Evolution and development of flowers, fruits and inflorescences of Phalaenopsis and other orchid species

Pramanik, D.

Chapter 2

Transcriptomics

Dewi Pramanik, Ozan Çiftçi, and Yannick Woudstra

In: de Boer, H. Marcella, M.O. Verstraete, B. and Gravendeel, B., (eds). *Molecular identification of plants: from sequence to species* (pp. 213-232). Pensoft publisher, Sofia-Bulgaria. *https://doi.org/10.3897/ab.e98875*

Dewi Pramanik[1,2,3], **Ozan Çiftçi**[4], **and Yannick Woudstra**[5,6,7,8]

1 Evolutionary Ecology Group, Naturalis Biodiversity Center, The Netherlands

2 Institute of Biology Leiden, Leiden University, The Netherlands

3 National Research and Innovation Agency (BRIN), Indonesia

4 Institute of Environmental Sciences, Leiden University, The Netherlands

5 Royal Botanic Gardens, Kew, United Kingdom

6 Natural History Museum Denmark, University of Copenhagen, Denmark

7 Gothenburg Global Biodiversity Center, Department of Biological and Environmental Sciences, University of Gothenburg, Sweden

8 Department of Plant Sciences, University of Oxford, United Kingdom

2.1 BACKGROUND

Transcriptomics is the study of the transcriptome, which is the entire set of all RNA molecules, including coding and noncoding RNA, that is expressed in a cell, tissue, or organism at a certain spatial, temporal, or developmental stage (Morozova et al., 2009; Piétu et al., 1999). Transcriptomics presents a convenient and cost-effective hybrid approach to study both genomes and biological function at the same time. It is especially useful for making direct links between the genotype and the phenotype as it allows for the accurate detection of gene expression levels in different tissues (Hrdlickova et al., 2017) under different environmental circumstances and even at different spatial scales (Burgess, 2019). Sequencing of only the exons of expressed coding genes (Van Verk et al., 2013) represents a simpler alternative to whole genome sequencing that makes the assembly of nuclear coding genes more attainable.

In plant research, transcriptomics is widely used for studying differential expression, identifying novel genes, and general expression patterns (Shakya et al., 2019). It is also widely used to study genetic diversity in environmental samples ranging from the human (and other animal) microbiomes, to microbes found in or on plants, within soil and in aquatic environments which is referred to as metatranscriptomics (Poretsky et al., 2005; Shakya et al., 2019). For instance, metatranscriptomics can be used to understand how plant-microbiome interactions evolve through time and under different environmental conditions (Shakya et al., 2019).

The first publication studying individual transcripts used Northern blotting for RNA detection, which is a hybridization-based method (Alwine et al., 1977). Further developments included gene expression quantification by sequence/sequencing-based methods including the expressed sequence tag (EST) (Adams et al., 1991), serial analysis of gene expression (SAGE) (Velculescu et al., 1997), massively parallel signature sequencing (MPSS) (Brenner et al., 2000), and cap analysis of gene expression (Shiraki et al., 2003). Microarrays were the first high-throughput method developed for transcriptomics to achieve widespread use due to their affordability and highly sensitive transcript detection (Wang et al., 2019). However, with the introduction of second generation RNA sequencing platforms (RNA-seq), microarrays are no longer widely used (McGettigan, 2013). RNA-seq offers many advantages over microarrays in plant studies including higher genomic coverage (Kukurba and Montgomery, 2015), better resolution of expression differences amongst paralogs (Sundell et al., 2017) and higher precision in co-expression networks for estimating the exact expression levels for lowly expressed genes (Fu et al., 2014). RNA-seq data can accurately quantify expression levels, is inexpensive to acquire, and does not require highly

skilled labor (Wang et al., 2009). The first application of RNA-seq in plant science was with the Arabidopsis transcriptome (Weber et al., 2007). Non-model plants can also be sequenced with RNA-seq since it does not require existing genomic data (Wang et al., 2009).

Currently, RNA-seq data is often acquired using technologies that allow for long read data. Long read RNA-seq data allows reading full transcripts, finding new isoforms, identifying fusion transcripts, identifying long noncoding RNA, simplifying the computational analysis, and reducing PCR biases (Depledge et al., 2019). The two technologies that currently dominate long-read sequencing are Pacific Biosciences (PacBio), single-molecule real-time (SMRT), isoform sequencing (Iso-Seq), and Oxford Nanopore Technologies (ONT) (Amarasinghe et al., 2020). Nevertheless, this technology is not without drawbacks, including high experimental cost, low throughput, and higher read error rate (Stark et al., 2019). This chapter thus emphasises short-read RNA-seq, though protocols for long-read RNA seq are included (Stark et al., 2019).

2.2 EXPERIMENTAL DESIGN

2.2.1 RNA isolation

Experimental design considerations

Isolating a sufficient quantity of high-quality RNA is critical for conducting transcriptome sequencing experiments and their analyses. When designing a protocol, a number of biological replicates should be considered. Biological replication represents RNA harvested from different plants or different sets of independent samples treated under the same conditions. This biological replication is important for assessing variation between samples, and more biological replicates can increase statistical power during analysis. In general, the minimum number of samples for transcriptomics studies is three biological replicates. Once the minimum number of samples and replications is achieved, the following steps are sample treatment and handling, RNA isolation, and RNA quality and quantity testing.

Tissue preparation and homogenization

RNA to be used in transcriptomic experiments is most commonly isolated from a maximum of 100 mg of fresh plant tissue. If not used immediately, collected plant tissue should be frozen in liquid nitrogen and stored at -80 °C. If it is not possible to homogenize the fresh material or to snap-freeze it in liquid nitrogen immediately (e.g., in the field), it should be kept in a preservation buffer that maintains a constant pH to preserve proteins and protect the RNA. RNA stabilization and storage solutions available from manufacturers (e.g., Ambion, Applied Biosystem or RNAlaterTM, Invitrogen, ThermoFisher Scien-

tific, USA) or other preservatives such as a sulfate salt solution (e.g., ammonium sulfate) preserve tissue samples after harvesting in order to retain the quality and quantity of RNA for long periods (Allewell and Sama, 1974). Samples stored in a stabilization solution can be disrupted and homogenized without the use of liquid nitrogen. A similar treatment can also be carried out for soil samples (metatranscriptomics) with an additional sample screening step by sieving the soil sample to separate the sample from organic debris, roots and rocks (mesh 2 mm) before storing at -80 °C or in an RNA preservative solution (e.g., LifeGuard™ Soil Conservation Solution, MO BIO Laboratories Inc., USA) (Carvalhais and Schenk, 2013; Carvalhais et al., 2012, 2013) (Fig 2.1).

A crucial step in tissue preparation is finding the most appropriate method to homogenize the tissue in order to maximize the yield and quality of the RNA. The most common method to homogenize the tissue is snap-freezing in liquid nitrogen and subsequent homogenization/disruption of the tissue by manually grinding with a mortar and pestle or with glass/metal beads and a tissue lyser. However, this is challenging for hard tissue like wood, roots or plant tissues with thick cuticles such as succulent leaves. The combination of snap-freezing in liquid nitrogen, disruption of the tissue by manually grinding, and second grinding with glass/metal beads and a tissue lyser can be a solution to optimize the tissue homogenization of hard tissues. Once the tissue samples are powdered, they can be stored at -80 °C or used immediately for RNA isolation. It is advised to thaw a frozen tissue sample only once and add the lysis buffer immediately to obtain high-quality isolated RNA. It is important that the lytic agent or denaturant comes into contact with the cellular contents when the cells are disrupted. The RNA lysis buffer (e.g., Buffer RLT, Qiagen-USA) is usually composed of phenol and guanidine isothiocyanate. This buffer has two functions as a denaturing agent and stabilizes nucleic acid by preventing the activity of the RNase enzyme.

RNA isolation

Compared to DNA, RNA is less stable due to its chemical structure: RNA is single-stranded and can easily be enzymatically degraded by the abundant amounts of ribonuclease (RNase) that are present in the environment. RNases are secreted through our skin and in the air we breathe out. RNA isolations therefore need to be conducted in RNase-free conditions. Gloves must be worn at all times and the RNA isolation should take place in a fume hood. Designated working spaces and equipment should be cleaned with RNase inhibitors. Common RNase inhibitors to use are strong denaturants such as guanidinium, sodium dodecyl sulfate (SDS), diethyl pyrocarbonate (DEPC), or phenol-based compounds. Additionally, commercially available products include DNase/

RNase AWAY™ (Merck BV, The Netherlands) or bleach (sodium hypochlorite). Keep in mind to also use RNase-free plastics and glassware. The main steps for RNA isolation are similar to the DNA isolation protocol (Przelomska et al., 2022). RNA can be extracted and purified by following protocols described in the literature such as an acidic phenol-chloroform RNA extraction (Chomczynski and Sacchi, 2006) or by using commercial kits (e.g., Qiagen Plant RNA kit, Turbo DNA-free kit, etc.). Commercial extraction kits come with the advantage that there is reduced handling of hazardous reagents and less time needed to prepare the reagents. For woody tissues or other tissue types with high phenolic compounds, lysis buffers with high molecular weight polymers (e.g., polyvinylpyrrolidone (PVP)) that can bind and remove polyphenols and polysaccharides may be required (Maceda-López et al., 2021). The purity of the extracted RNA may be improved by digesting genomic DNA (for example with the RNase-Free DNase Set, Qiagen, USA) and by RNA precipitation to enrich RNA over DNA (for example by using ethanol-glycogen or LiCl precipitation). This can be especially important for obtaining pure RNA from plant samples with high amounts of alkaloids (Leh et al., 2019).

Tissue-specific RNA and single-cell RNA isolation

Single-cell RNA-seq (scRNA-seq) is an advanced method to profile transcriptomes from individual cells. scRNA-seq can be used for cell type identification, transcriptome profiling, and inference of gene regulatory networks across the cell (Rich-Griffin et al., 2020). Several methods are available for tissue-specific or single-cell RNA isolations.

One method for tissue-specific isolation is laser microdissection (LMD), which is based on a histological identification that isolates specific cell types by laser capture and laser cutting (Kivivirta et al., 2019). An area of at least 10 μm^2 with a section thickness of 10 µm must be captured to obtain a sufficient amount, which is approximately 10 pg to 1 ng of RNA for cDNA synthesis (Kivivirta et al., 2019).

Single-cell sequencing can further provide high-resolution functional information on an individual cell. In order to capture single cells for scRNA-seq experiments, fluorescence-activated cell sorting (FACS) with the use of protoplasts is commonly used. This is both a high-throughput and highly specific method (Efroni and Birnbaum, 2016), but it is also labour intensive. Using protoplasts for FACS is challenging as the enzymatic digestion and cleanup process during protoplast isolation results in a stress response that must be accounted for in subsequent data analysis, and generating protoplast cells from plant tissues remains challenging (Long et al., 2021). Recently, an isolated nuclei approach was developed as an alternative to using plant protoplasts. With this approach, it has been shown that it is possible to design single-cell RNA libraries and obtain meaningful trans-

criptomic information from plant cells (Thibivilliers et al., 2020).

RNA quality and quantity

The quality and quantity evaluation of RNA is essential to the success of sequencing experiments and the downstream analysis. The RNA quality and quantity can be evaluated by measuring the UV absorption of a sample. The optical density (OD) ratios at A260/A280 and A260/A230 can be used to determine the RNA purity. Pure RNA has an A260/A280 ratio of 2.1, and an A260/A230 ratio in the range of 2.0-2.2 (Wilfinger et al., 1997). A low A260/A230 ratio may suggest contamination from carbohydrate carry over or residual phenol, while a low A260/280 ratio can indicate contamination from residual phenol, guanidine, or reagents associated with the extraction. These contaminants can affect the downstream application and bias the expression results from qPCR (Carvalhais and Schenk, 2013) (e.g., uneven gene coverage or 3′–5′ transcript bias) (Kukurba and Montgomery, 2015).

Measuring the RNA integrity in order to determine its degradation level is also recommended. Traditionally, RNA integrity was determined by visualizing total RNA using gel electrophoresis and ethidium bromide staining. Intact RNA gives sharp and clear 28S and 18S rRNA bands with an intensity ratio of 28S/18S at 2.0 or higher, in addition to a messenger RNA (mRNA) smear that should be visible between these two distinct bands. A more recent and standardized RNA integrity determination method is determining the RNA integrity number (RIN) with Agilent Bioanalyzer Systems instruments (Agilent Technologies, USA) (Schroeder et al., 2006). The RIN is calculated from total RNA sample characteristics that are based on records of electrophoretic trace data including the ratio of 28S/18S rRNA, the height of the 28S and 18S rRNA peak, and the area between the 18S and 5S rRNA peaks. The RIN software algorithm classifies RNA integrity from 1 to 10, with 1 being the most degraded and 10 being the most intact. A RIN of 7 or higher indicates that the RNA is sufficiently intact for RNA-seq (Jahn et al., 2008). If the isolated RNA is of low quality and quantity, additional precipitation steps may be required to improve the purity of RNA. For large tissue samples, the total amount of harvested RNA is usually between 100 ng and 1 µg. However, for tissue specific RNA, the necessary amounts are between 10 pg to 1 ng. For scRNA-seq, 1000–8000 cells per single-cell suspension are needed (Rich-Griffin et al., 2020) or 300 ng of total RNA for the SMRT PacBio Iso-Seq platform.

2.2.2 Library preparation

The selection of library preparation methods depends on the fragment size, presence of structural features, and sequencing platform. In the Illumina short-read RNA-seq protocol, the library preparation entails four main steps: (1) RNA molecule selection (mRNA enrichment or rRNA depletion), (2) fragmenting the targeted sequence to the desired length and converting fragmented RNA into cDNA, (3) attaching the adapters and PCR amplification to create the cDNA library, and (4) quantifying the library product for sequencing. The library preparation for long-read sequencing is somewhat simpler than for short-read sequencing. The PacBio Iso-Seq protocol consists of three main steps: (1) cDNA synthesis, (2) cDNA amplification, and (3) library construction. With the Oxford Nanopore platform, the sequencing can be done directly from RNA or by using the amplified (or non-amplified) cDNA input (Fig. 2.1).

mRNA enrichment or rRNA depletion

A total RNA sample after extraction contains ribosomal RNA (rRNA), precursor mRNA (pre-mRNA), mRNA, small noncoding RNA (sRNA/sncRNA), and long ncRNA (transcripts longer than 200 nucleotides), where the majority of material is rRNA (Hrdlickova et al., 2017). Total RNA sequencing or whole transcriptome sequencing refers to the sequencing of all RNA molecules, both coding and noncoding. A selection of the mature polyadenylated (poly(A)) mRNA (mRNA with poly(A) tail) can be made in order to sequence the protein-coding regions only. The insertion of a poly(A) tail to the mRNA molecule improves the stability of the molecule and allows the mRNA to be exported from the nucleus and translated into the protein. Since a high percentage of rRNA (> 80%) can interfere with the analysis of mRNA transcripts, an additional step to enrich mRNA or deplete rRNA may be necessary. rRNA depletion is most commonly used to capture unique transcriptome features. In contrast mRNA enrichment is mainly used to increase exonic coverage (Zhao et al., 2018) or for expression profiling studies. mRNA enrichment, also known as poly(A) enrichment, can be done by selecting only polyadenylated mRNA from total RNA. During this procedure, the total RNA is mixed with oligo (dT) primers and a high-salt binding buffer to promote binding to paramagnetic beads. Oligo dT bound to the bead's surface hybridizes to the poly(A) containing mRNA. Precipitate the mRNA bound to beads with a magnet, followed by application of a high-salt washing buffer to discard unbound RNA while retaining oligo (dT) bound poly(A) mRNAs (Green and Sambrook, 2019). Similarly, rRNA can be depleted by rRNA hybridization to complementary biotinylated oligo probes followed by extraction with streptavidin-coated paramagnetic beads

(Kraus et al., 2019). The selection of a mRNA enrichment or rRNA depletion protocol depends on the aim of the study and other factors such as sample quantity and sample type.

cDNA synthesis

The conversion of RNA into cDNA is an essential step for RNA-seq. This conversion is necessary because DNA is biologically more stable than RNA. PCR amplification can only be done with DNA, and most sequencing protocols are designed for sequencing DNA. The first step in converting RNA to cDNA is the fragmentation of the RNA into an appropriate size for sequencing (i.e., 100–600 bp). Several approaches are available for RNA fragmentation, including physical approaches (e.g., acoustic shearing and sonication), chemical approaches (i.e., heating and divalent metal cation addition), and enzymatic methods (i.e., non-specific endonuclease cocktails and transposase tagmentation reactions) (Marine et al., 2011). The fragmented RNA is then converted to single-stranded cDNA using mRNA as the template, reverse transcriptase, and random primers or oligo (dT) primers (depending on the kit). The first strand of cDNA then can be used as a template for PCR. After this, double-stranded cDNA is produced by second-strand synthesis. The second strand cDNA synthesis is catalyzed by *Escherichia coli* DNA polymerase I combined with *E. coli* RNase H and *E. coli* DNA ligase. *E. coli* RNase H degrades the RNA to produce 3' -hydroxyl and 5' -phosphate terminated products, which are necessary for the DNA polymerase to function. The *E. coli* DNA polymerase I has two activities: the 5'-3' exonuclease activity removes RNA strands in the direction of synthesis, and in the meantime, it replaces RNA with deoxyribonucleotides. *E. coli* DNA ligase then joins the single strands into double-stranded cDNA.

Adapter ligation and cDNA amplification

Adapters are ligated to one or both ends of the cDNA fragment. Adapters consist of sequences that allow library fragments to bind to the flow cell, sequencing primer binding sites, and index sequences. Index/barcode sequences are sequence identifiers that enable the pooling of several samples (multiplexing) in a single sequencing run or flow cell lane. Products from the ligation reaction are purified using agarose gel electrophoresis prior to PCR amplification to create the cDNA library.

Library preparation kits

Several library preparation kits based on the Illumina platform are available. The "TruSeq Stranded Total RNA with Ribo-Zero Plant" kit is useful for large tissue samples (0.1–1 µg total RNA). While for low quantities of RNA, the "NEBNext® Ultra™ II Directional

RNA Library Prep with Sample Purification Beads" kit (10 ng–1 µg total RNA for polyA mRNA workflow and 5 ng–1 µg total RNA for rRNA depletion workflow) (New England Biolabs Inc., UK) can be used. These kits incorporate Illumina library preparation steps, including bead-based rRNA depletion or mRNA enrichment, cDNA synthesis, adding adaptors, indexing, and PCR. For a tissue sample that yields smaller amounts of RNA, like a single cell (1–25 ng), the "Collibri stranded RNA Library Prep kit" (ThermoFisher Scientific, USA) can be applied.

For the PacBio Iso-Seq platform for long-read RNA-seq, the "NEBNext Single Cell/ Low Input cDNA Synthesis & Amplification Module" kit (New England Biolabs Inc., UK) can be used for cDNA synthesis and its amplification from a single cell or ultra-low input RNA (as low as 1 pg–200 ng). The "SMRTbell Express Template Prep Kit 2.0" (Pacific Bioscience, USA) can be used to detect full-length transcripts up to 10 kb.

The ONT platform provides a starter pack for direct RNA-seq, PCR-cDNA sequencing kit, and direct cDNA sequencing kit (Oxford Nanopore Technologies Ltd., UK) with necessary inputs for RNA or Poly-A+(poly(A) on the present of the polyadenylated 3'-ends) 500 ng for direct RNA-seq, 1 ng for PCR-cDNA sequencing, and 100 ng for direct cDNA sequencing.

Quality control of library preparation

A very sensitive method for checking the quantity of a library preparation is with fluorometric methods (i.e., Qubitt™ Fluorometer, ThermoFisher Scientific, USA) or by qPCR. qPCR library quantification is based on the amplification of cDNA fragments with the adapters. The qPCR machine measures the intensity of fluorescence emitted by the probe at each cycle. In this approach, only templates that have both adapter sequences on either end will be measured and subsequently form clusters in a flow cell. Other methods include the use of electrophoresis-based quantification methods such as fragment analyzer systems that use automated parallel capillary electrophoresis to assess the library size distribution (e.g., Tapestation, Agilent Technologies, USA). A critical aspect in the quality check from the fragment analyzer is the library size distribution in the expected range. The peaks near the lower marker on library electrophoresis show contaminants, including primer and adapter dimers. An additional clean-up of the sample is recommended to increase the quality.

2.2.3 RNA sequencing

cDNA sequencing can be performed on several different platforms (see Rydmark et al. (2022)). Overall, RNA sequencing does not differ from the sequencing of genomic DNA. The sequencer reads cDNA fragments in one of two ways: using a single-end or paired-ends. In single-end reading, the sequencer reads the cDNA from the 3' or 5' end of only one strand of the insert. This method can produce large volumes of high-quality data especially for differential gene expression studies where an important factor is determining where the reads in transcripts come from (Stark et al., 2019). In the paired-ends reading, the sequencer reads two ends of a cDNA fragment and then combines the forward and reverse reads as reading pairs. Longer overlapping reads are advantageous for detecting splicing variants (Chhangawala et al., 2015). The sequencing length is between 100–600 bp, which will generate at least 20–100 million reads per sample. The requirements for sequence coverage and depth varies depending on the scientific questions to be answered, with complex studies perhaps needing greater sequencing depth and coverage. For example, a differential expression study using the Illumina platform requires 10-30 million reads per sample (Stark et al., 2019), while for a more in-depth transcriptome study or de novo transcriptome assembly, the recommended sequencing depth is 100 million reads per sample (Sims et al., 2014). To study isoforms, identify novel transcripts, and detect fusions, long-read sequencing is the appropriate choice. The ONT platform produces 1 to 12+ million reads, with the highest number of reads resulting from amplified cDNA input (7–12+ million reads). At the same time, the PacBio Iso-Seq platform produces up to 3 million full-length reads. Yet, long-read sequencing has some disadvantages compared to the short-read RNA-seq regarding throughput, number of reads, and error reads. In the end, the key in selecting the sequencing platform depends on your budget, research questions, the experimental strategies, technical aspects, data availability on the target organism, and the availability of bioinformatics pipelines.

2.3 BIOINFORMATICS

Prior to the development of high-throughput methods, individual transcriptome studies were performed using hybridization-based methods such as Northern blotting and microarrays (see above) or amplification-based methods including Sanger sequencing and RT-qPCR.

Hybridization-based methods require visual inspection or image processing analyses to interpret the output, while in qPCR, it is the amplification that must be monitored. In

qPCR, the expression levels are represented by cycle threshold (Ct) values and further normalization steps and statistical analyses need to be used for the estimation of relative or absolute abundances. Neither hybridization methods nor qPCR require labor-intensive post-processing.

On the other hand, EST/SAGE/MPSS or RNA-seq methods rely on sequence data and require several post-processing steps such as clustering, assembly, and functional annotation. As RNA-seq allows characterization of whole transcriptomes and currently is the most widely used method, we outline the bioinformatic analysis steps for high-throughput RNA-seq data. Long read sequencing methods such as ONT and SMRT allow full-length characterization of transcripts and can be used to study complex transcriptomes. Although one common concern regarding these technologies is high error rates, their accuracy has dramatically increased recently and the development of long-read specific error correction approaches are providing further improvements (Amarasinghe et al., 2020). These technologies have already been used for e.g. isoform identification (Wang et al., 2017) and long noncoding RNA (lncRNA) discovery (S. Li et al., 2016). Preprocessing of the reads obtained from these platforms requires specific tools, while common short-read analysis tools can be used for downstream analyses (e.g., differential expression) after reconstruction of transcripts. Most long-read isoform detection tools cluster aligned and error-corrected reads and collapse these into isoforms (Amarasinghe et al., 2020). PacBio provides an open source bioinformatics suite "SMRT Analysis" that includes tools for classification of reads, clustering, polishing, alignment, and visualization of isoforms (Oikonomopoulos et al., 2020). Several tools can be used to align and visualize polished Iso-Seq reads such as MiniMap2 and Iso-Seq Browser (Hu et al., 2017; Li et al., 2017). Similar to PacBio, ONT reads can be analyzed with publicly available software and the same tools can be used for these steps.

2.3.1 Quality control and pre-processing

After obtaining raw RNA-seq data, the quality of the reads should be checked and sequencing errors should be corrected in order to improve the accuracy and efficiency of the assembly process. It is also recommended to mask low complexity regions and repetitive sequences that might generate hits that are artefacts. DUST and SEG modules of BLAST can be used for this purpose on nucleotide and amino acid sequences, respectively. Bacterial and viral contaminants can be removed by running similarity searches against public databases or using tools such as DeconSeq (Schmieder and Edwards, 2011a). rRNA sequences can be removed by mapping the reads to an rRNA database (e.g., SILVA, https://www.arb-silva.de) using tools such as bowtie2 (Langmead and Salzberg, 2012). FastQC

(Andrews 2010) performs an initial assessment of raw high-throughput sequence reads and reports quality metrics that might be useful to determine issues with library preparation or sequencing protocols. There are several other tools that can perform quality control and/or remove artefacts such as sequencing adapters, including Fastx-toolkit (Gordon and Hannon, 2009), Prinseq (Schmieder and Edwards, 2011b), and Trimmomatic (Bolger et al., 2014).

Most short-read assemblers first divide reads into subsequences of length k (i.e., k-mers) and generate a graph representing the overlap between them (Compeau et al., 2011; Heydari et al., 2017). Sequencing errors introduce problematic k-mers into this process. Several tools with k-mer-based error correction strategies have been developed to overcome this challenge such as Fiona (Schulz et al., 2014) and BFC (Li, 2015). Although these tools were originally developed for genomic data, they can correct RNA-seq reads as well. Rcorrector is another tool developed specifically for correcting Illumina RNA-seq reads (Song and Florea, 2015). While k-mer-based error correction methods additionally remove reads that originate from rare transcripts, shallow sequencing depth typically does not give accurate assemblies of those transcripts regardless (Martin and Wang, 2011).

2.3.2 Transcriptome assembly

Depending on whether a reference genome/transcriptome is available or not, there are different strategies for transcriptome assembly (Fig. 2.1).

De-novo assembly

De-novo assembly is solely based on RNA-seq data and uses the k-mer composition by subdividing the reads into shorter segments of a given length k. This composition and the overlaps between these k-mers are represented on a de Bruijn graph, which is finally resolved to reconstruct transcripts (Pevzner et al., 2001). In general, de-novo assembly requires higher sequencing depths compared to reference-guided assembly; around 30X coverage is needed to reconstruct full length transcripts de-novo, while the same task can be achieved at 10X coverage with a reference (Martin and Wang, 2011).

Commonly used de-novo assemblers include Trans-ABySS (Robertson et al., 2010), Trinity (Grabherr et al., 2011), and Oases (Schulz et al., 2012). Trinity is developed specifically for de-novo transcriptome assembly and is widely used (Kerr et al., 2019). Most of the other available tools are simply extensions of de-novo genome assembly tools. Trinity relies on a single k-mer length, while other assemblers can use multiple values of k-mer length (Hölzer and Marz, 2019). The choice of the k-mer length can affect the quality of an assembly drastically. The optimal value depends on several factors such as read

length, sequencing depth, error rate, and complexity of the target species transcriptome (Góngora-Castillo and Buell, 2013). At shorter lengths the possibility of overlap between k-mers is higher, while at longer lengths there are fewer overlaps and reconstruction of transcripts can be comparatively easier. It is also easier to resolve repetitive regions at longer lengths, while the assembly of transcripts with low expression levels becomes more challenging. It should also be noted that memory requirements increase significantly at longer k-mer lengths. In a case from allopolyploid plants, for example, a k-mer length of 41 produced the highest number of full length transcripts, and researchers suggested to consider a broad range of k-mer lengths and coverages for avoiding chimeric assemblies of homologous and paralogous gene copies in polyploid taxa (Gruenheit et al., 2012). Some assemblers, such as Trans-ABySS, post-process an assembly to merge contigs and identify isoforms, while other assemblers, such as Trinity, directly use the de Bruijn graph to assemble each isoform (Martin and Wang, 2011). Shannon (Kannan et al., 2016) uses a different approach by analyzing read abundance information together with the de Bruijn graph in order to resolve complex isoforms and paralogues. After assembly, the longest isoform can be selected as a single representative transcript to simplify the downstream steps (Góngora-Castillo et al., 2012).

There are also combined de-novo assembly approaches such as EvidentialGene (Gilbert, 2016) and Oyster River Protocol (MacManes, 2018). These tools aim to improve the completeness and accuracy of the assembly by providing an optimized consensus approach that combines several k-mer lengths and transcripts coming from different assemblers.

Reference-guided assembly

Genome-guided assemblers map RNA-seq data to a reference genome and avoid constructing de Bruijn graphs by merging the reads based on their overlapping regions. The quality of the reference genome is critical here, as a high-quality assembly can provide accurate transcript predictions and expression profiles, while using a fragmented or incomplete assembly as reference might aggravate this process. When mapping RNA-seq reads to a reference genome, introns should be accounted for. Therefore genome-guided assemblers allow splitting the reads during mapping. This is achieved by using a splice aware alignment strategy where the downstream regions of a read can map to a downstream exon on the reference. Such splice aware aligners include TopHat2 (Kim et al., 2013) and STAR (Dobin et al., 2013). After mapping, the reads are merged to reconstructed transcripts and isoforms using genome-guided assemblers such as Cufflinks (Trapnell et al., 2012) and StringTie (Pertea et al., 2015). All these tools create a graph representing splice junctions to merge the mapped reads, but they produce different results depending

on their different approaches to transcript reconstruction (Hsieh et al., 2019; Voshall and Moriyama, 2018). De-novo assemblers can also be used to reconstruct the transcripts at each locus after mapping with splice aware aligners. Trinity offers a genome-guided assembly option using this approach as well. Another genome-guided assembler, RefShannon (Mao et al., 2020), exploits abundance data to reconstruct transcripts with a similar approach to de-novo assembler Shannon.

RNA-seq reads can also be mapped to a transcriptome, if a high-quality assembly is available for the target or a closely related species. This transcriptome-guided approach can improve the contiguity and completeness of the assembly (Garber et al., 2011; Ungaro et al., 2017). Aligners, such as bowtie2 and BWA (Li and Durbin, 2009) can be used in this approach, however, it is not possible to identify splicing events in new junctions when using a transcriptome as reference (Garber et al., 2011).

Combined approach

High-sensitivity reference-guided assemblers can be combined with de-novo assemblers in order to detect novel and missing transcripts as well. If the reference genome is incomplete, fragmented, or from a distantly related species, the de-novo assembly should be conducted first in order to prevent the potential errors in the reference. This approach can also be useful for extending incomplete transcripts to full-length by merging these based on a reference (Martin and Wang, 2011). On the other hand, if a good quality reference genome is available, the combined approach should start by aligning the reads to a reference, followed by de-novo assembly of the reads that cannot be mapped. Thus, this method can also be used to filter out unwanted sequences before de-novo assembly.

2.3.3 Assembly quality, annotation, and quantification

The average length of assembled contigs in an RNA-seq experiment will vary based on the actual mRNA fragments that are sequenced. Thus, metrics based on assembled contigs do not necessarily indicate the quality of a transcriptome assembly. Transcriptome-specific metrics have been suggested such as ExN50, which computes transcript lengths as expression-weighted means of isoform lengths. Another method to assess the assembly quality is by checking the read percentage that can concordantly align to the final assembly in order to understand if the full complement of paired-end reads are represented in the assembled transcripts. Tools such as bowtie2 or BWA can be used for this type of mapping. Other tools for evaluating the quality of an assembled transcriptome include DETONATE (Li et al., 2014) and TransRate (Smith-Unna et al., 2016).

Transcripts can also be translated into protein sequences and mapped against well annotated databases such as UniProt/Swiss-Prot, Pfam, or NCBI. If the sequenced organism is closely related to a model organism, a high proportion of the contigs should have potential homologs in these databases. Another tool, BUSCO, assesses the completeness of the assembly by comparing it with universal single-copy gene databases specific to different lineages such as bacteria, fungi, or plants.

Expression quantification is a critical step for most RNA-seq experiments. There are two main sources of systematic variability which might introduce errors to this process; (i) longer transcripts generate more reads than shorter transcripts at the same abundance due to RNA fragmentation during library construction, and (ii) the number of fragments mapped across samples are different due to varying number of reads produced for each run. Therefore, read counts need to be normalized in order to obtain accurate gene expression estimates. Inter-sample normalization methods have been developed for differential expression analysis, such as DeSeq2 (Anders and Huber, 2010) and trimmed mean of M-values (TMM) which is implemented in edgeR (Robinson et al., 2010). For next generation sequencing research, reads per kilobase per million mapped reads (RPKM) is the most widely used method. It also accounts for gene lengths and is implemented in Salmon (Patro et al., 2017). Methods based on machine learning algorithms such as RSEM (Li et al., 2014) and Sailfish (Patro et al., 2014) can consider additional variables such as library size. Reference-based quantification approaches need to map the assembled transcripts to the reference genome first and then quantify the annotated genes. Functional annotations of the top expressed transcripts can be quickly examined at this step to check if tissue specific genes are abundantly expressed (e.g., genes known to be important for photosynthesis in leaf tissue).

Assembled transcripts from de-novo or reference-guided assemblies are expected to represent real biological differences such as expression levels, alternative splice forms, and paralogous or allelic transcripts (Schliesky et al., 2012). However, assembling plant transcriptomes with short reads can be more challenging compared to bacteria or lower eukaryotes, due to factors such as polyploidy, diversity in alternative splice isoforms and the heterozygosity of alleles (Góngora-Castillo and Buell, 2013; Martin and Wang, 2011). Thus, experimental strategies and bioinformatics pipelines should be developed specifically for each individual study and take the target organism and research questions into consideration.

2.4 APPLICATIONS OF TRANSCRIPTOMICS IN SPECIES IDENTIFICATION

2.4.1 Marker discovery for phylogenomic inference

Transcriptomes have been used for plant phylogenomic inference as they contain abundant information from the nuclear genome. Famously, the generation of > 1000 transcriptomes across the plant kingdom led to new evolutionary insights for land plants (One Thousand Plant Transcriptomes Initiative, 2019). However, the application of RNA-seq is limited to fresh tissue with low levels of degradation, making it less applicable to studies with large taxonomic sampling.

An emerging phylogenomic approach that partly relies on transcriptomics uses targeted next-generation sequencing (Woudstra et al., 2022) to obtain specific genes for high-coverage DNA sequencing in large numbers of samples with varying taxonomic breadth. Target capture is very efficient in recovering hundreds of genes, regardless of the degradation level in the source DNA (Brewer et al., 2019; Hart et al., 2016) making this technique ideally suited for studies of plant systematics. Transcriptomes are the most commonly used source for nuclear marker discovery (Chamala et al., 2015), particularly when no reference genome sequence is available (Woudstra et al., 2022). They can be compared against curated databases of low-copy nuclear genes (De Smet et al., 2013) and form the basis for designing the short oligonucleotide 'baits' used to capture the target genes.

2.4.2 Metatranscriptomics

Metatranscriptomics is the application of transcriptome sequencing to environmental samples such as water, soil, or sediments. It gives an overview of the actual metabolic activity and taxonomic diversity within a community. The protocol involves HTS of reverse-transcribed cDNA obtained from an environmental mRNA isolate. While reverse transcriptase PCRs can only detect a single gene at a time, metatranscriptomics gives a whole gene expression profile of a diverse community of organisms playing various functional roles in the ecosystem (Carvalhais et al., 2012; Mason et al., 2012). Coupling these analyses with taxonomically informative rRNA offers the possibility to gather information on the community composition as well.

Some of the main challenges of metatranscriptomics are the presence of PCR inhibitors in environmental samples (e.g., humic acid, polysaccharides; (Crump et al., 2018) and the low fraction of mRNA in the total RNA isolates (less than 5%) (Creer et al., 2016).

An additional PCR step is often used to increase the total amount of genetic material. However, this might result in biased detection of diversity and quantification estimates in downstream steps (Porter and Hajibabaei, 2018). Another challenge is to assign mRNA sequences to a specific function, as existing databases contain the most abundant genes in a limited number of environmental samples, or the genes from cultured species representing a limited proportion of environmental diversity (Prosser, 2015).

There are various applications of metatranscriptomics such as revealing the composition of freshwater bacterioplankton communities (Poretsky et al., 2005) and animal/plant microbiomes (Crump et al., 2018; Pérez-Losada et al., 2015), understanding the dynamics of cyanobacterial blooms and viral-host relationships (Berg et al., 2018; Moniruzzaman et al., 2017; Shi et al., 2009), identifying important biochemical pathways (Franzosa et al., 2014; Saminathan et al., 2018), and understanding the mechanisms behind infection and disease (Hayden et al., 2018).

2.5 CONCLUSION AND FUTURE PERSPECTIVES

Plant transcriptomics studies have undergone huge advances over the past few years as the costs of the second generation of sequencing, such as Illumina, have declined, third generation sequencing has become more accurate, and a wider range of analysis tools and pipelines have become available and become more accurate (Heather and Chain, 2016). It is now possible to sequence large numbers of genes, or even whole genomes, for phylogenomic analyses (Soltis and Soltis, 2020). Transcriptomics studies are an integral part of plant research and it's widely used for marker identification, phylogenetic inference, species diversification, genetic response to abiotic and biotic stresses, evolution and development, metatranscriptomics to reveal the relation between plant and microbiome etc

Studies using comparative transcriptomics to understand interactions between different organisms (Hayden et al., 2018), as well RNA-seq for single-cell work in particular are at the forefront of transcriptomic applications in functional studies, and open up the possibility of looking into the complex network of gene regulation, with significant implications in both fundamental science as well as in more applied fields such as crop development (Rich-Griffin et al., 2020).

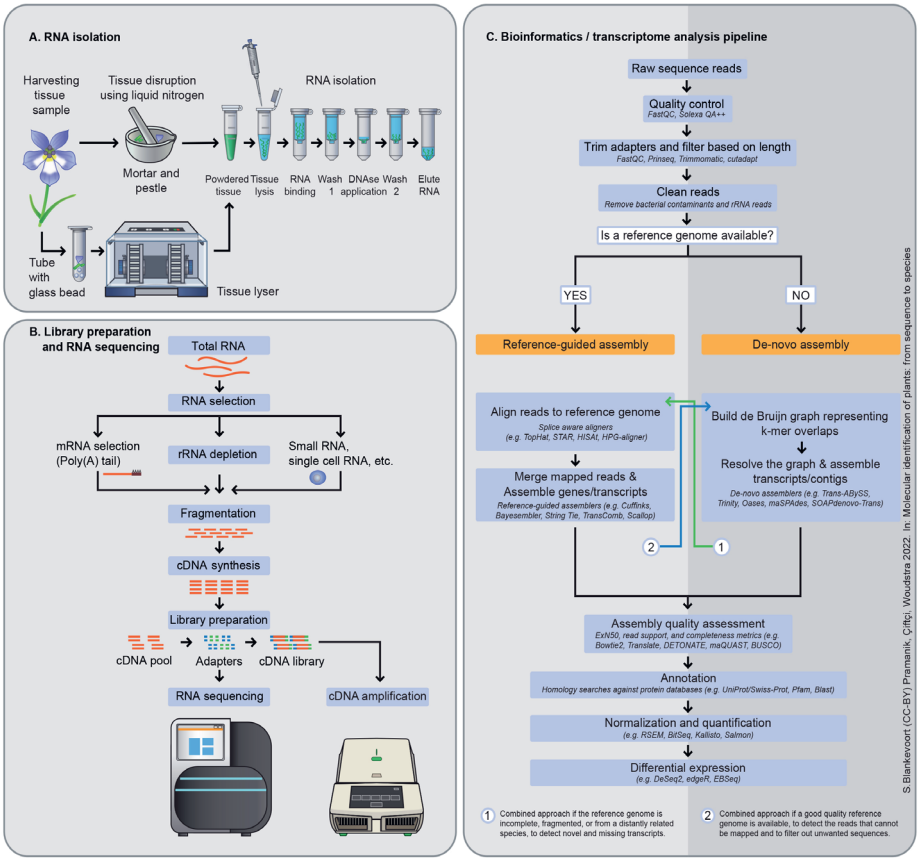

Figure 2.1. Overview scheme of transcriptomics in plants with emphasis on the RNA-seq method. (**A**) Sample preparation and RNA isolation. (**B**) Library preparation starts by selecting the target RNA from the total RNA, followed by fragmentation of the RNA sequence, cDNA synthesis, adapter ligation, cDNA amplification, and RNA sequencing. (**C**) The first step in transcriptome analysis is assessing the quality and quantity of reads. The clean reads are assembled to the reference genome or through a de novo assembly or by combining these two approaches. The assemble reads are then annotated, followed by quantification and normalization of the annotated results. The final step is differential expression analysis to quantify the difference in the expression level of genes between the samples or treatments.