Chapter 3

# Transcriptome Analysis for Non-Model Organism: Current Status and Best-Practices

Vahap Eldem, Gokmen Zararsiz, Tunahan Taşçi,
Izzet Parug Duru, Yakup Bakir and Melike Erkan

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.68983

**Abstract**

Since transcriptome analysis provides genome-wide sequence and gene expression information, transcript reconstruction using RNA-Seq sequence reads has become popular during recent years. For non-model organism, as distinct from the reference genome-based mapping, sequence reads are processed via *de novo* transcriptome assembly approaches to produce large numbers of contigs corresponding to coding or non-coding, but expressed, part of genome. In spite of immense potential of RNA-Seq–based methods, particularly in recovering full-length transcripts and spliced isoforms from short-reads, the accurate results can be only obtained by the procedures to be taken in a step-by-step manner. In this chapter, we aim to provide an overview of the state-of-the-art methods including (i) quality check and pre-processing of raw reads, (ii) the pros and cons of *de novo* transcriptome assemblers, (iii) generating non-redundant transcript data, (iv) current quality assessment tools for *de novo* transcriptome assemblies, (v) approaches for transcript abundance and differential expression estimations and finally (vi) further mining of transcriptomic data for particular biological questions. Our intention is to provide an overview and practical guidance for choosing the appropriate approaches to best meet the needs of researchers in this area and also outline the strategies to improve on-going projects.

**Keywords:** whole transcriptome, *de novo* assembly, genome-wide expression, non-model organism

## 1. Introduction

The on-going advances in sequencing technologies and a drastic drop in the cost of sequencing allow us to obtain genome-wide genetic information for virtually all kingdoms of life.

Particularly, making large-scale DNA sequencing more affordable and accessible for small-scale laboratories has greatly promoted genomic research studies on non-model organisms genetically linked to a specific biological question of interest [1, 2]. Despite huge effort, *de novo* sequencing of an entire genome is not an easy task, even now, and this also makes 'RNA sequencing (hereafter, RNA-Seq)-based transcriptomic analysis' appealing for non-model organisms that are generally described as having no or limited genomic resources and transcriptomic datasets as well as molecular tools [3–6]. In the field of '-omics' disciplines, RNA-Seq is among high-throughput experimental methods and widely used for identifying all functional elements in the genome. In other words, RNA-Seq data are directly derived from functional genomic elements, mostly protein-coding genes. Therefore, analysing the expressed part of genome by RNA-Seq gives substantial information about the genome-wide transcriptome structure, profile and dynamics for non-model organism at genome-wide scale. Currently, large-scale sequencing efforts such as 'Fish-T1K (Transcriptomes of 1000 fishes)', '1KITE (1K insect transcriptome evolution)' and '1KP (1000 Plants Project)' have been initiated to serve as valuable source of transcriptome composition and dynamics. In spite of immense potential of RNA-Seq–based methods, particularly in recovering full-length transcripts and spliced isoforms from short-reads, the accurate results can be only obtained by the procedures to be taken in a step-by-step manner.

Compelling evidence show that a number of factors *de novo* transcript construction procedure were reported, such as error-prone and biased (e.g. GC%) nature of sequencing technologies, limitations of assembler algorithm and multi k-mer approaches [7–9], read length [10], coverage depth of reads [11], pre-processing options of raw reads [12, 13] and transcript complexity of organism (e.g. sequence variations at terminal regions, alternative splicing, antisense transcription, overlapping genes) [14]. Therefore, the state-of-the-art advancements in methodologies

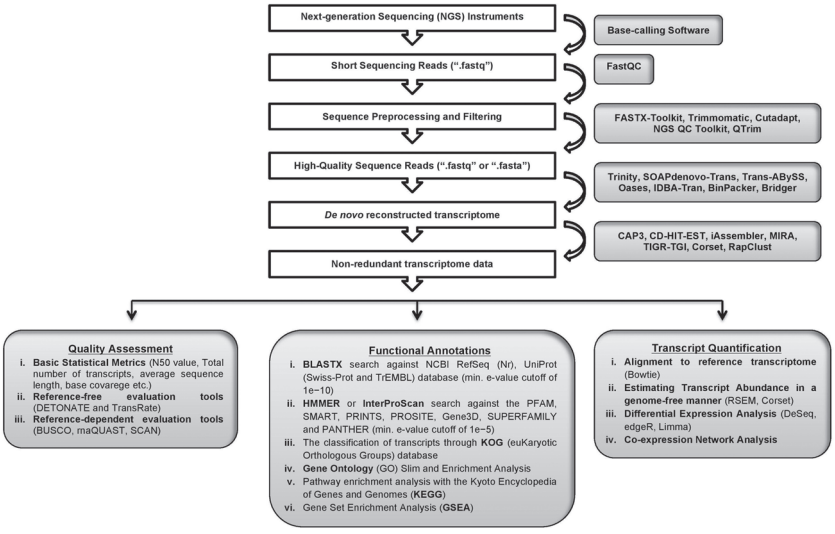


**Figure 1.** An overview of *de novo* transcriptome analysis pipelines from assembly to quality checking and pre-processing to assembly and transcript quantification.

and applications for transcriptome assembly should be meticulously considered while planning a project. As no consensus procedure exists, researchers mainly in the field of ecology and evolution use many different approaches and tools from sequence pre-processing to functional annotations (**Figure 1**). In this context, establishing a guideline that facilitates and standardizes the transcriptome assembly and post-assembly analysis provides a good starting point.

## 2. *De novo* transcriptome assembly methods and mining transcriptome data for non-model organism

### 2.1. Quality check and pre-processing of raw reads

Following sequencing reaction and initial processing, next-generation sequencing instruments generate raw image files that are automatically processed via instrument base calling software to output a massive quantity of raw sequence data in ".fastq" format. The ".fastq" is a text format containing both sequence read and base calling information encoded in ASCII characters. The read quality at each base or quality score can be obtained by converting the ASCII characters into Phred score ($Q$) indicating the probability of an erroneous base call. Compelling evidences show that a minimum threshold of Phred score for assembly and alignment is 20 (equivalent to 99% probability of being correct) for each base in raw read. Despite remarkable progress in sequencing chemistry and base detection approaches, the instruments can still produce incomplete, erroneous and ambiguous reads. Therefore, a pre-processing step (quality checking and read filtering) is considered an essential prerequisite prior to *de novo* transcriptome assembly because erroneous and ambiguous bases can often lead to fragmented and misassembled transcripts.

Quality checking and visualization of raw reads (in fastq) start with the FastQC tool (a standalone Java program available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC generates a HTML output containing a number of graphical illustrations providing the number and length of raw reads and duplication rate, but two main component of the FastQC tool: (i) *per base sequence content* and (ii) *per base sequence quality* are particularly useful in guiding pre-processing step. The most popular pre-processing tools are FASTX-Toolkit [15], Trimmomatic [16], Cutadapt [17], NGS QC Toolkit [18] and Qtrim [19], and regardless of the tools used, common pre-processing steps include: (i) removing adapter sequences, (ii) discarding the low quality reads ($Q \leq 20$) and ambiguous nucleotides (Ns), (iii) removing the short-read length sequences (length below 50 base pair (bp)) and (iv) trimming low quality bases at the both ends of reads (generally first 10 bp) (**Figure 1**) [20]. After pre-processing, resulting high-quality reads are ready for downstream analysis; *de novo* transcriptome assembly.

### 2.2. A brief glance at *de novo* transcript assemblers

Currently, the length of sequence reads from NGS instruments (e.g. sequencing by synthesis from Illumina HiSeq Models) is ranged from 150 to 250 base pairs (bp) and, following quality checking and filtering step, the high-quality sequence reads have to be *de novo* assembled for

transcript reconstruction. The sequence read length is shown to be one of the key parameters in determining *de novo* assembly strategy. While the overlap-layout consensus (OLC) approach has been used for the assembly of long reads generated from the third-generation sequencing instruments such as PacBio Sequel or Oxford Nanopore, *de Bruijn* graph approach has been used in both *de novo* genome and transcriptome assembly because this computationally effective algorithm can process billions of short reads to reconstruct the transcriptome as complete as possible. In the *de Bruijn* methods, the graphs are constructed from short reads and then paths in this graph are used to generate contigs. In graph construction, a given read is broken into k-mer seeds (nodes) and edges are added between consecutive k-mers (in manner; the suffix of length k−1 of one node is the prefix of length k−1 of the other) and then, these k-mers are arranged into a *de Bruijn* graph structure (**Figure 2**). Contigs are obtained by inversely transforming the optimal path in the *de Bruijn* graph into sequences [21]. However, *de Bruijn* graph-based strategy between *de novo* genome and transcriptome assembly is slightly modified because of the following reasons: (i) while the DNA sequencing depth is expected to be uniform across the genome (except in repetitive regions), the sequencing depth of transcripts can vary considerably, (ii) Genome assembly graph is considered as linear (theoretically one graph for each chromosome), but due to alternative splicing, transcriptome assembly is more complex than genome and requires a graph to represent the multiple alternative transcripts per locus [1, 21]. By considering these challenges, several *de novo* assembly tools such as Trinity [1], SOAPdenovo-Trans [22], Trans-AbySS [23], Oases [24], IDBA-Tran [25], BinPacker [26] and Bridger [27] have been developed so far (Box 1). Most of these tools, which are initially developed for *de novo* genome assembly (except for Trinity) use *de Bruijn* graph-based assembly strategy and have their own pros and cons in transcript reconstruction.

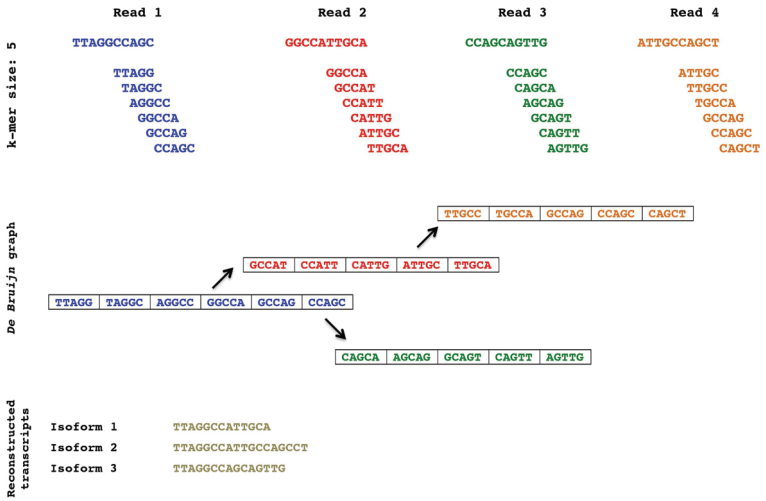


**Figure 2.** The *de Bruijn* graph approach is instrumental for reference-free transcriptome assembly and *de Bruijn* graphs are built from the short reads. These short reads are split into short k-mers (here, k-mer length, 5) and then k-mers are connected by overlapping prefix and suffix (k–1)-mers. When the *de Bruijn* graph is built from reads, the optimal paths are obtained in the graphs and reconstructed transcripts (or contigs) are recovered by inversely transforming the optimal path in the *de Bruijn* graph.

The quality of assemblies in terms of transcript number and length generated by such assemblers is highly influenced by k-mer length or hash length. Schulz et al. [24] reported that although assemblies generated using short k-mer have the risk of introducing misassemblies, rare transcripts can only be retrieved by selecting short k-mers while longer k-values perform best on high expression genes. In order to identify the full spectrum of transcript abundance and isoforms, *de novo* assemblers utilize an iterative multi-kmer approach from 21 to 71, except for Trinity whose k-mer length is fixed to 25. Due to its apparent importance, an informed k-mer selection tool, KREATION, has been recently developed using fit-based algorithm, limiting the number of k-mer values without significant loss in assembly quality but with saving in assembly time [28]. KREATION first clusters the assemblies generated from single k-mer to determine "*extended clusters*" showing the assembly quality and then, a heuristic model is applied to predict the optimal stopping threshold for a multi k-mer assembly study.

Box 1. A general overview of *de novo* transcriptome assembly tools from short-reads.

## Trinity

Trinity's main difference from other transcriptome assembly programs is that it is directly manufactured for *de novo* RNA assembly. It uses the parallel calculation method to create alternate spliced isoforms and transcripts with *de Bruijn* method [1]. Trinity has three functional modules; *Inchworm*, *Chrysalis* and *Butterfly* of which work in succession and perform different tasks [29]. *Inchworm* uses greedy extension model based on k-mer overlap and reports full-length transcripts for a dominant isoform. Then, *Chrysalis* clusters overlapping contigs and constructs *de Bruijn* graphs. Finally, *Butterfly* process these graphs in parallel and reconstructs full-length transcripts for each isoform. In addition to reconstruct accurate transcripts from RNA-Seq data, Trinity exhibit superior performance in recovering isoforms. Trinity requires extensive computational resources and running time, but it performs best in terms of assembly quality such as N50 value, fewer chimeras and transcript coverage.

## SOAPdenovo-Trans

SOAPdenovo-Trans is *de Bruijn* graph-based assembler, which derived from its genome assembler version SOAPdenovo2 [22]. In SOAPdenovo-Trans algorithm, two module error-removal and heuristic graph traversal methods are borrowed from Trinity and Oases, respectively. The algorithm has two main steps: (i) contig assembly and (ii) transcript assembly. Contigs are generated using SOAPdenovo after globally and locally error removal. SOAPdenovo-Trans uses both single-end reads and paired-end reads which mapped back onto the contigs to build scaffolds and then it applies a strict transitive reduction method to simplify the scaffolding graphs, and provide more accurate results. SOAPdenovo-Trans uses less memory and shortest running time than other assembler programs. Although SOAPdenovo-Trans performed best in base coverage, the minimum, first quartile, median, mean and third quartile length of transcripts obtained from SOAPdenovo-Trans is shorter than that in BinPacker, Bridger, IDBA-Tran and Trinity.

## Trans-AbySS

Trans-abyss is a method and pipeline for the collection and analysis of short transcriptomic data. Abyss assembly process consists of single-ended and double-ended stages. The single-ended stage is also based on the *de Bruijn* graph structure; when parameter k is given, it is transformed into tiled k-mer represented as read nodes and (k-1) bases are superimposed as directed edges. Allelic differences, minor changes in the sequence and repetitive random base invocation errors lead to 'bubbles' throughout the graph. Once these errors have been removed in the k-mer space, the single-ended contigs defined by the 'walk' clear across the graph. In the matched tier phase, the pairs aligned in the single-ended contigs define the empirical distribution of the distances of the pairs. Single-ended readings of different contigs to the co-aligned pairs and empirical distribution then intercontig distance and combined to form contigs are paired end contigs that can be combined [23]. Trans-AbySS reaches the end by creating direct sequenced readings with Bruijn graphics, removing possible errors from the middle and solving each connected Bruijn graph for each connected component. Compared to other assembler programs the lowest percentage of chimera is seen in Trans-AbySS [30]. Comparative studies showed that with Trinity, Trans-ABySS performed best in gene coverage and number of recovered full-length transcripts [31].

## Oases

Oases is a RNA transcriptome assembler that contains many developmental constructs. Combines multiple k-mers and topological analysis methods. In addition, it uses the dynamic error correction feature developed for RNA-Seq data. Assembly process of Oases takes place by creating independent assemblies, which vary according to the length of the k-mers, and then assembling them all together in one assembly. In each assembly, readings are used to generate *de Bruijn*, and then faults are simplified, organized into a scaffold, divided into loci and eventually analysed. Then dynamic correction is performed and Oases creates contigs sets of clusters called loci. Since it is more likely to be unique, long contigs treated first when the scaffold is constructed and faults that may arise from alternative splices are eliminated. Oases provide a robust pipeline from RNA-Seq readings to generate full-length assemblies of transcripts. Especially designed for dealing with RNA-Seq condition, unequal coverage and alternative spliced situations [24]. Oases-Velvet produced the highest number of chimeric transcripts at different k-mer sizes and it has the highest RAM (i.e. random access memory) usage among all assemblers.

## IDBA-Tran

IDBA-Tran uses a different approach. Firstly, it produces small *de Bruijn* graphs and enlarges the graph with larger k values. Subsequently, transcripts are found on a large Bruijn graph,

where the same genetic transcripts usually form a single component [25]. IDBA-Tran modulates the products of the k-mers of the same composition with a very normal distribution, which depends on the expression levels of the corresponding isoforms. IDBA-Tran obtains a large number of small components, each representing a single gene. For each small component, IDBA-Tran retrieves the isoform sequences with matched-ended reads by looking for compound pathways. Based on more than one normal distribution and contig length, IDBA-Tran calculates a local threshold to determine whether a k-mer or contigs in error. Using the probabilities and depths that connect the two components together, taking into account the length of the path, the graphics that make up the IDBA-Tran components detect and remove faulty paths. For this reason, IDBA-Tran produces more contigs for low-expressed transcripts and performs better than Oases and Trinity [25].

## BinPacker

BinPacker reshapes the problems and generates full-length transcripts by following the aggregated graph line generated by various techniques used in Bridger. Some advantages of BinPacker: (i) BinPacker allows the use of user-defined k-mer values for best performance and (ii) BinPacker uses a strict mathematical model. This allows the BinPacker to achieve a lower false positive rate at the same sensitivity level. (iii) BinPacker makes full use of the step depth applied to graphics, so that the assembly results are more accurate. BinPacker combines transcripts on every merging graph it creates [26]. BinPacker is more unsuccessful than other programs on chimeric data [31].

## Bridger

Using a multi-k strategy to achieve high sensitivity leads to more false positives. However, identifying the optimal set of paths that represent the potential isoform can significantly reduce false positive estimates. Bridger's basic idea is to build a bridge between two popular assemblers, Cufflinks (reference-based assembler) and Trinity (*de novo* assembler). Bridger uses a rigorous mathematical model called the minimum path envelope to search for the lowest path set (transcript) supported by RNA-Seq readings. Bridger runs very fast and requires less memory space and CPU (i.e. Central Processing Unit) time than other methods and generates splicing graphics for all genes [27].

### 2.3. Generating non-redundant transcript data

As described in the previous section in detail, a reference transcriptome for non-model organism can be built using various types of *de novo* transcriptome assemblers. All these assemblers are successful to some extent in recovering expressed transcripts; however, constructing full-length transcripts from short reads remains a daunting and complicated task. Therefore, to obtain more accurate data, researchers performed several studies to optimize a number of

key parameters affecting assembly results such as optimal sequencing depth [11], the read length [10], multi k-mer approaches [7–9], the quality score and error correction of sequence reads [12, 13]. However, transcriptome software themselves follow a multi-stage procedure to avoid introducing misassemble, chimeric assembly and transcript artefacts and to obtain all spliced isoforms from the same gene. For instance, the Inchworm module of Trinity assembles short-reads using greedy extension based on k-mer overlap and reports full-length transcripts for a dominant isoform. Then, the final module, Butterfly, processes the individual graphs in parallel and reconstructs full-length transcripts for each isoform after Chrysalis clusters overlapping contigs, and constructs de Bruijn graphs. Despite all these efforts, *de novo* assembly of short-reads, regardless of software used, results in hundreds of thousands of contigs, a set of contiguous transcript sequences. Without any further analysis such as clustering or post-assembling, the final set of contigs includes (i) partial transcripts and rudimentary isoforms (splice variants), (ii) redundant transcripts (different lengths of the same transcripts, mostly fragments) and (iii) chimeric (fusion) and misassembled sequences [3].

Creating non-redundant transcript dataset with various bioinformatics approaches is a first step after *de novo* transcript assembly. Because, eliminating redundant transcripts and retaining one representative of each transcript isoform (generally, correct and longest in each transcript cluster) are particularly important for downstream applications such as the analysis of transcript structure, gene expression, phylogenomics and identification of SNP variants [8, 30, 32]. To date, several clustering algorithm and post-assembly implementations were developed and used in a significant number of articles for the purpose of creating a non-redundant consensus dataset. The most popular tools used to reduce redundancy in the assembled dataset are CAP3 [33], CD-HIT-EST [34], iAssembler [35], MIRA [36] and TIGR-TGI Clustering tool [37] as well as Corset [32], if performing a differential gene expression analysis. In addition to these tools, some assemblers such as Oases and Trans-ABySS have their own "merging tools" to generate a consensus transcript set when applied multiple k-mer approaches.

So far, all studies using *de novo* transcriptome assembly procedure have included either post-assembly or clustering analysis. Among the assembly-based approaches, CAP3 [33] is one of the first large-scale EST-based assembly tool, which filters for redundant information by detecting overlaps between the contigs and generate the consensus sequence for each transcript. As an overlap-layout-consensus (OLC)-based assembly pipeline, TIGR gene indices clustering tool (TGICL) [37] was developed for producing larger and more complete consensus sequences. In this pipeline, a final set of contigs is first clustered based on pairwise sequence similarity and then each cluster is assembled so that consensus sequences (or non-redundant unigenes) are generated. Yet these methods are successful in removing redundancy, the methods have failed to satisfy the needs of generating a contig per transcripts. It was suggested that there are two type problems, which might be responsible for such failure. The problems frequently observed during assembly are (i) the misassembly of spliced transcripts or paralogs and (ii) contigs derived from the same transcript fail to be assembled together. The iAssembler [35] specially developed to overcome these problems encountered and it consists of seven modules grouped into three functional phases: general controller (input, output and assembly parameters), assembler and error corrector phases. The iAssembler utilizes the approaches of

CAP3 and MIRA assemblers for initial assembly of transcripts, and subsequently, the pairwise alignment information of overlapped transcripts is obtained using Megablast to assemble them into one contig if those transcripts fail to be assembled by either MIRA or CAP3. The assembly process finishes after correcting the above-mentioned errors via error corrector phases, which is the main contribution of iAssembler. A comparison showed that iAssembler has a superior performance over CAP3, MIRA and TGICL in terms of generating much less assembly errors in assembling [35].

Another widely used approach to reduce redundancy in contig assembly is clustering sequences. In this regard, by far the most popular tool is CD-HIT-EST [34]. The CD-HIT-EST is generally used to remove the shorter redundant transcripts and duplicate contigs in large-scale transcriptome datasets. Compared to assembly-based approaches, the CD-HIT-EST is dramatically faster in practice due to its novel parallelization strategy. Corset [32] as a state- of-the-art approach was proposed for hierarchically clustering contigs using information about shared reads. The performance evaluation showed that Corset outperformed CD-HIT-EST in recall (i.e. true positives/(true positives + false negatives)) for genes with no fragmentation and the authors suggested that CD-HIT-EST is not the most effective contig clustering tool while Corset gives a convenient method to cluster contigs [32]. More recently, a clustering tool, RapClust [38] has been developed for *de novo* transcriptome clustering based on the relationships exposed by multi-mapping sequencing fragments and it generates clusters of comparable or better quality than current clustering approaches and does so substantially faster. Although accumulating evidences have indicated that the sequence identity threshold should be set above 90% in both assembly and clustering approaches, a detailed comparison analysis is required for those approaches in terms of accuracy and capability for removing redundant sequences.

### 2.4. Quality assessment tools for *de novo* transcriptome assemblies

Quality assessment of *de novo* assembled transcripts using reference-free or evidence-based tools seems to be a prerequisite for meaningful interpretation of downstream analysis such as discovery of novel transcripts and correct identification of differentially expressed genes. From a practical point of view, the quality assessment of assembled transcriptome sequences can be handled in three different ways: (i) basic statistical metrics, (ii) reference-free evaluation tools and (iii) reference-dependent or sequence homology-based approaches. Generally, calculating basic statistical metrics is considered as first step in the evaluation of assembled transcriptome. These metrics include total number of transcripts, total base coverage, transcript coverage, N50 value, the presence of chimeric transcripts, longest transcript length, average length of transcripts, etc. These metrics are simple and useful to obtain information about the transcript numbers and coverage at a first glance, but provides no information about accuracy or reliability of transcripts. For instance, N50 value is a median length of a set of contigs (assembled transcripts), but it measures the continuity of contigs but not their accuracy. Recently, reference-free evaluation tools were developed for the accuracy and completeness of *de novo* transcriptome assemblies (see Box 2, i.e. RSEM-EVAL and TransRate). These approaches only process high-quality sequence reads and assembled transcriptome

based on their strong background models and producing scores indicating assembly quality. As for sequence homology-based quality metric, it is seen as standard evaluation criteria for transcriptome assemblies. In this approach, each contig in the assembled transcriptome set was aligned against a reference database (rnaQUAST) or publicly available databases using BLAST, BLAT or SCAN methods (Box 2). Besides, now it is well known that the genome of all living organisms from bacteria to mammals contains evolutionary conserved and phylogenetic clades characteristic of single-copy orthologous gene sets. Therefore, it is considered as an indicator of quality and completeness of transcriptome assembly (see BUSCO in Box 2).

Box 2. A general overview and framework of *de novo* transcriptome assembly evaluation tools.

---

## DETONATE

Li et al. [39] proposed a software package called DETONATE (DE novo TranscriptOme rNa-seq Assembly with or without the Truth Evaluation) which is a methodology for assessing and ranking of *de novo* transcriptome assemblies obtained from various assemblers. DETONATE software is consisted of two parts: RSEM-EVAL and REF-EVAL. As a reference-free evaluation method, RSEM-EVAL is considering as main contribution of the software and uses a probabilistic model that requires only an assembly and the RNA-Seq reads to compute the joint probability. RSEM-EVAL provides a score obtained from calculation of three components; maximum likelihood (ML) estimate, an assembly prior and a Bayesian information criterion (BIC) penalty, reflecting whether resulting contigs are supported by RNA-Seq reads or not. Then, RSEM-EVAL ranks these scores in descending order (from highest to lowest) and highest-scoring assembly is considered as ground truth, in other words, most reliable and compact assembly.

## rnaQUAST

Bushmanova et al. [40] developed a quality evaluation tool for transcriptome assemblies. The tool, rnaQUAST, basically maps assembled transcripts to reference genome using BLAT [41] or GMAP [42] and comparing resulting alignments to gene database for measuring quality metrics. In addition to the basic descriptors for contig continuity such as total length, average length of assembled transcripts, longest transcripts and N50 value, the principal contribution of rnaQUAST is arised from the alignments of transcripts to isoforms' positions and analyses them to estimate how well the isoforms are covered by the assembly. For *de novo* quality assessment, rnaQUAST takes advantage of other tools like BUSCO.

## BUSCO

In an evolutionary context, Simao et al. [43] presented a software package, BUSCO (Benchmarking Universal Single-Copy Orthologs) for assessment of transcriptome assembly and completeness.

For that purpose, BUSCO scans transcriptome assembly for the presence of near-universal single-copy orthologous gene-sets generated from OrthoDB database of orthologs (http://www.orthodb.org). Covering a high proportion of single-copy orthologous gene-sets indicates completeness of assembled transcripts. BUSCO sets are generated for six major phylogenetic clades; 3023 genes for vertebrates, 675 for arthropods, 843 for metazoans, 1438 for fungi and 429 for eukaryotes. Accumulating evidence showed that above 90% covering of single-copy orthologous gene-sets indicates a good completeness of transcriptome assembly.

## TransRate

Despite relative success in generating *de novo* transcriptome assemblies from short-reads, due to wide range of multiple and flexible parameters of *de novo* assembly methods, this methods can generate different assemblies, even if same data were used. These assemblies include chimeras, structural errors, incomplete assembly (e.g. hybrid assembly of gene families, spurious insertions in contigs) and base errors. To overcome frequently occurring problems and filtering, optimization as well as comparison of assemblies, Smith-Unna et al. [44] developed a reference-free transcriptome assembly evaluation tool for the accuracy and completeness of *de novo* transcriptome assemblies using only input reads and assembled contigs. TransRate first aligns the input reads to final assembly, processes those alignments, and calculates contig scores using the full set of processed read alignments. Following these processes, TransRate classifies contigs into two classes; well assembled and poorly assembled, by learning a score cut-off from the data that maximizes the overall assembly score. TransRate gives two types of reference-free statistics; TransRate contig score and assembly score which are calculated by considering these errors. Therefore, TransRate is seen as a diagnostic quality score tool while RSEM-EVAL, another reference-free transcriptome assembly evaluation tool.

## SCAN

Comparing assembled transcripts against a reference nucleotide or proteome is a routine task for annotating transcripts. By utilizing this information, Misner et al. [45] described an analytical R package called SCAN (sequence comparative analysis using networks) which generates gene-similarity networks illustrating sequence similarities between transcript assemblies and reference data. The SCAN differs from other software such as BLAST [46] or BLAT [41] in that it provides a robust statistical support in a biological context.

### 2.5. Current approaches for transcript quantification from RNA-Seq

Following to the assembly procedures, next step is to map the reads to a reference genome or transcriptome, quantify the transcript abundances and detect the differentially expressed transcripts among interested biological conditions. In this section, we give a brief overview of algorithms used is each analysis procedure (**Figure 3**).

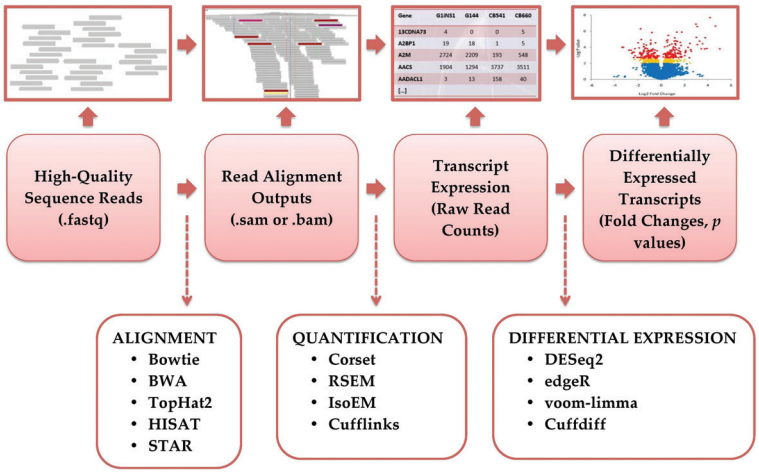


**Figure 3.** Schematic representation of transcript quantification from alignment to differential expression analysis.

Alignment is an important step in RNA-Seq analysis, which refers to the mapping of the reads to a reference genome or transcriptome, if it is available. Aligners can be classified as spliced and unspliced aligners. Unspliced aligners, e.g. Burrows-Wheeler alignment tool (BWA) and Bowtie, align the reads to the transcriptome by using Burrows-Wheeler or seed methods. These aligners do not properly control intron-sized gaps since they are not designed for spliced alignments. For accurate and fast alignment of the sequence reads over exon/intron boundaries, spliced aligners are proposed which use either exon-first or seed-extended methods [47]. While mapping reads using splice-aware aligners such as HISAT [48], TopHat2 [49] and STAR [50] are generally preferred for genome alignment, the software that is particularly developed for differential gene expression analysis for *de novo* assembled transcriptome uses Bowtie alignment program with '—best' option [32, 51]. Alignment process can be complicated due to several factors: sequencing errors, polymorphisms, imperfect annotation, intron-sized gaps, intron signal, alternative splicing and pathological splicing. Moreover, alignment results directly affect the results of downstream analysis, e.g. transcript quantification, differential expression, gene ontology and pathway analysis [52].

After mapping, the next step is the quantification of each transcript for each sample. It has been reported that the number of reads aligned to the reference genome is linearly related to the abundance of transcripts. Large number of transcript quantification algorithms is available in the literature. rSeq models the sequence reads assumed to follow Poisson distribution with parameters related to the transcript abundances [53]. RSEM is a widely used approach that uses expectation-maximization (EM) algorithm to compute the maximum-likelihood estimates of $\theta$ parameters, where $\theta_i$ is the probability of a fragment derived from $i$th transcript. Gibbs sampling is used as well to estimate the posterior means and confidence intervals of transcript abundances. RSEM does not require reference genome or transcriptome files from the users. RSEM conducts a quality score data within the scope of its statistical model or uses a position-dependent error model based on the FASTQ or FASTA input data

types, respectively [51, 54]. Scripture uses the gapped alignments of the reads across splice junctions and the annotated transcripts and produces transcript expressions as RPKM (read per kilobase per million mapped reads) values [55]. Cufflinks assume the sequence reads are sampled independently with uniform probability along transcripts and proportional to the abundances among transcripts. A Bayesian method is used in parameter estimation [56]. IsoEM method exploits information from the distribution of insert sizes and estimates the isoform abundances using an EM algorithm [57]. MMSeq estimates haplotype, isoform and gene-specific expression using a Poisson-based model and EM algorithm. The priors of transcript abundances are assumed to follow a Gamma distribution [58]. BitSeq models the posterior probabilities sequence reads with Markov chains and estimates the transcript expressions using a Bayesian approach [59]. eXpress has a similar methodology to cufflinks. However, it can determine the transcript abundances real-time, and can model indels and errors [60]. CEM identifies the RNA-Seq biases, i.e. positional, sequencing and mapping biases, with quasi-multinomial distribution model and estimates the isoform abundances with component elimination EM approach [61]. Sailfish is an alignment-free approach that is based on indexing and counting k-mers of sequence reads. EM method is used in maximum-likelihood estimation of the transcript abundances. Sailfish is reported as the fastest quantification method as compared to other methods [62]. TIGAR2 models the insertion, deletion and substitution errors in a probabilistic framework, given the gapped alignment of reads to the reference genome. TIGAR2 uses a generative model, including alignment state, nucleotides, the read length distribution and read qualities at first and second positions, to estimate the transcript isoform expressions [63].

Kanitz et al. [64] benchmarked these methods on both simulated and an experimental datasets. The performances are found to be very similar for all algorithms. Teng et al. [65] described several evaluation metrics and compared 7 quantification algorithms and reported that Flux Capacitor and eXpress underperformed, while RSEM outperformed other methods. We believe that RSEMs accuracy may result from its ability to properly handling short transcripts, poly (A) tails and the reads that map to multiple genes. Moreover, this method does not require a reference genome, which is stated to be challenging mostly for eukaryotic species, whose RNA transcripts are spliced and polyadenylated [51]. Beyond these methods, Corset has shown to be another powerful method, which clusters the transcripts into genes and calculates the counts for each gene in a single step [32].

After mapping, per transcript read counts can be used as a relative measure of transcript abundance. In a perfect world, transcript abundance of steady-state mRNA should be directly proportional to the number of reads: a transcript from gene A with twice the cellular concentration of transcript B should have twice as many reads. This relationship should hold across a large range of expression levels spanning several orders of magnitude. Generated transcript abundances can be input to various analysis pipelines. In most cases, the objective is to identify the differentially expressed transcripts between given biological conditions. A key data assumption here is that the data should not contain any technical biases, which may arise from sequence composition, transcript length, sequence depth, sampling bias in library preparation, presence of majority fragments, etc. To enable comparison of genes across samples, these technical biases should be identified and corrected before starting

differential expression analysis. Total count (TC), upper quartile (UQ) and median methods are quantile-based methods, which divide transcript read counts by total number of reads, 3rd quartile and median, respectively. The disadvantage of these methods is that the greater counts can dominate the lower counts in downstream analysis, e.g. differential expression analysis. Reads per kilobase per million mapped reads (RPKM) adjusts read counts both for sequence depth and gene length. RPKM produces unbiased estimates of number of reads; however, this affects the variance. Trimmed mean of M values (TMM) and DESeq2 median ratio approaches are considered as effective library size approaches. These methods assume that a majority of transcripts is not differentially expressed and thus minimize the effect of majority sequences. TMM trims the data based on the log-fold-changes and absolute intensities, then computes the weighted average of genewise log-fold-changes using delta method [66]. DESeq2 median ratio approach generates a pseudo reference sample, which is the geometric mean across samples. Size factors are obtained from the counts and the pseudo reference sample across all genes [67]. An important problem in differential expression analysis is to statistically model the obtained RNA-Seq counts. The preceding studies applied microarray-based methods to log-transformed counts [68, 69]. Some of the studies preferred modelling these data using Poisson distribution [61, 70]. Poisson distribution has a single parameter that represents both mean and variance. Nagalakshmi et al. [71] stated that the presence of biological replicates leads the variance exceeds the mean. This problem is referred to as overdispersion, which led to the development of novel approaches using negative binomial (NB) distribution. DESeq2 and edgeR are the two popular and NB-based approaches to model RNA-Seq data. Both approaches are based on the estimation of mean and variance relationship based on NB distribution. DESeq2 conducts local regression, while edgeR uses a single proportionality constant in this estimation [72, 73]. More recently, Law et al. [74] proposed the voom method, which estimates the mean and variance relationship from log-counts at observational level. Voom provides both gene expression estimates and the corresponding precision weights for downstream analysis. Integration of this method with limma (linear models for microarray and RNA-Seq data) method provided the best control of type-I error, best power and lowest false discovery rate. Wang and Gribskov [31] points out that there may be differences on the differential expression results, between reference genome-based and *de novo* transcriptome assembly approaches. Incomplete and incorrect reference annotation, exon level expression differences and fragmentation of low coverage transcripts are pointed as the reasons of these differences. The authors suggest to perform both approaches even the reference genome is present.

### 2.6. Transcriptomics tells more: focusing on specific annotation tools and guidelines

The general analysis framework of *de novo* assembled transcripts has three phases: (i) generating non-redundant transcripts and quality assessment, (ii) basic sequence annotations including homology-based sequence annotations (BlastX), gene ontology (GO Slim and Enrichment), pathway analysis (KEGG Enrichment) and (iii) transcript quantifications (**Figure 1**). Although annotation process (beyond the scope of this chapter) provides significant information regarding cellular component, molecular functions and biological process in which transcripts involved, more information can be obtained if transcriptomic data can be further analysed and

interpreted in line with the study objectives and research questions. For instance, in evolutionary perspective, transcriptome data can be used for detecting positively selected or fast evolving genes (PSG, FEG) and are increasingly used in genome-wide phylogenetic studies [75–77] following the steps: orthologs gene detection (particularly single copy genes), multiple sequence alignment of coding regions with PRANK and GUIDANCE pipeline (PRANK algorithm is based on an exhaustive search of the best pairwise solution; the guidance assigning a confidence score for each residue, column and sequence in a multi-alignment from Prank [78], so Guidance [79] can be used for weighting, filtering or masking unreliably aligned positions in sequence alignments before positive selection using the branch-site dN/dS test). Following a multiple sequence alignment, the phylogeny is inferred by Phyml [80] based on proteins residues translated from multi-alignments of single copy orthologous. Then, multiple sequence alignment is used to detect positive selection using the branch-site model with the CodeML program of the PAML [81].

In the context of genome-wide sequence polymorphism within species, mining *de novo* constructed transcripts by appropriate variant calling tools may help us to elucidate the nucleotide-level organismal differences. Among the genetic markers, single nucleotide polymorphisms (SNPs) are the most frequent DNA variation across genome and these genetic markers are widely used for characterising genetic diversity and population structure at genome level, construction of linkage and QTL mapping and association mapping due to their high density/frequency and low mutation rate over generations. In non-model organism, lack of genome sequence information, the standard approach for identification of SNPs or insertion-deletion (InDels) starts by mapping high-quality reads against a reference transcript set constructed *de novo* and detect variations. Briefly, the high-quality reads were aligned against reference transcript set using unspliced aligners such as Burrows-Wheeler alignment tool (BWA) [82] or Bowtie2 [83] and then mapped file '.bam' is obtained for variant calling. After sorting aligned reads and removing duplicates and merging '.bam' alignment results, GATK2 (genome analysis tool kit) [84] is used to perform SNP calling. GATK2 software first filters, realigns and recalibrates reads using its standard filter and data pre-processing methods. The resulting analysis ready reads are parsed to detect SNPs using GATK-UnifiedGenotyper tool with parameters of "-stand_call_conf 30" and "-stand_emit_conf 10". Following this step, SNP calls are hard-filtered using GATK-VariantFiltration tool with parameters of "quality by depth > 5", "unfiltered read depth ≥ 10" and "read mapping quality ≥ 40" to obtain reliable and accurate SNPs [85–87].

The eukaryotic genome harbours a large number of non-coding RNAs, which include small and long non-coding RNAs (lncRNAs). LncRNAs are RNA molecules that are longer than 200 nucleotides in length and do not contain protein-encoding sequences. Recent studies have shown that although human genome contains about 19,000 protein-encoding genes (approximately 2% of the genome) [88], 58,684 high-quality lncRNAs have been identified in the genome using a large-scale transcriptome analysis [89]. Accumulating evidence showed that the protein-coding genes are accounted for only 50% of final assembled transcriptome data. Mining final non-redundant transcriptome data via long non-coding RNA identification tools such as PLEK [90], lncRScan-SVM [91], FEELnc [92] or measuring protein coding potential of transcripts using various tools such as coding potential calculator (CPC) [93], coding potential

assessment tool (CPAT) [94], coding-non-coding index (CNCI) [95] provides us more information about the transcriptome landscape of non-model organism.

## Acknowledgements

All authors contributed to the editing of the manuscript and the content is solely the responsibility of the authors. This work was partly supported by the Istanbul University Scientific Research Project (Project No. 46473 and 29506) and also partly supported by Marmara University Research Fund (Grant Number: FEN-A-100616-0275).

## Author details

Vahap Eldem[1]*, Gokmen Zararsiz[2], Tunahan Taşçi[3], Izzet Parug Duru[4], Yakup Bakir[5] and Melike Erkan[1]

*Address all correspondence to: vahap.eldem@istanbul.edu.tr

1 Department of Biology, Faculty of Sciences, Istanbul University, Istanbul, Turkey

2 Department of Medical Statistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey

3 Department of Medical Imaging Techniques, Vocational School of Health Services, Istanbul Bilgi University, Istanbul, Turkey

4 Department of Physics, Faculty of Science and Art, Marmara University, Istanbul, Turkey

5 Department of Biology, Faculty of Science and Art, Marmara University, Istanbul, Turkey

## References

[1] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;**29**(7):644-652. DOI: 10.1038/nbt.1883

[2] Ekblom R, Galindo J. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity. 2011;**107**(1):1-15. DOI: 10.1038/hdy.2010.152

[3] Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, et al. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Molecular Ecology Resources. 2012;**12**(5):834-845. DOI: 10.1111/j.1755-0998.2012.03148.x

[4] Todd EV, Black MA, Gemmell NJ. The power and promise of RNA-seq in ecology and evolution. Molecular Ecology. 2016;**25**(6):1224-1241. DOI: 10.1111/mec.13526

[5] da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrigal J, Sibbesen JA, Maretty L, et al. Next-generation biology: Sequencing and data analysis approaches for non-model organisms. Marine Genomics. 2016;**30**:3-13. DOI: 10.1016/j.margen.2016.04.012

[6] Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, Edger PP, et al. Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant fenome. PloS one. 2016;**11**(1):e0146062. DOI: 10.1371/journal.pone.0146062

[7] Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Research. 2010;**20**(10):1432-1440. DOI: 10.1101/gr.103846.109

[8] Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: A comparative study. BMC Bioinformatics. 2011;**12**(Suppl 14):S2. DOI: 10.1186/1471-2105-12-S14-S2

[9] Haznedaroglu BZ, Reeves D, Rismani-Yazdi H, Peccia J. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinformatics. 2012;**13**:170. DOI: 10.1186/1471-2105-13-170

[10] Chang Z, Wang Z, Li G. The impacts of read length and transcriptome complexity for de novo assembly: A simulation study. PloS one. 2014;**9**(4):e94825. DOI: 10.1371/journal.pone.0094825

[11] Francis WR, Christianson LM, Kiko R, Powers ML, Shaner NC, Haddock SH. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics. 2013;**14**:167. DOI: 10.1186/1471-2164-14-167

[12] Macmanes MD, Eisen MB. Improving transcriptome assembly through error correction of high-throughput sequence reads. PeerJ. 2013;**1**:e113. DOI: 10.7717/peerj.113

[13] Mbandi SK, Hesse U, Rees DJ, Christoffels A. A glance at quality score: Implication for de novo transcriptome reconstruction of Illumina reads. Frontiers in Genetics. 2014;**5**:17. DOI: 10.1186/s12859-015-0492-5

[14] Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature Communications. 2016;**7**:11708. DOI: 10.1038/ncomms11708

[15] Gordon A, Hannon GJ. FastX-Toolkit. FASTQ/A Short-reads Preprocessing Tools [Internet]. 2010. Available from: http://hannonlab.cshl.edu/fastx_toolkit/ [Accessed: 01-01-2017]

[16] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;**30**(15):2114-2120. DOI: 10.1093/bioinformatics/btu170

[17] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.[Internet]. 2011 [Accessed: 01-01-2017]

[18] Patel RK, Jain M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. PloS one. 2012;**7**(2):e30619. DOI: 10.1371/journal.pone.0030619

[19] Shrestha RK, Lubinsky B, Bansode VB, Moinz MB, McCormack GP, Travers SA. QTrim: A novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. BMC Bioinformatics. 2014;**15**:33. DOI: 10.1186/1471-2105-15-33

[20] Eldem V, Zararsiz G, Erkan M, Bakir Y. De novo assembly and comprehensive characterization of the skeletal muscle transcriptomes of the European anchovy (*Engraulis encrasicolus*). Marine Genomics. 2015;**20**:7-9. DOI: 10.1016/j.margen.2015.01.001

[21] Martin JA, Wang Z. Next-generation transcriptome assembly. Nature Reviews Genetics. 2011;**12**(10):671-682. DOI: 10.1038/nrg3068

[22] Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;**30**(12):1660-1666. DOI: 10.1093/bioinformatics/btu077

[23] Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nature Methods. 2010;**7**(11):909-912. DOI: 10.1038/nmeth.1517

[24] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;**28**(8):1086-1092. DOI: 10.1093/bioinformatics/bts094

[25] Peng Y, Leung HC, Yiu SM, Lv MJ, Zhu XG, Chin FY. IDBA-tran: A more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics. 2013;**29**(13):i326-i334. DOI:10.1093/bioinformatics/btt219

[26] Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R, et al. BinPacker: Packing-Based de novo transcriptome assembly from RNA-seq data. PLOS Computational Biology. 2016;**12**(2):e1004772. DOI: 10.1371/journal.pcbi.1004772

[27] Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: A new framework for de novo transcriptome assembly using RNA-seq data. Genome Biology. 2015;**16**:30. DOI: 10.1186/s13059-015-0596-2

[28] Durai DA, Schulz MH. Informed kmer selection for de novo transcriptome assembly. Bioinformatics. 2016;**32**(11):1670-1677. DOI: 10.1093/bioinformatics/btw217

[29] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;**8**(8):1494-1512. DOI: 10.1038/nprot.2013.084

[30] Yang Y, Smith SA. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics. 2013;**14**:328. DOI: 10.1186/1471-2164-14-328

[31] Wang S, Gribskov M. Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics. 2017; **33**(3):327-333. DOI: 10.1093/bioinformatics/btw625

[32] Davidson NM, Oshlack A. Corset: Enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biology. 2014;**15**(7):410. DOI: 10.1186/s13059-014-0410-6

[33] Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Research. 1999;**9**(9):868-877

[34] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;**28**(23):3150-3152. DOI: 10.1093/bioinformatics/bts565

[35] Zheng Y, Zhao L, Gao J, Fei Z. iAssembler: A package for de novo assembly of Roche-454/Sanger transcriptome sequences. BMC Bioinformatics. 2011;**12**:453. DOI: 10.1186/1471-2105-12-453

[36] Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Research. 2004;**14**(6):1147-1159. DOI:10.1101/gr.1917404

[37] Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, et al. TIGR gene indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. Bioinformatics. 2003;**19**(5):651-652

[38] Srivastava A, Sarkar H, Malik L, Patro R. Accurate, fast and lightweight clustering of de novo transcriptomes using fragment equivalence classes. arXiv preprint arXix. 2016:1604.03250

[39] Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology. 2014;**15**(12):553. DOI: 10.1186/s13059-014-0553-5

[40] Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. rnaQUAST: A quality assessment tool for de novo transcriptome assemblies. Bioinformatics. 2016;**32**(14):2210-2212. DOI:10.1093/bioinformatics/btw218

[41] Kent WJ. BLAT--the BLAST-like alignment tool. Genome Research. 2002;**12**(4):656-664. DOI: 10.1101/gr.229202

[42] Wu TD, Watanabe CK. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;**21**(9):1859-1875. DOI: 10.1093/bioinformatics/bti310

[43] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;**31**(19):3210-3212. DOI: 10.1093/bioinformatics/btv351

[44] Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. Genome Research. 2016;**26**(8):1134-1144. DOI: 10.1101/gr.196469.115

[45] Misner I, Bicep C, Lopez P, Halary S, Bapteste E, Lane CE. Sequence comparative analysis using networks: Software for evaluating de novo transcript assembly from next-generation sequencing. Molecular Biology and Evolution. 2013;**30**(8):1975-1986. DOI: 10.1093/molbev/mst087

[46] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 1997;**25**(17):3389-3402

[47] Heras Saldana S, Al-Mamun HA, Ferdosi MH, Khansefid M, Gondro C. RNA sequencing applied to livestock production. In: Kadarmideen HN, editor. Systems Biology in Animal Production and Health. 1st ed. Switzerland: Springer; 2016. pp. 63-94. DOI: 10.1007/978331943335.ch4

[48] Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nature Methods. 2015;**12**(4):357-360. DOI: 10.1038/nmeth.3317

[49] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 2013;**14**(4):R36. DOI: 10.1186/gb-2013-14-4-r36

[50] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;**29**(1):15-21. DOI: 10.1093/bioinformatics/bts635

[51] Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;**12**:323. DOI: 10.1186/1471-2105-12-323

[52] Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature Methods. 2017;**14**(2):135-139. DOI: 10.1038/nmeth.4106

[53] Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;**25**(8):1026-1032. DOI: 10.1093/bioinformatics/btp113

[54] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010;**26**(4):493-500. DOI: 10.1093/bioinformatics/btp692

[55] Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotechnology. 2010;**28**(5):503-510. DOI: 10.1038/nbt.1633

[56] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010;**28**(5):511-515. DOI: 10.1038/nbt.1621

[57] Nicolae M, Mangul S, Mandoiu, II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms for Molecular Biology. 2011;**6**(1):9. DOI: 10.1186/1748-7188-6-9

[58] Turro E, Su SY, Goncalves A, Coin LJ, Richardson S, Lewin A. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biology. 2011;**12**(2):R13. DOI: 10.1186/gb-2011-12-2-r13

[59] Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinformatics. 2012;**28**(13):1721-1728. DOI: 10.1093/bioinformatics/bts260

[60] Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods. 2013;**10**(1):71-73. DOI: 10.1038/nmeth.2251

[61] Li W, Jiang T. Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. Bioinformatics. 2012;**28**(22):2914-2921. DOI: 10.1093/bioinformatics/bts559

[62] Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology. 2014;**32**(5):462-464. DOI: 10.1038/nbt.2862

[63] Nariai N, Kojima K, Mimori T, Sato Y, Kawai Y, Yamaguchi-Kabata Y, et al. TIGAR2: Sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads. BMC Genomics. 2014;**15**(Suppl 10):S5. DOI: 10.1186/1471-2164-15-S10-S5

[64] Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biology. 2015;**16**:150. DOI: 10.1186/s13059-015-0702-5

[65] Teng M, Love MI, Davis CA, Djebali S, Dobin A, Graveley BR, et al. A benchmark for RNA-seq quantification pipelines. Genome Biology. 2016;**17**:74. DOI: 10.1186/s13059-016-0940-1

[66] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;**11**(3):R25. DOI: 10.1186/gb-2010-11-3-r25

[67] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;**11**(10):R106. DOI: 10.1186/gb-2010-11-10-r106

[68] Parikh A, Miranda ER, Katoh-Kurasawa M, Fuller D, Rot G, Zagar L, et al. Conserved developmental transcriptomes in evolutionarily divergent species. Genome Biology. 2010;**11**(3):R35. DOI: 10.1186/gb-2010-11-3-r35

[69] Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, et al. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. PLoS Genetics. 2009;**5**(7):e1000569. DOI: 10.1371/journal.pgen.1000569

[70] Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010;**26**(1):136-138. DOI: 10.1093/bioinformatics/btp612

[71] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;**320**(5881):1344-1349. DOI: 10.1126/science.1158441

[72] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;**15**(12):550. DOI: 10.1186/s13059-014-0550-8

[73] Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;**26**(1):139-140. DOI: 10.1093/bioinformatics/btp616

[74] Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014;**15**(2):R29. DOI: 10.1186/gb-2014-15-2-r29

[75] Yang Y, Wang L, Han J, Tang X, Ma M, Wang K, et al. Comparative transcriptomic analysis revealed adaptation mechanism of Phrynocephalus erythrurus, the highest altitude Lizard living in the Qinghai-Tibet Plateau. BMC Evolutionary Biology. 2015;**15**:101. DOI: 10.1186/s12862-015-0371-8

[76] Yang L, Wang Y, Zhang Z, He S. Comprehensive transcriptome analysis reveals accelerated genic evolution in a Tibet fish, *Gymnodiptychus pachycheilus*. Genome Biology and Evolution. 2014;**7**(1):251-261. DOI: 10.1093/gbe/evu279

[77] Shao Y, Wang LJ, Zhong L, Hong ML, Chen HM, Murphy RW, et al. Transcriptomes reveal the genetic mechanisms underlying ionic regulatory adaptations to salt in the crab-eating frog. Scientific Reports. 2015;**5**:17551. DOI: 10.1038/srep17551

[78] Loytynoja A, Goldman N. webPRANK: A phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics. 2010;**11**:579. DOI: 10.1186/1471-2105-11-579

[79] Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. GUIDANCE: A web server for assessing alignment confidence scores. Nucleic Acids Research. 2010;**38**(Web Server issue):W23-W28. DOI: 10.1093/nar/gkq443

[80] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biology. 2010;**59**(3):307-321. DOI: 10.1093/sysbio/syq010

[81] Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;**24**(8):1586-1591. DOI: 10.1093/molbev/msm088

[82] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;**25**(14):1754-1760. DOI: 10.1093/bioinformatics/btp324

[83] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;**9**(4):357-359. DOI: 10.1038/nmeth.1923

[84] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;**20**(9):1297-1303. DOI: 10.1101/gr.107524.110

[85] Lopez-Maestre H, Brinza L, Marchet C, Kielbassa J, Bastien S, Boutigny M, Monnin D, El Filali A, Carareto CM, Vieira C, Picard F, Kremer N, Vavre F, Sagot MF, Lacroix V. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. Nucleic Acids Research. 2016;**44**(19). DOI: 10.1093/nar/gkw655

[86] Li Y, Zhou Z, Tian M, Tian Y, Dong Y, Li S, Liu W, He C. Exploring single nucleotide polymorphism (SNP), microsatellite (SSR) and differentially expressed genes in the jellyfish (*Rhopilema esculentum*) by transcriptome sequencing. Marine Genomics. 2017. DOI: 10.1016/j.margen.2017.01.007

[87] Humble E, Thorne MA, Forcada J, Hoffman JI. Transcriptomic SNP discovery for custom genotyping arrays: Impacts of sequence data, SNP calling method and genotyping technology on the probability of validation success. BMC Research Notes. 2016;**9**(1):418. DOI: 10.1186/s13104-016-2209-x

[88] Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Human Molecular Genetics. 2014;**23**(22):5866-5878. DOI: 10.1093/hmg/ddu309

[89] Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics. 2015;**47**(3):199-208. DOI: 10.1038/ng.3192

[90] Li A, Zhang J, Zhou Z. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics. 2014;**15**:311. DOI: 10.1186/1471-2105-15-311

[91] Sun L, Liu H, Zhang L, Meng J. lncRScan-SVM: A tool for predicting long non-coding RNAs using support vector machine. PloS one. 2015;**10**(10):e0139654. DOI: 10.1371/journal.pone.0139654

[92] Wucher V, Legeai F, Hedan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Research. 2017. DOI: 10.1093/nar/gkw1306

[93] Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Research. 2007;**35**(Web Server issue):W345-W349. DOI: 10.1093/nar/gkm391

[94] Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Research. 2013;**41**(6):e74. DOI: 10.1093/nar/gkt006

[95] Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research. 2013;**41**(17):e166. DOI: 10.1093/nar/gkt646