# Advances of transcriptomics in crop improvement: A Review

Sougata Bhattacharjee[1]
[1]Scientist, Plant Biotechnology, Crop Improvement Division, ICAR-VPKAS, Almora, Uttarakhand, India-263601

***Abstract:***
The transcriptome defines the whole set of transcripts available in a cell for a specific point of time and condition. Interpreting these functional elements of the genome is key to reveal the phenotype or response of an individual or cell. Broadly, to understand the physiological and molecular events in terms of genome sequence, gene differentiation, gene expression regulation, posttranscriptional modifications, editing and gene splicing, transcriptome study is essential to answer the fundamental biological queries. Over the years, to deduce and quantify the transcriptome there have been a drastic shift from simple PCR (RT-PCR)/hybridization (Northern blot) dependent candidate gene-based detection to random end sequencing-based (tag-based; ESTs, SAGE, MPSS) approaches followed by more recent, complicated genome-wide high throughput global expression profiling (Microarray & sequencing based). The advancement of next generation sequencing platforms has revolutionized not only the knowledge of dynamic transcriptomics but also provides multi-dimensional analysis in terms of novel transcript discovery, structural detection at single base resolution, functional annotation, mapping, differential gene expression, enrichment analysis etc. Furthermore, cDNA based and tag-based sequencing approaches are not amenable for short/degraded RNAs or non-coding RNAs, and for those, RNA-seq or single molecule direct RNA sequencing have been developed to overcome these limitations. Transcriptomics in recent years have vast applicability in every aspect in the field of agriculture, medicine, environment for overall strategic improvements. In near future, enough scope is there to introduce more advance technologies to bridge the gap of recent technological shortcomings.

***Key words:*** Next Generation Sequencing (NGS), transcriptome, e-QTL, RNA-seq

**Introduction:**
The term transcriptome sometimes used to define the total set of transcripts in a given organism, or to the specific subset of transcripts expressed (mostly mRNA) in a particular cell type for the ease of explanation depending on experiments. Transcriptome can vary with degree of expression of coding sequences depending on various factors unlike genome, which defines total genetic material of an organism, is roughly fixed for a given cell line irrespective of conditions. Since long 'RNA world hypothesis' there have been an established fact about diversity of RNA structure and functionality. This led to the birth of 'transcriptomics' and this is a multidisciplinary science encompasses the study of transcriptome as a whole and broadly a part of functional genomics. In the era of high throughput technologies, many tools are now available to integrate the knowledge of sequencing data with gene structure and function. To analyse genome structure i.e. elucidation of transcription units (open reading frames, exon-intron boundaries, editing, alternative gene structure, untranslated region, promoter region etc.), forming a reference sequence for genome mapping, chromosome walking followed by positional cloning, elucidation of functional marker (e.g. any genic marker designed from exon sequences), identification of structure & conserved regions of tRNA, rRNA, lncRNA and their cellular localization etc. transcriptome study is crucial. Moreover, to understand the global as well as the specific gene expression study is needed to understand the snapshot of genome functionality and its response towards differentiation, stress or other environment (e.g. RNAi/Antisense/Co-suppression/Ribozyme etc.). Transcriptome analysis helps to identify potential target for any disease (e.g., ssRNA Virus can be targeted using CRISPR/Cas12a/DNA ternary complex) and sometimes helps in identification of gene regulatory proteins (e.g., ChIP-Seq helps in identification of Transcription Factor and their exact binding site at DNA). Now a days, the concept of DNA free genome editing also relies on RNA rather than DNA for making transgenic using CRISPR-Ribonucleoprotein complexes. So, understanding of this diverse arena is very essential to not only assign functionality of gene rather answer fundamental biological questions which ultimately leads to understand the phenomics of an organism.

**Transcriptome techniques and advances:**
Since 1970, after discovery of reverse transcriptase enzyme, studies on transcripts were being not familiarized for three decades before any high throughput transcriptomics approaches were accessible. First generation Sanger sequencing was available in 1980 and became very much popular for its usage in sequencing of random individual transcripts from cDNA libraries, which is called expressed sequence tags (ESTs). Though this technology was predominant until any next generation sequencing technology arrived, but the read length of this technology was small and cost is also was high. Advent of Sequencing by synthesis, ESTs became more popular due to its efficiency to reflect the expressed gene content of an organism when complete genome sequence was a challenge as well as unnecessary. Several other technologies also hit the market consecutively in the arena of transcriptomics. Quantification through real time PCR and identification of specific transcript from a pool of transcripts through northern blotting was also very popular and sophisticated tool and widely used, but these methods are laborious and used only to capture a portion a transcriptome. Moreover, the high throughput sequencing platforms in last one decade have revolutionized the transcriptomics in terms of efficacy and quality of data generated. RNA-seq is the most advanced tool till date and uses all the modern technologies from sequencing to analysis.

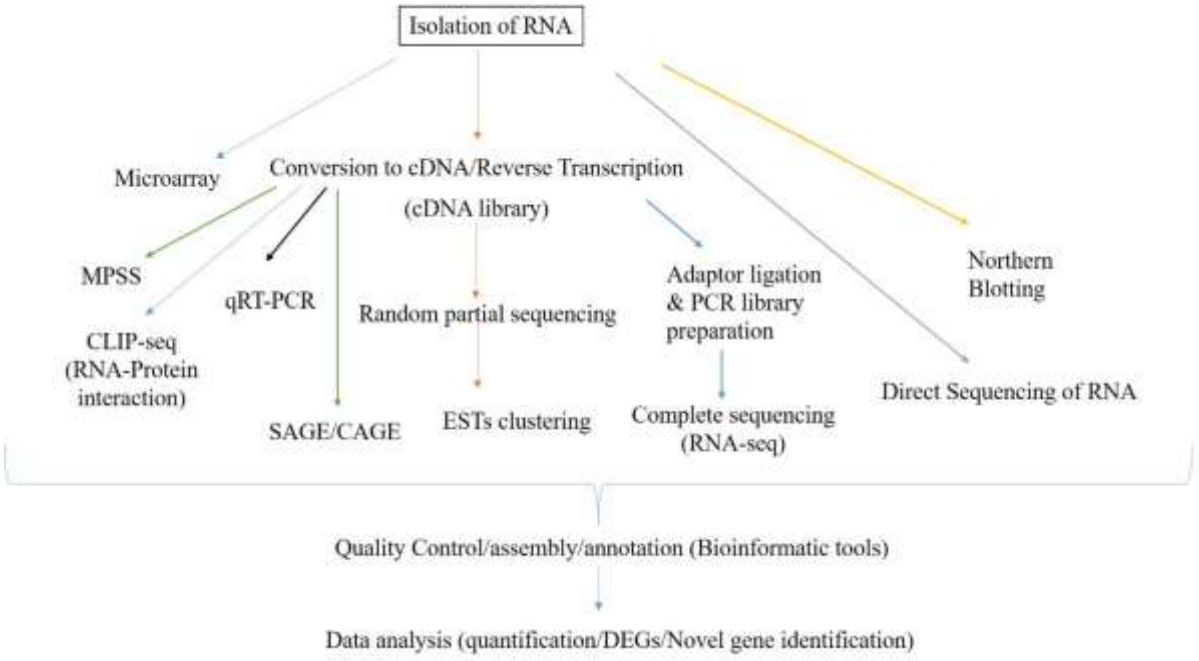



**Fig.1. Flow of transcriptome technology**

Now study on transcriptome has been shifted from gene specific and PCR based approach to high throughput sequencing based and global transcript-based strategies (Fig.1). Multiplexing and simultaneous data analysis is also possible with recent transcriptomics tools with the help of computational biology. Since 1977, so many technologies have been shown their potential uses and till date there are so many remarkable improvisations happed (Table 1).

**Reverse transcription Polymerase Chain Reaction (RT-PCR)/Taminism**:
RT-PCR is the basic process of transcriptomics which is universal to any tool require conversion of mRNA into cDNA using reverse transcriptase enzyme (an RNA dependent DNA polymerase) and oligo T (for mRNA having poly-A tail)/random hexamer/t-RNA as primer, before applying any other kind of technique to store, quantify, detect, clone and analyse RNA further. In order to better access the biological information of genome including location of open reading frames (ORF), 5′ and 3′ untranslated regions (UTR), and splicing sites, full-length and high-quality sequences of cDNAs is needed. Answering the biological events responsible for alternate gene structure like RNA editing, trans-splicing and alternate splicing also can be possible only if the complete cDNA sequence information is available.

**ESTs (Expressed Sequence Tags):**
ESTs are randomly selected single pass reads of clones sequenced from cDNA libraries, which is constructed from total RNA derived from a specific tissue or cell. Thus, the library indeed represents subtotal of genes expressed in that cell. In general EST consists of 300–1000 base pairs (bp) of DNA and is often deposited in a database to establish the identity of the expressed gene. EST analysis is very rapid and the massive sets of gene sequences that are expressed can be characterized efficiently. This approach was first used by Adams et al. to the screen a human brain cDNA library in 1991[1]. In comparison with whole genome sequencing techniques, EST is less costly and simpler, especially in organism with large genome size, it is very informative. The ESTs derived from any cell type of any organism also used as reference sequence for EST analysis of other organisms. ESTs have tremendous role in novel gene discovery and helps in unigene construction [2] which helps in reference genome assembly also. EST has potential pitfalls because it does not represent the whole transcript rather a partial gene expression profile is defined and sometimes miss rare transcripts also [3]. Moreover, EST data do not reveal many information regarding chromosome position, genome organization, gene structure, evolutionary origin, gene family or regulatory regions.

**Serial Analysis of Gene Expression (SAGE):**
Unlike ESTs, this is another method of gene expression analysis based on cDNA sequencing but using specific restriction cleavage followed by tagging those sequences, which in turn can be used to quantify the level of expression based on the quantity of tag present. SAGE has the major advantage over ESTs is the large amount of sequencing is not required (Supplementary figure 1). This technique was addressed by Veculescu et al. [4]. In this technique, mRNA with polyA tail is converted to cDNA using an oligo-dT primer (biotinylated) and digested with a frequent cutter restriction enzyme, such as NlaIII (tetra-cutter). The 3′ end of cDNA is then seized streptavidin by affinity and resulting in a pool of representative cDNA ends that is used to generate SAGE tags (9–14 nucleotide/short ESTs). Sequence similarity present in the pool also be captured since 3’ end sequence is diverse for all conserve gene family also. Two types of linkers (containing recognition sites for a type IIs restriction enzyme which cleaves outside the recognition site) are then ligated to the pool after splitting the sample into two groups. SAGE tags (attached as a part of the linker) are generated by cleavage. The tags can be ligated in “tail to tail” fashion to generate ‘ditags’ followed by PCR amplification (linkers act as primer annealing sites). NlaIII cleaves the amplified products to eliminate the linkers, and the ditags are concatemerized and cloned. Each tag represents the cDNA from which it originates and provides a guide to the comparative abundance of the different mRNA level. The major potentiality of SAGE is that the obtained data are absolute expression and are digitally represented. Now a days, Super-SAGE is used using NGS based sequencing of SAGE tags to increase throughput. Other modifications have also been created to increase the tag length up to 26 bp through different approaches, like, changing tagging enzyme, reverse SAGE (rSAGE approach), 3’ end cDNA amplification, and PATs (polyA tags) using restriction digestion. The rSAGE uses primers of 64 nucleotides including 30 Ts as linkers for RT-PCR followed by digestion with NlaIII enzyme and ligation with 5’ adapters. Both

5' adapter and 3' linker sequences are used to design primers that can amplify the long tags for sequencing further. The 3' end cDNA amplification process utilizes a 2-base anchored oligo(dT) primer with a heel (like a linker sequence) for first-strand cDNA synthesis, followed by the second-strand cDNA synthesis. The cDNA products are then cleaved with restriction enzymes and ligated to an adapter (Y-shaped), which completely blocks amplification of the Y-Y ligated products. Therefore, only 3' ends can be amplified for sequencing. In PAT method, digestion by restriction enzymes and switching primers which contain the cleavage sites are used in cDNA preparation along with linker. The cDNA products then are cleaved with NlaIII or TaiI followed by ligation with new adapter molecules that have overhangs complementary to the enzyme cleavage sites [5].

**Cap Analysis of Gene Expression (CAGE):**
Most of the transcriptome analysis rely on 3' end related sequence the transcripts and hence to determine and analyse the 5' end sequence, transcriptional start site as well as mapping of promoter region, Cap Analysis of Gene Expression (similar with 5'-SAGE) is a reliable option [6]. In this technique, 5' end of mRNA is captured and converted to cDNA through RT-PCR and cDNA is again trapped by biotinylated linkers having two restriction enzyme cleavage sites (Xmaj1 & Mme1). After second strand synthesis of cDNA 5'end get cleavage by restriction endonuclease MmeI (cuts DNA at 18-20 nucleotide downstream of recognition site with two base overhang). Second linker (having Xmaj1 cleavage site) is now ligated. PCR amplification using linkers specific primer sequences followed by cleavage using Xmaj1 RE helps in release of linkers from cDNA sequences. CAGE tags can be now concatenated to clone and sequence for further analysis. This tool is extremely useful in gene annotation, gene discovery and transcript expression profiling.

**Massively Parallel Signature Sequence (MPSS):**
MPSS is basically a hybrid of microarray technology and sequence sampling, in which a flow cell consists of millions of DNA tagged microbeads, are aligned and analysed by fluorescence-based sequencing technique (Supplementary figure 2). The principle [7] behind it is, cDNA clones are sequenced after rounds of cleavage with a restriction enzyme (type II), followed by adapter ligation, which (each adopter) is able to "decode" the rest of the bases at the overhang left by the restriction enzyme. Decoding is generally achieved by using conjugated labels, which is further analysed by flow cytometry technique. The potentiality of this technique lies on highly parallel nature which gives the ability to simultaneous analysis of fluorescent signal coming from thousands of microbeads in a flow cell and resulting the high degree of throughput.

**Rapid Amplification of cDNA Ends (RACE):**
Cloning of full length of a transcript is always a challenging task and RACE is one of those powerful technique which allows amplification of a messenger RNA template between a well-defined (known sequence) internal site and to sequences at either the 3' or the 5' -end of the mRNA is unknown. In this procedure [8], cDNA is prepared from mRNA transcripts using RT-PCR with oligo-dT primer followed by amplification of cDNA using an internal gene-specific primer (GSP) that complements to a region of known sequences and a primer that complements to the poly(A) tail region. This allows to capture the region between unknown 3'-mRNA end sequence (called 3' RACE). Similarly, 5' RACE (also called anchored" PCR) also facilitates the isolation and characterization of 5' ends sequences. This is very useful in full length cloning and transcript characterization in when high GC content at 5' end does not allow typical PCR reaction to amplify fully.

**Northern Blot:**
This is one of the reliable transcriptome techniques was developed by James Alwine, David Kemp, and George Stark at Stanford University in 1977 [9, 10] to measure the size and amount of a specific mRNA transcript expressed in a tissue by hybridization using radioactivity or fluorescence probe. In this blotting technique, denaturing gel is used to separate RNA molecules according to their size followed by transferring them to a nylon membrane using 20X SSC (Saline Sodium Citrate) buffer and capillary action, keeping the same orientation of the gel. RNA is then fixed to the membrane and hybridization is carried out with immobilized RNA and labelled probe (designed using complementary sequence of target mRNA/gene of interest) using hybridization buffer. Non-specifically bound probes are then removed through washing and the membrane is then dried for analysed. This technique can be effectively used to detect different transcript variants of a gene without involving PCR or sequencing techniques. However, if the quantity of total RNA of interest present in the sample is low this technique cannot be used rather more sensitive tool is needed like such as semi-RT-PCR or quantitative RT-PCR etc.

**Differential Display Reverse Transcription PCR (DDRT-PCR):**
This is basically a PCR-based method that allows widespread analysis of gene expression among several cell populations at a time for a comparison-based study. Using this tool identify, isolation and cloning of differentially expressed genes (mRNA) can be effectively carried out by means of reverse transcription followed by PCR (supplementary figure 3) from tissue under altered conditions (stress/any treatment). The key element in this technique is the use a set of oligonucleotide primers. The first step is RT-PCR of samples (mRNA) isolated from different conditions using a set of degenerate oligo-(dT) primers to anchor the polyadenylate tail of mRNAs to generate a pool of cDNA. The second step involves PCR amplification of random partial sequence from that cDNA pools using short arbitrary primers (generally using arbitrary decamers) and resolved in denaturing polyacrylamide gel. The reproducible patterns of amplified fragments can be obtained when multiple primer sets were used (multiplexing) and it is due to sequence specificity of either primer during PCR. In general, the anchored oligo-dT primers consist of 11-12 no of Ts with two additional bases at 3' region of primer (i.e., four set primers (T12XY) where X, Y can be G, A, T or C) to provide specificity of amplification. To increase the sensitivity and specificity, completely DNA free RNA with only ploy A tail was recommended for to reduce false positive and robust use of this technique [11].

**Suppression Subtractive Hybridization (SSH):**
This technique was first conceptualized by Siebert and Larrick in 1992 [12] and called competitive PCR. Later some modification was carried out to increase efficiency and SSH was developed. This is a highly effective method used in construction of subtracted cDNA libraries [13], which combines two steps of normalization and subtraction in a single process to equalize the

cDNA abundance within target population and exclude common sequences of target/tester and driver population respectively (Supplementary figure 4). SSH method is based on reassociation kinetics of DNA molecule. This technique drastically enriches the low abundance cDNAs many folds that obtained from differentially expressed mRNAs. This subtracted cDNA library can be used effectively as probe for capturing homologous sequences from any cDNA/genomic library and can be useful in positional cloning also. As per protocol of SSH, control samples are considered as 'driver' and treated samples as 'tester'. The mRNA first isolated from both the samples and cDNA are prepared using RT-PCR. The tester cDNA is then digested using restriction enzyme to generate blunt end and tester cDNA get divided into two equal parts followed by ligation of two adapters (each part is ligated with one kind of adopter). Tester cDNA is amplified using adaptor specific primers followed by denaturation and driver in high quantity is then added to the reaction to generate driver-tester hybrid of over expresses/equally expressed transcripts (normalization). Further in the second cycle of PCR, both tester (having different adaptors) is mixed and amplified using adaptor specific primers. This leads to amplification of differential expressed genes. The reason behind applying excess quantity of driver cDNA is to suppress the amplification of common cDNA present in driver and tester to allow only amplification novel/unique transcript present in tester.
Another similar and cost-effective method was developed called Negative Subtraction Hybridization (NSH) method [14], in which a plasmid library is prepared from cDNAs of cells grown under a particular condition and are screened by probes (these probes are basically cDNAs sequences prepared from transcripts isolated from cells that are under another physiological condition). Plasmids with inserts that are not able to do hybridization are representative of differentially expressed/rare transcripts. This method is very useful in novel gene identification also.

**Real Time Polymerase Chain Reaction (qRT-PCR):**
Real-time PCR has become one of the most reliable, tremendous sensitive and widely used methods of gene quantitation, transcript profiling, gene copy number detection, allele discrimination, pathogen detection, SNP genotyping etc. It is highly sequence-specific and amenable for increasing throughput as per need. This technique involves quantification of amplicon (PCR products) during PCR cycles only unlike conventional PCR reactions (that is why the term 'Real time' is used). The theory of quantification of mRNA transcripts in Realtime PCR can be defined in two ways; i.e., one-step reaction, where the complete process of cDNA synthesis as well as PCR amplification is done in a single tube otherwise, in a two-step reaction, where cDNA synthesis and PCR amplification occur at separate reactions. One step qRT-PCR is thought to be more precise since it minimizes the experimental variation, but one major drawback is that, mRNA is susceptible to rapid degradation. However, two-step process is also very sensitive and highly reproducible. Chances of primer dimer detection is also reduced through manipulation of melting temperature (Tm) in two step processes.
In real time PCR, quantification/detection depends on two most important factors, types of probes and method of quantification.

**Probes:**
**SYBR Green Dye:** It is the most routinely used dye that binds at double stranded nucleic acid and emits light upon excitation and gives information about the quantity of double stranded DNA molecule present in the samples after each cycle of amplification. SYBR Green is relatively less expensive compared to other fluorescent probes. However, since SYBR dye also gives signal in non-desired PCR product and primer-dimers [15], sometimes overestimated result give rise to inaccuracy and this problem can be somehow solved by analysis of dissociation curve/melting curve (fluorescence vs. melting temperature graph).
**TaqMan probes:** It is more sensitive and effective fluorescently labelled probe where a reporter and a quencher are attached to the 5' end and to 3' end respectively. During denaturation step in thermal cycler, this probe binds with target template and during strand elongation DNA polymerase cleaves the probe by 5' exonuclease activity, which leads to detachment of quencher and reporter and fluorescent is detected through FRET (Fluorescence Resonance Energy Transfer; distance-dependent interaction between two fluorophores) [16].
**Molecular beacon probes:** Unlike TaqMan probe, this probe does not get cleaved during amplification reaction rather emits fluorescence when do hybridization with target molecule at denaturation step of PCR [17].
**Scorpion primer**: The reaction kinetics is so fast in this kind of probe. In this case, a hairpin-loop structure with a reporter at 5' end and an internal quencher, which directly linked to the 5' end of the PCR primer through a blocker. The role of blocker is to prevent the strand elongation. At the first cycle of PCR, the DNA polymerase extends the complementary strand but at next cycle, the hairpin-loop unfolds while denaturation and the loop-region of the florescence probe binds with newly synthesized target molecule at complementary bases. Now that the reporter and quencher are separated, and fluorescence emission is detected which is directly proportional to the amount of target DNA in the sample [18].

**Types of Real-Time Quantification**
**Absolute Quantitation**
Absolute quantitation involves serially diluted standards (concentration is known) to generate a standard curve. The standard curve generates a linear relationship between $C_t$ (cycle threshold value of a reaction is the number of PCR cycle when the fluorescence is detected above the background signal), and initial amounts of the target cDNA. This allows to determine the unknown sample's initial concentration as per graph plotted against Ct value detected during real time PCR.
**Relative Quantitation**
In relative quantitation, comparative gene expression is measured based on a reference sample (any housekeeping gene can be taken as reference), also called calibrator. Results can be expressed as target/reference ratio. There are various analytical models available to calculate the mean normalized gene expression but $C_t$ ($^{2-\Delta\Delta Ct}$) method [19] is widely used and most accepted model where difference in Ct values between reference gene expression as well as target gene expression with respect to control is calculated to determine $\Delta\Delta$Ct value.

**Microarray:**
Microarray based transcriptome analysis is an important tool for characterization and understanding of the gene expression and analysing the transcriptional profile of genes at a genome level. This is a hybridization-based tool which combines simultaneous analysis of the RNA expression of thousands of genes at a go using probe hybridization and fluorescent dye. The potentiality of this

technique lies on throughput, automated and acquisition of quantitative data from multiple samples quickly. DNA microarrays can be assayed in two ways: genomic library/cDNA library/ESTs representing each gene in each dot of solid surfaces is fixed prior to hybridization (called probe) which is called spotted/DNA microarray and 2) in situ synthesis of oligonucleotides, when identity and sequences of each transcript are pre-known, using photo-lithographic technique which is called oligonucleotide microarray/DNA-chip/bio-chip assay. In the chip, each probe has assigned location and gene ID. cDNA/mRNA molecules from experimental sample are now hybridized with DNA template (or probe) present on surface of chip. Upon hybridization fluorescent signal is detected which is representative of the amount of cDNA/mRNA bound to each site on the array which indeed explains the degree of expression of these genes. Detector then generates a profile of each gene got expressed. Though the first computerized image-based analysis was invented by Brown. P., 1981, DNA-microarray was first discovered in 1988 by Edward and in 1995 Mark Schena was the first to use it for human gene expression studies [20]. For differential gene expression analysis also, it is a widely used tool. The comparative effects on gene expression of different conditions like, any treatment, disease, stresses and developmental stages can be easily detected and analysed (multiplexing). Detection of SNPs are also possible through microarray technology.

**RNA-seq and data analysis:**

RNA-seq one of the most powerful, high throughput transcriptomics tool which provides the sequence information as well as exact quantification of expression level of transcripts. In this technique, transcripts are converted to cDNA and are adopter ligated followed by sequenced using next-generation sequencing (NGS) platforms, like, e Illumina IG, Applied Biosystems SOLiD22 and Roche 454 Life Science, e Helicos Biosciences tSMS system etc. The sequencing process may follow either single-read or paired-end sequencing methods depending on depth requirement, time, and cost. Millions of short reads produced are then assembled either using de novo or reference sequence algorithms using alignment tools (Bowtie, TopHat, STAR, BWA, Novoalign, HMMSplicer, Olego, BLAT etc.) and quality control tool (FastQC, Picard, RNA-SeQC, or Qualimap etc. to check PCR Bias, contamination, GC bias, length optimization etc.) followed by remediation of poor quality read-ends (Trimmomatic, fastx-toolkit, etc). But, the analysis of NGS data generated in RNA-seq is one of the tedious and time-consuming tasks. After alignment (Supplementary figure 5), analysis tool like Cufflinks, CLC genomics, Sailfish, RSEM and BitSeq etc. help in transcript identification/quantification of expression level and other tools, like MISO, can be useful for determination of alternate splice site or any other advance analysis. The inherent difficulty in NGS data is non-uniform read coverage and biased read mapping (because transcripts which are longer and highly expressed are more likely be biasedly represented among RNA-seq reads), when expression level is same for both. Hence, software is used to transform the data into RPKM (Reads Per Kilobase per Million mapped reads) or FPKM (Fragments Per Kilobase per Million mapped reads) to normalize the read counts [21]. The most useful information one can get from transcriptome profiling is about differentially expressed genes (DEG) when comparing RNA-seq data between two or more samples. For differential expression analysis, Cuffdiff, DESeq EdgeR etc. package is useful. But all the packages rely on statistical analysis based on Poisson or negative-binomial models and are heavily influenced by “outliers” in the data; if the sample library size is not same, a biased result may come out. Therefore, tools based on non-parametric statistical methods have also been proven to encounter these problems. Fold Change based analysis of gene expression pattern gives more reproducible results. So, NOISeq [22], Samseq [23], LFCseq etc. packages are more reliable in differential gene expression study. However, till date, RNA-seq is the most useful, widely used and irreplaceable tool, which provides the highest level of fidelity in analysis. Potentiality of RNA-seq lies in detection and quantification the expression of many novel/unknown transcript, eQTL mapping [24] and fusion gene detection [25] that might not have been reported previously. It has selective advantage over microarray technique is that, in RNA-seq read count are non-negative integers and thus inherently follow a discrete distribution, in contrast, in microarray the degree of expression is recorded as continuous measurements of intensity and follow a log-normal distribution. Beside this, RNA-seq is superior in respect of low background noise and it is not limited to unknown genomic regions (in microarray, probe designing is from known sequence only). Hence, RNA-Seq provides a comprehensive, quantitative, as well as unbiased view of transcript sequences in the sample.

**Single-Molecule Direct RNA Sequencing:**

Helicos BioSciences has developed and commercialized by the Helicos® Genetic Analysis System. This technique involves completely unbiased direct sequencing of single RNA molecules in a massively parallel sequencing pattern, providing an accurate quantitation as well as sequence information in real time. In this technology, conversion of RNA to cDNA is not required and provides deep sequence coverage. For expression profiling of transcribed mRNA (protein coding genes) or to map the polyadenylation sites this approach is widely used. In this process, mRNA molecules with polyA tails are hybridized with the poly(dT) primers, which are immobilized onto the flow cell and allow sequencing-by-synthesis of transcripts in parallel. The average read length is 35 nucleotides and one complete run may yield 800,000–8,000,000 reads per channel, and there are up to 50 channels in the 2 flow cells that can be run in parallel. The error rate can be up to 5 % (majorly deletions and insertions).

**CLIP-Seq (Crosslinking and Immunoprecipitation-Sequencing):**

Almost all RNA molecules are subject to some degree of post-transcriptional gene regulation (PTGR) involving capping, polyadenylation, splicing, editing, degradation and transport to other organelle, translation, stability etc. The high throughput sequencing technologies enabled the invention of new methods to map the interaction sites between RNA-binding proteins (RBPs) and RNA sequences. Hence, crosslinking and immunoprecipitation (CLIP) methods (Supplementary figure 6) are widely adapted tool in transcriptomics after RNA-seq for identification of target RNA-binding sites in large-scale to study the protein or RNA in question. In this process, after crosslinking of RNA-Protein complexes (UV crosslinked), fragmentation using RNase1 digestion and immobilization in antibody-coupled beads, dephosphorylation is done using phosphatase as initial step followed by ligation of barcoded adaptors at 3’end of RNA and transferring to the protein gel for electrophoresis. Then it is transferred to nitrocellulose membrane and after treating with Proteinase K (to separate RNA from protein), reverse transcribed into cDNA. RNA is then removed from cDNA-RNA hybrid and adaptor is ligated to cDNA at 3’ end and product is PCR amplified for sequencing library preparation and sequence is further analysed for size detection and identification of protein binding site. So many modifications have been done in this technique to reduce the PCR duplication and background noise in order to detection of specific single base

interaction. This is very powerful tool for elucidation of different transcriptional events where Protein-RNA interactions are involved including sub-cellular localization study.

Another recent technique was invented to detect RNA-binding proteins (RBPs) in living cells, called CARPID (CRISPR-Assisted RNA–Protein Interaction Detection) is also very useful, robust and fast tool, developed by [26] based at City University of Hong Kong. This technique utilizes CRISPR-dCasRx for specifically targeting lncRNA-Protein based detection as well as proximity labelling. Fusion of the dCasRx (for specific lncRNA targeting) along with engineered biotin ligase BASU (for capturing in streptavidin beads), self-cleaving T2A peptide and an enhanced green fluorescent protein (eGFP) helps in tracking the expression of transcripts in living cells. This method has already been utilized in human transcriptomics but still must be optimized in other organisms including plant system.

**Sequencing technologies for Transcriptomics:**

A number of methods can be grouped broadly to fit into each sequencing technologies so far like, preparation of template (library), sequencing (by synthesis/ligation) and imaging followed by data analysis. But, some point of uniqueness exits in each technology and that determines the cost and quality of data produced from these platforms.

The template is prepared from whole transcript, cDNA, single mRNA etc. and can be randomly sheared by nebulization, sonication, or enzymatic digestion to produce fragments of suitable size (if transcript size is large) followed by adaptor ligation (these are attached to each fragment at both 3' and 5' ends and they facilitate purification, amplification, and sequencing methods). The fragments are now attached to a single capture bead along with PCR enzymes and reagents. Then these are enclosed in mixture of water-in-oil droplets (called emulsion), which forms an independent isolated micro-reactor, where PCR will take place.

**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 beads. Pyrosequencing principle (luciferase-based light detection on pyrophosphate release when a base is added in sequencing process) is used for sequencing [27]. From template preparation to data processing the FLX system (read length 450-500 bases) takes 10 hours per run (generates 400-Mb sequence data) with consensus accuracy is more than 99.99 %. Recently developed GS FLX Titanium XL+ platform can generate 1 Gb sequence data with read length of 1000bp in same time.

**Illumina/Solexa:**

Illumina was commercialized in 20 and it is the most widely used NGS technology till date. Few recent platforms like, Illumina Genome Analyzer 1 Gb and HiSeq 600 Gb are very popular. In this technology, two different adapters are ligated to their 5' and 3' ends of the transcripts (cDNA fragments) and these fragments are then attached to a substrate on a flow cell (contains a dense lawn of primers which are complementary to adopter ligated). The solid phase PCR is then carried out inside flow cell, also called fold-back PCR or bridge PCR [28], which is very fast and works on reversible terminator technology (the four fluorophores linked dNTPs used for PCR amplification and these fluorophores serve as chain terminators). The CCD camera keeps the record of dNTPs added each time at 3' end of the growing chain primers. Illumina read length generally varies from 35 to 150 bases, and the accuracy is more than 98.5 %. Illumina HiSeq 2000 platform yields 400 Gb of sequence data in a single run (takes 7–8 days). The HiSeq X Ten is a widely used population-scale whole-genome sequencing platform (launched in 2014) and it generates 1.8 Tb sequence data in 3 days. The system can be used for de novo transcriptome sequencing when reference sequence is not available, resequencing to genotype SNPs/InDels, copy number variation (CNV) can also be detected, and structural variation of gene isoforms, transcript expression profiling, etc.

**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 of in 2005 as SOLiD 3.0 platform [29, 30]. In this technology, beads are prepared in the same manner as discussed earlier and are immobilized in a single layer 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; e.g., "TT" is assigned to blue, "TC" is assigned to green, and so on for all 16 unique oligos. During sequencing, each base in the template is sequenced twice and hence SOLiD technology is said to be 100% accurate. The SOLiD 3.0 platform yields read-length of 50 bases only and can generates approx. 20 Gb sequence data per run in 6–7 days. SOLiD 5500 and SOLiD 5500 XL systems were introduced to increase the sequence data of up to 300 Gb per run at 99.9 % accuracy [31].

**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 [32]. 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 sequential addition of each nucleotides allows base discrimination through non-optical scanning and therefore, speeding up the sequencing process dramatically and reduce cost also. In 2012, another high throughput technology was released, called 'Ion Proton', which increases output an order of magnitude of 10X but read length was drastically reduced in comparison to Ion Torrent (200 bp instead of 400bp).

**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). In this technology, 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 [33]. Fluorophore is automatically removed and the signal is detected when a dNTP is added to the primer

3'end at growing chain. The cDNA molecule can be sequenced so many times to increasing the accuracy of sequencing result. Moreover, direct sequencing instead of clonal multiplication allows the sequence to be read in real-time [34]. Each SMRT cell can generate ~50 k reads and up to 1 Gb of data in 4 hr and less sensitive to GC contents of sequence but its error rate is little high (~11%) due to single pass sequencing approach.

**Oxford Nanopore Technologies:**

In this advance sequencing technology, a sequencing flow cell composed of hundreds of micro-wells containing a synthetic bilayer and punctured by biologic nanopores [35]. Sequencing is achieved simply by precise measuring the changes in current induced as a result of incorporation of bases through the nano-pore with the help of a molecular motor protein. Library is prepared by ligating adopters to cDNA ends in a manner that, first adopter can bind with motor enzyme and second adopter (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 hr) 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) [36].

**Application of advance transcriptomics in agriculture:**

Study of transcriptome, as a branch of omics-science, offers a crucial platform to link the genotype with phenotype and provides a clear understanding about underlying metabolic network that controls cell fate, development, and disease progression etc. It has great impetus in different field of biological science including medicine, system biology, environmental study, agriculture, metagenomics and many more. In the agriculture, transcriptomics plays a key role in understanding the fundamentals of molecular biology led to the improvement of crop genetic architecture and characterization of genes for desired phenotype (Fig.2).

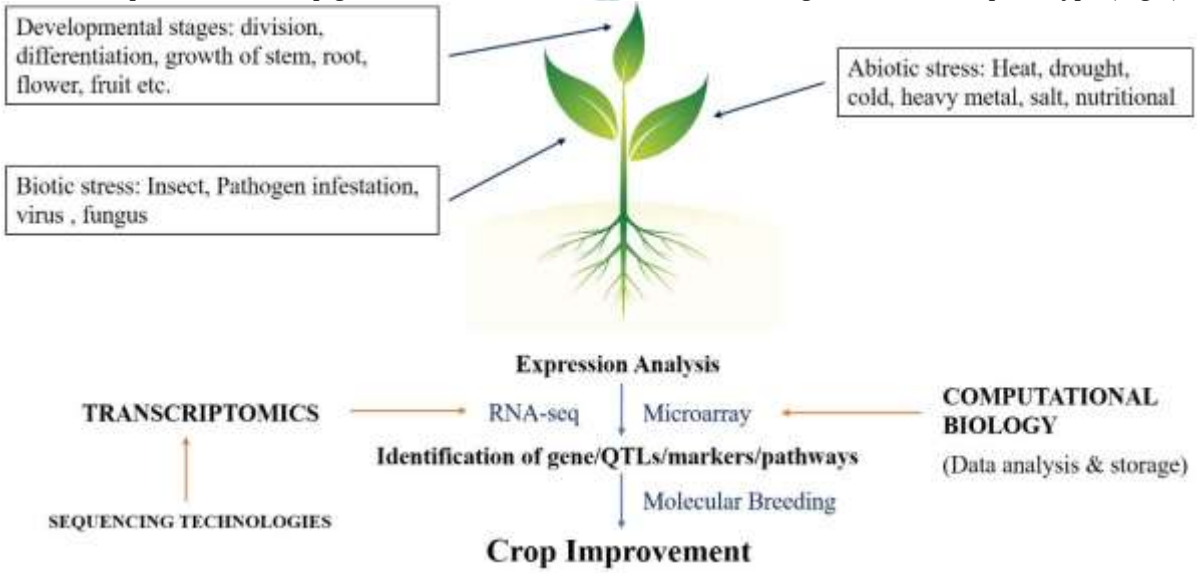



**Fig.2. Transcriptome analysis in plants**

**Genome assembly, annotation and differential gene expression, gene isoform study using RNA-seq**

The complexity in gene regulation at transcriptional level has a great impact on phenotype, especially in eukaryotic system. Zhang et al., tried to elucidate the global view of transcriptome of different organs of cultivated rice through deep sequencing approaches by using high-throughput paired-end RNA-seq [37]. A total of 28 Gb sequencing data was obtained through Illumina technology, which was ~67-fold of the size of rice genome (67X coverage). After alignment of reads onto the reference (*Oryza sativa* subsp. *Indica*) genome, it was found that ~73% of the reads were mapped uniquely to the genome. They revealed that deep sequencing of the rice transcriptome covered more than 99% of the available rice full-length cDNA data. They followed reads per kilobase per million mapped reads (RPKM) for quantification of gene expression. After completion of mapping, 38,650 transcript units including ~180 non-coding RNA had been found out. They also identified more than 10000 new exons using this pair-end approach with the precise 3' and 5' UTR (untranslated region of genes) of more than 29000 genes, that was previously unknown. Computational analysis revealed that 33% of rice gene undergoes alternate splicing of which more than 50% have multiple alternate splicing sites indicating the high complexity in rice transcriptional regulation. Full length cDNA and ESTs analysis was unable to detect these alternate splicing sites before unlike RNA-seq data. Transcriptional fusion events (give rise to fusion protein), that generates from combining transcripts originating from different gene, was also be traced using RNA-seq data of rice (Fig.3). They identified more than 230 transcriptional fusion events and this information helps to understand the exon-shuffling mechanism in rice.

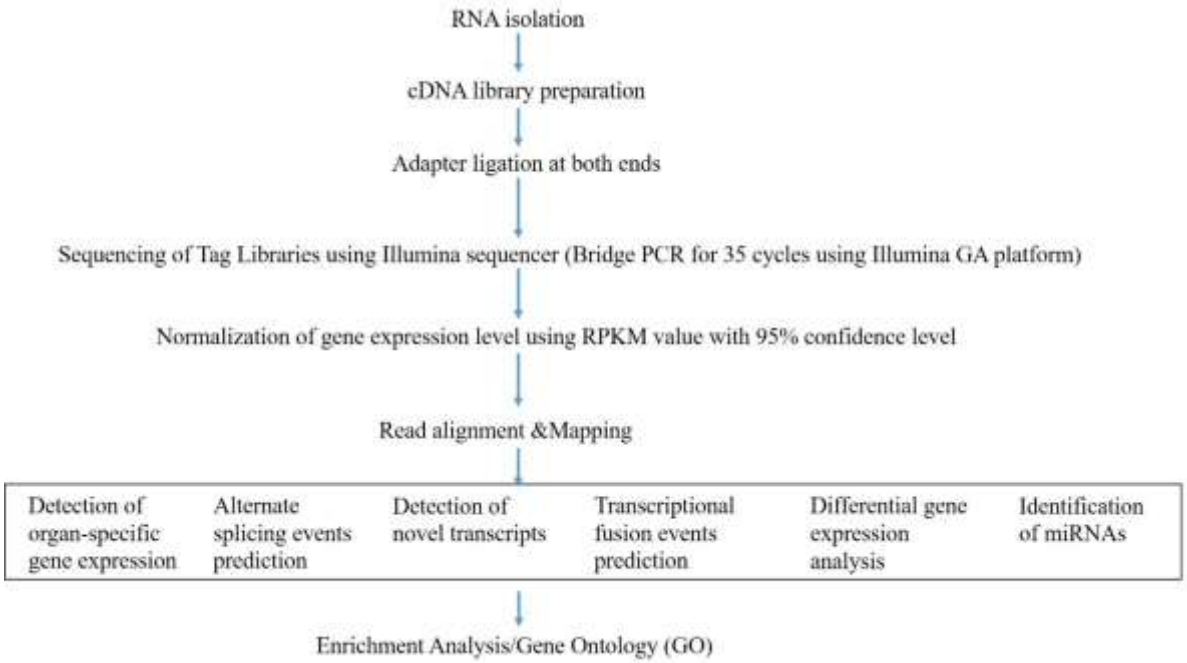



**Fig.3. Genome assembly, annotation and differential gene expression, gene isoform study using RNA-seq**

The organ-specific gene expression was calculated using index τ value developed by Yanai et al. [38]. Then reads were aligned using SOAP [39]. To detect novel transcripts, bases that was supported by minimum of four reads with 35 bp or longer size were considered to be transcriptionally active region (TAR). The TARs that were connected by at least one set of sequenced paired-end reads were joined into a transcript unit (TU) (TUs with length <150 bp or whose average expression of <10 reads per base were excluded in analysis to increase sensitivity of detection). Out of 38,650 TUs detected, 7232 TUs were considered as novel TUs, because they were not overlapping with genomics transcripts previously reported using annotated gene model and tiling array. To check whether these TUs were due to transposal activity or not, they blasted those TUs against the repetitive DNA sequences (Repbase database) and TUs with <50% of repetitive sequences were further analysed for gene prediction using by Augustus tool [40]. To identify candidate non-coding miRNAs, they used MIREAP tool (https://sourceforge.net/projects/mireap/) and compared with MiRNA database (miRbase: http://microrna.sanger.ac.uk/sequences/, release 13.0) for elucidate novel miRNAs. Alternate splicing events was predicted by detecting splice junction sites followed by aligning them when reads cannot be mapped in genome. Gene Ontology (GO) or gene enrichment was achieved using Cytoscape software [41] (http://www.cytoscape.org, version 2.5.2). After that, BLAST was performed to find homologous protein to check the conserve nature of alternate splicing events. Transcriptional fusion events were detected using trans-splicing junction alignments of those reads which neither be mapped in genome nor can be aligned in alternate splicing site. These sites can be validated using real time PCR.

**1. eQTL detection and identification of genic markers (SNPs/InDels):**

Analysis of expression QTLs (eQTLs) led to understand the transcriptional regulation in many plant species including interaction of pathogens or insects with host plants. Unlike traditional QTLs mapping, where genome sequence directly associated with trait, the eQTLs expression is linked with the degree of expression of a sequence not directly linked the phenotype as such. eQTL can be cis-acting or trans-acting depending on distance of eQTL from gene in question. Co-expression and co-location analysis help in eQTLs identification. Identification of eQTLs has been routinely followed in molecular biology to detect exiting interactions between genes and to determine the regulatory mechanism of gene expression (Fig. 4).

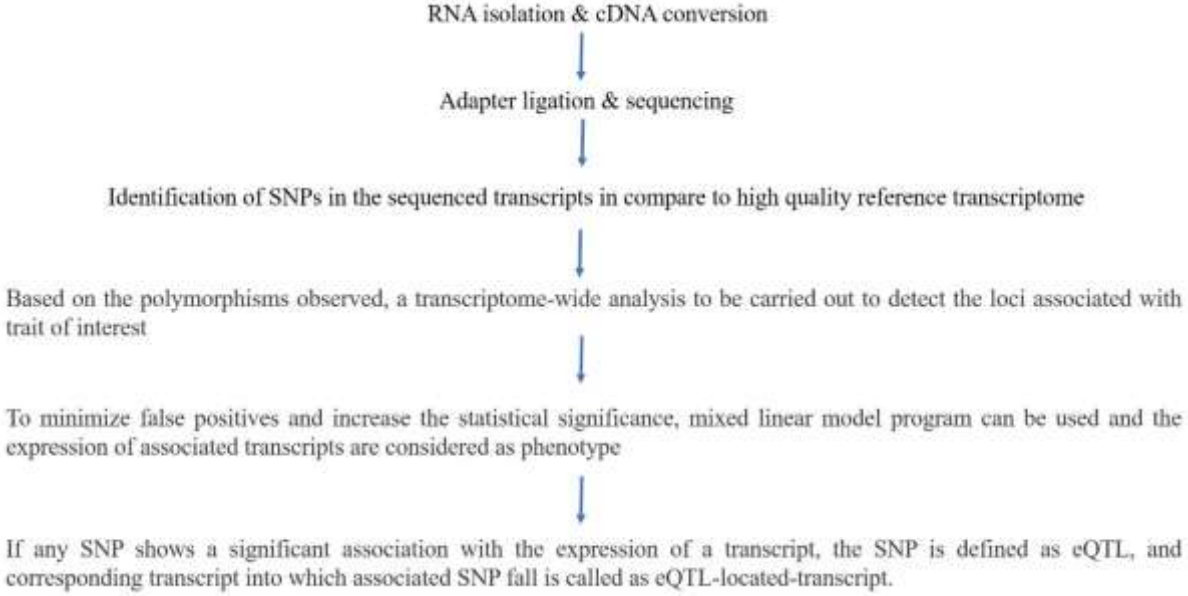



**Fig.4. eQTL detection process**

When both the sequence and transcript expression level are associated with trait and a positive co-relation exists between the SNPs (any marker fall in the transcript sequence) and the expressed target transcript, they are grouped together as co-expression module. eQTL helps to identify function of coding sequences in terms of enzymatic activity, epigenetic study, transcription factor function, ribonucleoprotein function etc. Those transcripts then can be characterized in wet lab and be used in crop improvement programme. Holloway reported the genome-wide eQTL analysis of hydroponically grown maize to study the cis & trans acting eQTLS regulating the gene expression in roots [42]. Identification of a candidate gene underlying a trans-eQTL demonstrated the feasibility of eQTL cloning in maize and could help to understand the mechanism of gene expression regulation.

In terms of genic SNP marker and InDel marker development, transcriptomics plays a crucial role even when whole genome sequence information is not available. A genome-wide novel genetic markers were identified by RNA sequencing and assembly of reads from ten diverse *Aegilops tauschii* (wild Triticeae) accessions in comparison with available barley and *A. tauschii* reference sequence [43]. To efficiently utilize the high natural variation present in D genome of common wheat and to eliminate the complexity exist in the genome, to capture important agricultural traits and utilize them in cereal breeding, advance NGS-based transcriptome analysis along with computational biology are indispensable. More detail genetic map can be constructed using more no of SNPs detected. One drawback of using transcript sequencing information in InDel identification is that, InDel causes frame-shift mutation resulting transcript variation from single ORF, and detection of InDel is more robust when whole genome sequence is available.

**2. Understanding the molecular mechanism behind disease establishment, insect feeding and potential drug discovery:**

Transcriptome profiling and comparative transcriptome analysis helps to understand the molecular changes happen when any pathogen or insect attacks a host plant and the response in terms of host-pathogen/insect interactions. For example, a comparative transcriptomic profile was analysed and was reported that, the passive defence strategies exist against the insect and phytoplasma in the susceptible cultivars of grapevine against yellows disease which is caused by phytoplasmas and transmitted by leafhopper (*Scaphoideus titanus)* [44].

Discovery of biomarkers, that helps in determination of an exposure/changes in a cell and therefore opens the arena of drug discovery and targeting, can also be achieved through whole genome transcriptome profiling. Several findings have already been reported regarding target molecules in plants and the response against elicitor upon infection/infestation by biotic factors. In a publication r et al. identified the induction of the chitin elicitor receptor kinase gene (*HvCERK1*) which confer resistance against *Fusarium graminearum* in barley [45]. Similarly, transcriptome data with metabolome data can generate more stringent information. Virulence and survival in Xoo pathogen (leaf blight in rice) is modulated by Xoo flux released based on carbohydrate metabolism state and when plants are exposed to nitrogenous fertilizer. In 2020, group of scientists tried to identify few anti-bacterial targets of leaf blight through metabolome and transcriptome data analysis [46].

**Table 1. Progress in transcriptomics as a functional genomics in plant science**

| Technology Intervened | Year of intervention | Reference | Comment(s) |
|---|---|---|---|
| Northern Blot | 1977 | [9] | Gene specific detection, not applicable for global gene profiling |
| Sanger sequencing | 1977 | [47] | First sequencing platform but very slow and costly |
| RT-PCR | 1984 | After PCR discovery by Kary Mullis [48] | For cDNA synthesis from mRNA, routine transcriptome work |
| Microarray/Affymetrix/DNA-chips | 1988 | [20] | Gene expression profiling & differential gene expression study but expensive and prior sequence information required, so much background noise |
| RACE | 1990 | [8] | For cDNA end information, not useful for global transcript profiling |
| ESTs | 1991 | [1] | High throughput single pass partial cDNA sequencing; now EST-clusters (unigene) used |
| Competitive PCR | 1992 | [12] | For differential gene expression analysis, not used recently |
| Antisense/Co-Suppression | 1992 | [49] | Functional transcript knocked down, targeted approach, now become obsolete |
| Improved DDRT-PCR | 1993 | [11] | Differential gene expression study, target specific approach, not useful in organism level |
| SAGE/CAGE | 1995 | [4] | Representative partial sequencing of transcripts, tags give useful information about cell/tissue specific transcript profile. |
| SSH | 1996 | [13] | Identify novel gene, very useful tool but not amenable for whole transcriptome level |
| RNAi | 1998 | [50] | Targeting mRNA for functional validation, gene specific approach |
| MPSS | 2000 | [7] | Sequencing throughput accelerated, useful for cell level when using NGS technology |
| qRT-PCR/Real Time PCR-based analysis | 2001 | [19] | Quantification of mRNA expression, gene specific approach and widely used. |
| Other Next Generation Sequencing (NGS) platforms | 2004 onwards | - | High throughput sequencing platforms, few are obsoleting like Rosche and few are cost demanding like SOLiD, but Illumina is still cost-effective, demanding and widely used |
| 454 sequencing | 2005 | [51] | Used for transcriptome study, not much useful now a days |
| Single molecule transcript sequencing using Helicos platform | 2009 | [52] | No need of library preparation, fast |

| | | | |
|---|---|---|---|
| Direct RNA-seq | 2009 | [53] | Advance RNA-seq approach, no need of library preparation, very fast and cost effective |
| Third generation sequencing platforms | 2009 onwards | - | Pacific Biosciences, Oxford Nanopore Technology, Quantapore (CA-USA), and Stratos (WA-USA) |
| STTM (Short Tandem Target mimic) | 2012 | [54] | Targeting mi/si-RNA to assign function of them, transcript specific approach, not widely used |
| Advanced CLIP-seq | 2016 | [55] | RNA-Protein interaction, fast, very useful but not for global transcript level |
| CRISPR-Assisted RNA–Protein Interaction Detection (CARPID) | 2020 | [56] | Advance, High throughput detection, fast, but still no report in plant transcriptome analysis |

**Reference:**


1. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., & Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252(5013), 1651-1656.
2. Zhang, B. H., Pan, X. P., Wang, Q. L., George, P. C., & Anderson, T. A. (2005). Identification and characterization of new plant microRNAs using EST analysis. *Cell research*, *15*(5), 336-360.
3. Parkinson, J., & Blaxter, M. (2009). Expressed sequence tags: an overview. Expressed sequence tags (ESTs), 1-12.
4. Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, *270*(5235), 484-487.
5. Shen, C.-H. (2019). *Genome and Transcriptome Analysis. Diagnostic Molecular Biology*, *303–329.* doi:10.1016/b978-0-12-802823-0.00012-2
6. Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., ... & Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nature methods*, *3*(3), 211-222.
7. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., ... & Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, *18*(6), 630-634.
8. A, Michael A. "RACE: rapid amplification of cDNA ends." *PCR protocols: A guide to methods and applications* 28 (1990).
9. Alwine, J. C., Kemp, D. J., & Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences*, *74*(12), 5350-5354.
10. Wang, Z., & Yang, B. (2010). Northern blotting and its variants for detecting expression and analyzing tissue distribution of miRNAs. In *MicroRNA Expression Detection Methods* (pp. 83-100). Springer, Berlin, Heidelberg.
11. Liang, Peng, and Arthur Pardee. "Distribution and cloning of eukaryotic mRNAs by mean of differential display: refinements and optimization." *Nucleic Acids Research* 21, no. 14 (1993): 3269-3275.
12. Siebert, Paul D., and James W. Larrick. "Competitive Pcr." *Natur* 359, no. 6395 (1992): 557-558.
13. Diatchenko, Luda, Y. F. Lau, Aaron P. Campbell, Alex Chenchik, Fauzia Moqadam, Betty Huang, Sergey Lukyanov et al. "Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries." *Proceedings of the National Academy of Sciences* 93, no. 12 (1996): 6025-6030.
14. Ray, Anamika, Sunita Macwana, Patricia Ayoubi, Leo T. Hall, Rolf Prade, and Andrew J. Mort. "Negative subtraction hybridization: an efficient method to isolate large numbers of condition-specific cDNAs." *BMC genomics* 5, no. 1 (2004): 1-11.
15. Morrison, T. B., Weis, J. J., & Wittwer, C. T. (1998). Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. *Biotechniques*, *24*(6), 954-8.
16. Dooley, J. J., Paine, K. E., Garrett, S. D., & Brown, H. M. (2004). Detection of meat species using TaqMan real-time PCR assays. *Meat science*, *68*(3), 431-438.
17. Elsayed, S., Chow, B. L., Hamilton, N. L., Gregson, D. B., Pitout, J. D., & Church, D. L. (2003). Development and validation of a molecular beacon probe–based real-time polymerase chain reaction assay for rapid detection of methicillin resistance in Staphylococcus aureus. *Archives of pathology & laboratory medicine*, *127*(7), 845-849.
18. Arya, M., Shergill, I. S., Williamson, M., Gommersall, L., Arya, N., & Patel, H. R. (2005). Basic principles of real-time quantitative PCR. *Expert review of molecular diagnostics*, *5*(2), 209-219.
19. Livak, Kenneth J., and Thomas D. Schmittgen. "Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method." *methods* 25, no. 4 (2001): 402-408.
20. Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467-470.
21. Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods* 5, no. 7 (2008): 621-628.
22. Tarazona, Sonia, Fernando García, Alberto Ferrer, Joaquín Dopazo, and Ana Conesa. "NOIseq: a RNA-seq differential expression method robust for sequencing depth biases." *EMBnet. journal* 17, no. B (2011): 18-19.
23. Li, Jun, and Robert Tibshirani. "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data." *Statistical methods in medical research* 22, no. 5 (2013): 519-536.
24. Sun, Wei, and Yijuan Hu. "eQTL mapping using RNA-seq data." *Statistics in biosciences* 5, no. 1 (2013): 198-219.

25. Heyer, Erin E., Ira W. Deveson, Danson Wooi, Christina I. Selinger, Ruth J. Lyons, Vanessa M. Hayes, Sandra A. O'Toole et al. "Diagnosis of fusion genes using targeted RNA sequencing." *Nature communications* 10, no. 1 (2019): 1-12.
26. Yi, W., Li, J., Zhu, X., Wang, X., Fan, L., Sun, W., & Yan, J. (2020). CRISPR-assisted detection of RNA–protein interactions in living cells. *Nature methods*, *17*(7), 685-688.
27. Ronaghi, Mostafa, Samer Karamohamed, Bertil Pettersson, Mathias Uhlén, and Pål Nyrén. "Real-time DNA sequencing using detection of pyrophosphate release." *Analytical biochemistry* 242, no. 1 (1996): 84-89.
28. Fedurco, Milan, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies." *Nucleic acids research* 34, no. 3 (2006): e22-e22.
29. Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., & Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728-1732.
30. Hedges, D. J., Guettouche, T., Yang, S., Bademci, G., Diaz, A., Andersen, A., ... & Gilbert, J. R. (2011). Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PloS one*, *6*(4), e18595.
31. Edwards, D., Batley, J., & Snowdon, R. J. (2013). Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics*, *126*(1), 1-11.
32. Rothberg, Jonathan M., Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon et al. "An integrated semiconductor device enabling non-optical genome sequencing." *Nature* 475, no. 7356 (2011): 348-352.
33. Levene, Michael J., Jonas Korlach, Stephen W. Turner, Mathieu Foquet, Harold G. Craighead, and Watt W. Webb. "Zero-mode waveguides for single-molecule analysis at high concentrations." *science* 299, no. 5607 (2003): 682-686.
34. Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso et al. "Real-time DNA sequencing from single polymerase molecules." *Science* 323, no. 5910 (2009): 133-138.
35. Wang, Yue, Qiuping Yang, and Zhimin Wang. "The evolution of nanopore sequencing." *Frontiers in genetics* 5 (2015): 449.
36. Ashton, Philip M., Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'grady. "MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island." *Nature biotechnology* 33, no. 3 (2015): 296-300
37. Zhang, Guojie, Guangwu Guo, Xueda Hu, Yong Zhang, Qiye Li, Ruiqiang Li, Ruhong Zhuang et al. "Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome." *Genome research* 20, no. 5 (2010): 646-654.
38. Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even et al. "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification." *Bioinformatics* 21, no. 5 (2005): 650-659.
39. Li, Ruiqiang, Yingrui Li, Karsten Kristiansen, and Jun Wang. "SOAP: short oligonucleotide alignment program." *Bioinformatics* 24, no. 5 (2008): 713-714.
40. Stanke, Mario, Oliver Schöffmann, Burkhard Morgenstern, and Stephan Waack. "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources." *BMC bioinformatics* 7, no. 1 (2006): 62.
41. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498-2504.
42. Holloway, Beth, Stanley Luck, Mary Beatty, J-Antoni Rafalski, and Bailin Li. "Genome-wide expression quantitative trait loci (eQTL) analysis in maize." *BMC genomics* 12, no. 1 (2011): 336.
43. Nishijima, Ryo, Kentaro Yoshida, Yuka Motoi, Kazuhiro Sato, and Shigeo Takumi. "Genome-wide identification of novel genetic markers from RNA sequencing assembly of diverse Aegilops tauschii accessions." *Molecular Genetics and Genomics* 291, no. 4 (2016): 1681-1694.
44. Bertazzon, Nadia, Paolo Bagnaresi, Vally Forte, Elisabetta Mazzucotelli, Luisa Filippin, Davide Guerra, Antonella Zechini, Luigi Cattivelli, and Elisa Angelini. "Grapevine comparative early transcriptomic profiling suggests that Flavescence dorée phytoplasma represses plant responses induced by vector feeding in susceptible varieties." *BMC genomics* 20, no. 1 (2019): 526.
45. Karre, Shailesh, Arun Kumar, Dhananjay Dhokane, and Ajjamada C. Kushalappa. "Metabolo-transcriptome profiling of barley reveals induction of chitin elicitor receptor kinase gene (HvCERK1) conferring resistance against Fusarium graminearum." *Plant molecular biology* 93, no. 3 (2017): 247-267.
46. Koduru, Lokanand, Hyang Yeon Kim, Meiyappan Lakshmanan, Bijayalaxmi Mohanty, Yi Qing Lee, Choong Hwan Lee, and Dong-Yup Lee. "Genome-scale metabolic reconstruction and in silico analysis of the rice leaf blight pathogen, Xanthomonas oryzae." *Molecular Plant Pathology* 21, no. 4 (2020): 527-540.
47. Roe, B. A. (2014). Frederick Sanger (1918–2013). *Genome research*, *24*(4), xi-xii.
48. Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic acids research*, *29*(9), e45-e45.
49. Napoli, C., Lemieux, C., & Jorgensen, R. (1990). Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *The plant cell*, *2*(4), 279-289.
50. Fire, Andrew, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." *nature* 391, no. 6669 (1998): 806-811.
51. Rothberg, J. M., & Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nature biotechnology*, *26*(10), 1117-1124.
52. Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., & Causey, M. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. *Nature biotechnology*, *27*(7), 652-658.
53. Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., & Milos, P. M. (2009). Direct RNA sequencing. *Nature*, *461*(7265), 814-818.

54. Yan, J., Gu, Y., Jia, X., Kang, W., Pan, S., Tang, X., ... & Tang, G. (2012). Effective small RNA destruction by the expression of a short tandem target mimic in Arabidopsis. *The Plant Cell*, *24*(2), 415-427.
55. Wheeler, E. C., Van Nostrand, E. L., & Yeo, G. W. (2018). Advances and challenges in the detection of transcriptome-wide protein–RNA interactions. *Wiley Interdisciplinary Reviews: RNA*, *9*(1), e1436.
56. Yi, W., Li, J., Zhu, X., Wang, X., Fan, L., Sun, W., ... & Yan, J. (2020). CRISPR-assisted detection of RNA–protein interactions in living cells. *Nature methods*, *17*(7), 685-688.

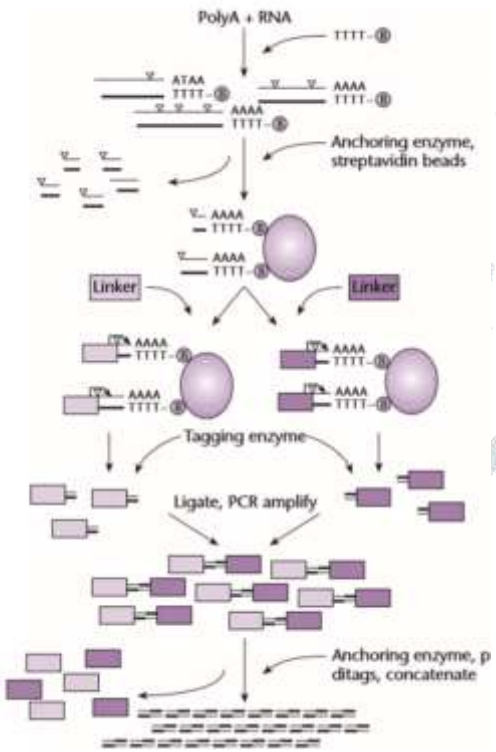
**List of figures:**

**Main Text Figure List:**

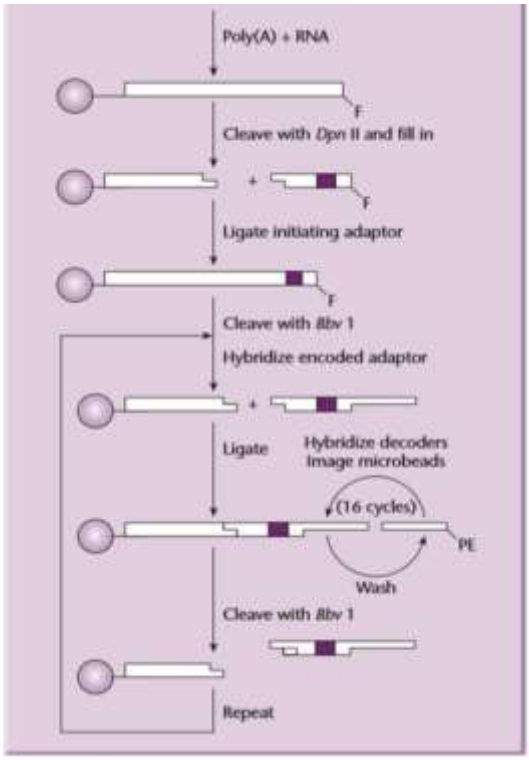
| | |
|---|---|
| Fig.1 | Flow of transcriptome technology |
| Fig.2 | Transcriptome analysis in plants |
| Fig.3 | Genome assembly, annotation and differential gene expression, gene isoform study using RNA-seq |
| Fig.4 | eQTL detection process |

**Supplementary Figure List:**

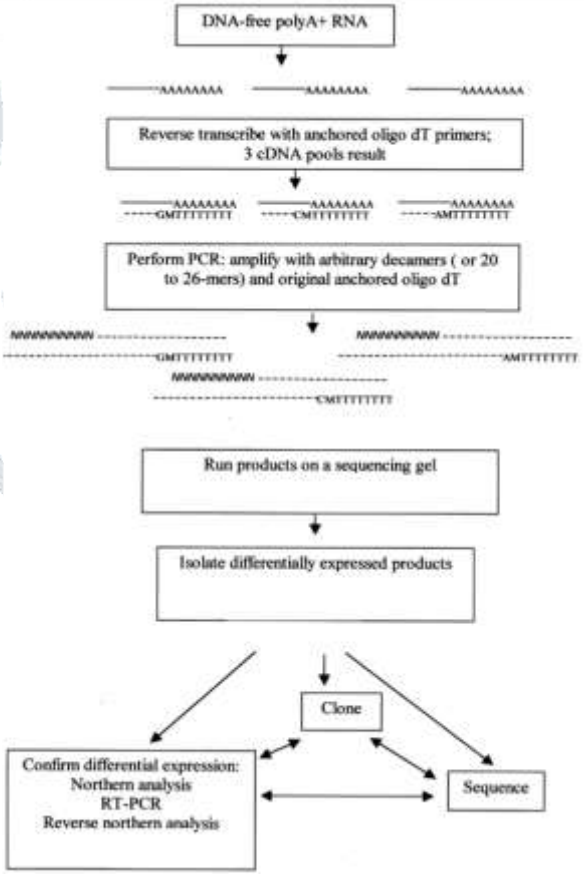
| | |
|---|---|
| Supplementary figure 1 | SAGE (Serial Analysis of Gene Expression) workflow |
| Supplementary figure 2 | MPSS (Massively Parallel Signature Sequence) workflow |
| Supplementary figure 3 | DDRT-PCR Workflow |
| Supplementary figure 4 | Suppression Subtractive Hybridization workflow |
| Supplementary figure 5 | RNA-seq data processing and analysis |
| Supplementary figure 6 | CLIC-seq work flow |



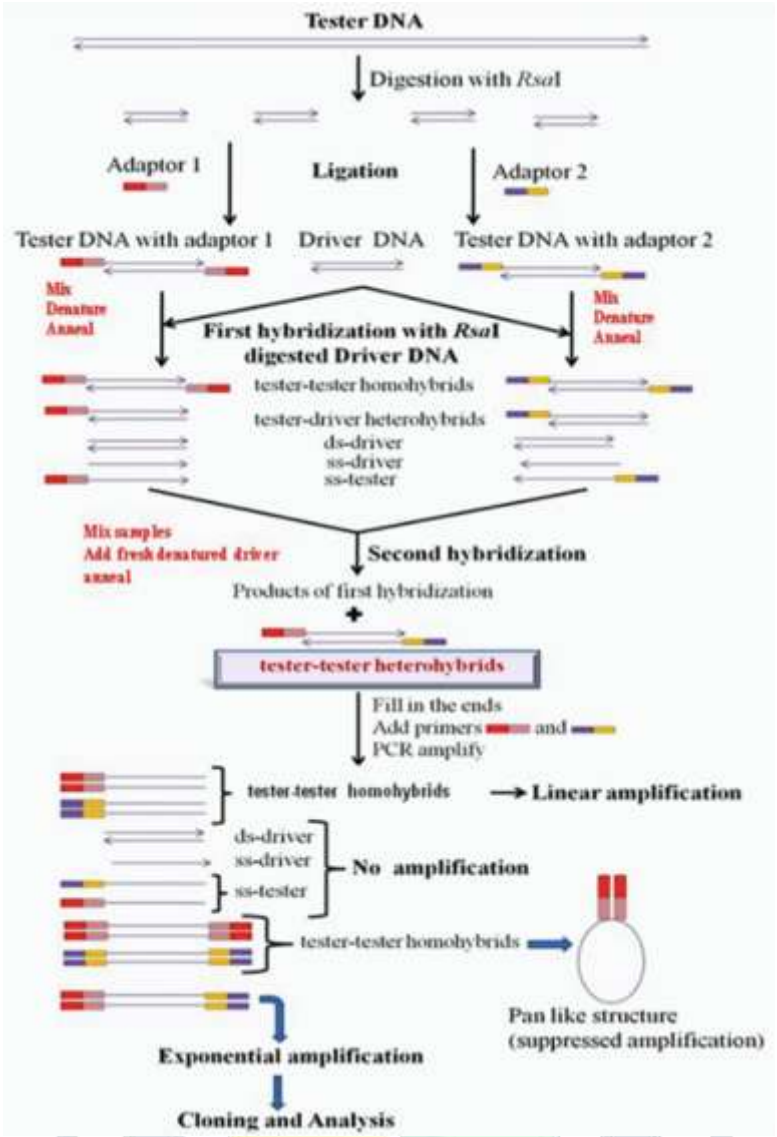

**Supplementary figure 1. SAGE (Serial Analysis of Gene Expression) workflow**



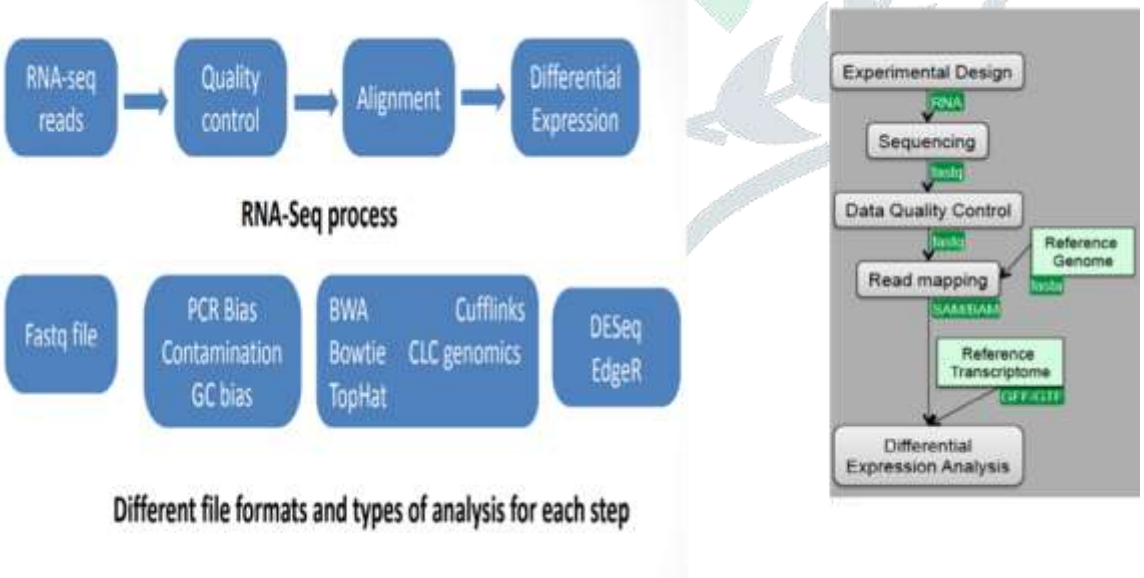

**Supplementary figure 2. MPSS (Massively Parallel Signature Sequence) workflow**



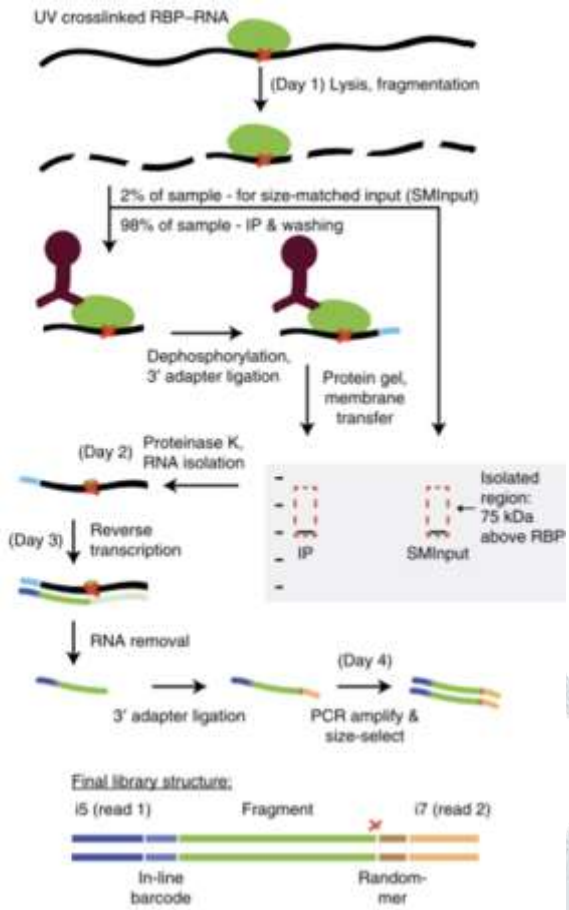

**Supplementary figure 3. DDRT-PCR Workflow**




**Supplementary figure 4. Suppression Subtractive Hybridization workflow**




**Supplementary figure 5. RNA-seq data processing and analysis**




**Supplementary figure 6. CLIC-seq work flow**