



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/216315/>

Version: Published Version

Article:

Kara, Mehmet Fatih, Guo, Wenbin, Zhang, Runxuan et al. (2024) LsRTDv1, a reference transcript dataset for accurate transcript-specific expression analysis in lettuce. The Plant journal. ISSN: 1365-313X

<https://doi.org/10.1111/tpj.16978>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:



<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

RESOURCE

LsRTDv1, a reference transcript dataset for accurate transcript-specific expression analysis in lettuce

Mehmet Fatih Kara^{1,†}, Wenbin Guo^{2,†}, Runxuan Zhang^{2,*}  and Katherine Denby^{1,*} 

¹Biology Department, Centre for Novel Agricultural Products (CNAP), University of York, Wentworth Way, York YO10 5DD, UK, and

²Information and Computational Sciences, James Hutton Institute, Dundee DD2 5DA, UK

Received 25 January 2024; revised 20 June 2024; accepted 31 July 2024.

*For correspondence (e-mail runxuan.zhang@hutton.ac.uk and katherine.denby@york.ac.uk).

[†]These authors contributed equally to this research.

SUMMARY

Accurate quantification of gene and transcript-specific expression, with the underlying knowledge of precise transcript isoforms, is crucial to understanding many biological processes. Analysis of RNA sequencing data has benefited from the development of alignment-free algorithms which enhance the precision and speed of expression analysis. However, such algorithms require a reference transcriptome. Here we generate a reference transcript dataset (LsRTDv1) for lettuce (cv. Saladin), combining long- and short-read sequencing with publicly available transcriptome annotations, and filtering to keep only transcripts with high-confidence splice junctions and transcriptional start and end sites. LsRTDv1 identifies novel genes (mostly long non-coding RNAs) and increases the number of transcript isoforms per gene in the lettuce genome from 1.4 to 2.7. We show that LsRTDv1 significantly increases the mapping rate of RNA-seq data from a lettuce time-series experiment (mock- and *Botrytis cinerea*-inoculated) and enables detection of genes that are differentially alternatively spliced in response to infection as well as transcript-specific expression changes. LsRTDv1 is a valuable resource for investigation of transcriptional and alternative splicing regulation in lettuce.

Keywords: lettuce, transcriptome, alternative splicing, RNA-seq analysis, transcript isoforms, gene expression, transcription start and end sites, *Lactuca sativa*.

INTRODUCTION

RNA sequencing (RNA-seq) technology has become a cornerstone of modern biosciences research, enabling profiling of global gene expression in a high-throughput and quantitative manner and driving major breakthroughs in understanding a wide range of biological phenomena. Analysis of RNA-seq data can be initiated by mapping raw sequencing reads to a reference genome assembly and quantifying transcript abundance based on the reads mapped to annotated coordinates (Stark et al., 2019). Alternatively, a pseudo-alignment algorithm, KALLISTO (Bray et al., 2016), and lightweight alignment algorithm, SALMON (Patro et al., 2017), are highly accurate and fast and outperform reference-based methods. These efficient methods require, and their accuracy depends upon, a comprehensive and well-annotated transcriptome, which is lacking for most plant species (Brown et al., 2017). Transcriptome

annotations are often generated *in silico* (from known transcript features and homology of coding sequences to those in other species) and/or with short-read RNA-seq data. Although a large proportion of the coding sequences in a genome can be annotated in this way, these annotations do not typically provide sufficient information on individual transcript isoforms of a gene (Brown et al., 2017). This limits the ability of Kallisto and Salmon to quantify different transcript isoforms and for downstream analysis of RNA-seq data to generate information on transcript usage and alternative splicing (AS).

Alternative splicing of precursor messenger RNAs (pre-mRNAs) plays a key role in shaping the dynamic landscape of gene expression in eukaryotic organisms. AS events produce multiple transcript variants from a single intron-containing gene by retaining introns in the mature mRNA or selecting different splice sites during intron

excision from nascent RNAs (Black, 2003; Lee & Rio, 2015). While the majority of intron-containing genes in plants undergo AS events, intron retention predominates as opposed to exon skipping in animals (Chamala et al., 2015; Filichkin et al., 2010; Marquez et al., 2012; Ner-Gaon et al., 2004, 2007). AS significantly amplifies transcript diversity and influences gene function by generating transcript isoforms that produce proteins with distinct functions or unproductive isoforms with premature termination codons (PTCs) that incur rapid degradation through the nonsense-mediated decay (NMD) RNA surveillance pathway (Drechsel et al., 2013; Kalyna et al., 2012). In plants, AS is a key mechanism in a number of biological processes including root differentiation (Zhang & Mount, 2009), circadian rhythm and light perception (Sanchez et al., 2010; Shikata et al., 2014) and the vegetative-to-reproductive switch (Capovilla et al., 2017). Additionally, AS appears to be instrumental in the adaptive potential of plants facing environmental challenges, regulating responses to cold (Calixto et al., 2018, 2019), salinity (Ding et al., 2014) and heat stress (Ling et al., 2018). AS of key defence genes is required for the regulation of immunity (Zhang et al., 2014), with pathogens employing effectors to subvert host AS machinery (Huang et al., 2017). Hence, AS is a fundamental regulatory process in plants, modulating responses in cell-type-specific, stimuli-specific, or genotype-specific manners (Martín et al., 2021; Vaneechoutte et al., 2017). *Lactuca sativa* L. (lettuce) is a leafy vegetable crop of global importance nutritionally and economically. Although RNA-seq is widely used in lettuce research (Guo et al., 2023; Kumar et al., 2022; Pink et al., 2022; Smoleń et al., 2023; Zhang, Su, et al., 2017), our understanding of AS in lettuce remains very limited with only a few case studies to date and two of these investigating AS of a single gene (Ner-Gaon et al., 2007; Sawada et al., 2012; Zhang, Qian, et al., 2022).

The study of AS and differential transcript usage (DTU) is facilitated by transcript-specific expression analysis, requiring a transcriptome annotation which captures the full complexity of the transcriptome, including the diversity resulting from post-transcriptional processing of pre-mRNAs as well as alternative transcription start sites (TSS) and transcription end sites (TES). Such reference transcript datasets (RTDs) were constructed for Arabidopsis [AtRTD1 (Zhang et al., 2015) followed by AtRTD2 (Zhang, Calixto, et al., 2017)] and barley [BaRTv1.0 (Rapazote-Flores et al., 2019)] unifying transcriptome assemblies derived from multiple libraries of short-read RNA-seq data. Leveraging the high read coverage across exon-intron boundaries, these RTDs offered enhanced precision in determining splice junctions (SJs). However, despite stringent filtering and quality control, they contain misassembled transcripts owing to the inherent limitations of short-read assemblies. Reduced read coverage at the

end of transcripts leads to errors in determining untranslated regions (UTRs), heterogeneous coverage along the transcript can lead to segmentation, and it is very difficult to determine which AS events occur in the same transcript isoform (Hayer et al., 2015; Kainth et al., 2023; Stark et al., 2019; Steijger et al., 2013). Long-read RNA-seq can deliver full-length transcripts, providing empirical evidence of transcript structure rather than relying on inferential transcript reconstruction from multiple short reads. In this way, long-read RNA-seq can resolve the precise combination of splicing events in individual isoform variants with a study in Arabidopsis demonstrating Pacific Biosciences Iso-seq outperformed Oxford Nanopore Technology in distinguishing diverse isoforms (Cui et al., 2020). Long-read data are also superior to short-read for determining TSS and TES due to homogeneous coverage along the entire transcript, including the termini (Coulter et al., 2022; Kainth et al., 2023; Zhang, Kuo, et al., 2022). Iso-seq technology has been applied to various crop species, including maize and sorghum (Wang, Regulski, et al., 2018), coffee (Cheng et al., 2017), red clover (Chao et al., 2018), cotton (Wang, Wang, et al., 2018), rice (Zhang et al., 2019) and grapevine (Minio et al., 2019). These studies consistently unveiled novel genes and isoforms, particularly non-coding RNAs.

The genome assembly of cultivated lettuce was published in 2017 (Reyes-Chin-Wo et al., 2017) with an updated genome version (version 11) available on NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002870075.4/). Here, we introduce the first lettuce reference transcript dataset (LsRTDv1) integrating long-read Iso-seq and short-read RNA-seq of diverse tissue and treatment samples from lettuce (cv. Saladin) with the GenBank and RefSeq transcript annotations, using stringent quality measures. The final LsRTDv1 included 179 404 non-redundant transcripts encoded by 65 724 genes, greatly expanding the existing lettuce transcriptome and increasing the number of transcripts per gene from 1.4 to 2.7. LsRTDv1 identifies 3696 novel gene models, predominantly long non-coding RNAs, absent in both GenBank and RefSeq annotations. Re-analysis of a previously published time-series data set from lettuce after pathogen infection, demonstrated the ability of LsRTDv1, compared to the GenBank and RefSeq annotations, to identify differential AS and DTU.

RESULTS

Construction of LsRTDv1

We generated a lettuce Reference Transcript Dataset (LsRTDv1) by integrating transcript assemblies from short- and long-read RNA sequencing data with existing lettuce genome annotations. RNA sequencing data was generated from 23 different lettuce samples capturing different tissues, ages of plant and treatments (Table S1). The 23

Table 1 Comparison of the gene and transcript statistics of the different lettuce transcriptomes

	LsRTDv1	Iso-seq	RNA-seq	RefSeq	GenBank
Genome covered bases	94 100 358	48 281 951	49 287 718	69 956 771	60 348 835
Gene number	65 724	26 952	23 735	47 870	44 232
Multi-isoform gene number	27 866	19 189	13 095	8142	0
Mono-exon gene number	15 139	4016	2083	13 447	5509
Multi-exon gene number	50 585	22 936	21 652	34 423	38 723
Transcript number	179 404	107 032	53 298	69 048	44 232
Mono-exon transcript number	31 676	18 847	3501	13 501	5509
Multi-exon transcript number	147 728	88 185	49 797	55 547	38 723
Transcript number per gene	2.7	4.0	2.3	1.4	1.0
Exon number	1 013 775	641 837	333 034	380 151	230 831
Exon number per transcript	5.7	6.0	6.2	5.5	5.2
Exon average length	282.6	274.8	306.8	318.4	261.4
Intron number	834 371	534 805	279 736	311 103	186 599
Intron number per transcript	4.7	5.0	5.2	4.5	4.2
Intron average length	304.4	245.7	266.9	406.4	252.8
Intron min length	10	30	60	1	10
Intron max length	149 269	31 061	14 950	149 269	13 097
Transcript N50	1963	1870	2126	2255	1916
Transcript N90	944	1032	1196	1014	666
Transcript average length (exonic)	1597.0	1648.0	1917.0	1752.7	1364.4
Transcript total length (exonic)	286 513 371	176 385 656	102 173 305	121 023 193	60 348 835
Non-canonical SJ	0	0	0	666	942
Non-canonical transcripts	0	0	0	626	821

LsRTDv1, Iso-seq and RNA-seq transcriptomes were generated in this study. RefSeq and GenBank transcriptomes are available at https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002870075.4/.

samples, all from *L. sativa* cv. Saladin (a European selection of cv. Salinas) were combined equally into seven samples prior to sequencing.

Three hundred and ten million pairs of short reads were obtained from the seven samples (Table S1b). Pre-processing to remove adapters and low-quality reads resulted in 309 million reads (99.4% of the total) being retained. These were mapped to the most recent lettuce reference genome (Lsat_Salinas_v11, https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002870075.4/) for transcriptome assembly, with an average mapping rate of 86.7%. To maximise coverage, we assembled the transcripts using STRINGTIE (Pertea et al., 2015) and SCALLOP (Shao & Kingsford, 2017) from mapped read data across all seven samples, resulting in a preliminary count of 787 665 transcripts. Given the potential for misassemblies (Zhang, Calixto, et al., 2017; Zhang, Kuo, et al., 2022), we employed RTDMAKER (v1.0.0, <https://github.com/anoconda/RTDmaker>) to eliminate problematic assemblies including redundant transcripts, fragmented transcripts, transcripts with incorrect SJs, and low-expressed transcripts. The same methods were used to generate high-quality reference transcriptomes from short-read sequencing data for Arabidopsis and barley (Rapazote-Flores et al., 2019; Zhang, Calixto, et al., 2017). The final RNA-seq lettuce transcriptome consisted of 23 735 genes and 53 298 transcripts, exhibiting an average of 2.3 transcripts per gene (Table 1).

For the Iso-seq long-read data, we used the Isoseq3 pipeline and obtained an average of 3 151 311 CCS reads across the seven samples (Table S1c). Following barcode trimming and read refinement, we obtained an average of 2 641 236 full-length, non-concatemer (FLNC) reads. These were mapped to the reference genome independently for each sample using Minimap2 (Li, 2018). After collapsing redundant transcripts and merging transcript assemblies from all seven samples through TAMA-COLLAPSE & TAMA-MERGE (Kuo et al., 2020) we obtained a total of 248 376 transcripts. We applied stringent filters to exclude transcripts with problematic SJs, low-confidence TSS and TES respectively, and transcripts supported by only a single read. The final Iso-seq transcriptome consisted of 26 952 genes and 107 032 transcripts, with an average of 4.0 transcripts per gene (Table 1).

The RefSeq annotation of the lettuce reference genome includes 47 870 genes and 69 048 transcripts while the GenBank annotation comprises 44 232 genes, each with a single transcript. Both the RefSeq and GenBank annotations contain transcripts with very rare (typically <0.1%, Pucker & Brockington, 2018) non-canonical SJs (666 in RefSeq and 942 in GenBank, Table 1) which could be technical effects of mismapping, thus these were excluded from downstream analysis. Comparing the Iso-seq, RNA-seq, RefSeq and GenBank transcriptomes, 89 306 distinct introns are common across all four, with a further

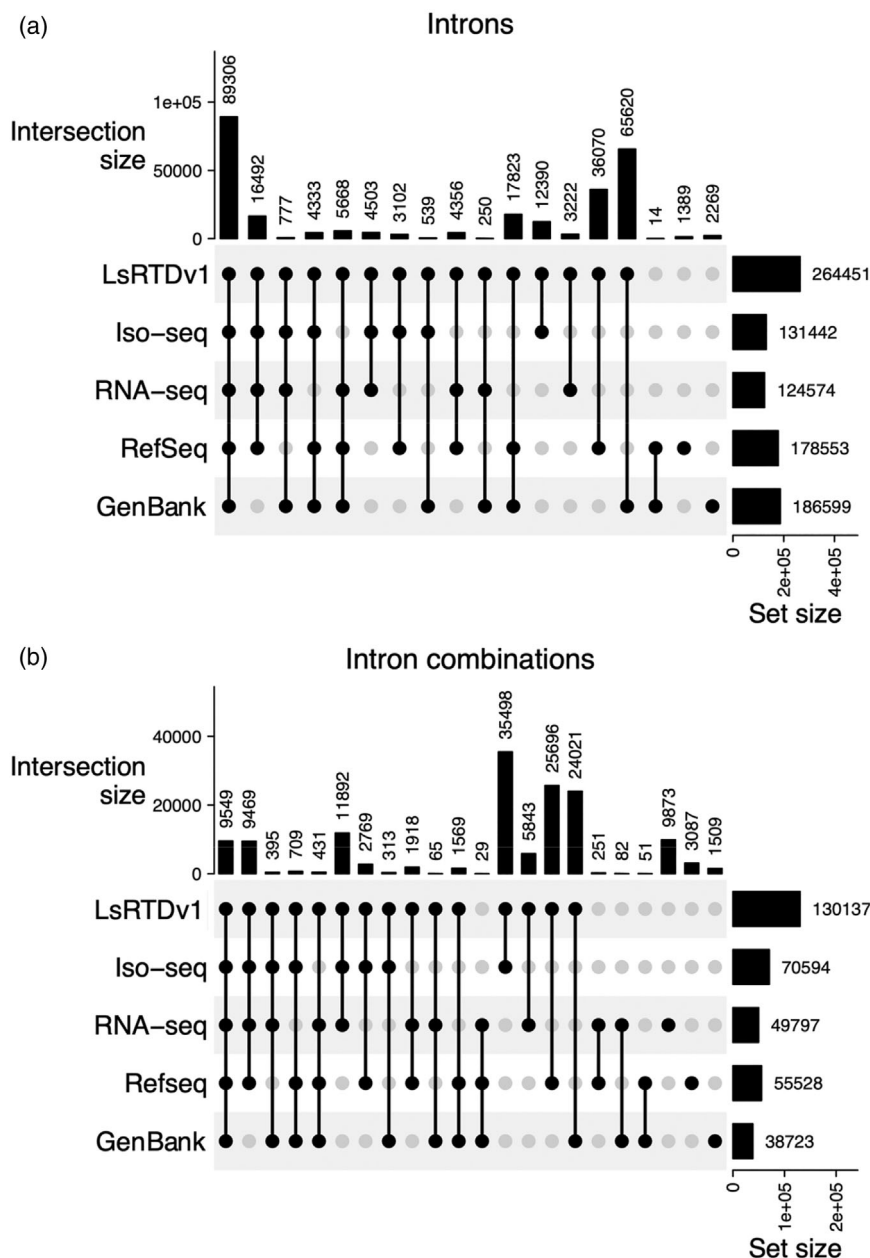


Figure 1. Comparison of introns and splice junction combinations across the LsRTDv1, Iso-seq, RNA-seq, RefSeq and GenBank transcriptome annotations. The number of distinct introns (a) and intron combinations (b) shared between the different transcriptome annotations is indicated. The total number of intron combinations is lower than the total number of transcripts as only multi-exon transcripts are included and transcription start sites/transcription end sites variation is not considered.

27 270 shared across any three (Figure 1a). In terms of intron combinations (a way of comparing individual transcripts), only 20 553 were shared across three or more of the transcriptomes, with the majority of intron combinations present in the Iso-seq assembly (Figure 1b) highlighting the improved transcript diversity captured by long-read sequencing. Distinct transcript structures were observed between the Iso-seq and RNA-seq transcriptomes, despite the data coming from the same samples, illustrating the

power of long-read RNA sequencing to identify novel transcripts (Figure 1).

The RNA-seq, RefSeq and GenBank transcriptomes were compared to the Iso-seq assembly to identify transcripts that introduced novel SJs or gene loci. These transcripts were subsequently combined with the Iso-seq RTD (Figure S1), collapsing overlapping transcripts, to generate LsRTDv1 and ensure that LsRTDv1 captured a more comprehensive transcriptome than our initial 23 samples (Figure 1). The final

LsRTDv1 contains 65 724 genes and 179 404 transcripts, surpassing the scope of the individual transcriptomes (Table 1). On average, each gene in LsRTDv1 has 2.7 transcripts. Furthermore, despite stringent filtering of transcripts in the short- and long-read assemblies, LsRTDv1 offers the broadest genomic coverage with 94 287 718 bases, ~50–100% greater than that of the Iso-seq, RNA-seq, RefSeq, or GenBank transcriptomes (Table 1).

In this study, we have adopted similar methods and pipelines to those we developed for transcript assembly of *Arabidopsis* AtRTD3 (Zhang, Kuo, et al., 2022) and barley BaRTv2 (Coulter et al., 2022). This includes two innovative computational approaches to improve the accuracy and completeness of transcript assemblies. Firstly, we have identified that mismatches to the genome reference sequence around SJs are highly indicative of erroneously mapped SJs. Thus, analysis of mismatches around the SJs is an effective measure to identify high-confidence SJs and remove transcripts containing low-quality SJs. Secondly, assuming TSS and TES of degraded mRNAs would be more randomly distributed than complete transcripts, we developed a probability-based method to effectively remove transcript fragments. The inclusion of only high-confidence TSS and TES should enhance the accuracy of these sites in the Iso-seq transcriptome. To assess this in lettuce, we calculated the occurrence of sequence motifs associated with TSS [T/A rich cis-regulatory TATA box (Morton et al., 2014); transcription activation Initiator motif (Nakamura et al., 2002); plant promoter pyrimidine (Y)-patch (Yamamoto et al., 2007); Kozak translation start site (Kozak, 1987)] in the four individual transcriptome assemblies/annotations and the final LsRTDv1. A TATA box motif 25–35 bp upstream of the TSS is present in the Iso-seq RTD approximately 2.8-fold more often than in the RNA-seq and RefSeq annotations, and 1.2-fold more often than in GenBank (Figure S2a). Similarly, the Initiator motif, which peaks in occurrence immediately around the TSS, is present 2.4, 4 and 1.6 fold more frequently in the Iso-seq transcript assembly, compared to the RNA-seq, RefSeq and GenBank respectively (Figure S2b). The Y-patch motif, which has a wide distribution between –100 and the TSS in plant genomes, shows a significant peak in the Iso-seq transcriptome at –10 bp (Figure S2c). The Kozak motif functions as a translation initiation site and would be expected to be at a variable distance downstream of the TSS (reflecting 5' UTR sequences). The GenBank annotation exhibits a pronounced Kozak motif peak at position 0 bp, indicating a significant absence of UTRs in its transcripts (Figure S2d).

A similar analysis for motifs associated with the TES-cleavage factor Im binding site motif (Neve et al., 2017) found 60–100 nucleotides upstream of the TES in plant mRNA (Li & Hunt, 1997) and AAUAAA, a eukaryotic 3' polyadenylation signal motif typically 10–30 nucleotides

upstream of the TES (Proudfoot, 2011) – demonstrated higher peaks in occurrence for both motifs in the Iso-seq transcriptome compared to RNA-seq, RefSeq and GenBank (Figure S2e,f). Overall, the increased frequency of these TSS/TES motifs in the Iso-seq assembly highlight its superior accuracy in detecting TSS and TES over the other transcriptomes. When constructing LsRTDv1, we incorporated the high-confidence TSS and TES from the Iso-seq assembly resulting in a robust resource for studying transcriptional regulation and gene expression.

LsRTDv1 genes and transcript diversity

We assessed the gene-level overlap of LsRTDv1 with the GenBank and RefSeq annotations using BEDTOOLS (intersect-wao) (Quinlan & Hall, 2010) (Table S2a). A gene from LsRTDv1 was classified as a pre-existing gene if its overlap with a gene from GenBank or RefSeq exceeded 30% (of both gene lengths). LsRTDv1 includes 95% of genes from GenBank and 97% from RefSeq transcriptomes. The remaining genes in GenBank and RefSeq that were not kept in LsRTDv1 were typically excluded due to being fragmentary to longer gene models or the presence of non-canonical SJs. In addition, 457 genes in LsRTDv1 were split into two or more genes in the GenBank and/or RefSeq annotations (Table S2c,d). Finally, a gene model was tagged as a novel gene if it had no overlap or the overlap covers <30% of a pre-existing gene in GenBank or RefSeq assemblies. LsRTDv1 includes 3696 novel genes that encode 6306 transcripts (Table S2e). The majority (80%) of these are supported by full-length Iso-seq reads.

The TRANSUITE software (Entizne et al., 2020) was used to annotate structural features of genes and transcripts in LsRTDv1 as well as to perform *in silico* translations of all transcripts (Table S3). We also provide predicted functional annotation of all LsRTDv1 genes (Table S4). Of the 65 724 gene models present in LsRTDv1, 50 985 (77.6%) were classified as protein-coding (Table S5). The remaining 14 739 (22.4%) were categorised as non-protein-coding due to either lacking an open read frame (ORF) with absence of an authentic AUG start codon, or the longest ORF from the transcript was shorter than 100 amino acids. 80% of the protein-coding (and 48.8% of non-protein coding) genes are comprised of multiple exons with only 55% of these producing multiple transcript isoforms (Table S5). This is lower than previous estimates in *Arabidopsis* (75.0%, Zhang, Kuo, et al., 2022) and barley (73.1%, Coulter et al., 2022), likely reflecting the fact that only 26 952 genes were captured in the Iso-seq data despite the inclusion of samples from a range of tissues, ages and treatments. Sampling of additional tissues and treatments would likely capture more genes and their transcript variants.

The 65 724 genes in LsRTDv1 produce a total of 179 404 transcripts. Among these, 86.4% (155 028) originate from protein-coding genes (Table S5), with 110 749

Type	Structure	Genes	Events
	5' → 3'		
Retained intron (RI)		10,292 (57.8%)	28,888 (36.8%)
Alternative 3' splice-site (A3)		10,208 (57.3%)	21,040 (26.8%)
Alternative 5' splice-site (A5)		7,881 (44.3%)	13,655 (17.4%)
Skipping exon (SE)		4,442 (24.9%)	8,171 (10.4%)
Alternative first exon (AF)		1,436 (8.1%)	3,889 (5.0%)
Alternative last exon (AL)		865 (4.9%)	2403 (3.1%)
Mutually exclusive exons (MX)		265 (1.5%)	450 (0.6%)
Total		17,808	78,496

Figure 2. Frequency of different forms of alternative splicing (AS) events detected in LsRTDv1.

Local AS events were generated by SUPPA2 (Alamancos et al., 2015; Trincado et al., 2018). Constitutive exons are represented as grey boxes, alternative exons or regions as red and orange boxes, and introns as black lines. The thick blue line represents a retained intron. Dashed lines indicate the AS events. The type, structure and number of events as well as number of genes involved are indicated.

transcripts capable of producing full-length proteins; the remaining 44 279 transcripts are predicted to be unproductive transcripts (Table S5). The 110 749 protein-coding transcripts originate from 50 985 genes, hence in addition to the 50 985 primary transcripts (longest ORF), there are an additional 59 764 transcripts from these genes. Forty-three percent of these arise from AS at tandem splice acceptor (NAGNAG) sites or within the 5' or 3' UTR with minimal or no impact on coding sequence. In contrast, 57% encode protein variants (Table S5). The vast majority (88%) of unproductive transcripts exhibit a PTC either alone or with one other nonsense-mediated mRNA decay (NMD) feature (Lykke-Andersen & Jensen, 2015) with the remaining 12% labelled "unproductive unclassified" due to their short ORF (<100 amino acids) or the absence of an ORF (Table S5).

We used the SUPPA2 software (Alamancos et al., 2015; Trincado et al., 2018) to determine the major forms of AS events and their frequencies in LsRTDv1 (Table S6). 17 808 genes in LsRTDv1 – accounting for 64% of multi-isoform genes (Table S5) – contain a total of 78 496 AS events (Figure 2). As seen in Arabidopsis and barley (Coulter et al., 2022; Zhang, Kuo, et al., 2022) intron retention was the most common type of AS event, constituting 36.8% of all events (Figure 2), followed by alternative 3'

splice sites (26.8%), alternative 5' splice sites (17.4%) and exon skipping (10.4%). This is consistent with the observation that intron retention is more prevalent in plants, in contrast to the dominance of exon skipping observed in animals (McGuire et al., 2008).

Identification of lettuce long non-coding RNAs

76.4% of the novel genes in LsRTDv1 (compared to GenBank and RefSeq) produce non-protein-coding transcripts, characterised by a CDS of <100 amino acids. Given this, we investigated the full complement of long non-coding RNAs (lncRNAs) within LsRTDv1. Transcripts designated as from non-protein-coding genes by TRANSSUITE (Table S3) were further evaluated using CPC (Kang et al., 2017), CPAT (Wang et al., 2013) and FEELNC (Wucher et al., 2017), with each tool employing different algorithms to calculate the coding probability of a transcript. To exclude transcripts encoding small peptides, we conducted a BLAST search of reading frames within each candidate lncRNA against the SwissProt database, with transcripts showing no protein hit considered non-coding. By combining the results of the above analyses (CPC, CPAT, FEELNC, TRANSSUITE and SwissProt), we assigned a confidence level for each lncRNA transcript and gene: high-confidence lncRNAs were identified by at least four out of five analyses, moderate

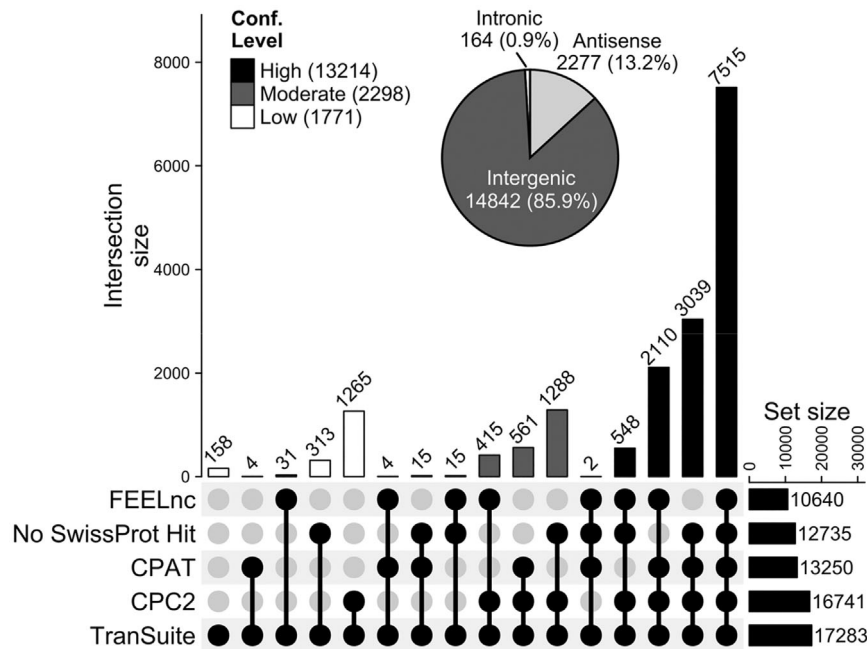


Figure 3. Identification of long non-coding RNAs in LsRTDv1.

The coding potential of candidate lncRNAs was assessed using the CPC2 (Kang et al., 2017), CPAT (Wang et al., 2013) and FEELnc (Wucher et al., 2017) tools. The plot highlights the intersection of transcripts recognised as non-coding by each tool and TranSuite and the set of transcripts showing no similarity to any SwissProt protein. Black, grey and white bars denote the confidence level of lncRNAs predictions based on the above analysis (high, non-coding predicted by four or more analyses; moderate, predicted by three analyses; and low, predicted by only one or two analyses). Pie charts illustrate the classification of lncRNAs based on their genomic positions relative to adjacent protein-coding genes.

confidence lncRNAs were predicted from three of the analyses, and low-confidence lncRNAs were only identified by one or two analyses (Table S7). LsRTDv1 contains 8708 lncRNA genes (encoding 17 283 distinct lncRNAs); by contrast, RefSeq contains 5271 lncRNA genes which yield 12 010 lncRNAs and the GenBank annotation lacks any non-coding entities. Of the LsRTDv1 lncRNAs, the majority (76.5%; 13 214) were deemed high-confidence lncRNAs (Figure 3). Evaluating the genomic positions of the lncRNAs in relation to the closest protein-coding genes revealed that over 85.9% of the lncRNAs are situated in intergenic regions, with 13.2% located in antisense orientation to a protein-coding gene, and 0.9% located within an intron of a protein-coding gene.

LsRTDv1 identifies novel DEGs and transcripts in a time-series RNA-seq dataset

To test the value of the LsRTDv1 in lettuce gene expression analysis, we benchmarked its performance against the GenBank and RefSeq annotations when analysing a time-series RNA-seq dataset. The dataset consists of 14 time points harvested between 9 and 48 h after inoculation of lettuce leaves (cv. Saladin) with the fungal pathogen, *Botrytis cinerea*, or mock inoculation. Three biological replicates were profiled at each time point, generating 84 samples which were sequenced using Illumina short-read

sequencing (Pink et al., 2023). When the reads from each sample were mapped to each transcriptome annotation using Salmon (Patro et al., 2017), the mapping efficiency to LsRTDv1 was markedly higher than to GenBank and slightly (but significantly, $P < 0.0001$) higher than to the RefSeq annotation (Figure S3).

The software 3D RNA-seq (Guo et al., 2021) was used for expression analysis with the transcript quantification data from SALMON, and the same parameters for filtering low-expressed transcripts and genes were used for all three transcriptome annotations. The use of LsRTDv1 increased the number of expressed genes in the time-series data set by 4678 and 1540 compared to GenBank and RefSeq, respectively (Figure 4). Furthermore, the number of transcripts expressed in the time-series dataset increased more than four-fold when using LsRTDv1 compared to GenBank, and more than two-fold compared to RefSeq (Figure 4), highlighting the ability of LsRTDv1 to enable expression of diverse transcript variants to be detected.

Of the 3696 novel genes in LsRTDv1 (compared to GenBank and/or RefSeq annotations) 2167 (58.6%) were found to be expressed in the time-series RNA-seq data. These 2167 genes exhibited a median expression across the time-series data set that was a quarter of the median expression of expressed pre-existing genes (Figure S4),

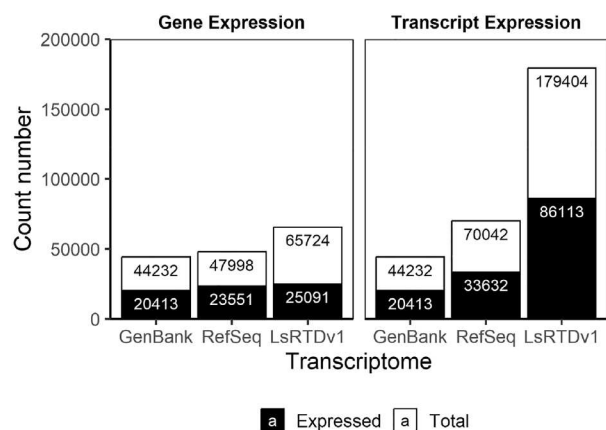


Figure 4. A comparison of expressed genes and transcripts from the use of each reference transcriptome.

Based on mean–variance trend analysis, ≥ 1 CPM (count per million reads) in ≥ 3 samples provided an optimal expression cut-off, and a transcript fulfilling these criteria was considered as expressed. A gene was considered as expressed if any of its transcript isoforms was expressed with the above criteria. Black bars denote expressed genes and transcripts, white bars denote the total number of genes and transcripts in each reference transcriptome.

suggesting LsRTDv1 has annotated, and can detect, rarer transcripts, many of which are lncRNAs. lncRNAs are known to be able to mediate transcriptional regulation despite low expression levels (Chekanova, 2015) and although upon inoculation with *B. cinerea*, the transcript per million reads (TPM) values of the lowest expressed novel genes increased, the difference in median expression between the two groups remained.

We identified differentially expressed genes (DEGs) and transcripts (DETs) (adjusted P -value < 0.01) between the *B. cinerea*- and mock-inoculated samples at each time point, again using the quantifications generated from each transcriptome annotation. With a cumulative total of 14 955 DEGs over the full time-series dataset (Table S8a), LsRTDv1 detected an additional 2373 and 970 DEGs compared to GenBank and RefSeq, respectively (Figure S5; Table S8b,c). At almost all time points, LsRTDv1 identified a higher number of DEGs, both up- and down-regulated, during *B. cinerea* infection. Furthermore, LsRTDv1 resulted in a doubling of the cumulative total of DETs, reaching 29 251 (Table S9a; Figure S5), substantially adding to our understanding of the transcriptional response to pathogen infection.

During *B. cinerea* infection, a total of 1184 genes unique to LsRTDv1 (novel genes) were found to be differentially expressed at one or more time point. Figure S6 shows the expression profiles of selected up- and down-regulated novel genes. It is evident that the changes in expression of these novel genes are substantial upon infection, and the chlorophyll *a/b* binding protein orthologue is known to be down-regulated after *B. cinerea* infection in *Arabidopsis* (Windram et al., 2012). Hence, LsRTDv1 has

enabled us to identify new genes that may possess a biologically relevant role in the lettuce defence response.

Approximately 12% (1727) of the DEGs detected using LsRTDv1 are lncRNA genes, of which 792 are unique to LsRTDv1 and 1238 (72%) are high-confidence lncRNAs. The differentially expressed lncRNAs are both up- and down-regulated during *B. cinerea* infection (Figure 5) with similar dynamics to those seen across all DEGs (Figure S5), suggesting a coordinated response with protein-coding genes. The temporal expression profiles of these differential expression (DE) lncRNAs hint at their potential involvement in mediating defence mechanisms against *B. cinerea*.

Overall LsRTDv1 enables an increase in the number of DEGs identified from this time-series data set, compared to GenBank or RefSeq annotations, providing a more comprehensive insight into the molecular interplay between lettuce and *B. cinerea*. Crucially, LsRTDv1 identifies not only novel DEGs (many of which are lncRNAs) but also novel transcripts, many of which are likely to encode protein variants. Thus, through LsRTDv1, we have access to a richer landscape of transcriptome reprogramming during *B. cinerea* infection underpinning lettuce defence against pathogens.

LsRTDv1 detects AS and isoform switches following *B. cinerea* inoculation

Alternative splicing produces diverse transcript isoforms from a single gene by selectively including or excluding different splice sites in a pre-mRNA transcript. Using LsRTDv1, with an average of 2.7 transcripts per gene (179 404 transcripts in total), we investigated regulation at the splicing level, using three metrics: differential AS at the gene level (DAS, a change in AS of a gene between mock- and *B. cinerea*-inoculated samples); DTU (a change in the proportion of each transcript isoform from a single gene); and isoform switching (IS, a point in the *B. cinerea* inoculation time series when the relative abundance of two transcripts from the same gene swaps). At the gene level, LsRTDv1 detected 4265 DAS genes after *B. cinerea* inoculation (Table S10a). In contrast, the RefSeq annotation, limited by its isoform diversity, identified only 245 DAS genes (Table S10b), whilst GenBank (with no isoform variants) could not detect any. Many of the DAS genes show no overall change in expression of the gene with others showing both up- and down-regulation of the gene as well as DAS. At the transcript level, our analysis using LsRTDv1 identified 5807 DTU transcripts in the time-series dataset (Table S11a; Figure S7). In comparison, RefSeq revealed 378 DTU transcripts (Table S11b), and as expected, no DTU transcripts could be detected using the GenBank annotation.

Interestingly, the dynamics of DAS and DTU differ from that of DEGs. Genes/transcripts start to exhibit changes in AS more rapidly after *B. cinerea* inoculation

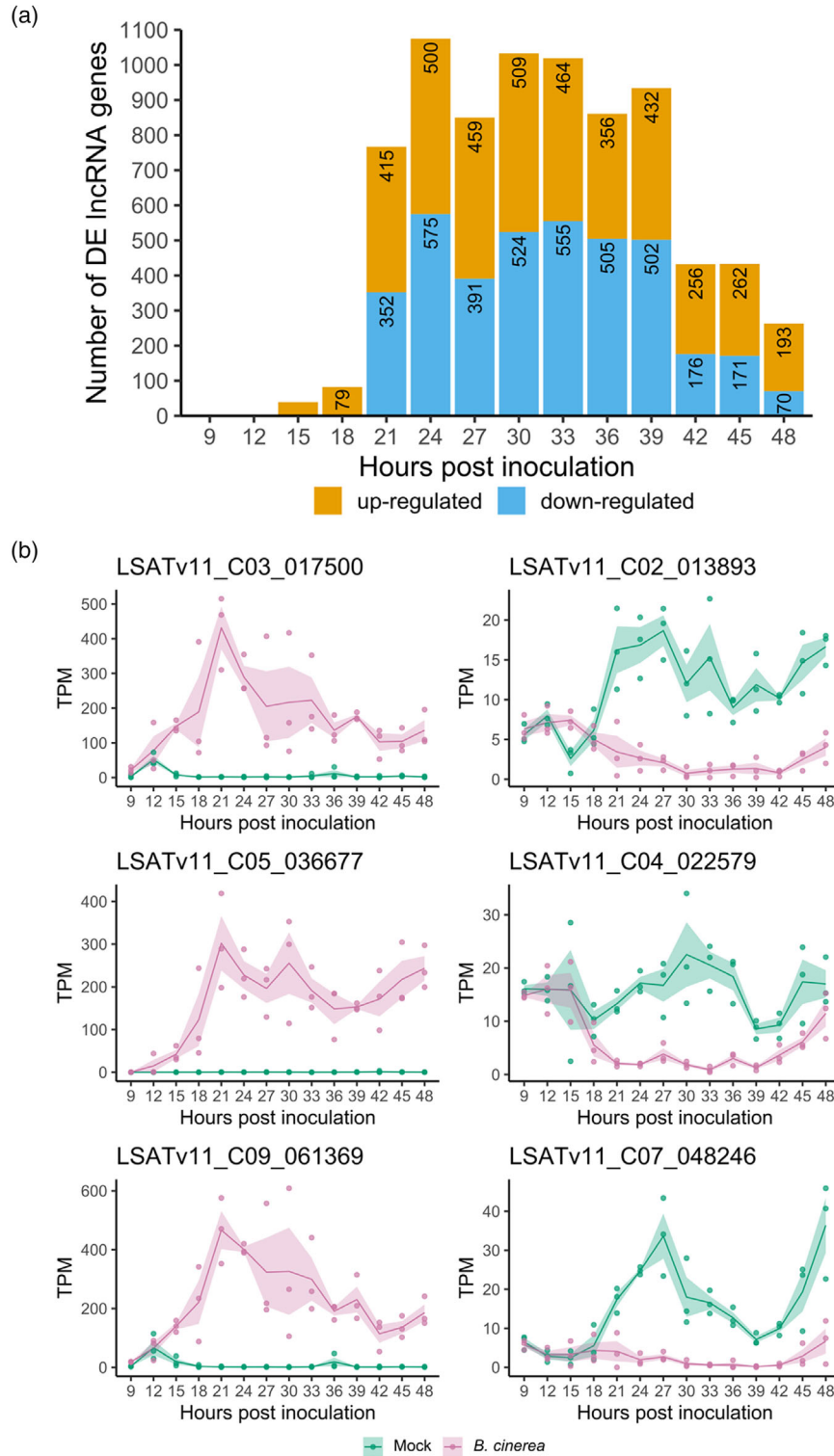


Figure 5. Differential expression of lncRNAs during *Botrytis cinerea* infection. (a) X-axis denotes hours post-inoculation. Orange bars signify up-regulation, while blue bars indicate down-regulation in the *B. cinerea*-inoculated samples compared to the mock-inoculated. (b) Time-series expression profile of selected lncRNA genes differentially regulated in response to *B. cinerea* infection. Pink colour denotes expression after *B. cinerea* inoculation and green mock inoculation. Each dot represents the expression in one of the three individual samples at each time point, with the line showing the mean expression and faded colour indicating standard error. TPM, transcripts per million.

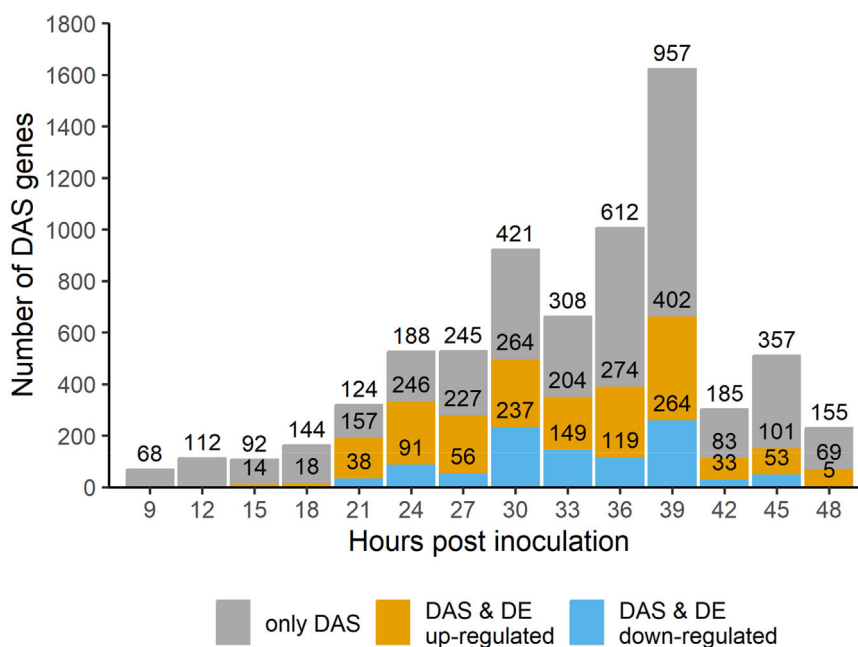


Figure 6. Differentially alternative spliced (DAS) genes in lettuce gradually increase over the course of *Botrytis cinerea* infection. The X-axis denotes hours post-inoculation. Grey bars signify genes that are DAS but show no overall change in expression, orange bars indicate up-regulated DAS genes and blue bars down-regulated DAS genes in the *B. cinerea*-inoculated samples compared to the mock-inoculated.

compared to changes in gene expression, with DAS genes and DTU transcripts detected at 9 and 12 hours post-inoculation (hpi), compared to the first DEGs at 15 hpi (Figure 6; Figure S5). Furthermore, the extent of DAS/DTU increases as infection proceeds, peaking at 39 hpi (Figure 6) whereas maximum numbers of DEGs were seen by 24 hpi and sustained until 39 hpi (Figure S5). Given the importance of early changes in the transcriptome during pathogen infection (Ingle et al., 2015) and that changes in AS appear to happen more quickly after infection than transcriptional change, LsRTDv1 has significant value in being able to detect such post-transcriptional regulation.

Given the scale of DAS/DTU detected, we investigated IS within the time-series dataset using the Time-Series Isoform Switching (TSIS) algorithm (Guo et al., 2017) which detects time points at which two transcript isoforms shift in relative abundance. We assessed IS for the high-abundance transcripts (transcript expression at least 20% of the overall gene expression) of the 4265 DAS genes separately for the mock- and *B. cinerea*-inoculated time-series data. 65 ISs (from 51 genes) were identified exclusively in the mock inoculation data (Table S12a), and 215 ISs (from 171 genes) were unique to *B. cinerea* inoculation (Table S12b). Six genes exhibited ISs in both mock and *B. cinerea*-inoculated samples with all but one of the genes showing an IS within the same time interval in both mock- and *B. cinerea*-inoculated samples (Table S12c). An important observation was that over half of the *B. cinerea*-specific ISs occurred prior to 21 hpi (Figure S8), the critical

time point at which the number of DE genes and transcripts dramatically increases (Figure S5), again suggesting that differential AS is a rapid response to infection. Gene ontology term enrichment analysis of the genes involved in these early ISs (<21 hpi) revealed no significant enrichment; however, the annotations of these genes suggest their potential participation in cellular signalling processes, including protein phosphorylation/de-phosphorylation, methylation, MAPK cascade and hormone signalling (Table S12d).

Examples of the *B. cinerea*-specific ISs are shown in Figure S8(b) and highlight the dynamics of ISs and the potential importance of transcript versus gene expression changes in the lettuce immune response. For example, LSATv11_C01_002961, which encodes a glucan endo-1,3-beta-D-glucosidase, exhibits a marked rise in gene expression in response to infection. Yet, the change in expression is driven by an IS event where the relative expression of an unproductive isoform (i.e. non-protein coding) surpassed that of a protein-coding isoform after 18 hpi (Figure S8b). Similar profiles are seen for the genes LSATv11_C05_038031 and LSATv11_C08_059159 as the relative abundance of the coding isoform diminishes over time, even as the overall gene expression increases. LSATv11_C08_053557 and LSATv11_C09_059900 are examples of genes where IS results in different coding isoforms (with differing protein size) changing in relative abundance likely resulting in differential protein abundance and activity (Figure S8b). These findings highlight the value of a

comprehensive reference transcript dataset to uncover transcript-level regulatory dynamics and provide a more nuanced view of the transcriptome compared to existing annotations.

LsRTDv1 enhances RNA-seq data analysis across multiple lettuce accessions

The time-series dataset analysed above was from the lettuce cultivar used for the transcriptome sequencing and closely related to that used for genome sequencing. Considerable biological insight can be gained from analysis of multiple lettuce accessions (Walley et al., 2017). High-resolution PCR analysis in barley has shown that the BaRTv2 transcriptome (built on the Barke cultivar) was able to improve quantification accuracy in the Morex cultivar (Coulter et al., 2022). Hence, we examined the performance of LsRTDv1 against an RNA-seq dataset from a lettuce diversity set, which encompasses 20 different lettuce accessions as well as one wild relative *Lactuca serriola* (Pink et al., 2022). The RNA-seq data was generated from leaf samples 42 h after inoculation with the fungal pathogen, *Sclerotinia sclerotiorum*, with three biological replicates of each sample. The Salmon mapping efficiency using LsRTDv1, although lower than the time-series data from the Saladin accession, was still higher than that using GenBank or RefSeq transcript annotations (Figure S9). The use of LsRTDv1 increased the number of expressed genes in the samples by 4422 and 1102 compared to GenBank and RefSeq, respectively. As with the time-series data set, the number of transcripts expressed in the diversity set RNA-seq increased by more than 2.5-fold when using LsRTDv1 compared to RefSeq (Figure S9). Hence, LsRTDv1 appears to be of value across multiple lettuce accessions, paving the way for in-depth exploration into genotype-specific isoform usage.

DISCUSSION

Accurate transcript datasets are essential for transcript-level quantification of RNA sequencing data, a fast, accurate and computationally efficient approach (Bray et al., 2016; Patro et al., 2017). Accurate annotation of SJs, TSS and TES also provides essential information to study transcriptional and post-transcriptional regulatory mechanisms such as identification of cis-regulatory elements, polyadenylation and AS. We have generated a new reference transcript dataset for lettuce (LsRTDv1) that integrates data from long- and short-read sequencing of lettuce gene expression from diverse tissues and treatments, with existing publicly available transcript annotations in a non-redundant manner. We have shown experimentally in both Arabidopsis and barley (Coulter et al., 2022; Zhang, Kuo, et al., 2022) that our methods generate a comprehensive transcriptome enabling more accurate quantification of transcript abundance and AS events from RNA-seq data.

LsRTDv1 contains transcripts from 3696 novel genes but more importantly, it identifies an increased number of transcript isoforms from existing genes. LsRTDv1 contains 179 404 distinct transcripts compared to 69 048 in the RefSeq annotation and 44 232 in the GenBank annotation, raising the average number of transcripts per gene from 1.4 to 2.7. 85 284 of these transcripts are predicted to be productive transcripts encoding proteins. Hence our transcript dataset enhances our knowledge of the transcriptome complexity of lettuce, captures a high level of protein diversity and markedly improves our ability to analyse transcript-specific expression and differential AS, as well as providing an improved protein database to guide analysis of proteomic data.

The importance of this resource to existing and future gene expression analyses was illustrated by analysis of a time-series RNA-seq data set of lettuce leaves inoculated with the fungal pathogen *B. cinerea* (Pink et al., 2023). The significantly enhanced mapping rate (compared to RefSeq and GenBank annotations, Figure S3) meant a higher proportion of the RNA-seq reads contributed to the expression analysis, and LsRTDv1 identified an increased number of DEGs compared to the existing annotations (Figure 4) as well as new information on transcript-level expression. The majority of genes showing differential alternative splicing (DAS) did not display any change in overall expression, and hence would be missed in gene-level expression analysis. Being able to examine transcript-specific expression also provides a more accurate insight into the potential downstream impacts of changes in expression, with the isoform switch analysis providing clear examples of where transcripts coding for different proteins from the same gene show very different expression patterns (Figure S8), or where the observed change in gene-level expression is due to an increase in expression of an unproductive transcript (Figure S8). Clearly, this type of transcript-specific information is critical in gaining a more accurate insight into the transcriptome and downstream impacts on protein expression.

A significant proportion of the novel genes annotated in LsRTDv1 are lncRNAs. As seen in other plants (Chekanova, 2015) the majority of these are intergenic, with ~13% antisense transcripts to protein-coding genes and a small proportion (0.9%) located entirely within an intron of a protein-coding gene (Figure 3). lncRNAs in plants are known to play a role in multiple developmental processes, as well as plant responses to biotic and abiotic stress (Domínguez-Rosas et al., 2023) impacting gene expression in multiple ways such as mimicry of miRNA targets, modulating AS and chromatin remodelling (Chekanova, 2015). A well-known example of lncRNA regulation is that of *Flowering Locus C (FLC)*, an Arabidopsis gene which represses the transition from vegetative growth to flowering. An antisense lncRNA (*COOLAIR*) is up-regulated in response to

cold and aids the down-regulation of FLC. In addition, an intronic lncRNA and lncRNA from the promoter region of *FLC* (*COLDAIR* and *COLDWRAP*) recruit an epigenetic silencing complex to the promoter (Kim & Sung, 2017). Another lncRNA, *FLAIL*, acts in trans targeting the spliceosome and modulating AS of target genes to repress flowering (Jin et al., 2023). The annotation of lncRNAs in LsRTDv1 will clearly help in elucidating the role of these versatile regulators in lettuce. lncRNAs typically have low, highly specific expression and are poorly conserved across species making their identification more challenging than that of mRNAs. PLncDB (Jin et al., 2021) is one of the latest lncRNA databases predicting lncRNAs based on publicly available short-read sequencing data. This identified 13 394 lncRNAs mapped to an earlier version of the lettuce genome (Lsat_Salinas_v7). In LsRTDv1 we identified 17 283 distinct lncRNAs originating from 8708 lncRNA loci. Sequence similarity searches indicated that 9808 of these had best hits to *L. sativa* lncRNAs in the RNAcentral database (<https://rnacentral.org/>, which includes PLncDB) (Table S7). The study of lncRNAs in lettuce and plant-pathogen interaction is still in its infancy and the lncRNAs we have shown to be differentially expressed in response to *B. cinerea* inoculation, and their target genes in the case of antisense lncRNAs, need to be investigated further.

As expected, the long-read Iso-seq data were instrumental in identifying transcript isoforms (Figure 1) with the Iso-seq transcriptome annotation having an average of four transcripts per gene. Crucially, long-read sequencing data provides information on the combination of AS events in a single transcript and accurate TSS/TES. In barley, BaRTv2 (generated using Iso-seq and short-read data from the cv. Barke) achieved higher accuracy compared to the earlier BaRTv1 (generated using only short-read data from cv. Morex) when analysing RNA-seq data from Morex (Coulter et al., 2022). Hence, as with lettuce, the improvement in RNA-seq analysis of our RTD methodology and Iso-seq data is evident even across cultivars. In the final LsRTDv1, 57% of protein-coding genes have multiple transcript isoforms. This is on par with the RTD from barley (Coulter et al., 2022) but lower than the latest Arabidopsis RTD (Zhang, Kuo, et al., 2022) in which 79% of protein-coding genes exhibit AS. The Iso-seq transcriptome in Arabidopsis was generated from less than half the number of reads compared to our study (7.36 M mapped reads versus 17.9 M) but consisted of 33 154 genes compared to the lettuce Iso-seq RTD of 26 952. It is likely that the increased transcript diversity in AtrTDv3 reflects the greater breadth of tissue/treatment sampling, and the scale of transcript isoforms in lettuce would be further enhanced by sequencing of additional samples to detect tissue-specific and condition-dependent transcripts. For example, although we included a variety of developmental stages and tissues as well as abiotic and biotic

stress treatments, we did not include flowers or seed in our current analysis.

LsRTDv1 is a valuable transcriptome resource able to improve the accuracy of gene expression analysis in lettuce, enabling quantification of protein-coding isoforms and detection of changes in AS and transcript usage that occur with or without overall changes in gene expression. LsRTDv1 can be further enhanced with the inclusion of additional (particularly long read) sequencing data. Employing single-cell RNA sequencing could detect transcript heterogeneity only present in some cell types, and sequencing of different lettuce accessions could expand LsRTDv1 into a pan-transcriptome able to distinguish between core, variable and unique transcript isoforms within lettuce.

EXPERIMENTAL PROCEDURES

Plant material

Plant samples for PacBio isoform sequencing (Iso-seq) and Illumina mRNA sequencing were all collected from lettuce (*L. sativa* cv. Saladin, a crisphead lettuce selected from cv. Salinas). Lettuce seeds were sown onto 7-cm² pots containing Levington F2 + S compost mix and stratified for 3 days at 4°C in darkness. Plants were grown in individual pots in trays under controlled environment conditions (16 h/8 h day/night photoperiod, 120 $\mu\text{mol m}^{-2} \text{sec}^{-1}$ photosynthetic photon flux density, 20 \pm 2°C constant temperature and 60 \pm 10% relative humidity). A total of 23 tissues were sampled from at least five plants either in different developmental stages or post-stress treatments. All samples were collected in 2 ml RNase-free tubes and flash-frozen in liquid nitrogen.

Cotyledon, hypocotyl and radicle samples were collected from 1-week-old seedlings. First true leaves and primary roots were harvested from 2-week-old young seedlings. Mature leaf and root samples were collected from 6-week-old plants. Later, when plants started cupping/heading, leaf 1–5, leaf 6–10, leaf 11–15 and apical meristem including young leaves forming heads were sampled separately. Chilling, heat-shock, waterlogging, drought and wounding treatments were applied on 6-week-old plants. For chilling, plants were exposed to 4°C for 24 h at dusk, and leaf and root samples collected after treatment. For heat-shock, plants were kept at 38°C for 24 h in the light (135 $\mu\text{mol m}^{-2} \text{sec}^{-1}$) and 90% relative humidity, with leaf and root samples harvested from wilting plants. For waterlogging, plants were submerged under water 1 cm above the soil surface, leaves and roots were collected after 7 days of treatment. For drought, plant irrigation was stopped for 10 days, with leaves and roots collected from partially wilted plants. For wounding, leaves were pierced by needles and leaf samples were collected 3 h afterwards. For pathogen infection, leaf 3 was detached from 6-week-old lettuce, placed onto an agar tray, and inoculated with 4 \times 10 μl droplets of 5 \times 10⁵ *B. cinerea* pepper isolate (Denby et al., 2004) spores per millilitre in 0.5 \times grape juice. Trays were sealed and incubated at 20°C, 16 h/8 h light/dark 80% humidity. Leaf disks enclosing *B. cinerea* lesions were excised 48 hpi.

RNA extraction and sequencing

Total RNA was extracted using the CTAB method (Chang et al., 1993) with RNA samples column-purified using Monarch[®] Total RNA

Miniprep Kit (NEB, including on-column DNase I treatment). RNA samples were pooled into seven libraries in equimass amounts, and pooled libraries were used for Iso-seq™ full-length isoform RNA sequencing and Illumina short-read RNA sequencing. Full-length isoform RNA sequencing (Iso-seq™) was performed by Novogene (UK) Company Limited (Cambridge, UK) using PacBio SMRT Sequel II platform. Illumina short-read transcript sequencing was performed by Novogene using Illumina NovaSeq 6000 platform.

Short-read assembly

The RNA-seq reads of the seven pooled samples were pre-processed with FASTP (Chen et al., 2018) to remove adapters and filter low-quality reads (quality score <20, length <30). Trimmed reads were mapped to the latest lettuce reference genome assembly in NCBI (Lsat_Salinas_v11) using STAR aligner in the 2-pass mode to increase the mapping sensitivity at SJs (Dobin & Gingeras, 2015). Mismatch was set to 1 with minimum and maximum intron sizes of 60 and 15 000 bp respectively. Two transcript assemblers, STRINGTIE (Pertea et al., 2015) and SCALLOP (Shao & Kingsford, 2017), were used to assemble transcripts for each sample. The assemblies were then merged and refined using RTDMAKER (<https://github.com/anoconda/RTDmaker>) to remove low-quality transcripts, including redundant transcripts with identical intron combinations to longer transcripts, fragmented transcripts with length <70% of gene length, transcripts with non-canonical SJs, transcripts with SJs only supported by <5 spliced reads in one sample and low-expressed transcripts with <1 TPM in one sample.

Long-read assembly

We employed the IsoSeq pipeline (<https://github.com/PacificBioSciences/IsoSeq>) to pre-process the Iso-seq data from the seven samples. The circular consensus sequences (CCS) method was used to generate CCS from raw subreads and reads with minimum predicted accuracy <90% were discarded (--min-rq=0.9). Barcodes associated with the CCS reads were eliminated using the lima method. To further refine the reads, Isoseq3 was applied to trim poly(A) tails and identify and remove concatemers. The output of FLNC reads was mapped to the reference genome using MINIMAP2 (Li, 2018). TAMA-COLLAPSE was used to collapse redundant transcript models in each sample with variation at the 5' and 3' ends and at SJs not allowed (-a = 0, -m = 0 and -z = 0) to ensure high accuracy of boundaries. Reads with errors within the 10 bp up- or downstream of a SJ were removed. TAMA-MERGE was used to merge transcript models from the seven samples (Kuo et al., 2020). To improve the quality of the assembly, we implemented well-established methods for SJ and TSS and TES analyses previously used for Arabidopsis AtrTD3 and barley BaRTv2 (Coulter et al., 2022; Zhang, Kuo, et al., 2022). We removed low-quality transcripts that exhibited non-canonical SJs (donor and acceptor motifs deviated from GT/AG, GC/AG and AT/AC) and low-quality SJs (mapping errors \pm 10 nt to the SJ) unless they were also present in the short-read assembly. We applied a binomial test to distinguish high-confidence TSS and TES with a false discovery rate <0.05. For genes with limited read support, statistical testing becomes challenging, hence we also kept TSS/TES if they were supported by at least 2 Iso-seq reads. Redundancy merge was applied to transcripts if they only differed \pm 50 nucleotides at their TSS/TES. In addition, transcripts only supported by a single Iso-seq read were removed from the final dataset.

Integration of multiple annotations

We integrated four transcript annotations: the long-read assembly, short-read assembly and two versions of Lsat_Salinas_v11

genome annotations GenBank (GCA_002870075.4) and RefSeq (GCF_002870075.4). The Iso-seq long-read assembly served as the reliable backbone, while the other three annotations were incorporated in a step-wise manner to improve the RTD completeness (Figure S1). Firstly, the transcripts in the short-read assembly that introduce novel SJs and/or novel gene loci were integrated into the long-read assembly (Figure S1a). Subsequently, we added transcripts from GenBank and RefSeq annotations that contributed novel SJs or gene loci to build the lettuce RTD (LsRTDv1). In cases where two transcripts from GenBank and RefSeq had identical SJ combinations or were mono-exonic transcripts with overlapping regions exceeding 30% of both transcripts, we collapsed them to a single transcript, and the longest TSS and TES were used as the start and end point of the collapsed transcript (Figure S1b). In LsRTDv1, the overlapped transcripts were assigned the same gene ID (Figure S1c). However, if a set of overlapped transcripts entirely resided within the intron region of other transcripts, they were treated as intronic transcripts and assigned with a different gene ID. Where the overlapped transcripts can be divided into multiple groups and the adjacent groups overlapped less than 5% of the group lengths, they were assigned separate gene IDs (Figure S1d). Gene-related motifs were assessed using the genome sequences \pm 500 nucleotides from the TSS and TES. These sequences were scanned for motif sequences using the vmatchPattern function within the Biostrings R package (<https://github.com/Bioconductor/Biostrings>).

Gene and transcript annotation

Structural features of the genes and transcripts in LsRTDv1 were characterised via TRANSUITE (Entizne et al., 2020) software, by running on auto mode with default parameters. Independent built-in modules (FINDLORF, TRANSFIX and TRANSFEAT) (1) determine the longest ORF of each transcript, (2) generate protein translations of the transcripts with iterated translation start site (AUG) fixing/translation cycles, and (3) annotate structural features of genes/transcripts. The major forms of AS events were determined for multi-isoform transcripts utilising SUPPA2 (v2.3) with the 'generateEvents' option, employing the default parameters (Trincado et al., 2018).

We used the AHRD (v3.3.3, <https://github.com/groupschoof/AHRD>) pipeline to assign functional annotations for the protein-coding genes in LsRTDv1 (The Tomato Genome Consortium [TGC], 2012). First, the longest protein translation of each coding gene was searched for protein similarity against Araport11 (Release September 2022), UniProtKB/SwissProt (Release 2023_01) and UniProtKB/TrEMBL (Release 2023_01) protein databases via BLASTP [BLAST+ v2.13.0 (Camacho et al., 2009)]. Protein domains in the InterPro database (<https://www.ebi.ac.uk/interpro/>) (Paysan-Lafosse et al., 2023) were searched via INTERPROSCAN v5.46-81.0 (Jones et al., 2014). Protein homology searches were processed in the AHRD pipeline, and short functional descriptions assigned for each protein-coding gene.

To identify long non-coding RNAs, non-protein coding (i.e. any predicted ORF is <100 aa) transcripts identified by TRANSUITE and longer than 200 nt were considered as candidate lncRNAs. These candidate lncRNA transcripts were searched against the Rfam database v14.9 (Kalvari et al., 2021) using INFERNAL CMSCAN v1.1.4 (Nawrocki & Eddy, 2013), and any transcripts showing homology to tRNA, rRNA, snRNA and snoRNA RNA biotypes with hits of *E*-value <1e-5 were excluded. The coding potential of candidate lncRNAs was calculated using FEELNC v0.2 (Wucher et al., 2017), CPAT v3.0.4 (Wang et al., 2013) and CPC v2.0 (Kang et al., 2017). Sequence similarity of candidate lncRNAs to known proteins was determined via BLASTX against UniProtKB/SwissProt

database (v2023_01) with an E-value threshold set to $1e-10$. lncRNA homology searches were performed using nhmmer search v3.3.2 (Wheeler & Eddy, 2013) against the RNAcentral (v22) database. Intergenic (u), antisense (x) and intronic (i) lncRNAs classes were determined by GFFCOMPARE tool v0.12.6 (Perrea & Perrea, 2020), and transcripts with class code “o”, “p”, “y”, “c” and “s” were removed.

Benchmarking LsRTDv1

Read mapping and quantification processes were carried out using SALMON v1.10.0 (Patro et al., 2017). Prior to read quantification, reference transcripts in GenBank, RefSeq and LsRTDv1 were concatenated with the *B. cinerea* reference transcripts (ASM14353v4) for the time-series RNA-seq data, and with the *S. sclerotiorum* reference transcripts (ASM185786v1) for the diversity set RNA-seq data. Salmon transcript indices were generated using the joint lettuce + pathogen transcripts. Read alignment and transcript quantification was performed with Salmon's mapping-based mode.

Following transcript quantification, the 3D RNA-seq web-tool (Guo et al., 2021) was employed to compare gene/transcript expression between *B. cinerea* and mock-inoculated samples at each time point. To generate read counts and TPM values, Salmon quantification outputs were imported via tximport R package using lengthScaledTPM method (Soneson et al., 2016). Low-expressed transcripts and genes were filtered based on data mean-variance trend analysis (Law et al., 2014): a transcript was considered as expressed, providing it had ≥ 1 count per million (CPM) reads in ≥ 3 samples, and a gene was expressed if any of its transcripts met the above expression criteria. The gene and transcript read counts were normalised to \log_2 CPM using the TMM (weighted trimmed mean of *M*-values) method (Bullard et al., 2010).

In the time-series RNA-seq data, DE, DTU and DAS analyses were conducted using the limma-voom pipeline (Law et al., 2014; Ritchie et al., 2015). All *p*-values were corrected for multiple testing using the Benjamini-Hochberg method (Benjamini & Yekutieli, 2001). A gene/transcript was considered significantly differentially expressed if it had an adjusted *P*-value < 0.01 and a \log_2 fold change ≥ 1 . A transcript was considered a significant DTU transcript if it had adjusted *P*-value < 0.01 and a change in the proportion of spliced transcript within expressed transcripts (Δ PS) ≥ 0.1 . A gene was considered a significant DAS gene if it had an adjusted *P*-value < 0.01 and any of its transcripts were DTU transcripts with Δ PS ratio ≥ 0.1 .

To detect isoform switches in the time-series data, a computational analysis was performed on high-abundance transcripts from all DAS genes using the TSIS R package (Guo et al., 2017). Transcripts with expression less than 20% of the total gene expression at all time points were excluded to avoid including transcripts with noisy low expression levels. The mean expression approach, where the intersection points between average expression of two isoforms of a gene are searched for across time points, was adopted to score each switch point, and the following parameters applied for significance filtering: (1) probability cut-off > 0.5 , (2) difference cut-off > 1 , (3) adjusted *P*-value cut-off < 0.01 , and (4) minimum time in interval > 2 .

AUTHOR CONTRIBUTIONS

The study was conceived and designed by WG, RZ and KD with input from MFK. Experimental work and analysis of lncRNAs as well as the time series and diversity set data

sets was carried out by MFK, construction of the RTD by WG, and further data interpretation performed by RZ and KD. The manuscript was written by MFK and KD with input from WG and RZ. All authors have approved the submission of this manuscript.

ACKNOWLEDGEMENTS

This work was jointly supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) BB/S020160/1 to RZ and the Scottish Government Rural and Environment Science and Analytical Services Division (RESAS) to RZ and WG. MFK was supported by a postgraduate scholarship from the Milli Eğitim Bakanlığı (Ministry of National Education of the Republic of Türkiye) under grant number MEB1416.

CONFLICT OF INTEREST

None of the authors declare a conflict of interest.

DATA AVAILABILITY STATEMENT

The RNA sequencing data is available in the NCBI Sequencing Read Archive under BioProject PRJNA1018253. LsRTDv1 is available in Dryad at <https://doi.org/10.5061/dryad.xwdbrv1m8>. All code used in this study is available at [10.5281/zenodo.11658411](https://zenodo.org/record/11658411).

SUPPORTING INFORMATION

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

Figure S1. Overview of the integration of multiple transcript annotations.

Figure S2. Enrichment of sequence motifs associated with transcription start sites (TSS) and transcription end sites (TES).

Figure S3. The use of LsRTDv1 as reference transcriptome improves the mapping rate of time-series RNA-seq data compared with GenBank and RefSeq annotations.

Figure S4. Expression level of novel and pre-existing genes in LsRTDv1 that are expressed in the time-series dataset.

Figure S5. LsRTDv1 enables detection of a larger number of differentially expressed (DE) genes (a) and transcripts (b) compared to the GenBank and RefSeq annotations over the course of *B. cinerea* infection of lettuce.

Figure S6. Expression profile of selected novel genes in LsRTDv1 over the course of *B. cinerea* infection.

Figure S7. The number of transcripts that exhibit differential transcript usage (DTU) in the *B. cinerea*-inoculated samples compared to mock inoculated at each time point.

Figure S8. Isoform switches (ISs) in the transcriptome profiles of *B. cinerea*- and mock-inoculated lettuce.

Figure S9. LsRTDv1 enhances analysis of RNA-seq data across diverse lettuce accessions.

Table S1. (a) Descriptions of the lettuce samples (cv. Saladin) used for PacBio Iso-seq and Illumina RNA-seq. Read statistics for (b) Illumina RNA-seq libraries and (c) PacBio Iso-seq libraries.

Table S2. Gene-level overlap between (a) GenBank annotation and LsRTDv1, (b) RefSeq annotation and LsRTDv1. Genes fused in LsRTDv1 but fragmented in (c) GenBank, (d) RefSeq. (e) Novel gene models and their transcripts in LsRTDv1.

Table S3. Translation and feature characterisation of LsRTDv1 transcripts.

Table S4. Functional annotation of LsRTDv1 genes using AHRD.

Table S5. Statistics of LsRTDv1.

Table S6. SUPPA2 output identifying alternative splicing events and associated transcript isoforms in LsRTDv1 multi-exonic genes.

Table S7. (a) Long non-coding RNAs identified in LsRTDv1. (b) RefSeq gene IDs for the lncRNAs in LsRTDv1 with overlapping genomic locations. (c) Results of nhmmer sequence homology between lncRNAs and RNAcentral DB.

Table S8. Genes differentially expressed (DE) in lettuce inoculated with *B. cinerea* compared to mock inoculated, identified using (a) LsRTDv1, (b) RefSeq annotation, and (c) GenBank annotation.

Table S9. Transcripts differentially expressed (DE) in lettuce inoculated with *B. cinerea* compared to mock inoculated, identified using (a) LsRTDv1, (b) RefSeq annotation, and (c) GenBank annotation.

Table S10. (a) Genes exhibiting differential alternative splicing (DAS) during infection of lettuce by *B. cinerea* (compared to mock inoculation) detected using (a) LsRTDv1 and (b) RefSeq annotation.

Table S11. Transcripts exhibiting differential transcript usage (DTU) during infection of lettuce by *B. cinerea* detected using (a) LsRTDv1 and (b) RefSeq annotation.

Table S12. (a) Transcript pairs exhibiting isoform switch events in (a) mock-inoculated time-series samples as detected by TSIS analysis, (b) *B. cinerea*-inoculated time-series samples as detected by TSIS analyses, and (c) both *B. cinerea*- and mock-inoculated samples as detected by TSIS analyses. (d) Gene description and GO term annotations of the genes exhibiting isoform switching prior to 21 hpi in response to *B. cinerea* inoculation.

REFERENCES

- 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**(9), 1521–1531. Available from: <https://doi.org/10.1261/rna.051557.115>
- Benjamini, Y. & Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**(4), 1165–1188. Available from: <https://doi.org/10.1214/aos/1013699998>
- Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, **72**(1), 291–336. Available from: <https://doi.org/10.1146/annurev.biochem.72.121801.161720>
- Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5), 525–527. Available from: <https://doi.org/10.1038/nbt.3519>
- 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. *The New Phytologist*, **213**(2), 525–530. Available from: <https://doi.org/10.1111/nph.14208>
- Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**(1), 94. Available from: <https://doi.org/10.1186/1471-2105-11-94>
- 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. *The Plant Cell*, **30**(7), 1424–1444. Available from: <https://doi.org/10.1105/tpc.18.00177>
- 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. Available from: <https://doi.org/10.3389/fpls.2019.00235>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**(1), 421. Available from: <https://doi.org/10.1186/1471-2105-10-421>
- Capovilla, G., Symeonidi, E., Wu, R. & Schmid, M. (2017) Contribution of major FLM isoforms to temperature-dependent flowering in *Arabidopsis thaliana*. *Journal of Experimental Botany*, **68**(18), 5117–5127. Available from: <https://doi.org/10.1093/jxb/erx328>
- 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. Available from: <https://doi.org/10.3389/fbioe.2015.00033>
- Chang, S., Puryear, J. & Cairney, J. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter*, **11**(2), 113–116. Available from: <https://doi.org/10.1007/bf02670468>
- Chao, Y., Yuan, J., Li, S., Jia, S., Han, L. & Xu, L. (2018) Analysis of transcripts and splice isoforms in red clover (*Trifolium pratense* L.) by single-molecule long-read sequencing. *BMC Plant Biology*, **18**(1), 300. Available from: <https://doi.org/10.1186/s12870-018-1534-8>
- Chekanova, J.A. (2015) Long non-coding RNAs and their functions in plants. *Current Opinion in Plant Biology*, **27**, 207–216. Available from: <https://doi.org/10.1016/j.pbi.2015.08.003>
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**(17), i884–i890. Available from: <https://doi.org/10.1093/bioinformatics/bty560>
- Cheng, B., Furtado, A. & Henry, R.J. (2017) Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *GigaScience*, **6**(11), 1–13. Available from: <https://doi.org/10.1093/gigascience/gix086>
- Coulter, M., Entizne, J.C., Guo, W., Bayer, M., Wonneberger, R., Milne, L. *et al.* (2022) BaRTv2: a highly resolved barley reference transcriptome for accurate transcript-specific RNA-seq quantification. *The Plant Journal*, **111**(4), 1183–1202. Available from: <https://doi.org/10.1111/tpj.15871>
- Cui, J., Shen, N., Lu, Z., Xu, G., Wang, Y. & Jin, B. (2020) Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods*, **16**(1), 85. Available from: <https://doi.org/10.1186/s13007-020-00629-x>
- Denby, K.J., Kumar, P. & Kliebenstein, D.J. (2004) Identification of *Botrytis cinerea* susceptibility loci in *Arabidopsis thaliana*. *The Plant Journal*, **38**(3), 473–486. Available from: <https://doi.org/10.1111/j.0960-7412.2004.02059.x>
- Ding, F., Cui, P., Wang, Z., Zhang, S., Ali, S. & Xiong, L. (2014) Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. *BMC Genomics*, **15**(1), 431. Available from: <https://doi.org/10.1186/1471-2164-15-431>
- Dobin, A. & Gingeras, T.R. (2015) Mapping RNA-seq reads with STAR. *Current Protocols in Bioinformatics*, **51**(1), 11.14.1–11.14.19. Available from: <https://doi.org/10.1002/0471250953.bi1114s51>
- Domínguez-Rosas, E., Hernández-Onate, M.A., Fernández-Valverde, S.-L. & Tiznado-Hernández, M.E. (2023) Plant long non-coding RNAs: identification and analysis to unveil their physiological functions. *Frontiers in Plant Science*, **14**, 1275399. Available from: <https://doi.org/10.3389/fpls.2023.1275399>
- Drechsel, G., Kahles, A., Kesarwani, A.K., Stauffer, E., Behr, J., Drewe, P. *et al.* (2013) Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *The Plant Cell*, **25**(10), 3726–3742. Available from: <https://doi.org/10.1105/tpc.113.115485>
- Entizne, J.C., Guo, W., Calixto, C.P.G., Spensley, M., Tzioutziou, N., Zhang, R. *et al.* (2020) TranSuite: a software suite for accurate translation and characterization of transcripts. *bioRxiv*. Available from: <https://doi.org/10.1101/2020.12.15.422989>
- Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E. *et al.* (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Research*, **20**(1), 45–58. Available from: <https://doi.org/10.1101/gr.093302.109>
- Guo, W., Calixto, C.P.G., Brown, J.W.S. & Zhang, R. (2017) TSIS: an R package to infer alternative splicing isoform switches for time-series data. *Bioinformatics*, **33**(20), 3308–3310. Available from: <https://doi.org/10.1093/bioinformatics/btx411>
- Guo, W., Tzioutziou, N.A., Stephen, G., Milne, I., Calixto, C.P., Waugh, R. *et al.* (2021) 3D RNA-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists. *RNA Biology*, **18**(11), 1574–1587. Available from: <https://doi.org/10.1080/15476286.2020.1858253>

- Guo, Z., Li, B., Du, J., Shen, F., Zhao, Y., Deng, Y. *et al.* (2023) LettuceGDB: the community database for lettuce genetics and omics. *Plant Communications*, **4**(1), 100425. Available from: <https://doi.org/10.1016/j.xplc.2022.100425>
- Hayer, K.E., Pizarro, A., Lahens, N.F., Hogenesch, J.B. & Grant, G.R. (2015) Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, **31**(24), 3938–3945. Available from: <https://doi.org/10.1093/bioinformatics/btv488>
- Huang, J., Gu, L., Zhang, Y., Yan, T., Kong, G., Kong, L. *et al.* (2017) An oomycete plant pathogen reprograms host pre-mRNA splicing to subvert immunity. *Nature Communications*, **8**(1), 2051. Available from: <https://doi.org/10.1038/s41467-017-02233-5>
- Ingle, R.A., Stoker, C., Stone, W., Adams, N., Smith, R., Grant, M. *et al.* (2015) Jasmonate signalling drives time-of-day differences in susceptibility of Arabidopsis to the fungal pathogen *Botrytis cinerea*. *Plant J.*, **84**(5), 937–948. Available from: <https://doi.org/10.1111/tpj.13050>
- Jin, J., Lu, P., Xu, Y., Li, Z., Yu, S., Liu, J. *et al.* (2021) PLOCDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. *Nucleic Acids Research*, **49**(D1), D1489–D1495. Available from: <https://doi.org/10.1093/nar/gkaa910>
- Jin, Y., Ivanov, M., Dittrich, A.N., Nelson, A.D. & Marquardt, S. (2023) LncRNA FLAIL affects alternative splicing and represses flowering in Arabidopsis. *The EMBO Journal*, **42**(11), e110921. Available from: <https://doi.org/10.15252/emboj.2022110921>
- Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9), 1236–1240. Available from: <https://doi.org/10.1093/bioinformatics/btu031>
- Kainth, A.S., Haddad, G.A., Hall, J.M. & Ruthenburg, A.J. (2023) Merging short and stranded long reads improves transcript assembly. *PLoS Computational Biology*, **19**(10), e1011576. Available from: <https://doi.org/10.1371/journal.pcbi.1011576>
- Kalvari, I., Nawrocki, E.P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M. *et al.* (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, **49**(D1), D192–D200. Available from: <https://doi.org/10.1093/nar/gkaa1047>
- 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**(6), 2454–2469. Available from: <https://doi.org/10.1093/nar/gkr932>
- Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L. *et al.* (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, **45**(W1), W12–W16. Available from: <https://doi.org/10.1093/nar/gkx428>
- Kim, D.-H. & Sung, S. (2017) Vernalization-triggered intragenic chromatin loop formation by long noncoding RNAs. *Developmental Cell*, **40**(3), 302–312.e4. Available from: <https://doi.org/10.1016/j.devcel.2016.12.021>
- Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, **15**(20), 8125–8148. Available from: <https://doi.org/10.1093/nar/15.20.8125>
- Kumar, V., Sugumaran, K., Al-Roumi, A. & Shajan, A. (2022) De-novo transcriptome assembly and analysis of lettuce plants grown under red, blue or white light. *Scientific Reports*, **12**(1), 22477. Available from: <https://doi.org/10.1038/s41598-022-26344-2>
- Kuo, R.I., Cheng, Y., Zhang, R., Brown, J.W.S., Smith, J., Archibald, A.L. *et al.* (2020) Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*, **21**(1), 751. Available from: <https://doi.org/10.1186/s12864-020-07123-7>
- Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15**(2), R29. Available from: <https://doi.org/10.1186/gb-2014-15-2-r29>
- Lee, Y. & Rio, D.C. (2015) Mechanisms and regulation of alternative pre-mRNA splicing. *Annual Review of Biochemistry*, **84**(1), 291–323. Available from: <https://doi.org/10.1146/annurev-biochem-060614-034316>
- Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100. Available from: <https://doi.org/10.1093/bioinformatics/bty191>
- Li, Q. & Hunt, A.G. (1997) The polyadenylation of RNA in plants. *Plant Physiology*, **115**(2), 321–325. Available from: <https://doi.org/10.1104/pp.115.2.321>
- Ling, Y., Serrano, N., Gao, G., Atia, M., Mokhtar, M., Woo, Y.H. *et al.* (2018) Thermopriming triggers splicing memory in Arabidopsis. *Journal of Experimental Botany*, **69**(10), 2659–2675. Available from: <https://doi.org/10.1093/jxb/ery062>
- Lykke-Andersen, S. & Jensen, T.H. (2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews. Molecular Cell Biology*, **16**(11), 665–677. Available from: <https://doi.org/10.1038/nrm4063>
- 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**(6), 1184–1195. Available from: <https://doi.org/10.1101/gr.134106.111>
- Martín, G., Márquez, Y., Mantica, F., Duque, P. & Irimia, M. (2021) Alternative splicing landscapes in *Arabidopsis thaliana* across tissues and stress conditions highlight major functional differences with animals. *Genome Biology*, **22**(1), 35. Available from: <https://doi.org/10.1186/s13059-020-02258-y>
- McGuire, A.M., Pearson, M.D., Neafsey, D.E. & Galagan, J.E. (2008) Cross-kingdom patterns of alternative splicing and splice recognition. *Genome Biology*, **9**(3), R50. Available from: <https://doi.org/10.1186/gb-2008-9-3-r50>
- Minio, A., Massonnet, M., Figueroa-Balderas, R., Vondras, A.M., Blanco-Ulate, B. & Cantu, D. (2019) Iso-Seq allows genome-independent transcriptome profiling of grape berry development. *G3: Genes, Genomes, Genetics*, **9**(3), 755–767. Available from: <https://doi.org/10.1534/g3.118.201008>
- 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. *The Plant Cell*, **26**(7), 2746–2760. Available from: <https://doi.org/10.1105/tpc.114.125617>
- 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), 1–10. Available from: <https://doi.org/10.1046/j.0960-7412.2001.01188.x>
- Nawrocki, E.P. & Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**(22), 2933–2935. Available from: <https://doi.org/10.1093/bioinformatics/btt509>
- Ner-Gaon, H., Halachmi, R., Savaldi-Goldstein, S., Rubin, E., Ophir, R. & Fluhr, R. (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *The Plant Journal*, **39**(6), 877–885. Available from: <https://doi.org/10.1111/j.1365-313x.2004.02172.x>
- Ner-Gaon, H., Leviatan, N., Rubin, E. & Fluhr, R. (2007) Comparative cross-species alternative splicing in plants. *Plant Physiology*, **144**(3), 1632–1641. Available from: <https://doi.org/10.1104/pp.107.098640>
- Neve, J., Patel, R., Wang, Z., Louey, A. & Furger, A.M. (2017) Cleavage and polyadenylation: ending the message expands gene regulation. *RNA Biology*, **14**(7), 865–890. Available from: <https://doi.org/10.1080/15476286.2017.1306171>
- 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**(4), 417–419. Available from: <https://doi.org/10.1038/nmeth.4197>
- Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G. *et al.* (2023) InterPro in 2022. *Nucleic Acids Research*, **51**(D1), D418–D427. Available from: <https://doi.org/10.1093/nar/gkac993>
- Perteau, G. & Perteau, M. (2020) GFF utilities: GffRead and GffCompare. *F1000Research*, **9**, 304. Available from: <https://doi.org/10.12688/f1000research.23297.2>
- Perteau, M., Perteau, 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**(3), 290–295. Available from: <https://doi.org/10.1038/nbt.3122>
- Pink, H., Talbot, A., Carter, R., Hickman, R., Cooper, O., Law, R. *et al.* (2023) Identification of *Lactuca sativa* transcription factors impacting resistance to *Botrytis cinerea* through predictive network inference. *bioRxiv*. Available from: <https://doi.org/10.1101/2023.07.19.549542>
- Pink, H., Talbot, A., Graceson, A., Graham, J., Higgins, G., Taylor, A. *et al.* (2022) Identification of genetic loci in lettuce mediating quantitative resistance to fungal pathogens. *Theoretical and Applied Genetics*, **135**(7), 2481–2500. Available from: <https://doi.org/10.1007/s00122-022-04129-5>
- Proudfoot, N.J. (2011) Ending the message: poly(A) signals then and now. *Genes & Development*, **25**(17), 1770–1782. Available from: <https://doi.org/10.1101/gad.17268411>

- Pucker, B. & Brockington, S.F. (2018) Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics*, **19**(1), 980. Available from: <https://doi.org/10.1186/s12864-018-5360-z>
- Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842. Available from: <https://doi.org/10.1093/bioinformatics/btq033>
- 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**(1), 968. Available from: <https://doi.org/10.1186/s12864-019-6243-7>
- Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikiti, S., Song, C. *et al.* (2017) Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, **8**(1), 14953. Available from: <https://doi.org/10.1038/ncomms14953>
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), e47. Available from: <https://doi.org/10.1093/nar/gkv007>
- Sanchez, S.E., Petrillo, E., Beckwith, E.J., Zhang, X., Rugnone, M.L., Hernandez, C.E. *et al.* (2010) A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature*, **468**(7320), 112–116. Available from: <https://doi.org/10.1038/nature09470>
- Sawada, Y., Umetsu, A., Komatsu, Y., Kitamura, J., Suzuki, H., Asami, T. *et al.* (2012) An unusual spliced variant of DELLA protein, a negative regulator of gibberellin signaling, in lettuce. *Bioscience, Biotechnology, and Biochemistry*, **76**(3), 544–550. Available from: <https://doi.org/10.1271/bbb.110847>
- Shao, M. & Kingsford, C. (2017) Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*, **35**(12), 1167–1169. Available from: <https://doi.org/10.1038/nbt.4020>
- Shikata, H., Hanada, K., Ushijima, T., Nakashima, M., Suzuki, Y. & Matsushita, T. (2014) Phytochrome controls alternative splicing to mediate light responses in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(52), 18781–18786. Available from: <https://doi.org/10.1073/pnas.1407147112>
- Smoleń, S., Czernicka, M., Keška-Izworska, K., Kowalska, I., Grzebelus, D., Pitala, J. *et al.* (2023) Transcriptomic and metabolic studies on the role of inorganic and organic iodine compounds in lettuce plants. *Scientific Reports*, **13**(1), 8440. Available from: <https://doi.org/10.1038/s41598-023-34873-7>
- Soneson, C., Love, M.I. & Robinson, M.D. (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**, 1521. Available from: <https://doi.org/10.12688/f1000research.7563.2>
- Stark, R., Grzelak, M. & Hadfield, J. (2019) RNA sequencing: the teenage years. *Nature Reviews. Genetics*, **20**(11), 631–656. Available from: <https://doi.org/10.1038/s41576-019-0150-2>
- Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., Hubbard, T.J., Guigó, R. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, **10**(12), 1177–1184. Available from: <https://doi.org/10.1038/nmeth.2714>
- The Tomato Genome Consortium (TGC). (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**(7400), 635–641. Available from: <https://doi.org/10.1038/nature11119>
- Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J. *et al.* (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, **19**(1), 40. Available from: <https://doi.org/10.1186/s13059-018-1417-1>
- Vanechoutte, D., Estrada, A.R., Lin, Y.-C., Loraine, A.E. & Vandepoele, K. (2017) Genome-wide characterization of differential transcript usage in *Arabidopsis thaliana*. *The Plant Journal*, **92**(6), 1218–1231. Available from: <https://doi.org/10.1111/tpj.13746>
- Walley, P.G., Hough, G., Moore, J.D., Carder, J., Elliott, M., Mead, A. *et al.* (2017) Towards new sources of resistance to the currant-lettuce aphid (*Nasonovia ribisnigri*). *Molecular Breeding*, **37**(1), 4. Available from: <https://doi.org/10.1007/s11032-016-0606-4>
- Wang, B., Regulski, M., Tseng, E., Olson, A., Goodwin, S., McCombie, W.R. *et al.* (2018) A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. *Genome Research*, **28**(6), 921–932. Available from: <https://doi.org/10.1101/gr.227462.117>
- Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P. & Li, W. (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research*, **41**(6), e74. Available from: <https://doi.org/10.1093/nar/gkt006>
- 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**(1), 163–178. Available from: <https://doi.org/10.1111/nph.14762>
- Wheeler, T.J. & Eddy, S.R. (2013) Nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, **29**(19), 2487–2489. Available from: <https://doi.org/10.1093/bioinformatics/btt403>
- Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R., Cooke, E. *et al.* (2012) Arabidopsis defense against *Botrytis cinerea*: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *The Plant Cell*, **24**(9), 3530–3557. Available from: <https://doi.org/10.1105/tpc.112.102046>
- Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T. *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, **45**, gkw1306. Available from: <https://doi.org/10.1093/nar/gkw1306>
- Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M. *et al.* (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics*, **8**(1), 67. Available from: <https://doi.org/10.1186/1471-2164-8-67>
- Zhang, G., Sun, M., Wang, J., Lei, M., Li, C., Zhao, D. *et al.* (2019) PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in Rice. *The Plant Journal*, **97**(2), 296–305. Available from: <https://doi.org/10.1111/tpj.14120>
- Zhang, L., Qian, J., Han, Y., Jia, Y., Kuang, H. & Chen, J. (2022) Alternative splicing triggered by the insertion of a CACTA transposon attenuates LsGLK and leads to the development of pale-green leaves in lettuce. *The Plant Journal*, **109**(1), 182–195. Available from: <https://doi.org/10.1111/tpj.15563>
- Zhang, L., Su, W., Tao, R., Zhang, W., Chen, J., Wu, P. *et al.* (2017) RNA sequencing provides insights into the evolution of lettuce and the regulation of flavonoid biosynthesis. *Nature Communications*, **8**(1), 2264. Available from: <https://doi.org/10.1038/s41467-017-02445-9>
- Zhang, R., Calixto, C.-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**(9), 5061–5073. Available from: <https://doi.org/10.1093/nar/gkx267>
- Zhang, R., Calixto, C.P.G., Tzioutziou, N.A., James, A.B., Simpson, C.G., Guo, W. *et al.* (2015) AtrTD – a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in *Arabidopsis thaliana*. *The New Phytologist*, **208**(1), 96–101. Available from: <https://doi.org/10.1111/nph.13545>
- 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**(1), 149. Available from: <https://doi.org/10.1186/s13059-022-02711-0>
- Zhang, X.-N. & Mount, S.M. (2009) Two alternatively spliced isoforms of the Arabidopsis SR45 protein have distinct roles during normal plant development. *Plant Physiology*, **150**(3), 1450–1458. Available from: <https://doi.org/10.1104/pp.109.138180>
- Zhang, Z., Liu, Y., Ding, P., Li, Y., Kong, Q. & Zhang, Y. (2014) Splicing of receptor-like kinase-encoding SNC4 and CERK1 is regulated by two conserved splicing factors that are required for plant immunity. *Molecular Plant*, **7**(12), 1766–1775. Available from: <https://doi.org/10.1093/mp/ssu103>