2 Modern Approaches for Transcriptome Analyses in Plants

Diego Mauricio Riaño-Pachón, Hector Fabio Espitia-Navarro, John Jaime Riascos, and Gabriel Rodrigues Alves Margarido

Abstract

The collection of all transcripts in a cell, a tissue, or an organism is called the transcriptome, or meta-transcriptome when dealing with the transcripts of a community of different organisms. Nowadays, we have a vast array of technologies that allow us to assess the (meta-)transcriptome regarding its composition (which transcripts are produced) and the abundance of its components (what are the expression levels of each transcript), and we can do this across several samples, conditions, and time-points, at costs that are decreasing year after year, allowing experimental designs with ever-increasing complexity. Here we will present the current state of the art regarding the technologies that can be applied to the study of plant transcriptomes and their applications, including differential gene expression and coexpression analyses, identification of sequence polymorphisms, the application of machine learning for the identification of alternative splicing and ncRNAs, and the ranking of candidate genes for downstream studies. We continue with a collection of examples of these approaches in a diverse array of plant species to generate gene/transcript catalogs/atlases, population mapping, identification of genes related to stress phenotypes, and phylogenomics. We finalize the chapter with some of our ideas about the future of this dynamic field in plant physiology.

Keywords

RNA-Seq · Crops · Transcription · Gene expression · Polyploidy · Next-generation sequencing · Long reads · Short reads · Assembly

D. M. Riaño-Pachón (✉)
Laboratory of Computational, Evolutionary and Systems Biology, Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil
e-mail: diego.riano@cena.usp.br

H. F. Espitia-Navarro
School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: hspitia@gatech.edu

J. J. Riascos
Centro de Investigación de la Caña de Azúcar de Colombia, CENICAÑA, Cali, Valle del Cauca, Colombia
e-mail: jjriascos@cenicana.org

G. R. A. Margarido
Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, Brazil
e-mail: gramarga@usp.br

F. V. Winck (ed.), *Advances in Plant Omics and Systems Biology Approaches*, Advances in Experimental Medicine and Biology 1346, https://doi.org/10.1007/978-3-030-80352-0_2

2.1 Introduction

The transcriptome is the collection of all RNA molecules found at a given time in an organism, in a tissue, or in a cell. Researchers today can study the full transcriptome, or a targeted transcriptome (a defined subset of transcripts under a certain condition) using an array of different technologies, like microarrays, reverse transcription quantitative PCR (RT-qPCR), and nucleic acid sequencing. In most approaches, the population of RNA molecules should be first converted into the more stable cDNA, but recent advances and the development of new sequencing platforms are allowing the direct sequencing of the RNA molecules, removing biases that could be introduced by the synthesis of cDNA (Garalde et al. 2018; Keller et al. 2018). Assessing the transcriptome offers an overview of the functional component of a genome and of the genes that must be active in order to achieve a given transcriptional state. Transcriptomics studies have been employed to develop catalogs of expressed sequences, by the identification of mRNAs, small-RNAs (e.g., miRNA, snoRNAs), long-non-coding RNAs (lncRNAs) among others. Also, to aid in the annotation of newly sequenced genomes, improving the inference and definition of gene structure, like start and end sites of the transcription, position of introns and exons, and alternative splicing patterns. Perhaps the most prevalent use of transcriptomics is the quantification of gene expression levels under different conditions aiming at revealing the molecular mechanisms underlying the establishment of phenotypes and responses to stresses. Transcriptomics is increasingly being used to infer the function of genes, by exploiting co-expression, under the assumption of "guilt-by-association," and for the identification of coordinated expression modules. The rapidly decreasing costs and wide availability of the diverse transcriptomics technologies are allowing studies in diverse groups of plants and addressing evolutionary questions about the evolution of expression patterns, gene expression and regulation networks, at a scale without precedent.

The earliest approaches that can be called transcriptomics studies relied on sequencing expressed sequence tags (ESTs) using the low-throughput Sanger chain-termination sequencing technology and started in the 1980s (see Fig. 2.1). EST sequencing projects were expensive and laborious but allowed assessing the functional fraction of a genome sequence at a fraction of the effort and cost. The wealth of sequence information generated in these projects could be leveraged with the development of array-based hybridization technologies (macroarrays used nylon membranes and microarrays used glass slides), which offered higher throughput and had lower application costs than EST projects, once the development of the membranes/slides had been deduced. The first use of the words microarray or macroarray in the scientific literature dates back to 1996, but their use really takes off in the 2000s (Fig. 2.1). The use of ESTs and array-based technologies was superseded by high-throughput sequencing-based methods, first exploiting small transcript signatures (tags) and later the sequencing of complete or close to complete transcripts.

In this chapter, we will introduce you to the basics of transcriptome studies, applications, and some examples in non-model plants.

2.2 Transcriptomics Approaches

2.2.1 Array-Based Approaches

Large-scale characterization of transcriptomes was made possible with the use of microarrays. In this technology, an array of oligonucleotide probes that are complementary to known transcripts is immobilized on a glass slide. Next, cDNA molecules synthesized from RNA are hybridized with the probes, and signal intensities are assessed to provide a measure of transcript abundance. This provides an economical way of analyzing transcriptomes on a genome-wide scale. Microarrays are used nowadays for model species and economically important crops, primarily due to low cost and laboratory routine.

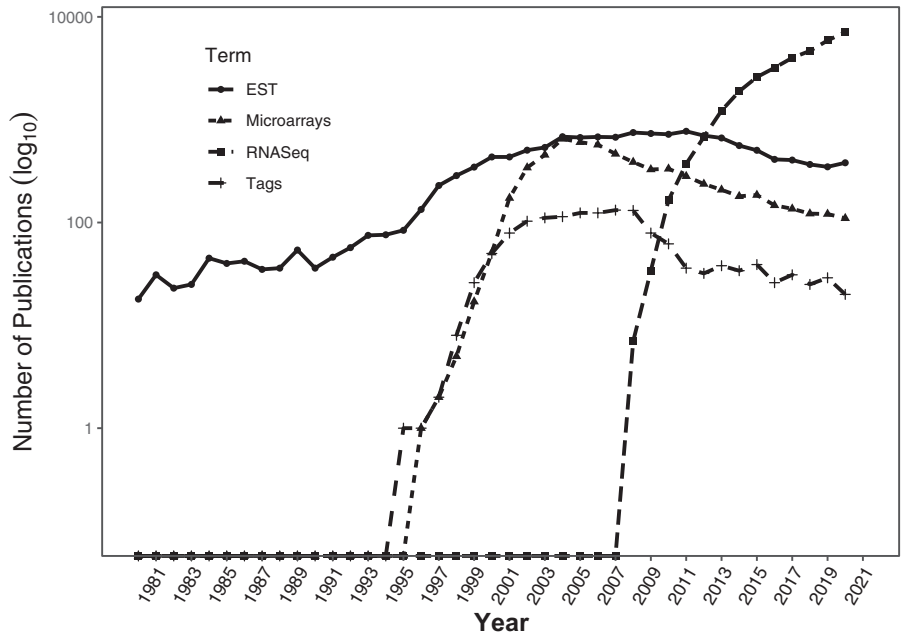

Fig. 2.1 Number of publications in the last four decades for different transcriptome technologies

However, this approach presents a number of disadvantages that have relevant practical implications. First, previous knowledge about the transcripts of interest is required for designing the array chip, which hinders application for non-model species. This may introduce bias toward the specific sequences used to obtain the probes, which is particularly important for genes with multiple isoforms. Second, transcript abundance estimation is not accurate for lowly expressed genes, owing to background noise from nonspecific hybridization, or for very highly expressed genes, due to probe saturation. The dynamic range of detection is thus limited. Third, cross-hybridization of transcripts with similar sequence can adversely affect expression estimates. Finally, intrinsic differences in hybridization exist between probes because of their sequence content (Marioni et al. 2008; Wang et al. 2009; Zhao et al. 2014).

Sequencing-based approaches resolve many of these issues and are now the method of choice for large-scale transcriptome profiling in a variety of scenarios. From now on, we will focus on these more recent strategies.

2.2.2 Sequencing-Based Approaches

In-depth knowledge and understanding of a plant genome, or any organism for that matter, involves the elaboration of a catalog of the genes present in the genome and information about the expression levels of the transcripts derived from these genes under a wide array of conditions. In both cases, one requires sequence data.

The most widely used technology in early genome projects was Expressed Sequence Tag (EST) sequencing (reviewed by Parkinson and Blaxter 2009). EST sequencing was employed to generate gene catalogs, both in model plants (Delseny et al. 1997; Weng et al. 2005; Asamizu et al. 1999; Banks et al. 2011) and in crops (e.g., Yamamoto and Sasaki 1997; Vettore et al. 2003; Ma et al. 2004; Pavy et al. 2005). In many cases, ESTs also served as a basis for the development of cDNA microarrays to query gene expression under different plant conditions or developmental stages (Lembke et al. 2012; Pavy et al. 2008). In most projects, ESTs were derived from normal-

ized libraries, which meant that all transcripts have approximately the same probability to be sequenced. This readily reduces costs for gene discovery, but the gene expression levels and the dynamics of transcription regulation cannot be assessed.

With the creation and advance of high-throughput sequencing (HTS) technologies toward the end of the 1990s and in the early 2000s, new approaches were applied to discover plant genes and transcripts and to assess the dynamics of transcription, and its regulation, like alternative transcription starting sites (TSS) and alternative splicing form usage. Among these approaches, one could mention Cap Analysis of Gene Expression—CAGE (de Hoon and Hayashizaki 2008) and Serial Analysis of Gene Expression—SAGE (Velculescu et al. 1995; Matsumura et al. 2005), to name just a few, which are collectively known as tag sequencing approaches (Harbers and Carninci 2005) (see "Tags" in Fig. 2.1). These technologies started by exploiting the traditional Sanger DNA sequencing method to assess transcription, but moved soon to exploit the newer, highly parallel and HTS technologies, and thus gained suffixes like –deep or –seq and prefixes like ultra–, to differentiate them from their older lower throughput versions. Briefly, tag sequencing approaches aim to generate short sequence tags from the transcript ends, either the 5′ or the 3′ end. These short tags should unequivocally identify each transcript or genomic region, although it was not uncommon that a single tag could be mapped to more than one transcript/gene, particularly in cases of large gene families which are common in plants. In addition, the number of tags sequenced for each transcript is directly related to the transcript abundance in the original sample. Being based on short sequence tags from the transcript ends, these approaches were better suited for organisms whose genomes were already sequenced.

On the one hand, one of the main advantages of either EST or tag-sequencing approaches is the generation of a digital measure of gene expression, the number, or count, of a certain event, i.e., the sequencing of a complete, or part of a, RNA molecule. In contrast to an analogous measure, such as that offered by cDNA microarrays which is subject to probe saturation and thus has a low dynamic range, this digital measure is not saturated in the case of highly abundant transcripts. For the case of lowly expressed transcripts, the trivial alternative is to continue counting events until a certain number of rare events (lowly expressed transcripts) have been achieved, although this could have an important impact on the overall cost of the experiment. If lowly expressed transcripts are the focus of the study, then alternative approaches can be employed, such as targeted sequencing and reverse transcription quantitative PCR (RT-qPCR). On the other hand, the main drawback of both approaches (ESTs and tag-sequencing) is that neither of them provides the full representation of the underlying transcripts. Additionally, tag-sequencing and microarray approaches require preexisting knowledge about the transcript space of the species of interest, which impose serious limitations to its application in non-model organisms.

2.2.2.1 RNA-Seq

The sequencing of transcriptomes employing HTS technologies, without focus on any particular region of the mRNA, in contrast to CAGE or SAGE, is known as RNA-Seq. The first publications using the word RNA-Seq appeared between 2006 and 2008 applied to few organisms (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Bainbridge et al. 2006; Wilhelm et al. 2008; Cloonan et al. 2008), also including *Arabidopsis thaliana*, the model land plant (Lister et al. 2008) (see "RNA-Seq" in Fig. 2.1).

The synthesis and maturation of transcripts is a finely regulated process that allows the plant cell to produce the required gene products in the proper quantities and at the proper times and places. Within a single experiment, RNA-Seq allows the discovery of expression levels, splicing events (Marquez et al. 2012; Shang et al. 2017; Brown et al. 2017), RNA editing (Hackett and Lu 2017), and mutations (Peng et al. 2016; Serin et al. 2017). RNA-Seq paves the way for the understanding of the rules governing RNA

regulation and the underlying regulatory networks, thus generating new insights on plant development and the response to biotic and abiotic (Imadi et al. 2015) stresses at the cellular and molecular levels.

The main steps in any RNA-Seq project are (1) sample preparation, (2) library preparation and (3) sample sequencing.

(1) Sample preparation consists on the isolation of RNA from the biological samples of interest. Plant cells have different types of RNA molecules, like messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and other types of non-coding RNA (ncRNA). Over 95% of the transcript population in a cell consists of rRNA and tRNA species (Rosenow et al. 2001). Thus, to assess, via HTS technologies, the other transcript species, samples must be processed in special ways. For instance, if the objective of the project is to assess mRNAs transcribed by RNA pol II (which are mostly genes that will eventually undergo translation), one can exploit the fact that these eukaryotic mRNAs are polyadenylated, by fishing for these transcripts using poly-dT oligonucleotides, effectively excluding the large fraction of rRNA and other ncRNAs. On the other hand, if one is interested in evaluating the whole transcriptome (mRNA + all types of ncRNAs, only excluding rRNA), then there are approaches to specifically remove rRNA from the sample, usually employing hybridization techniques, methods that are usually referred to as ribo-depletion (O'Neil et al. 2013). Additionally, the goal of the study could be to focus on small ncRNAs, in that case one would perform a size fractionation and selection step.

As part of (2) library preparation, for short-read HTS technologies (see below for long-read HTS technologies), the isolated RNA must be converted into double-stranded cDNA and fragmented. Fragments should be ligated to adapters to allow amplification and sequencing. At this point, it is important to remember that a given message in the genome is encoded in one of the two strands of the DNA double helix, and thus it is important in most cases to keep the information of which strand was transcribed. In general, one can divide the library preparation methods in two groups, those that keep the strand information (strand-specific protocols) and those that do not (often called unstranded protocols). Today, most RNA-Seq datasets are still being generated using library preparation protocols that do not keep the strand information. For instance, from 219,832 green plant datasets using RNA as source in RNA-Seq experiments in the Short Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/; July 2020), only 5995 have 'strand-specific' in their description.

(3) Sample sequencing is carried out in massively parallel sequencing instruments, paying attention to the dependence between library preparation method and sequencing instrument. The most widely available technologies for RNA-Seq are those released by Illumina Inc, i.e., using reversible-terminators sequencing-by-synthesis technology (Bentley et al. 2008; Illumina 2010), within their sequencing instruments MiSeq, HiSeq, NextSeq, or NovaSeq. Samples prepared with Illumina library construction methods are compatible with any of their instruments, the only difference being on the throughput obtained, e.g., number of sequenced fragments and number of samples that can be analyzed simultaneously.

Before you start your RNA-Seq project, you must develop the experimental design that will allow you to answer biologically relevant questions with a predefined level of certainty. Here we will only highlight two factors among the many that must be taken into account during the experimental design phase: (1) number of biological replicates and (2) number of sequenced fragments per sample. The number of replicates depends on your final goal. On the one hand, if your goal is to make a catalog of genes present in an organism's genome, typical when sequencing a new genome and preparing for annotating it, then preparing a single, or few, library from a pool of tissues and/or conditions might be enough. On the other hand, if you plan to evaluate the statistically significant differences in gene expression values between different conditions, then a higher number of replicates is required. Depending on the size of the effects that are desired to be detected, if only changes around two to threefold are sought, then a number of bio-

logical replicates around five should suffice in most cases; a higher number of replicates would be required to detect smaller changes in expression values (Schurch et al. 2016). Regarding the number of sequenced fragments, you should keep in mind that RNA-Seq is basically a random sampling process. If your goal is to assess statistical differences among conditions, you must check whether your sampling is deep enough to support your conclusions. A few approaches have been proposed to check for this, all of them are based on resampling your reads, and counting a feature of interest for each subsample for increasingly large subsamples. If the sequencing depth is high enough, you would expect that the number of a given feature is close to saturation with increasing number of resampled reads. There are a few approaches to achieve this. First you could count the number of transcripts that are detected at different fractions of the original datasets, e.g., 5%, 10%, 20, of the original reads; if sampling is deep enough, you would expect to find a plateau (Garcia-Ortega and Martinez 2015). Similarly, instead of looking at the number of transcripts, you can look at the number of exon–exon junctions detected with increasingly large samples of the reads; again you expect to achieve a plateau if your sequencing depth was saturated. This can be achieved with the junction-saturation.py script part of RSeQC (Wang et al. 2012). It is important to note that, despite sequencing depth being important, especially for lowly expressed genes, the number of biological replicates is much more important, and if you have to choose between more depth or more biological replicates, you should always choose the latter (Liu et al. 2014; Lamarre et al. 2018; Baccarella et al. 2018).

Regarding the sequencing depth, it is important to keep in mind that under several conditions, a large fraction of the reads would originate from one or a few transcripts. For instance, when doing sequencing of total RNA, you will have a large fraction of sequencing reads originating from rRNA transcripts, which can be up to 90% of the total RNA in the cell (Conesa et al. 2016). In these cases, you should try to deplete your sample from rRNA transcripts, for which several options are available in the market (Conesa et al. 2016; Hrdlickova et al. 2017; NuGen n.d.; siTOOLsBiotech 2018). However, not only rRNA transcripts exhibit such high abundance. A recent study of the *A. thaliana* transcriptome identified over 4000 ubiquitously and extremely highly expressed transcripts (Sun et al. 2014). If your specific project aims at assessing the expression of lowly expressed and rare transcripts, it might be important to deplete these ubiquitous and highly expressed transcripts, for such case, some alternatives for library preparation are available, as the AnyDeplete or riboPools technologies (NuGen n.d.; siTOOLsBiotech 2018).

2.2.2.2 Strand-Specific RNA-Seq

The existence of overlapping genes (genes whose transcripts are encoded—completely, or most frequently partially—in opposite strands of the same genomic region) in plants has been known for some years (Quesada et al. 1999; Xiao et al. 2005). Natural antisense transcripts (NATs) are RNA molecules that can have regions of sequence complementary to other RNAs and that can regulate the expression level of their target genes. Particularly, cis-NATs are pairs of transcripts that overlap on the genome. Disambiguating the expression levels of the two overlapping transcripts requires data that keep the information about which strand was transcribed (see for example, Britto-Kido Sde et al. 2013; Li et al. 2013a; Jin et al. 2008; Riano-Pachon et al. 2016). Between 7% and 8% of genes in rice (Osato et al. 2003) and *Arabidopsis* (Wang et al. 2005; Jen et al. 2005), respectively, are cis-NATs, recent studies suggesting even higher rates of cis-NATs (Oono et al. 2017; Zhao et al. 2018). Figure 2.2 illustrates the importance to have strand information for transcriptome analyses.

Currently, three technologies are widely available that can maintain strand information: Illumina's TruSeq Stranded library preparation kits, Pacific Biosciences's IsoSeq, and Oxford Nanopore Technologies's direct rRNA sequencing. Perhaps the most pervasive of the three in the market is the one commercialized by Illumina in their TruSeq Stranded library preparation kits, which use the deoxy-UTP strand-marking strategy. The Illumina instruments are capable of

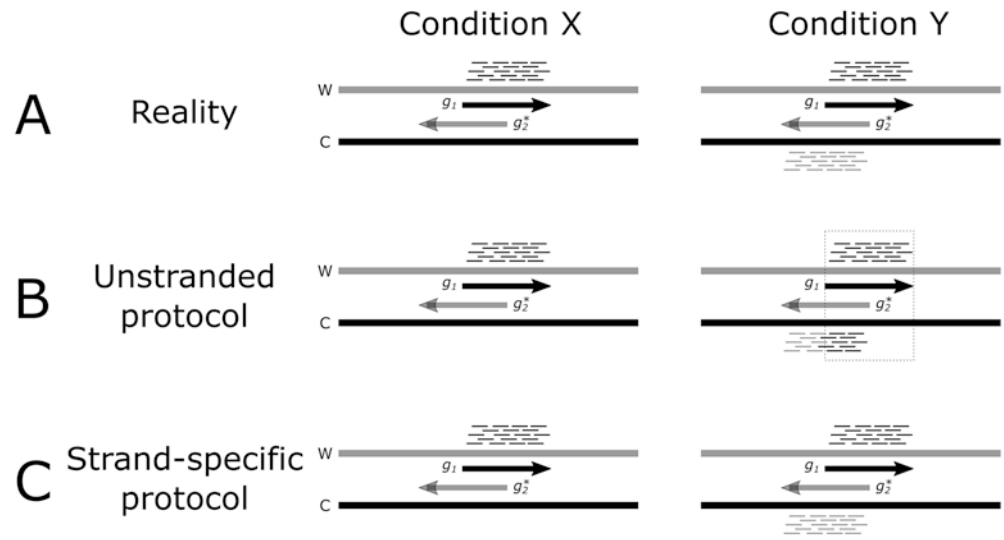

Fig. 2.2 Use of strand-specific information to disambiguate the expression of overlapping genes. Two overlapping genes g_1 in the Watson strand and g_2 in the Crick strand shown in two different experimental conditions, *X* and *Y*. The symbol * indicates that g_2 is an unknown (unannotated gene). Short sequencing reads appear either above or below the DNA strands as short line, each line representing a sequencing fragment. (**a**). The real case: g_1 is expressed in both conditions *X* and *Y*, with similar or identical abundances, while g_2 is only expressed in condition *Y*. (**b**) Sequencing results obtained with a protocol that ignores (or loses during library preparation) the information about which strand generated the reads. Only reads that overlap with annotated features are counted (dashed line in condition *Y*). In condition *Y*, many of the reads originated from the gene g_2 will counted as if they were from gene g_1 (reads shown in black). This will lead to the wrong conclusion that the expression of g_1 in condition *Y* is higher than in condition *X*. (**c**) Using a protocol that keeps strand information (strand-specific), in condition *Y* only the reads in black will be assigned to g_1, and the additional reads in gray will hint toward the existence of an additional gene in the same locus that is only expressed in condition *Y*. The abundances of g_1 in condition *X* and *Y* will be similar and will not lead to a differential expression call, as in (**b**)

sequencing double-stranded DNA molecules (dsDNA), but not single-stranded RNA molecules (ssRNA), so transcript sequences, which are made of ssRNA, must be transformed into dsDNA molecules by a process called cDNA synthesis. Briefly, the RNA molecules are fragmented, and each resulting fragment will be used for the synthesis of dscDNA in a two-step process. The first step, called First-Strand Synthesis (FSS), uses random primers, reverse transcriptase, and all the four deoxy nucleotides (dATP, dTTP, dCTP, and dGTP), resulting in a hybrid double-stranded RNA-DNA molecule. After FSS, the RNA molecule is degraded. In the second step, called Second-Strand Synthesis, the dTTP is replaced with dUTP. At the end of SSS, there is a dsDNA molecule, in which the strand with dTTP is the reverse complement of the sequence that was transcribed, and the strand with dUTP corresponds to the transcribed sequence. At this stage, the information about which strand was transcribed is already encoded in the chemistry of the created dscDNA. In the following step, the typical asymmetric Illumina Y-adapters are ligated to the dscDNA fragments. The incorporation of dUTP will quench the synthesis of the second strand during downstream amplification steps (Illumina 2017) or could be selectively degraded by Uracil-DNA-Glycosylase (UDG) (Borodina et al. 2011). Deciding whether an RNA-Seq dataset is stranded or not is quite easy and can be achieved by visual inspection of the reads mapped to either the genome or the transcriptome. However, some packages can aid inferring this, and are very useful when dealing with tens or hundreds of samples, some examples are the infer_experiment.py module part of RSeQC (Wang et al. 2012), or the option --lib-

Type A in Salmon (Patro et al. 2017), to name just a couple.

Data obtained from sequencing libraries prepared in such a way can be exploited either to map directly to a reference genome or transcriptome or build a de novo transcriptome assembly, in both cases exploiting the strand information and leading to correct directionality of the identified transcripts, with the potential for the identification of novel transcripts.

2.2.2.3 Long Read RNA Sequencing

Next-generation sequencing (NGS) technologies afforded the most widely used tools for transcriptome analysis in the recent years and are likely to remain pervasively used for many years to come. Still, RNA-Seq is not devoid of biases and limitations, notably about transcript identification and isoform disambiguation, as well as expression-level estimation. Short reads can be ambiguous, map to multiple locations, and originate from low-complexity sequences that hamper alignment.

The ability to sequence full-length transcripts, from the 5′ end to the poly-A tail, in principle allows complete differentiation of isoforms, with no ambiguity in assigning fragments to transcripts. It also eliminates the need for (de novo) transcript assembly. Third-generation sequencing (TGS) technologies already provide the means for achieving this goal, at least for a large fraction of the transcripts, with long reads that completely cover molecules with lengths upwards of 10 kbp. Besides facilitating transcript identification, long reads boost transcriptome analyses through the discovery of novel genes, novel isoforms, and detection of fusion transcripts (Rhoads and Au 2015; Shi et al. 2016). Even previously annotated sequences can be enhanced with these technologies, through correction of existing gene models (Liu et al. 2017). Furthermore, PCR-free protocols get rid of amplification biases that affect expression quantification.

One such technology is the Iso-Seq method (Rhoads and Au 2015) from Pacific Biosciences (PacBio). This isoform sequencing strategy has shown power to discriminate transcript isoforms in some important species (Abdel-Ghany et al. 2016; Li et al. 2018), including some with very complex genomes, such as cotton (Wang et al. 2018b), coffee (Cheng et al. 2017), and even the highly polyploid sugarcane (Hoang et al. 2017; Piriyapongsa et al. 2018). These studies collectively show that RNA-Seq based exclusively on short reads renders a limited view of the transcriptome, because of partial isoform identification and inaccuracies in expression quantification.

Long reads can also be obtained with the Oxford Nanopore technology. In addition to sequencing cDNA molecules, this approach allows direct RNA sequencing (Garalde et al. 2018), an alternative that removes reverse transcription biases and helps in identifying other types of RNA molecules, such as long non-coding and antisense RNAs (Jenjaroenpun et al. 2018). These technologies can also be applied for characterizing transcriptomes of individual cells (Byrne et al. 2017).

Despite these benefits, a series of practical concerns still limit the widespread application of third-generation sequencing technologies. Even though success in sequencing full-length transcripts is highly advantageous for cataloging the transcriptome of cells, quantitation is a different matter. Although potentially less biased for transcript abundance estimation (Byrne et al. 2017), the current lower throughput of these approaches prevents accurate quantification of transcripts in the wide dynamic range of expression levels, with more pronounced effects on lowly expressed transcripts. Increasing sequencing depth can circumvent this issue, but this is presently limited by the higher cost of long reads, such that efforts in improving throughput and lowering costs are vital.

Another obstacle is that sequencing errors rates are substantially higher for state-of-the-art long read technologies (Jenjaroenpun et al. 2018). Error rates in Iso-Seq reads can be greatly reduced by the so-called circular consensus sequence (CCS), in which the same molecule is repeatedly sequenced (Rhoads and Au 2015; Liu et al. 2017). However, this is not yet feasible for long, single-pass transcripts, which still suffer from lower sequencing accuracy. Hybrid strategies that combine the transcript identification power of TGS with the massive read volume of

NGS enable error correction and abundance estimation for a more complete and trustworthy transcriptome characterization (Li et al. 2018; Jenjaroenpun et al. 2018).

2.2.3 Transcriptome Assembly

2.2.3.1 Genome-Guided Transcriptome Assembly

When the genome sequence of the species under study is available, one can choose to try assembling the transcriptome from raw data (short reads) using the genome as a guide. This procedure consists of mapping the RNA-Seq reads onto the reference genome sequence and then looking for clusters of sequencing reads representing putative isoform transcripts that should be assembled. During the mapping step, the read mapper employed must be aware of spliced-reads, that is reads that span exon–exon borders, like HiSAT2 (Kim et al. 2015, 2019), STAR (Dobin et al. 2013), or GSNAP (Wu and Nacu 2010), among others. After reads have been mapped and clustered along the genome sequence, these clusters of reads are usually represented as a graph (Florea and Salzberg 2013). The graph model could be a splice graph, where exons or parts of exons are represented as nodes and edges represent possible splice variants, implemented in the software Stringtie (Pertea et al. 2015), or an overlap graph, where nodes represent sequence fragments or reads (*k*-mers) and edges connect sequence fragments if they overlap and have a compatible splice pattern, implemented in software such as Cufflinks, Scripture, and Trinity (Trapnell et al. 2010; Haas et al. 2013; Guttman et al. 2010). Alternatively the genome sequence could be just used to cluster reads together to be then de novo assembled, using software such as Trinity (Haas et al. 2013; Grabherr et al. 2011).

Genome-guided transcriptome assembly is usually more precise than de novo transcriptome assembly (see below), as it is less sensitive to sequencing errors, polymorphisms, and paralogous loci (Ungaro et al. 2017; Zhao et al. 2011). It is important to note, though, that it could only help in recovering/assembling the transcripts that are present in the sequence used as reference, so variation between individual, ecotypes, cultivars, etc. would be missed. This has been highlighted in recent studies about the pan-transcriptome and pan-genome of diverse plant species (Gao et al. 2019; Ma et al. 2019). Also, if the genome sequence used as reference is fragmented, exons or whole transcripts could be located in sequencing gaps. An alternative to overcome these limitations would be the generation of a comprehensive, or non-redundant, transcriptome, that leverages the information of the genome-guided transcript assembly and of de novo transcript assemblies (Visser et al. 2015; Jain et al. 2013). The PASA pipeline (Haas et al. 2003) and CD-HIT-EST (Fu et al. 2012) can generate such non-redundant transcriptome representations, by controlling the minimum fraction identity, and length aligned to create transcript clusters. Clustering at 100% identity would be the most basic level of clustering, and lower values, like 99% or 95% identity, could be useful to cluster transcripts originating from the same locus via alternative splicing, allelic versions, or closely related paralogous genes. GET-HOMOLOGUES-EST could enhance the generation of a comprehensive transcriptome, while taking into account coding potential, the presence of conserved protein domains, and information from closely related species or individuals within a polymorphic species (Contreras-Moreira et al. 2017).

2.2.3.2 De Novo Transcriptome Assembly

The availability of an annotated reference genome sequence eases the analysis of RNA-Seq data, by dividing the problem of transcript assembly and quantification into substantially smaller subsets. In this situation, sets of reads aligning against a particular genomic region can be analyzed independently of the remainder of the sequencing data.

It is nevertheless possible to carry out a thorough transcriptome analysis for non-model plant species lacking a reference genome (Collins et al. 2008). When available, the genome sequence of a closely related species can be used as a reference.

Alternatively, instead of aligning the reads against genomic sequences, a transcriptome reference can be assembled de novo based on the RNA-Seq reads alone. This provides a cost-effective means of applying functional genomics tools to less well-studied organisms. It can also shorten the path to biological insight because any species can potentially be studied without the need for previous genomic knowledge. However, de novo transcript assembly is one of the most difficult tasks in bioinformatics (Garg and Jain 2013).

The most widely used de novo transcriptome assemblers are based on a *de Bruijn* graph, a data structure that compactly represents the sequences of hundreds of millions of short sequencing reads. Construction of a *de Bruijn* graph involves parsing the collection of reads and extracting *k*-mers of a certain size. A *k*-mer is a subsequence of length *k* contained in any biological sequence segment, such as a read, a transcript, or even an entire chromosome. In a standard *de Bruijn* graph, each existing *k*-mer is represented by a node, or vertex. If a suffix of length $k - 1$ of a given node matches the $k - 1$ prefix of another node, an edge connecting these vertices is used to represent this overlap. After obtaining this graph, assembly software packages usually perform several (combinations of) steps of error correction, graph simplification and collapsing, scaffolding, and gap closure. Finally, graph traversal based on sequencing read information can be used to reconstruct contigs representing transcripts.

Contig assembly algorithms based on *de Bruijn* graphs were initially devised for genome assembly based on high depth sequencing data. Indeed, many of the currently available transcriptome assemblers were built relying on previously existing genome assemblers. For example, Oases (Schulz et al. 2012) is a pipeline built on top the Velvet genome assembler (Zerbino and Birney 2008). Similarly, Trans-ABySS (Robertson et al. 2010) is based on ABySS (Simpson et al. 2009), and SOAPdenovo-Trans (Xie et al. 2014) uses the *de Bruijn* graph from SOAPdenovo2 (Luo et al. 2012) as a starting point. Following a more widespread adoption of RNA sequencing studies, proper de novo assemblers such as Trinity (Grabherr et al. 2011; Haas et al. 2013), were also developed from scratch to tackle the challenges posed by these datasets.

Despite using an underlying data structure similar to genome assemblers, these software packages take into account unique features of the RNA-Seq data to drive the assembly strategy and address several particular issues. While the goal in genome assembly is to produce a few large (chromosome-sized) sequences, transcriptome assembly aims to reconstruct tens of thousands of sequences, each representing a different transcript. Also, coverage depth in RNA sequencing is heavily dependent on gene expression levels, such that approaches for assembling lowly or highly expressed genes can differ.

These de novo assembly methods can naturally handle alternative splicing arising from RNA processing after transcription. Ideally, a transcriptome assembly should contain full-length transcripts accurately representing different isoforms, while also separating paralogs from large gene families. For polyploid species, the presence of multiple alleles and homeologs adds another layer of complexity that makes assembly an even harder exercise. In this context, it is noteworthy that long-range information from paired-end and/or longer sequencing reads provide a valuable resource that can greatly enhance assembly quality by simplifying the recovery of full-length transcripts.

Even though the current transcriptome assemblers are based on similar basic concepts and share many features, they differ widely in running time and required memory. They also stand apart in their ability to recover full-length transcripts from datasets with varying sequencing depth, obtained from species with distinct transcriptome complexity. Comparisons among assemblers can reveal scenarios in which particular combinations of software and parameters show superior performance (Zhao et al. 2011).

Finally, functional annotation of the assembled transcripts is commonly done to provide meaningful biological information about each resulting sequence. This usually entails adding gene ontology terms (Ashburner et al. 2000; Gene Ontology Consortium 2017) and pathway

information from KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto 2000; Kanehisa et al. 2016), to the transcripts, as well as searching for protein domains. Pipelines for performing such annotation include Blast2GO (Conesa et al. 2005) and Trinotate (Bryant et al. 2017).

2.2.3.3 Assessment of Transcript Assemblies

The goal of transcriptome assembly, either genome-guided or de novo, is to generate a truly complete collection of all the transcripts produced by an organism. However, attaining that goal is in most real cases unlikely, some of the reasons for this include: (1) Sequencing depth is limiting, and lowly abundant transcripts are not represented in the sequencing data. (2) Biases of the sequencing depth limit the observation of certain transcripts, e.g., problems with high GC content sequences. (3) Not all possible transcripts are expressed at a given moment, a good transcriptome coverage should include a survey of samples from different developmental stages, growing conditions, tissues, and organs. Thus, we need tools to assess the quality and completeness of a generated transcriptome assembly (Honaas et al. 2016; Moreton et al. 2015; Li et al. 2014; Smith-Unna et al. 2016). In the following, we describe some of the most important metrics to evaluate a transcriptome assembly.

Evaluation of Sequencing Depth

There are two related questions that are often asked at the beginning of any transcriptome study using NGS. (1) How many reads should be generated to capture most/all of the transcripts? (2) Are the reads generated enough to make statistical inferences or to get a complete overview of the transcriptome? In order to answer these, one can evaluate the degree of read saturation present in the assembly as a function of sampling effort, using an approach analogous to that of species accumulation curves (rarefaction curves) in biodiversity studies. This approach will allow to decide whether sequencing depth has been enough to capture all transcripts in the sample (Hale et al. 2009). At the beginning of a study, before generating the data, one could carry out a pilot study with shallow sequencing depths, that could help estimating the depth required to capture all or most of the transcripts. Alternatively, and if a genome reference is available, one could evaluate the saturation of orthogonal features, for instance the number of exon–exon junctions supported by the sequencing reads at different levels of sequencing effort, this approach has been implemented in the tool junction_saturation.py in the package RSeQC (Wang et al. 2012).

Percent Reads Mapped

The proportion of reads that map back to the assembly is also a measure of assembly and data quality. In principle one wants most of the original read data (after quality trimming) mapping to the transcriptome assembly. However, when using a genome as a reference (or the transcriptome derived from the genome sequence), a low percent of reads mapping could also be indicative of large diversity between the reference and the sample, or of contamination, and further analyses would be required.

Identification of Sets of Conserved Genes

Genes that appear in all of the best-known genomes can be exploited to evaluate the completeness of a transcriptome assembly. The tool Benchmarking Universal Single-Copy Ortholog (BUSCO) has sets of conserved single-copy orthologous genes present at diverse taxonomic levels, e.g., Viridiplantae (green plants), Embryophyta (land plants) (Waterhouse et al. 2017). A transcriptome that was assembled from samples representing different developmental stages, growth conditions, tissues and organs, should have a good representation of these conserved single-copy gene sets. On the other hand, a transcriptome representing a single condition could have a low value for this metric, corroborating its specificity. Alternatively one could also compare the assembled transcripts to the transcripts (or proteins) of a related species, these are usually called reference-based or comparative metrics and are implemented in tools such as TransRate (Smith-Unna et al. 2016) or Detonate's REF-EVAL (Li et al. 2014).

Contamination Screening and Filtration

NGS data can easily be contaminated, but it is important to note that there are different sources of contaminants. There can be internal contaminants, for instance, mitochondrial and plastid sequences, or ribosomal RNA sequences. Or there could be external contaminants, genetic material from other organism present in the sample, e.g., symbionts, pests, fungi, or bacteria. In general, contamination should be removed as early as possible, in order to reduce computational costs, fragmentation of the assembly and the chance to generate chimeric transcripts (Zhou et al. 2018). For example, BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/) can be used to efficiently remove rRNA reads by comparing them against the SILVA database (Quast et al. 2013). A similar approach could be followed to eliminate reads from other contaminants if they have been previously identified. The presence of rRNA could be exploited to identify which contaminants (if any) are present in the sample.

2.2.4 Transcript Quantification

2.2.4.1 Alignment/Mapping-Based Approaches

Transcriptome characterization via RNA-Seq not only provides a catalog of transcripts present in a particular sample of cells, but also yields quantitative information that allows expression levels to be assessed. This is true both for species with and without a reference genome. A major step for obtaining expression estimates is to assign sequencing reads to genes or transcripts, which is commonly accomplished by first aligning them to a reference genome or transcriptome sequence.

Development and application of alignment algorithms has been one of the most active research areas in bioinformatics, and consequently, there is a wide range of tools available for various purposes. The majority of alignment algorithms tailored for short reads use indexing strategies that can be categorized into two main approaches: a seed-and-extend strategy based on hash tables or alignment based on a Burrows-Wheeler transform (Flicek and Birney 2009; Trapnell and Salzberg 2009; Li and Homer 2010).

Short read sequence aligners were initially developed for aligning genomic reads against a reference genome. In this situation, reads are expected to align contiguously against the reference, except for minor gaps which may stem from small indels or sequencing errors. Reads from RNA-Seq libraries, on the other hand, originate from cDNA molecules synthesized from mature mRNA templates, from which introns have been stripped off. Aligning RNA-Seq reads against a reference genome then requires splice-aware aligners, which appropriately handle reads that span exon junctions, without penalizing long gaps corresponding to introns. This class of aligners includes TopHat2 (Trapnell et al. 2009; Kim et al. 2013), which has been superseded by HISAT2 (Kim et al. 2015) and STAR (Dobin et al. 2013). An interesting quality of these aligners is that they can not only use previously annotated splice junctions, but also discover novel junctions and isoforms.

Following alignment, mapped reads can be assigned to annotated features in the genome. A simple and widely used way to measure expression levels is to count the number of reads overlapping a feature of interest. This is the approach implemented in programs such as HTSeq (Anders et al. 2015) and the featureCounts (Liao et al. 2014) component of the Subread package (Liao et al. 2013).

Reflecting the nature of gene expression, feature annotation follows a hierarchy of terms, with a gene frequently corresponding to the highest-level term. Any given gene may originate one or more transcripts, which in turn may contain one or more exons and compose one or more coding sequences. Read counts can be obtained for features at any level desired, but it is frequent to count reads overlapping exons. Depending on the goals of the study, features may then be grouped to obtain expression levels for meta-features. For instance, counts for all exons of a given transcript may be combined to get a transcript-level expression estimate, or all exons of all transcripts of a gene may be used to yield a gene-level read count. It is important to realize that, when working with paired-end read information, both reads of a pair come from a single molecule fragment, such that they should contribute only once to the expression count.

It is not always possible to uniquely assign a read to a feature or meta-feature. In some cases, there are overlapping features in an annotated genome reference, as a consequence of the structural organization of genes in the species of interest. Reads that align to a genomic region covered by two or more genes may not unequivocally be assigned to any one of them. Much of this ambiguity can be worked out by using stranded RNA-Seq library preparation, because overlapping genes may be transcribed in opposing directions.

Additionally, different gene isoforms can share a common exon, such that reads overlapping this exon are ambiguous. Lastly, the aligner may report multiple possible mappings for some reads, due to sequence similarity between members of a gene family, conserved protein domains and sequencing errors. The researcher can decide whether to simply discard multimapping or ambiguous reads, count them for all overlapping features or assign them heuristically. It should be noted that ambiguities at a given annotation level may not represent ambiguities at a higher level (e.g., a read mapping to an exon shared by multiple isoforms is ambiguous at the transcript level, but not at the gene level).

When using a de novo assembled transcriptome, introns are virtually absent from the reference, and therefore, one may use standard sequence aligners, such as BWA-MEM (Li and Durbin 2009; Li 2013) and Bowtie2 (Langmead et al. 2009; Langmead and Salzberg 2012). Splice-aware aligners also have modes for aligning reads against a splice junction-free reference sequence. For expression level quantification, in this case each contig can independently be treated as a feature. In fact, some assemblers such as Trans-ABySS may internally leverage the alignment of reads to contigs and automatically provide a measure of the per-contig expression level. The simplicity of the feature annotation in an assembled transcriptome does not mean that alignment and quantification are an easier endeavor. In fact, the issue of multiply aligned reads can be even more challenging in this situation, as it can be hard to distinguish between paralogs of the same gene.

These ambiguity issues have prompted alternative approaches for obtaining expression estimates to be devised. Because of the uncertainty in determining the transcript of origin of sequencing reads, one such possibility is to use mixture-model procedures that probabilistically assign reads to features, instead of simply counting overlapped fragments. As an example, the RSEM method (Li et al. 2010; Li and Dewey 2011) generates maximum likelihood or Bayesian expression estimates based on several variables of the annotated feature set and of the aligned reads, such as length, orientation, and quality scores. The main underlying principle is that uniquely aligned reads can also provide information for the (probabilistic) assignment of ambiguous reads. For example, suppose that two isoforms of a gene share one common exon, but also contain one exclusive exon each. If a large number of fragments align to one of the exclusive exons, while the other shows no overlapping reads, it is likely that fragments overlapping the common exon also originate from the isoform with a higher expression level based on the uniquely aligned reads.

Similarly, the Stringtie package formulates the simultaneous estimation of isoform assembly and abundance as a maximum network flow problem (Kovaka et al. 2019; Pertea et al. 2015). This maximum flow approach has been shown to be as accurate as the maximum likelihood approach in cufflinks (Trapnell et al. 2010), but it is able to recover a larger fraction of bona fide transcripts (Kovaka et al. 2019). In the maximum flow approach, a path in the splice graph with the heaviest coverage is used to build a flow network, this path represents a transcript, which is then removed from the splice graph, and a new path with the heaviest coverage is sought, until no more transcripts are assembled. The coverage for each assembled transcript is used to represent expression values as FPKM (fragments per kilobase million) and TPM (transcripts per million).

These difficulties in estimating expression levels are substantial enough for diploid model species. The situation may be considerably harder for researchers dealing with polyploid organisms, because of the added complexity from homeologs and multiple alleles. It is reasonable to

assume that probabilistic strategies for read assignment may provide more accurate estimates of transcript abundance in this case.

Finally, a brief comment on expression-level normalization is needed. Transcript read counts are influenced by the length of the transcript and the size of the sequenced library, i.e., the number of fragments obtained from a given sample. Read counts are expected to be higher for longer transcripts and larger libraries. Many downstream application packages directly handle raw read counts, but it is not always straightforward to interpret raw values. For reporting expression levels, for instance, it is useful to use normalized values, such as the TPM (*transcripts per million*) value (Li et al. 2010; Wagner et al. 2012). It represents the number of transcripts of a certain type present in a total of one million sequenced transcripts from a given sample and thus estimates the fraction of that transcript in a pool of RNA molecules. The TPM is normalized by the length of the transcript, the sequencing depth, and the mean transcript length in the sample. Relative expression levels represented by TPM values do not depend on the expression levels of other genes in the transcriptome and appropriately measure the fraction of fragments from a given gene or isoform. Other measure of gene expression includes the RPKM (reads per kilobase million) and FPKM (fragments per kilobase million), but they have been largely superseded by the TPM.

2.2.4.2 Alignment-Free Approaches

Recent methods have tried to let go of the traditional strategy of mapping reads to a reference and then count, to arrive at estimates of gene expression levels, approach described above. The main reason for this is that these traditional approaches require large computational resources, and do not scale well with the amount of available data. These newer approaches implement what they call as pseudo-alignment, lightweight mapping, or quasi-mapping (Patro et al. 2017, 2014; Bray et al. 2016) and are known as alignment-free methods. Another important difference to the traditional approach is that instead of using reference genomes, these approaches use reference and well-annotated transcriptomes, including transcript isoforms, allowing the accurate estimation of isoform expression levels. Expression-level estimates at the level of isoforms are important given that most plant genes are interrupted (i.e., they have introns), and the removal of introns is a regulated process that can generate alternative splicing forms, which can have different, even antagonistic functions (Shang et al. 2017). In order to estimate isoform expression levels, tools like Kallisto or Salmon, let go of the idea of knowing where a read aligns in a given transcript, with base-to-base correspondence, and instead try to identify a transcript, or a set of transcripts, that could have originated such read, without keeping track of base-to-base correspondences. Such approaches have been shown to be extremely fast and accurate (Zhang et al. 2017). Some of these methods, besides their speed, can model different sources of sample-specific biases that can affect transcript quantification, like sequence-specific, fragment GC-content and positional biases (Patro et al. 2017; Bray et al. 2016). Refinement of the initial lightweight mapping of reads to the transcriptome, using Selective Alignment, allows the elimination of most mapping errors, by providing alignment scores that allow to distinguish alternative mapping locations that otherwise would appear the same (Srivastava et al. 2019).

2.3 Applications

Figure 2.3 shows some of the paths that can be followed in RNA-Seq studies. Table 2.1 lists some of the main software packages to carry out the operations shown in Fig. 2.3.

2.3.1 Differential Gene Expression

RNA sequencing is frequently done with the goal of detecting differences in expression levels between two or more contrasting groups of samples. One may be interested in evaluating the effect of different experimental treatments, genotypes, or stress conditions, for instance, on the

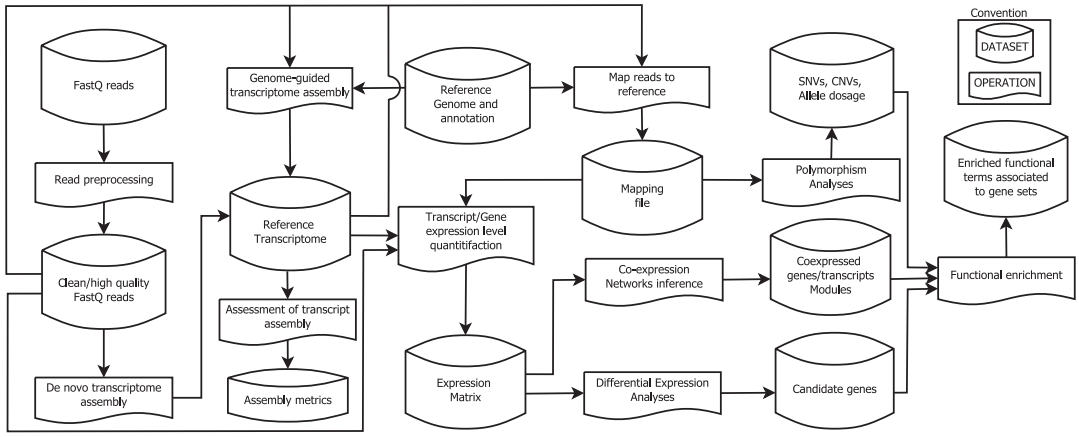

Fig. 2.3 General steps in an RNA-Seq analysis pipeline. Not all steps/paths are taken in a given study

transcriptome of particular cells. Gene or isoform expression measures are thus often used for identifying transcripts that are significantly up- or downregulated in a condition of interest, in comparison to a distinct condition.

Differences in the expression levels of two (groups of) samples can be represented by the fold change, which is simply the ratio of the expression levels estimated for both cases. Usually the expression estimate of a control or reference condition is used in the denominator, whereas the expression level of the treatment group is used in the numerator. As a result, genes that are upregulated in the treatment samples show a fold change greater than one (with no upper boundary), while downregulated genes display a fold change between zero and one. This discrepancy in scale led to the representation of these ratios in the log2 scale, such that fold changes in both directions are symmetric around zero.

Several methodologies are available for testing whether an observed fold change is statistically significant. Many of these methods use read count data directly, which calls for modeling of the expression levels with discrete distributions. The first statistical approaches proposed for such tests used the Poisson distribution to model read counts, assuming that the variance in the estimates was directly proportional to the mean expression level (Wang et al. 2010). This proved to be appropriate for technical replicates of the same sample (Marioni et al. 2008), but variance for biological replicates was shown to be higher than expected based on the mean alone (Robinson and Smyth 2008).

An alternative to the Poisson distribution is the negative binomial, which adds a second parameter (often denoted dispersion), allowing the sample variance to be different from the mean; hence, it corresponds to a Poisson distribution with overdispersion. This is the approach taken by most of the modern differential expression analysis packages (Wang et al. 2010; Robinson et al. 2010; Trapnell et al. 2013; Love et al. 2014).

The need to estimate sample variances makes it clear that biological replication is necessary in RNA-Seq experiments. Appropriate design planning is required, and all treatment combinations should be replicated, as alternatives devised for data without replicates are far from ideal. Yet, despite continual reduction in sequencing costs, RNA-Seq for large numbers of samples may still be impractical for many research goals. In order to increase reliability of variance estimates obtained from small numbers of replicates, techniques that share information between genes were proposed and implemented (Robinson and Smyth 2007).

Software packages edgeR (Robinson et al. 2010), DESeq (Anders and Huber 2010; Love et al. 2014), and Cuffdiff (Trapnell et al. 2010, 2013) are among the most extensively used tools

Table 2.1 Some of the software packages for different steps in RNA-Seq analysis pipelines

Activity		Software	Reference
Read pre-processing	Sequencing diagnostics	FastQC	bioinformatics.babraham.ac.uk/projects/fastqc
		RSeQC	Wang et al. (2012)
		RNA-SeQC	DeLuca et al. (2012)
	Removal of adapters and low-quality bases	Trimmomatic	Bolger et al. (2014)
		Atropos	Didion et al. (2017)
		BBDuk	sourceforge.net/projects/bbmap/
	Removal of ribosomal RNA	SortMeRNA	Kopylova et al. (2012)
		BBDuk	sourceforge.net/projects/bbmap/
	Identification of duplication artifacts	dupRadar	Sayols et al. (2016)
De novo transcriptome assembly		Trinity	Grabherr et al. (2011)
		Trans-ABySS	Robertson et al. (2010)
		Velvet/Oases	Schulz et al. (2012)
Genome-guided transcriptome assembly		Trinity	Grabherr et al. (2011)
		Stringtie	Kovaka et al. (2019)
		PASA	Haas et al. (2003)
Assessment of transcriptome assembly		BUSCO	Waterhouse et al. (2017)
		DETONATE	Li et al. (2014)
		Transrate	Smith-Unna et al. (2016)
Functional annotation		Trinotate	Bryant et al. (2017)
		Blast2GO	Conesa et al. (2005)
Read mapping		STAR	Dobin et al. (2013)
		GSNAP	Wu and Nacu (2010)
		HISAT2	Kim et al. (2015), Kim et al. (2019)
Transcript/gene expression-level quantitation		Stringtie	Kovaka et al. (2019)
		featureCounts	Liao et al. (2014)
		kallisto	Bray et al. (2016)
		Salmon	Patro et al. (2017)
Differential expression analyses		Limma	Ritchie et al. (2015)
		edgeR	Robinson et al. (2010)
		Ballgown	Frazee et al. (2015)
		Sleuth	Pimentel et al. (2017)
		DESeq2	Love et al. (2014)
Co-expression network inference		WGCNA	Langfelder and Horvath (2008)
		HRR	Liesecke et al. (2018)
		HCCA	Mutwil et al. (2010)
Polymorphism analyses		GATK	DePristo et al. (2011), McKenna et al. (2010)
		NGSEP V4.0	sourceforge.net/p/ngsep/
Functional enrichment		goseq	Young et al. (2010)
		topGO	Alexa et al. (2006)
		Blast2GO	Conesa et al. (2005)

for differential expression analyses. In more detail, edgeR uses raw read counts and models sample variation in terms of the biological coefficient of variation, which corresponds to the square root of the dispersion. It allows estimating a common dispersion for all genes, or a trended dispersion via a locally weighted adjusted profile likelihood for genes with similar average read count. It further allows moderated gene-wise dispersion estimates to be obtained by a weighted

likelihood method combining individual and trended or common estimates (McCarthy et al. 2012). Normalization is carried out with a trimmed mean of log2 fold changes (Robinson and Oshlack 2010).

Similarly, the DESeq2 package uses size factors estimated based on the median of ratios of observed read counts to normalize expression levels. It empirically estimates the relationship between mean and variance of the negative binomial distribution, fitting a smooth curve of the dispersion as a function of the average expression of genes with similar means. Finally, it employs empirical Bayes approaches to shrink gene-wise dispersion estimates and also the fold changes, which is particularly relevant for lowly expressed genes and/or those with highly variable expression levels. Both edgeR and DESeq were initially designed for performing differential expression analyses of simple experiments, commonly involving pairwise comparisons of contrasting conditions. More recent implementations of edgeR and DESeq2 allow fitting generalized linear models for analysis of more complex designs, with the inclusion of experimental blocking factors and modeling of interactions, for example.

Cuffdiff 2 was developed for testing differential expression at both the isoform and the gene levels. Instead of using raw read counts, it models variability across replicated expression estimates by jointly considering overdispersion and uncertainty in the assignment of reads to their possible originating transcripts. Because of differences in the normalization procedures and model assumptions, these methods differ in their statistical power to detect differential expression over the range of expression values, as well as in the occurrence of false positives. Note also that conducting differential expression analyses at the transcript level may have important implications for statistical power. Greater uncertainty in expression estimates, because of more ambiguously mapped reads, negatively influences statistical power. Differential isoform expression analyses may require higher coverage depth, as more reads are needed to provide accurate estimates of individual isoform expression levels, especially for genes with many isoform variants and many shared exons. On the other hand, failure to adequately model uncertainty in read to transcript assignment can result in higher rates of false positives, even at the gene level.

RNA-Seq is a high-throughput screen that yields quantitative information for tens of thousands of genes (or hundreds of thousands of transcripts). Consequently, statistical tests are applied for multiple comparisons, which can result in many false positives if liberal significance levels are used for individual tests. Multiple testing correction is generally used to control for the occurrence of such false positives. One of most well-known corrections is the Benjamini and Hochberg (Benjamini and Hochberg 1995) false discovery rate (FDR) correction, aimed at controlling the proportion of false discoveries among the rejected hypotheses, while minimizing the drop in statistical power.

The output of these analyses is a list of significantly differentially expressed genes. Because of the large number of genes studied, this list may be quite long, which complicates summarization and reporting of the results. More easily interpretable biological meaning can be extracted from such lists through functional enrichment analyses, that look for overrepresented groups of genes among the statistically significant ones. Groupings of interest are usually obtained by categorizing genes according to their functional annotation, including gene ontology terms and/or biological pathways. Each functional group is tested for overrepresentation in the gene list against a background set, which includes all (expressed) genes in the transcriptome.

2.3.2 Co-expression Networks

Networks have recently emerged as a robust and holistic approach to understand complex cellular processes that comprise multiple and parallel interactions between cellular constituents such as DNA, RNA, and proteins. The network approach allows analyzing components and interactions as a system instead of analyzing them as separate entities. In a general way, a network, or graph, is defined as a set of elements called nodes, which

are related through connections called edges. When edges have a direction, that is, they have source and target nodes, the network is called directed; otherwise, the network is undirected. These simple definitions are used to create biological networks that model cellular processes by taking nodes to represent molecules such as genes, proteins, or metabolites, and edges to represent physical, functional, or chemical interactions (Barabasi and Oltvai 2004). Depending on the molecules and interactions used, biological networks can be gene co-expression networks (GCN), genetic interaction networks, gene regulatory networks, protein–protein interaction (PPI) networks, metabolic networks, and signaling networks (Serin et al. 2016; Vital-Lopez et al. 2012). This section will focus on gene co-expression networks, in which each node corresponds to a gene, and edges represent co-expression relationships.

An advantageous feature of GCNs is the ability to reduce data complexity drastically. Nodes in a GCN, rather than solely representing a gene per se, represent its whole expression profile when the studied organism is under a condition, such as a treatment or biotic/abiotic stress. Edges in a GCN represent associations between gene expression profiles and can be interpreted as the simultaneous and coordinated expression of two or more genes under the studied perturbations. Thus, GCNs reduce the complexity of expression data of multiple samples from one or multiple experiments.

GCNs can be constructed from expression data derived from DNA microarrays and RNA-Seq. Traditionally DNA microarrays were the primary source of data expression for constructing GCNs, as this technology has been used intensively for almost two decades in gene expression studies. Recently, with the advent of next-generation sequencing (NGS) technologies, RNA-Seq has turned in a natural source for constructing GCNs. Among the advantages that microarrays had over RNA-Seq for the reconstruction of GCNs, we can name the considerable amount of information available in public databases, the well-established and mature data normalization approaches, and data homogeneity. Although RNA-Seq was shown as a promising source of data for GCNs (Iancu et al. 2012), some limitations related to normalization methods used for this technology were also demonstrated (Giorgi et al. 2013). However, with the increased number of RNA-Seq samples publicly available, more recent studies have shown that bigger datasets can overcome those caveats (Ballouz et al. 2015; Huang et al. 2017a) and highlight multiple advantages of RNA-Seq over microarrays for GCNs.

GCN inference comprises three main steps: similarity calculation, filtering, and edges construction (Serin et al. 2016). In the first step, a measure of similarity (or relatedness) is computed for each pair of genes. Multiple measures can be used in this step, such as mutual information (MI) (Meyer et al. 2008, 2007), or the prevalent correlation coefficients. The latter category includes the Pearson correlation coefficient (PCC), Spearman's correlation coefficient (SCC), and biweight midcorrelation (bicor) (Langfelder and Horvath 2008). Although MI is useful for finding nonlinear relationships between genes (Langfelder and Horvath 2008), it has been shown that it has several caveats and can be outperformed in many situations by correlation measures (Liesecke et al. 2018; Song et al. 2012). In the second step, the pairs of genes (edges) are either filtered based on a relatedness threshold that specifies the minimum level of similarity between expression profiles to define if a pair of genes is connected, or weighted. When using a threshold, it can be defined as a simple cutoff (hard threshold) (Tsaparas et al. 2006; Qiao et al. 2017), or as a result of more elaborated approaches. Some of these approaches include selecting a subset of the most positive/negative correlations (Lee et al. 2004), relying on topological features of co-expression networks like the clustering coefficient (Elo et al. 2007) or a power law distribution of the number of edges per node (Zhang and Horvath 2005), or applying models such as the Random Matrix Theory (Luo et al. 2007). Finally, in the third step, edges of the GCN are defined based on the resultant list of genes after filtering.

Depending on the type of connection between nodes, GCNs can be unweighted and weighted.

In the unweighted networks, edges indicate whether there is an association between a pair of nodes. They are derived from applying a hard threshold, i.e., an edge is present if the similarity measure between nodes is above the cutoff value. In the weighted networks, the degree of association between nodes is quantified by an attribute called weight, which commonly corresponds to a value in the range [0, 1]. This weight can result from applying a soft similarity threshold (Langfelder and Horvath 2008; Zhang and Horvath 2005) or from assigning a value derived from correlations such as the coefficient rank (Ballouz et al. 2015).

After constructing a GCN, a wide repertoire of analyses from graph theory, computer science, and engineering can be applied for elucidating valuable information hidden in the expression data. For example, by applying clustering algorithms like the hierarchical clustering (Langfelder and Horvath 2008), or the Markov Cluster (Zhang et al. 2012), it is possible to identify groups of highly coexpressed nodes (modules) with similar functions or involved in common biological processes. Modules are annotated with functional and metabolic information publicly available in databases such as Gene Ontology (GO, http://www.geneontology.org), Reactome (https://reactome.org/), and the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/).

Another example of methods applied to GCNs are the topological analyses that examine the structural properties of networks. One of the most used topological properties is the node degree, which indicates how many connections each node has. It has been suggested that some biological networks are scale-free, which means that their degree distribution $P(k)$ approximates a power law $P(k) \sim k^{-\gamma}$ (Barabasi and Oltvai 2004). However, in many cases, proper statistical tests have revealed otherwise (Arita 2005; Broido and Clauset 2019; Lima-Mendez and van Helden 2009; Khanin and Wit 2006; Stumpf and Ingram 2005), and methods that strongly rely on the power-law distribution of the node degree must be assessed critically. In general, biological networks, including co-expression networks, exhibit many nodes poorly connected (low degree) and a relatively small number of nodes with many connections. Highly connected nodes (hubs) are usually representative of the biological function associated with a module and also have been associated with interesting processes like regulation (Hollender et al. 2014), and evolution (Masalia et al. 2017). Another biologically relevant topological property is the betweenness centrality that indicates the level to which a node works as a bridge between other nodes and allows to detect bottlenecks (genes with high centrality). Since high connectivity and betweenness centrality tend to be related to essentiality in functional processes (Carlson et al. 2006), they can be used to identify key genes with biological relevance. Other topological properties with biological relevance, including clustering coefficient, density, centralization, and heterogeneity, have also been explored (Dong and Horvath 2007; Horvath and Dong 2008).

GCNs have been used mainly for two purposes, gene function prediction, and the selection and prioritization of genes associated with specific phenotypes like diseases or traits. The first application is derived from module identification and annotation, which infer functions for uncharacterized genes following the "guilt by association" principle (Oliver 2000). For instance, functions for unknown genes have been predicted in yeast (Luo et al. 2007) and grapevine (Liang et al. 2014) using GCNs. The second application is perhaps the most popular of GCNs, and it is derived from exploiting network centrality properties (e.g., degree and betweenness) combined with module information. For example, several studies have used GCNs to identify genes associated with traits of interest in plants, such as heat shock recovery in grapevine (Liang et al. 2014), aluminum stress response in soybean (Das et al. 2017), sugar/acid ratio in sweet orange (Qiao et al. 2017), regulation of cell wall biosynthesis in sugarcane and bamboo (Ferreira et al. 2016; Ma et al. 2018), wood formation in *Populus trichocarpa* (Shi et al. 2017), the regulation of catechins, theanine, and caffeine metabolism in the tea plant *Camellia sinensis* (Tai et al. 2018), and plant height in maize (Wang et al. 2018a).

GCNs also have some caveats that are worth mentioning. GCNs provide only direct information for co-expression and not of direct interactions between its components like in PPIs. Additional information such as functional relationships or the essentiality of genes is elucidated by applying analyses that can be prone to biases, for example, clustering or annotation methods. Biologically meaningful conclusions are only supported by reliable networks that sometimes are difficult to obtain due to multiple factors in the construction like the amount and quality of the expression data, or the appropriate selection of similarity measures, parametrization (e.g., thresholds), and clustering methods.

Despite the caveats and difficulties in their inference, it has been shown that GCNs remain useful tools in gene expression analysis. They allow to reduce the complexity of the currently growing expression data, suggest functions of unknown genes, and identify essential genes involved in biological processes of interest.

2.3.3 Polymorphisms

Sequencing reads from RNA-Seq studies are often used for identifying polymorphisms in the expressed regions of the genome. The principles of variant identification from transcriptomic data are similar to those involved in variant calling from DNA sequencing and many important applications are possible. Briefly, software such as GATK (McKenna et al. 2010; DePristo et al. 2011) and BCFtools (Li et al. 2009; Li 2011) traverse genomic positions from a reference sequence and compare the aligned reads to identify single-nucleotide polymorphisms (SNPs) and insertions and deletions (indels). However, there are important particularities when working with RNA-Seq data and care must be taken when interpreting the results.

If these aligned reads are originated from transcriptomics datasets, polymorphic sites can only be identified between expressed transcripts. This is useful, for instance, if the goal is to search for imbalance of expression levels among different alleles of the same gene, or allele-specific expression (Pham et al. 2017; Shao et al. 2019). Accuracy for detecting polymorphisms and estimating allele expression ratios depends on the depth of coverage. This can be improved by increasing the sequencing depth but also depends on the expression level of each gene (Castel et al. 2015). Highly expressed genes naturally draw on a larger proportion of the sequencing data and thus offer more power to identify variants and higher accuracy of allelic expression estimates. On the other hand, lowly expressed genes are more prone to false negatives and require deeper sequencing to accurately identify polymorphisms.

Also, the fact that identified variants are constrained to expressed exons can limit the scope of the study. Polymorphic sites in introns, regulatory and intergenic sequences, which can be more numerous and may have key biological significance, cannot be identified from RNA-Seq data alone (Cubillos et al. 2012; Magalhaes et al. 2007). Genomic variants located in alleles that are not expressed in a given transcriptome will also be missed. Finally, many possible posttranscriptional modifications may negatively impact variant calling results and lead to flawed conclusions (Lee et al. 2013).

Variant calling efforts and studies of allelic imbalance are even more complicated in polyploid organisms, where more than two different alleles can be found (Cai et al. 2020). First, for allopolyploids, it can be difficult to differentiate between true alleles and homeologous sequences, which may not be polymorphic within each subgenome (Yang et al. 2018a). Additionally, it is important to note that allele ratio information from RNA-Seq data is not appropriate for quantitative genotyping (estimating genomic dosage) in autopolyploids, because of differences in the expression levels of different alleles. In other words, while the variation in allelic expression levels does provide valuable biological information, these ratios are affected by expression control mechanisms and do not necessarily reflect allele dosage at the DNA level (Pham et al. 2017).

Considering these complications and limitations, in most scenarios a combination of variant calling with other strategies is more valuable, such as identifying polymorphisms from both RNA-Seq and whole-genome sequencing (WGS) data, for instance.

2.3.4 Machine Learning Technologies for Transcriptomics

The advent of high-throughput technologies like microarrays and next-generation sequencing has led researchers in biosciences to face the challenges of analyzing large amounts of data. These challenges include heterogeneity, high dimensionality, noisiness, incompleteness, and computational expensiveness, among others. Machine learning (ML) has emerged as a suitable solution for analyzing massive data while dealing well with its challenges. ML has been extensively applied for large-scale data analysis in fields such as genetics (Libbrecht and Noble 2015), biomedicine (Mamoshina et al. 2016; Leung et al. 2016), genomics, transcriptomics, proteomics, and systems biology (Larranaga et al. 2006; Min et al. 2017). This section presents an overview of ML that includes basic concepts and applications on transcriptomics in plants.

ML can be defined as the computational process of automatically learning from experience to make predictions on new data (Murphy 2012). The process of learning is carried out by extracting knowledge from exemplary data by identifying hidden patterns. ML methods are classified into two main groups, supervised and unsupervised learning. Supervised learning is a predictive approach that comprises data examples with inputs and outputs. This approach uses evidence from the example data to make a model that generates reasonable predictions for new unseen datasets. More formally, the example data corresponds to a set of input–output pairs D called training set and defined as,

$$D = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N},$$

where x_i is a training input of the set x, y_i is the response variable that represents an output from the set y, and N is the number of training examples. Hence, the model is trained to learn how to map each x_i to a corresponding output y_i.

Supervised learning methods can be subdivided into two categories according to the nature of predictions. When the response variable is discrete or categorical, e.g., male or female, healthy or diseased, the method falls into the classification category. General applications of classification algorithms are voice and handwriting recognition, and document and image classification. Common algorithms of this category include support vector machines (SVM) Support Vector Regression (SVR), k-nearest neighbor (KNN), decision trees, logistic regression, and neural networks. When the response variable is continuous, e.g., the height of a person, or a temperature, the method corresponds to the regression category. Regression algorithms include linear and nonlinear models, neural networks, and regularization. A variation of the late category is the ordinal regression, which comprises methods whose response variable has a natural ordering.

The second main group of ML, unsupervised learning, uses data examples with just inputs, i.e., the set

$$D = \left\{x_i\right\}_{i=1}^{N}.$$

This type of ML tries to elucidate hidden patterns in data, which can be considered "interesting" to the researcher. In this case, there is no information about the kind of patterns that are expected to be found in the data. Unsupervised learning, also called knowledge discovery, is more commonly used than unsupervised techniques. Two notorious categories within unsupervised learning are clustering and dimensionality reduction. Clustering algorithms are intended to group data by looking for similarities among the features of each element from the input. Standard clustering algorithms include k-means, self-organized maps (SOM), hierarchical clustering, and hidden Markov models. Dimensionality reduction algorithms try to extract the "essence" of data (Murphy 2012) by selecting a subset of features that represents better the dataset (feature selection) or by transforming the high-dimensional space of the original data into a lower one (feature extraction). Usual algorithms for dimensionality reduction are principal component analysis (PCA), linear discriminant analysis (LDA), and generalized discriminant analysis (GDA).

Supervised ML techniques have been applied in transcriptomics-related tasks such as assembly, identification, and abundance estimation of transcripts, splicing sites/events detection, non-coding

RNA identification, and gene selection. Transcriptome assembly is one of the essential tasks in RNA-Seq-based studies that is followed by analyses such as, the estimation of gene expression levels or differential gene/trnascript expression. IsoLasso is a reference-based RNA-Seq transcriptome assembler that uses an ML regression algorithm called Least Absolute Shrinkage and Selection Operator (LASSO) and has the interesting feature of identifying and quantifying novel isoforms (Li et al. 2011b). Another ML-based tool for transcript identification and abundance estimation is SLIDE, which uses a linear model that models the sampling probability of RNA-Seq reads from mRNA isoforms, and a modified LASSO algorithm for estimating parameters (Li et al. 2011a). Unlike IsoLasso, SLIDE requires the coordinates of transcripts and exons previously assembled with other tools.

Identifying splicing sites and splicing events is crucial for determining isoforms and, thus, for estimating the abundance of transcripts. TrueSight is a tool developed for detecting splice junctions (SJs) based on an iterative regression algorithm that uses RNA-Seq mapping information and splicing signals from the DNA sequence of a reference genome (Li et al. 2013b). TrueSight was tested using simulated and real datasets from humans, *D. melanogaster*, *C. elegans*, and *A. thaliana*, and showed better specificity and sensitivity compared to other SJs detection applications. A recently developed tool called DeepBound also uses alignment information to determine SJs and infer boundaries of expressed transcripts from RNA-Seq data (Shao et al. 2017). DeepBound utilizes deep convolutional neural fields (DeepCNF), a technique that belongs to an emerging ML branch referred to as deep learning (Mamoshina et al. 2016; Min et al. 2017; Angermueller et al. 2016). All the described applications for transcript abundance and SJ detection can be used in plants. However, except for SLIDE, these tools are not suitable for being applied directly to non-model species, as they depend on a reference genome.

In plants, supervised learning methods have also been used for detecting alternative splicing (AS) events. SVM classifiers were employed to detect two types of AS events, exon skips and intron retentions, in *A. thaliana* from tiling arrays data (Eichner et al. 2011). EST and cDNA data were used for training with two SVM layers: one for classifying sequence segments as introns or exons, assigning probabilities of being included in mature mRNA, and a second layer to predict AS events by using the probabilities from the first layer. In addition to SVM, Random Forest (RF) has been used to detect intron retention in *A. thaliana*, the most common type of alternative splicing in this species. These RF were created using a hybrid approach that combines essential features (i.e., length, nucleotide occurrence probabilities, AT and GC content) with additional features (i.e., common motifs, splice sites, and flanking sequences) to differentiate retained introns from constitutively spliced introns. These RFs had a better classification performance than SVM (Mao et al. 2014).

Noncoding RNAs (ncRNAs) are determinant in cellular processes like regulation and alternative splicing. Several ML methods have been applied to discover ncRNAs, including micro RNAs (miRNA) and long non-coding RNAs (lncRNA), using NGS datasets. In the case of miRNAs, decision trees (based on the C4.5 algorithm) combined with genetic algorithms, allowed the prediction of miRNA targets in humans from datasets that comprise genomic and transcriptomic information (Rabiee-Ghahfarrokhi et al. 2015). miRNAs were predicted in 18 different plant species from data extracted from RNA-Seq, chromosome sequences, or ESTs, exploiting decision trees (C5.0 algorithm) (Williams et al. 2012). An SVM approach was employed to identify miRNAs associated with cold stress in *A. thaliana* (Zhou et al. 2008). Multiple Kernel Learning has been applied to the identification of circularRNA, a type of lncRNA, in humans, which can identify them with high accuracy in de novo assembled transcriptomes (Pan and Xiong 2015).

Gene selection from expression data is a problem in which ML methods can be used naturally. Given an expression dataset that usually comprises thousands of genes, the goal here is to select a handful of relevant genes associated with

a specific condition of interest, e.g., a disease or a treatment. A common ML-based approach for gene selection from expression datasets is variable ranking, in which genes (variables) are prioritized according to a value derived from the applied classification algorithm. This value is a proxy for the importance or relevance of each gene among the whole dataset. In this way, genes at the top of the rank are more relevant to the condition of interest, e.g., healthy/diseased tissue, treated/untreated tissue, and genes at the lower positions are redundant and less relevant. Following this approach, ML algorithms such as RFs, SVMs, and decision trees have been used with microarray data to select subsets of cancer-related genes which can be used as markers in diagnosis (Diaz-Uriarte and Alvarez de Andres 2006; Horng et al. 2009; Guyon et al. 2002).

Although most of the proposed ML-based gene selection methods are tested in cancer expression datasets, some studies have applied similar approaches to plants using gene expression data from microarrays. An SVM with Recursive Feature Elimination (SVM-RFE) and a Radial Basis Function (RBF) was used to identify four genes related to resistance to tungro disease in rice (Ren et al. 2010). This was a modification of the application of the same technique to cancer (Guyon et al. 2002). A caveat in this study is the small dataset used (21 samples), as the amount of data for training is a decisive factor to get revealing results in ML. A further study refined the same SVM-RFE approach to identify genes related to drought resistance in *A. thaliana* (Liang et al. 2011). Although authors of this study used a dataset with only 22 samples, they mitigated the small sample size effect by implementing a Leave One Out Cross Validation (LOOCV) scheme to select the training dataset and bootstrapping strategy to iterate the variable ranking process. In such a way, a subset of ten genes were identified, seven of which have previous biological information that links them to processes involved in drought resistance. ML and GCN were combined into the R package "machine learning-based differential network analysis" (mlDNA), which implements a two-phase ML method for selecting genes from expression data. In the first phase, the method identifies and discards irrelevant genes from the dataset using an RF classifier with the Positive Sample only Learning algorithm (PSoL), a technique that discriminates positive from negative data after using only positive samples for training. The second phase involves the construction of GCNs from the filtered genes, the extraction of topological features from the GCNs, and an RF algorithm to select the candidate genes based on the extracted features. This approach proved to successfully select candidate genes in *A. thaliana* responding to drought, cold, heat, wound, and genotoxic stress conditions (Huang et al. 2011).

2.4 Case/Examples of Transcriptomics in Non-model Plants

Perhaps the most notable quality of transcriptomics is the possibility of producing robust amounts of data for a reduced representation of the genome, which is of importance in non-model plant species and species with complex genomes. This quality allows for a diverse series of biological questions to be asked and for which answers can be obtained. In this section we will exemplify the most relevant uses of recent transcriptomics studies.

2.4.1 Construction of Improved Transcripts Catalogs

Although, in principle, transcriptomic studies derived from RNA-Seq do not require any prior genetic information, it is true that having a high-quality reference transcriptome undoubtedly favors high-quality research. Current assembly tools and sequencing technologies have advanced our capacity to produce de novo assemblies. In constructing high-quality transcriptomes for polyploid (allopolyploid) species, where two or more sub-genomes are present, one particular challenge is the identification of homeologous

copies of the same genes which tend to be highly similar and difficult to separate in a de novo assembly. Classical assemblers such as SOAPdenovo-Trans Trinity and TransAByss have been tested for this task. This is exemplified in the study by (Chopra et al. 2014) aiming at reconstructing the transcriptome of tetraploid and diploid peanut species, using RNA-Seq data. After examining several variables including contig length and number, results showed that Trinity and TransAByss performed in a similar way for the diploid species, while Trinity performed better for the tetraploid genotype. In addition, the transcriptome produced for the tetraploid genotype almost doubled in number of contigs, total size and transcript N50 compared to the existing resources. It also produced at least 40% more full-length sequences.

Others have searched to develop specific software to tackle the problem. Such is the case of the software HomeoSplitter which takes into consideration the elevated rates of heterozygosity of certain contigs (alleles) to target possible homeoalleles. Once identified, the software uses a likelihood model-based method to disentangle the mixed alleles taking into consideration their expression levels. For durum wheat (*Triticum turgidum*) HomeoSplitter showed capacity to separate homeologous sequences, as assessed by comparison to the diploid progenitors, and allowed to recover a greater number of SNPs for the population genotyped (Ranwez et al. 2013).

From the sequencing-and-assembly point of view, this issue has been approached through the use of normalized libraries, which increases the likelihood of seeing rare or less abundant transcript, and the use of single-molecule long read sequencing technologies, which can produce near complete transcript sequences represented in a single-sequencing read. The protocol called Iso-Seq has been applied to several crop species, including sorghum (Abdel-Ghany et al. 2016), maize (Wang et al. 2016), cotton (Wang et al. 2018b), coffee (Cheng et al. 2017), *Salvia miltiorrhiza* (Xu et al. 2015), grape wine (Minio et al. 2019), the Chinese herb *Astragalus membranaceus* (Li et al. 2017a), *Arabidopsis pumila* (Yang et al. 2018b), the shrub *Zanthoxylum bungeanum* (Tian et al. 2018), the giant timber bamboo native to China (Zhang et al. 2018), wild strawberry (Li et al. 2017b), and the highly complex sugarcane (Hoang et al. 2017). Iso-Seq has been shown to recover full-length isoforms, which was not possible with short-read technologies, but also it has allowed the detection of alternative start sites, alternative splicing and alternative polyadenylation (Zhao et al. 2019). In the case of sugarcane, Iso-Seq was further complemented with short RNA-Seq reads in order to correct errors present in long reads. The same dataset also served to compare the transcriptomes created by the hybrid approach and a de novo approach based solely on RNA-Seq reads. The hybrid transcriptome recovered more full-length transcripts, with a longer N50, more ORFs and predicted transcripts, and higher average length of the largest 1000 proteins, compared to the de novo contigs. Importantly, RNA-Seq covered more gene content, and more RNA classes than Iso-Seq, which was attributed to the greater sequencing depth (Hoang et al. 2017).

Oxford Nanopore Technologies (ONT) have a platform option that allows for the direct sequencing of RNA molecules, which in addition to producing full-length transcript sequences, study of alternative polyadenylation and splice and start sites, reveals the status of RNA modifications, and could revolutionize the transcriptomics field (Hussain 2018). This approach is still very recent and has not yet been applied to many plant species. Direct RNA sequencing was performed on seeds of soybean to quantify transcript degradation as a proxy of seed viability (Fleming et al. 2018). Eukaryotic transcripts are usually modified on their 5′-end by the addition of a 7-methylguanylate (m^7G) cap which protects mRNA from decay and has several implications in mRNA-downstream processes. However, a recent study, using direct RNA sequencing, showed that in *A. thaliana*, up to 5% of the transcripts of several thousand genes have instead a NAD+ cap (Zhang et al. 2019a), an RNA modification that had been reported before in bacteria (Chen et al. 2009), yeast (Walters et al. 2017), and humans (Jiao et al. 2017).

Overall, despite current advances in the construction of de novo transcriptomes, there is still

room for improvement in assemblers tailored to polyploid genomes. Also, given the current rate of innovation in high-throughput sequencing, and provided a decrease in costs, the construction of novel transcriptomes through the use of long RNA molecules are expected to increase rapidly.

2.4.2 Populations Mapping

Transcriptomics can also be used to identify polymorphisms to map populations of interest. Two alternative strategies are often followed: In the first, the genetic variants are identified from transcriptomic data, from a diverse group of individuals. The variants identified are then used to design probes to test DNA samples from the same or an alternative, bigger, population. Contrary to the classic DNA mapping studies, this strategy increases the probability of identifying causal mutations given that the majority of the selected variants will be located within coding sequences. This is specially the case of species with big genomes and a high percentage of repetitive sequences which, for mapping studies, require a considerable number of markers to increase the probability of having a significant association. Markers, particularly SNP and SSR, derived from transcriptomic data have been produced for different crops including, but not limited to soybean (Guo et al. 2018), sugarcane (Bundock et al. 2009), grasspea (Hao et al. 2017), peanut (Chopra et al. 2015), and oilseed rape (Trick et al. 2009). More recently, and through the implementation of the Bulk Segregant RNA-Seq analyses (BSR-Seq) principle, which requires the formation of pooled samples contrasting for the phenotype of interest, markers linked to traits of interest have been mapped in crop species such as wheat (Wang et al. 2017; Ramirez-Gonzalez et al. 2015; Wu et al. 2018) and Chinese cabbage (Huang et al. 2017b).

In the second strategy, transcriptomics data is produced for a biparental population, and the markers identified (SNP markers) are directly used for construction of genetic maps. The value of these maps lies in the fact that "unlike sequence assembly, linkage analysis is essentially unaffected by allopolyploidy and repeated sequences as long as homeologous recombination is rare and genome-specific alleles can be identified" (reviewed in McKay and Leach 2011). This strategy, to the best of our knowledge, has been only used in the tetraploid *Brassica napus* (oilseed rape) (Bancroft et al. 2011). In this case, twin genetic maps were constructed for the two progenitor species (*B. oleracea* and *B. rapa*) of the modern *B. napus* genotypes, which also served as parents for the population tested. These genetic maps were next aligned to the existing genome of *B. napus* and that of *A. thaliana*. The whole strategy allowed to identify genome rearrangements between *B. oleracea* and *B. rapa* and therefore helped to refine the existing assemblies for these species. Likewise, it helped to pinpoint genomic regions involved in the recent breeding history of the crop. Considering these implications and the urgent necessity of genomic tools to tackle polyploid genomes, it is expected that linkage maps derived from transcriptomic data will be on the rise.

2.4.3 Stress-Related Studies

As sessile organisms, plants must deal with a variety of environmental conditions that can impact on their potential for growth and reproduction. In order to study the molecular mechanisms underlying the response to such conditions plant transcriptomics is being widely used. The most common approach consists of comparing gene expression levels of a specific genotype under a control and a stress-induced treatment. Oftentimes, contrasting genotypes (tolerant and susceptible) for the trait of interest are used. By identifying the changes in gene expression between control and treatment conditions, it is possible to determine the mRNAs activated by the stress under consideration. This in turn allows for exploring the mRNAs that are differentially expressed among the genotypes selected (tolerant vs. susceptible). Following this approach, it has been possible to study the molecular regulation of salt stress tolerance in cotton (Zhang et al. 2016a), the roles of the photosynthetic system

during drought in upland rice (Zhang et al. 2016b), the molecular mechanisms driving copper stress tolerance in grapevine (Leng et al. 2015), the mechanisms for lipid accumulation in response to nitrogen deprivation in the green algae *Chlamydomonas reinhardtii* (Park et al. 2015), the molecular responses underlying drought tolerance in sugarcane (Pereira-Santana et al. 2017; Belesini et al. 2017), just to mention a few.

Perhaps, one of the most studied traits through comparative transcriptomic is drought. When "drought" and "RNA-Seq" are used as keywords in PubMed, 217 different titles, excluding reviews, show up as a result. Studies have been performed on nearly every major crop (Zhang et al. 2014; Chen et al. 2016; Divya Bhanu et al. 2016; Mofatto et al. 2016), but also on non-major crops and other plants whose original habitat are water-deprived locations and thus can contribute to better understanding of the physiological bases of this condition (Gross et al. 2013; Yang et al. 2015; Li et al. 2015). In polyploids, the challenge resides on having a high-quality reference transcriptome that allows to distinguish among isoforms derived from different sub-genomes. In fact, in hexaploid wheat, where different genomic resources have been recently developed (Pearce et al. 2015), it has been found that a large proportion of wheat homeologs exhibited expression partitioning under normal and abiotic stresses, indicating a specialized gene expression coordination among genomes.

2.4.4 Phylogenomics

Phylogenomics is a new biological discipline focusing on the resolution of relationships among taxa and the reconstruction of evolutionary histories through the use of genomic data. It involves the analysis of entire genomes, transcriptomes, or specific sequences that can be targeted (Yu et al. 2018) through the mining of already published information (Washburn et al. 2017).

In order to resolve relationships among species, phylogenomics relies heavily on the identification of single-copy genes to reduce the possibility of paralogy and thus limiting to conclusions based solely on orthologous genes. However, information on single-copy genes is difficult to obtain especially for non-model, polyploid species, where the entire genome is expected to be duplicated. Chloroplast genes are often targeted for phylogenomics; however, this part of the plant genome has its own problems such as a low recombinant nature, and thus low polymorphism levels, exclusive maternal inheritance, and these genes are subject to processes such as chloroplast capture and hybrid speciation which reduce its resolution capacity. Still, due to its high-throughput nature, transcriptomics offers the possibility to mine for nuclear single-copy markers in a rich set of genic sources. This is even possible in the case of polyploids and despite their repetitive nature. Due to evolutionary mechanisms such as gene conversion and loss, the number of retained duplicates in polyploids decreases over the time, allowing single-copy signals (coding and non-coding sequences) to arise (Wen et al. 2015). In the case of ferns, for example, which have a long history of polyploidy, 20 new nuclear regions spanning ten coding sequences have been identified by comparative transcriptomics which has increased significantly the taxonomic resolution across these group of plants (Rothfels et al. 2013).

Comparative transcriptomics can also contribute to detect and characterize polyploidy speciation. Although ancient polyploidy could be reconstructed through the comparison of high-quality, chromosome-level genomes, the lack of high-quality assemblies for the vast majority of polyploid species has positioned transcriptomics as a viable alternative. For this purpose, the rate of synonymous substitution (Ks), in coding sequences, derived from transcriptomics is widely used. This is possible because whole-genome duplications produce peaks in the cumulative distributions of pairwise Ks between paralogs within a genome. By evaluating the distribution of Ks among evolutionary lineages, it has been possible to better understand polyploidy speciation in the flax genus (Sveinsson et al. 2014), the evolution of gene families like CYP75 after the events of whole-genome duplication

(Zhang et al. 2019b), the redistribution of the seed plants in phylogenetic trees explaining the origin of angiosperms (Ran et al. 2018), the evolutionary patterns of agricultural traits in strawberry (Qiao et al. 2016), or the origin and early diversification of green (One Thousand Plant Transcriptomes Initiative 2019) and land plants (Wickett et al. 2014), among others.

2.5 Future Directions in the Field

Over the past decades, transcriptomics has seen a revolution. The technologies employed to produce expression data are nowadays much more efficient and with their regular decrease in costs, they are a realistic possibility even for small labs, and so it has become practical to be applied to non-model exotic plant species, and to perform more complex experimental designs. Nonetheless, the cost of sequencing is still not at reach for projects in which hundreds to thousands of samples need to be sequenced. This level of sequencing capacity is a reality for consortiums and greater collaborative efforts but not for smaller groups, which commonly have the possibility of greater access to genetically diverse samples but smaller budgets. Further decrease in library preparation and sequencing costs will ameliorate this though.

Technical advances have made it possible to directly sequence RNA molecules, and together with PCR-free protocols, they aid in eliminating potential sources of bias that could be introduced during library preparation. In addition to building comprehensive transcript catalogs, these advances will allow more reliable estimation of transcript abundances when it becomes affordable to sequence at higher depths of coverage. Recently published genome assemblies are increasingly resolving the different sequence haplotypes in organims with ploidy levels greater than one in these cases long-read RNA sequencing will allow the study of allele-specific expression with unprecedented levels of detail.

Along with this new technological capacity to produce data, the questions that may be answered with transcriptomics-based strategies have also matured. However, for many of these questions, their answers are limited by the available bioinformatic software. For example, all the efforts that have been made to confidently identify orthologous genes and in general to filter out the noise caused by polyploidy are encouraging because, among other reasons, this has increased our understanding of complex genomes. Nonetheless, only a handful of genes or a small portion of the transcriptomes are used for these purposes. It is then reasonable to believe that further efforts in software development are necessary to truly take advantage of the level of information being produced in transcriptomics studies. A similar situation happens with all the studies aiming at better understanding of specific phenomena (e.g., stress-related studies) that after producing high-quality, robust data are still left with lists of hundreds to thousands of differentially expressed genes, from which it is difficult to define the key players for the process under study. Perhaps this type of studie could benefit from the integration of different OMICs approaches to the same problems, with a more integrative approach which requires further advances in tool development, for instance including machine learning algorithms, necessary to mine for the most relevant transcripts.

Overall, we can confidently say that the last decade has been a defining one for plant transcriptomics thanks to the greater access to sequencing data. However, the same breakthrough has yet to impact data analyses and storage. Our data processing capabilities are being surpassed by our capacity to produce data, and it is imperative to face this challenge if we want to further increase our ability to address the challenges posed by climate change, speed up the efforts to breed crop plants, and deepen our understanding of the history of evolution of plants.

References

Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy AS (2016) A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun 7:11706. https://doi.org/10.1038/ncomms11706

Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22(13):1600–1607. https://doi.org/10.1093/bioinformatics/btl140

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106

Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2):166–169. https://doi.org/10.1093/bioinformatics/btu638

Angermueller C, Parnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol 12(7):878. https://doi.org/10.15252/msb.20156651

Arita M (2005) Scale-freeness and biological networks. J Biochem 138(1):1–4. https://doi.org/10.1093/jb/mvi094

Asamizu E, Nakamura Y, Sato S, Fukuzawa H, Tabata S (1999) A large scale structural analysis of cDNAs in a unicellular green alga, Chlamydomonas reinhardtii. I. Generation of 3433 non-redundant expressed sequence tags. DNA Res 6(6):369–373

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29. https://doi.org/10.1038/75556

Baccarella A, Williams CR, Parrish JZ, Kim CC (2018) Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. BMC Bioinformatics 19(1):423. https://doi.org/10.1186/s12859-018-2445-2

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJ (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics 7:246. https://doi.org/10.1186/1471-2164-7-246

Ballouz S, Verleyen W, Gillis J (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics 31(13):2123–2130. https://doi.org/10.1093/bioinformatics/btv118

Bancroft I, Morgan C, Fraser F, Higgins J, Wells R, Clissold L, Baker D, Long Y, Meng J, Wang X, Liu S, Trick M (2011) Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat Biotechnol 29(8):762–766. https://doi.org/10.1038/nbt.1926

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, dePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, Ashton NW, Axtell MJ, Barker E, Barker MS, Bennetzen JL, Bonawitz ND, Chapple C, Cheng C, Correa LG, Dacre M, DeBarry J, Dreyer I, Elias M, Engstrom EM, Estelle M, Feng L, Finet C, Floyd SK, Frommer WB, Fujita T, Gramzow L, Gutensohn M, Harholt J, Hattori M, Heyl A, Hirai T, Hiwatashi Y, Ishikawa M, Iwata M, Karol KG, Koehler B, Kolukisaoglu U, Kubo M, Kurata T, Lalonde S, Li K, Li Y, Litt A, Lyons E, Manning G, Maruyama T, Michael TP, Mikami K, Miyazaki S, Morinaga S, Murata T, Mueller-Roeber B, Nelson DR, Obara M, Oguri Y, Olmstead RG, Onodera N, Petersen BL, Pils B, Prigge M, Rensing SA, Riano-Pachon DM, Roberts AW, Sato Y, Scheller HV, Schulz B, Schulz C, Shakirov EV, Shibagaki N, Shinohara N, Shippen DE, Sorensen I, Sotooka R, Sugimoto N, Sugita M, Sumikawa N, Tanurdzic M, Theissen G, Ulvskov P, Wakazuki S, Weng JK, Willats WW, Wipf D, Wolf PG, Yang L, Zimmer AD, Zhu Q, Mitros T, Hellsten U, Loque D, Otillar R, Salamov A, Schmutz J, Shapiro H, Lindquist E, Lucas S, Rokhsar D, Grigoriev IV (2011) The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science 332(6032):960–963. https://doi.org/10.1126/science.1203810

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113. https://doi.org/10.1038/nrg1272

Belesini AA, Carvalho FMS, Telles BR, de Castro GM, Giachetto PF, Vantini JS, Carlin SD, Cazetta JO, Pinheiro DG, Ferro MIT (2017) De novo transcriptome assembly of sugarcane leaves submitted to prolonged water-deficit stress. Genet Mol Res 16(2):gmr16028845. https://doi.org/10.4238/gmr16028845

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57(1):289–300

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA,

Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59. https://doi.org/10.1038/nature07517

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borodina T, Adjaye J, Sultan M (2011) A strand-specific library preparation protocol for RNA sequencing. Methods Enzymol 500:79–98. https://doi.org/10.1016/B978-0-12-385118-5.00005-0

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527. https://doi.org/10.1038/nbt.3519

Britto-Kido Sde A, Ferreira Neto JR, Pandolfi V, Marcelino-Guimaraes FC, Nepomuceno AL, Vilela Abdelnoor R, Benko-Iseppon AM, Kido EA (2013) Natural antisense transcripts in plants: a review and identification in soybean infected with Phakopsora pachyrhizi SuperSAGE library. ScientificWorldJournal 2013:219798. https://doi.org/10.1155/2013/219798

Broido AD, Clauset A (2019) Scale-free networks are rare. Nat Commun 10(1):1017. https://doi.org/10.1038/s41467-019-08746-5

Brown JW, Calixto CP, Zhang R (2017) High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. New Phytol 213(2):525–530. https://doi.org/10.1111/nph.14208

Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Leigh ND, Kuo TH, Davis FG, Bateman J, Bryant S, Guzikowski AR, Tsai SL, Coyne S, Ye WW, Freeman RM Jr, Peshkin L, Tabin CJ, Regev A, Haas BJ, Whited JL (2017) A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. Cell Rep 18(3):762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol J 7(4):347–354. https://doi.org/10.1111/j.1467-7652.2009.00401.x

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun 8:16027. https://doi.org/10.1038/ncomms16027

Cai M, Lin J, Li Z, Lin Z, Ma Y, Wang Y, Ming R (2020) Allele specific expression of Dof genes responding to hormones and abiotic stresses in sugarcane. PLoS One 15(1):e0227716. https://doi.org/10.1371/journal.pone.0227716

Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 7:40. https://doi.org/10.1186/1471-2164-7-40

Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T (2015) Tools and best practices for data processing in allelic expression analysis. Genome Biol 16:195. https://doi.org/10.1186/s13059-015-0762-6

Chen YG, Kowtoniuk WE, Agarwal I, Shen Y, Liu DR (2009) LC/MS analysis of cellular RNA reveals NAD-linked RNA. Nat Chem Biol 5(12):879–881. https://doi.org/10.1038/nchembio.235

Chen W, Yao Q, Patil GB, Agarwal G, Deshmukh RK, Lin L, Wang B, Wang Y, Prince SJ, Song L, Xu D, An YC, Valliyodan B, Varshney RK, Nguyen HT (2016) Identification and comparative analysis of differential gene expression in soybean leaf tissue under drought and flooding stress revealed by RNA-Seq. Front Plant Sci 7:1044. https://doi.org/10.3389/fpls.2016.01044

Cheng B, Furtado A, Henry RJ (2017) Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. GigaScience 6(11):1–13. https://doi.org/10.1093/gigascience/gix086

Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, Burow MD (2014) Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (*Arachis* spp.) RNA-Seq data. PLoS One 9(12):e115055. https://doi.org/10.1371/journal.pone.0115055

Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, Wilkins TA, Baring MR, Puppala N, Chamberlin KD, Burow MD (2015) Next-generation transcriptome sequencing, SNP discovery and validation in four market classes of peanut, *Arachis hypogaea* L. Mol Gen Genomics 290(3):1169–1180. https://doi.org/10.1007/s00438-014-0976-4

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM (2008) Stem cell transcriptome

profiling via massive-scale mRNA sequencing. Nat Methods 5(7):613–619. https://doi.org/10.1038/nmeth.1223

Collins LJ, Biggs PJ, Voelckel C, Joly S (2008) An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Inform 21:3–14

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13. https://doi.org/10.1186/s13059-016-0881-8

Contreras-Moreira B, Cantalapiedra CP, Garcia-Pereira MJ, Gordon SP, Vogel JP, Igartua E, Casas AM, Vinuesa P (2017) Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. Front Plant Sci 8:184. https://doi.org/10.3389/fpls.2017.00184

Cubillos FA, Coustham V, Loudet O (2012) Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. Curr Opin Plant Biol 15(2):192–198. https://doi.org/10.1016/j.pbi.2012.01.005

Das S, Meher PK, Rai A, Bhar LM, Mandal BN (2017) Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: an application to aluminum stress in soybean (Glycine max L.). PLoS One 12(1):e0169605. https://doi.org/10.1371/journal.pone.0169605

Delseny M, Cooke R, Raynal M, Grellet F (1997) The *Arabidopsis thaliana* cDNA sequencing projects. FEBS Lett 405(2):129–132

DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28(11):1530–1532. https://doi.org/10.1093/bioinformatics/bts196

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498. https://doi.org/10.1038/ng.806

Diaz-Uriarte R, Alvarez de Andres S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7:3. https://doi.org/10.1186/1471-2105-7-3

Didion JP, Martin M, Collins FS (2017) Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ 5:e3720. https://doi.org/10.7717/peerj.3720

Divya Bhanu B, Ulaganathan K, Shanker AK, Desai S (2016) RNA-seq analysis of irrigated vs. water stressed transcriptomes of Zea mays Cultivar Z59. Front Plant Sci 7:239. https://doi.org/10.3389/fpls.2016.00239

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Syst Biol 1:24. https://doi.org/10.1186/1752-0509-1-24

Eichner J, Zeller G, Laubinger S, Ratsch G (2011) Support vector machines-based identification of alternative splicing in *Arabidopsis thaliana* from whole-genome tiling arrays. BMC Bioinformatics 12:55. https://doi.org/10.1186/1471-2105-12-55

Elo LL, Jarvenpaa H, Oresic M, Lahesmaa R, Aittokallio T (2007) Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics 23(16):2096–2103. https://doi.org/10.1093/bioinformatics/btm309

Ferreira SS, Hotta CT, Poelking VG, Leite DC, Buckeridge MS, Loureiro ME, Barbosa MH, Carneiro MS, Souza GM (2016) Co-expression network analysis reveals transcription factors associated to cell wall biosynthesis in sugarcane. Plant Mol Biol 91(1–2):15–35. https://doi.org/10.1007/s11103-016-0434-2

Fleming MB, Patterson EL, Reeves PA, Richards CM, Gaines TA, Walters C (2018) Exploring the fate of mRNA in aging seeds: protection, destruction, or slow decay? J Exp Bot 69(18):4309–4321. https://doi.org/10.1093/jxb/ery215

Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods 6(11 Suppl):S6–S12. https://doi.org/10.1038/nmeth.1376

Florea LD, Salzberg SL (2013) Genome-guided transcriptome assembly in the age of next-generation sequencing. IEEE/ACM Trans Comput Biol Bioinformatics 10(5):1234–1240

Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT (2015) Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat Biotechnol 33(3):243–246. https://doi.org/10.1038/nbt.3172

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu Y, van der Knaap E, Huang S, Klee HJ, Giovannoni JJ, Fei Z (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51:1044. https://doi.org/10.1038/s41588-019-0410-2

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ,

Turner DJ (2018) Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods 15(3):201–206. https://doi.org/10.1038/nmeth.4577

Garcia-Ortega LF, Martinez O (2015) How many genes are expressed in a transcriptome? Estimation and results for RNA-Seq. PLoS One 10(6):e0130262. https://doi.org/10.1371/journal.pone.0130262

Garg R, Jain M (2013) RNA-Seq for transcriptome analysis in non-model plants. Methods Mol Biol 1069:43–58. https://doi.org/10.1007/978-1-62703-613-9_4

Gene Ontology Consortium T (2017) Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res 45(D1):D331–D338. https://doi.org/10.1093/nar/gkw1108

Giorgi FM, Del Fabbro C, Licausi F (2013) Comparative study of RNA-seq- and microarray-derived coexpression networks in *Arabidopsis thaliana*. Bioinformatics 29(6):717–724. https://doi.org/10.1093/bioinformatics/btt053

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652. https://doi.org/10.1038/nbt.1883

Gross SM, Martin JA, Simpson J, Abraham-Juarez MJ, Wang Z, Visel A (2013) De novo transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana. BMC Genomics 14:563. https://doi.org/10.1186/1471-2164-14-563

Guo Y, Su B, Tang J, Zhou F, Qiu LJ (2018) Gene-based SNP identification and validation in soybean using next-generation transcriptome sequencing. Mol Gen Genomics 293(3):623–633. https://doi.org/10.1007/s00438-017-1410-5

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28(5):503–510. https://doi.org/10.1038/nbt.1633

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422. https://doi.org/10.1023/A:1012487302797

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31(19):5654–5666. https://doi.org/10.1093/nar/gkg770

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8(8):1494–1512. https://doi.org/10.1038/nprot.2013.084

Hackett JB, Lu Y (2017) Whole-transcriptome RNA-seq, gene set enrichment pathway analysis, and exon coverage analysis of two plastid RNA editing mutants. Plant Signal Behav 12(5):e1312242. https://doi.org/10.1080/15592324.2017.1312242

Hale MC, McCormick CR, Jackson JR, Dewoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery. BMC Genomics 10:203. https://doi.org/10.1186/1471-2164-10-203

Hao X, Yang T, Liu R, Hu J, Yao Y, Burlyaeva M, Wang Y, Ren G, Zhang H, Wang D, Chang J, Zong X (2017) An RNA sequencing transcriptome analysis of grasspea (Lathyrus sativus L.) and development of SSR and KASP markers. Front Plant Sci 8:1873. https://doi.org/10.3389/fpls.2017.01873

Harbers M, Carninci P (2005) Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2(7):495–502. https://doi.org/10.1038/nmeth768

Hoang NV, Furtado A, Mason PJ, Marquardt A, Kasirajan L, Thirugnanasambandam PP, Botha FC, Henry RJ (2017) A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genomics 18(1):395. https://doi.org/10.1186/s12864-017-3757-8

Hollender CA, Kang C, Darwish O, Geretz A, Matthews BF, Slovin J, Alkharouf N, Liu Z (2014) Floral transcriptomes in woodland strawberry uncover developing receptacle and anther gene networks. Plant Physiol 165(3):1062–1075. https://doi.org/10.1104/pp.114.237529

Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, Edger PP, Altman NS, Pires JC, Leebens-Mack JH, dePamphilis CW (2016) Selecting superior de novo transcriptome assemblies: lessons learned by leveraging the best plant genome. PLoS One 11(1):e0146062. https://doi.org/10.1371/journal.pone.0146062

de Hoon M, Hayashizaki Y (2008) Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. BioTechniques 44(5):627–628., 630, 632. https://doi.org/10.2144/000112802

Horng JT, Wu LC, Liu BJ, Kuo JL, Kuo WH, Zhang JJ (2009) An expert system to classify microarray gene expression data using gene selection by decision tree. Expert Syst Appl 36(5):9072–9081. https://doi.org/10.1016/j.eswa.2008.12.037

Horvath S, Dong J (2008) Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4(8):e1000117. https://doi.org/10.1371/journal.pcbi.1000117

Hrdlickova R, Toloue M, Tian B (2017) RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA 8(1):e1364. https://doi.org/10.1002/wrna.1364

Huang Q, Lin B, Liu H, Ma X, Mo F, Yu W, Li L, Li H, Tian T, Wu D, Shen F, Xing J, Chen ZN (2011) RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. PLoS One 6(10):e26168. https://doi.org/10.1371/journal.pone.0026168

Huang J, Vendramin S, Shi L, McGinnis KM (2017a) Construction and optimization of a large gene coexpression network in maize using RNA-Seq data. Plant Physiol 175(1):568–583. https://doi.org/10.1104/pp.17.00825

Huang Z, Peng G, Liu X, Deora A, Falk KC, Gossen BD, McDonald MR, Yu F (2017b) Fine mapping of a clubroot resistance gene in Chinese cabbage using SNP markers identified from bulked segregant RNA sequencing. Front Plant Sci 8:1448. https://doi.org/10.3389/fpls.2017.01448

Hussain S (2018) Native RNA-sequencing throws its hat into the transcriptomics ring. Trends Biochem Sci 43(4):225–227. https://doi.org/10.1016/j.tibs.2018.02.007

Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S (2012) Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics 28(12):1592–1597. https://doi.org/10.1093/bioinformatics/bts245

Illumina (2010) Illumina sequencing technology

Illumina (2017) TruSeq stranded total RNA - reference guide

Imadi SR, Kazi AG, Ahanger MA, Gucel S, Ahmad P (2015) Plant transcriptomics and responses to environmental stress: an overview. J Genet 94(3):525–537

Jain P, Krishnan NM, Panda B (2013) Augmenting transcriptome assembly by combining de novo and genome-guided tools. PeerJ 1:e133. https://doi.org/10.7717/peerj.133

Jen CH, Michalopoulos I, Westhead DR, Meyer P (2005) Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation. Genome Biol 6(6):R51. https://doi.org/10.1186/gb-2005-6-6-r51

Jenjaroenpun P, Wongsurawat T, Pereira R, Patumcharoenpol P, Ussery DW, Nielsen J, Nookaew I (2018) Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of Saccharomyces cerevisiae CEN.PK113-7D. Nucleic Acids Res 46(7):e38. https://doi.org/10.1093/nar/gky014

Jiao X, Doamekpor SK, Bird JG, Nickels BE, Tong L, Hart RP, Kiledjian M (2017) 5' end nicotinamide adenine dinucleotide cap in human cells promotes RNA decay through DXO-mediated deNADding. Cell 168(6):1015–1027.e1010. https://doi.org/10.1016/j.cell.2017.02.019

Jin H, Vacic V, Girke T, Lonardi S, Zhu JK (2008) Small RNAs and the regulation of cis-natural antisense transcripts in *Arabidopsis*. BMC Mol Biol 9:6. https://doi.org/10.1186/1471-2199-9-6

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44(D1):D457–D462. https://doi.org/10.1093/nar/gkv1070

Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, Stark TJ, Neuhaus EB, Dugan VG, Wentworth DE, Barnes JR (2018) Direct RNA sequencing of the coding complete influenza A virus genome. Sci Rep 8(1):14408. https://doi.org/10.1038/s41598-018-32615-8

Khanin R, Wit E (2006) How scale-free are biological networks. J Comput Biol 13(3):810–818. https://doi.org/10.1089/cmb.2006.13.810

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–360. https://doi.org/10.1038/nmeth.3317

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4

Kopylova E, Noe L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28(24):3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol 20(1):278. https://doi.org/10.1186/s13059-019-1910-1

Lamarre S, Frasse P, Zouine M, Labourdette D, Sainderichin E, Hu G, Le Berre-Anton V, Bouzayen M, Maza E (2018) Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size. Front Plant Sci 9:108. https://doi.org/10.3389/fpls.2018.00108

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559. https://doi.org/10.1186/1471-2105-9-559

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. https://doi.org/10.1038/nmeth.1923

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25. https://doi.org/10.1186/gb-2009-10-3-r25

Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles V (2006) Machine learning in bioinformatics. Brief Bioinform 7(1):86–112. https://doi.org/10.1093/bib/bbk007

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14(6):1085–1094. https://doi.org/10.1101/gr.1910904

Lee JH, Ang JK, Xiao X (2013) Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. RNA 19(6):725–732. https://doi.org/10.1261/rna.037903.112

Lembke CG, Nishiyama MY Jr, Sato PM, de Andrade RF, Souza GM (2012) Identification of sense and antisense transcripts regulated by drought in sugarcane. Plant Mol Biol 79(4–5):461–477. https://doi.org/10.1007/s11103-012-9922-1

Leng X, Jia H, Sun X, Shangguan L, Mu Q, Wang B, Fang J (2015) Comparative transcriptome analysis of grapevine in response to copper stress. Sci Rep 5:17749. https://doi.org/10.1038/srep17749

Leung MKK, Delong A, Alipanahi B, Frey BJ (2016) Machine learning in genomic medicine: a review of computational problems and data sets. Proc IEEE 104(1):176–197. https://doi.org/10.1109/Jproc.2015.2494198

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323. https://doi.org/10.1186/1471-2105-12-323

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11(5):473–483. https://doi.org/10.1093/bib/bbq015

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2010) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493–500. https://doi.org/10.1093/bioinformatics/btp692

Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ (2011a) Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proc Natl Acad Sci U S A 108(50):19867–19872. https://doi.org/10.1073/pnas.1113972108

Li W, Feng J, Jiang T (2011b) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol 18(11):1693–1707. https://doi.org/10.1089/cmb.2011.0171

Li S, Liberman LM, Mukherjee N, Benfey PN, Ohler U (2013a) Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. Genome Res 23(10):1730–1739. https://doi.org/10.1101/gr.149310.112

Li Y, Li-Byarlay H, Burns P, Borodovsky M, Robinson GE, Ma J (2013b) TrueSight: a new algorithm for splice junction detection using RNA-seq. Nucleic Acids Res 41(4):e51. https://doi.org/10.1093/nar/gks1311

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol 15(12):553. https://doi.org/10.1186/s13059-014-0553-5

Li H, Yao W, Fu Y, Li S, Guo Q (2015) De novo assembly and discovery of genes that are involved in drought tolerance in Tibetan Sophora moorcroftiana. PLoS One 10(1):e111054. https://doi.org/10.1371/journal.pone.0111054

Li J, Harata-Lee Y, Denton MD, Feng Q, Rathjen JR, Qu Z, Adelson DL (2017a) Long read reference genome-free reconstruction of a full-length transcriptome from Astragalus membranaceus reveals transcript variants involved in bioactive compound biosynthesis. Cell Discov 3:17031. https://doi.org/10.1038/celldisc.2017.31

Li Y, Dai C, Hu C, Liu Z, Kang C (2017b) Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. Plant J 90(1):164–176. https://doi.org/10.1111/tpj.13462

Li Y, Wei W, Feng J, Luo H, Pi M, Liu Z, Kang C (2018) Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina- and SMRT-based RNA-seq datasets. DNA Res 25:61. https://doi.org/10.1093/dnares/dsx038

Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D (2011) Prediction of drought-resistant genes in *Arabidopsis thaliana* using SVM-RFE. PLoS One 6(7):e21750. https://doi.org/10.1371/journal.pone.0021750

Liang YH, Cai B, Chen F, Wang G, Wang M, Zhong Y, Cheng ZM (2014) Construction and validation of a gene co-expression network in grapevine (Vitis vinifera. L.). Hortic Res 1:14040. https://doi.org/10.1038/hortres.2014.40

Liao Y, Smyth GK, Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41(10):e108. https://doi.org/10.1093/nar/gkt214

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321–332. https://doi.org/10.1038/nrg3920

Liesecke F, Daudu D, Duge de Bernonville R, Besseau S, Clastre M, Courdavault V, de Craene JO, Creche J, Giglioli-Guivarc'h N, Glevarec G, Pichon O, Duge de Bernonville T (2018) Ranking genome-wide correlation measurements improves microarray and

RNA-seq based global and targeted co-expression networks. Sci Rep 8(1):10885. https://doi.org/10.1038/s41598-018-29077-3

Lima-Mendez G, van Helden J (2009) The powerful law of the power law and other myths in network biology. Mol BioSyst 5(12):1482–1493. https://doi.org/10.1039/b908681a

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133(3):523–536. https://doi.org/10.1016/j.cell.2008.03.029

Liu Y, Zhou J, White KP (2014) RNA-seq differential expression studies: more sequence or more replication? Bioinformatics 30(3):301–304. https://doi.org/10.1093/bioinformatics/btt688

Liu X, Mei W, Soltis PS, Soltis DE, Barbazuk WB (2017) Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. Mol Ecol Resour 17(6):1243–1256. https://doi.org/10.1111/1755-0998.12670

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550. https://doi.org/10.1186/s13059-014-0550-8

Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK, Zhou J (2007) Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics 8:299. https://doi.org/10.1186/1471-2105-8-299

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1(1):18. https://doi.org/10.1186/2047-217X-1-18

Ma HM, Schulze S, Lee S, Yang M, Mirkov E, Irvine J, Moore P, Paterson A (2004) An EST survey of the sugarcane transcriptome. Theor Appl Genet 108(5):851–863. https://doi.org/10.1007/s00122-003-1510-y

Ma X, Zhao H, Xu W, You Q, Yan H, Gao Z, Su Z (2018) Co-expression gene network analysis and functional module identification in bamboo growth and development. Front Genet 9:574. https://doi.org/10.3389/fgene.2018.00574

Ma Y, Liu M, Stiller J, Liu C (2019) A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication. BMC Genomics 20(1):12. https://doi.org/10.1186/s12864-018-5357-7

Magalhaes JV, Liu J, Guimaraes CT, Lana UG, Alves VM, Wang YH, Schaffert RE, Hoekenga OA, Pineros MA, Shaff JE, Klein PE, Carneiro NP, Coelho CM, Trick HN, Kochian LV (2007) A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. Nat Genet 39(9):1156–1161. https://doi.org/10.1038/ng2074

Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. Mol Pharm 13(5):1445–1454. https://doi.org/10.1021/acs.molpharmaceut.5b00982

Mao R, Raj Kumar PK, Guo C, Zhang Y, Liang C (2014) Comparative analyses between retained introns and constitutively spliced introns in *Arabidopsis thaliana* using random forest and support vector machine. PLoS One 9(8):e104049. https://doi.org/10.1371/journal.pone.0104049

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18(9):1509–1517. https://doi.org/10.1101/gr.079558.108

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res 22(6):1184–1195. https://doi.org/10.1101/gr.134106.111

Masalia RR, Bewick AJ, Burke JM (2017) Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. PLoS One 12(7):e0182289. https://doi.org/10.1371/journal.pone.0182289

Matsumura H, Ito A, Saitoh H, Winter P, Kahl G, Reuter M, Kruger DH, Terauchi R (2005) SuperSAGE. Cell Microbiol 7(1):11–18. https://doi.org/10.1111/j.1462-5822.2004.00478.x

McCarthy DJ, Chen Y, Gordon KS (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40(10):4288–4297

McKay JK, Leach JE (2011) Linkage illuminates a complex genome. Nat Biotechnol 29(8):717–718. https://doi.org/10.1038/nbt.1945

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303. https://doi.org/10.1101/gr.107524.110

Meyer PE, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol 2007:79879. https://doi.org/10.1155/2007/79879

Meyer PE, Lafitte F, Bontempi G (2008) minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 9:461. https://doi.org/10.1186/1471-2105-9-461

Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18(5):851–869. https://doi.org/10.1093/bib/bbw068

Minio A, Massonnet M, Figueroa-Balderas R, Vondras AM, Blanco-Ulate B, Cantu D (2019) Iso-Seq allows genome-independent transcriptome profiling of grape berry development. G3 (Bethesda) 9(3):755–767. https://doi.org/10.1534/g3.118.201008

Mofatto LS, Carneiro Fde A, Vieira NG, Duarte KE, Vidal RO, Alekcevetch JC, Cotta MG, Verdeil JL, Lapeyre-Montes F, Lartaud M, Leroy T, De Bellis F, Pot D, Rodrigues GC, Carazzolle MF, Pereira GA, Andrade AC, Marraccini P (2016) Identification of candidate genes for drought tolerance in coffee by high-throughput sequencing in the shoot apex of different Coffea arabica cultivars. BMC Plant Biol 16:94. https://doi.org/10.1186/s12870-016-0777-5

Moreton J, Izquierdo A, Emes RD (2015) Assembly, assessment, and availability of de novo generated eukaryotic transcriptomes. Front Genet 6:361. https://doi.org/10.3389/fgene.2015.00361

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7):621–628. https://doi.org/10.1038/nmeth.1226

Murphy KP (2012) Machine learning: a probabilistic perspective. Adaptive computation and machine learning series. MIT Press, Cambridge, MA

Mutwil M, Usadel B, Schutte M, Loraine A, Ebenhoh O, Persson S (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol 152(1):29–43. https://doi.org/10.1104/pp.109.145318

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320(5881):1344–1349. https://doi.org/10.1126/science.1158441

NuGen (n.d.) AnyDeplete

O'Neil D, Glowatz H, Schlumpberger M (2013) Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol Chapter 4:Unit 4.19. https://doi.org/10.1002/0471142727.mb0419s103

Oliver S (2000) Guilt-by-association goes global. Nature 403(6770):601–603. https://doi.org/10.1038/35001165

One Thousand Plant Transcriptomes Initiative (2019) One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574(7780):679–685. https://doi.org/10.1038/s41586-019-1693-2

Oono Y, Yazawa T, Kanamori H, Sasaki H, Mori S, Matsumoto T (2017) Genome-wide analysis of rice cis-natural antisense transcription under cadmium exposure using strand-specific RNA-Seq. BMC Genomics 18(1):761. https://doi.org/10.1186/s12864-017-4108-5

Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K, Matsubara K, Kikuchi S, Hayashizaki Y (2003) Antisense transcripts with rice full-length cDNAs. Genome Biol 5(1):R5. https://doi.org/10.1186/gb-2003-5-1-r5

Pan X, Xiong K (2015) PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. Mol BioSyst 11(8):2219–2226. https://doi.org/10.1039/c5mb00214a

Park JJ, Wang H, Gargouri M, Deshpande RR, Skepper JN, Holguin FO, Juergens MT, Shachar-Hill Y, Hicks LM, Gang DR (2015) The response of Chlamydomonas reinhardtii to nitrogen deprivation: a systems biology analysis. Plant J 81(4):611–624. https://doi.org/10.1111/tpj.12747

Parkinson J, Blaxter M (2009) Expressed sequence tags: an overview. Methods Mol Biol 533:1–12. https://doi.org/10.1007/978-1-60327-136-3_1

Patro R, Mount SM, Kingsford C (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 32(5):462–464. https://doi.org/10.1038/nbt.2862

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14(4):417–419. https://doi.org/10.1038/nmeth.4197

Pavy N, Paule C, Parsons L, Crow JA, Morency MJ, Cooke J, Johnson JE, Noumen E, Guillet-Claude C, Butterfield Y, Barber S, Yang G, Liu J, Stott J, Kirkpatrick R, Siddiqui A, Holt R, Marra M, Seguin A, Retzel E, Bousquet J, MacKay J (2005) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. BMC Genomics 6:144. https://doi.org/10.1186/1471-2164-6-144

Pavy N, Boyle B, Nelson C, Paule C, Giguere I, Caron S, Parsons LS, Dallaire N, Bedon F, Berube H, Cooke J, Mackay J (2008) Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses. New Phytol 180(4):766–786. https://doi.org/10.1111/j.1469-8137.2008.02615.x

Pearce S, Vazquez-Gross H, Herin SY, Hane D, Wang Y, Gu YQ, Dubcovsky J (2015) WheatExp: an RNA-seq expression database for polyploid wheat. BMC Plant Biol 15:299. https://doi.org/10.1186/s12870-015-0692-1

Peng Z, Gallo M, Tillman BL, Rowland D, Wang J (2016) Molecular marker development from transcript sequences and germplasm evaluation for cultivated peanut (Arachis hypogaea L.). Mol Gen Genomics 291(1):363–381. https://doi.org/10.1007/s00438-015-1115-6

Pereira-Santana A, Alvarado-Robledo EJ, Zamora-Briseno JA, Ayala-Sumuano JT, Gonzalez-Mendoza VM, Espadas-Gil F, Alcaraz LD, Castano E, Keb-Llanes MA, Sanchez-Teyer F, Rodriguez-Zapata LC (2017) Transcriptional profiling of sugarcane leaves and roots under progressive osmotic stress reveals a regulated coordination of gene expression in a spatio-temporal manner. PLoS One 12(12):e0189271. https://doi.org/10.1371/journal.pone.0189271

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33(3):290–295. https://doi.org/10.1038/nbt.3122

Pham GM, Newton L, Wiegert-Rininger K, Vaillancourt B, Douches DS, Buell CR (2017) Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated

potato. Plant J 92(4):624–637. https://doi.org/10.1111/tpj.13706

Pimentel H, Bray NL, Puente S, Melsted P, Pachter L (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods 14(7):687–690. https://doi.org/10.1038/nmeth.4324

Piriyapongsa J, Kaewprommal P, Vaiwsri S, Anuntakarun S, Wirojsirasak W, Punpee P, Klomsa-Ard P, Shaw PJ, Pootakham W, Yoocha T, Sangsrakru D, Tangphatsornruang S, Tongsima S, Tragoonrung S (2018) Uncovering full-length transcript isoforms of sugarcane cultivar Khon Kaen 3 using single-molecule long-read sequencing. PeerJ 6:e5818. https://doi.org/10.7717/peerj.5818

Qiao Q, Xue L, Wang Q, Sun H, Zhong Y, Huang J, Lei J, Zhang T (2016) Comparative transcriptomics of strawberries (*Fragaria* spp.) provides insights into evolutionary patterns. Front Plant Sci 7:1839. https://doi.org/10.3389/fpls.2016.01839

Qiao L, Cao M, Zheng J, Zhao Y, Zheng ZL (2017) Gene coexpression network analysis of fruit transcriptomes uncovers a possible mechanistically distinct class of sugar/acid ratio-associated genes in sweet orange. BMC Plant Biol 17(1):186. https://doi.org/10.1186/s12870-017-1138-8

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41(Database issue):D590–D596. https://doi.org/10.1093/nar/gks1219

Quesada V, Ponce MR, Micol JL (1999) OTC and AUL1, two convergent and overlapping genes in the nuclear genome of *Arabidopsis thaliana*. FEBS Lett 461(1–2):101–106

Rabiee-Ghahfarrokhi B, Rafiei F, Niknafs AA, Zamani B (2015) Prediction of microRNA target genes using an efficient genetic algorithm-based decision tree. FEBS Open Bio 5:877–884. https://doi.org/10.1016/j.fob.2015.10.003

Ramirez-Gonzalez RH, Segovia V, Bird N, Fenwick P, Holdgate S, Berry S, Jack P, Caccamo M, Uauy C (2015) RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. Plant Biotechnol J 13(5):613–624. https://doi.org/10.1111/pbi.12281

Ran JH, Shen TT, Wang MM, Wang XQ (2018) Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. Proc Biol Sci 285(1881):20181012. https://doi.org/10.1098/rspb.2018.1012

Ranwez V, Holtz Y, Sarah G, Ardisson M, Santoni S, Glemin S, Tavaud-Pirra M, David J (2013) Disentangling homeologous contigs in allo-tetraploid assembly: application to durum wheat. BMC Bioinformatics 14(Suppl 15):S15. https://doi.org/10.1186/1471-2105-14-S15-S15

Ren Y, Wang D, Wang Y, Zhou J, Zhang H, Zhou Y, Liang Y (2010) Prediction of disease-resistant gene in rice based on SVM-RFE. In: 2010 3rd International Conference on Biomedical Engineering and Informatics, 16–18 October 2010, pp 2343–2346. https://doi.org/10.1109/bmei.2010.5640583

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genom Proteom Bioinformatics 13(5):278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Riano-Pachon DM, Mattiello L, Cruz LP (2016) Surveying the complex polyploid sugarcane genome sequence using synthetic long reads. Laboratório Nacional de Ciência e Pesquisa do Bioetanol, Centro Nacional de Pesquisa em Energia e Materiais, Campinas. https://doi.org/10.13140/RG.2.1.3468.0565

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43(7):e47. https://doi.org/10.1093/nar/gkv007

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I (2010) De novo assembly and analysis of RNA-seq data. Nat Methods 7(11):909–912. https://doi.org/10.1038/nmeth.1517

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25

Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23(21):2881–2887. https://doi.org/10.1093/bioinformatics/btm453

Robinson MD, Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9(2):321–332. https://doi.org/10.1093/biostatistics/kxm030

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140. https://doi.org/10.1093/bioinformatics/btp616

Rosenow C, Saxena RM, Durst M, Gingeras TR (2001) Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches. Nucleic Acids Res 29(22):E112. https://doi.org/10.1093/nar/29.22.e112

Rothfels CJ, Larsson A, Li FW, Sigel EM, Huiet L, Burge DO, Ruhsam M, Graham SW, Stevenson DW, Wong GK, Korall P, Pryer KM (2013) Transcriptome-mining for single-copy nuclear markers in ferns. PLoS One 8(10):e76957. https://doi.org/10.1371/journal.pone.0076957

Sayols S, Scherzinger D, Klein H (2016) dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. BMC Bioinformatics 17(1):428. https://doi.org/10.1186/s12859-016-1276-2

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22(6):839–851. https://doi.org/10.1261/rna.053959.115

Serin EA, Nijveen H, Hilhorst HW, Ligterink W (2016) Learning from co-expression networks: possibilities and challenges. Front Plant Sci 7:444. https://doi.org/10.3389/fpls.2016.00444

Serin EAR, Snoek LB, Nijveen H, Willems LAJ, Jimenez-Gomez JM, Hilhorst HWM, Ligterink W (2017) Construction of a high-density genetic map from RNA-Seq data for an Arabidopsis Bay-0 x Shahdara RIL population. Front Genet 8:201. https://doi.org/10.3389/fgene.2017.00201

Shang X, Cao Y, Ma L (2017) Alternative splicing in plant genes: a means of regulating the environmental fitness of plants. Int J Mol Sci 18(2):432. https://doi.org/10.3390/ijms18020432

Shao M, Ma J, Wang S (2017) DeepBound: accurate identification of transcript boundaries via deep convolutional neural fields. Bioinformatics 33(14):i267–i273. https://doi.org/10.1093/bioinformatics/btx267

Shao L, Xing F, Xu C, Zhang Q, Che J, Wang X, Song J, Li X, Xiao J, Chen LL, Ouyang Y (2019) Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. Proc Natl Acad Sci U S A 116(12):5653–5658. https://doi.org/10.1073/pnas.1820513116

Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, Lintner KE, Ding Q, Wang Z, Hu J, Wang D, Wang F, Wang L, Lyon GJ, Guan Y, Shen Y, Evgrafov OV, Knowles JA, Thibaud-Nissen F, Schneider V, Yu CY, Zhou L, Eichler EE, So KF, Wang K (2016) Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun 7:12065. https://doi.org/10.1038/ncomms12065

Shi R, Wang JP, Lin YC, Li Q, Sun YH, Chen H, Sederoff RR, Chiang VL (2017) Tissue and cell-type co-expression networks of transcription factors and wood component genes in Populus trichocarpa. Planta 245(5):927–938. https://doi.org/10.1007/s00425-016-2640-1

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123. https://doi.org/10.1101/gr.089532.108

siTOOLsBiotech (2018) riboPOOL: affordable ribosomal/custom RNA depletion for any species

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26(8):1134–1144. https://doi.org/10.1101/gr.196469.115

Song L, Langfelder P, Horvath S (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics 13:328. https://doi.org/10.1186/1471-2105-13-328

Srivastava A, Malik L, Sarkar H, Zakeri M, Soneson C, Love MI, Kingsford C, Patro R (2019) Alignment and mapping methodology influence transcript abundance estimation. bioRxiv:657874. https://doi.org/10.1101/657874

Stumpf MPH, Ingram PJ (2005) Probability models for degree distributions of protein interaction networks. Europhys Lett 71(1):152–158

Sun X, Yang Q, Deng Z, Ye X (2014) Digital inventory of Arabidopsis transcripts revealed by 61 RNA sequencing samples. Plant Physiol 166(2):869–878. https://doi.org/10.1104/pp.114.241604

Sveinsson S, McDill J, Wong GK, Li J, Li X, Deyholos MK, Cronk QC (2014) Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (Linum) using transcriptomics. Ann Bot 113(5):753–761. https://doi.org/10.1093/aob/mct306

Tai Y, Liu C, Yu S, Yang H, Sun J, Guo C, Huang B, Liu Z, Yuan Y, Xia E, Wei C, Wan X (2018) Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (Camellia sinensis). BMC Genomics 19(1):616. https://doi.org/10.1186/s12864-018-4999-9

Tian J, Feng S, Liu Y, Zhao L, Tian L, Hu Y, Yang T, Wei A (2018) Single-molecule long-read sequencing of zanthoxylum bungeanum maxim. transcriptome: identification of aroma-related genes. Forests 9(12):765

Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nat Biotechnol 27(5):455–457. https://doi.org/10.1038/nbt0509-455

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515. https://doi.org/10.1038/nbt.1621

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31(1):46–53. https://doi.org/10.1038/nbt.2450

Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. Plant Biotechnol J 7(4):334–346. https://doi.org/10.1111/j.1467-7652.2008.00396.x

Tsaparas P, Marino-Ramirez L, Bodenreider O, Koonin EV, Jordan IK (2006) Global similarity and local divergence in human and mouse gene co-expression networks. BMC Evol Biol 6:70. https://doi.org/10.1186/1471-2148-6-70

Ungaro A, Pech N, Martin JF, McCairns RJS, Mevy JP, Chappaz R, Gilles A (2017) Challenges and advances for transcriptome assembly in non-model species. PLoS One 12(9):e0185020. https://doi.org/10.1371/journal.pone.0185020

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270(5235):484–487

Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL, Nobrega MP, Carrer H, Franca SC, Bacci Junior M, Goldman MH, Gomes SL, Nunes LR, Camargo LE, Siqueira WJ, Van Sluys MA, Thiemann OH, Kuramae EE, Santelli RV, Marino CL, Targon ML, Ferro JA, Silveira HC, Marini DC, Lemos EG, Monteiro-Vitorello CB, Tambor JH, Carraro DM, Roberto PG, Martins VG, Goldman GH, de Oliveira RC, Truffi D, Colombo CA, Rossi M, de Araujo PG, Sculaccio SA, Angella A, Lima MM, de Rosa Junior VE, Siviero F, Coscrato VE, Machado MA, Grivet L, Di Mauro SM, Nobrega FG, Menck CF, Braga MD, Telles GP, Cara FA, Pedrosa G, Meidanis J, Arruda P (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res 13(12):2725–2735. https://doi.org/10.1101/gr.1532103

Visser EA, Wegrzyn JL, Steenkmap ET, Myburg AA, Naidoo S (2015) Combined de novo and genome guided assembly and annotation of the Pinus patula juvenile shoot transcriptome. BMC Genomics 16:1057. https://doi.org/10.1186/s12864-015-2277-7

Vital-Lopez FG, Memišević V, Dutta B (2012) Tutorial on biological networks. Wiley Interdiscipl Rev Data Mini Knowl Discov 2(4):298–325. https://doi.org/10.1002/widm.1061

Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 131(4):281–285. https://doi.org/10.1007/s12064-012-0162-3

Walters RW, Matheny T, Mizoue LS, Rao BS, Muhlrad D, Parker R (2017) Identification of NAD+ capped mRNAs in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 114(3):480–485. https://doi.org/10.1073/pnas.1619369114

Wang XJ, Gaasterland T, Chua NH (2005) Genome-wide prediction and identification of cis-natural antisense transcripts in *Arabidopsis thaliana*. Genome Biol 6(4):R30. https://doi.org/10.1186/gb-2005-6-4-r30

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. https://doi.org/10.1038/nrg2484

Wang L, Feng Z, Wang X, Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26(1):136–138. https://doi.org/10.1093/bioinformatics/btp612

Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics 28(16):2184–2185. https://doi.org/10.1093/bioinformatics/bts356

Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun 7:11708. https://doi.org/10.1038/ncomms11708

Wang Y, Xie J, Zhang H, Guo B, Ning S, Chen Y, Lu P, Wu Q, Li M, Zhang D, Guo G, Zhang Y, Liu D, Zou S, Tang J, Zhao H, Wang X, Li J, Yang W, Cao T, Yin G, Liu Z (2017) Mapping stripe rust resistance gene YrZH22 in Chinese wheat cultivar Zhoumai 22 by bulked segregant RNA-Seq (BSR-Seq) and comparative genomics analyses. Theor Appl Genet 130(10):2191–2201. https://doi.org/10.1007/s00122-017-2950-0

Wang H, Gu L, Zhang X, Liu M, Jiang H, Cai R, Zhao Y, Cheng B (2018a) Global transcriptome and weighted gene co-expression network analyses reveal hybrid-specific modules and candidate genes related to plant height development in maize. Plant Mol Biol 98(3):187–203. https://doi.org/10.1007/s11103-018-0763-4

Wang M, Wang P, Liang F, Ye Z, Li J, Shen C, Pei L, Wang F, Hu J, Tu L, Lindsey K, He D, Zhang X (2018b) A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. New Phytol 217(1):163–178. https://doi.org/10.1111/nph.14762

Washburn JD, Schnable JC, Conant GC, Brutnell TP, Shao Y, Zhang Y, Ludwig M, Davidse G, Pires JC (2017) Genome-guided phylo-transcriptomic methods and the nuclear phylogentic tree of the paniceae grasses. Sci Rep 7(1):13528. https://doi.org/10.1038/s41598-017-13236-z

Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543. https://doi.org/10.1093/molbev/msx319

Wen J, Egan AN, Dikow RB, Zimmer EA (2015) Utility of transcriptome sequencing for phylogenetic inference and character evolution. In: Hörandl E, Appelhans MS (eds) Next-generation sequencing in plant systematics. International Association for Plant Taxonomy (IAPT), Bratislava, pp 51–91

Weng JK, Tanurdzic M, Chapple C (2005) Functional analysis and comparative genomics of expressed sequence tags from the lycophyte *Selaginella moellendorffii*. BMC Genomics 6:85. https://doi.org/10.1186/1471-2164-6-85

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafula E, Der JP, Graham SW, Mathews S, Melkonian M, Soltis DE, Soltis PS, Miles NW, Rothfels CJ, Pokorny L, Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H, dePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchan TM, Augustin MM, Wang J, Zhang Y, Tian Z, Yan

Z, Wu X, Sun X, Wong GK, Leebens-Mack J (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A 111(45):E4859–E4868. https://doi.org/10.1073/pnas.1323926111

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453(7199):1239–1243. https://doi.org/10.1038/nature07002

Williams PH, Eyles R, Weiller G (2012) Plant microRNA prediction by supervised machine learning using C5.0 decision trees. J Nucleic Acids 2012:652979. https://doi.org/10.1155/2012/652979

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881. https://doi.org/10.1093/bioinformatics/btq057

Wu P, Xie J, Hu J, Qiu D, Liu Z, Li J, Li M, Zhang H, Yang L, Liu H, Zhou Y, Zhang Z, Li H (2018) Development of molecular markers linked to powdery mildew resistance gene Pm4b by combining SNP discovery from transcriptome sequencing data with bulked segregant analysis (BSR-Seq) in wheat. Front Plant Sci 9:95. https://doi.org/10.3389/fpls.2018.00095

Xiao YL, Smith SR, Ishmael N, Redman JC, Kumar N, Monaghan EL, Ayele M, Haas BJ, Wu HC, Town CD (2005) Analysis of the cDNAs of hypothetical genes on Arabidopsis chromosome 2 reveals numerous transcript variants. Plant Physiol 139(3):1323–1337. https://doi.org/10.1104/pp.105.063479

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam TW, Li Y, Xu X, Wong GK, Wang J (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30(12):1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, Zhu Y, Ji A, Zhang B, Hu S, Au KF, Song J, Chen S (2015) Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. Plant J 82(6):951–961. https://doi.org/10.1111/tpj.12865

Yamamoto K, Sasaki T (1997) Large-scale EST sequencing in rice. Plant Mol Biol 35(1–2):135–144

Yang Y, Dong C, Yang S, Li X, Sun X (2015) Physiological and proteomic adaptation of the alpine grass Stipa purpurea to a drought gradient. PLoS One 10(2):e0117475. https://doi.org/10.1371/journal.pone.0117475

Yang G, Liu Z, Gao L, Yu K, Feng M, Yao Y, Peng H, Hu Z, Sun Q, Ni Z, Xin M (2018a) Genomic imprinting was evolutionarily conserved during wheat polyploidization. Plant Cell 30(1):37–47. https://doi.org/10.1105/tpc.17.00837

Yang L, Jin Y, Huang W, Sun Q, Liu F, Huang X (2018b) Full-length transcriptome sequences of ephemeral plant Arabidopsis pumila provides insight into gene expression dynamics during continuous salt stress. BMC Genomics 19(1):717. https://doi.org/10.1186/s12864-018-5106-y

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11(2):R14. https://doi.org/10.1186/gb-2010-11-2-r14

Yu X, Yang D, Guo C, Gao L (2018) Plant phylogenomics based on genome-partitioning strategies: progress and prospects. Plant Divers 40(4):158–164. https://doi.org/10.1016/j.pld.2018.06.005

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5):821–829. https://doi.org/10.1101/gr.074492.107

Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4:17. https://doi.org/10.2202/1544-6115.1128

Zhang L, Yu S, Zuo K, Luo L, Tang K (2012) Identification of gene modules associated with drought response in rice by network-based analysis. PLoS One 7(5):e33748. https://doi.org/10.1371/journal.pone.0033748

Zhang N, Liu B, Ma C, Zhang G, Chang J, Si H, Wang D (2014) Transcriptome characterization and sequencing-based identification of drought-responsive genes in potato. Mol Biol Rep 41(1):505–517. https://doi.org/10.1007/s11033-013-2886-7

Zhang F, Zhu G, Du L, Shang X, Cheng C, Yang B, Hu Y, Cai C, Guo W (2016a) Genetic regulation of salt stress tolerance revealed by RNA-Seq in cotton diploid wild species, Gossypium davidsonii. Sci Rep 6:20582. https://doi.org/10.1038/srep20582

Zhang ZF, Li YY, Xiao BZ (2016b) Comparative transcriptome analysis highlights the crucial roles of photosynthetic system in drought stress adaptation in upland rice. Sci Rep 6:19349. https://doi.org/10.1038/srep19349

Zhang C, Zhang B, Lin LL, Zhao S (2017) Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics 18(1):583. https://doi.org/10.1186/s12864-017-4002-1

Zhang H, Wang H, Zhu Q, Gao Y, Zhao L, Wang Y, Xi F, Wang W, Yang Y, Lin C, Gu L (2018) Transcriptome characterization of moso bamboo (Phyllostachys edulis) seedlings in response to exogenous gibberellin applications. BMC Plant Biol 18(1):125. https://doi.org/10.1186/s12870-018-1336-z

Zhang H, Zhong H, Zhang S, Shao X, Ni M, Cai Z, Chen X, Xia Y (2019a) NAD tagSeq reveals that NAD^+-capped RNAs are mostly produced from a large number of protein-coding genes in *Arabidopsis*. Proc Natl Acad Sci 116(24):12072–12077. https://doi.org/10.1073/pnas.1903683116

Zhang T, Liu C, Huang X, Zhang H, Yuan Z (2019b) Land-plant phylogenomic and pomegranate transcriptomic analyses reveal an evolutionary scenario of CYP75 genes subsequent to whole genome duplications. J Plant Biol 62(1):48–60. https://doi.org/10.1007/s12374-018-0319-9

Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12(Suppl 14):S2. https://doi.org/10.1186/1471-2105-12-S14-S2

Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 9(1):e78644. https://doi.org/10.1371/journal.pone.0078644

Zhao X, Li J, Lian B, Gu H, Li Y, Qi Y (2018) Global identification of *Arabidopsis* lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. Nat Commun 9(1):5056. https://doi.org/10.1038/s41467-018-07500-7

Zhao L, Zhang H, Kohnen MV, Prasad K, Gu L, Reddy ASN (2019) Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and nanopore-based direct RNA sequencing. Front Genet 10:253. https://doi.org/10.3389/fgene.2019.00253

Zhou X, Wang G, Sutoh K, Zhu JK, Zhang W (2008) Identification of cold-inducible microRNAs in plants by transcriptome analysis. Biochim Biophys Acta 1779(11):780–788. https://doi.org/10.1016/j.bbagrm.2008.04.005

Zhou Q, Su X, Jing G, Chen S, Ning K (2018) RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC Genomics 19(1):144. https://doi.org/10.1186/s12864-018-4503-6