Transcriptomics in Plant

5

Pratik Satya, Sougata Bhattacharjee, Debabrata Sarkar, Suman Roy, Laxmi Sharma, and Nur Alam Mandal

Abstract

Within a span of about 20 years, transcriptomics has established itself as an indispensible tool in almost all the areas of plant research. This chapter provides information on the rapid development of this important research area over a short period. Here, we present an overview of plant transcriptomics with an outline of the basic processes and tools including study design, RNA isolation, library preparation, sequencing platforms and bioinformatics analyses for annotation, pathway mapping and differential gene expression. A brief overview of the current status of transcriptomics in plants is presented followed by examples from a fibre producing plant, jute (*Corchorus* spp., Malvaceae), where transcriptomic researches have been proved very useful to understand biology and genetics of economically important traits.

Keywords

Transcriptomics · Transcriptome · NGS · Sequencing · Library · Gene expression · Jute

P. Satya (✉) · D. Sarkar · S. Roy · L. Sharma · N. A. Mandal
ICAR-Central Research Institute for Jute and Allied Fibres, Barrackpore, Kolkata, West Bengal, India

S. Bhattacharjee
ICAR-Vivekananda Parvatiya Krishi Anusandhan Sansthan, Almora, Uttarakhand, India

R. L. Singh et al. (eds.), *Plant Genomics for Sustainable Agriculture*,
https://doi.org/10.1007/978-981-16-6974-3_5

5.1 Introduction

Analysis of genes and genomes is a major research arena for understanding life process. A genome encompasses the whole set of inherited genetic material, carrying genes, regulatory sequences, repetitive elements and other components. Only a small part of it carries active genes, many of which are differentially expressed in tissues during various developmental phases or/and in response to external stimulations. If any of these conditions (time, cell or tissue type or environment) change, the cell adjusts to the new condition by changing the pattern and degree of gene expression. The term transcriptome, first used by Charles Auffray in 1996 (Piétu et al. 1999), refers to the total set of expressed mRNA molecules in a particular cell/cell type at a given physiological state in a specific environment. Such dynamic nature of transcriptome, which is not observed at genome level, provides opportunities to study the response of the organism to the change in environment or to study its growth and development. By analogy, transcriptomics is a collection of tools and techniques used to study the transcriptome. Over time, the use of the term 'transcriptomics' has expanded to include other coding and non-coding RNAs expressed in the cell, such as long non-coding RNA, as the same techniques can be used for mRNA or other RNA characterization by tweaking crucial steps. In plant science, transcriptomics is employed in various research arenas, such as to study environmental responses of plants under biotic or abiotic stresses, to understand basic biological processes like germination or fertilization, to identify genes and metabolic pathways, to decipher biological basis of crop productivity or to mine for novel phytochemicals from plant sources. Transcriptomics also helps to identify potential targets for a disease (e.g. ssRNA virus can be targeted using CRISPR/Cas12a/DNA ternary complex) and discovery of gene regulatory proteins (e.g. ChIP-Seq helps in identification of transcription factor and their exact binding sites on DNA). Its application has been extended to the areas of genomic manipulation, such as DNA free genome editing, which relies on RNA rather than DNA for making transgenic using CRISPR-Ribonucleoprotein complexes. Thus, transcriptomics has applications beyond identification of genes and characterization of their functionality and is an indispensable tool for solving fundamental biological questions.

While the sequencing technologies are same for both the genome and the transcriptome, the output sequence information has some basic differences. First of all, genome sequencing captures all the coding and non-coding sequences by first sequencing raw reads (contigs), and then stitches these contigs to a full length genome sequence. Transcriptome sequencing, on the other hand, captures all the mRNA sequences that are synthesized in a specific tissue/cell, where multiple copies of one particular gene are captured, each transcript being a 'raw read'. These are then aligned and matched with the gene sequences present in the genome (reference based) or de novo (based on robust gene identification algorithms). This allows a quantitative evaluation of transcript abundance by comparing the relative copy number of a read. Thus, relative expression of a gene in two or more transcriptomes, ideally from same plant and sequenced using same platform can be done to

understand the biological role of the gene. This ability to interpret gene function is the most important strength of transcriptomics, which cannot be obtained from genome sequencing. Second, genome sequence is individual-level information; thus it is fixed for a genotype. Transcriptome sequence, on the other hand, exhibits variations in different cells within a genotype. Even the same cell exhibits difference in transcriptome sequence under different environmental stimuli, making the transcriptome much more variable and informative than the genome. This inherent variability allows a wider application of transcriptomics in biological sciences, particularly to study development and responses to environmental changes. Third, gene expression is controlled by a variety of regulators, including small molecules, other genes (transcriptional factors), methylation (epigenomic modification), and external factors. This multi-dimensional cross-talk makes interpretation of transcriptome data more complex. Obviously, such dimensional complexity of transcriptome data requires specific robust mathematical analysis, big data analysis platforms and trained human resources to find the desired 'needle' from the transcriptomic haystack. To provide a better overview of transcriptomics, some specific terminologies are presented in Table 5.1.

5.2 Historical Development

Although expression analysis has long been utilized as a technique to establish gene functionality, large scale cDNA analysis was first undertaken under the Human Genome Project. In a seminal work published in the journal 'Science', Adams et al. (1991) generated 600 expressed sequence tags (EST) after cloning randomly selected cDNA from human brain tissue and showed that 337 of these coded for novel genes. They predicted that their approach would allow mapping most of the human genes within a few years. Within 5 years, the first human transcript map carrying 16,000 genes was generated (Schuler et al. 1996). Since then, this approach of large-scale gene characterization through cloning and sequencing of ESTs has been proved to be extremely useful for gene identification and characterization. A variety of subsequent methods were developed, ultimately bypassing the cloning step (direct sequencing of cDNA fragments). Development of DNA microarray technique (Schena et al. 1995) was the first milestone for large-scale gene expression analysis. Several large-scale platforms of microarray based gene expression systems emerged rapidly, including serial analysis of gene expression (Velculescu et al. 1995) and cDNA fingerprinting (Clark et al. 1999). At the same time, new clustering and multivariate algorithms for robust statistical analysis of large-scale gene expression data started to appear. One such clustering technique (Eisen et al. 1998) came from David Botstein's group, who is well-known as a pioneer researcher in DNA marker development. The research in plant transcriptomics gained momentum when Zhu and Wang (2000) designed the first large-scale expression array containing 8835 probes for *Arabidopsis* genes and generated over 500 transcriptome profiles. Since then, almost all the branches of plant science have resorted to transcriptome analysis for solving research problems, which can be envisaged from the sharp rise

Table 5.1 Terminologies associated with transcriptomics

Terminology	Explanation
Adapter	Short oligonucleotide sequences that are ligated to the 5′ and 3′ ends of DNA fragment during library preparation for sequencing. They match to the sequences present on the surface of the flow cells
Alignment	The process of matching two sequences. Two types of alignment strategies, local alignment and global alignment are used for aligning two nucleotide sequences
Barcode (tag)	A unique DNA sequence attached to template sequence before sequencing. Useful for multiplex sequencing, pooling of libraries, post-sequencing analyses, etc.
Cluster	Multiple copies of a sequence around a template, formed by bridge amplification. Each cluster grows in size as sequencing proceeds until a desired size of about 1000 copies are reached, and represents a single template sequence
Contig	A stretch of continuous nucleotide sequence
Coverage level	The average number of sequenced nucleotides that match with the reference nucleotide
De novo Assembly	Assembly of a set of RNA sequences without the support of a reference sequence
FASTQ file	A text output file of NGS sequencing containing the sequence and quality information of every sequenced base
Flow cell	A specially designed glass slide containing lanes for sequencing. The templates are fixed (immobilized) on the flow cell surface, so that enzymes can synthesize multiple copies using the template as source
Indels	Insertions and deletions in DNA sequences. Indels identified from a transcriptome analysis may be due to sequencing error or due to true mutations
Kmer length	A sequence can be broken down into small sequences (words) that can be overlapping or non-overlapping. These are used for rapid matching of sequences during matching with reference genome or matching between multiple sequences. The length of the word is the kmer length
Paired-end sequencing	Sequencing a fragment of DNA from both end
Q-score	A measure for error in base calling during sequencing. A Phred score is a quality score defined by the negative logarithm of the error probability
Reference-based assembly	Assembly of a set of RNA sequences based on a reference sequence
RNA-seq	An abbreviation of 'RNA-sequencing', a technique for sequence analysis of RNA from a sample. The sample can contain full spectrum of the RNA of a cell, tissue or organism (transcriptome), specific components of RNA (mRNA, snRNA, etc.), or partial sequences
Variant discovery	Identification in variation in genetic material between two cell, tissue or individual. Detects single nucleotide polymorphism (SNP), InDel (insertion-deletion) and variation in RNA secondary structure

in research publications. A search in Pubmed Central of the US National Library of Medicine (https://www.ncbi.nlm.nih.gov/pmc) with keyword 'plant transcriptome' retrieved only 22 hit in 2000, which increased rapidly to 10,408 in 2020. By 2019,

the One Thousand Plant Transcriptome Initiative sequenced the transcriptomes of 1124 species of Viridiplantae, and reconstructed the phylogeny of the major clades. To date, this is the most exhaustive documentation of plant transcriptomics. Analysis of major gene families revealed the role of gene and genome duplications in evolution although some components of the species tree remain still unresolved. The same group is now working on transcriptomics of ten thousand plants to develop a more robust phylogenetic species tree.

Advances in next-generation sequencing as well as biocomputing technologies during the past two decades resulted in development of several approaches for sequencing of genome and transcriptome. The first generation sequencers developed by Applied Biosystem Instruments (ABI) employed Sanger sequencing with fluorescent probes and used early-generation computers to collect and analyse data. In 1982, GenBank, the first public repository for sequence data was established, and a number of genome sequences were deposited by 2000. Two very important technologies, polymerase chain reaction (PCR) and shotgun sequencing revolutionized the field of genome sequencing and analysis during this period. However, post-2000 period was dominated by various new chemistry-based sequencing technologies, collectively referred as next-generation sequencing (NGS) technologies. The first wave came with the development of sequencing by synthesis (SBS), a technology based on massively parallel signature sequencing (MPSS) on microbeads. The commercial venture of MPSS was started by Lynx Therapeutics. Shankar Balasubramanian and David Klenerman developed the SBS technology and formed Solexa. Lynx Therapeutics merged with Solexa in 2005 and Solexa was acquired by Illumina in 2007. In 2004, 454 Life Science (now acquired by Roche) offered a pyrosequencing based NGS platform, and new models based on this system came in 2005-06 (454 GS 20), 2007 (454 GS FLX) and was further improved such as 454 GS FLX+, but was discontinued in 2013. By 2005, another commercial venture, Life Technologies developed SOLiD (Sequencing by Oligonucleotide Ligation and Detection). Illumina Inc. after acquiring Solexa started its own sequencing platform in 2009 and developed three very popular sequencing technologies, Hiseq, Miseq and Novaseq, and later developed Genome Analyzer platform, which is also based on SBS. A third generation of sequencing platforms like Pac Bio, nanopore and electron microscopy-based systems are currently being developed and utilized for large-scale sequencing, filling of gaps in existing sequences and resequencing of hundreds and thousands of samples. These improvements led to drastic reduction in cost and time of sequencing whole genome and transcriptome. A timeline of various events in transcriptome analysis is presented in Table 5.2.

5.3 Pipeline for Transcriptome Analysis in Plants

A pipeline or workflow of transcriptome analysis is an outline of the sequential processes to be followed to generate a transcriptome sequence and further analyse it as per the researcher's requirement. The processes can be divided in few major steps,

Table 5.2 A timeline of different transcriptomics technologies

Technology intervened	Year of intervention	Reference	Comment(s)
Northern Blot	1977	Alwine et al.	Gene specific detection, not applicable for global gene profiling
Sanger sequencing	1977	Frederick Sanger	First sequencing platform but very slow and costly
RT-PCR	1984	After PCR discovery by Kary Mullis	For cDNA synthesis from mRNA, routine transcriptome work
Microarray/ Affymetrix gene chip	1990s	Fodor et al. (1991), Schena et al. (1995)	Gene expression profiling and differential gene expression study
RACE	1989	Frohman and Martin	For cDNA end information, not useful for global transcript profiling
ESTs	1991	Adams et al.	High throughput single pass partial cDNA sequencing; now EST-clusters (unigene) used
Competitive PCR	1992	Siebert and Larrick	For differential gene expression analysis, not used recently
Antisense/Co-Suppression	1992	Richard Jorgensen	Functional transcript knocked down, targeted approach, now become obsolete
Improved DDRT-PCR	1993	Liang et al.	Differential gene expression study, target specific approach, not useful in organism level
Microarray system	1995	P Brown and R Davis	cDNA sequences on glass slides
SAGE/CAGE	1995	Veculescu et al.	Representative partial sequencing of transcripts, tags gives useful information about cell/tissue specific transcript profile.
Two-dimensional microarray	1995–96	P Brown's group	Fluorescent detection, high speed
Initiation of the concept of 'sequencing by synthesis'	Mid 1990s	S. Balasubramanian, and D. Klenerman at Cambridge	Detected motion of DNA polymerase during synthesis by fluorescent labelling
Patent filed for nanopore sequencing	1995	Church, Deamer, Branton and colleagues	The concept of nanopore sequencing developed
SSH	1996	Diatchenko et al.	Identify novel gene, very useful tool but not amenable for whole transcriptome level

(continued)

Table 5.2 (continued)

Technology intervened	Year of intervention	Reference	Comment(s)
RNAi	1998	Fire et al.	Targeting mRNA for functional validation, gene specific approach
Clustering of microarray data	1998	Eisen et al.	Improved statistical analysis and interpretation
Oligonucleotide microarray system/ GeneChip platform	1999	Affymerix	In situ synthesis of oligos on chip
Massively Parallel Signature Sequencing (MPSS)	2000	Brenner et al.	Sequencing throughput accelerated, useful for cell level when using NGS technology
qRT-PCR/Real-Time PCR-based analysis	2001	Livak et al.	Quantification of mRNA expression
Next Generation Sequencing (NGS) platforms	2004 onwards	Various commercial ventures	See previous section for historical development
454 sequencing	2005	Life Sciences (Roche Diagnostics)	Used for transcriptome study, technology withdrawn in 2013
SOLiD (Sequencing by Oligonucleotide Ligation and Detection)	2006	Applied Biosystems Inc. (later Life Technologies)	Can generate 60 Gb data per run
Genome Analyzer	2006	Solexa	Sequenced 1 GB per run
Single molecule detected by nanopore	2008	Gundlach's group	Used MspA nanopore
Single Molecule Real-Time (SMRT) sequencing	2009	Craighead, Korlach, Turner and Webb	Sequencing is performed in a SMRT cell containing nanowells
Third generation sequencing (TGS) platforms	2009 onwards	Various commercial ventures	Pacific Biosciences, Oxford Nanopore Technology, Quantapore (CA-USA), and Stratos (WA-USA)
Single Molecule Real-Time (SMRT) sequencing commercialized	2011	Pacific Biosciences	Can sequence longer reads, base calling less accurate than Illumina short read sequencing
MinION sequencer	2014	Oxford Nanopore	Portable device, up to 30 GB
NovaSeq platforms	2017	Illumina	Up to 6 TB read capacity
HiFi (High Fidelity)	2019	Pacific Biosciences	Can generate Circular Consensus Sequences (CCSs) approximately 10–20 kbp-long
Sequel II sequencer	2019	Pacific Biosciences	Contains 8 million nanowells SMRT Cell, capacity 160 GB.
R10 Nanopore sequencing	2019	Oxford Nanopore	Double sensor for more efficient base calling

namely, design of the study, RNA isolation, prepare a sequencing library, sequencing of the library, processing of the raw reads to obtain clean reads, assembly of the sequences and annotation of the transcriptome (Fig. 5.1). Further bioinformatics or wet-lab analyses are performed based on the research need. A transcriptome analysis pipeline can be objective specific. For example, several pipelines have been developed for differential gene expression analysis.

5.3.1 Transcriptomic Study Design

Any experiment needs to be planned methodically by applying appropriate tool (s) for testing the hypothesis. Testing a biological hypothesis using transcriptomics requires selecting an appropriate sequencing platform, determining the number of replicates and use of a statistically robust design. In plant RNA-seq experiments, at least three biological replicates are recommended by the European Molecular Biology Laboratory (EMBL). Biological replicates are, however, only required when inferences are to be drawn on population rather than the individual organism itself. In case of studies with plant, we primarily draw inference on population, so determining the number of biological replicates is essential for drawing a robust conclusion. In some cases, however, biological replicates may be avoided. For example, if in some experiment two tissues from the same plant are compared, or two plants of same genotype cultured in same flask are compared, one may not use biological replicate. But usually in an experiment, particularly for differential gene expression analysis, plants are grown in different conditions, where biological replicates are required. Selecting technical replicates depends on the technical reproducibility of the sequencing platform, which is high for most of the advanced sequencing systems although several processes during library preparation can introduce bias in output. Cost is another important issues, because with each replicate sequencing cost is increased, thus the researcher has to sacrifice some accuracy in case of budget constraint. A number of statistical techniques are available for interpreting un-replicated transcriptome data. Software like NOISeq (Tarazona et al. 2011) and GFOLD are effective for expression analysis of genes that have strong biological response (Khang and Lau 2015). Use of three or more replicates improves the power of the study, allowing identification of genes with weak biological response.

Another important issue is the read depth/read coverage of the transcriptome. If the experiment is a pilot scale study, or high quality reference sequence information is available, one may select low read depth and more number of replicates. But if the RNA-seq sequencing is de novo, more read depth would be preferable. Although such benchmark studies are rare, Liu et al. (2013) observed that an increase in number of DE genes with sequencing depth has diminishing returns after 10 million reads and suggested increasing replication over read depth. Lamarre et al. (2018) observed that the optimal threshold to control the false discovery rate (FDR) is approximately 2^{-r} (r = replicate number). They showed that 20 million reads per

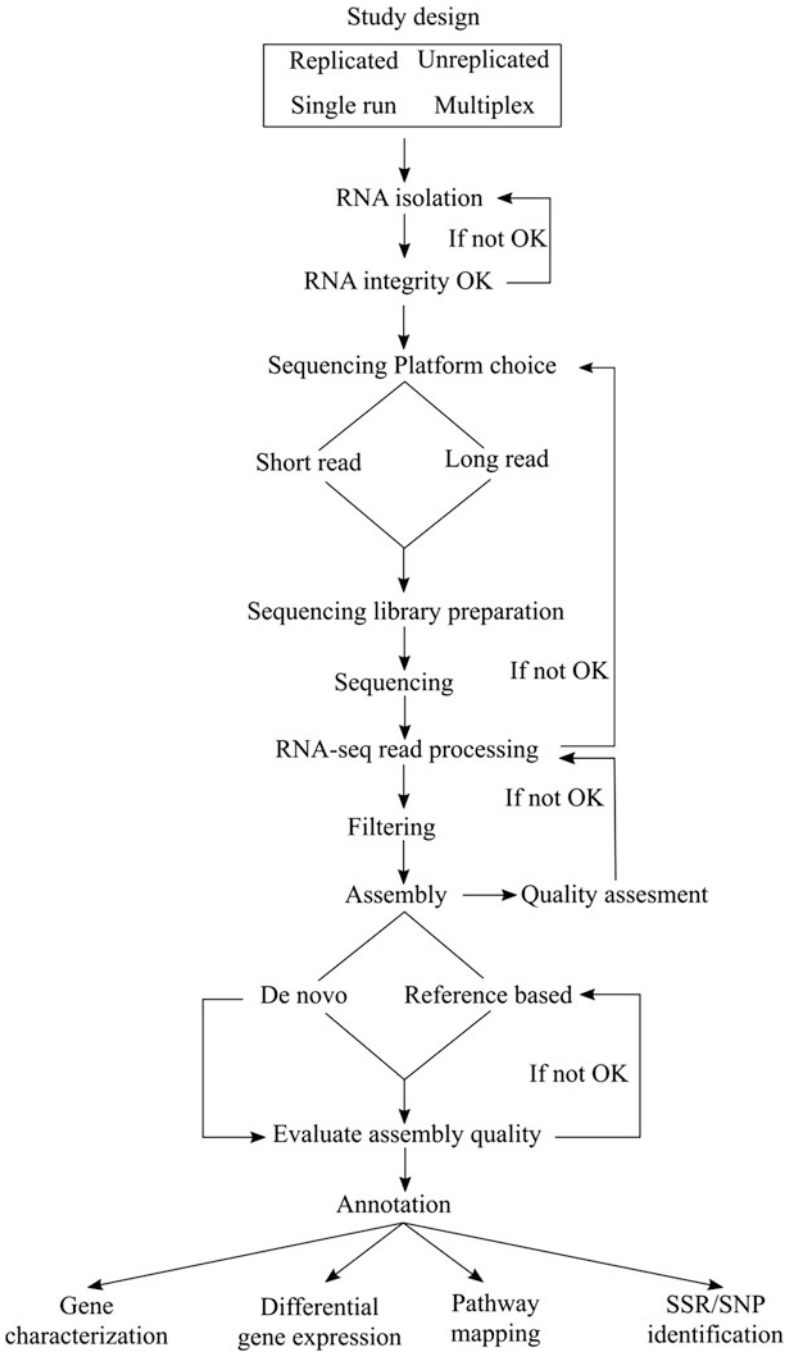

Fig. 5.1 A simplified pipeline for transcriptomics in plant

sample and four biological replicates would be required to capture 1000 differentially expressed genes in tomato.

5.3.2 RNA Isolation and Processing

Transcriptomics starts with quality RNA isolation. Since there are various kinds of RNA in cell that differ in length, the procedure for isolation of RNA will vary based on the experimental requirement. Standard RNA-seq captures only the protein mRNAs that are generally >200 bp, thus the RNA isolation procedure is standardized in such a way that short RNAs (<100 bp) are washed out, and the isolated RNA is enriched with mRNAs. To isolate short RNAs, specific silica-based membranes are used. The quality of RNA, in addition to standard spectrophotometric quality assessment, is evaluated by RNA integrity number (RIN), which is determined in an Agilent BioAnalyzer by 18s/28s rRNA electrophoresis. A RIN value of >7 (range 1–10) is well accepted for RNA-seq analysis. The next step is to remove the rRNA and tRNA, which together constitutes 96–98% of the total RNA sample, and to retain only mRNA (2–3%) in the sample. The mRNA portion is recovered from the total RNA pool by poly-dT primers that specifically bind to the poly-A tail of mRNA. Alternatively, the rRNA and tRNA can be removed by binding to probes specific to these RNAs. In case of mRNA capture, long non-coding RNAs (lncRNA) are also removed, so this method cannot be used to capture and sequence lncRNA. On the other hand, probe based methods are not full-proof and some rRNA and tRNA remnants are always present in the sample after processing. This method also requires probe information, which requires prior sequence knowledge.

5.3.3 Library Preparation

The preparation of sequencing library is the most important step of RNA-seq. It depends on the sequencing platform and sequencing strategies used. Principally, a RNA-seq library is a pool of cDNA fragments (in case of sequencing by synthesis). On an Illumina platform, the RNA pool is fragmented to a size of 50–300 bp (read length) either enzymatically, chemically or mechanically. The cDNA is synthesized either by single end sequencing or paired-end sequencing using reverse transcriptase, using a specific PCR system called bridge amplification. For the first strand synthesis, oligo-dT primer, random primer or adaptor ligated primers can be used, each of which has its own advantages and limitations. The oligo-dT primers are biased towards 3′-end, and will miss all the fragments that lack poly-A tail. Random primers capture all these fragments but suffer from drawbacks like non-random binding and loss of strand information. The ligated primers are better than the other two systems for capturing the mRNA pool. The second strand is synthesized by DNA polymerase using specially designed primers.

5.3.4 Library Sequencing

Different sequencing platforms employ different strategies for sequencing the cDNA library. The following are the major sequencing platforms widely used in plant transcriptomics. Technological advances in each of these systems have resulted in tremendous improvement in sequencing power, output quality and cost reduction.

5.3.4.1 Roche (454) FLX

454 Life Sciences (Roche Diagnostics) was first commercialized in 2005 and currently Genome Sequencer (GS) FLX System and GS FLX Titanium series platforms are available. After preparing template as discussed above, beads along with the attached DNA fragments are removed from the emulsion and loaded into the wells (PicoTiter Plate). Each well contains only one bead. Pyrosequencing principle (luciferase-based light detection on pyrophosphate release when a base is added in sequencing process) is used for sequencing (Ronaghi et al. 1996). From template preparation to data processing the FLX system (read length 450–500 bases) takes 10 h per run (generates 400-Mb sequence data). Recently developed GS FLX Titanium XL+ platform can generate 1 Gb sequence data with read length of 1000 bp.

5.3.4.2 Illumina/Solexa

The Illumina sequencing system is based on the principle of sequencing by synthesis (SBS). The solid phase PCR is then carried out inside flow cell, which is also called fold-back PCR or bridge PCR (Fedurco et al. 2006). The system works on reversible terminator technology. The templates are immobilized on a proprietary flow cell array and are ligated with adaptors carrying barcode or inline index (both are unique short sequences to discriminate reads of different pools). After the first cycle of cDNA elongation, the 5′ end of the single strand DNA bends and binds to a functional group on the flowcell, and the original templates are washed away. The bridge fragments are made double stranded and PCR is performed on these bridges to generate several millions of dense clusters. The first sequencing cycle is performed by adding four fluorescently labelled terminator nucleotides with primers and DNA polymerase. After incorporation of each dNTP, the polymerization is terminated to image the fluorescence tag, the dye is enzymatically removed and the next dNTP is incorporated to extend the chain, which allows recording of every fluorescent signal, thereby determining the sequence of the template. Few recent platforms like, Illumina Genome Analyzer 1 Gb and HiSeq 600 Gb are very popular. Illumina read length generally varies from 35 to 150 bases. IlluminaHiSeq 2000 platform yields 400 Gb of sequence data in a single run (takes 7–8 days). Another model, HiSeq X Ten can generate 1.8 Tb sequence data. In 2017, Illumina introduced Novaseq platforms which are more efficient, generating up to 6 TB sequence data and claims to complete sequencing of 48 genomes in less than two days.

5.3.4.3 ABI SOLiD

SOLiD (sequencing by oligonucleotide ligation detection) platform utilizes oligonucleotide probes (8 bp long, each having two unique nucleotides at 3′end and labelled with fluorophore at the 5′ end) ligation for detecting the base of transcripts while sequencing. It was commercialized by Applied Biosystems in 2005 as SOLiD 3.0 platform (Shendure et al. 2005). In this technology, single layer beads are immobilized in an acrylamide matrix on a glass slide along with attached DNA molecules. A set of 16 oligos (for 4 bases of nucleic acid) are required for hybridization with template cDNA while sequencing in each reaction. While encoding base in sequencing, each unique base pair of 3′ end of the probe is assigned one out of four possible colours for ease of detection and analysis. During sequencing, each base in the template is sequenced twice and hence SOLiD technology is said to be highly accurate. The SOLiD 3.0 platform yields read length of 50 bases only and can generate approx. 20 Gb sequence data per run. SOLiD 5500 and SOLiD 5500 XL systems were introduced to increase the sequence data of up to 300 Gb per run (Edwards et al. 2013).

5.3.4.4 Ion Torrent (Semiconductor-Based Life Technologies)

This technology was developed by Ion Torrent Systems Inc. and was commercialized in 2010. It utilizes a semiconductor-based device, also called ion chip, that senses the H+ ions generated during DNA extension by DNA polymerase (measures the induced pH changes by the release of hydrogen ions (Rothberg et al. 2011)). The ion chip, having wells of 3.5-μm-diameter, is located directly over the electronic sensor. The voltage signal is proportional to the number of bases incorporated in the new strand synthesized by DNA polymerase and the detection system is non-optical scanning, which eliminates use of fluorophores, thereby reducing cost and increasing speed of detection. In 2012, another high throughput technology was released, called 'Ion Proton', which increased output by an order of magnitude of 10× but the read length was drastically reduced in comparison with Ion Torrent (200 bp instead of 400 bp).

5.3.4.5 Pacific Biosciences

Single molecule real-time (SMRT) sequencing was developed by Nanofluidics, Inc. and commercialized by Pacific Biosciences, USA. In this technology, template is prepared through ligation of single-stranded hairpin structured adaptor to the cDNA ends (thereby generating a bell-shaped structure called SMRT-bell). Single molecules of DNA polymerase are immobilized at the bottom using biotin–streptavidin interaction in zepoliter-sized wells, also called zero-mode waveguides (ZMWs), and four dNTPs in high concentration with different fluorophore labelled are used for rapid DNA synthesis using strand displacing polymerase (Levene et al. 2003). One advantage is that a cDNA molecule can be sequenced multiple times. Moreover, direct sequencing instead of clonal multiplication allows the sequence to be read in real-time (Eid et al. 2009). Each SMRT cell can generate ∼50 k reads and up to 1 Gb of data in 4 h.

5.3.4.6 Oxford Nanopore Technologies

In a nanopore system, a sequencing flow cell composed of hundreds of micro-wells containing a synthetic bilayer and punctured by biologic nanopores (Wang et al. 2015). Sequencing is achieved simply by precise measuring the changes in current induced as a result of incorporation of bases through the nanopores with the help of a molecular motor protein. Library is prepared by ligating adapters to cDNA ends in a manner that first adapter can bind with motor enzyme and second adapter (a hairpin oligonucleotide) can bind with another HP motor protein. Therefore, simultaneously two strands can be sequenced from a single molecule and increase the accuracy in comparison with SMRT technology. This is a highly throughput technology where a single run (18 h) can generate more than 90 Mb of sequencing data with maximum read lengths of more than 60 kb using MinION platform (USB-powered, portable sequencer) (Ashton et al. 2015).

5.3.5 Quality Control

The raw sequence data output from the system is obtained in 'FASTQ' format. A quality score, known as Phred quality score (Q) determines the quality of the sequence. Generally, Q >28 indicates good quality of the transcriptome, while Q <20 has a poor quality. Several other parameters, such as technical artefacts (adaptor, primer dimer, etc.) and biological artefacts (other sequence contamination) can interfere with the quality. To test these parameters, number of overrepresented sequence, duplicate reads and kmer count (a measure for technical artefact) are examined. Once such artefacts are determined, the contaminated sequences are removed by filtering and trimming using processing software to generate processed reads.

5.3.6 Read mapping, assembly and annotation

Once the reads are generated, they are to be assembled to identify the genes. Since the transcriptome reads are of very small length (30–100 nt) (though some platforms produce longer reads) and are to be matched ideally against genome sequence of the same organism (which is in case of plants can go up to thousands of megabases), a robust annotation system is required. Mostly, a compression algorithm is applied to reduce the computational load. Burrows–Wheeler algorithm is one such compression tool that helps in fast annotation of the sequences. Several annotation pipelines are available for de novo and reference-based annotation of transcriptomes and differential gene expression.

5.4 Bioinformatics Software for Transcriptome Analysis

Bioinformatics software is at the core of transcriptomics. These software filter raw data, assembly the filtered sequences into transcripts, annotate their biological function and mine the transcriptome for various information including SSRs, SNPs, regulatory genes, differentially expressed genes, metabolic pathways, genetic causes and responses of disease or stress, transposable elements and many more. While using software depends on use of platform and purpose of experimentation, some software are often preferred due to their high reliability and accuracy.

5.4.1 Filtering

As described earlier, filtering involves cleaning and trimming of unwanted sequences from the reads and quality assessment. FASTX-Toolkit (Gordon and Hannon 2010) is widely used for filtering of transcriptome raw reads. For quality inspection of a transcriptome, FastQC (Andrews 2010) is a good choice. During sequencing, the raw reads are stored in 'FASTQ' format by the sequencer, which merge the sequence (FASTA) with a quality score, called Phred score, which is determined by error probability of base calling. A higher Phred score indicates more confidence in base calling, i.e., sequencing quality.

5.4.2 Assembly

Errors in assembly can seriously impair transcriptome quality. A single transcript may be fragmented and scored as multiple transcripts, causing loss in information, or multiple transcripts may be erroneously joined together constructing a chimera, creating problems in annotation. Many genes exist as duplicates or gene families having high sequence similarity. Correct assembly of fragmented reads of these genes is extremely difficult, which is another source of error. Several tools are available for assembly, some of which are bundles of software or assembly pipeline. Often it is better to use more than one assembly for finding out the best one, which obviously depends on the sequence type, sequence quality and method of assembly (reference-based/de novo). Since a de novo assembly generates transcripts only based on RNA-seq data, it is more erroneous than reference-based assembly. The basis of de novo assembly is generation of a de Bruijn Graph based on *kmer* decomposition of the read. Therefore, *kmer* length is an important factor for de novo assembly. A shorter *kmer* has more coverage, but at the same time has more chance to be read from multiple transcripts. For de novo assembly, several tools are available, of which Trinity (Grabherr et al. 2011), SOAPdenovo-Trans (Xie et al. 2014), Trans-ABySS (Robertson et al. 2010) and rnaSPAdes (Bankevich et al. 2012) are more popular. Trinity is a well trusted assembly pipeline for de novo assembly and is recommended by various researchers as it has high transcript recovery and accuracy (Freedman and Weeks 2020). A software, TransRate can compare various

assemblies by giving a quality score (Smith-Unna et al. 2016), which can be used for selecting appropriate assembly. Wang and Gribskov (2017) compared eight de novo assembly tools (BinPacker, Bridger, IDBA-tran, Oases-Velvet, SOAPdenovo-Trans, SSP, Trans-ABySS and Trinity) at different kmer length (25-71) and observed that SOAPdenovo-Trans had the highest base coverage, while Trans-ABySS was best in gene coverage and recovery of full-length transcripts. They recommended performing de novo assembly even when reference genome is available, as transcript fragmentation, incorrect/incomplete gene annotation and exon level differences are major reasons for difference in annotation and differential gene expression. Holzer and Marz (2019) observed that for short read sequences, Trinity, SPAdes, and Trans-ABySS, were better than other tools, but no tool was best for all data sets. These results show that evaluation of different assemblies is a critical step for good assembly construction. Another tool that can be used for de novo transcriptome analysis for gene expression is RSEM (RNA-Seq by Expectation-Maximization), which uses Bowtie/Bowtie2/STAR for read alignment and EBseq for differential gene expression (Li and Dewey 2011). New methods like principles of information theory and abundance of alternate spliced transcripts are being applied to improve the efficiency of de novo assembly (Mao et al. 2020).

For reference-based genome guided assembly, the chance of error is less, but the quality of transcriptome depends on the reference genome/transcriptome quality. Several reference based assemblers are available, such as Cufflinks (Trapnell et al. 2010), StringTie (Pertea et al. 2015), TransComb (Liu et al. 2016), Bayesembler (Maretty et al. 2014), CLASS2 (Song et al. 2016) and Scallop (Shao and Kingsford 2017). In addition, Trinity has options for genome guided de novo assembly. Comparative estimations show that StringTie produces more accurate assembly than Cufflinks or Bayesembler, but results may vary depending on sequence quality. An updated version of StringTie, StringTie2 is now available that can assemble longer reads (>200) efficiently (Kovaka et al. 2019). The RNA-seq reads are first aligned using a spliced aligner such as HISAT/HISAT2 (Kim et al. 2015) or STAR (Dobin and Davis 2013). Alignment outputs are stored as SAM (Sequence Alignment/Map) or BAM (Binary Alignment/Map format) file format, which are used as input files for differential gene expression analysis tools. New alignment-free assemblers, based on kmer matching, for example, Salmon (Patro et al. 2017) and Kallisto (Bray et al. 2016) are faster than the alignment-dependent methods like StringTie, but have lower efficiency in detecting low-abundance transcripts and novel transcripts. Another assembler, Necklace (Davidson and Oshlack 2018) is useful when the reference sequence is incomplete. It requires the RNA-seq read to be assembled, the incomplete reference genome and one or more well-annotated genome from related species, and builds a super Transcriptome merging all inputs.

5.4.3 Annotation

For de novo transcriptome assemblies, annotation is required to identify the function of the transcript, while in reference based assemblies, the transcripts are matched to

annotated reference, so further annotation is not required. One may, however, improve over the previous annotation, as the databases used for annotation (for example, BLAST databases) are updated frequently. The tools of genome and transcriptome annotation are same, based on BLAST databases, which have a variety of algorithms and tools for annotation of RNA-seq data.

5.4.4 Differential Gene Expression

Perhaps the most common use of plant transcriptome analysis is study of differential expression of genes (DEG or DGE) of tissues having different treatments or stress conditions. The basic principle is to identify the number of sequenced reads mapped to a single gene, which is a measure of expression of the gene in the sample. Several other factors influence this count, such as gene length (longer transcripts have more fragments mapped), sequencing depth and expression level of other genes. Therefore, a normalized measure is required to estimate DEG. A couple of such measures are widely used in DEG analyses. The RPKM (Reads PerKilobase per Million mapped reads) is a measure where mapped reads are first normalized to reads per million (RPM) with a scaling factor of 106, which is then divided by the length of the gene. For paired-end sequencing, FPKM (Fragment Per Kilobase per Million mapped reads) is used, which follows the same normalization RPKM, with the difference that that two paired reads are considered as a single unit. Another measure is TPM (Transcript Per Million), where the mapped reads are first normalized with the length of the gene followed by with the total of the normalized reads scaled by the factor 106. Significance of gene expression can be tested by estimating mean and variance of expression of a gene over replicates, which means that replicated data should be generated for DEG. A number of other measurements and plots, such as false discovery rate, MA plot and volcano plot can be generated to understand DEG data. Software like DeSeq2 (Love et al. 2014) and NOIseq (Tarazona et al. 2011) provide these normalized read counts for comparing gene expression and perform clustering or other multivariate techniques to study relationship of samples in terms of gene expression. Most commonly, the clustering is described with a heat map showing gene expression values.

5.4.5 Pathway and Gene Ontology Mapping

Once the DEGs are identified, the next step is to understand their biological roles. While annotations using blast identify the closest homolog from the database, more meaningful biological information can be derived by DEG. For this, two approaches, pathway mapping and gene ontology (GO) mapping are very helpful. The Kyoto Encyclopaedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/pathway.html) maintains databases and tools for mapping a gene onto metabolic pathways, which is extensively used by researchers for assigning annotated genes to metabolic pathways, and in case of DEG, helps to identify reaction paths

overexpressed and underexpressed within a metabolic pathway under two or more different conditions. However, only a small fraction of the genes identified in a transcriptome or from DEG are annotated by KEGG. The GO Consortium (http://geneontology.org), on the other hand, can provide biological meaning to more number of genes, assigning them under broad categories of cellular component, molecular function and biological process, under which several sub-categories are available, which sequentially describe the 'ontology' of the gene via a GO map. Another approach, cluster of orthologous groups (COG) (Tatusov et al. 2000) classify the annotated genes into several cluster of orthologous groups. The query protein sequences can be searched using 'blastp' against the COG database. Most gene annotation pipelines have capacity for searching these databases and assign biological meaning to the RNA-seq transcripts. The European Molecular Biology Laboratory (EMBL) hosts another gene ontology search tool, EggNOG (Huerta-Cepas et al. 2019), that includes non-supervised orthologs (NOGs) for functional characterization of a gene.

5.5 Transcriptomics of Plants

5.5.1 *Arabidopsis thaliana*

A. thaliana, the mouse ear cress, is a model plant species for biological researches on plant. Consequently, *Arabidopsis* transcriptomes are the most researched transcriptomes. Before the advent of NGS technologies, large scale gene expression experiments were carried out using microarray, which still provides useful information on expression pattern of the A. thaliana genes. The Unite´ de Recherche en Ge´nomique Ve´ge´tale (URGV), France hosts a publicly available database of *Arabidopsis* transcriptomes, CATdb (Complete *Arabidopsis* Transcriptome Database) (http://urgv.evry.inra.fr/CATdb), a collection of 281 *Arabidopsis* projects mainly obtained from the microarray data resources generated by the URGV transcriptome platform. It provides access to CATMA (Complete *Arabidopsis* Transcriptome MicroArray), developed by a European consortium. The CATMA probes cover over 85% of the genes present in *Arabidopsis* providing gene sequence tags for individual genes. It has further been extended to 20 other species and presently contains data on 353 projects. The Salk Institute hosts a *Arabidopsis* Transcriptome Genomic Express Database (http://signal.salk.edu/cgi-bin/atta), containing data from the *Arabidopsis* transcriptome Tilling array, exosome, At-TAX (a whole genome tilling array) and DNA methylome. It provides a pictorial description of the expression pattern of the genes. The *Arabidopsis* Information resource (Tair) (https://www.*Arabidopsis*.org/index.jsp) also contains exhaustive functional genomics resources on *Arabidopsis*. Various other databases are publicly available to researchers for transcriptomics studies in *Arabidopsis*, making it the most researched plant species (Table 5.3).

Table 5.3 A list of *Arabidopsis* functional genomics databases

Database	Description	url
Arabidopsis Transcriptome Genomic Express Database	Contains information from tilling, methylome, expression analysis. Provides gene specific expression profile	http://signal.salk.edu/cgi-bin/atta
Arabidopsis RNA-seq Database	Gene expression levels from 20,000+ public *Arabidopsis* RNA-Seq libraries	http://ipf.sustc.edu.cn/pub/athrna/
ARTADE -- *Arabidopsis* Tiling-Array-based Detection of Exons	Annotation of genome-wide tiling-array data	http://omicspace.riken.jp/ARTADE/
Arabidopsis Gene Regulatory Information Server (AGRIS)	Contains promoter sequences, transcription factors and their target genes	http://*Arabidopsis*.med.ohio-state.edu/
Arabidopsis Small RNA Project database (ASRP)	Information on small nuclear RNA	http://asrp.danforthcenter.org/
Arabidopsis Next Gen sequence database	A part of Next Gen sequence database at Donald Danforth Plant Science Centre	https://mpss.meyerslab.org/
AthaMap	Genome-wide map of potential transcription factor and small RNA binding sites	http://www.athamap.de/
CATMA	Provides high quality Gene-specific Sequence Tags (GSTs) covering most *Arabidopsis* genes	http://www.catma.org/
ePLANT	Multiple visualization tools for gene expression	http://bar.utoronto.ca/eplant/
Expression Atlas	Contains results of 962 experiments including *Arabidopsis*, rice and maize	https://www.ebi.ac.uk/gxa/plant/experiments
SeedGenes	Information on genes with essential function during seed development	http://seedgenes.org/
TraVA	A database of gene expression profiles based on RNA-seq	http://travadb.org/

5.5.2 Current Status of Transcriptomics in Crop Plants

Use of transcriptomics in understanding the biology and cultivation of the crop plants is rising sharply in the present century. However, the sequence read archive (SRA) deposits in NCBI (https://www.ncbi.nlm.nih.gov/sra/) for the most important 25 crops of the world show a skewed pattern. Out of these, the number of SRA deposits for ten crops (maize, rice, wheat, Brassica, soybean, tomato, cotton, tea, potato and sugarcane) are about 0.35 million, which is ten times higher than the SRA deposits for the other 15 crops (0.036 million) (Fig. 5.2), indicating that more transcriptomics research is needed for harvesting the benefit of this technology in minor crops. Maize and rice, two principal food crops have received maximum attention to the transcriptomics researchers, comprising about 53% of the total SRA deposits.

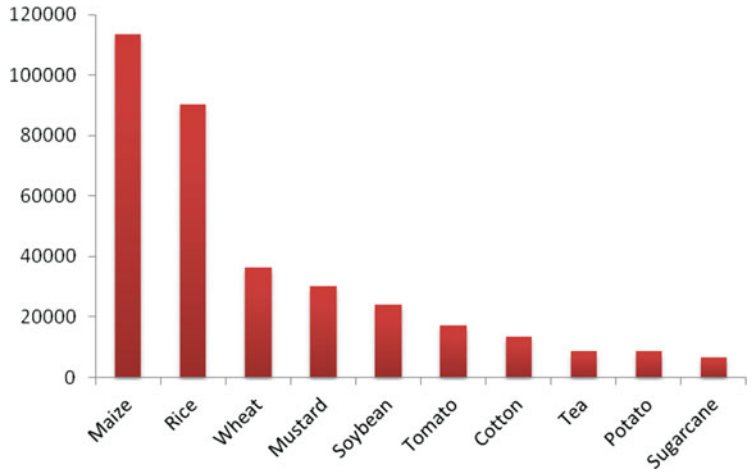

Fig. 5.2 Crop-wise top ten species with SRA (sequence read archive) deposits in NCBI SRA database (as on 11.02.2021)

Similarly, in Pubmed Central (https://www.ncbi.nlm.nih.gov/pmc/), a total of 0.156 million hits were recorded for these 25 cultivated crops (Fig. 5.3, data up to 31.12.2020). The distribution shows that over 50% of these hits are from four principal crops, rice (18.6%), wheat (11.8%), maize (13.3%) and soybean (9.1%), while another 26.4% are from tomato, potato and cotton. More emphasis on transcriptomics of other economically important crops would be required for having a better understandin
g of genetic basis of economically important traits in these crops. In the next section, we will give some examples of use of transcriptomics in jute (*Corchorus* spp.). Despite being the second most important fibre crop (after cotton), jute transcriptomics has received comparatively less attention and support than the food crops or even the beverage crops like tea and coffee. However, within a short time frame, transcriptomics has helped to understand a number of biological processes in jute, which is an inspiring example of the benefits of transcriptomics in crop plants.

5.6 Transcriptomics in Jute: An Overview

The jute plant represented by two species *Corchorus olitorius* L. and *Corchorus capsularis* L. (Malvaceae, subfamily Grewoideae) is cultivated for production of long, tough fibre synthesized in bast (phloem) tissue. The fibre, known as jute fibre is a lignocellulosic fibre is used for production of sacks, bags, burlaps, geotextiles, fibre composites and various other diversified products. It is valued globally as the most important non-textile fibre. While the principal producers of jute are India, Bangladesh and China, it is globally used to pack food grains and is in high demand

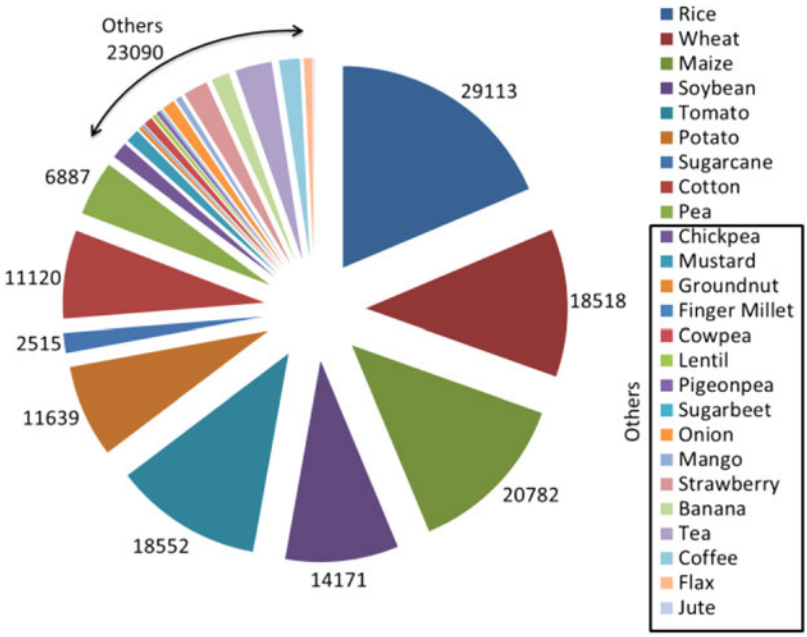

Fig. 5.3 Comparative crop-wise research focus in transcriptomics as indicated by number of hits returned by Pubmed Central on search with keywords "Transcriptome"+"respective crop name" between 2000 and 2020. Most of the researches (85%) are focused on major crops (rice, wheat, maize, soybean, tomato, potato, sugarcane, cotton and pea)

for replacement of synthetic polypropylene bags. Apart from producing natural fibres, jute plant consumes over 14 ton of CO_2/ha during its vegetative growth period of about 120 days and fixes nutrition to the soil by addition of leaf litters. It, therefore, is a climate-friendly crop that produces climate-friendly natural fibre. Consequently, research in jute genomics and transcriptomics has attracted considerable attention in recent decades in the wake of the rising concerns over climate change. Due to low genetic variability in jute at population level, researchers have concentrated more on transcriptomics to understand the genetics of economically important traits rather than using genomic tools like linkage mapping and genomic selection. This has generated a large amount of sequence information, identifying genes, regulatory sequences, metabolic pathways and genic markers. Till 11.02.2021, the number of SRA deposits for both the jute species was 714, which included full length high coverage transcriptomes of various tissues, as well as low coverage sequences from mapping experiments. The earliest reference of jute sequence was deposited to NCBI SRA archive in 2015 by the Central Research Institute for Jute and Allied Fibres, India, which were the RAD-seq (restriction-site associated DNA sequence) of jute cv. Sudan Green (SRX591273) and mutant bast fibre shy (*bfs*) and their F_2 plants. These were sequenced using IlluminaHiseq 2000 platform generating 2.2 Gb and 1.8 Gb sequences for Sudan Green and *bfs*,

respectively. In this study, RAD-seq data for 330 F_2 genotypes were also deposited. The first SRA deposit for the first whole transcriptome sequence of jute was submitted by the same institute in 2015, providing transcriptome data for bast tissue of a mutant deficient in lignified phloem fibre production (*dlpf*) and its wild-type cv. JRC-212. Transcriptomes of different tissues including bast, hypocotyl, developing stem, root, fibre cell, leaf and flower have been generated in jute. In addition, expression of genes under different conditions such as salt-stressed and GA_3-treated plants has been investigated.

5.6.1 Transcriptome Assembly

A number of assemblers have been used for annotation and functional characterization of jute genes. Chakraborty et al. (2015) and Satya et al. (2018) performed de novo assembly of bast transcriptome using three assemblers CLC Genomics Workbench (v6.0; CLC bio, Aarhus), SOAPdenovo-Trans and Trinity. Islam et al. (2017) performed a reference-based assembly of the fibre cell transcriptome using Cufflinks. Yang et al. (2020) also developed reference genome based assembly using Bowtie and TopHat. The first transcriptome assembly with publicly available TSA (Transcriptome shotgun assembly) was generated by Chakraborty et al. (2015) using Illumina^{TM}Hiseq 2000 platform generating a total of 72,750,724 raw read and 67,424,930 clean reads for cv. JRC-212. After de novo assembly using CLC workbench, SOAPdenovo-Trans and Trinity, a total of 34,163 genes were identified. Among the three assemblers, Trinity was found to be the best performing with maximum percentage of unigene recovery.

5.6.2 Gene Discovery from Transcriptome Data

Wide variations have been reported for the number of genes expressed in different tissues of both the jute species. The earliest transcriptome study (Chakraborty et al. 2015) reported presence of 29,000-34,000 genes in the bast tissue of *C. capsularis*, which can be publicly accessed from the TSA database of NCBI. Some reports contain an exorbitantly high number of genes expressed (over 72,000) in jute, which probably needs to be verified. Overall, jute has an estimated number of 35,000-40,000 annotated genes (Table 5.4).

5.6.3 Orthologous Group Identification and Gene Ontology

Chakraborty et al. (2015) used four annotation databases, Nr, SwissProt, KEGG and COG for functional interpretation of the bast transcriptome assembly (Fig. 5.4), and reported that Nr-annotation was superior to the other three systems. Further, they identified gene ontology using Blast2GO (Conesa et al. 2005) and obtained the GO functional classifications using WEGO (Ye et al. 2006). Satya et al. (2018) followed

Table 5.4 Major transcriptomics studies in jute

Tissue (s)	Sequencing platform	Species	Unigenes identified	Reference
Tissues from Vegetative growth period, flowering period, bast of technical mature period, fruit	Illumina™ Hiseq 2500	*C. olitorius*	33,312 in total, 15,491 common in all the tissues	Yang et al. (2020)
Whole plant	Illumina™ HiSeq 4000	*Corchorus* sp. L.	72,674	Tao et al. (2020)
Shoot apiece	Illumina NextSeq 500	*C. olitorius*	14,050	Choudhary et al. (2019)
Hypocotyl	Illumina™ Hiseq 2000	*C. capsularis*	32,821-39,076 (annotated)	Satya et al. (2018)
Fibre cell	Illumina™ Hiseq 2500	*C. capsularis and C. olitorius*	37,031 (*C. olitorius*), 30,096 (*C. capsularis*)	Islam et al. (2017)
PEG-treated tissue	Illumina HiSeq X Ten	*C. olitorius*	45,831	Yang et al. (2017a, b)
Salinity-stressed tissue	Illumina HiSeq 4000 platform	*C. olitorius*	72,278	Yang et al. (2017a, b)
Pooled RNA from various tissues	Illumina™ Hiseq 2000	*C. capsularis*	48,914	Zhang et al. (2015)
Bast	Illumina™ Hiseq 2000	*C. capsularis*	34,163–29,463	Chakraborty et al. (2015)

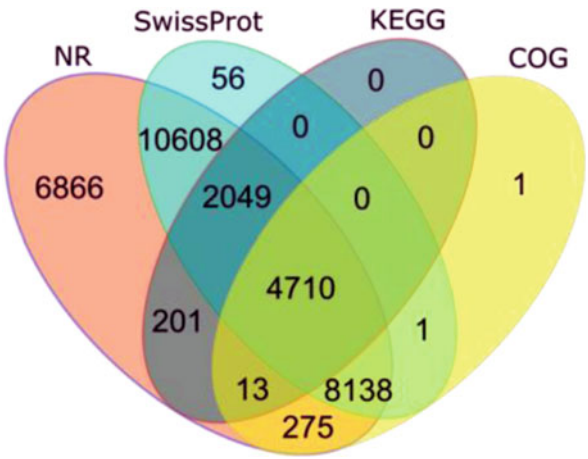

Fig. 5.4 A venn diagram representing annotations of the bast transcriptome of *Corchorus capsularis* cv. JRC-212 (Chakraborty et al. 2015) using four annotation databases. Note that SwissProt, KEGG and COG annotations did not add much information over Nr-annotations

similar methodology for annotation of hypocotyl transcriptomes and confirmed that Nr-annotation performed better. Both the studies screened the COG database to retrieve and classify COG functional categories of the genes. In case of reference-based assemblies, orthology and ontology analysis are not required, as the transcriptome is assembled based on a reference genome having genes with assigned functions.

5.6.4 Identification of Novel genes

Often, transcriptomics leads to discovery of novel genes that were unknown to exist in that species, family or even may be unknown to plant Kingdom. During examination of the role of β-galactosidases in hypocotyl development in jute, Satya et al. (2018) discovered a novel class of beta-galactosidases that are similar to prokaryotic β-galactosidase (Fig. 5.5). The prokaryotic β-galactosidase (a member of Glycosyl Hydrolase-2 or GH-2 family of enzymes) converts lactose to glucose and galactose, and was thought to be lost in higher eukaryotes. The domains of the GH-2 β-galactosidases are highly conserved in prokaryotes, consisting of three protein domains Glyco-hydro_2_N, Glyco-hydro_2 and Glyco-hydro_2_C that are linked to a Bgal_Small_N domain by another β-sandwich domain of unknown function (DUF4981). As such, plant cannot utilize lactose as a food source, which was thought to be due to absence of the prokaryotic GH-2 β-galactosidase. They, on the other hand, contain a number of β-galactosidases of GH-35 family, which function in cell wall formation by breaking galactose linked with other molecules. Satya et al. (2018) observed that a homolog of *E. coli* lacZ gene (codes for β-galactosidase) with this five-domain architecture is present not only in jute but in all the plants starting from algae to woody perennials. Phylogenetic study revealed that the plant GH-2 β-galactosidases evolved from the prokaryotic β-galactosidases. It was transferred from prokaryotes to lower plants (Marchantiophyta and Bryophyta) via Charophytic green algae and from lower plants to higher plants via Lycophyta. Protein modelling revealed remarkable similarity between the plant and prokaryotic GH-2 β-galactosidases despite having low sequence similarity.

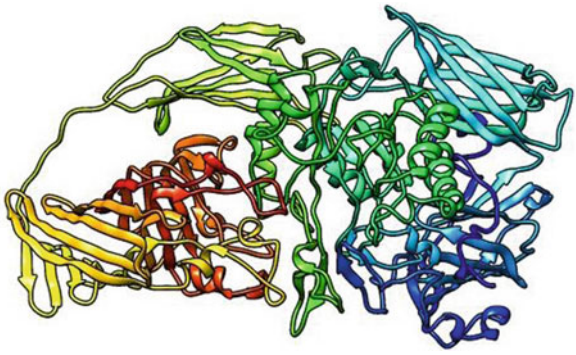

Fig. 5.5 A 3-D predicted protein structure of a novel prokaryotic β-galactosidase gene of jute discovered in plant lineage by hypocotyl transcriptomics. The structure was generated using Phyre2

Particularly, the catalytic residues that cause a nucleophile attack on the β-1,4 linkage of glucose and galactose were found to be conserved in higher plants.

5.6.5 Metabolic Pathway Identification

For metabolic pathway analysis, annotated genes are mapped to the KEGG database (Kanehisa et al. 2008) using the KEGG Automatic Annotation Server (Moriya et al. 2007). Chakraborty et al. (2015) characterized the phenylpropanoid biosynthesis pathways in jute that lead to monolignol formation and genes involved in the secondary cell wall development. They identified a total of sixteen genes with multiple isoforms which were involved in lignin biosynthesis and jute fibre formation. Islam et al. (2017) described the major genes involved in fibre formation in jute from genomic and transcriptomic datasets and observed that *C. capsularis* exhibits higher ATPase activity, oxidoreductase activity, transmembrane transport, vacuolar transport and homeostasis, suggesting that it has wider environmental adaptability. In another study, Satya et al. (2020) characterized the pectin biosynthesis pathways in jute, identifying 18 genes involved in interconversion of nucleotide-sugars, salvage biosynthesis of sugar-acids and polymerization of pectin monomers. Of these, 17 were involved in nucleotide-sugar interconversion and one, galacturonosyltransferase (GAUT) for polymerization of pectin monomers. A total of 12 *GAUT* genes were identified from both the species, which phylogenetically were distributed in seven subclades. Two of these, CcGAUT3 and CcGAUT12 were identified as the primary pectin homo-polymerizing enzymes. Both the CcGAUT3 and CcGAUT12 had an N-terminal transmembrane domain that carried a consensus motif ((R)–(X)2–(R)) for proteolytic cleavage. It was predicted that a CcGAUT3-CcGAUT12 complex may be involved in polymerization of galacturonic acid monomers in jute. The study also reported that the core pectin biosynthesis pathway is conserved in higher plants. Species that produce high mucilage, such as *Ziziphus jujube* exhibited high conservation with the jute pectin biosynthesis genes.

5.6.6 DEG Analysis

Only a few DEG experiments have been conducted in jute. Choudhary et al. (2019) identified a total of 240 differentially expressed transcripts between delayed flowering mutants under short-day (*pfr59*) in comparison with cv. JRO-524 and observed that 10 transcripts showed homology to known photoperiodic genes of *Arabidopsis*. DEG analysis was also used for identification of drought-stress associated genes in *C. olitorius* by Yang et al. (2017a, b). A drought sensitive cultivar exhibited 794 DEGs under drought stress, while in a drought tolerant cultivar only 39 genes were differentially expressed. Recently, Yang et al. (2020) identified 576/379, 291/227, 2367/255 and 1766/736 genes (upregulated/downregulated), respectively, in the stem bast, fruit, flower, and leaf compared to other tissues. They observed that 26 genes of the secondary metabolite biosynthesis

pathway were consistently upregulated in the bast and the phenylpropanoid biosynthesis pathway genes were significantly upregulated in flower.

5.6.7 Marker Development

5.6.7.1 SSR

One of the major applications of plant transcriptomics is to identify EST-SSRs or genic SSRs. In jute, Zhang et al. (2015) discovered 1906 EST-SSRs with a frequency of di-, tri-, tetra-, and penta-nucleotide repeat types of 12.0%, 56.9%, 21.6% and 9.5%, respectively. They identified 113 transcription factor associated SSRs and 3 SSRs for cellulose synthase. Later, more SSRs (12,772) were identified from a bast transcriptome, with an average frequency of one SSR per 3.86 Kb (Satya et al. 2017). About 45.4% of the sequences exhibited repeat length between 10 and 15 nt. and 46.2% of the SSR loci were about 300–2000 nt. The number of repeats varied from 6 to 15 for dinucleotides, 5–8 for trinucleotides and 5–6 for tetranucleotides. Of the dinucleotide repeats, (TA/AT)6 was the most frequent (9.3%). They also identified 961 compound SSRs (7.5%). They also designed 39 phenylpropanoid biosynthesis pathway gene-specific SSR markers, seven SSR markers for peroxidase genes, and 24 SSRs for the genes involved in bast fibre formation. They also reported 4457 transcription factors (TF) and identified 2163 TF-SSRs. The study designed 1079 SSR primers and validated 120 of them using gel electrophoresis studies. Saha et al. (2017) identified 4509 SSRs and developed a set of 2079 flanking primer-pairs. They also developed a web-based SSR repository of jute (http://jutemarkerdb.icar.gov.in/). All these SSRs were found to show moderate polymorphism and were able to generate high intra-specific and inter-specific diversity.

5.6.7.2 SNP and InDel

Zhang et al. (2015) identified a total of 12,518 SNPs in jute with transition and transversion frequencies of 59.2 and 22.3%, respectively. Most of the SNPs were of the synonymous SNP type (99.37 %). Yang et al. (2018) identified 51,172 InDel sites in 18,800 unigenes of jute, which were distributed in 94 InDel types. Mononucleotide InDels were more (23,028) than bi-nucleotide (9824) or tri-nucleotide (9182) ones. The polymorphism information content of InDel markers in jute varied from 0.340 to 0.680, with an average of 0.491.

5.7 Conclusion

During the past 20 years, transcriptomics has established itself as an essential tool in plant biology. Advances in next generation sequencing have opened up new avenues for in-depth investigations of the sequence of events in a biological process at single cell level. The cost of transcriptomics studies have been reduced by several folds in recent years, allowing its wider application in plant biology and crop improvement. Moreover, publicly deposited transcriptomics studies not only benefit the researchers

working for specific crops, but also enrich the public databases, allowing better precision for gene annotation in future researches. The role of transcriptomics will be invaluable for future plant research, particularly to battle various abiotic stresses escalating due to climate change, soil degradation, higher population pressure and water stress.

References

Adams MD, Kelley JM, Gocayne JD et al (1991) Complementary DNA sequencing: Expressed sequence tags and human genome project. Science 252(5013):1651–1656

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Ashton PM, Nair S, Dallman T et al (2015) MinIONnanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol 33(3):296–300

Bankevich A, Nurk S, Antipov D et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19(5):455–477

Bray NL, Pimentel H, Melsted P et al (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527

Brenner S, Johnson M, Bridgham J (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18(6):630–634

Chakraborty A, Sarkar D, Satya P et al (2015) Pathways associated with lignin biosynthesis in lignomaniac jute fibres. Mol Genet Genom 290(4):1523–1542

Choudhary SB, Saha D, Sharma HK et al (2019) Transcriptional analysis of a delayed-flowering mutant under short-day conditions reveal genes related to photoperiodic response in tossa jute (*Corchorus olitorius* L.). Ind Crop Prod 132:476–486

Clark MD, Panopoulou GD, Cahill DJ et al (1999) Construction and analysis of arrayed cDNA libraries. In: Weissman SM (ed) Methods enzymol, vol 303. Academic Press, San Diego, CA, pp 205–233

Conesa A, Gotz S, García-Gómez JM et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676

Davidson NM, Oshlack A (2018) Necklace: combining reference and assembled transcriptomes for more comprehensive RNA-Seq analysis. Gigascience 7(5):giy045

Diatchenko L, Lau YF, Campbell AP et al (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci 93(12):6025–6030

Dobin A, Davis CA, Schlesinger Fet al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126(1):1–11

Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138

Eisen MB, Spellman PT, Brown PO et al (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95(25):14863–14868

Fedurco M, Romieu A, Williams S et al (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34(3):e22

Fire A, Xu S, Montgomery MK et al (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391(6669):806–811

Fodor SPA, Read JL, Pirrung MC et al (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251:767–773

Freedman A, Weeks N (2020) Best practices for *de novo* transcriptome assembly with Trinity. Available via https://informatics.fas.harvard.edu/best-practices-for-de-novo-transcriptome-assembly-with-trinity.html.
Frohman MA, Martin GR (1989) Rapid amplification of cDNA ends using nested primers. Techniques 1:165–173
Gordon A, Hannon GJ (2010) FASTX-Toolkit. FASTQ/A short-reads pre-processing tools http://hannonlab.cshl.edu/fastx_toolkit
Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652
Hölzer M, Marz M (2019) *De novo* transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience 8(5):giz039
Huerta-Cepas J, Szklarczyk D, Heller D et al (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47(D1):D309–D314
Islam MS, Saito JA, Emdad EM et al (2017) Comparative genomics of two jute species and insight into fibre biogenesis. Nat Plants 3(2):16223
Kanehisa M, Araki M, Goto S (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:D480–D484
Khang TF, Lau CY (2015) Getting the most out of RNA-seq data analysis. PeerJ 3:e1360
Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–360
Kovaka S, Zimin AV, Pertea GM et al (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol 20(1):278
Lamarre S, Frasse P, Zouine M et al (2018) Optimization of an RNA-seq differential gene expression analysis depending on biological replicate number and library size. Front Plant Sci 9:108
Levene MJ, Korlach J, Turner SW et al (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299(5607):682–686
Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12(1):323
Liang P, Pardee A (1993) Distribution and cloning of eukaryotic mRNAs by mean of differential display: refinements and optimization. Nucleic Acids Res 21(14):3269–3275
Liu Y, Ferguson JF, Xue C et al (2013) Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. PLoS One 8(6):e66883
Liu J, Yu T, Jiang T et al (2016) TransComb: genome-guided transcriptome assembly via combing junctions in splicing graphs. Genome Biol 17(1):213
Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2-\Delta\Delta CT$ method. Methods 25(4):402–408
Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550
Mao S, Pachter L, Tse D et al (2020) RefShannon: A genome-guided transcriptome assembler using sparse flow decomposition. PLOS One 15(6):e0232946
Maretty L, Sibbesen JA, Krogh A (2014) Bayesian transcriptome assembly. Genome Biol 15:501
Moriya Y, Itoh M, Okuda S et al (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–W185
Patro R, Duggal G, Love MI et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14(4):417–419
Pertea M, Pertea GM, Antonescu CM et al (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33(3):290–295
Piétu G, Mariage-Samson R, Fayein NA et al (1999) The Genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. Genome Res 9(2):195–209

Robertson G, Schein J, Chiu R et al (2010) *De novo* assembly and analysis of RNA-seq data. Nat Methods 7(11):909-912

Ronaghi M, Karamohamed S, Pettersson B et al (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242(1):84–89

Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352

Saha D, Rana RS, Chakraborty S et al (2017) Development of a set of SSR markers for genetic polymorphism detection and interspecific hybrid jute breeding. Crop J 5(5):416–429

Satya P, Chakraborty A, Jana S et al (2017) Identification of genic SSRs in jute (*Corchorus capsularis*, Malvaceae) and development of markers for phenylpropanoid biosynthesis genes and regulatory genes. Plant Breed 136(5):784–797

Satya P, Chakraborty A, Sarkar D et al (2018) Transcriptome profiling uncovers β-galactosidases of diverse domain classes influencing hypocotyl development in jute (*Corchorus capsularis* L.). Phytochemistry 156:20–32

Satya P, Sarkar D, Vijayan J et al (2020) Pectin biosynthesis pathways are adapted to higher rhamnogalacturonan formation in lignocellulosic jute (*Corchorus* spp.). Plant Growth Regul 14: 1–7

Schena M, Shalon D, Davis RW et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):460–470

Schuler GD, Epstein JA, Ohkawa H et al (1996) Entrez: molecular biology database and retrieval system. Methods Enzymol 266:141–162

Shao M, Kingsford C (2017) Accurate assembly of transcripts through phase-preserving graph decomposition. Nat Biotechnol 35(12):1167–1169

Shendure J, Porreca GJ, Reppas NB et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 5741:1728–1732

Siebert PD, Larrick JW (1992) Competitive PCR. Nature 359:557–558

Smith-Unna R, Boursnell C, Patro R et al (2016) TransRate: reference-free quality assessment of *de novo* transcriptome assemblies. Genome Res 26(8):1134–1144

Song L, Sabunciyan S, Florea L (2016) CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. Nucleic Acids Res 44(10):e98

Tao AF, You ZY, Xu JT et al (2020) Development and verification of CAPS markers based on SNPs from transcriptome of jute (*Corchorus* L.). Acta Agronomica Sinica 46(7):987–996

Tarazona S, García F, Ferrer A et al (2011) NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. EMBnet J 17(B):18–19

Tatusov RL, Galperin MY, Natale DA et al (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36

Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515

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

Wang S, Gribskov M (2017) Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics 33(3): 327–333

Wang Y, Yang Q, Wang Z (2015) The evolution of nanopore sequencing. Front Genet 5:449

Xie YL, Wu GX, Tang JB et al (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666

Yang Z, Dai Z, Lu R et al (2017a) Transcriptome analysis of two species of jute in response to polyethylene glycol (PEG) induced drought stress. Sci Rep 7(1):1–11

Yang Z, Yan A, Lu R (2017b) *De novo* transcriptome sequencing of two cultivated jute species under salinity stress. PLoS One 12(10):e0185863

Yang Z, Dai Z, Xie D, Chen J, Tang Q, Cheng C, Xu Y, Wang T, Su J (2018) Development of an InDel polymorphism database for jute via comparative transcriptome analysis. Genome 61:323–327

Yang Z, Wu Y, Dai Z et al (2020) Comprehensive transcriptome analysis and tissue-specific profiling of gene expression in jute (*Corchorus olitorius* L.). Ind Crop Prod 146:112101

Ye J, Fang L, Zheng H et al (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34:W293–W297

Zhang L, Ming R, Zhang J et al (2015) *De novo* transcriptome sequence and identification of major bast-related genes involved in cellulose biosynthesis in jute (*Corchorus capsularis* L.). BMC Genomics 16**:**1062

Zhu T, Wang X (2000) Large-scale profiling of the *Arabidopsis* transcriptome. Plant Physiol 124(4):1472–1476