Received: 24 November 2025 | Accepted: 23 March 2026

DOI: 10.1002/tpg2.70240

The Plant Genome

RESOURCES

Transcriptome assemblies for two drug-type cannabis chemotypes by long-read RNA sequencing

Oliver Berkowitz[1,2,3,4] | **Ricarda Jost**[1,2,3,4] | **Sophia Ng**[1,2] | **Antony Bacic**[1,2,3] | **James Whelan**[1,5] | **Mathew G. Lewsey**[1,2,3,4,6]

[1]Australian Research Council Research Hub for Medicinal Agriculture, Department of Ecological Plant and Animal Sciences, School of Agriculture, Biomedicine and Environment, La Trobe University, Bundoora, Victoria, Australia

[2]La Trobe Institute for Sustainable Agriculture and Food, Department of Ecological Plant and Animal Sciences, School of Agriculture, Biomedicine and Environment, La Trobe University, Bundoora, Victoria, Australia

[3]Australian Research Council Research Hub for Protected Cropping, Department of Ecological Plant and Animal Sciences, School of Agriculture, Biomedicine and Environment, La Trobe University, Bundoora, Victoria, Australia

[4]Australian Plant Phenomics Network, La Trobe University, Bundoora, Victoria, Australia

[5]College of Life Science, Zhejiang University, Hangzhou, P. R. China

[6]Australian Research Council Centre of Excellence in Plants for Space, Department of Ecological Plant and Animal Sciences, School of Agriculture, Biomedicine and Environment, La Trobe University, Bundoora, Victoria, Australia

Correspondence
Oliver Berkowitz, Australian Research Council Research Hub for Protected Cropping, Department of Ecological Plant and Animal Sciences, School of Agriculture, Biomedicine and Environment, La Trobe University, Bundoora, VIC, Australia.
Email: o.berkowitz@latrobe.edu.au

Assigned to Associate Editor Manuel Spannagl.

Funding information
Australian Research Council (ARC) Centre of Excellence in Plants for Space, Grant/Award Number: CE230100015; Australian Research Council (ARC) Research Hub for Medicinal Agriculture, Grant/Award Number: IH180100006; Australian Research Council (ARC) Research Hub for Protected Cropping, Grant/Award Number: IH240100024

Abstract

Cannabis sativa has undergone over 10,000 years of domestication, resulting in extensive genetic and phenotypic diversity among cultivated chemotypes. Increased medical and recreational use of specialized metabolites accumulating in cannabis glandular trichomes—primarily the cannabinoids Δ9-tetrahydrocannabinol (THC) and cannabidiol (CBD)—has amplified research interest and a need to develop supporting genomic tools. Here, we present PacBio Iso-Seq-based transcriptome assemblies for two contrasting cannabis drug-type chemotypes (THC- and CBD-dominant) and their characterization. These assemblies encompass approximately 60% of the annotated loci in the cs10 reference genome, consistent with the commonly expressed fraction of the genome, and identify 1145 novel transcribed loci not present in the cs10 reference. Each assembly defines >50,000 transcripts and 15,000 alternative splicing events. Their accuracy is exemplified by confirming the conservation of alternative splicing events for Rubisco activase and a serine/argine-rich protein (SR45). We further highlight their utility by characterizing a novel cis-regulatory long non-coding RNA associated with the transcription factor NITRATE REGULATORY GENE 2. In addition, alternative splicing events for *SPX DOMAIN*

Abbreviations: AS, alternative splicing; CBD, cannabidiol; FI, female inflorescence; lncRNA, long non-coding RNA; LR, lateral roots; THC, Δ9-tetrahydrocannabinol; TPM, transcripts per million; YL, young leaf; YS, young stems.

4, a key regulator of phosphate homeostasis, identified expression of transcripts encoding proteins with altered domain structure. Quantification of transcript abundances of these genes across different organs and varying phosphate supplies revealed isoform-specific expression patterns that differ between chemotypes, suggesting novel regulatory mechanisms for nitrogen and phosphate acquisition not previously described in either model or crop plant species. Our transcriptome assemblies provide a rich resource for the functional characterization of transcript and protein diversity in cannabis.

Plain Language Summary

We provide transcriptome assemblies of two *Cannabis sativa* drug-type chemotypes generated by PacBio Iso-Seq. These provide a resource to interrogate the diversity of transcript isoforms and their contribution to the expression of gene function and encoded proteins. We demonstrate the utility of the transcript assemblies by providing evidence for alternative splicing events with importance for nutrient acquisition not shown previously in any other plant species.

1 | INTRODUCTION

Cannabis (*Cannabis sativa* L.) is one of the oldest domesticated plants. Human selection for the production of fiber, seed, and drugs has led to a diverse genetic and phenotypic variation within the species (Clarke & Merlin, 2016). Decarboxylated forms of specialized metabolites accumulating within the trichomes of female flowers, that is, the cannabinoids Δ9-tetrahydrocannabinol (THC) and cannabidiol (CBD), are used as pharmaceutical and recreational drugs (Lu & Mackie, 2016). This makes cannabis a high-value crop with increased production across the globe and a heightened interest in research to improve plant performance, yields, and drug mode of action on the human endocannabinoid pathway (UNODC, 2023). Specific selection for the genes of the cannabinoid pathway, especially tetrahydrocannabinolic acid synthase (THCAS) and cannabidiolic acid synthase (CBDAS), combined with their targeted (re)introgression, has generated drug-type cannabis varieties termed chemotypes with high specific accumulation of either THC, CBD, or both (Clarke & Merlin, 2016; Small, 2015).

Over recent years genomic resources to understand the interaction of valuable traits and their genetic determinants in cannabis have been generated by the assembly and annotation of genomes (Gao et al., 2020; Grassa et al., 2021; Laverty et al., 2019; Ryu et al., 2024). In addition, large-scale whole genome (re)sequencing of diverse accessions has provided insight into the genetic diversity across the pangenome and the domestication history of cannabis (Lynch et al., 2025; Ren et al., 2021). These studies also provided resources to interrogate the expression of genes by generating information on the cannabis transcriptome largely based on short-read sequencing (RNA-seq) and subsequent read mapping to the reference genome. This has the advantage of providing the high read coverage essential for quantitative studies, tracing transcriptionally active loci and their functional annotation. However, transcript assembly using short-read RNA-seq-based methods by bioinformatic approaches is challenging. Determining transcript diversity and complexity is also limited by frequent artifacts (Steijger et al., 2013).

Recent advances in long-read RNA sequencing methodologies (PacBio Iso-Seq and Oxford Nanopore Technologies RNA sequencing), which directly sequence single molecules with high accuracy, have become the gold standard to generate plant transcriptomes with highly resolved information on sequence diversity (Amarasinghe et al., 2020). This complexity is largely due to alternative splicing (AS), where the posttranscriptional processing of pre-mRNA leads to mature transcript isoforms with varying sequence composition (Alhabsi et al., 2025; Tognacca et al., 2022). In contrast to other phyla, the most common AS event in plants is the retention of introns (~40%–50%), which is likely due to either slow maturation or a form of storage as part of a regulatory mechanism (Petrillo, 2023). The usage of alternative splice sites at the 5′ or 3′ ends of exons (~30%–40%) and exon skipping (~10%–20%) are other frequently observed AS events. A potential consequence of AS is the generation of different coding sequences for proteins and thus an increased potential to diversify protein function. While this alters domain composition or the occurrence of catalytic residues within proteins, important mechanisms for the control of metabolic pathways or their regulation, the impact of

AS on the protein sequence level is often not investigated (Alhabsi et al., 2025; Tognacca et al., 2022). The number of annotated transcripts for current cannabis genome assemblies, with 1.3 and 1.4 transcripts per gene for the cs10 and Pink Pepper reference assemblies (Grassa et al., 2021; Ryu et al., 2024), is likely an underestimation of the isoform diversity. This is expected given that annotation pipelines are largely focused on obtaining the main isoform(s) supported by RNA-seq and covering gene loci. Two studies based on transcript assembly from short-read sequencing predicted about 10,000 and 28,000 AS events, respectively (Severson & Adams, 2023; Wu et al., 2021), while another study using direct long-read RNA sequencing identified only 1170 AS events (Jiang et al., 2022). Thus, current knowledge on the diversity and complexity of the cannabis transcriptome is limited.

We have performed PacBio Iso-Seq to generate a high-quality compendium of transcript isoforms with >15,000 AS events for each of the two cannabis chemotypes (THC-dominant, CBD-dominant) analyzed. We demonstrate the utility of these chemotype-specific assemblies by characterization of transcript isoforms, novel regulatory non-coding RNAs, and quantification of their abundances using RNA-seq data for selected genes. We also provide examples for changes in domain composition of proteins caused by AS events. Thus, our transcriptome assemblies are a rich resource for the analysis and quantification of transcript diversity in cannabis.

Core Ideas

- Generation of transcriptomes for two contrasting cannabis chemotypes by PacBio Iso-Seq serve as crucial resource for further functional studies.
- More than 50,000 transcripts covered 60% of gene loci for each chemotype.
- >15,000 alternative splicing events for each chemotype were identified.
- There is evidence for functional diversity of proteins encoded by specific transcript isoforms.

2 | MATERIALS AND METHODS

2.1 | Plant material, RNA isolation, and PacBio sequencing

Plant material, growth conditions, and sample collection were described in detail previously (Jost et al., 2025). Briefly, two cannabis chemotypes with contrasting cannabinoid profiles (CBD-dominant chemotype Cannatonic, THC-dominant chemotype Northern Lights) were grown in a glasshouse under standard growth conditions for about 2 months before tissue harvest at the mature flowering stages. In addition, a leaf sample was also taken from a 4-month-old mother plant of each chemotype. About 100 mg of each tissue (lateral root [LR], old stem [OS], young stem [YS], senescing leaf [SL], old leaf [OL], young leaf [YL], terminal flower [TF], and leaf, mother plant [LMP]) was collected, snap frozen in liquid nitrogen, and stored at −80°C until further processing (Jost et al., 2025). Total RNA isolation and genomic DNA removal were performed with the Plant Total RNA including DNase kit according to the manufacturer's protocol (Sigma–Aldrich). PacBio Iso-Seq library generation, quality control, and sequencing were carried out by the Australian Genomic Resource Facility (AGRF Brisbane Node) using a PacBio Sequel II SMRT Cell 8 M.

2.2 | Bioinformatic analyses

See Figure 1 for an overview of the complete bioinformatic pipeline. Initial preprocessing of HiFi reads (consensus calling, demultiplexing, refining, and clustering) to the Iso-Seq HQ and LQ transcript stage was performed by the Australian Genome Research Facility following the default IsoSeq3 pipeline (https://github.com/PacificBiosciences/IsoSeq). To further remove transcript redundancy, HQ transcripts were then collapsed using default parameters with the "pbmm2" (with –sort –preset ISOSEQ parameters) and "isoseq collapse" (with default parameters) tools to generate chemotype-specific transcripts. These were then mapped to the cannabis cs10 reference genome, downloaded from the Ensembl Plants database (https://plants.ensembl.org/Cannabiscannabis_sativa_female), with minimap2 (-ax splice:hq -t 32 -G 100000 -uf –secondary = no parameters) (Li, 2018), and consensus transcripts of both chemotypes were called and classified with gffcompare with -C -V -e 500 -d 500 parameters (Pertea & Pertea, 2020) (for classification codes see Table S1). Transcripts of the cs10 assembly and Iso-Seq transcripts were merged into a non-redundant assembly using gffcompare. Functional annotation of novel transcripts was performed via BLAST and Mercator (Schwacke et al., 2019). The longest open reading frames were predicted for each transcript using Transcoder (Haas et al., 2013) (with -G 'Universal' –retain_long_orfs_mode dynamic –single_best_only -T 500 parameters), and the corresponding proteins then aligned for every locus using ClustalOmega (Sievers & Higgins, 2018). Transcript abundances as transcripts per million (TPM) and estimated counts were quantified by pseudo-aligning RNA-seq reads of previously published datasets consisting of three biological replicates per tissue and/or treatment (PRJNA884161, [Jost

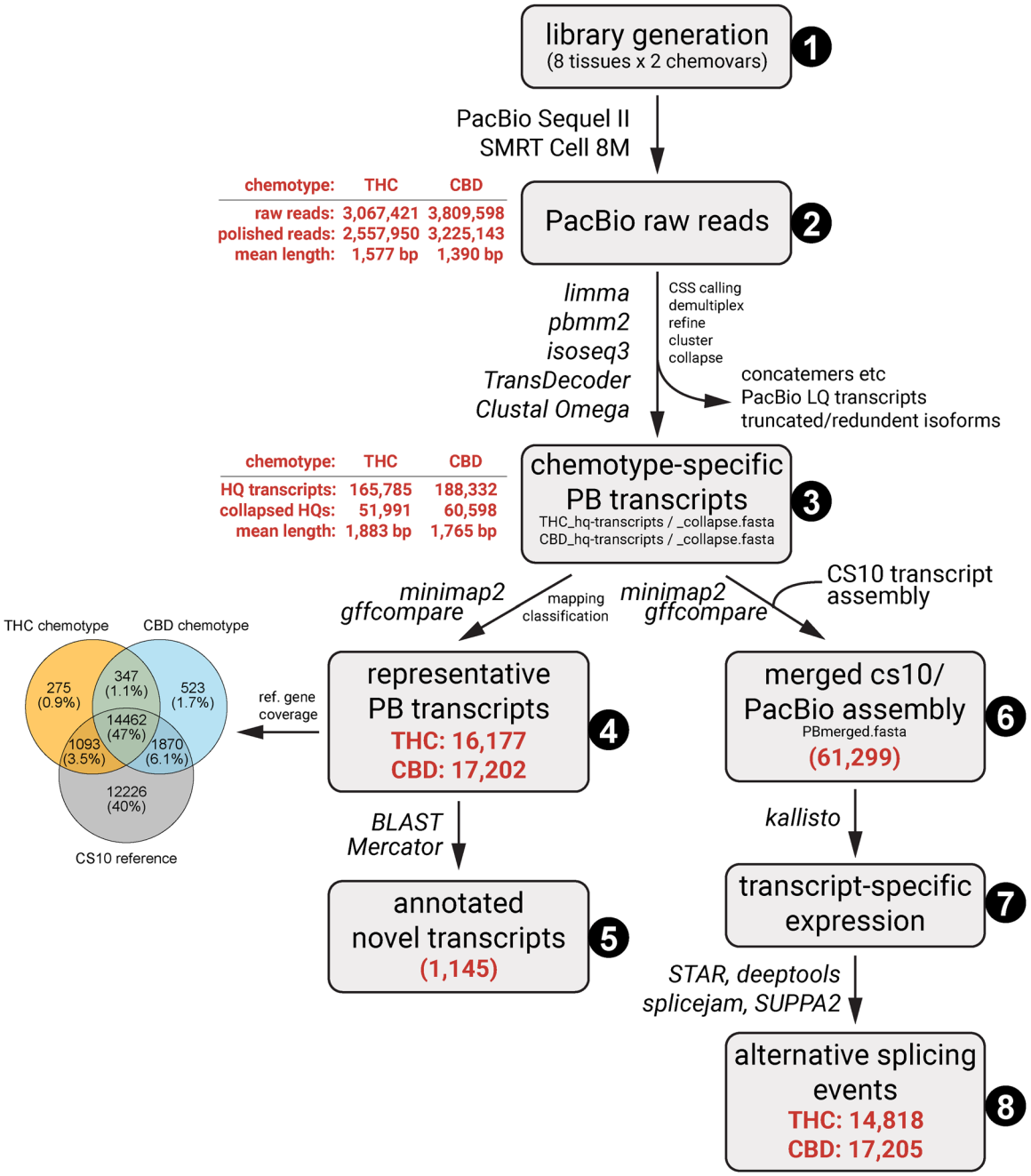

FIGURE 1 PacBio Iso-Seq analysis pipeline for eight tissues of two cannabis chemotypes. PacBio Iso-seq was separately conducted for eight tissues of two cannabis chemotypes (Δ9-tetrahydrocannabinol [THC]- or cannabidiol (CBD)-dominant) (step 1). Raw reads (step 2) were processed, refined, and clustered before merging for each chemotypes to high-quality transcripts (limma, pbmm2, and IsoSeq3) and subsequent further collapsing (IsoSeq3). For each transcript, the longest open reading frame and the corresponding encoded protein sequences were also determined. These protein sequences were aligned by Clustal Omega for every locus (step 3). Transcripts were then mapped to the cs10 reference genome, and for each gene locus, the representative transcript was also determined for the two chemotype-specific PacBio Iso-seq data sets (step 4). The overlap of these representative transcripts with annotated reference gene loci is indicated by the Venn diagram. Transcripts identified as novel, that is, not matching any annotated loci of cs10, were further characterized by BLAST and Mercator for functional annotation (step 5). In parallel, the chemotype-specific PB transcripts (step 3) and the cs10 transcriptome were merged. For this, non-redundant PacBio transcripts of the two chemotypes replaced transcripts of the cs10 assembly for all covered loci (step 6). Isoform-specific expression (step 7) and alternative splicing (step 8) were analyzed using independently generated RNA-seq data sets (NCBI SRA Project ID PRJNA884161, PRJNA884162, and PRJNA1278883). Numbers of transcripts resulting from each analysis step are highlighted in read. Bioinformatic tools used to process data are indicated in italics.

et al., 2025]; PRJNA1278883 [Wee et al., 2025]). For this, a k-mer index built from the transcript models for the merged cs10-PacBio transcript assemblies using the kallisto program with default parameters and 100 bootstraps (Bray et al., 2016). For further analysis of AS, splice junctions determined by STAR (Dobin et al., 2013) and read coverage quantified by deeptools (Ramírez et al., 2016) were used for visualization by splicejam (Farris et al., 2019). AS events were categorized using SUPPA2 (Trincado et al., 2018) (for category codes, see Table S1). Transcript assemblies in fasta format are deposited at the European Nucleotide Archive (ENA) under Project ID PRJEB90527.

3 | RESULTS AND DISCUSSION

3.1 | PacBio Iso-Seq and transcript assemblies for two contrasting drug-type cannabis chemotypes

To increase knowledge on transcript diversity in cannabis, we carried out PacBio RNA sequencing (Iso-Seq) on eight tissue samples (see Methods for details) collected from two contrasting chemotypes, that is, one THC-dominant and one CBD-dominant chemotype with high accumulation of these two cannabinoids, respectively.

PacBio Iso-Seq (Figure 1; step 1) resulted in about 3 million and 3.8 million raw reads for the THC- and CBD-dominant chemotype, respectively (Figure 1; step 2). Further read processing (CSS calling, refining, clustering, and collapsing; Figure 1; step 3) generated a final 51,991 and 60,598 high-quality transcripts for the THC- and CBD-dominant chemotypes, respectively (Table S1). The longest open reading frame for every transcript was also identified and translated to obtain the encoded protein sequences and their alignments for each locus. For the two assemblies, we also determined the representative PB transcripts for each locus covered (Figure 1; step 4). Together, the two PacBio Iso-Seq assemblies covered 60% of all genes annotated in the cs10 reference genome, consistent with the proportion of genes typically detected as expressed in RNA-seq experiments (Klepikova & Penin, 2019). There was a 47% overlap between the two chemotypes and the cs10 reference, and some chemotype-specific, partially overlapping transcripts (Figure 1, Venn diagram). In both chemotypes, we identified novel transcripts that were not part of the cs10 reference annotation, with 347 novel transcripts detected in both chemotypes, 275 transcripts only detected in the THC-dominant chemotype, and 523 transcripts detected only in the CBD-dominant chemotype (Figure 1; step 5). About one-third of these novel transcripts had a hit using BLAST searches ($E < 10^{-50}$) and a similar number were also expressed with TPM > 5 querying available RNA-seq data we generated previously (Jost et al., 2025), confirming these novel transcripts are active transcriptional units (Table S2). To allow for additional analyses, the two chemotype transcript assemblies and the cs10 transcript assembly were combined into a non-redundant assembly of all transcripts serving as a reference transcriptome (Figure 1; step 6).

The length distributions of the two assemblies were similar with an average length of 1883 bp and a longest transcript of 7982 bp for the THC-dominant chemotype and an average length of 1765 bp and a longest transcript of 9590 bp for the CBD-dominant chemotype (Figure 2A). The number of transcripts per gene locus was similar for both chemotypes, with an average of 3.2 for the THC-dominant chemotype and 3.6 for the CBD-dominant chemotype, as was their frequency, which resembled a one-sided normal distribution (Figure 2B). This represents a threefold higher number than the current reference genomes cs10 and Pink Pepper (1.3 and 1.4 transcripts per gene, respectively). In comparison to the cs10 transcripts, the chemotype-specific PacBio transcripts overlapping annotated gene loci had a higher number of exons, with averages of 4.9 for the cs10 reference, 6.6 for the THC-dominant, and 6.4 for the CBD-dominant chemotype (Figure 2C), also suggesting that our PacBio assemblies captured a higher transcript diversity. The novel transcripts had lower exon numbers, with averages of 1.4 for both chemotypes (Figure 2D). Using RNA-seq data for different tissues previously generated in our labs showed that differences in transcript abundances between the two chemotypes were only minor (Jost et al., 2025) (Figure 1; step 7). Across the analyzed tissues, the majority of transcripts were lowly expressed with TPM < 2, with another set of transcripts more highly expressed around a secondary peak between 10 and 30 TPM (Figure 2E). Novel transcripts were generally expressed at lower levels (Figure 2E).

We next classified each PacBio transcript by characterizing their exon structure according to their match with overlapping cs10 reference transcripts (Pertea & Pertea, 2020) (Table S1). For both chemotypes, about 40% of their transcripts were an exact match and in agreement with the cs10 references (Figure 3A; "exact match" class). Another 10% had the same exon structure but were longer or shorter than the reference—likely due to alternative transcriptional start sites or degradation of RNA ends (Figure 3A; "intron compatible" class). In addition, AS events in the form of exon skipping and alternative splice sites made up 25% of the difference between the representative cs10 and chemotype assemblies (Figure 3A; "multi-exon" class). Interestingly, about 5% of the transcripts were classified as overlapping with a gene but transcribed from the opposite strand, suggesting these are natural antisense transcripts transcribed from the same locus (cis-NATs; Figure 3A; "exonic overlap on opposite strand" class). The distribution of transcript abundances for each reference comparison class was similar, mostly skewed to lower expression. The exceptions were comparison classes with

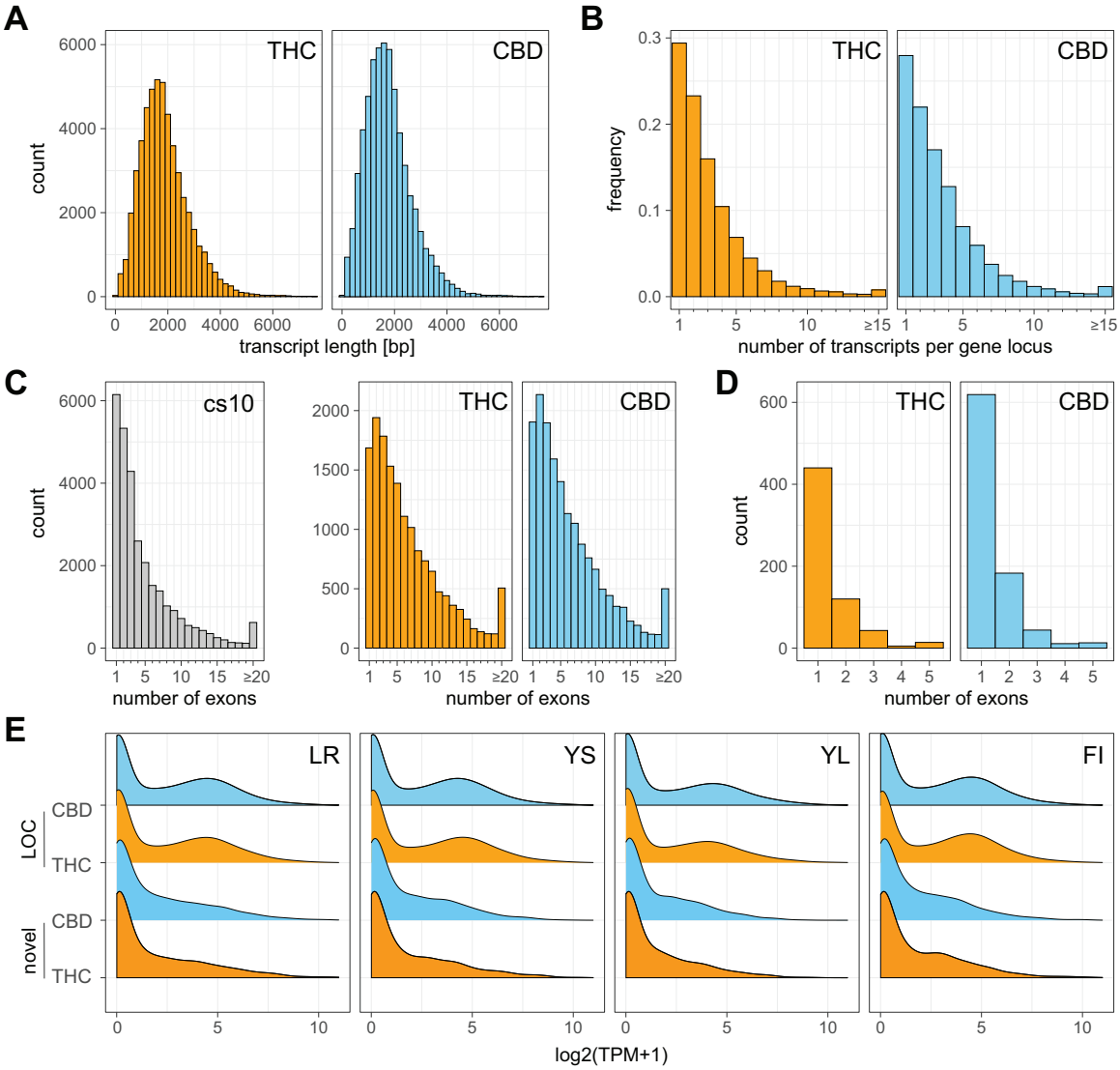

FIGURE 2 Characterization of PacBio Iso-seq transcripts identified in the two cannabis chemotypes. Key characteristics of the PacBio Iso-seq transcript assemblies for the Δ9-tetrahydrocannabinol (THC)-dominant (THC) and cannabidiol (CBD)-dominant (CBD) cannabis chemotypes. (A) Transcript length distribution for the two transcript assemblies. (B) Number of PacBio Iso-seq transcripts overlapping with gene loci annotated in the cs10 genome. For each chemotype, the frequency distribution is shown. (C) Shown is the distribution of exon numbers per transcript annotated in cs10 reference genome, in PacBio transcripts overlapping with these genes, and for novel transcripts not annotated in cs10 for the two chemotypes (THC and CBD). (E) Expression of PacBio Iso-seq transcripts not annotated in the cs10 genome (novel) compared to those overlapping with annotated gene loci (LOC). Shown are density plots representing the distribution of the log-transformed transcripts per million (TPM) values in four tissues (lateral root [LR], young stem (YS), young leaf [YL], and female inflorescence [FI]) of the two chemotypes (THC and CBD). RNA-seq data were obtained from Jost et al. (2025).

good match to the cs10 reference and a broader distribution toward higher expression (Figure 3C, subplots A, C, D). This suggests that the other comparison classes with substantial difference to the cs10 reference represent rarer transcripts, as might be expected.

AS events for the two PacBio transcript assemblies were also further characterized in detail with SUPPA2 (Trincado et al., 2018) (Figure 3B; Table S3). Consistent with read coverages (Figure 1), this identified about 15,000 and 17,000 AS events for the THC-dominant and CBD-dominant chemotype, respectively (Figure 3B). For both chemotypes, the largest fraction was intron retention with about 40% of all events. Intron retention represents splicing intermediates of not (yet) fully processed transcripts, and these are usually found at low expression levels (Zhang et al., 2022). Events with different splice site positions in 5′ or 3′ of an exon made up about 12% and 25%, respectively (Figure 3B). Variation in first and last exons was detected for 10% and 2% of transcripts, respectively, and exon skipping for about 8%. While the read depth for our PacBio Iso-Seq was moderate, the overall distribution

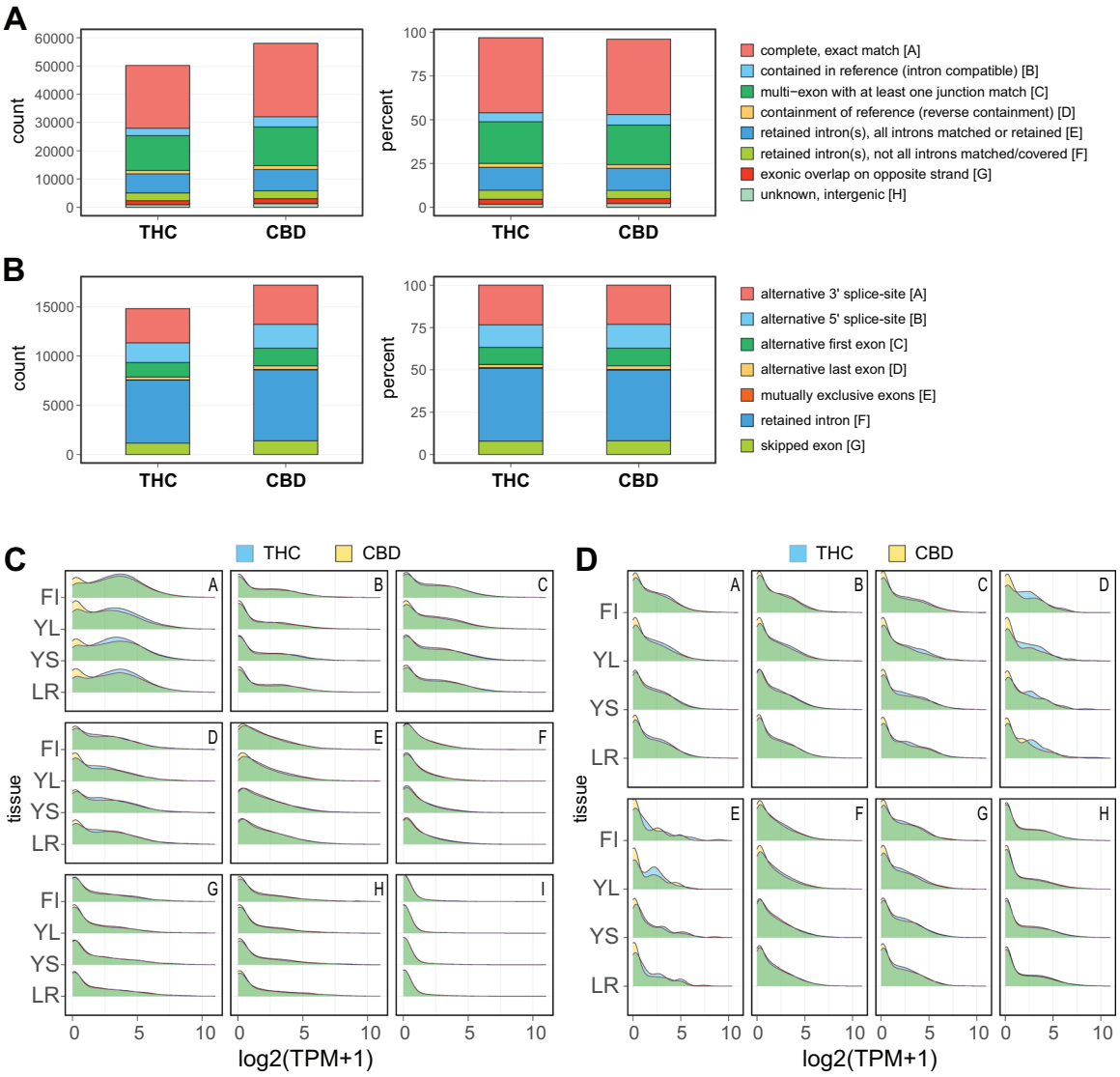

FIGURE 3 Comparison of PacBio transcripts with the cs10 reference transcriptome and alternative splicing analysis. (A) Comparison of the exon/intron structures of two chemotype-specific PacBio Iso-seq transcript sets (Δ9-tetrahydrocannabinol [THC] and cannabidiol [CBD]-dominant) with the cannabis cs10 reference transcripts. Codes for comparison classes, either in absolute numbers (left) or percent (right), were assigned using gffcompare (Pertea & Pertea, 2020); see Table S1 for details. Only the eight major classes are shown and letters in brackets in the legend identify classes for plots in (C). The 'unknown, intergenic' class indicates transcripts not annotated in the reference genome and hence represents novel transcriptional units. (B) Categorization of alternative splicing events determined across PacBio Iso-seq transcripts. All transcripts were analyzed for their exon structure against each other using the SUPPA2 pipeline (Trincado et al., 2018; see Table S3). Number of splicing events for transcripts of the two cannabis chemotypes is given in absolute numbers (left panel) and percent (right panel). Letters in brackets in the legend identify classes for plots in (D). C. Expression of transcripts in each transcript comparison class. Letters indicating classes are the same as in the legend of (A). Shown are density plots representing the distribution of the log-transformed transcripts per million (TPM) values for transcripts in each class for four tissues of the two chemotypes. (D) Expression of transcripts in alternative splicing category (letters indicating events are the same as in the legend of [B]), with addition of transcripts with no observed alternative splicing in subplot (H) Shown are density plots representing the distribution of the log-transformed TPM values for transcripts in each class for the four tissues of the two chemotypes. RNA-seq data used for (C) and (D) was obtained from a previous publication (Jost et al., 2025). Tissue lateral root (LR), young stem (YS), young leaf (YL), and female inflorescence (FI).

of these events aligns with reports in other plant species (Feng et al., 2019; Li et al., 2017; Martín et al., 2021; Zhang et al., 2022). The distribution of expression levels of transcripts in each event category was similar, and transcripts with lower expression were found more often (Figure 3D). Only the rare AS event categories had secondary peaks in their distribution and at higher TPM ("alternative last exon," "mutually exclusive exons"; Figure 3D, subplots D and E), which were based on only a few individual genes with higher expression levels (Tables S2 and S3).

In summary, the generation of our PacBio assemblies for two contrasting cannabis chemotypes, identification of non-redundant transcripts, their comparison with the reference transcriptome, and characterization of AS events serve as crucial resource for further functional studies. This includes the analysis of novel transcripts, variation of transcript-level expression, and impact of AS on gene function.

Below we provide examples of such analyses as "proof-of-concept" for the utility of these assemblies, which also provide novel insights not previously reported in other plant species.

3.2 | A novel antisense lncRNA involved in the regulation of nitrogen assimilation

In our assembly, 1382 transcripts were classified as transcribed opposite to the corresponding gene at the same locus (denoted with "x" in ST1), suggesting these are potential regulatory natural antisense transcripts (cis-NATs, [Reis & Poirier, 2021]). Among them were two long non-coding RNAs (lncRNAs; defined as > 200 nucleotides or longer and not translated into a protein [Palos et al., 2023]) matching the two loci encoding *NITRATE REGULATORY GENE 2* (*NRG2-1*, LOC115717108; *NRG2-2*, LOC115715863). In Arabidopsis, *NRG2* is a key regulator of the nitrogen starvation responses by controlling the expression of transporters involved in nitrate uptake (Xu et al., 2016). Recently we generated tissue-specific RNA-seq data on hemp-type cannabis grown under different phosphorus supplies (NCBI SRA project ID PRJNA1278883; Jost et al., 2025). Given phosphorus limitation also affects nitrogen homeostasis (Krouk & Kiba, 2020), we used this RNA-seq dataset to investigate the abundance of sense and antisense transcripts for the *NRG2* loci. For both lncRNAs, this confirmed their antisense expression, overlapping with the first two exons of the annotated *NRG2* genes (Figure 4A,B). In addition, there were tissue-specific differences in expression, with highest expression of the *NRG2-1* and *NRG2-2* genes in the YS and the female inflorescence (FI) and lowest expression in the LR and mature leaves (ML) (Figure 4C; Table S4). While the *lncNRG2-1* and *lncNRG2-2* transcript abundance was also low in the LR, for the other tissues, they contrasted with the *NRG2* genes. For both *lncNRG2*s, expression was high in ML and low in FI. For the YS, the contrast between expression of *NRG2* genes and their corresponding *lncNRG2* was more intricate, with higher transcript abundance of the lncRNAs in cannabis plants under P-limitation (0 mM P supplied) and lower *NRG2* transcript abundance. This reversed with increasing P supply, that is, higher NRG2 and lower lncNRG2 transcript abundance, respectively. It has been suggested that cis-NATs have a regulatory role on the corresponding gene through gene silencing, although experimental evidence for this is limited (Reis & Poirier, 2021). Quantification of their interactions is complicated, as RNA-seq experiments usually capture steady-state transcript abundances, as in our case, and not their dynamic relationship. Only about 4% of sense/antisense transcript show positive or negative correlation of their abundances in Arabidopsis (Deforges et al., 2019). Hence, the cannabis *NRG2* genes might provide a rare example where data suggest a regulatory function for cis-NATs, that is, the two novel lncNRG2 transcripts, under varying phosphate supplies. This indicates an additional level for the regulation of nutrient acquisition not reported before in any other plant species.

3.3 | AS leads to functional diversity in encoded proteins

AS increases the number of transcripts emanating from the same gene and is therefore a mechanism to regulate gene expression on the post-transcriptional level. In addition, encoded proteins may have altered amino acid sequence and domain structure. This can change their activity and/or post-translational regulation; that is, AS also represents an important mechanism to diversify protein function (Kashkan et al., 2022).

To investigate if such events are represented in our assemblies, we interrogated the PacBio transcripts first by focusing on genes with well-established AS. One of these is the *RIBULOSE BISPHOSPHATE CARBOXYLASE/OXYGENASE ACTIVASE* (*RCA*) gene, the first plant gene identified in Arabidopsis and spinach showing AS of its transcripts (Werneke et al., 1989). Further analyses revealed that the two transcript isoforms, the longer Rca-α isoform and the shorter Rca-β isoform, encode proteins with varying C-terminal ends. Rca-α contains two conserved cysteine residues in the C-terminal extension (CTE), which confer redox-sensitive regulation of its enzymatic activity (Zhang & Portis, 1999). RCA is expressed extremely highly in photosynthetic tissues of plants, and, correspondingly, in Arabidopsis 27 RCA transcript isoforms are detectable (Zhang et al., 2022). Our PacBio Iso-Seq identified an even higher number with 45 transcript variants for cannabis RCA (LOC115708105). Of these, the majority had only minor variations at the 5′ and/or 3′ terminal exons, while eight transcript isoforms were showing considerable variation in their exon structure and encoded proteins missing substantial parts of functional domains, for example, ATPase or RBCS-interacting domain (Figure 5). Three isoforms showed sufficient support by RNA-seq data generated previously for the same chemotypes and tissues (Figure 5; Table S5) (Jost et al., 2025), suggesting that the other transcripts are rare. Two of the major three transcripts, RCA.2 and RCA.26, showed only minor differences at their 5′ and 3′ ends and were encoding proteins including a CTE, that is, these

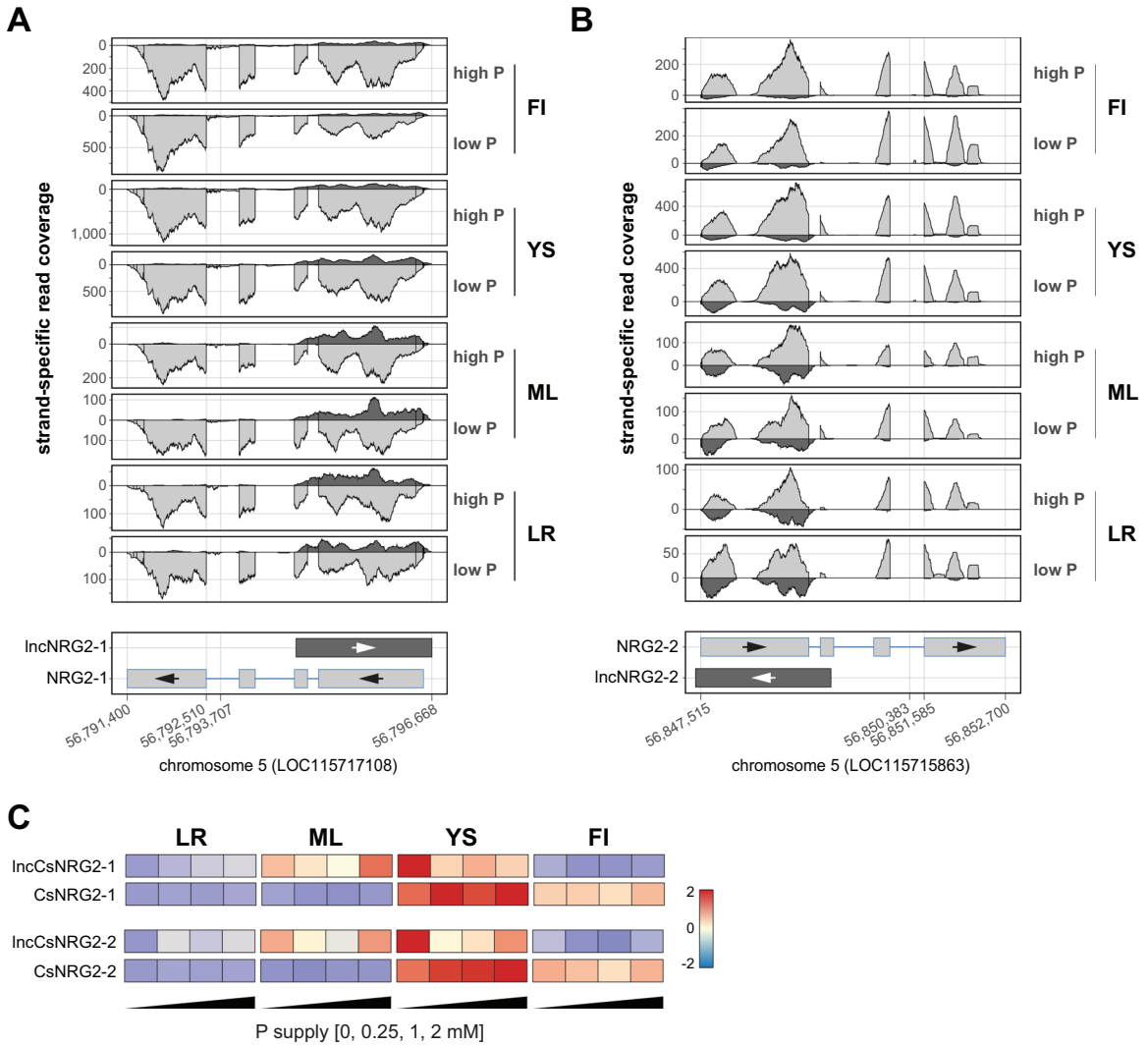

FIGURE 4 Identification and expression of lncRNAs transcribed in antisense to the two cannabis *NITRATE REGULATORY GENE 2*-encoding loci. For the two orthologs of the *NITRATE REGULATORY GENE 2* (*NRG2*) in cannabis (*CsNRG2-1*, LOC115717108; *CsNRG2-2*, LOC115715863), PacBio Iso-seq identified long non-coding RNAs (lncRNAs) transcribed in antisense. The strand-specific RNA-seq read coverage for the *CsNRG2-1* locus (A) and for the *CsNRG2-2* locus (B) in tissues (lateral roots [LR], mature leaves [ML], young stems [YS], and female inflorescences [FI]) of hemp-type plants grown with two phosphorus supplies (low P: 0.025 mM, high P: 1 mM) is shown. The bottom panels indicate the exon-intron structure of transcripts for both loci. Numbers indicate the position in base pairs on chromosome 5. Light and dark gray shading highlight strand-specific coverage for *NRG2* genes and their lncRNAs. (C) Quantification of transcript abundances for both *NRG2* transcripts and their corresponding lncRNAs across tissues of hemp-type plants with four different phosphorus supplies. Shown are heatmaps for TPM (transcripts per million) values quantified from RNA-seq after z-scoring. RNA-seq data from NCBI SRA Project ID PRJNA1278883).

isoforms are homologous to the Rca-α isoform of other plant species (Figure 5). By contrast, RCA.1 was homologous to the shorter Rca-β isoform lacking the CTE extension. As expected, expression of these isoforms was lower in the root (R) than in the photosynthetically active tissues (FI, YL, and YS) (Figure 5). Interestingly, for the latter, the expression of the Rca-β isoform (RCA.1) was higher in the THC-dominant than in the CBD-dominant chemotype, while the Rca-α isoforms (RCA.2 and RCA.26) were expressed at similar levels in both chemotypes. The ratio of Rca-α to Rca-β controls the redox sensitivity of Rubisco activation under varying irradiance (Carmo-Silva & Salvucci, 2013). Thus, the expression of both isoforms observed suggests a difference in control of photosynthetic performance for the two chemotypes via AS under the growth conditions at time of harvest.

The serine/arginine-rich (SR) proteins are a family of key regulators of AS, and their transcripts themselves are widely alternatively spliced. For example, in Arabidopsis 95 transcripts have been detected for the 15 SR genes in the genome. Many transcripts contain a premature stop codon and thus are not encoding functional proteins. These are processed through nonsense-mediated decay, which has been postulated as

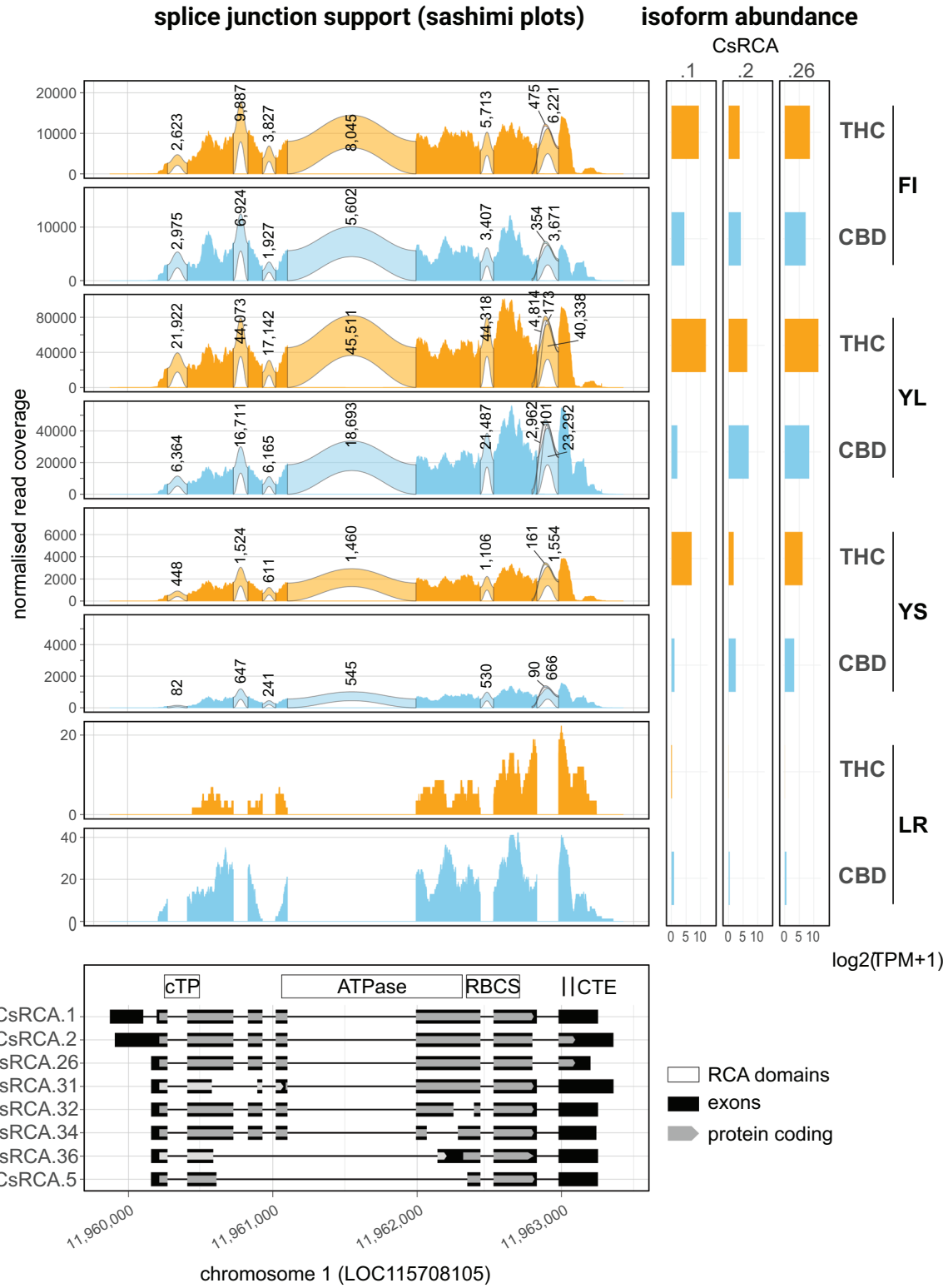

FIGURE 5 Quantitative analysis of alternative splicing for transcripts encoding cannabis Rubisco activase. Transcript isoforms for the cannabis Rubisco activase gene (RCA, LOC115708105) identified by PacBio Iso-seq was confirmed in the two cannabis chemotypes (Δ9-tetrahydrocannabinol [THC] and cannabidiol [CBD]) using RNA-seq data of four tissues (lateral root [LR], young stem [YS], young leaf [YL], and female inflorescence [FI]) (Jost et al., 2025). The sashimi plot (top left panel) shows read coverage across exons after normalization for total read number of individual samples and numbers indicate support of exon-exon junctions by intron-spanning reads. The panel below indicates the exon structure for representative isoforms (black boxes), with the sequences encoding RCA protein isoforms highlighted in gray and positions of corresponding functional domains as boxes (chloroplastic transit peptide, cTP; ATPase; RBCS-binding, RBCS; C-terminal extension, CTE). Numbers indicate the position in base pairs on chromosome 1. The panel on the right shows the expression level of each of these isoforms across tissues and chemotypes quantified from the same RNA-seq data set.

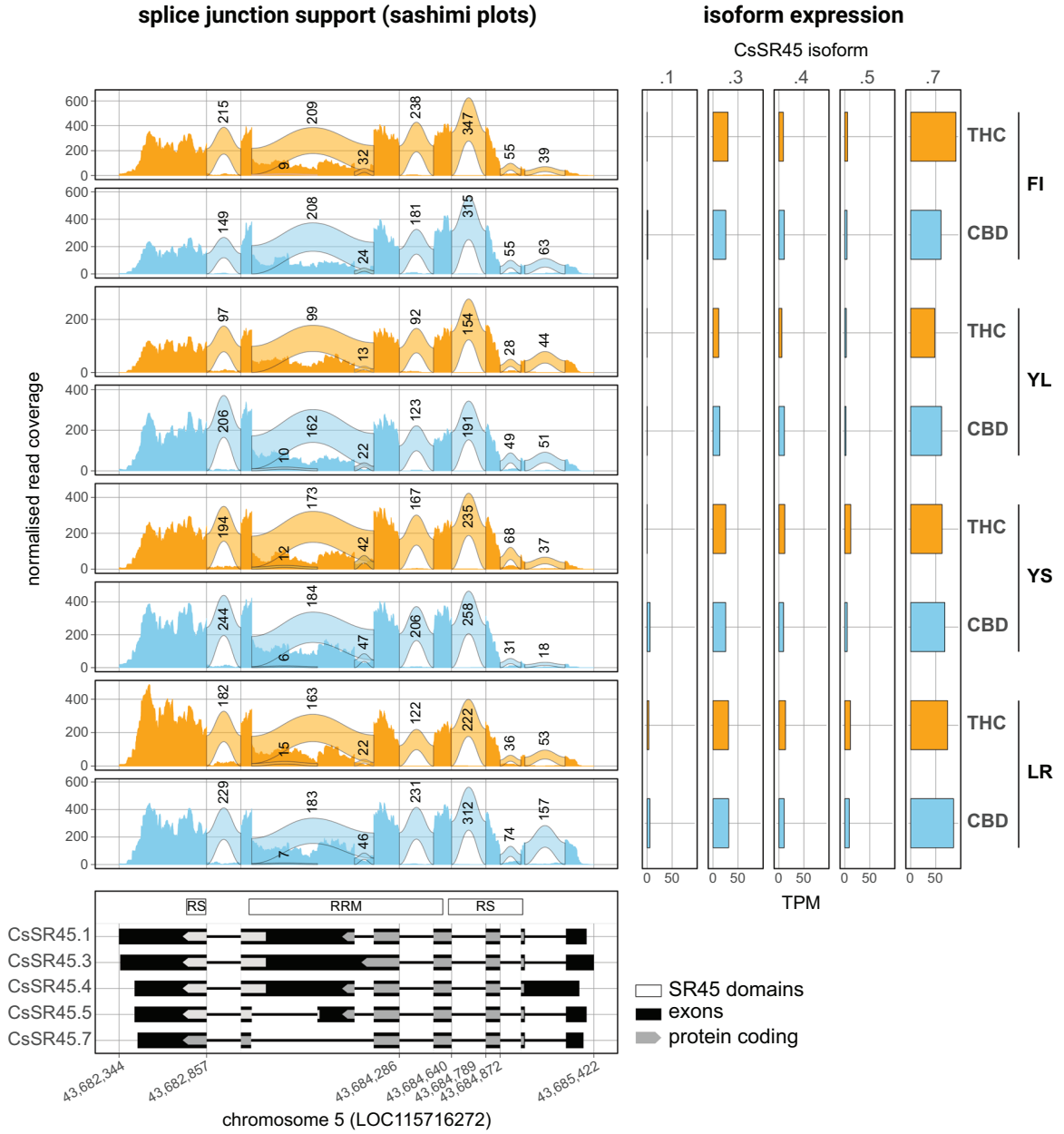

FIGURE 6 Quantitative analysis of alternative splicing for transcripts encoding a cannabis homolog of the *SERINE/ARGININE-RICH 45* gene. Transcript isoforms identified by PacBio Iso-seq for a cannabis homolog of the *SERINE/ARGININE-RICH 45* gene (*SR45*, LOC115716272) resulting from alternative splicing were confirmed in two cannabis chemotypes (Δ9-tetrahydrocannabinol [THC] and cannabidiol [CBD]) using RNA-seq data of four tissues (lateral root [LR], young stem [YS], young leaf [YL], and female inflorescence, [FI] [Jost et al., 2025]). The sashimi plot (top left panel) shows read coverage across exons after normalization for total read number of individual samples, and numbers indicate support of exon-exon junctions by intron-spanning reads. The panel below indicates the exon structure for representative isoforms detected by PacBio Iso-seq (black boxes), with the sequences encoding SR45 protein isoforms highlighted in gray and positions of corresponding functional domains (RS: arginine/serine-rich domain, RRM: RNA recognition motif) as boxes. Numbers indicate the position in base pairs on chromosome 5. The panel on the right shows the expression level of each of these isoforms across tissues and chemotypes as quantified from the same RNA-seq data set.

regulatory mechanism to fine-tune SR protein abundance and subsequently the processing of their target transcripts (Palusa et al., 2007). We screened our PacBio transcripts encoding SR proteins and found LOC115716272, with sequence similarity to Arabidopsis SR45, as an example for a cannabis SR gene showing extensive AS of its transcripts (Figure 6). Out of the seven PacBio transcripts in our assembly, expression and exon structure were sufficiently supported for five transcripts by RNAseq data. While none of these encoded for protein with the same length as Arabidopsis SR45, the cannabis SR45.7 transcript encodes a protein that contains the two characteristic, terminal arginine/serine-rich (RS) domains and an RNA

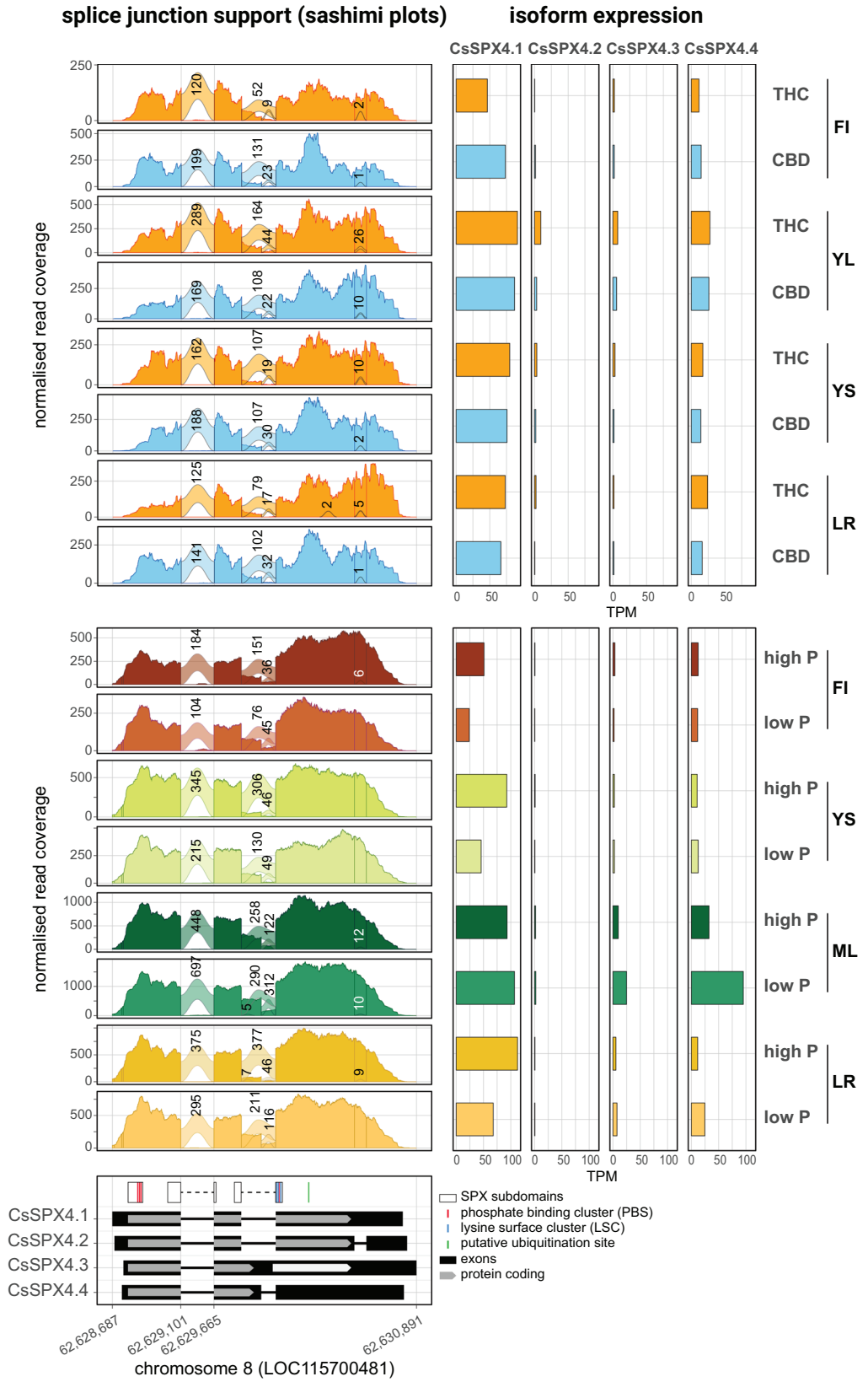

FIGURE 7 Quantitative analysis of alternative splicing for transcripts encoding the cannabis *SPX4* gene. Transcript isoforms identified by PacBio Iso-seq for the cannabis *SPX DOMAIN GENE 4* (*SPX4*, LOC115700481) resulting from alternative splicing were confirmed in the two cannabis chemotypes (Δ9-tetrahydrocannabinol [THC] and cannabidiol (CBD) dominant; RNA-seq data obtained from [Jost et al., 2025]) and hemp-type plants grown with differing phosphorus supplies (low P: 0.025 mM P; high P: 1 mM P; RNA-seq data from NCBI SRA Project ID

(Continues)

FIGURE 7 (Continued)
PRJNA1278883) across four tissues (lateral root [LR], young stem [YS], young leaf [YL], and female inflorescence [FI]). The sashimi plots (left panels) show read coverage across exons after normalization for total read number of individual samples, and numbers indicate support of exon-exon junctions by intron-spanning reads. The bar charts on the right show the expression levels of each these isoforms across tissues of the two chemotypes and P treatments of hemp-type quantified from the same RNA-seq data sets. The bottom panel indicates the exon structure for representative isoforms detected by PacBio Iso-seq (black boxes), with the sequences encoding SPX4 protein isoforms highlighted in gray and the positions of corresponding functional SPX subdomains are represented by boxes. Position of amino acids involved in sensing of the plant phosphorus status by binding phosphate or inositol polyphosphates are highlighted by red and blue lines, respectively. The position of a conserved lysine identified as ubiquitination site for rice SPX4 is indicated in green. Numbers indicate the position in base pairs on chromosome 8.

recognition motif (RRM), thus resembling a functional SR45 homolog. The RS domain facilitates interaction with splicing cofactors, and the RRM provides the interaction specificity for target transcripts to be processed by the splicing complex (Reddy, 2004). Domain composition and position in SR proteins are highly variable and suggests functional flexibility. Higher abundance of alternative spliced transcripts that contain a premature stop codon is common in plants, in contrast to animals and fungi (Barta et al., 2008). The SR45.7 transcript also showed the highest abundance of all isoforms, while there were only minor differences in expression of isoforms between the two chemotypes across analyzed tissues (Figure 6; Table S5).

The examples above confirmed conservation of AS events previously identified in other plant species, also for cannabis genes, and showed differential expression of transcript isoforms. In addition, we provided examples for AS events leading to diversification in encoded proteins.

3.4 | AS involving a key regulator of the phosphate starvation responses

We were also able to identify an AS event not described or characterized in any other plant species and with implications for the regulation of phosphate homeostasis. In plants, the SPX DOMAIN (SPX) proteins are key regulators of the phosphate starvation response (PSR), an acclimation mechanism that increases the plants' ability to cope with limited availability of phosphate in the soil. This includes the upregulation of phosphate transporter genes, changes in root architecture, and adjustments in primary metabolism (Paz-Ares et al., 2022). Binding of inositol pyrophosphates, which signal the cellular phosphorus status, to SPX modulates its interaction with the transcription factor PHOSPHATE STARVATION RESPONSE 1 (PHR1), the major regulator of the transcriptional responses. PHR1 directly binds to the promoters of downstream genes, leading to their induction (for review see [Collins et al., 2024]). In addition, SPX proteins also coordinate the integration of phosphorus with nitrogen metabolism (Hu et al., 2019; Ueda et al., 2020).

In our PacBio assembly we observed several transcripts for the cannabis homolog of SPX4 (LOC115700481), which showed distinct AS and covered the whole locus (Figure 7). To the best of our knowledge, AS of SPX transcripts has not been reported or analyzed for either their putative function or expression in any other plant species. To cross-reference, we also found similar AS events for the four SPX homologs, which are contained in an Arabidopsis transcript database (AtRTD; [Zhang et al., 2022]) (Figure S1). The CsSPX4.1 and CsSPX4.2 transcripts encode a canonical SPX family protein, with three SPX domains, a phosphate binding cluster (PBC), and a lysine surface cluster (LSC) (Figure 7) (Wild et al., 2016). Intron retention in the CsSPX4.3 transcript leads to two open reading frames, with the 5′ ORF encoding the PBC, LCS, and two-and-a-half SPX domains, while the 3′ ORF encodes another half SPX domain and the remainder of the full-length protein. The SPX4.4 transcript only encodes the 5′ ORF, with the 3′ ORF disrupted by a retained intron. The encoded protein is thus lacking the ubiquitination site and part of the sensory inositol pyrophosphate binding site (Figure 7; Figure S2). All PacBio transcripts were independently supported by our two RNA-seq data sets, with CsSPX4.1 showing highest expression across tissues, followed by CsSPX4.4, in drug-type chemotypes. The other two isoforms were expressed only at very low levels (Figure 7; Table S6). There were only minor differences between the two chemotypes for the analyzed tissues. Interestingly, for RNA-seq data of hemp-type cannabis plants receiving different P supplies, the CsSPX4.4 transcript abundance increased under P limitation in ML by about threefold, while for CsSPX4.1 the transcript level declined in FIs, YS, and LR (Figure 7). Given the different composition of functional domains in proteins encoded by the two transcript isoforms, changes in relative isoform abundances could be part of a regulatory mechanism to control the SPX-PHR1 interaction-dependent regulation of the downstream PSR (Collins et al., 2024).

4 | CONCLUSIONS

Our PacBio transcript assemblies for the two main cannabis chemotypes—THC- and CBD-dominant—expand the number of annotated and classified transcript isoforms by about threefold compared to the current reference annotations (cs10, Pink Pepper) and AS events by about 10-fold to a previous

study also using long-read sequencing methodology (Jiang et al., 2022). The examples of transcript characterization provided here demonstrate the value of our assemblies to analyze the diversity of isoforms, their expression and the potential impact of AS on protein function. These insights lay a foundation for future experimental investigations into the cannabis transcriptome and proteome landscapes.

AUTHOR CONTRIBUTIONS

Oliver Berkowitz: Conceptualization; data curation; formal analysis; investigation; methodology; visualization; writing—original draft; writing—review and editing. **Ricarda Jost**: Conceptualization; investigation; methodology; resources; writing—original draft. **Sophia Ng**: Investigation; methodology. **Antony Bacic**: Conceptualization; funding acquisition; writing—review and editing. **James Whelan**: Conceptualization; funding acquisition; writing—review and editing. **Mathew G. Lewsey**: Conceptualization; funding acquisition; writing—review and editing.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council (ARC) Research Hub for Medicinal Agriculture (IH180100006; Mathew G. Lewsey, James Whelan, and Antony Bacic), the Australian Research Council (ARC) Research Hub for Protected Cropping (IH240100024; Mathew G. Lewsey, Antony Bacic), and the Australian Research Council Centre of Excellence in Plants for Space (CE230100015; Mathew G. Lewsey). The authors would like to thank Benjamin Wee Y. for providing RNA-seq data of hemp.

Open access publishing facilitated by La Trobe University, as part of the Wiley - La Trobe University agreement via the Council of Australasian University Librarians

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Sequences of PacBio Iso-Seq assemblies were submitted to the European Nucleotide Archive under Project ID PRJEB90527.

ORCID

Oliver Berkowitz https://orcid.org/0000-0002-7671-6983
Ricarda Jost https://orcid.org/0000-0002-3819-6358
Sophia Ng https://orcid.org/0000-0002-1677-863X
Antony Bacic https://orcid.org/0000-0001-7483-8605
James Whelan https://orcid.org/0000-0001-5754-025X
Mathew G. Lewsey https://orcid.org/0000-0002-2631-4337

REFERENCES

Alhabsi, A., Ling, Y., Crespi, M., Reddy, A. S. N., & Mahfouz, M. (2025). Alternative splicing dynamics in plant adaptive responses to stress. *Annual Review of Plant Biology*, *76*, 687–717. https://doi.org/10.1146/annurev-arplant-083123-090055

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*, 30. https://doi.org/10.1186/s13059-020-1935-5

Barta, A., Kalyna, M., & Lorković, Z. J. (2008). Plant SR proteins and their functions. *Current Topics in Microbiology and Immunology*, *326*, 83–102.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525–527. https://doi.org/10.1038/nbt.3519

Carmo-Silva, A. E., & Salvucci, M. E. (2013). The regulatory properties of Rubisco activase differ among species and affect photosynthetic induction during light transitions. *Plant Physiology*, *161*, 1645–1655. https://doi.org/10.1104/pp.112.213348

Clarke, R., & Merlin, M. (2016). *Cannabis: Evolution and ethnobotany*. University of California Press.

Clarke, R. C., & Merlin, M. D. (2016). *Cannabis* domestication, breeding history, present-day genetic diversity, and future prospects. *Critical Reviews in Plant Sciences*, *35*, 293–327. https://doi.org/10.1080/07352689.2016.1267498

Collins, E., Shou, H., Mao, C., Whelan, J., & Jost, R. (2024). Dynamic interactions between SPX proteins, the ubiquitination machinery, and signalling molecules for stress adaptation at a whole-plant level. *Biochemical Journal*, *481*, 363–385. https://doi.org/10.1042/BCJ20230163

Deforges, J., Reis, R. S., Jacquet, P., Sheppard, S., Gadekar, V. P., Hart-Smith, G., Tanzer, A., Hofacker, I. L., Iseli, C., Xenarios, I., & Poirier, Y. (2019). Control of cognate sense mRNA translation by cis-natural antisense RNAs. *Plant Physiology*, *180*, 305–322. https://doi.org/10.1104/pp.19.00043

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Farris, S., Ward, J. M., Carstens, K. E., Samadi, M., Wang, Y., & Dudek, S. M. (2019). Hippocampal subregions express distinct dendritic transcriptomes that reveal differences in mitochondrial function in CA2. *Cell Reports*, *29*, 522–539.e526. https://doi.org/10.1016/j.celrep.2019.08.093

Feng, S., Xu, M., Liu, F., Cui, C., & Zhou, B. (2019). Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in Gossypium australe. *BMC Plant Biology*, *19*, Article 365. https://doi.org/10.1186/s12870-019-1968-7

Gao, S., Wang, B., Xie, S., Xu, X., Zhang, J., Pei, L., Yu, Y., Yang, W., & Zhang, Y. (2020). A high-quality reference genome of wild Cannabis sativa. *Horticulture Research*, *7*, 73 https://doi.org/10.1038/s41438-020-0295-3

Grassa, C. J., Weiblen, G. D., Wenger, J. P., Dabney, C., Poplawski, S. G., Timothy Motley, S., Michael, T. P., & Schwartz, C. J. (2021). A new Cannabis genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytologist*, *230*, 1665–1679. https://doi.org/10.1111/nph.17243

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*, 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hu, B., Jiang, Z., Wang, W., Qiu, Y., Zhang, Z., Liu, Y., Li, A., Gao, X., Liu, L., Qian, Y., Huang, X., Yu, F., Kang, S., Wang, Y., Xie, J., Cao, S., Zhang, L., Wang, Y., Xie, Q., … Chu, C. (2019). Nitrate-NRT1.1B-SPX4 cascade integrates nitrogen and phosphorus signalling networks in plants. *Nature Plants*, *5*, 401–413. https://doi.org/10.1038/s41477-019-0384-1

Jiang, H., Li, Y., Luan, M., Huang, S., Zhao, L., Yang, G., & Pan, G. (2022). Single-molecule real-time sequencing of full-length transcriptome and identification of genes related to male development in *Cannabis sativa*. *Plants*, *11*(24), 3559.

Jost, R., Berkowitz, O., Pegg, A., Hurgobin, B., Tamiru-Oli, M., Welling, M. T., Deseo, M. A., Noorda, H., Brugliera, F., Lewsey, M. G., Doblin, M. S., Bacic, A., & Whelan, J. (2025). Sink strength, nutrient allocation, cannabinoid yield, and associated transcript profiles vary in two drug-type Cannabis chemovars. *Journal of Experimental Botany*, *76*, 152–174. https://doi.org/10.1093/jxb/erae367

Kashkan, I., Timofeyenko, K., & Růžička, K. (2022). How alternative splicing changes the properties of plant proteins. *Quantitative Plant Biology*, *3*, e14 https://doi.org/10.1017/qpb.2022.9

Klepikova, A. V., & Penin, A. A. (2019). Gene Expression maps in plants: Current state and prospects. *Plants*, *8*, 309.

Krouk, G., & Kiba, T. (2020). Nitrogen and Phosphorus interactions in plants: From agronomic to physiological and molecular insights. *Current Opinion in Plant Biology*, *57*, 104–109. https://doi.org/10.1016/j.pbi.2020.07.002

Laverty, K. U., Stout, J. M., Sullivan, M. J., Shah, H., Gill, N., Holbrook, L., Deikus, G., Sebra, R., Hughes, T. R., Page, J. E., & van Bakel, H. (2019). A physical and genetic map of Cannabis sativa identifies extensive rearrangements at the THC/CBD acid synthase loci. *Genome Research*, *29*, 146–156. https://doi.org/10.1101/gr.242594.118

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, Y., Dai, C., Hu, C., Liu, Z., & Kang, C. (2017). Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. *The Plant Journal*, *90*, 164–176. https://doi.org/10.1111/tpj.13462

Lu, H. C., & Mackie, K. (2016). An introduction to the endogenous cannabinoid system. *Biological Psychiatry*, *79*, 516–525. https://doi.org/10.1016/j.biopsych.2015.07.028

Lynch, R. C., Padgitt-Cobb, L. K., Garfinkel, A. R., Knaus, B. J., Hartwick, N. T., Allsing, N., Aylward, A., Bentz, P. C., Carey, S. B., Mamerto, A., Kitony, J. K., Colt, K., Murray, E. R., Duong, T., Chen, H. I., Trippe, A., Harkess, A., Crawford, S., Vining, K., & Michael, T. P. (2025). Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. *Nature*, *643*(8073), 1001–1010. https://doi.org/10.1038/s41586-025-09065-0

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*, 35.

Palos, K., Yu, L., Railey, C. E., Nelson Dittrich, A. C., & Nelson, A. D. L. (2023). Linking discoveries, mechanisms, and technologies to develop a clearer perspective on plant long noncoding RNAs. *The Plant Cell*, *35*, 1762–1786. https://doi.org/10.1093/plcell/koad027

Palusa, S. G., Ali, G. S., & Reddy, A. S. (2007). Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: Regulation by hormones and stresses. *Plant Journal*, *49*, 1091–1107. https://doi.org/10.1111/j.1365-313X.2006.03020.x

Paz-Ares, J., Puga, M. I., Rojas-Triana, M., Martinez-Hevia, I., Diaz, S., Poza-Carrión, C., Miñambres, M., & Leyva, A. (2022). Plant adaptation to low phosphorus availability: Core signaling, crosstalks, and applied implications. *Molecular Plant*, *15*, 104–124. https://doi.org/10.1016/j.molp.2021.12.005

Pertea, G., & Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Res*, *9*, 304. https://doi.org/10.12688/f1000research.23297.1

Petrillo, E. (2023). Do not panic: An intron-centric guide to alternative splicing. *Plant Cell*, *35*, 1752–1761. https://doi.org/10.1093/plcell/koad009

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*, W160–W165.

Reddy, A. S. N. (2004). Plant serine/arginine-rich proteins and their role in pre-mRNA splicing. *Trends in Plant Science*, *9*, 541–547. https://doi.org/10.1016/j.tplants.2004.09.007

Reis, R. S., & Poirier, Y. (2021). Making sense of the natural antisense transcript puzzle. *Trends in Plant Science*, *26*, 1104–1115. https://doi.org/10.1016/j.tplants.2021.07.004

Ren, G., Zhang, X., Li, Y., Ridout, K., Serrano-Serrano, M. L., Yang, Y., Liu, A., Ravikanth, G., Nawaz, M. A., Mumtaz, A. S., Salamin, N., & Fumagalli, L. (2021). Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Science Advances*, *7*, eabg2286. https://doi.org/10.1126/sciadv.abg2286

Ryu, B.-R., Gim, G.-J., Shin, Y.-R., Kang, M.-J., Kim, M.-J., Kwon, T.-H., Lim, Y.-S., Park, S.-H., & Lim, J.-D. (2024). Chromosome-level haploid assembly of *Cannabis sativa* L. cv. Pink Pepper. *Scientific Data*, *11*, 1442. https://doi.org/10.1038/s41597-024-04288-8

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M. E., & Usadel, B. (2019). MapMan4: A refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, *12*, 879–892. https://doi.org/10.1016/j.molp.2019.01.003

Severson, T. F., & Adams, K. L. (2023). Transcriptome-wide characterization of alternative splicing in five drug-type cultivars of Cannabis sativa. *Botany*, *101*, 243–254. https://doi.org/10.1139/cjb-2022-0099

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*, 135–145. https://doi.org/10.1002/pro.3290

Small, E. (2015). Evolution and classification of *Cannabis sativa* (Marijuana, Hemp) in relation to human utilization. *The Botanical Review*, *81*, 189–294. https://doi.org/10.1007/s12229-015-9157-3

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., Harrow, J., & Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, *10*, 1177–1184. https://doi.org/10.1038/nmeth.2714

Tognacca, R. S., Rodríguez, F. S., Aballay, F. E., Cartagena, C. M., Servi, L., & Petrillo, E. (2022). Alternative splicing in plants: Current

knowledge and future directions for assessing the biological relevance of splice variants. *Journal of Experimental Botany*, *74*, 2251–2272. https://doi.org/10.1093/jxb/erac431

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, *19*, 40. https://doi.org/10.1186/s13059-018-1417-1

Ueda, Y., Kiba, T., & Yanagisawa, S. (2020). Nitrate-inducible NIGT1 proteins modulate phosphate uptake and starvation signalling via transcriptional regulation of SPX genes. *The Plant Journal*, *102*, 448–466. https://doi.org/10.1111/tpj.14637

UNODC. (2023). *UN world drug report 2023*. https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2023.html

Wee, Y. B., Berkowitz, O., Ng, S., Pegg, A., Whelan, J., & Jost, R. (2025). Metabolic and transcriptomic analyses identify coordinated resource reallocation in response to phosphate supply in hemp. *bioRxiv*. https://doi.org/10.1101/2025.09.18.677093

Werneke, J. M., Chatfield, J. M., & Ogren, W. L. (1989). Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenase activase polypeptides in spinach and Arabidopsis. *The Plant Cell*, *1*, 815–825.

Wild, R., Gerasimaite, R., Jung, J.-Y., Truffault, V., Pavlovic, I., Schmidt, A., Saiardi, A., Jessen, H. J., Poirier, Y., Hothorn, M., & Mayer, A. (2016). Control of eukaryotic phosphate homeostasis by inositol polyphosphate sensor domains. *Science*, *352*, 986–990. https://doi.org/10.1126/science.aad9858

Wu, B., Li, Y., Li, J., Xie, Z., Luan, M., Gao, C., Shi, Y., & Chen, S. (2021). Genome-wide analysis of alternative splicing and non-coding RNAs reveal complicated transcriptional regulation in *Cannabis sativa* L. *International Journal of Molecular Sciences*, *22*, 11989. https://doi.org/10.3390/ijms222111989

Xu, N., Wang, R., Zhao, L., Zhang, C., Li, Z., Lei, Z., Liu, F., Guan, P., Chu, Z., Crawford, N. M., & Wang, Y. (2016). The Arabidopsis NRG2 protein mediates nitrate signaling and interacts with and regulates key nitrate regulators. *The Plant Cell*, *28*, 485–504. https://doi.org/10.1105/tpc.15.00567

Zhang, N., & Portis, A. R. (1999). Mechanism of light regulation of Rubisco: A specific role for the larger Rubisco activase isoform involving reductive activation by thioredoxin-f. *Proceedings of the National Academy of Sciences*, *96*, 9438–9443. https://doi.org/10.1073/pnas.96.16.9438

Zhang, R., Kuo, R., Coulter, M., Calixto, C. P. G., Entizne, J. C., Guo, W., Marquez, Y., Milne, L., Riegler, S., Matsui, A., Tanaka, M., Harvey, S., Gao, Y., Wießner-Kroh, T., Paniagua, A., Crespi, M., Denby, K., Hur, A. B., Huq, E., … Brown, J. W. S. (2022). A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-seq analysis. *Genome Biology*, *23*, 149. https://doi.org/10.1186/s13059-022-02711-0

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Berkowitz, O., Jost, R., Ng, S., Bacic, A., Whelan, J., & Lewsey, M. G. (2026). Transcriptome assemblies for two drug-type cannabis chemotypes by long-read RNA sequencing. *The Plant Genome*, *19*, e70240. https://doi.org/10.1002/tpg2.70240