RESEARCH   **Open Access**

Trans2express – *de novo* transcriptome assembly pipeline optimized for gene expression analysis

Aleksandra M. Kasianova[1,2], Aleksey A. Penin[1], Mikhail I. Schelkunov[1,2], Artem S. Kasianov[1], Maria D. Logacheva[1,2] and Anna V. Klepikova[1*]

Abstract

Background As genomes of many eukaryotic species, especially plants, are large and complex, their *de novo* sequencing and assembly is still a difficult task despite progress in sequencing technologies. An alternative to genome assembly is the assembly of transcriptome, the set of RNA products of the expressed genes. While a bunch of *de novo* transcriptome assemblers exists, the challenges of transcriptomes (the existence of isoforms, the uneven expression levels across genes) complicates the generation of high-quality assemblies suitable for downstream analyses.

Results We developed Trans2express – a web-based tool and a pipeline of *de novo* hybrid transcriptome assembly and postprocessing based on rnaSPAdes with a set of subsequent filtrations. The pipeline was tested on *Arabidopsis thaliana* cDNA sequencing data obtained using Illumina and Oxford Nanopore Technologies platforms and three non-model plant species. The comparison of structural characteristics of the transcriptome assembly with reference *Arabidopsis* genome revealed the high quality of assembled transcriptome with 86.1% of *Arabidopsis* expressed genes assembled as a single contig. We tested the applicability of the transcriptome assembly for gene expression analysis. For both *Arabidopsis* and non-model species the results showed high congruence of gene expression levels and sets of differentially expressed genes between analyses based on genome and based on the transcriptome assembly.

Conclusions We present Trans2express – a protocol for *de novo* hybrid transcriptome assembly aimed at recovering of a single transcript per gene. We expect this protocol to promote the characterization of transcriptomes and gene expression analysis in non-model plants and web-based tool to be of use to a wide range of plant biologists.

Keywords *De novo* transcriptome assembly, Assembly pipeline, Online tool, Longest isoform, Trans2express

---

*Correspondence:
Anna V. Klepikova
annklepikova@gmail.com
[1]Institute for Information Transmission, Russian Academy of Sciences, Moscow, Russia
[2]Skolkovo Institute of Science and Technology, Moscow, Russia

Background

Despite great progress in sequencing technology and decrease of sequencing costs, large-scale characterization of complete genomes still remains challenging and resource consuming. However, for many types of biological studies complete genome assembly is not necessary, and the data on the sequence of expressed genes RNA products (transcriptome) is sufficient. These directions of research include (but not limited to) the search of genes involved in traits of interest and their functional validation [1, 2], the investigation of the evolution of gene families and genetic networks [3, 4], the search of horizontal gene transfer [5], phylogenetic and population genetic studies [6, 7]. The availability of transcriptome sequences also facilitates genetic studies enabling the development of markers for genetic mapping [8, 9] and design of guide RNA for genetic editing.

De novo transcriptome assembly is usually carried out by shotgun sequencing with short read technologies [10–12]. In this case the typical primary outcome of *de novo* assembly is the high number (several hundred thousand) of relatively short contigs (N50 ~ 1 kb) most of which represent only partial sequence of a transcript [13, 14]. At the same time, most applications outlined above require one sequence per gene (ideally, full-length transcript of the isoform with the longest CDS). In order to reduce the number of contigs several stages of filtration and/or clustering are usually done [10, 15]. This approach is error-prone and may result in missing transcripts by filtering them based on length or by collapsing (which is especially critical for polyploids). Many applications, especially ones dealing with quantitative analysis, are sensitive to these errors. In particular, it was shown that the use of short read-based assemblies for differential expression analysis leads to both false negative (the non-detection of differential expression, DE) and false positive (false detection of DE) errors [16].

Long reads generated by third-generation technologies such as PacBio and Oxford Nanopore Technologies (ONT) have potential to overcome these limitations because they allow to get complete transcript sequences. However, the accuracy of these technologies is still low (up to 10% of sequencing errors [17]). Moreover, the predominant type of errors are insertions and deletions which can affect ORF prediction [18]. Thus, by now the most prospective approach for the *de novo* assembly is the combination of short but accurate Illumina reads and long reads from PacBio or ONT [19]. It is widely used for genomes; there exists a number of tools for hybrid genome assembly (e.g. SPAdes [20], MaSuRCA [21], Unicycler [22]).

Transcriptome assembly however has its own challenges (the existence of isoforms, the uneven expression levels across genes), which makes them unsuitable and

calls for the development of new protocols. By now the number of hybrid *de novo* transcriptome assemblers is limited and includes IDP-denovo [23] and rnaSPAdes [24]. IDP-denovo algorithm is a multistep procedure that includes (1) assembly using short reads alone; (2) error correction of long reads using mapping of short reads; (3) mapping of corrected long reads on short reads-alone assembly and scaffolding of short reads contigs. The obvious limitation of this approach is the accuracy of long reads. The authors of the algorithm tested it on PacBio reads which are more accurate than ONT and have a more random error profile [17]. Though error correction using information from short reads greatly improves accuracy of long reads generated by PacBio [25], for ONT long reads correction is not that efficient due to multiple indels [18]. Besides this, recent benchmarking of error correction algorithms shows that it generates a number of artifacts, in particular removal of minor isoforms and lowly expressed genes and distortion of multigene families [26]. Also, IDP-denovo is highly demanding for computational resources.

An alternative approach is the use of long reads for the resolution of a graph obtained from the assembly of short reads. The program that realizes this approach is rnaSPAdes [24, 27]. SPAdes is a family of assembly tools which includes modules for standard genome, metagenome, transcriptome and single cell genome data. Specialized module for transcriptome assembly, rnaSpades, is an incorporation of a hybrid genome assembler hybrid-SPAdes [28] into a transcriptome assembly pipeline. rnaSPAdes allows the identification of potential gene isoforms and combination them into groups, reducing the transcriptome ambiguity, however, the number of resulting transcripts is still much higher than the number of genes [24].

We suggest a pipeline of *de novo* assembly of a transcriptome optimized for downstream gene-centered analyses – Trans2express. The pipeline includes *de novo* assembly using rnaSPAdes and several subsequent filtrations aiming to reduce the number of contigs per gene (these include incompletely assembled transcripts and isoforms). In order to test the efficiency of this approach, we sequenced and assembled the transcriptome of *Arabidopsis thaliana*, a model object of plant genetics with a very well characterized genome. We assembled and functionally annotated the transcriptome without reliance on the reference genome, and then compared the results with ones based on the genome. Since the goal of many transcriptome projects is the estimation of gene expression levels, we also compared the results of the analysis of differential gene expression based on reference and on *de novo* transcriptome assembly.

Trans2express is also available as an online tool (https://t2e.online) with easy-to-use interface.

Methods

Plant growth

A. thaliana (Columbia accession, CS70000) was grown in a climate chamber (16 h light/ 8 h dark, 22℃, 60% relative humidity) for 21 days after germination. For transcriptome assembly mature leaves were collected in three replicates (each replicate represented a single leaf from a single plant) at ZT10 (Zeitgeber 10, ten hours after turning the light on). For differential expression analysis mature leaves were harvested at ZT09 (light) and ZT21 (dark) in two replicates; each replicate consists of a single leaf from seven plants.

Sample preparation and sequencing

RNA was extracted using Qiagen RNeasy Mini kit (Venlo, Netherlands). RNA integrity was assessed using capillary electrophoresis on Agilent Bioanalyzer2100 (Santa Clara, CA, USA). NEBNext Ultra II RNA directional kit was used for Illumina library preparation (New England Biolabs, Ipswich, MA, USA) following manufacturer instructions. Libraries were sequenced on HiSeq4000 (Illumina, San Diego, CA, USA) in a paired mode with read length 150. For ONT libraries RNA was converted to cDNA using Mint cDNA synthesis kit (Evrogen, Moscow, Russia) with 18 cycles of amplification with the following modification: cDNA synthesis is primed with oligonucleotides that contain not only dT part and adapter part but also custom barcode, specific for each sample. Amplified cDNA was used as input for library preparation using standard protocol for genomic DNA with LSK-109 kit. Sequencing was performed on MinION with 9.4.1 flow cell.

Illumina read processing

Raw Illumina reads were trimmed using fastp 0.23.4 [29] with following parameters: "-l 50 --cut_front --cut_right".

ONT read processing

Basecalling was performed by Guppy 3.4.5 (Oxford Nanopore Technologies, Oxford, United Kingdom); reads with mean quality less than 7 were removed ("*--qscore_filtering --min_qscore 7*"). After that, reads were processed by a custom pipeline NTproc (https://github.com/shelkmike/NTproc). NTproc performs the following operations consecutively:

1. NTproc trims PCR adapters.
2. NTproc removes all reads that do not have PCR adapters on both ends. This allows to eliminate reads coming from cDNAs that were fragmented after the reverse transcription or were not sequenced completely.
3. NTproc trims remaining (non-PCR) adapters.
4. NTproc removes reads that contain barcode sequences on both ends. Such reads likely originate from molecular chimeras.
5. NTproc makes reverse-complements of reads that have barcodes on 5' ends. Since barcoded primers bind to 3' ends of RNAs, this operation makes all reads oriented in the 5'→3' direction of their RNAs. In other words, reads become oriented such that their poly-A tails are on the right.

Transcriptome assembly

The transcriptome was assembled using the rnaSPAdes v. 3.15.5 [27] with the following command: *rnaspades.py -m 1000 -t 100 -1 Illumina_R1.fastq -2 Illumina_R2.fastq --nanopore nanopore.fastq -o results*.

The selection of isoforms with longest CDS

ORF prediction was performed using TransDecoder v. 5.5.0 [30] and DIAMOND tool v. 2.1.8.162 [31]. At the first stage, module TransDecoder.LongOrfs with default parameters was used to define the longest ORF. The command used: *TransDecoder.LongOrfs -t transcripts.fasta*, where *transcripts.fasta* is an output of rnaSPAdes. Then, protein sequences from non-redundant NCBI protein database were used to create a database with the command *diamond makedb --in nr --taxonmap prot.accession2taxid --taxonnodes nodes.dmp --taxonnames names.dmp --db nr* and an output of TransDecoder.LongOrfs (amino acid sequences) were searched against created database with DIAMOND blastp (the command *diamond blastp --query longest_orfs.pep --db nr.dmnd --max_target_seqs 1 --outfmt 6 --threads 100 --out blastp.outfmt6*). At the last stage ORFs and results of DIAMOND search were used as input for another TransDecoder module, TransDecoder.Predict with parameter "*--single_best_only*" (the command *TransDecoder.Predict -t transcripts.fasta --retain_blastp_hits blastp.outfmt6 --single_best_only*). After that, for each group ("a gene") defined by rnaSPAdes the transcript (isoform) with longest protein-coding sequence was retained using a python custom script implemented in the pipeline.

Transcript clustering

The longest isoforms were clustered using CD-HIT-EST v. 4.8.1 [32] with the following command *cd-hit-est -i cds_longest_iso.fasta -o idy_0.98_overlap_0.6.fasta -c 0.98 -G 0 -aS 0.6 -T 32 -p 1 -g 1*, which returns a set of longest isoform from each cluster. A custom script was used to extract corresponding peptides and retain only isoforms with longest CDS in the gff3 annotation file.

Assembly filtration

The amino acid sequences of the assembly were searched against the non-redundant NCBI protein database (nr. dmnd) using the DIAMOND tool v. 2.1.7.162 [31] with the command *diamond blastp --db nr.dmnd --query prots_after_clustered.fasta --evalue 1e-5 --max-target-seqs 8 --threads 100 --out diamond_compairing_res. tsv --outfmt 6 qseqid sseqid staxids*. The python custom script implemented in the pipeline was used to identify foreign transcripts. If 50% or more of BLAST hits with taxon IDs (or three or more BLAST hits with taxon IDs) did not belong to the Embryophyta clade, a transcript was considered contaminant and excluded from the assembly.

Gene Ontology annotation of the transcripts

Gene Ontology annotation of transcriptome assembly was done using PANNZER2 [33]. To obtain GO annotation using Python queries and Beautiful Soup libraries, we uploaded amino acid sequences of transcripts to the PANNZER2 website (http://ekhidna2.biocenter.helsinki. fi/cgi-bin/sans/dump_panz.cgi).

For the emulation of a species from an unexplored taxon all Brassicaceae species were excluded from the protein database downloaded from Phytozome 13 [34] and remaining sequences were used to create DIAMOND blastp database with command *diamond --in phyto_go_db/phyto_prots_for_GO.faa --db eggnog_proteins.dmnd*. Next, eggNOG-mapper v2 [35] was used to create GO annotation with following commands *download_eggnog_data.py -y*, and *emapper.py -m diamond --itype proteins -i protein_fasta.fasta -o arabidopsis --cpu 30 --dmnd_db eggnog_proteins.dmnd*.

Illumina and ONT read mapping

Illumina paired reads were aligned on the reference *A. thaliana* genome (TAIR10.1) and reads per gene were counted using STAR v. 2.7.10b [36] with parameters *--sjdbGTFfeatureExon exon --sjdbGTFtagExonParentTranscript transcript_id --sjdbGTFtagExonParentGene gene_id --quantMode GeneCounts*.

ONT reads were aligned on the reference *A. thaliana* genome (TAIR10.1) using minimap2 v. 2.24 [37] with parameters *-ax splice*.

The coverage of *Arabidopsis* genes by assembled transcripts

Assembled transcripts were mapped on the reference *A. thaliana* genome (TAIR10.1) using minimap2 v. 2.24 [37] with parameter *"-x splice"*. Samtools view v. 1.15.1 [38] with parameters *"-Sb"* was used to convert minimap2 output to bam format and bam file was sorted using samtools sort. To match the assembled transcripts with *Arabidopsis* genes bedtools intersect v. 2.30.0 [39] was used

with parameters *"-wao"* and *A. thaliana* genome annotation. The number of transcripts covering each gene was extracted from bedtools intersect output.

Differential gene expression analysis

Differentially expressed genes were identified with the R package "DESeq2" v. 1.38.3 [40] and thresholds of |logFC| ≥ 1 and FDR < 0.05.

GO enrichment analysis

GO enrichment analysis was performed with the R package "topGO" v. 2.48.0 [41]. P-values of all tests were adjusted via FDR (R "stats" function p.adjust). Terms with FDR < 0.05 and fold enrichment value > 2 were considered statistically significant.

Pipeline availability

The Trans2express pipeline is available at GitHub via link: https://github.com/albidgy/trans2express. The web-based version of Trans2express is available at https://t2e. online/.

Results

Transcriptome assembly pipeline

The Trans2express pipeline (the overview is depicted at Fig. 1) is based on the rnaSPAdes tool [24] used in a hybrid transcriptome assembly mode. rnaSPAdes demonstrated an excellent performance both in short-read mode benchmark [42] and in comparison of hybrid assemblers [24]. As an output rnaSPAdes generates a set of contigs corresponding to transcripts (mRNA isoforms) grouped by alleged correspondence to a locus.

As multiple isoforms in an assembly are known to decrease gene expression analysis performance [16], a transcriptome filtration is a crucial next step. The longest isoform for each group is extracted using TransDecoder [30] and DIAMOND [43]. First, module TransDecoder. LongOrfs with default parameters defines the longest open reading frames (ORFs) of at least 100 amino acids in each transcript. The output amino acid sequences are searched against non-redundant NCBI protein database with DIAMOND and parameter *"-max_target_seqs 1"*. Next, ORFs and results of DIAMOND search are used as an input for another TransDecoder module, TransDecoder.Predict with parameter *"--single_best_only"*. As a result, lists of CDSs and proteins for each transcript are generated. Using a custom script, one transcript with the longest CDS is retained for each rnaSPAdes group.

The next step of assembly filtration includes clustering of transcripts with CD-HIT [32]. CD-HIT performs local alignment and clusters transcripts by sequence similarity. The parameters of CD-HIT are relaxed with respect to similarity and length overlapping in order to increase the isoform number reduction.
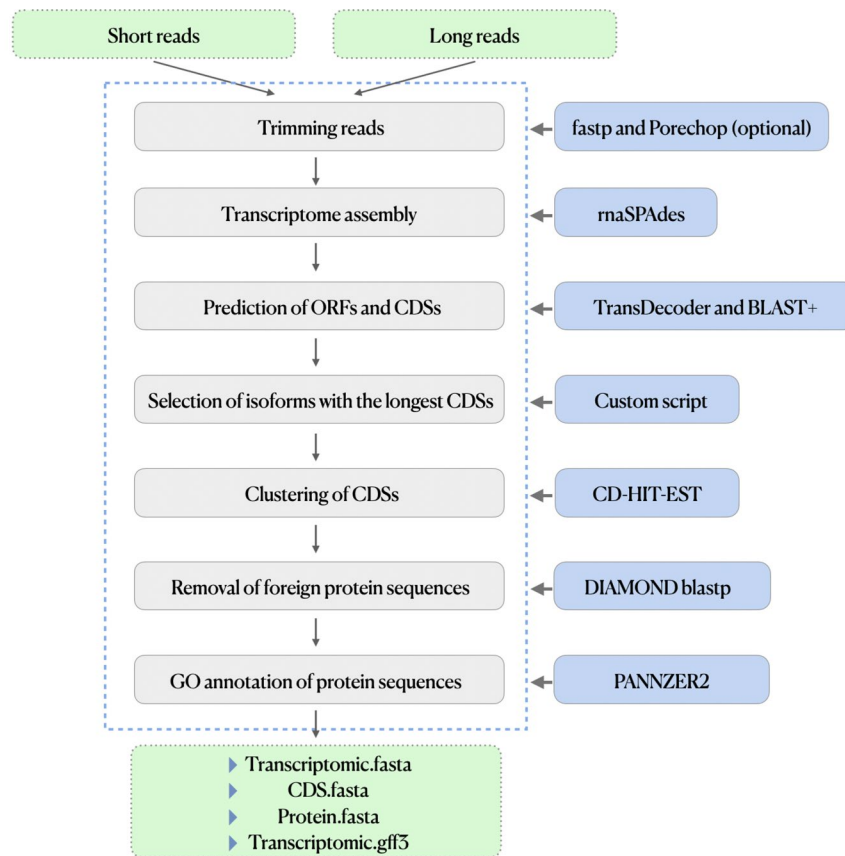
Fig. 1 The overview of the Trans2express pipeline

At the last step the search of contaminant transcripts is done. If such transcripts are present in the assembly, they will be filtered out. The amino acid sequences corresponding to CDS of the assembly transcripts are searched against the non-redundant NCBI protein database (nr.dmnd, accessed at April 9th, 2023) using blastp of the DIAMOND tool [31]. Taxon IDs of BLAST hits are stored for each transcript; if three or more out of five hits do not belong to Embryophyta clade, a transcript is considered foreign and is removed from the assembly. If less than five hits with taxon IDs are detected, the transcript is also considered as contaminant and removed from the assembly.

After these steps the resulting assembly consists mostly of single transcripts (represented by the longest CDS) per expressed gene of a studied organism.

In order to generate and check biological hypotheses on the process-of-interest it is essential to have functional annotation of the transcripts. Gene Ontology (GO) [44] is the most popular functional classification of genes and plenty of methods for GO annotation inference based on sequence similarity are available so far [45]. In the current pipeline we use command line version of PANNZER2 [33] in order to create a GO annotation of transcriptome assembly.

Trans2express pipeline can be easily download from GitHub (https://github.com/albidgy/trans2express), installed at Linux of MacOS machines and executed from the command line.

Trans2express online tool

We developed an online version of Trans2express pipeline (https://t2e.online/trans2express/). A user should upload short (both single- and paired-end reads are supported) and long reads and provide an e-mail address where notification of the job completion will be sent.

Assembly of *Arabidopsis thaliana* transcriptome

To test this pipeline we obtained RNA-seq data of *A. thaliana* Columbia accession (leaves, three replicates). Each replicate was sequenced both on Illumina and Oxford Nanopore platform resulting in 105 M of high-quality short reads and 1.4 M of filtered long reads (Additional file 1: Table S1).

To assess the congruence of technologies regarding gene expression Illumina and ONT reads were mapped separately on *Arabidopsis* reference genome (Additional file 1: Table S1). The number of genes that were identified as expressed based on Illumina data was 15 546 (number of reads greater or equal to 16 reads in each

replicate, strict threshold [46]), while ONT identified 8 526 expressed genes with threshold one read in each replicate. The majority of genes (97.4%) detected by ONT were also detected in Illumina data and their expression levels were congruent between platforms (Additional file 2: Fig. S1). Genes present in Illumina or ONT data only were mostly expressed at low level and can potentially be absent in sequencing library due to the stochastic effects during library preparation.

ONT and Illumina reads were used for transcriptome assembly via pipeline described above. The primary assembly generated by rnaSPAdes consisted of 55 850 raw contigs (transcripts) with N50=2 032 bp which were clustered by rnaSPAdes into 43 054 isoform groups. Raw transcripts were submitted to TransDecoder for ORF annotation, then the longest CDS of each rnaSPAdes-generated isoform group was selected. 19 540 transcripts corresponding to the longest CDSs were combined into a draft transcriptome assembly (N50=2 017 bp). At the next stage CDSs of the draft transcriptome assembly were clustered by sequence similarity using CD-HIT, which returned the longest CDS of each cluster. Transcripts of 19 110 CD-HIT-selected CDSs (N50=2 036 bp) were subjected to the search of contaminating transcripts. Transcripts with 50% or more non-plant hits in the first five blastp results were filtered out, as well as transcripts containing ambigious nucleotides (N) in their sequences. The final transcriptome assembly included 18 281 transcripts (N50=2 051 bp).

Comparison of reference-based and assembly-based results

In an ideal case, the results of transcriptome assembly of a certain organ/tissue/condition should resemble the results based on a reference genome for the same organ/tissue/condition in terms of the number and content of expressed genes.

At first step we checked the correctness of the transcriptome assembly. In an ideal case the contigs of the assembly should have 1-to-1 correspondence with *Arabidopsis* representative gene models and to align on their sequence on the whole length. Transcripts of both draft (after ORF prediction but before clustering with CD-HIT) and final assemblies were aligned on the *Arabidopsis* reference genome. 96.0% and 99.4% of transcripts for draft and final assemblies, respectively, were successfully mapped on *Arabidopsis* genes (TAIR10.1 genome annotation). The contigs of the final assembly covered 86.1% of *Arabidopsis* genes expressed in the samples in a single copy (while for the draft assembly before CD-HIT clustering 84.8% of contigs matched a single gene). The maximum number of contigs mapped on one gene (i.e. isoforms) dropped from 11 to 9 (prior to CD-HIT clustering draft assembly comparing with final assembly), indicating the importance of an additional clustering step with CD-HIT.

The identity between translated sequences of *Arabidopsis* CDS and corresponding contigs showed a perfect match for 82.2% of 17 679 gene-contig pairs having identity≥90 (Fig. 2a). The lengths of corresponding proteins arising from transcriptome assembly and reference genome were similar: 87.6% of protein pairs had length ratio close to one (Fig. 2b), indicating on that the contigs represent full-length transcripts.

The next step was the functional annotation using Gene Ontology. As Gene Ontology is a hierarchical network, the different GO annotation procedures may result in assignment of GO terms to a gene at different levels of the Ontology tree. To make annotations more comparable, GO annotations for both *A. thaliana* genome and transcriptome assembly were brought to the third level of
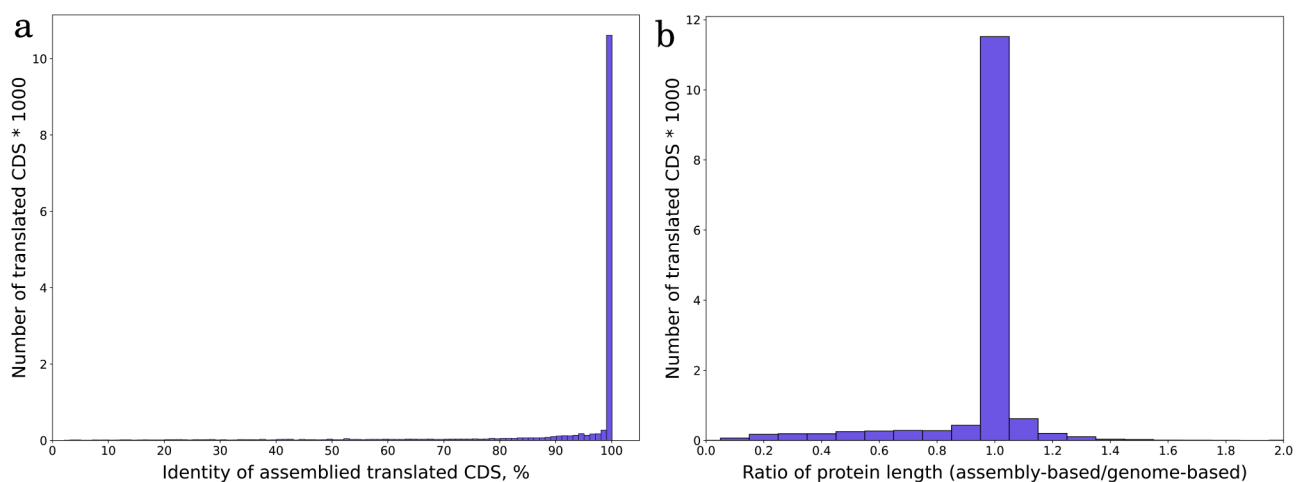

Fig. 2 Congruence between reference-based and transcriptome assembly-based transcripts. **(a)** The identity of reference-based and assembly-based CDS at amino acid level. **(b)** The length ratio of translated reference-based and assembly-based CDS

the Ontology tree counting from the top terms. 17 679 of genes with a single corresponding contig were used in the comparison of GO annotation. 85.6% of transcripts were annotated with at least one GO term. Annotation of 80.7% of contigs at least partially coincided with the annotation of corresponding genes and only 1 196 contigs had annotation that included the same categories as corresponding *Arabidopsis* genes (i.e. the annotation for genes was nested within that of contigs). 14.4% of contigs were not annotated with any GO terms; the majority (92.5%) of the corresponding genes had GO terms assigned and were enriched in growth processes, stress response, cell wall metabolism, ribonucleases, translation, and kinases (Additional file 1: Table S2). GO annotation inferred with exclusion of Brassicales from the protein database (in order to simulate the procedure for a species from a poorly explored clade) returned much poorer results. Only 56.1% of transcripts were annotated with at least one GO term. 45.5% of transcripts had somehow matching with genome ground truth annotation, with 3 174 transcripts having annotation that included the same categories as corresponding *Arabidopsis* genes.

Gene expression analysis using transcriptome assembly

To explore the applicability of transcriptome assembly for downstream analyses, in particular, for the investigation of gene expression, we performed a model experiment aimed at the investigation of gene expression profiles in response to the change of darkness to light. To do this we collected leaves of *A. thaliana* in two replicates at nine hours and at 21 h after Zeitgeber (switching the light on, ZT09 and ZT21, respectively). cDNA libraries were sequenced on Illumina, resulting in 37-46 M reads (Additional file 1: Table S3). High-quality reads were mapped on *Arabidopsis* genome and on transcriptome assembly. The comparison of expression levels of genes and contigs

(based on 17 679 genes with a 1-to-1 correspondence to contigs) revealed high correlation (Pearson r^2=0.77, Fig. 3a).

At the next stage, differential expression analysis was conducted using ZT21 ("dark") as a control sample and ZT09 ("light") as an "experimental" sample. The number of differentially expressed (DE) genes/contigs was slightly higher for *Arabidopsis* genome than transcriptome assembly: 1 514 and 1 695 genes were up- and down-regulated in genome-based results, when transcriptome assembly yielded 1 236 and 1 314 DE genes, respectively. Both up- and down-regulated genes highly overlapped between genome and transcriptome gene lists with 87–88% congruence (Fig. 3b). The majority of DE genes discovered by either reference genome or transcriptome assembly only had fold changes close to the threshold (Fig. 3c) and tend to have lower expression level (Additional file 1: Fig. S2).

As TAIR10 *Arabidopsis* annotation was shown to be more precise than TAIR10.1 in gene boundary coordinates [47], we repeated the comparison of DE results on transcriptome assembly and TAIR10 genome annotation obtaining similar results (Additional file 1: Table S4).

Gene Ontology enrichment analysis is a common first step to functional study of expression difference between samples. We conducted GO enrichment analysis for both genome and transcriptome assembly-based DE genes (up- and down-regulated genes were treated separately). Again, genome-based results were wider: 49 terms were found for assembly-based and 210 — for genome-based DE genes; the intersection was 59.2%. Up-regulated assembly-based DE genes had 93 overrepresented terms, and the intersection with 213 genome-based DE genes was 61%. While being less pronounced, GO enrichment of transcriptome assembly-based DE genes still reflects the main processes associated with circadian cycle and
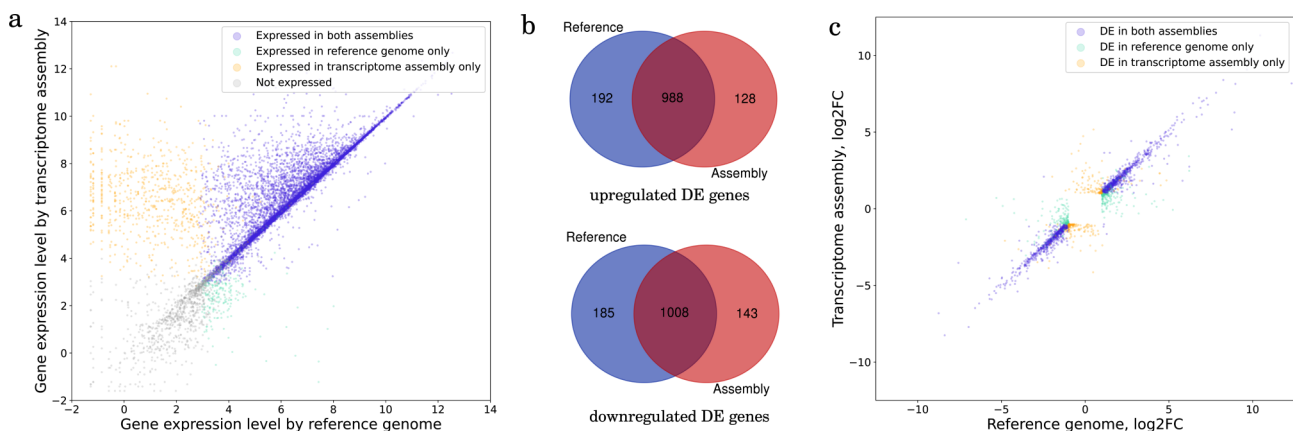

Fig. 3 Congruence of gene expression analysis based on genome reference and on transcriptome assembly of *Arabidopsis thaliana*. **(a)** Gene expression levels estimated using reference-based and assembly-based sequences. **(b)** The overlap of up- and downregulated DE genes between reference-based and assembly-based analysis. **(c)** Fold change of DE genes inferred from reference-based and assembly-based analysis

difference between light and dark plant metabolism (Additional file 1: Table S5).

Validation of the efficiency of Trans2express on non-model species

Analyses described above demonstrate high efficiency of Trans2express in reconstructing transcript sequences for further use as reference for gene expression analysis using data from *Arabidopsis thaliana*. However this might not be the case for other plants, especially ones with larger genomes shaped by polyploidy. To further expand the validation of Trans2express pipeline we applied it to three non-model species – *Zostera marina*, *Cucurbita maxima* and *Nelumbo nucifera* [48–50]. We generated datasets of long and short reads in the same way as described above and used available RNA-seq data for gene expression analysis. Similar to *A. thaliana*, the final assembly resulted in contigs aligning on reference genomes by more than 90% and most contigs had one-to-one correspondence with annotated genes (Additional file 1: Table S6). We used three publicly available RNA-seq datasets ([51, 52] Additional file 1: Table S7) for these species in order to test the efficiency of the obtained assemblies for gene expression analysis. The results showed high congruence of reference-based and assembly-based observations on gene expression (Additional file 1: Table S6). The genes that showed discordant patterns (were differentially expressed only in reference-based or assembly-based analysis) were confined to lower fold changes (Fig. 4).

Discussion

Non-model species, especially plants, often bear large and complex genomes demanding a great deal of resources for sequencing and assembly. Fortunately, a wide set of scientific problems may be addressed with sequences of RNAs, i.e. transcriptome. Among a large variety of *de novo* assembly approaches the hybrid assembly utilizing accurate short reads and transcript-spanning but less accurate long reads is a state-of-the-art approach resulting in high quality transcripts [24].

Many eukaryotic genes are prone to alternative splicing that generates several isoforms per locus. Theoretically *de novo* transcriptome assembly has potential to recognize all expressed isoforms of a gene, and some approaches (e.g. RATTLE [53] or isONform [54] are aimed to reconstruct as many isoforms as possible. However, for many types of downstream analyses the uncollapsed isoforms are disadvantageous (the same is true for allelic variants of transcripts). In case of the use of assembly for evolution-centered studies (phylogenetic analysis, investigation of gene families) isoforms or alleles can be incorrectly recognized as paralogs and thus lead to biases and incorrect biological conclusions. In case of gene expression analysis they will adversely affect the estimation of gene expression levels as many RNA-seq reads will be non-uniquely mapped resulting in a loss of data [16]. Thus, an extensive clustering and filtration of *de novo* transcriptome assembly are crucial steps. There are several pipelines that address this problem, for example based on the minimization of the fraction of duplicated genes in BUSCO gene sets [55]. This approach however might not be applicable for polyploid species (that are especially frequent in plant lineage) because it will collapse paralogs.

Here we present Trans2express – a pipeline for *de novo* transcriptome assembly optimized for one of the most frequent downstream applications – gene expression analysis. We tested our approach on *A. thaliana* leaf transcriptome sequenced on Illumina and ONT platforms. Almost all transcripts of the assembly were mapped on the *Arabidopsis* genome and the quality of transcripts was high in terms of protein identity and length ratio. 86.1% of expressed genes were covered by a single transcript, in contrast to what is typical for transcriptome assemblies. Additional validation comes from
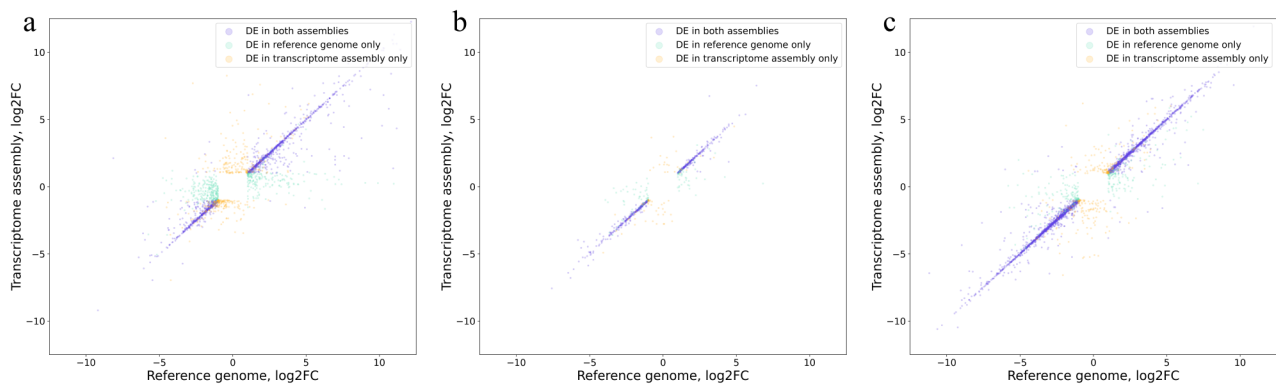

Fig. 4 Congruence of gene expression analysis based on genome reference and on transcriptome assembly of non-model species. Fold change of DE genes inferred from reference-based and assembly-based analysis of **(a)***Zostera marina*, **(b)***Cucurbita maxima*, and **(c)***Nelumbo nucifera* data

three non-model species, for which reference genomes and RNA-seq data were available; the results were also highly congruent.

The congruence of the results of differential expression analysis between reference-based and assembly-based variants indicates on the high potential of the pipeline for gene expression analysis in non-model species. It is important to note that the discrepancies that are still present between reference-based and assembly-based analyses are confined to the genes with lower expression values and with fold change close to the threshold (logFC=1). From the practical point of view this means that when dealing with the results of gene expression analysis based on *de novo* assembly it might be beneficial to set stricter thresholds on FC. This will recover less DE genes but with higher reliability.

The integral part of transcriptome characterization and the differential gene expression analysis is Gene Ontology (GO) annotation and further search of the enrichment of specific GO terms in DE genes. Despite that, quite few studies focus on (or at least discuss) the issues of GO annotation in non-model species. We show here that this step produces very dissimilar results depending of the datasets taken as a basis for homology search and does not work perfectly even for a model species. Indeed, surprisingly, even if the information on *Arabidopsis* is not excluded, the GO annotation of *Arabidopsis* contigs yields quite poor results – only 80.7% had annotations overlapping with the reference ones and ~14% remained unannotated. Under the simulation of the case of non-model species which was done by exclusion of information of *Arabidopsis* and other Brassicaceae only 56.1% annotations overlapped with the reference ones. This highlights the challenge of the GO-annotation in non-model species and calls for the development of more model species representing different plant lineages. Any analysis dealing with GO in non-model species should be treated with caution, especially if it involves cross-species comparisons. Despite these considerations, in case of strong and pronounced effect the results will be biologically relevant, as shows the model experiment done in this study.

Conclusion

Despite the existence of many pipelines for transcriptome assembly and filtering few of them were tested in downstream applications. Here we present Trans2express – a procedure that allows to get a de novo assembly out of combination of short and long reads optimized for the generation of a single isoform per gene. This is important for many applications, including the search of differential gene expression. Using experiments on model plant *Arabidopsis thaliana* we showed that our pipeline allows the assembly of more than 85% of expressed genes in a single isoform. The differential expression analysis showed high concordance between reference-based and *de novo* transcriptome assembly-based results. We expect that the method described here as well as online tool will promote studies on plant species with large genomes, where the transcriptome assembly is a feasible alternative to the genome assembly.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-024-01255-7.

> Supplementary Material 1Please, change Supplementary File 1 to the corrected one, attached in the system.
>
> Supplementary Material 2

Data availability
The *Arabidopsis de novo* transcriptome assembly is available at 10.6084/m9.figshare.24981996. The Illumina and ONT reads used for *de novo Arabidopsis* transcriptome assembly are available at SRA NCBI database (https://www.ncbi.nlm.nih.gov/sra/) under BioProject accession PRJNA1063591. The Illumina and ONT reads used for *de novo Zostera marina*, *Cucurbita maxima* and *Nelumbo nucifera* transcriptome assembly are available at SRA NCBI database (https://www.ncbi.nlm.nih.gov/sra/) under BioProject accession PRJNA1125281. The Illumina reads used for gene expression analysis are available at SRA NCBI database (https://www.ncbi.nlm.nih.gov/sra/) under BioProject accession PRJNA1063592.Trans2express pipeline is available at GitHub via link https://github.com/albidgy/trans2express.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

References
1. Hua X, Song W, Wang K, Yin X, Hao C, Duan B, et al. Effective prediction of biosynthetic pathway genes involved in bioactive polyphyllins in Paris polyphylla. Commun Biol. 2022;5:50.

2. Wang Z-L, Gao H-M, Wang S, Zhang M, Chen K, Zhang Y-Q, et al. Dissection of the general two-step di- *C* -glycosylation pathway for the biosynthesis of (iso) schaftosides in higher plants. Proc Natl Acad Sci USA. 2020;117:30816–23.

3. Neverov AM, Panchin AY, Mikhailov KV, Batueva MD, Aleoshin VV, Panchin YV. Apoptotic gene loss in Cnidaria is associated with transition to parasitism. Sci Rep. 2023;13:8015.

4. Kariñho-Betancourt E, Carlson D, Hollister J, Fischer A, Greiner S, Johnson MTJ. The evolution of multi-gene families and metabolic pathways in the evening primroses (Oenothera: Onagraceae): A comparative transcriptomics approach. Wang D, editor. PLoS ONE. 2022;17:e0269307.

5. Xi Z, Bradley RK, Wurdack KJ, Wong K, Sugumaran M, Bomblies K, et al. Horizontal transfer of expressed genes in a parasitic flowering plant. BMC Genomics. 2012;13:227.

6. Liu L, Wang Z, Su Y, Wang T. Population transcriptomic sequencing reveals allopatric divergence and local adaptation in Pseudotaxus chienii (Taxaceae). BMC Genomics. 2021;22:388.

7. Wen J, Yu Y, Xie D-F, Peng C, Liu Q, Zhou S-D, et al. A transcriptome-based study on the phylogeny and evolution of the taxonomically controversial subfamily Apioideae (Apiaceae). Ann Botany. 2020;125:937–53.

8. Karcı H, Paizila A, Topçu H, Ilikçioğlu E, Kafkas S. Transcriptome sequencing and development of novel genic SSR markers from Pistacia vera L. Front Genet. 2020;11:1021.

9. Xu Y, Zeng A, Song L, Li J, Yan J. Comparative transcriptomics analysis uncovers alternative splicing events and molecular markers in cabbage (Brassica oleracea L). Planta. 2019;249:1599–615.

10. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to *de novo* transcriptome assembly and annotation. Brief Bioinform. 2022;23:bbab563.

11. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20:92, s13059-019-1715–2.

12. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631–56.

13. Freedman AH, Clamp M, Sackton TB. Error, noise and bias in de novo transcriptome assemblies. Mol Ecol Resour. 2021;21:18–29.

14. Faherty SL, Villanueva-Cañas JL, Blanco MB, Albà MM, Yoder AD. Transcriptomics in the wild: hibernation physiology in free-ranging dwarf Lemurs. Mol Ecol. 2018;27:709–22.

15. Razo-Mendivil FG, Martínez O, Hayano-Kanashiro C. Compacta: a fast contig clustering tool for de novo assembled transcriptomes. BMC Genomics. 2020;21:148.

16. Ono H, Ishii K, Kozaki T, Ogiwara I, Kanekatsu M, Yamada T. Removal of redundant contigs from de novo RNA-Seq assemblies via homology search improves accurate detection of differentially expressed genes. BMC Genomics. 2015;16:1031.

17. Cui J, shen N, Lu Z, Xu G, Wang Y, Jin B. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. Plant Methods. 2020;16:85.

18. Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. Nat Biotechnol. 2019;37:124–6.

19. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microbial Genomics [Internet]. 2019 [cited 2023 Aug 15];5. https://www.microbiologyresearch.org/content/journal/mgen/https://doi.org/10.1099/mgen.0.000294

20. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a New Genome Assembly Algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.

21. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29:2669–77.

22. Wick RR, Judd LM, Gorrie CL, Holt KE, Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. Phillippy AM, editor. PLoS Comput Biol. 2017;13:e1005595.

23. Fu S, Ma Y, Yao H, Xu Z, Chen S, Song J et al. B Berger editor 2018 IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. Bioinformatics 34 2168–76.

24. Prjibelski AD, Puglia GD, Antipov D, Bushmanova E, Giordano D, Mikheenko A, et al. Extending rnaSPAdes functionality for hybrid transcriptome assembly. BMC Bioinformatics. 2020;21:302.

25. Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio Long Read Accuracy by Short Read Alignment. Xing Y, editor. PLoS ONE. 2012;7:e46679.

26. Lima L, Marchet C, Caboche S, Da Silva C, Istace B, Aury J-M, et al. Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. Brief Bioinform. 2020;21:1164–81.

27. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019;8:giz100.

28. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. Hybrid SPA Des: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016;32:1009–15.

29. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.

30. Haas BJ. TransDecoder [Internet]. [cited 2023 Mar 24]. https://github.com/TransDecoder/TransDecoder

31. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.

32. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.

33. Törönen P, Holm L. PANNZER —A practical tool for protein function prediction. Protein Sci. 2022;31:118–28.

34. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86.

35. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Tamura K, editor. Molecular Biology and Evolution. 2021;38:5825–9.

36. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

37. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34:3094–100.

38. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.

39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

40. Love MI, Huber W, Anders S. Moderated estimation of Fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

41. Alexa A, Rahnenfuhrer J, topGO. Enrichment Analysis for Gene Ontology. R package version 2.52.0. 2023.

42. Hölzer M, Marz M. De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience. 2019;8:giz039.

43. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.

45. Cozzetto D, Jones DT. Computational Methods for Annotation Transfers from Sequence. In: Dessimoz C, Škunca N, editors. The Gene Ontology Handbook [Internet]. New York, NY: Springer New York; 2017 [cited 2023 Aug 15]. pp. 55–67. http://link.springer.com/https://doi.org/10.1007/978-1-4939-3743-1_5

46. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing Quality Control Consortium. Nat Biotechnol. 2014;32:903–14.

47. Ivanov M, Sandelin A, Marquardt S. TrancriptomeReconstructoR: data-driven annotation of complex transcriptomes. BMC Bioinformatics. 2021;22:290.

48. Olsen JL, Rouzé P, Verhelst B, Lin Y-C, Bayer T, Collen J, et al. The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. Nature. 2016;530:331–5.

49. Sun H, Wu S, Zhang G, Jiao C, Guo S, Ren Y, et al. Karyotype Stability and unbiased fractionation in the Paleo-Allotetraploid Cucurbita genomes. Mol Plant. 2017;10:1293–306.

50. Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, et al. Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn). Genome Biol. 2013;14:R41.

51. Niu M, Xie J, Chen C, Cao H, Sun J, Kong Q, et al. An early ABA-induced stomatal closure, na + sequestration in leaf vein and K + retention in mesophyll confer salt tissue tolerance in Cucurbita species. J Exp Bot. 2018;69:4945–60.

52. Deng X, Yang D, Sun H, Liu J, Song H, Xiong Y, et al. Time-course analysis and transcriptomic identification of key response strategies of *Nelumbo nucifera* to complete submergence. Hortic Res. 2022;9:uhac001.

53. Johnson KA, Krishnan A. Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data. Genome Biol. 2022;23:1.

54. Petri AJ, Sahlin K. isONform: reference-free transcriptome reconstruction from Oxford Nanopore data. Bioinformatics. 2023;39:i222–31.

55. Rivera-Vicéns RE, Garcia-Escudero CA, Conci N, Eitel M, Wörheide G. TransPi—a comprehensive TRanscriptome ANalysiS PIpeline for *de novo* transcriptome assembly. Mol Ecol Resour. 2022;22:2070–86.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.