts were assembled using three separate assemblers: Cufflinks, Stringtie and Scallop. The resulting assemblies were merged and filtered using the RTDmaker software. The resulting annotation was further filtered to remove transcripts overlapping Ns and transcripts with no strand information to create the BaRTv2.0-Illumina dataset.
(c) The BaRTv2.0-Iso and BaRTv2.0-Illumina datasets were merged using TAMA merge (giving priority to BaRTv2.0-Iso based transcripts). Further filtering was carried out using custom code to remove redundant Illumina transcripts. Duplicate transcripts were removed using seqkit rmdup, as well as potential rRNA. The resulting transcriptome BaRTv2.10 went through further filtering steps with RTDmaker including removal of low-expressed mono-exon genes based on Salmon quantifications to generate the final transcriptome BaRTv2.18.

significantly enriched in read numbers relative to the total number of reads from the gene (Figure 3a). That is, *bona fide* TSS and TES will occur more frequently and will be significantly enriched, whereas the randomly distributed ends derived from degraded mRNAs are unlikely to be significantly enriched and can be removed. We found that many transcripts when compared had small variations in their 5′- or 3′-ends; such variation is likely to reflect the stochastic nature of most TSS/TES that show a distribution around a dominant site (Morton et al., 2014). To account

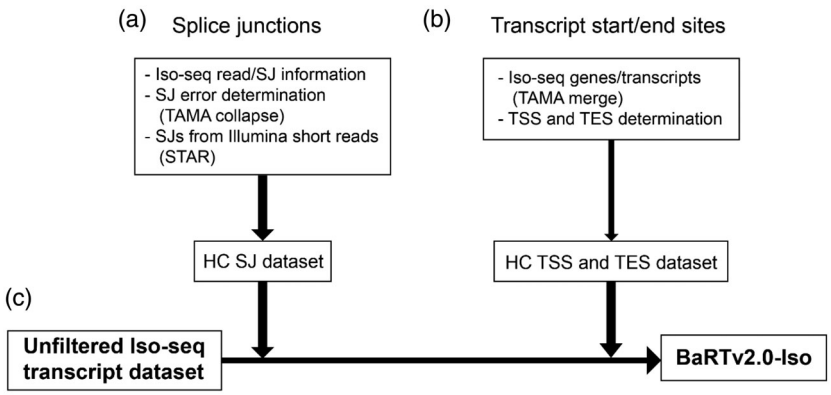


Figure 2. Overview of generation of BaRTv2.0-Iso using high-confidence (HC) splice junction (SJ), transition start sites (TSS) and transition end sites (TES) datasets.
(a) The HC dataset of SJs was generated from Iso-seq data using the error distribution profiles around SJs to remove false SJs and from Illumina SJ data.
(b) The HC TSS and TES dataset was generated from Iso-seq data using two methods dependent on read abundance.
(c) The unfiltered Iso-seq transcript dataset was filtered against the HC SJ, TSS and TES datasets, and only transcripts with HC information were retained in BaRTv2.0-Iso.

for this during determination of TSS and TES, we examined the distribution of read end sites around all end sites with high (> 10 reads) support. The majority of transcript 5′-ends occurred within a window ± 10 bp of the major site, while 3′-ends were more variable with the majority of them within a window of ± 30 bp (Supplementary Figure S1a,b). These windows were applied to allow for stochastic variation in transcript ends while removing transcripts with false ends (Figure 3a). A total of 43 302 and 59 944 significantly enriched TSS and TES, respectively, were identified in 14 589 genes (Table S7a).

For genes where the binomial probability approach failed to reveal enriched TSS/TES (normally due to the low expression of that gene), we require that at least two reads/transcripts had to have 5′- or 3′-ends within a window of ± 20 bp and ± 60 bp, respectively (Figure 3b). These windows were determined based on the distribution of ends around major end sites shown in Figure S1. By applying these criteria, HC TSS/TES were identified for a further 10 041 genes (Table S7a). Finally, the remaining 8739 genes had neither HC TSS nor TES sites predicted by either method. These genes had either a small number of mapped FLNC reads without similar ends or had only one FLNC read (Table S7b). As many of these genes had transcripts containing HC SJs, their transcripts were retained in a separate .bed file for comparison to Illumina short-read assembled transcripts to provide additional information on transcript features and, in some cases, were re-integrated into the RTD (Figure 1a; Table S7b).

Generation of Iso-seq transcripts in BaRT2.0-Iso

The datasets of HC SJs, HC TSS and TES were used to quality control the Iso-seq transcript dataset of 2 004 544 transcripts created using TAMA merge at single nucleotide resolution (Figures 1a and 2c). Only transcripts containing HC SJs, TSS and TES were retained giving 1 134 325 transcripts from 24 566 genes. Many of the differences among transcripts from the same gene were minor, with transcripts containing similar (but not identical) TSS and TES. To remove redundancy due to such variation in TSS and TES, transcripts were again run through TAMA merge. The settings 'm 0 -a 100 -z 100' ensured that transcripts with the same set of SJs and with similar 5′- and 3′-ends within 100 bp were merged together to the most abundant TSS and TES sites (Figure 3c). The final BaRTv2.0-Iso dataset contained 103 330 transcripts (reduced from 1 134 325) from 24 630 genes (Figure 1a).

Assembly of the Illumina transcriptome (BaRT2.0-Illumina)

Transcripts were assembled from alignments of the 20 short-read libraries (Figure 1b; Table S1). Illumina reads mapped by STAR from each library were assembled with three tools: Cufflinks (Trapnell et al., 2012); StringTie (Pertea et al., 2015); and Scallop (Shao & Kingsford, 2017), yielding one transcript assembly per sample per assembler (Figure 1b). The 60 individual assemblies were merged with RTDmaker (https://github.com/anonconda/RTDmaker) to give an Illumina transcript dataset. The amalgamated transcript assembly was quality controlled using RTDmaker, which is an automated and highly refined extension of the original protocol implemented to generate Arabidopsis AtRTD and AtRTD2 (R., Zhang et al., 2015a; Zhang et al., 2017). From the original total of > 3.6 M transcript models, 3 460 323 were rejected using the above criteria (Table S8). The resulting BaRT2.0-Illumina transcriptome consisted of 54 017 genes and 142 174 transcripts.

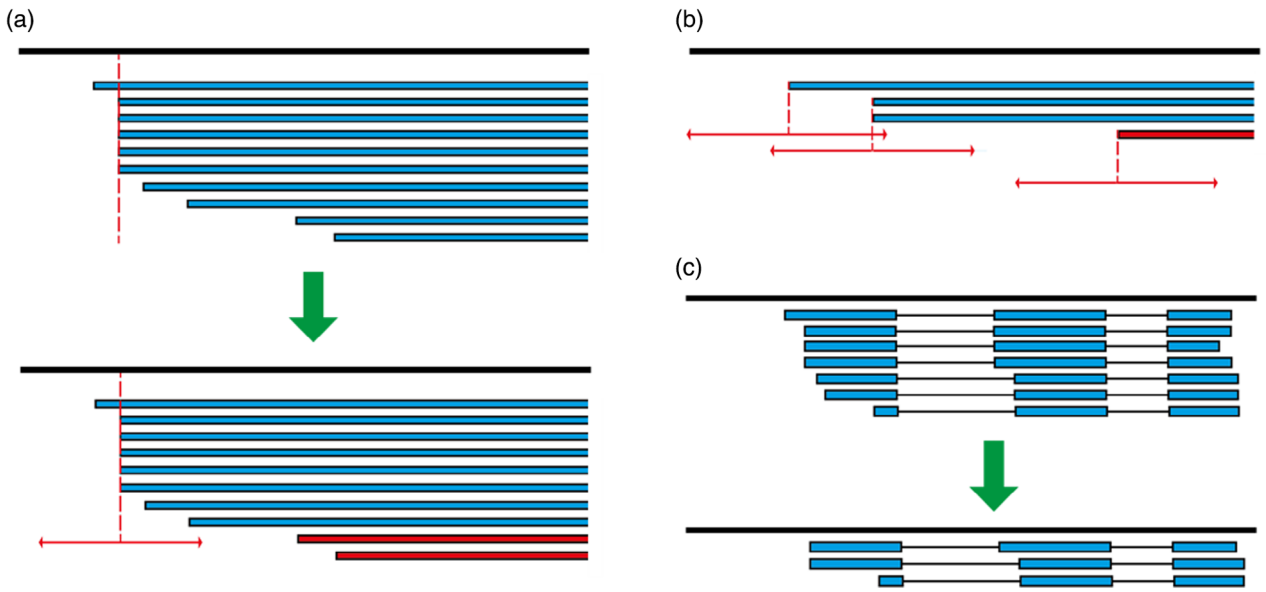


Figure 3. Schematic diagram of filtering and merging methods used to generate BaRTv2.0-Iso. Transcripts or transcript ends (blue) are shown aligned to the genome (black line).
(a) Binomial enrichment method: transcripts with significantly enriched end sites (top panel, dotted line) and other transcripts with similar end sites within a window (± 10 bp for TSS, ± 30 bp for TES sites; horizontal arrows) were kept (a, bottom panel). Transcripts with ends outside of the window from an enriched site were removed (red).
(b) For genes with lower Iso-seq coverage, the fixed window method was used where transcripts whose ends fell within a sliding window (± 20 bp for TSS, ± 60 bp for TES; horizontal arrows) were retained while those with unsupported ends were removed (red).
(c) Transcriptome Annotation by Modular Algorithms (TAMA) merge was used after Iso-seq filtering to merge transcripts with the same intron coordinates and similar ends (within ± 100 bp) to the length of the transcripts with highest abundance ends.

Merging Iso-seq and Illumina datasets

Gene, transcript and SJ identity of the BaRT2.0-Iso and BaRT2.0-Illumina transcriptomes were initially compared using TAMA merge with parameters where SJs had to be exact matches, and TSS/TES located within 100 bp. The overlap between Iso and Illumina genes was high, with 23 058 (95.1%) of the Iso-seq genes also present in Illumina data (Figure 4a). Only 1182 genes were unique to the Iso-seq dataset while 32 150 genes were unique to the Illumina assembly illustrating the lower gene coverage in the Iso-seq dataset. For SJs, 104 086 were common to both datasets (Figure 4b). The SJs unique to BaRT2.0-Illumina are largely from genes missing in the Iso-seq dataset. The 23 313 SJs unique to BaRT2.0-Iso potentially reflect novel transcript isoforms. In contrast to the SJ and gene overlaps, the transcript overlap was relatively low (Figure 4c). Only 7024 transcripts shared the same transcript structures between the Illumina and Iso-seq datasets. The transcript differences reflect 5′- and 3′-end variation, AS isoforms unique to one or other platform, transcript fragments and possible mis-assembly of some short-read transcripts. Many of these differences were resolved during merging and subsequent quality control. To examine 5′- and 3′-end differences, we identified transcripts with identical SJ sets in both BaRT2.0-Iso and Bart2.0-Illumina. Of the 41 079 transcripts, 34 055 (82.3%) had ends differing by > 100 bp, with only 7024 (Figure 4c) having ends within the 100-bp window. Thus, as expected, most of the defined TSS and TES in Iso-seq transcripts did not have matches among the short-read assembled transcripts.

In order to create a single RTD, the BaRTv2.0-Iso and BaRTv2.0-Illumina assemblies were merged and filtered using custom scripts to remove redundant transcripts, and: (a) prioritise Iso-seq transcripts as they contained HC SJs, TSS and TES; and (b) retain Illumina transcripts that represent genes absent from the Iso-seq dataset or which contained novel SJs (Figure 1c). This merging and filtering resulted in the development of an initial integrated RTD (BaRTv2.10), which contained a total of 56 029 genes and 173 635 transcripts. BaRTv2.10 contained an unexpectedly high number of mono-exonic genes including mono-exonic antisense genes, as well as some transcript fragments. The vast majority of the mono-exonic genes were derived from the short-read Illumina assembly, while some fragments and redundant transcripts still remained in both the Iso and Illumina assemblies. BaRTv2.10 was processed through RTDmaker in a series of steps to remove transcript fragments and mono-exonic genes with low expression. A high number of mono-exonic genes had low expression and were only observed in a single sample, likely due to

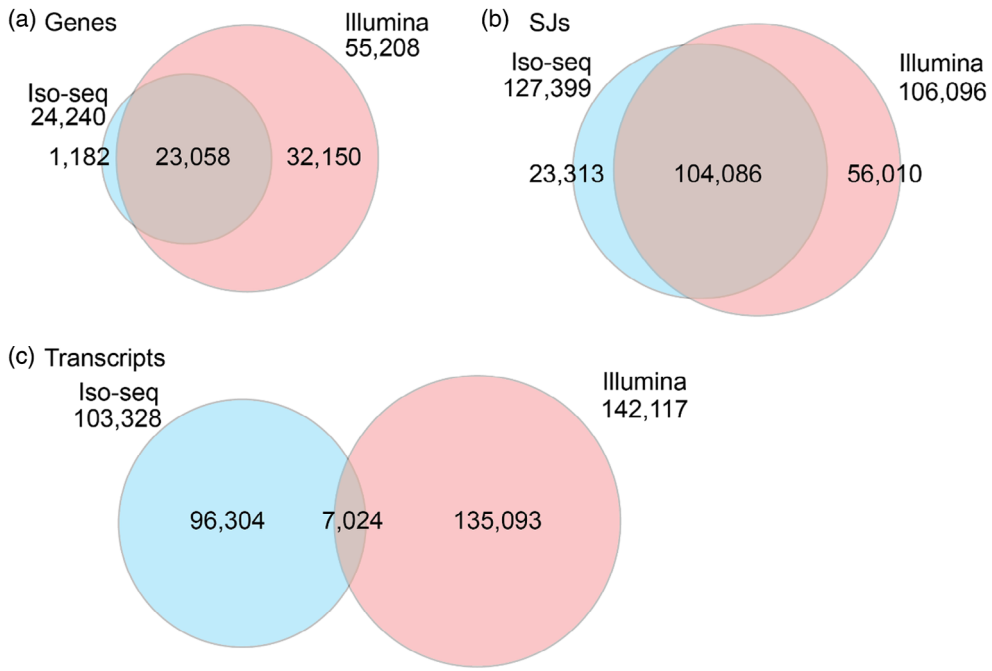


Figure 4. Venn diagrams showing overlap between genes, splice junctions (SJs) and transcripts produced by Iso-seq and Illumina.
(a) Genes, (b) SJs and (c) transcripts in the merged Iso-seq/Illumina dataset before filtering to create BaRTv2.18.

DNA contamination of samples. Thus, mono-exonic genes with expression detected in less than two samples were identified and removed unless they were protein-coding genes. The final barley RTD version, BaRTv2.18 consisted of 39 434 genes and 148 260 transcripts (Figure 1c; Table S9).

We also compared the frequency of different AS event types in BaRTv2.18 using SUPPA2 (Alamancos et al., 2015). The frequency of different AS events is similar to other plant species, with more alternative 3′ splicing events (24.8%) than alternative 5′ splicing events (16.4%), and relatively few exon skipping events (11.7%) (64). Intron retention (IR) is far more frequent in plants than in animals, with about 40% of plant AS events being IR (Marquez et al., 2012). IR events made up 41.5% of AS events in BaRTv2.18 (Table S10).

Characterisation of BaRTv2.18 genes and transcripts

The genes and transcripts in BaRTv2.18 were characterised using TranSuite (Entizne et al., 2020). TranSuite outputs detail on protein-coding and non-coding genes and transcripts, protein-coding and unproductive transcripts from protein-coding genes, a breakdown of genes by exon number (single versus multiple) and number of isoforms (single versus multiple), nonsense-mediated mRNA decay (NMD) predictions and translations of all transcripts in the RTD. These results are summarised in Figure 6, and Tables S9 and S11. Eighty-one percent of genes in BaRTv2.18 coded for proteins and 19% were non-protein-coding genes (Figure 6a; Table S9), with about 70% containing more than one exon and 30% single exon genes (Figure 6b; Table S9). Of the multi-exonic genes, 73% had more than one transcript isoform and 27% produced a single transcript (Figure 6c; Table S9). For protein-coding genes only, nearly 58% were multi-exonic with more than one transcript isoform and so are AS, agreeing with previous estimates of the prevalence of AS in Arabidopsis and other species (Chamala et al., 2015; Marquez et al., 2012; Zhang et al., 2017). The 7501 non-protein-coding genes generated 11 723 transcripts; 2160 genes were multi-exonic (spliced) but with a single transcript and 1179 genes were alternatively spliced (Table S9).

At the transcript level, 136 537 (92.1%) BaRTv2.18 transcripts came from protein-coding genes. Of these, 61.8% encoded protein isoforms while 38.2% were unproductive (Figure 6e; Table S11). AS transcripts that coded for proteins were divided into those where the AS events had little or no effect on the coding sequence (e.g. NAGNAG AS sites or AS in the UTR; 34.4%) and those that encoded protein variants (65.6%; Figure 6f; Table S11). NAGNAG AS events generate transcripts that code for protein variants differing by only one amino acid and transcripts of genes where AS events occur only in the 5′- and/or 3′-UTRs code for identical proteins. Breaking these down further (Figure 6g; Table S11), the most frequent AS events were in the 3′-UTR (42.3%), then 5′-UTR (37.3%) and NAGNAG events (10.4%), with lower numbers of combinations (Figure 6g). Across all transcripts, NAGNAG AS events were

present in 2.9%. Finally, the unproductive transcripts from protein-coding genes were classified by their NMD features: presence of a premature stop codon (PTC), downstream SJs, long 3′-UTR, or upstream ORF overlapping the authentic TSS (Kalyna et al., 2012; Figure 6h; Table S11); 67.2% of the unproductive transcripts contained the classical combination of NMD features of a PTC with downstream SJs and long 3′-UTRs, with a further 20.9% having one or two of these features. Overall, 13.3% of unproductive transcripts contained an overlapping uORF (Figure 6h; Table S11).

BaRTv2.18 represents a barley transcript dataset with improved integrity

To evaluate BaRTv2.18, six benchmarks were compared between BaRTv2.18 and BaRTv1.0 (Table S12). The highly conserved BUSCO gene set (Simão et al., 2015) was used to assess transcriptome completeness. BaRTv2.18 outperformed BaRTv1.0 in BUSCO benchmarking results, with 1530 complete BUSCO genes compared with 1501 in BaRTv1.0. BaRTv2.18 also contains fewer fragmented genes, with 38 compared with 70 in BaRTv1.0 (Table S12). No transcripts overlapping Ns in the genome or duplicated sequences were identified in BaRTv2.18 (Table S12) as they were removed during the filtering process (Figure 1). Further investigation of 38 fragmented BaRTv2.18 genes shows that only four were supported by Iso-seq evidence, the remainder were based on Illumina predictions. Twenty-one genes contained genomic gaps or partially overlapped these, or adjacent to gaps affecting the gene predictions. Of the remaining 17 genes, 16 all showed clear signs of reference mis-assembly with the vast majority of reads mapped near the edge of the gene model showing extensive soft-clipping, indicating that they could not be mapped fully due to the reference sequence being incorrect. To assess the relative level of chimeric transcripts in BaRTv1.0 and BaRTv2.18, transcripts were compared with barley Haruna Nijo full-length cDNAs (flncDNAs; Matsumoto et al., 2011) and the percentage of transcripts that overlapped with multiple flcDNAs determined. BaRTv2.18 contained about 250 chimeric transcripts, derived from Iso-seq. This was lower (1.3%) than in BaRTv1.0 (2.23%; Table S12), reflecting more possible artefactual chimeric transcripts caused by short-read transcript assembly. Thus, BaRTv2.18 outperforms BaRTv1.0 in the completeness and diversity of this new barley reference transcriptome.

We also used Mercator4 (Schwacke et al., 2019) (https://www.plabipd.de/portal/web/guest/mer) to classify plant protein sequences into categories (bins) of essential functions. We found that 91.67% are covered by at least one protein from BaRTv2.18. This highlights that most protein categories in BaRTv2.18 are well represented. The numbers of different bins covered with at least one gene in BaRTv1.0 is 4637 and 4785 in BaRT2v18, showing a slight improvement over BaRTv1.0.

BaRTv2.18 represents an improved barley transcript dataset for AS analysis

The impact of BaRTv2.18 versus BaRTv1.0 on transcript quantification accuracy was investigated by comparing splicing ratios of AS transcript isoforms using HR RT-PCR and those obtained directly from transcript quantification using RNA-seq data. We used BaRTv1.0-QUASI (a 'padded' version of BaRTv1.0 (Rapazote-Flores et al., 2019)) which was constructed using the cv. Morex genome and BaRTv2.18 based on the cv. Barke as comparative RTDs. To evaluate whether BaRTv2.18 can be used to accurately quantify RNA-seq data beyond the Barke cultivar, HR RT-PCR data were generated using 85 (Morex derived) and 42 (Barke derived) primer pairs to amplify across known AS events, and total RNA from both Morex (15 samples) or Barke (12 samples), respectively. Comparisons between RT-PCR data in Morex and Barke are compared with quantifications using Morex and Barke RNA-seq (Figure 7). The Morex HR RT-PCR dataset had more primer pairs, RT-PCR products and therefore more datapoints (Figure 7b and d). Using Barke RNA-seq, BaRTv2.18 showed superior performance over BaRTv1.0 producing higher Pearson and Spearman correlation coefficients between the splicing ratios derived from HR RT-PCR and RNA-seq (0.833 and 0.840, respectively) than BaRTv1.0-QUASI (0.777 and 0.780, respectively; Figure 7a, c and e; Table S13). Despite using Morex RNA-seq, BaRTv2.18 still achieved higher Pearson and Spearman correlation coefficients over BaRTv1.0, which is a Morex-based assembly (Figure 7b, d and e; Table S13). In summary, these results show the improvement of transcript models in BaRTv2.18 has led to a modest improvement in quantification accuracy when compared with BaRTv1.0, taking into account differences between cultivars Morex and Barke.

BaRTv2.18 represents a substantial improvement in defining TSS and TES

The predominant difference between BaRTv2.18 and the short-read-based BaRTv1.0 (Rapazote-Flores et al., 2019) is the high-quality Iso-seq forming a backbone with 94 247 (63.6%) transcripts from 20 969 (53.2%) genes (Table S14). These were supported and complemented by stringently quality-controlled transcripts from the short-read Illumina assembly. An important feature of BaRTv2.18 that is provided by the Iso-seq dataset is accurately defined TSS and TES. Out of the 20 969 genes with Iso-seq transcripts with defined TSS/TES, we identified 47 750 and 62 468 TSS and TES, respectively, an average of 2.28 TSS and 2.98 TES per gene. The majority of these genes have alternative TSS or TES; 52.5% (11 023 genes) have two or more TSS, whilst 63% (13217) of genes have two or more TES (Figure S3).

The accuracies of the TSS and TES identified in the Iso-seq were validated in two ways. Firstly, TSS and TES were validated *in silico* by scanning the region around transcript 5′- and 3′-ends for enrichment of transcription motifs (TATA box, Initiator – Inr and Y patch), 3′-end processing signals (3′ polyadenylation signal – PAS and CFIm) and TSS (Kozak; Figure 5; Table S2). The positions of these motifs were compared using the transcript 5′- and 3′-ends from the BaRTv2.18 and BaRTv1.0 datasets. The TATA box is a T/A rich *cis*-regulatory element 25–35 bp upstream of TSS that specifies where transcription begins (Joshi, 1987; Morton et al., 2014), while the Initiator (Inr) motif is pyrimidine-rich and overlaps the TSS and is important for activation of transcription (Nakamura et al., 2002). The Y patch is a pyrimidine-rich sequence upstream of the TSS unique to plants and found in more than 50% of annotated rice genes (Civáň & Švec, 2009), while the Kozak motif is found downstream of the TSS and includes the start AUG codon for translation initiation (Kozak, 1987). The 3′ polyadenylation signal (PAS) motif is required for 3′-end polyadenylation (Proudfoot, 2011), while the CFIm motif is the binding site of cleavage factor Im, an essential 3′ processing factor (Brown & Gilmartin, 2003; Neve et al., 2017). We found that the 5′ transcript ends from BaRTv2.18 were more enriched for TATA box, Inr, Y patch promotor motifs and the Kozak TSS motif compared with BaRTv1.0 (Figure 5a–d). The TATA box motif presented a peak density of 0.055 instances per site (5002 instances out of 91 123) ~30 bp upstream of TSS site in BaRTv2.18, compared with a maximum peak of 0.014 (1844 instances out of 129 835) in BaRTv1.0 (Figure 5a), approximately a fourfold enrichment (proportion test $P < 2.2e\text{-}16$), suggesting TSS positions in BaRTv2.18 are much more characteristic of a true TSS. The Inr and Y patch sequence motif showed similar results with 0.01029 and 0.0128 instances per site (938 and 1170 instances out of 91 123) in BaRTv2.18, and 0.00642 and 0.00791 instances per site (834 and 1027 instances out of 129 835) in BaRTv1.0 (proportion test $P < 2.2e\text{-}16$). The Kozak motif had a defined peak at ~150 nt downstream of the TSS at 0.0035 instances per TSS (322 instances out of 91 123) in BaRTv2.18 (Figure 5d), while BaRTv1.0 had a much broader distribution, with a peak of 0.0013 instances per TSS (168 instances out of 129 835). As a result, the Kozak motif is significantly enriched in BaRTv2.18 (proportion test $P < 2.2e\text{-}16$). Similar results were obtained for PAS and CFIm 3′-end processing motifs with greater enrichment in BaRTv2.18 compared with BaRTv1.0, with a peak of 0.062 instances per TES (6652 instances out of 107 977) for the CFIm motif in BaRTv2.18 compared with 0.038 (4875 instances out of 129 928) in BaRTv1.0 and a peak of 0.019 instances per TES (2095 instances out of 107 977) for the PAS motif compared with 0.0075 (969 instances out of 129 928) in BaRTv1.0 (Figure 5e and f). Proportion test for both TES related motifs shows significant enrichment in BaRTv2.18 ($P < 2.2e\text{-}16$). These results strongly support the determination of TSS and TES by the methods used to analyse Iso-seq data and give confidence to the identification of HC TSS/TES.

Secondly, to confirm enriched TSS determined for Iso-seq transcripts, four genes with evidence of alternative TSS were selected for analysis by 5′-RACE. Sequencing determined the positions of 5′-RACE products for each of the four genes. These products matched Iso-seq TSS sequences exactly or were in close proximity (within 5 bp; Figure S2). For example, BaRT2v18chr5HG238000 5′-RACE confirmed two of the three TSS including the most prominent TSS expressed in inflorescence and peduncle (Figure S2a). Similarly, 5′-RACE products coincided with predicted TSS in BaRT2v18chr1HG022010 and BaRT2v 18chr6HG297070 (Figure S2b,c) and BaRT2v18chr6HG 314960 (5′-RACE site 20 bp from most abundant TSS – not shown). BaRTv1.0 transcripts from the same BaRTv2.18 genes showed far greater variability in their 5′-ends between the transcripts. The majority of BaRTv1.0 transcripts were longer than the longest BaRTv2.18 transcript, and none of the TSSs matched the Iso-seq or 5′-RACE products. Thus, 5′-RACE analysis supported predicted TSS, often with exact matches, for the genes examined.

BaRTv2.18 allows the investigation of transcript isoform expression and differential TSS and TES usage across barley tissues and organs

To illustrate the utility of BaRTv2.18, RNA-seq data from the 20 different biological samples used in its construction were quantified using Salmon with BaRTv2.18 as the reference transcriptome. The full dataset of transcript quantifications is found in Table S15. Analysis of BaRT2v18chr3HG130540 illustrates different levels of expression of AS isoforms and differential usage of alternative TSS and TES among different tissues/organs (Figure 8). BaRT2v18chr3HG130540 is a 59-kD U11/U12 small nuclear ribonucleoprotein component of the spliceosome and is alternatively spliced to give nine Iso-seq transcript isoforms. The gene has two TSS and four TES (Figure 8a). Five of the transcripts appear to code for protein (.1, .2, .6, .7, .8; Figure 8a), while four are unproductive in containing PTCs and NMD features (.3, .4, .5 and .9; not shown). The composition and relative abundances of different transcript isoforms clearly differ among the various samples (Figure 8b), and the expression of some transcripts is limited to specific tissues. For example, BaRT2v18chr3HG130540.6 has an intron in the 3′-UTR and appears to be the predominant isoform expressed in 6-day-old embryonic tissue, heat-stressed coleoptiles and peduncle with low or no expression in other tissues (Figure 8b; Table S15). The other protein-coding transcripts .1, .2, .7 and .8 code for the same protein but have alternative TSS and TES (Figure 8a). The .1 and .2 transcripts

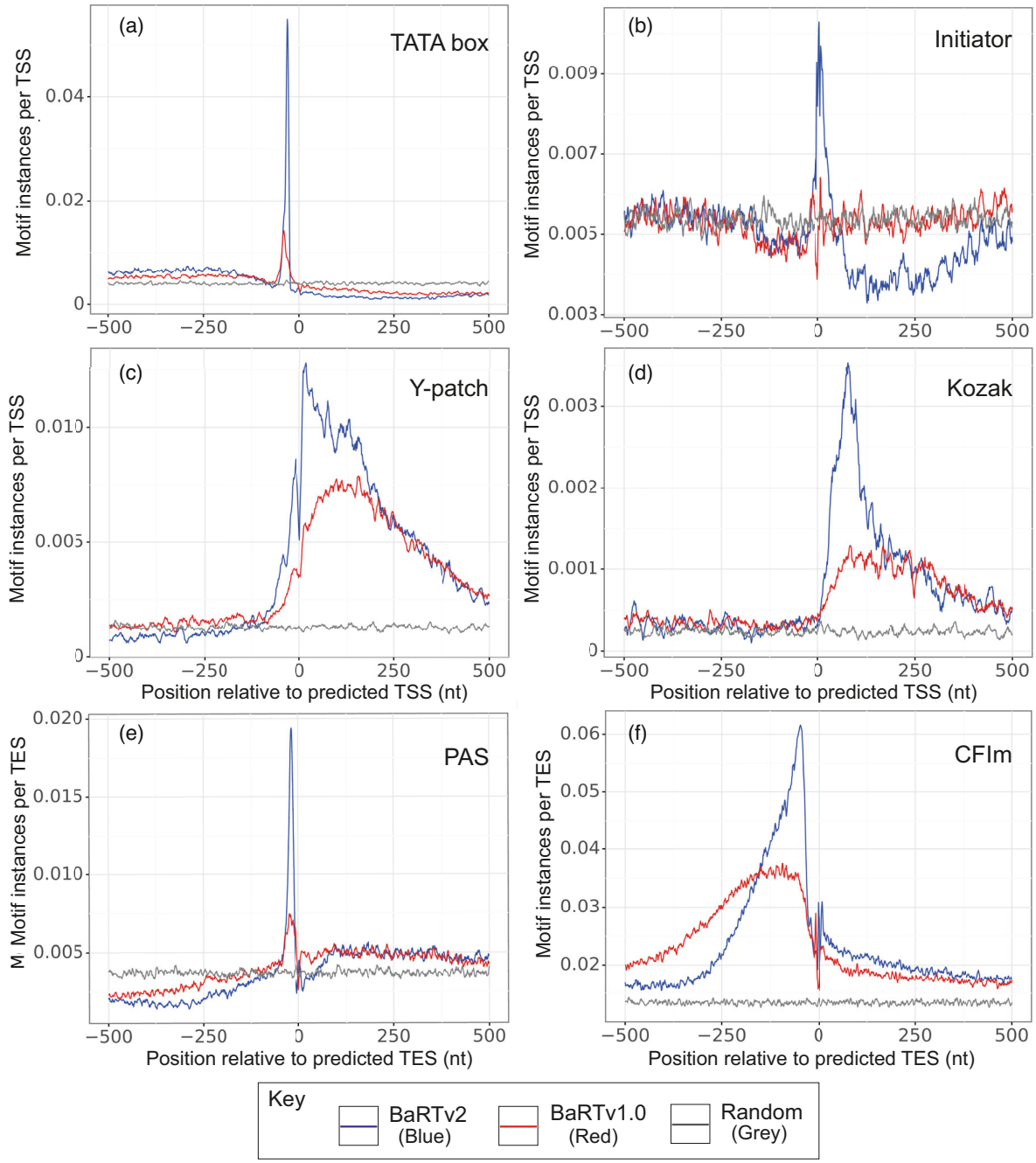


Figure 5. Enrichment of well-known sequence motifs associated with transcriptions start sites (TSS) and transcription end sites (TES).
(a–d) TSS-associated motifs and (e–f) TES-associated motifs. Lines indicate the frequency of instances of each motif in relation to the end sites, with red lines showing results from BaRTv1.0 and blue lines from BaRTv2.18; grey lines represent random control. TSS motifs: (a) TATA box; (b) Inr; (c) Y-patch; (d) Kozak motif; TES motifs: (e) PAS; (f) CFIm.

have the same TSS but different TES (Figure 8a), while the .7 and .8 pair have the same TSS (different from the TSS in .1 and .2) and different polyadenylation sites. Most samples expressed predominantly one or other of these pairs of transcripts, while tissues that expressed both usually expressed one more highly (Figure 8b). For example, while .7 was expressed in most tissues, the .8 transcript was found mainly in coleoptiles, day 55 lodicules, day 55 palea and 6-day-old embryos (Figure 8b). Variation in the levels of, for example, .2 and .8 in different tissues illustrates differential usage of alternative TSS. Similarly, variation in abundance of .1 and .2 transcripts among

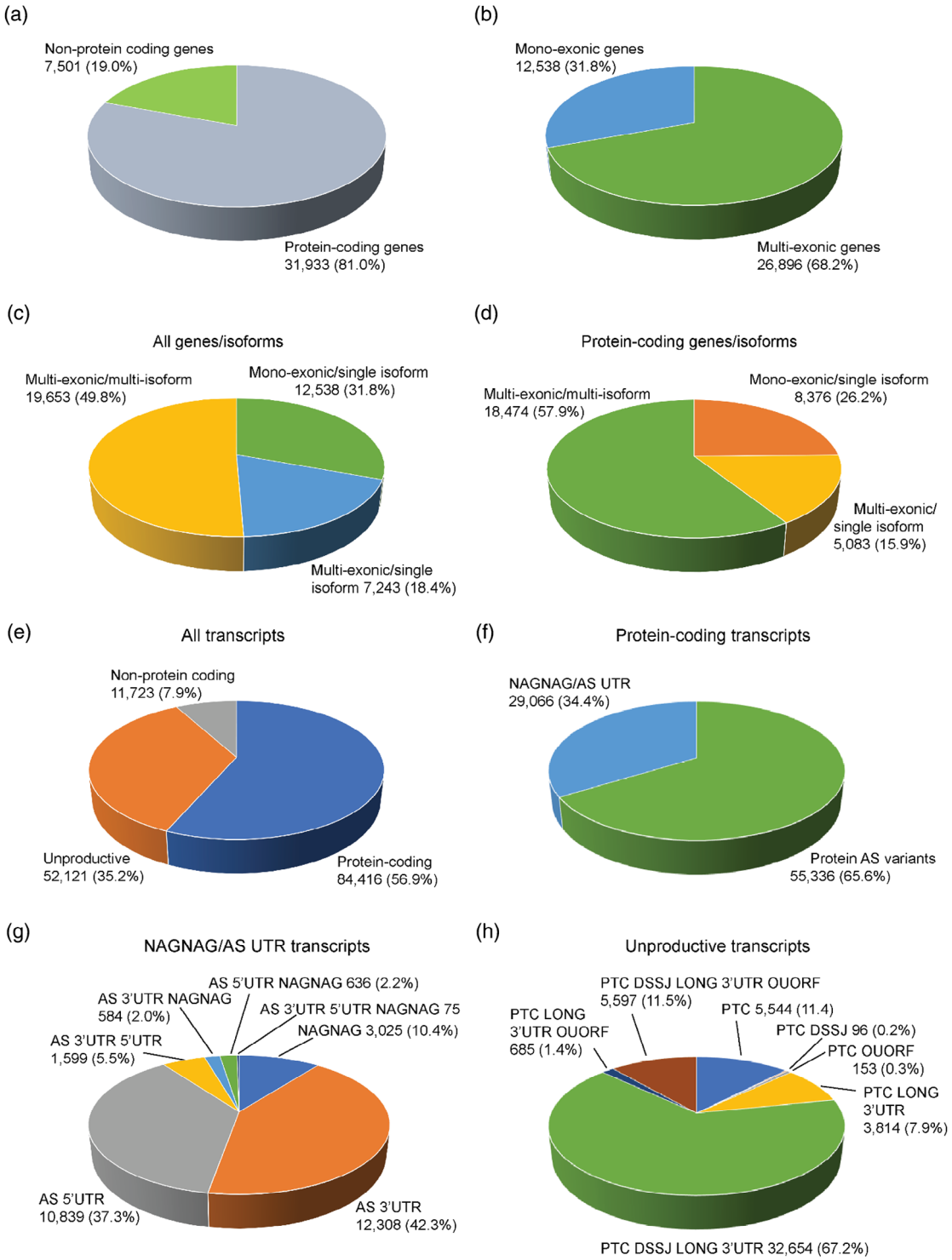
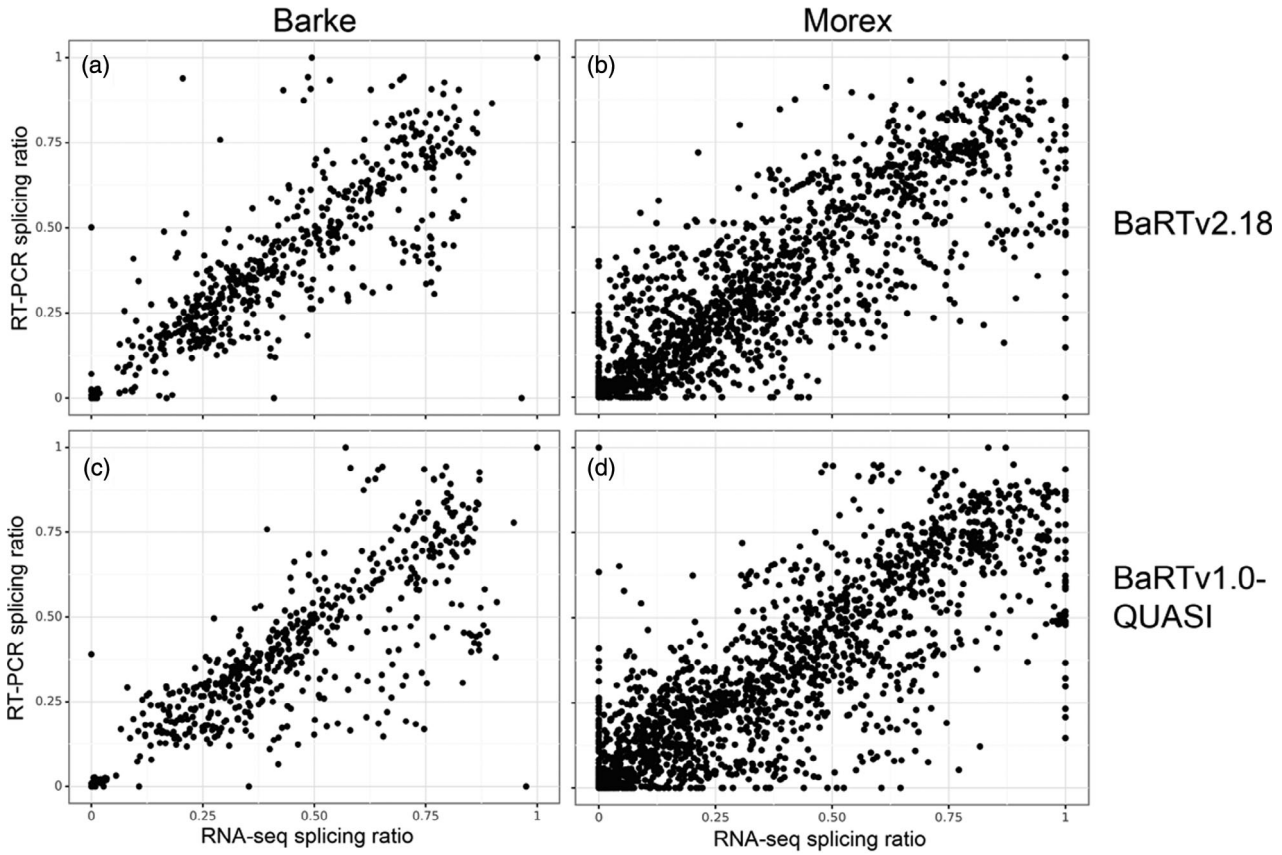


Figure 6. TranSuite analysis of BaRTv2.18 transcriptome.
The 39 434 genes of BaRTv2.18 are divided into: (a) protein-coding and non-protein-coding genes; and (b) mono-exonic and multi-exonic genes. The distribution of isoform numbers in mono-exonic and multi-exonic genes is provided for: (c) all genes; and (d) protein-coding genes. The 148 760 transcripts of BaRTv2.18 are divided into: (e) transcripts from protein-coding genes (both protein-coding and unproductive transcripts) and those from non-protein-coding genes; (f) protein-coding transcripts are separated into those transcripts with no or little change in protein sequence due to alternative splicing (AS) occurring only in the 5′- and/or 3′-UTR or NAGNAG AS, which alters the coding sequence by a single amino acid and those coding for protein variants; (g) NAGNAG and AS UTR transcripts are portioned in different combinations of AS events; (h) unproductive transcripts that contain PTCs are divided by the presence of different nonsense-mediated mRNA decay (NMD) features (DSSJ – downstream splice junction; long 3′-UTR; OUORF – overlapping uORF). The analysis is based on gene classification using a minimum of 100 amino acids to define a protein-coding transcript.

The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2022), doi: 10.1111/tpj.15871



(e)	BaRTv2.18 (Barke) with Barke HR RT-PCR	BaRTv1.0-QUASI (Morex) with Barke HR RT-PCR	BaRTv2.18 (Barke) with Morex HR RT-PCR	BaRTv1.0-QUASI (Morex) with Morex HR RT-PCR
Pearson correlation	0.833	0.777	0.825	0.821
Spearman ranked correlation	0.84	0.78	0.825	0.823

Figure 7. Comparison of accuracy of transcript quantification using BaRTv2.18 and BaRTv1.0-QUASI measured by correlation of splicing ratios from high-resolution (HR) reverse transcriptase-polymerase chain reaction (RT-PCR) and RNA-seq.
Splicing ratios were calculated for multiple alternative splicing (AS) events from different genes from relative fluorescence units from HR RT-PCR and transcript abundances (TPM) from RNA-seq data were quantified with Salmon. (a and b) BaRTv2.18 was used to calculate splicing ratios from RNA-seq data from 12 samples of Barke and 15 samples of Morex RNA and compare these with splicing ratios from HR RT-PCR using 42 and 85 primer pairs, respectively, from the same RNA samples. (c and d) The same was carried out using BaRTv1.0-QUASI. Pearson and spearman correlations for each of the combinations are shown (e).

different tissues reflect differential usage of alternative TES. The PTC-containing unproductive transcripts were generally less abundant than the other transcripts reflecting their likely sensitivity to NMD. These transcripts had either skipping of exon 6 or IR events. Thus, using BaRTv2.18 different transcript isoforms are readily detected and can be quantified at both the gene and transcript level.

DISCUSSION

The potential of RNA-seq experiments is often not fully realised due to the widespread lack of high-quality transcriptome annotations (Brown et al., 2017; Rapazote-Flores et al., 2019; Zhang et al., 2017). Stringently filtered, comprehensive RTDs permit fast and accurate gene and transcript level quantifications to be obtained from ultra-deep short-read RNA-seq data (Calixto et al., 2016; Calixto et al., 2019; Zhang et al., 2017) using Salmon and Kallisto (Bray et al., 2016; Patro et al., 2017). Here we present a new barley RTD resource, BaRTv2.18, based on parallel Iso-seq and Illumina data collected exclusively from the 2-row spring barley cv. Barke, a genotype directly relevant to present-day European agriculture and contemporary germplasm. The recently released chromosome-scale TRITEX assembly of the Barke genome (Figure 2a; Jayakodi et al., 2020) was used as the reference for mapping the Barke RNA-seq data. The previously published barley RTD, BaRTv1.0, was constructed from a diverse collection

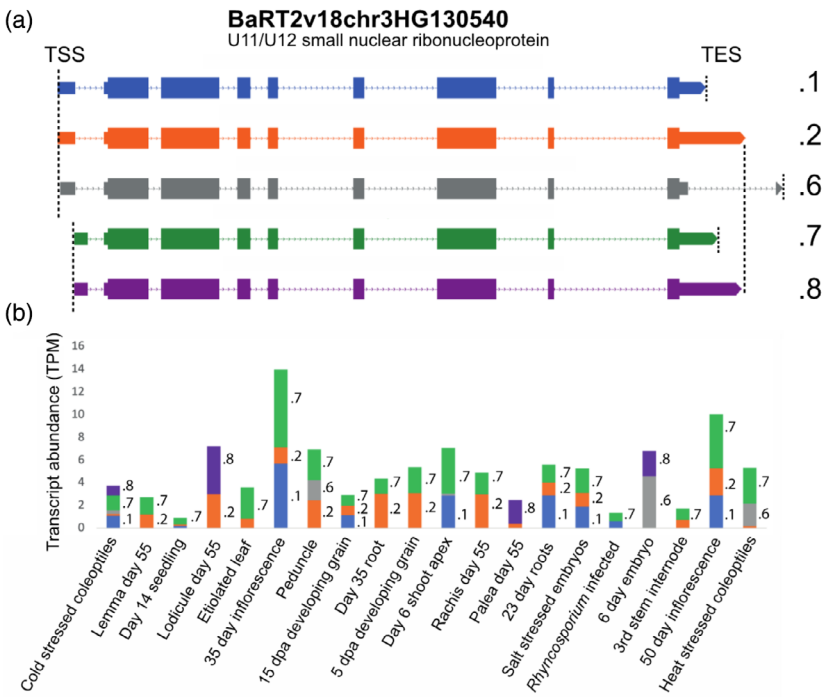


Figure 8. Differential transcription start sites (TSS) and transcription end sites (TES) usage in a range of barley tissues and conditions.
Abundance of transcript isoforms from gene BaRT2v18chr3HG130540, annotated as a U11/U12 small nuclear ribonucleoprotein 59-kDa protein across the 20 samples used to construct BaRTv2.18. (a) Transcript structures of protein-coding isoforms; dotted lines – introns; thick lines – UTR exons; and very thick lines – CDS regions. (b) Relative abundance of different isoforms (TPM) for each of the 20 samples are shown, with individual transcript quantifications indicated by different colours and labelled with isoform number within each bar.

of over 800 short-read RNA-seq samples in 11 datasets including different developmental stages and abiotic stresses from a range of barley cultivars and landraces using the genome sequence of barley cv. Morex as reference (Rapazote-Flores et al., 2019). Although BaRTv1.0 may contain greater transcript diversity (total of 177 240 transcripts) due to the range of genetic lines used, 5′- and 3′-ends of transcripts in BaRTv2.18 provide significant improvements over BaRTv1.0: (1) the majority of BaRTv2.18 reference transcripts (94247–64%) are derived from Iso-seq reads and have accurate TSS and TES as well as SJs; (2) the defined TSS and TES allow investigation of alternative TSS and TES usage reflecting enhanced resolution of transcriptional and post-transcriptional gene regulation, not possible with BaRTv1.0; (3) accurate transcript quantification using BaRTv2.18 does not require padding of transcripts (artificial extension of different transcripts of a gene to the length of the longest transcript – needed for BaRTv1.0 as 5′- and 3′-ends were not defined (Rapazote-Flores et al., 2019; Zhang et al., 2017)); and (4) SJs, TSS and TES can now be determined using single molecule sequencing data without the need for parallel experimental end determination techniques or hybrid correction. Iso-seq has been widely used to address transcriptome diversity in plant and crop species. However, issues still exist with false SJs caused by mapping errors around the introns, incomplete gene and transcript coverage, and accuracy of determination of TSS and TES. Methods to correct sequence errors impact transcript accuracy as both self-correction and hybrid correction of errors can lead to loss of AS events and introduce errors (over-correction). We have used the novel Iso-seq analysis pipeline based on that carried out for Arabidopsis AtRTD3 (Zhang et al., 2022) with some modifications for the construction of BaRTv2.18. We used TAMA for greater control of analysis, a SJ centric approach to identify and remove false SJs, and a probabilistic approach to determine TSS and TES, which take transcript abundance into account. Although the latest PacBio HiFi reads contain lower error rates (Wenger et al., 2019), false positive novel isoforms will still be an issue for CCS reads with a low number of passes. These problems are addressed by our pipeline. Short-read transcript assembly is used to provide gene and transcript coverage for gene regions where Iso-seq has little or no coverage. We applied a new software pipeline (RTDmaker) for stringent quality control of transcripts; RTDmaker filters transcripts to remove, for example, redundant and fragmentary transcripts and retain only transcripts with high-quality SJs. The combination of these methods ensured that BaRTv2.18 represents the most accurate and diverse barley transcriptome to date.

Other pipelines, besides the PacBio Iso-seq analysis pipelines, have been employed to improve determination of transcript features. For example, the Integrative Gene Isoform Assembler (IGIA) for cotton (Wang et al., 2019) combined both Iso-seq and RNA-seq to provide maximum coverage of genes and transcripts. The study also used CAGE-Seq and PolyA-Seq to define TSS and TES, respectively, which were used to correct Iso-seq sequence 5′- and 3′-ends (Wang et al., 2019). The number of CAGE-Seq and

PolyA-Seq sites over the expression threshold were found in 22 863 and 23 736 genes, respectively, representing about one-third of cotton genes (Wang et al., 2019). Here, 20 969 barley genes have Iso-seq support, each with both TSS and TES defined. Importantly, our method does not require experimental determination of ends by CAGE-Seq or PolyA-Seq data, rather it maximises the information that can be obtained from Iso-seq by determining TSS and TES directly from the Iso-seq dataset using our novel approaches. Furthermore, our method does not require hybrid error correction and so avoids over-correction (Kuo et al., 2020); instead, Iso-seq transcripts are removed if they are not supported by HC SJs, TSS and TES. The accuracy of TSS and TES for the Iso-seq derived transcripts in BaRTv2.18 was shown by the enrichment of characteristic sequence motifs of known TSS and TES in BaRTv2.18 in comparison to BaRTv1.0 (Figure 5) and by 5′-RACE of specific genes. In plants, TSS have been identified in Arabidopsis, maize and cotton (Mejía-Guerra et al., 2015; Morton et al., 2014; Wang et al., 2019), and TES in Arabidopsis and cotton (Parker et al., 2020; Wang et al., 2019; Zhang et al., 2020). CAGE-Seq analysis in maize has shown that ~70% of annotated gene models have mis-annotated TSS, reflecting the inaccuracy of short-read annotation (Mejía-Guerra et al., 2015). In many cases, the CAGE TSS mapped downstream of annotated TSS suggesting that the short-read assembled transcripts were artificially extended as seen for Arabidopsis in TAIR10 and Araport (Zhang et al., 2017). Comparing enrichment of promotor motifs between BaRTv1.0 and BaRTv2.18 confirmed the expected enhanced definition of TSS and TES in BaRTv2.18 (Figure 5). Quantitative CAGE-Seq in cotton found that 40% of gene loci had alternative TSS (Wang et al., 2019), whilst in maize, hundreds of significant differences in promotor usage were identified between tissues (Mejía-Guerra et al., 2015). Here, in the genes with HC TSS and TES sites, over a half had two or more TSS or TES sites. The advantage of BaRTv2.18 having well-defined TSS and TES for many genes is that differential TSS and TES usage can be addressed in RNA-seq analysis, as illustrated by BaRT2v18chr3HG130540 showing tissue-specific variation in alternative TSS or TES usage (Figure 8).

More than half of the transcripts in BaRTv2.18 are based on HC Iso-seq, and the genes with low or no Iso-seq coverage required complementation with short-read assembled transcripts. These transcripts have been generated by multiple assemblers to capture transcript diversity and stringent quality control to remove false transcripts and retain only well-supported transcripts. The accuracy of assembly is illustrated by the high number of short-read transcripts that are replaced by Iso-seq transcripts during merging of BaRTv2.0-Iso and BaRTv2.0-Illumina. Therefore, although the 5′- and 3′-ends of these transcripts should be treated with caution, their SJs are accurate and are a valuable contribution to BaRTv2.18. As further improvements in single molecule sequencing technology are made to significantly improve Iso-seq depth of coverage, ultimately, complementation with short-read assemblies will not be required. Similarly, our methods avoid the need for hybrid error correction, which also helps single molecule sequencing analysis to move towards being a self-contained process where complete transcript characterisation is achieved directly from the data.

BaRTv2.18 improved transcript quantification from RNA-seq data as demonstrated by the high correlations of AS splicing ratios between HR RT-PCR and RNA-seq data compared with BaRTv1.0 (Figure 7; Table S13). In this comparison, the QUASI (padded) version of BaRTv1.0 was compared with the unadjusted BaRT2.18 version. We showed previously in Arabidopsis AtRTD2 that variation in the 5′- and 3′-ends of transcripts from the same genes could skew transcript quantification, and that padding of transcripts to the length of the longest transcript overcame this problem (Zhang et al., 2017); the BaRTv1.0-QUASI version similarly improved transcript quantification in barley (Rapazote-Flores et al., 2019). The accurate determination of TSS and TES for many transcripts and the removal of short, fragmented transcripts in BaRTv2.18 should overcome the need for padding of the RTD. The unpadded BaRTv2.18 generated more accurate quantification than BaRTv1.0-QUASI as shown by the correlation of splicing ratios from HR RT-PCR and RNA-seq data. There was also an effect of cultivar on the correlations where comparing splicing ratios from HR RT-PCR and RNA-seq from the same cultivar had higher correlation values. This suggests that higher quantification accuracy is achieved when using the genome reference from same genotype in constructing the RTD. It will influence the development of a new pan-transcriptome to complement the recently published pan-genome (Jayakodi et al., 2020). Comparative benchmarking also confirmed that BaRTv2.18 provides a substantial improvement over BaRTv1.0 (Table S12). The reduction in fragmented BUSCO genes in comparison to BaRTv1.0 most likely reflects both the improved transcriptome *per se* and the higher quality of the Barke genome in comparison to the Morex genome that was used with BaRTv1.0 for read mapping (Jayakodi et al., 2020; Mascher et al., 2017). In other metrics, BaRTv2.0-Iso performed well, with relatively few duplicated sequences (35) despite no specific efforts to remove them, and low numbers of possible chimeras (0.3%).

High-quality RTDs form an important community resource for understanding the regulation of gene expression. Therefore, the BaRTv2.18 is also supported by extensive transcript characterisation and annotation (Tables S16, S17 and S18). Accurate translations of each transcript have been annotated using TranSuite (Entizne et al., 2020). The program fixes the authentic translation start site of each

gene and translates all the transcripts from the fixed AUG. This overcomes mis-annotation of ORFs and distinguishes protein-coding and unproductive isoforms (Brown et al., 2015; Zhang et al., 2017). The level of unproductive transcript isoforms of protein-coding genes in BaRTv2.18 (35.2%) compared well to the level in the recent Arabidopsis AtRTD3 (41%; Zhang et al., 2022). The translations from TranSuite also allow accurate identification of positions of PTCs and other NMD signals in unproductive transcripts. The largest group of potential NMD substrates in both BaRTv2.18 and AtRTD3 contained the classic features of a PTC with downstream splice junction (DSSJ) and long 3′-UTR, and consisted of 67.2% and 70.3% of unproductive transcripts, respectively. These figures reflect the close interaction of AS and NMD as a post-transcriptional mechanism for regulating gene levels of functional transcripts (that code for proteins; Kalyna et al., 2012). Transcripts are also annotated with potential functions, and coordinates of transcript features such as CDS, exons, 5′- and 3′-UTRs are provided. The methods that we have developed for both Iso-seq and Illumina analysis will allow any new long- and short-read datasets to be efficiently added as they become available, and will improve the construction of RTDs from other cultivars and feed into development of a barley pan-transcriptome. Clearly, to maintain their community value, improvements to transcriptome resources should evolve alongside parallel improvements to emerging genomic resources (Jayakodi et al., 2020; Mascher et al., 2017; Mascher et al., 2021; Monat et al., 2019). Importantly, BaRTv2.18 will play a valuable role in providing experimental support for annotation of the most recent barley genome assemblies.

In summary, BaRTv2.18 represents a new and improved RTD built upon a broad framework of highly informative Pacbio Iso-seq full-length transcripts incorporating accurate SJ, TSS and TES information. It is based upon the European 2-row spring barley cv. Barke, which itself is highly relevant to contemporary barley agriculture and current research. The combination of the novel Iso-seq analysis and filtering pipeline, the new automated quality control pipeline for short-read assembled transcripts and detailed transcript characterisation and annotation make BaRTv2.18 a high-quality community-wide reference dataset.

EXPERIMENTAL PROCEDURES

Plant material

Twenty-one tissues were sampled, all from barley cv. Barke, a European 2-row spring variety with a newly published genome (Jayakodi et al., 2020). Where possible, tissues were sampled from three or more plants. Third stem internode was sampled from one plant due to the large size of the tissue. Plant material was grown under controlled conditions (16 h light, 18°C, light intensity 400 μmol m^{-2} sec^{-2}, 8 h dark, 14°C, 0 μmol m^{-2} sec^{-2}, humidity 60%) at the James Hutton Institute, or at IPK, Gatersleben. All tissues were flash-frozen in liquid nitrogen and added to liquid-nitrogen-cooled tubes, before being stored at −80°C. Plant material is summarised in Table S1, and details of the sample preparations for each tissue can be found in Supplementary Materials.

RNA extraction, library preparation and sequencing

RNA extraction was carried out using Qiagen RNeasy Plant mini kits using the manufacturer's instructions. Total RNA was quality checked using a Bioanalyzer 2100 (Agilent), and RNA samples with the best RIN were used for Iso-seq and RNA-seq library preparation. RNA for Iso-seq and RNA-seq was aliquoted from the same sample. For the seven tissues collected at IPK Gatersleben, RNA was extracted using a Trizol extraction protocol (I.B.S.C., 2012) and purified using Qiagen RNAeasy miniprep columns following the manufacturer's instructions. RNA quality was checked on Agilent RNA HS screen tape.

Single-molecule real-time sequencing (SMRT) was performed at IPK Gatersleben using the Pacbio Iso-seq method. Full-length cDNA was generated from total RNA using the TeloPrime Full-Length cDNA Amplification Kit V2 from Lexogen (Vienna, Austria). Two separate cDNA synthesis reactions per sample, each with 2 mg total RNA as input, were performed following the manufacturer's recommendations for Iso-seq library preparation. The purified double-stranded cDNA from both reactions was pooled and amplified in a qPCR reaction to determine the optimal cycle number for large-scale endpoint PCR. The cDNA was amplified in 16 parallel PCR reactions using 2 μl cDNA per reaction with the optimal cycle number using the TeloPrime PCR Add-On Kit V2 (Lexogen). The PCR reactions were purified with AMPure PB Beads according to the post-PCR purification step described in the Iso-seq template preparation protocol for Sequel Systems for libraries without size selection (Pacific Biosciences, Menlo Park, CA, USA). Six of the PCR reactions were pooled to make up fraction 1 enriched in shorter fragments, and the other 10 PCR reactions were pooled to make up fraction 2 enriched in longer fragments. Equimolar quantities of the two cleaned-up fractions were then pooled, and 5 μg of cDNA was used for SMRTbell library preparation using the Sequel 3.0 chemistry (Pacific Biosciences) according to the manufacturer's protocol. Each library was sequenced on a separate lane of a SMRT cell 1 M LR on a PacBio Sequel machine with movie lengths of 1200 min. For tissues collected from IPK, pooled RNA from seven tissues (above) was used to prepare two libraries using the cDNA prepared from TeloPrime v1.0 kit (Lexogen, Vienna, Austria) following the manufacturer's instructions. Libraries were quantified and sequenced on a Pacbio Sequel at IPK Gatersleben (Jayakodi et al., 2020).

Illumina RNA-seq library preparation and RNA-seq was carried out by Novogene (HK) Company Limited, Hong Kong. The 20 libraries were prepared using NEBNext® Ultra™ Directional RNA Library Prep Kit and sequenced using Illumina NovaSeq 6000 S4 (PE 150). Accession numbers for accessing the raw data generated for each tissue are provided under Data Availability.

Iso-seq processing: Generation of the initial transcript assembly

The 20 Iso-seq libraries were processed individually. Raw subreads were initially processed using the CCS tool from Isoseq3 to create circular consensus sequences (CCS). The resulting CCS reads were stripped of adapter sequences and poly-A tails using lima and refine from Isoseq3. A further clean-up of PolyA tails was carried out using the Transcriptome Annotation by Modular

Algorithms (TAMA) tama_flnc_polya_cleanup.py script. The resulting reads were mapped to the Barke genome (Jayakodi et al., 2020) using Minimap2. TAMA collapse (https://github.com/GenomeRIK/tama/wiki/Tama-Collapse) was used to process the read mappings for each sample to generate a non-redundant set of transcripts. The TAMA collapse output files from 20 samples were merged together using TAMA merge (https://github.com/GenomeRIK/tama/wiki/Tama-Merge). Details of the parameters used for each step can be found in the Supplementary Materials.

Identification of HC SJs

Sequence errors in Iso-seq reads can lead to the identification of incorrect SJ coordinates. A collection of HC SJs was selected where they had: (1) canonical donor and acceptor sequence motifs; and (2) at least one supporting read that had a perfect match within 10 nucleotides upstream and downstream of the SJ (Zhang et al., 2022). SJ and read information was obtained from the TAMA collapse output using a customised script (https://github.com/maxecoulter/BaRT-2) for each of the 20 libraries. In addition, a dataset of HC SJs was created from Illumina short-read RNA-seq data using the output from the mapping tool STAR. HC SJs were selected based on having: (1) canonical motifs; (2) an overhang of $\geq$ 10 nt; and (3) more than 5 supporting reads.

To remove SJs that were potentially caused by template switching (Cocquet et al., 2006), the Hamming distance was calculated between the last 8 bases of the SJ's upstream exon and the last 8 bases of the intron for each SJ. The same was done for the first 8 bases of the intron and the first 8 bases of the downstream exon for each SJ. A Hamming distance of 1 or less (1 nucleotide difference between the two 8-nucleotide sequences being compared) in either comparison was considered a potential template switching event, and these SJs were removed from HC SJs.

Identification of TSS/TES

For 3′-ends, reads potentially with internal priming (Nam et al., 2002) were identified from the poly(A) output from TAMA collapse. Reads/transcripts with > 16 As in the 20 nucleotides following the 3′-end in the internal gene sequence in the genome were considered potential off-priming and were removed. For many genes, there were long reads representing full-length transcripts, but also 5′ and 3′ truncated reads likely due to degradation or incomplete cDNA synthesis. To distinguish between real and false transcript end sites, we used two methods depending on expression level (Zhang et al., 2022). We assumed that for genes with multiple reads with different ends, the start and ends of fragmentary reads would be distributed randomly and uniformly between the 5′- and 3′-ends, while true TSS/TES would have more read support than what would be expected by random. For each gene, we used a binomial discrete mass function to infer the probability that the read start and end position was likely to be random using the following formula (Loader, 2000):

$$Pr(Reads = h) = \binom{n}{h}\left(\frac{1}{t}\right)^{h}\left(\frac{t-1}{t}\right)^{n-h}$$

where n is the total number of reads in a gene, h is the number of reads with a start or end site being tested, and t is the total number of starting or end sites in the gene. TSS or TES within a gene were considered to be HC for downstream processing when the probability of having h reads at the site by random < 0.01.

To increase the gene coverage for low-abundance genes (typically < 10 reads) without at least one significantly enriched TSS and TES, we retained TSS/TES if they had start and end support from 2 or more reads within a sliding window of ± 20 and ± 60 for starts and ends, respectively.

Filtering the poorly supported transcript assembly

The transcripts from the initial assembly were filtered using the datasets of HC SJ and HC TSS/TES sites. For genes with HC TSS and TES, transcripts were kept if: (a) their starts locate within 10 nucleotides of the defined HC TSS and their ends locate within 30 nucleotides of the defined HC TES; and (b) all SJs in the transcript were present in the list of HC Iso-seq SJs. To remove redundancy due to variation at the 5′- and 3′-sites, transcripts were run through TAMA merge to create BaRTv2.0-Iso, using the settings 'm 0 -a 100 -z 100'.

Illumina RNA-seq processing and analysis

In conjunction with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) Trimmomatic version 0.39 was applied to the raw Barke RNA-seq reads (Bolger et al., 2014) for quality control and adapter trimming. The trimmed RNA-seq reads were aligned to the Barke reference genome with STAR (Spliced Transcripts Alignment to Reference; Dobin et al., 2012). A two-pass method in which the SJ outputs of a first STAR pass were used for indexing a second pass was applied to improve the accuracy and sensitivity of read alignment and SJ identification. Minimum and maximum intron sizes were 60 and 15 000 for the alignment. Mismatches were not allowed. Transcripts were assembled from the read alignments with three tools: Cufflinks version 2.2.1 (Trapnell et al., 2012), StringTie version 2.0 (Pertea et al., 2015) and Scallop version 0.10.4 (Shao & Kingsford, 2017) to take advantage of complementarity among the tools in capturing the diversity of transcripts. One transcript dataset was generated for each sample by each assembler. The 60 individual assemblies were merged and quality filtered into a single transcript assembly (BaRTv2.0-Illumina) with RTDMaker (https://github.com/anonconda/RTDmaker). RTDmaker identifies and removes: (1) redundant transcripts; (2) transcripts with SJs with low read support; (3) transcript fragments; (4) poorly supported antisense transcripts; (5) unstranded models; (6) antisense transcript fragments; and (7) low-expressed transcripts (TPM cut-off default = 1).

Merging BaRTv2.0-Iso and BaRTv2.0-Illumina

Transcripts from the HC Iso-seq dataset ('BaRTv2.0-Iso') and Illumina dataset ('BaRTv2.0-Illumina') were initially merged with TAMA merge using 'wobble' parameters of -m 0 -a 100 -z 100 (Kuo et al., 2020). This allowed the removal of redundant transcripts with exact matches for SJs with small variable ends (100 bp). During the merge, priority was given to 5′- and 3′-ends of Iso-seq transcripts by setting 'capped' flag for Iso-seq dataset and 'no_cap' for BaRTv2.0-Illumina dataset.

Custom code was used to carry out further filtering according to the following principles: (a) BaRTv2.0-Iso was considered to be the most accurate transcriptome assembly, so all transcripts with HC TSS/TES were kept; (b) Iso-seq transcripts without HC TSS and TES were only kept if they had Illumina support and were at least 80% of the length of the average length of Illumina transcripts in that gene; and (c) Illumina transcripts were retained for genes with no Iso-seq coverage or if they contained novel SJs.

Finally, to remove the high number of mono-exonic genes likely due to DNA contamination (Kuo et al., 2020), the mono-exonic genes were re-analysed using expression abundance, and those genes with expression higher than 1 TPM in less than 2 samples were removed (over 13.4 k genes). To keep *bona fide* single-exon

protein-coding genes that were tissue-, organ- or condition-specific, 2.4 k single-exon protein-coding genes with ORFs greater than 100 amino acids predicted by TranSuite (Entizne et al., 2020) were kept.

Motif enrichment analysis

To explore motifs commonly associated with TSS and TES (Table S2), for every predicted transcript 5′- and 3′-site in BaRTv1.0 and BaRTv2.18, the sequence ± 550 nucleotides either side of the TSS or TES was extracted from the genome, and after a regex search the position coordinates of all matching motifs relative to the sequence were extracted. As a control, the same number of random sites was taken from random chromosomes, and the above analysis was carried out.

High-resolution reverse transcriptase-polymerase chain reaction

RNA from 12 of the tissues used to construct BaRTv2.18 was used for HR RT-PCR validation. A list of samples is provided in Table S4. Forty-seven primer pairs used for covering 99 RT-PCR products designed previously for BaRTv1.0 (Rapazote-Flores et al., 2019) were identified where: (1) the primer pair had perfect matches to the Barke genome; and (2) the interval between the primer pairs had good RNA-seq coverage (> 1000 reads in at least two samples) identified with samtools bedcov (Li et al., 2009). The full list of primers, products and RT-PCR product proportions used in the analysis are in found Table S3a,b and c. The 5′-labelling, first-strand cDNA synthesis, HR RT-PCR and detection were carried out as described previously (Rapazote-Flores et al., 2019; Simpson et al., 2008). To compare with BaRTv1.0, taking into account possible differences in sequence, expression and AS due to genotypic variation, 86 primer sets (Tables S5 and S6) from the HR RT-PCR dataset of a previous study (Rapazote-Flores et al., 2019) were also used, and the data reanalysed according to the method described in Rapazote-Flores et al. (2019). A detailed description of the method along with the custom code used to carry out the analysis is described at https://github.com/maxecoulter/BaRT-2, and in the Supplementary Methods section.

5′-RACE

Transcription start sites were characterised by cloning and sequencing 5′-RACE products using the SMARTer RACE 5′/3′ kit (Takara Bio USA). In summary, antisense gene-specific primers were designed downstream of predicted Iso-seq TSS (Table S4). RNA samples from Barke inflorescence and peduncle were pooled, and 1 μg was used in the 5′-RACE assays. After reverse transcription, 3′-end tailing and annealing of the SMARTer II A oligonucleotide, the second strand was synthesised to form a double-stranded cDNA. RACE products were amplified using the universal primers and the gene-specific primers (Table S4), followed by in-fusion cloning into the vector pRACE using the 15 base overlap designed into the gene-specific primer. DNA from selected plasmids was extracted using Wizard Plus SV minipreps DNA purification system (Promega) and sequenced using the M13F primer. Sequences were aligned to the relevant gene transcripts in the BaRTv2.18 release using Clustal Omega on the EMBL-EBI platform (Madeira et al., 2019).

Annotation

TranSuite 0.1.2 was used to identify ORFs in BaRTv2.18 transcripts (Entizne et al., 2020), and the predicted protein sequences used for annotation using PANNZER2 (Koskinen et al., 2015) to produce predicted functions and GO terms. A minimum positive predictive value of 0.3 was used to include a predicted annotation in the dataset. TranSuite is a program that identifies coding and non-coding transcripts, generates accurate translations of transcripts, and identifies features of AS and NMD. SUPPA2 version 2.3 was used to characterise AS events that occur within BaRTv2.18 (Alamancos et al., 2015).

Benchmarking

Transcript lengths were computed from the candidate FASTA files using samtools faidx v. 1.9 (Li et al., 2009). The number of duplicated sequences was established using seqkit rmdup v0.12.0 (Shen et al., 2016). The number of genomic gaps denoted by Ns in the candidate transcripts was counted using custom Java code. BUSCO v4.0.6 (Simão et al., 2015) was used to benchmark transcript set completeness and transcript fragmentation.

To establish the percentage of assembled transcripts that was chimeric (either as a result of readthrough or bioinformatics artefacts), GMAP version 2018-07-04 (Wu & Watanabe, 2005) was used to map the Haruna Nijo full-length cDNA (flcDNA) sequences (Matsumoto et al., 2011) to the reference genomes of cultivars Morex (Mascher et al., 2017) and Barke (Jayakodi et al., 2020), using the parameters '--min-identity = 0.96' and '--min-trimmed-coverage = 0.95'. The resulting GFF output was converted to GTF format using gffread v0.11.6. (http://ccb.jhu.edu/software/stringtie/gff.shtml#gffread). Using both reference genomes was intended to prevent variation in genome composition from affecting the outcome of the mapping, as the previous version of BaRTv1.0 was based on the Morex genome. The resulting GTF files were then screened for positional overlap with the GTF files of the candidate assemblies using bedtools intersect (Quinlan & Hall, 2010), using the flcDNA mapping to the reference genome underlying the candidate assembly (BaRTv1.0 = Morex, BaRTv2.18 = Barke). Custom Java code was used to parse the output from bedtools intersect and count the number of flcDNA sequences spanning a candidate transcript, with those candidate transcripts overlapping more than one flcDNA being counted as chimeric.

ACKNOWLEDGEMENT

The authors acknowledge the Research/Scientific Computing teams at The James Hutton Institute and NIAB for providing computational resources and technical support for the "UK's Crop Diversity Bioinformatics HPC" (BBSRC grant BB/S019669/1), use of which has contributed to the results reported. We thank Susanne König and Axel Himmelbach from IPK Gatersleben for preparing and sequencing the Iso-seq libraries. The authors acknowledge the support of the Biotechnology and Biological Sciences Research Council (Grant Numbers BB/S020160/1, BB/S004610/1, BB/R014582/1) to MC, RZ, MB, WG, RWa and JWSB; The Rural & Environment Science & Analytical Services Division of the Scottish Government (Grant Number SRP WP2.1 and 2.2) to RZ, WG, MB, LM, NM, MS, CS, JF and RWa; a joint PhD scholarship from the James Hutton Institute and the University of Dundee/BBSRC DTP (BB/M010996/1) to JSWB, RZ and JCE; German Research Council (DFG) (Grant number STE 1102/15-1) to NS and RWo; and the National Science Foundation (Grant number ERA-CAPS 1844331) to GM and AH.

CONFLICTS OF INTEREST

The authors declare no conflicts or competing interests.

AUTHOR CONTRIBUTIONS

RZ, JWSB, RWa, NS and GM designed the experiments. MC, JCE, WG, RWo, NM, CS, JF carried out the experiments. MC, JCE, MB, MS, CS, JWSB, WG and RZ analysed the data. LM maintained and updated the BaRT annotation website. The manuscript was written by MC, RZ, JWSB and RWa with contributions from all other authors.

DATA AVAILABILITY STATEMENT

The BaRTv2.18 transcriptome and annotation files are available at https://ics.hutton.ac.uk/barleyrtd/bart_v2_18.html. Scripts used to produce BaRTv2.18 are available at the repository https://github.com/maxecoulter/BaRT-2. Raw Illumina and Iso-seq datasets have been uploaded to the Sequence Read Archive under the accession numbers PRJNA755156 and PRJNA755523.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. Plant material from cv. Barke for PacBio Iso-seq and Illumina RNA-seq.

Table S2. Motifs associated with TSS and TES sites used to validate TSS and TES sites in BaRTv2.18.

Table S3. (a) Sequences of primers used for RT-PCR. (b) Primers, 12 Barke samples and RT-PCR proportions with product sizes used in RT-PCR analysis. (c) Primers, 15 Morex samples and RT-PCR proportions with product sizes used in RT-PCR analysis

Table S4. Information on genes, TSS positions, primers used for 5′RACE validation of TSS.

Table S5. Iso-seq samples and results of preprocessing/TAMA annotation

Table S6. (a) SJ analysis in Iso-seq reads. (b) SJs removed from analysis

Table S7. (a) Genes with HC SJs, TSS and TES in Bart2.0-Iso. (b) Other genes without HC TSS and TES

Table S8. RTDmaker srQC quality control of BaRT2.0-Illumina transcripts

Table S9. Summary of TranSuite analysis of BaRT2.18 for coding/non-coding and mono-exonic/multi-exonic genes with single or multiple transcript isoforms

Table S10. Frequency of different AS events in BaRT2.18

Table S11. Summary of TranSuite analysis of BaRT2.18 transcript characterisation

Table S12. Benchmarking assessments of BaRT2.18 and BaRTv1.0

Table S13. Comparison of splicing ratios between HR RT-PCR and RNA-seq using BaRTv1.0 and BaRTv2.18

Table S14. Category and source of genes and transcripts in BaRTv2.18

Table S15. Salmon quantification of 20 samples using BaRTv2.18 (TPM)

Table S16. TransFeat characterisation of BaRT2.18 transcripts

Table S17. BaRTv2.18 annotation

Table S18. BaRTv2.18 annotation (gene level)

Data S1. Supplementary materials and methods

Figure S1. Histogram of end read support to determine window sizes for filtering of Iso-seq dataset. a) distribution of 5′ read ends of transcripts with high coverage (>= 10 reads), b) distribution of 3′ read ends of transcripts with high coverage (>= 10 reads). Window sizes used for filtering 5′ and 3′ ends of transcripts are indicated in C (+/− 10) and D (+/− 30) by vertical lines.

Figure S2. Confirmation of predicted TSS by 5′ RACE. a) BaRT2v18chr5HG238000. TSS of transcript isoforms are visualised on Integrated Genome Browser (IGB) along with the genomic sequence and the last digits of the co-ordinates are indicated. 5′ RACE products are shown by green arrows. The right hand panel shows the relative abundance (TPM) of expression of isoforms from the different different TSS in inflorescence (INF) and peduncle (PED) tissues. Orange arrow – direction of transcription. b) BaRT2v18chr6HG297070 – the two 5′ RACE sites correspond to the most abundant transcripts (legend as in a). c) BaRT2v18chr1HG022010 has a TSS (1552) 250 bp upstream of a second TSS (legend as in a). The 5′ RACE signal at 1552 coincides exactly with the major TSS. No signal was detected at 1291 but transcripts using this TSS are much more abundant in other tissues (e.g. roots, rachis, internode – not shown).

Figure S3. Histogram of TSS (a) and TES (b) per gene in BaRTv2.18 genes with Iso-seq support.

REFERENCES

Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F. *et al.* (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, **7**, 11706.

Alamancos, G.P., Pagès, A., Trincado, J.L., Bellora, N. & Eyras, E. (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, **21**, 1521–1531.

Au, K.F., Underwood, J.G., Lee, L. & Wong, W.H. (2012) Improving PacBio long read accuracy by short read alignment. *PLoS One*, **7**, e46679.

Barakate, A., Orr, J., Schreiber, M., *et al.* (2020) Time-resolved transcriptome of barley anthers and meiocytes reveals robust and largely stable gene expression changes at meiosis entry. *bioRxiv*, 2020.04.20.051425.

Baralle, F.E. & Giudice, J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*, **18**, 437–451.

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**, 525–527.

Brown, J.W.S., Calixto, C.P.G. & Zhang, R. (2017) High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. *New Phytologist*, **213**, 525–530.

Brown, J.W.S., Simpson, C.G., Marquez, Y., Gadd, G.M., Barta, A. & Kalyna, M. (2015) Lost in translation: pitfalls in deciphering plant alternative splicing transcripts. *The Plant Cell*, **27**, 2083–2087.

Brown, K.M. & Gilmartin, G.M. (2003) A mechanism for the regulation of pre-mRNA 3′ processing by human cleavage factor Im. *Molecular Cell*, **12**, 1467–1476.

Calixto, C.P.G., Guo, W., James, A.B., Tzioutziou, N.A., Entizne, J.C., Panter, P.E. *et al.* (2018) Rapid and dynamic alternative splicing impacts the arabidopsis cold response transcriptome. *Plant Cell*, **30**, 1424–1444.

Calixto, C.P.G., Simpson, C.G., Waugh, R. & Brown, J.W.S. (2016) Alternative splicing of barley clock genes in response to low temperature. *PLoS One*, **11**, e0168028.

Calixto, C.P.G., Tzioutziou, N.A., James, A.B., Hornyik, C., Guo, W., Zhang, R. *et al.* (2019) Cold-dependent expression and alternative splicing of Arabidopsis long non-coding RNAs. *Frontiers in Plant Science*, **10**, 235.

Chamala, S., Feng, G., Chavarro, C. & Barbazuk, W.B. (2015) Genome-wide identification of evolutionarily conserved alternative splicing events in flowering plants. *Frontiers in Bioengineering and Biotechnology*, **3**, 33.

Chen, L., Shi, X., Nian, B. *et al.* (2020) Alternative splicing regulation of anthocyanin biosynthesis in camellia sinensis var. assamica Unveiled by PacBio Iso-Seq. *G3: Genes|Genomes|Genetics*, **10**, 2713–2723.

Chow, K.-S., Khoo, J.-S., Mohd.-Zainuddin, Z., Ng, S.-M. & Hoh, C.-C. (2019) Utility of PacBio Iso-Seq for transcript and gene discovery in Hevea latex. *Journal of Rubber Research*, **22**, 169–186.

Civáň, P. & Švec, M. (2009) Genome-wide analysis of rice (Oryza sativa L. subsp. japonica) TATA box and Y Patch promoter elements. *Genome*, **52**, 294–297.

Cocquet, J., Chong, A., Zhang, G. & Veitia, R.A. (2006) Reverse transcriptase template switching and false alternative transcripts. *Genomics*, **88**, 127–131.

Dawson, I.K., Russell, J., Powell, W., Steffenson, B., Thomas, W.T. & Waugh, R. (2015) Barley: a translational model for adaptation to climate change. *New Phytologist*, **206**, 913–931.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. *et al.* (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Entizne, J.C., Guo, W., Calixto, C.P.G., Spensley, M., Tzioutziou, N., Zhang, R. and Brown, J.W.S. (2020) TranSuite: a software suite for accurate translation and characterization of transcripts. *bioRxiv*, 2020.12.15.422989.

Hackl, T., Hedrich, R., Schultz, J. & Förster, F. (2014) Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, **30**, 3004–3011.

Holmes, I. & Durbin, R. (1998) Dynamic programming alignment accuracy. *Journal of Computational Biology*, **5**, 493–504.

I.B.S.C. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–717.

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C. *et al.* (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**, 284–289.

Joshi, C.P. (1987) An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucleic Acids Research*, **15**, 6643–6653.

Kalyna, M., Simpson, C.G., Syed, N.H., Lewandowska, D., Marquez, Y., Kusenda, B. *et al.* (2012) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Research*, **40**, 2454–2469.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**, R36.

Kintlová, M., Blavet, N., Cegan, R. & Hobza, R. (2017) Transcriptome of barley under three different heavy metal stress reaction. *Genomics Data*, **13**, 15–17.

Koskinen, P., Törönen, P., Nokso-Koivisto, J. & Holm, L. (2015) PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics*, **31**, 1544–1552.

Kozak, M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, **15**, 8125–8148.

Kuo, R.I., Cheng, Y., Smith, J., Archibald, A.L. & Burt, D.W. (2020) Illuminating the dark side of the human transcriptome with TAMA Iso-Seq analysis. *BMC Genomics*, **21**, 751.

Laloum, T., Martín, G. & Duque, P. (2018) Alternative splicing control of abiotic stress responses. *Trends in Plant Science*, **23**, 140–150.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Loader, C. (2000) Fast and accurate computation of binomial probabilities. *VASA*, **2**, 1–10.

Madeira, F., Park, Y.M., Lee, J. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, **47**, W636–W641.

Mao, S., Pachter, L., Tse, D. & Kannan, S. (2020) RefShannon: a genome-guided transcriptome assembler using sparse flow decomposition. *PLoS One*, **15**, e0232946.

Maretty, L., Sibbesen, J.A. & Krogh, A. (2014) Bayesian transcriptome assembly. *Genome Biology*, **15**, 501.

Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A. & Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research*, **22**, 1184–1195.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T. *et al.* (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433.

Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C.S. *et al.* (2021) Long-read sequence assembly: a technical evaluation in barley. *The Plant Cell.*, **33**, 1888–1906.

Matsumoto, T., Tanaka, T., Sakai, H., Amano, N., Kanamori, H., Kurita, K. *et al.* (2011) Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiology*, **156**, 20–28.

Mejía-Guerra, M.K., Li, W., Galeano, N.F., Vidal, M., Gray, J., Doseff, A.I. *et al.* (2015) Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. *The Plant Cell*, **27**, 3309–3320.

Monat, C., Padmarasu, S., Lux, T., Wicker, T., Gundlach, H., Himmelbach, A. *et al.* (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biology*, **20**, 284.

Morton, T., Petricka, J., Corcoran, D.L., Li, S., Winter, C.M., Carda, A. *et al.* (2014) Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *Plant Cell*, **26**, 2746–2760.

Nakamura, M., Tsunoda, T. & Obokata, J. (2002) Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator. *The Plant Journal*, **29**, 1–10.

Nam, D.K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T. *et al.* (2002) Oligo (dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6152–6156.

Neve, J., Patel, R., Wang, Z., Louey, A. & Furger, A.M. (2017) Cleavage and polyadenylation: ending the message expands gene regulation. *RNA Biology*, **14**, 865–890.

Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D. *et al.* (2020) Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife*, **9**, e49658.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, **14**, 417–419.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. & Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290–295.

Proudfoot, N.J. (2011) Ending the message: poly(A) signals then and now. *Genes & Development*, **25**, 1770–1782.

Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rapazote-Flores, P., Bayer, M., Milne, L., Mayer, C.D., Fuller, J., Guo, W. *et al.* (2019) BaRTv1.0: an improved barley reference transcript dataset to determine accurate changes in the barley transcriptome using RNA-seq. *BMC Genomics*, **20**, 968.

Ren, P., Meng, Y., Li, B., Ma, X., Si, E., Lai, Y. *et al.* (2018) Molecular mechanisms of acclimatization to phosphorus starvation and recovery underlying full-length transcriptome profiling in barley (Hordeum vulgare L.). *Frontiers in Plant Science*, **9**, 500.

Rhoads, A. & Au, K.F. (2015) PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, **13**, 278–289.

Salmela, L. & Rivals, E. (2014) LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, **30**, 3506–3514.

Salmela, L., Walve, R., Rivals, E. & Ukkonen, E. (2017) Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, **33**, 799–806.

Schwacke, R., Ponce-Soto, G.Y., Krause, K., Bolger, A.M., Arsova, B., Hallab, A. *et al.* (2019) MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, **12**, 879–892. https://www.sciencedirect.com/science/article/pii/S1674205219300085

Shao, M. & Kingsford, C. (2017) Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*, **35**, 1167–1169.

Shen, W., Le, S., Li, Y. & Hu, F. (2016) SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, **11**, e0163962.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Simpson, C.G., Fuller, J., Maronova, M., Kalyna, M., Davidson, D., McNicol, J. *et al.* (2008) Monitoring changes in alternative precursor messenger RNA splicing in multiple gene transcripts. *Plant Journal*, **53**, 1035–1048.

Soneson, C., Love, M.I. & Robinson, M.D. (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*, **4**, 1521.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, **7**, 562–578.

Ushijima, T., Hanada, K., Gotoh, E., Yamori, W., Kodama, Y., Tanaka, H. *et al.* (2017) Light controls protein localization through phytochrome-mediated alternative promoter selection. *Cell*, **171**, 1316–1325.e12.

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y. *et al.* (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, **7**, 11708.

Wang, K., Wang, D., Zheng, X., Qin, A., Zhou, J., Guo, B. *et al.* (2019) Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. *Nature Communications*, **10**, 4717.

Wang, M., Wang, P., Liang, F., Ye, Z., Li, J., Shen, C. *et al.* (2018) A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *The New Phytologist*, **217**, 163–178.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T. *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, **37**, 1155–1162.

Wu, T.D. & Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Yi, L., Pimentel, H., Bray, N.L. & Pachter, L. (2018) Gene-level differential analysis at transcript-level resolution. *Genome Biology*, **19**, 53.

Zhang, R., Calixto, C.P.G., Marquez, Y., Venhuizen, P., Tzioutziou, N.A., Guo, W. *et al.* (2017) A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Research*, **45**, 5061–5073.

Zhang, R., Calixto, C.P.G., Tzioutziou, N.A., James, A.B., Simpson, C.G., Guo, W. *et al.* (2015a) AtRTD – a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in Arabidopsis thaliana. *New Phytologist*, **208**, 96–101.

Zhang, R., Kuo, R., Coulter, M., Calixto, C.P.G., Entizne, J.C., Guo, W. et al. (2022). A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. *Genome Biology*, **23**. https://doi.org/10.1186/s13059-022-02711-0

Zhang, S., Li, R., Zhang, L., Chen, S., Xie, M., Yang, L. *et al.* (2020) New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. *Nucleic Acids Research*, **48**, 7700–7711.

Zhang, Y., Gu, L., Hou, Y., Wang, L., Deng, X., Hang, R. *et al.* (2015b) Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. *Cell Research*, **25**, 864–876.