Quantitative Biology 2016, 4(2): 94–105
DOI 10.1007/s40484-016-0069-y

REVIEW

De novo assembly of transcriptome from next-generation sequencing data

Xuan Li[1,*], Yimeng Kong[1], Qiong-Yi Zhao[2], Yuan-Yuan Li[3] and Pei Hao[3,4,*]

[1] Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China
[2] The University of Queensland, Queensland Brain Institute, St Lucia, QLD 4072, Australia
[3] Shanghai Center for Bioinformation Technology, Shanghai Industrial Technology Institute, Shanghai 201203, China
[4] Key Laboratory of Molecular Virology and Immunology, Institute Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai 200031, China
* Correspondence: lixuan@sippe.ac.cn, phao@sibs.ac.cn

Received November 15, 2015; Revised February 4, 2016; Accepted February 4, 2016

Reconstruction of transcriptome by *de novo* assembly from next generation sequencing (NGS) short-sequence reads provides an essential mean to catalog expressed genes, identify splicing isoforms, and capture the expression detail of transcripts for organisms with no reference genome available. *De novo* transcriptome assembly faces many unique challenges, including alternative splicing, variable expression level covering a dynamic range of several orders of magnitude, artifacts introduced by reverse transcription, etc. In the current review, we illustrate the grand strategy in applying *De Bruijn* Graph (DBG) approach in *de novo* transcriptome assembly. We further analyze many parameters proven critical in transcriptome assembly using DBG. Among them, *k*-mer length, coverage depth of reads, genome complexity, performance of different programs are addressed in greater details. A multi-*k*-mer strategy balancing efficiency and sensitivity is discussed and highly recommended for *de novo* transcriptome assembly. Future direction points to the combination of NGS and third generation sequencing technology that would greatly enhance the power of *de novo* transcriptomics study.

Keywords: transcriptome; *de novo* assembly; *De Bruijn* Graph; next generation sequencing; *k*-mer length; RNA splicing; performance

HISTORY OF NGS AND ITS APPLICATION IN TRANSCRIPTOME STUDY

First generation versus second generation sequencing technologies

Debuted in 1977, the Sanger method [1] and hereafter capillary-based automated sequencing technology [2] represent the first generation sequencing technology that contribute to a series of break-through discoveries, including the completion of the human genome project [3]. Despite its many successes, the impact of the first generation technology is limited by its low throughput, high expense, and sample requirements. To overcome its limitation, a number of so called "second generation sequencing" or "next generation sequencing" (NGS) technologies emerged a decade ago, which employed the massive parallel sequencing scheme. Compared to the first generation sequencing, NGS technologies are characterized with high-throughput but short sequence reads. As the traditional Sanger method can obtain sequence reads of 800–1000 bp, the read length of NGS technologies typically varies from 35 bp (SOLiD) to 700 bp (Roche 454), depending on the platform used [4–6]. The read length of sequences would have serious impact on the reconstruction of transcriptome from RNA-seq data. The Roche 454 platform uses the emulsion PCR

This article is dedicated to Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

for DNA fragments isolation and amplification, and pyrophosphate-based single-nucleotide addition sequencing method on a micro-fabricated array of picoliter-scale wells [4]. Initially, its average read length was 108 bases [4], yet increased gradually to more than 300 bp [7,8]. In the Illumina/solexa platform, the template amplicon is achieved through bridge PCR, followed by four-color cyclic reversible termination steps in sequencing and imaging process [5]. The read length of Illumina/solexa platform is usually 35–150 bp, shorter compared to 454 platform. The reversible terminator chemistry enabled the Illumina/solexa platform with a higher throughput and lower cost [5,7,9]. The LifeTech SOLiD platform takes use of DNA ligase, rather than polymerase, to drive the sequencing by synthesis [6]. These three platforms are the most popular and commercially available technologies, which are widely used in genomics and transcriptomics studies because of their advantage in lower cost and higher throughput compared to the first generation sequencing [4,5,7,9,10].

Second generation sequencing versus other technologies in transcriptomics study

In multi-cellular organisms, the differentiation of cell types and their functions are defined by the constitute and quantity of transcripts, so called transcriptome, inherited from identical genetic make-up. Understanding the transcriptome is essential for interpreting the function and regulation of genes that offer insights into the mechanism of development and diseases [11–15]. The missions of transcriptomics studies are: (i) reconstructing/assembling all transcripts, including mRNA and noncoding RNA [16–18], small RNA [19], etc; (ii) identifying transcript structures, e.g., transcript start/end sites [15], exon-intron structure [20], alternative splicing [21–23], etc; and (iii) quantifying the expression levels of transcripts under certain biological conditions, e.g., development, and stress [24].

Before the development of NGS and its application in transcriptome study (RNA-seq) technologies, a variety of technologies have been used to study transcriptomics, mostly based on the Sanger sequencing or the hybridization technology (Table 1). These traditional methods were often designed for a specific aspect of transcriptome with severe limitations. The advantage of high-throughput and low-cost offered by RNA-seq technology makes it feasible to fully assess the transcriptome of organisms, with or without their genome sequences.

The expressed sequence tags, or EST, which were derived from cDNA libraries, had been proven to be useful in applications of expressed gene identification [25,26], and gene structure determination [27,28]. Because EST, which relies on the Sanger sequencing, is generally low throughput and costly in production, it can rarely be used to quantitative transcripts or discriminate gene expression between tissues or developmental samples. Notably, tag-based methods were developed for respective usage. Serial analysis of gene expression (SAGE) method [29], which counts sequence tags flanking the restriction sites of endonucleases to quantify gene expression, has been used in study of cancers, and other human diseases [30–32]. Its shortage was also noted to analyze changes in the regulatory regions of the transcripts. Another tag-based approach, cap analysis of gene expression (CAGE), captures and counts the 5′-cap region of full-length cDNA [33], and use similar protocol to quantify gene expression by sequencing, in analogy to the SAGE method [33,34]. Although CAGE has advantage in both gene expression analysis and transcription start site identification [35,36], it also failed to reveal variation in gene transcripts. Massively parallel signature sequencing (MPSS) was developed as another approach for gene expression quantification, which isolates template fragments through digestion with type IIs restriction endonuclease and determines their sequence after fixing them on microbead arrays by ligation [37]. Although the tag-based approaches are relatively high-throughput

Table 1. Comparison of methods in transcriptomic study.

Technology	EST	Tag-based method	Microarray	RNA-seq
Principle	Sanger sequencing	Tag-based sequencing	Hybridization	NGS
Throughput	Low	Relative high	High	High
Cost	High	High	High	Low
Reliance on genomic sequence	No	Yes	Yes	No
Background noise	Low	Low	High	Low
Construct full length transcript	No	No	No	Yes
Gene expression quantification	Limited	Yes	Yes	Yes
Alternative splicing identification	Yes	No	Limited	Yes
Dynamic range to quantify gene expression level	Not practical	More than two orders of magnitude	Two or three orders of magnitude	More than five orders of magnitude

compared to EST technology, their disadvantages pertaining to read length (less than 20 bp), cost, and dependence on restriction endonuclease recognition sites, are also obvious [34,38,39]. The hybridization-based microarray technology has been used for years to analyze gene expression using gene probes fixed on glass or silicon chip surface [40]. Although microarray technology is highly advantageous in throughput, the requirements for gene probes and predetermined gene sequences often limit its application in model organisms. Its high background noise and limited expression dynamic range are also factors restricting its usage in transcriptomics study [39,41].

The NGS technology has revolutionized the field of transcriptomics study. In contrast to the traditional methods, RNA-seq provides a comprehensive solution to transcriptomics study in full spectrum. Firstly, because RNA-seq does not rely on the existing gene annotation, it can catalog more transcripts. Not only has RNA-seq been used to identify novel genes such as lincRNA [42], it has also enabled more isoforms (alternative splicing) discovered. For instance, more than 94% genes in human and 61% genes in *Arabidopsis* are found to undergo alternative splicing by RNA-seq study, compared to 35% and 3% through previous methods [20,43–45]. The strand-specific RNA-seq technique offers a unique approach to study and distinguish sense and antisense transcripts [46], one of novel aspects of transcriptome discovered in recent years [47]. Secondly, RNA-seq has been applied in definition of gene structures, including transcriptional start sites [48], regulatory elements [49] and polyadenylation [50]. Thirdly, compared to the limited magnitude in gene expression change that can be detected by microarray, the dynamic range of gene expression can be analyzed by RNA-seq is unprecedented [51]. Moreover, RNA-seq are used to characterize the single nucleotide variation (SNV) [52,53] and RNA-editing activity in gene transcripts [54,55]. Notably, in regard to the non-model organisms that lack genome sequences, the RNA-seq technology is an extremely valuable tool. There are an increasing number of non-model species that have undergone transcriptomics study before their genome sequences were determined, especially for those crops of polyploidy [13,56,57].

RNA-seq has been applied in transcriptomics studies in many plant and animal species. It leads to new discoveries in alternative splicing and gene structure in model organisms, including human [51,58], *Drosophila* [11], *Arabidopsis* [44], and rice [59], etc. It proves to be a powerful tool in study of unusual *trans*-splicing genes [21,60–62]. Long noncoding RNA [63], new ORFs within UTR [15], antisense transcripts [47], and gene fusion [64] are some of the new fronts attributed to RNA-seq. Currently, the submitted RNA-seq data to public databases are exponentially increased each year [65]. There are more than 60 plants [66] and a large number of animals, including insects [67–69], fishes [42], birds [22], mammals [16,70] that are subjected to *de novo* transcriptome study by RNA-seq, yielding insights into the mechanism of development and gene regulation. *De novo* assembly of transcriptome is a major challenge in using RNA-seq technology for transcriptomics study. Another strategy, the reference-based assembly, which relies on a reference genome to first align all the reads to the genome and then cluster those overlapping reads into transcripts (such as Cufflinks [71]), will be covered in details in a separate chapter. In this review, we focus on the development of algorithms and computation details in the *de novo* assembly of transcriptome.

De novo ASSEMBLY OF TRANSCRIPTOME

Application of *De Bruijn* graph in *de novo* assembly of short-sequence reads

The use of *De Bruijn* graph (DBG) in assembly from short-sequence reads was first applied in EULER assembler [72]. Different from the overlap-layout-consensus approach [73], sequencing data are dissected into *k*-mers (words of *k* nucleotides) and organized into graph, consisting of paths. Paths are formed from *k*-mers in a certain order. Utilities that assembly from short-sequence reads have only been developed more recently [74–80] after the emerging of next-generation sequencing technologies. These DBG assemblers often consist of several programs that perform error correction, merging sequences, path building, repeat resolution, paths separation, and scaffolding with paired-end/mate-pair reads. The major *de novo* transcriptome assemblers and softwares are summarized in Table 2.

The challenge of *de novo* transcriptome assembly

Reconstruction of the full transcriptome by *de novo* assembly from sequence reads, will help catalog expressed genes, identify splicing isoforms, and capture the expression detail of all transcripts. Without reference genomes, the *de novo* assembly approach is considered to be more difficult than the assembly of *de novo* genome using short sequence reads [86]. In comparison to *de novo* genomics assembly, *De novo* transcriptome assembly faces many unique challenges [87]. Among them, transcripts are expressed from low, medium to high level, which can cover a dynamic range of gene expression in several orders of magnitude [78]. In plants, the range of gene expression in leaves was reported to spans more than five orders of magnitude [88,89]. The NGS platforms have associated biases, i.e., sequence

Table 2. Existing *de novo* transcriptome assemblers.

Assembler	Support for multiple *k*-mer	Support for paired end reads	Support for standard reads	Ref.
Multiple-*k*	Yes	Yes	Yes	[81]
SOAPdenovo	No	Yes	Yes	[80]
ABySS	No	Yes	Yes	[82]
Trans-ABySS	Yes	Yes	No	[83]
Oases	(Oases-M)	Yes	Yes	[84]
Trinity	No	Yes	Yes	[85]

reads redundancy, error rates tendency, etc., which can further skew the transcript data [90]. Transcript isoforms due to alternative splicing, which is only pertaining to gene transcripts from eukaryotes, is another critical issue that *de novo* assembly has to address [81]. Forming contigs from isolating paths can be hindered by sequence repeats and nucleotide variations, and also by alternatively spliced isoforms. Artifacts introduced during reverse transcription is proven to be another serious concern, as the noise unique to RNA-seq experimental process is not encountered in genome study. Taken together, these accompany factors grossly compound the difficulty in *de novo* transcriptome assembly.

Strategies for preprocessing and filtering sequence reads

High-throughput sequencing data are often contaminated with artificial elements, generated from library construction and/or PCR amplification. Thus, preprocessing of sequence reads to remove artifacts from RNA-seq data sets before assembly is essential to improve computational efficiency and assure the accuracy of assemblies. For RNA-seq data, this step targets mainly four types of sequences: adaptors [91,92], low-complexity reads [91], PCR products of non-biological origin [93], and rRNA. Tools recommended for the tasks include hardwired solution [94] to programmed tools [91,92,95].

The high-throughput sequencing (NGS) data are found to contain sequence errors pertaining to the new technologies. For example, the Illumina platform has, in general, an error rate between 1% and 3%, with overwhelmingly substitution errors [96]. They are distributed non-randomly with the error rate increasing from 5′- to 3′-end [97]. Other NGS technologies display similar characteristics [90]. Sequencing errors are serious compounding factor affecting the performance of *De Bruijn* graph, increasing its size and complexity, and demanding extended memory space and processing powers of computers [98,99]. There is added difficulty to separate sequence errors from genuine variations for RNA-seq data, due to the complication of post-transcription processing of RNA. Thus, sequencing error correction becomes essential in preparing transcriptome assembly.

Removal of sequence of low quality score is a common practice in filtering sequencing reads (a useful review by Yang *et al*. [96]). Two approaches are typically employed. Low-quality scored nucleotides are trimmed off at either one or both ends of the sequencing reads. Alternatively, average quality score is computed for a sliding window of fixed number of bases, and sequence regions with score below certain threshold are removed. Many tools have been developed for such tasks, namely FASTX TOOLKIT [100], Sickle [101], FastQC [102], TRIMMOMATIC [103], BIO-PIECES [104], and UrQt [105]. Although aggressive quality filtering is often employed to ensure the quality of data to be used in follow-up analysis, sometimes this can result in discarding a substantial portion of sequence data, thus may disproportionately affecting some transcripts with biased nucleotide content or lower expression level [106,107]. To determine the optimal approach in filtering sequence reads, especially for RNA-seq data, MacManes and Eisen performed some carefully designed study [108,109]. Their results indicated that significant improvement on assembly accuracy was achieved by applying the error correction process [109]. However, they noted stringent trimming of nucleotides with quality scores $\leqslant 20$ produced poorer transcriptome assembly, measured with several different metrics [108]. Thus, researchers interested in *de novo* transcriptome assembly are advised to use more gentle quality trimming scheme, or no trimming to achieve the most favorable results [108,110].

The efficiency of different *K*-mer length

The *K*-mer length is a critical parameter in assembly using *De Bruijn* graph, even more so for transcriptome assembly. The assembly quality of a *De Bruijn* graph is highly variable depending on the *k*-mer length, which defines sequence overlap between two contiguous reads. For genomics assembly, there is generally uniform reads coverage across the genome, so the optimal *k*-mer length

is determined as a function related to sequencing depth [111]. However, for transcriptomics assembly, the system is complicated with the complexity of an organism's gene contents, the highly variable gene expression levels, and multiple isoforms from alternative splicing of the same gene, in addition to the variables such as sequence depth, error rate, etc [78,83]. On top of these, the widely existing isoforms of transcribed genes in animals and plants [112–114] prevent the use of coverage depth for resolving repeated motifs. Hence, the optimal *k*-mer length in transcriptome assembly is affected by a lot of more factors [79]. In practice the *k*-mer value for transcriptome assembly is sometimes determined based on the particular study. When a more contiguous assembly is the primary goal and the loss of lowly expressed transcripts is not a concern, a large *k*-mer length is preferred. On the other hand, small *k*-mer length is often used to capture poorly expressed transcripts, resulting in more fragmented and diverse transcripts. Such theoretical scheme and inter-dependency of variables in transcriptome assembly were tested and best defined in the experimental or simulated studies using model organisms [87,115]. In common practice, the length of *k*-mer is arbitrarily decided to use an intermediate value, often as the result of a compromise between the conflicting goals of transcript diversity and transcript contiguity. Gruenheit *et al*. noted the close correlation between the *k*-mer size and the coverage depth cutoff of a transcriptome assembly, and that both parameters need to be optimized in a balanced approach for the best outcome [116]. Zhao *et al*. extensively analyzed the efficiency of *k*-mer size from small to large in capturing transcripts at different expression quantiles [87]. Interestingly, they have shown that with the same *k*-mer length, the efficiency for capturing transcripts at different expression quantiles varied greatly. In the single *k*-mer settings, when measuring with percentage of full-length transcript constructed, Trinity [71] performed well cross the full spectrum of gene expression levels, whereas SOAPdenovo had the worst outcome in both low and high expression quantiles. The outstanding performance of Trinity, is due in part to its implementation of an enumeration algorithm after construction of *De Bruijn* graph from RNA-seq data. The algorithm scores all possible paths and branches, recovers paths supported by actual reads and removes ambiguous/erroneous edges, so to retain those plausible ones. Its broad applicability was demonstrated in recovering full-length transcripts and isoforms in yeast, mouse, whitefly, and other non-model organisms [71,117]. From the above analysis [87,116], it was suggested that an extensive pre-testing and evaluation of *k*-mer length for different transcriptome assemblers is needed for each individual case to determine an appropriate *k*-mer size. Even so there is no guarantee that an optimal outcome can be achieved for a given organism. People are advised to use an approach of multiple *k*-mer size, which will be the focus of next section. *De novo* transcriptome assembly taking advantage of multiple *k*-mer length is highly desirable.

The multi-*k*-mer strategy balancing efficiency and sensitivity

The strategy to use multiple *k*-mers of different length was initially proposed by Robertson *et al*. [83], and later by Surget-Groba and Montoya-Burgos [81]. The principle of the multi-*k*-mer approach is to assemble transcriptome with various *k*-mer lengths at first. Then the outputs from the first step are merged to form a final assembly. Using the testing RNA-seq data set of *A. aegypti* [115], the assembly with single *k*-mer (= 21) gave a good compromise between the number of contigs and their average length (measured with N50 value), which was determined by comparing to the reference transcriptome from the work of Gibbons *et al*. [115]. Impressively, the final assembly from multi-*k*-mer approach vastly outperformed each single *k*-mer assembly, marked by substantial improvement in contiguity and increased number of contigs with length over 100 bp [81]. The authors also noted the multi-*k*-mer result achieved the highest coverage of the reference transcriptome, in which the base coverage of reference transcripts was doubled compared to the single *k*-mer assemblies.

The multi-*k*-mer approach has been applied to a number of different assemblers, including Trans-Abyss [83], SOAPdenovo-MK [87], and Oases-MK [87]. Robertson *et al*. observed that transcripts with lower or higher read depth were represented more effectively with smaller or larger *k*-mer values, respectively. They concluded that assembly across a range of *k*-mer values may be essential to recover transcripts with very different expression levels [83]. Their multi-*k*-mer version of the Abyss assembler, Trans-ABySS reported the numbers of transcripts were comparable to that produced by Cufflinks [118] that uses the output of the read aligner TopHat [119] to reconstruct transcripts. They believed that *de novo* assembly using multi-*k*-mer approach offered a sensitive and effective method to address the issues of variable expression levels and multiple transcript isoforms. They noted that for genes with contig-to-exon coverage ratio $\geqslant 0.8$, Trans-ABySS and ALEXA-seq [120] (a tool for expression analysis by sequencing) had well correlated expression estimates (Pearson's correlation coefficient = 0.921) [83]. In order to evaluate the efficiency and performance of single *k*-mer verse multi-*k*-mer conditions, Zhao *et al*. built a matrix of performance, including the number of transcripts $>$100 bp, N50 value, total number of transcripts, total transcript length, and number of full length transcripts captured at different expression quantiles [87].

Using the RNA-seq data sets from yeast and *Drosophila*, they observed, for all tested assemblers, the multi-*k*-mer method had a significant improvement in the full range of coverage depth over their single *k*-mer peer. This holds true for both *S. pombe* (~6,000 genes) and *D. melanogaster* (more complex; ~30,000 genes) transcriptomes. Impressively, their work illustrated the efficiency of assembly in capturing transcripts in full spectrum of expression quantiles [87]. They showed that transcripts at both ends (low and high expression quantiles) were not efficiently recovered by any single *k*-mer length. The major improvement by multi-*k*-mer approach over their single *k*-mer peers was observed in the high-quantiles range, but less significant in the low quantiles [87]. However, the multi-*k*-mer method is hindered by its complication in computation. Melicher *et al*. used cloud computing to build a bioinformatics pipeline to assemble and analyze the transcriptome with off-site data management and processing [121]. They employed Velvet-Oases (using various *k*-mer length) or Trinity (*k*-mer = 21) for the initial assembly and performed a secondary assembly with CAP3. By reconstructing transcriptomes from three non-model organisms, they demonstrated that their pipeline and the multi-*k*-mer method can be used broadly to assemble higher quality transcriptomes than any single *k*-mer approach [121].

The coverage depth of RNA-seq reads

While genomic sequencing coverage is generally uniform across the genome, transcriptomics sequencing coverage is highly variable that prevents the use of coverage information from resolving repeated motifs [78]. Zhao *et al*. showed that with increasing coverage depth, generally a larger number of transcripts and more total bases were assembled. However, the transcripts' mean length and N50, after an initial increase, peaked at a threshold and started to decrease [87]. On the other hand, the percentage of RMBT (reads mapped back to assembled transcripts) had a pattern reversely correlated to increasing coverage depth for all assemblers except Trinity. The percentage of RMBT is an important benchmark for evaluating the performance of an assembler. An optimal program should use as many reads as possible to reconstruct high-quality transcriptome. Trinity reached almost 90% in RMBT, which may be attributed to its greedy *k*-mer based approach at the Inchworm step. Oases-MK came in second for this measure. Given the lower value in RMBT, the performance of SOAPdenovo was not satisfactory [87]. The peak of the mean transcript length and N50 seems to be correlated to the complexity of genome for a species. Similar pattern was observed with the number of constructed full-length transcripts. A peak was reached after initial increase with the increase of coverage depth, before the number of full-length transcripts started to drop. The turning points appeared to be related to the complexity of the genome, for which it was 3G (sequencing data) for fruit fly, and between 1G and 3G for fission yeast [87]. Others found the quantitative difference in the assemblies using whole-animals RNA-seq data versus tissues data [122]. In assemblies from whole-animal data, increasing reads led to rapid increase of short transcripts and discovery of conserved genes. But single-tissue assemblies showed a slower discovery of conserved genes but often with longer transcripts. Additional study showed, in the mouse assemblies, more reads also led to more frequent assembly errors which must be mitigated using more stringent parameters [122]. Gruenheit *et al*. noted that *k*-mer size and read coverage depth are interacting factors that need to be considered simultaneously [116]. Their analysis showed that varying *k*-mer length with the coverage cutoff had a significant impact on the success of gene assemblies, and both parameters, *k*-mer length and reads coverage cut-off, need to be optimized together for the best outcomes.

Other considerations and future direction

De novo transcriptome assembly facilitates the study of organisms whose genome sequences are not available. However, such tasks also create new challenges for accurately assessing the quality of an assembly. Commonly, many parameters used in genomic assembly are referred in transcriptome assemblies, such as median contig length, number of contigs, and N50 [123,124]. However, these measures were proven insufficient and unreliable [125,126]. With available reference genome, the reference-based approach is helpful to estimate the accuracy and completeness of an assembly. By comparing assembled transcripts to a reference transcript set, the fraction of assemblies matching a reference, the fraction of reference being matched, and the fraction of assemblies containing complete CDS can be estimated with high confidence [87,123,125,127]. More recently, methods for evaluating *de novo* transcriptome assembly not relying on reference genome were developed [128,129]. They, instead, use a probabilistic model to assess an assembly and its underlining sequencing read data. Although these statistics-model based methods were powerful tool and showed to accurately reflect assembly quality in many tested cases, care must be taken as discrepancy was also noted when compared to traditional measures or reference-based approach.

The recent advances in *de novo* transcriptome assembly have enabled the expansion of RNA-seq studies to many organisms, with or without high-quality reference genome available. In light of such broad application of RNA-seq technology, there are other factors warranting considera-

tion. While it is often critical in assembly of large genome, resources usage for transcriptome assembly bears some equal importance for practical reason. Zhao *et al*. recorded the dramatic difference in performance among Oases, Trinity, ABySS, and SOAPdenovo when the same *Drosophila* RNA-seq data set was used for *de novo* assembly [87]. They noted that Oases was the most sensitive, and ABySS was the least sensitive in response to increasing data size, although generally memory usage displayed a good correlation with data size. The *k*-mer length also had a great impact on both memory usage and runtime. While runtimes for ABySS, Oases, and SOAPdenovo were reversely correlated with the *k*-mer length, memory usage remained almost constant for SOAPdenovo and ABySS, except for Oases whose memory usage had a reverse correlation with *k*-mer length. They also found that processing of a large data set by Trinity can exceed reasonable execution time and hence becomes impractical [87]. Trinity was initially built for reconstruction of full-length transcripts with maximum sensitivity [71]. Its efficiency was later improved by halving memory requirements and increasing processing speed via parallelization [130]. Currently, its newer release was recommended to have ~1 GB of memory per 1 million paired-end reads. A common multi-core server with 256 GB to 1 TB of memory would be sufficient for a set-up at departmental core facility [117]. Recently, new *de novo* transcriptome assemblers were developed. For instance, SOAPdenovo-Trans [131] took the advantage of the error-removal model from Trinity [85] and the robust heuristic graph traversal method from Oases [84]. Bridger [132] incorporated the key ideas of *de novo* assembler Trinity [85] and reference-based assembler Cufflinks [118] to construct the splicing graphs and the full-length transcripts.

Requirements for computation resources by assembling large transcriptome data sets can be mitigated with high-performance cluster computing. To take use of high-performance computing with thousands of CPU cores, many transcriptome assemblers, like Trinity [117], Oases [84], Trans-ABySS [83], Rnnotator [93], and SOAPdenovo [133], employ parallel computing methods of different levels. More recently, cloud computing (refer to review [134]) becomes increasingly popular with the bioinformatics community, as resources are rented as a service per a user's need. The Hadoop-Based project in developing MapReduce programming paradigm is underway as a community effort, attributing to the effectiveness of MapReduce in parallelization of bioinformatics algorithms, particularly those as the leading application in the area of NGS data analysis [135]. To solve the large transcriptome assembly problem, a scalable cloud-based solution is deemed to be the destination to meet future computation needs.

Transcriptome analysis has seen the transition from microarray technology to high-throughput NGS technologies. The RNA-seq approach provides transcriptome profiling and analysis as a "comprehensive" solution that is superior to other methods we have mentioned in the introductory section. Meanwhile, as the RNA-seq technology and experimental protocols continue to evolve, we foresee the emerging of new challenges for bioinformaticians. Many groups and commercial vendors are currently developing different flavors of the third generation sequencing technology [136,137], which are characterized with longer reads, single molecule, or realtime data. For example, RNA-seq reads from PacBio [138] have much longer reads (several kilobases) to enable it to sequence a single transcript to its full length, but are accompanied with high error rates (~15%). The PacBio's long reads are more advantageous when the error rate issue is mitigated with circular consensus sequencing mode [139]. PacBio technology has been applied in updating genome sequence [140] and in evaluating the assembly efficiency of *de novo* assembly [141]. In such scenario, bioinformatics tools resolving sequence errors by combining second and third generation sequencing data would become most valuable in transcriptomics profiling and analysis.

ABBREVIATIONS

CAGE, cap analysis of gene expression; cDNA, complementary DNA; CDS, coding sequence; CPU, central processing unit; DBG, *De Bruijn* Graph; MPSS, massively parallel signature sequencing; NGS, next generation sequencing; ORF, open reading frame; RMBT, reads mapped back to assembled transcripts; SAGE, serial analysis of gene expression; SNV, single nucleotide variation; SOLiD, sequencing by oligonucleotide ligation and detection; UTR, untranslated region

ACKNOWLEDGEMENTS

This work is supported in part by grants from the National Basic Research Program of China (Nos. 2012CB316501, 2012CB517905 and 2013CB127000) and the National Natural Science Foundation of China (Nos. 31571310 and 31271409).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xuan Li, Yimeng Kong, Qiong-Yi Zhao, Yuan-Yuan Li and Pei Hao declare they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA, 74, 5463–5467
2. Kheterpal, I., Scherer, J. R., Clark, S. M., Radhakrishnan, A., Ju, J.,

Ginther, C. L., Sensabaugh, G. F. and Mathies, R. A. (1996) DNA sequencing using a four-color confocal fluorescence capillary array scanner. Electrophoresis, 17, 1852–1859

3. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature, 431, 931–945
4. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437, 376–380
5. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456, 53–59
6. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., *et al.* (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res., 18, 1051–1063
7. Metzker, M. L. (2010) Sequencing technologies — the next generation. Nat. Rev. Genet., 11, 31–46
8. Morozova, O., Hirst, M. and Marra, M. A. (2009) Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genomics Hum. Genet., 10, 135–151
9. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. Nat. Biotechnol., 26, 1135–1145
10. Mardis, E. R. (2008) The impact of next-generation sequencing technology on genetics. Trends Genet., 24, 133–141
11. Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., *et al.* (2011) The developmental transcriptome of *Drosophila melanogaster*. Nature, 471, 473–479
12. Li, C.-F., Zhu, Y., Yu, Y., Zhao, Q.-Y., Wang, S.-J., Wang, X.-C., Yao, M.-Z., Luo, D., Li, X., Chen, L., *et al.* (2015) Global transcriptome and gene regulation network for secondary metabolite biosynthesis of tea plant (*Camellia sinensis*). BMC Genomics, 16:560
13. Wang, X. C., Zhao, Q. Y., Ma, C. L., Zhang, Z. H., Cao, H. L., Kong, Y. M., Yue, C., Hao, X. Y., Chen, L., Ma, J. Q., *et al.* (2013) Global transcriptome profiles of *Camellia sinensis* during cold acclimation. BMC Genomics, 14, 415
14. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 464, 768–772
15. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 320, 1344–1349
16. Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grtzner, F. and Kaessmann, H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 505, 635–640
17. Nam, J. W. and Bartel, D. P. (2012) Long noncoding RNAs in *C. elegans*. Genome Res., 22, 2529–2540
18. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J. L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev., 25, 1915–1927
19. Chen, X., Gao, C., Li, H., Huang, L., Sun, Q., Dong, Y., Tian, C., Gao, S., Dong, H., Guan, D., *et al.* (2010) Identification and characterization of microRNAs in raw milk during different periods of lactation, commercial fluid, and powdered milk products. Cell Res., 20, 1128–1137
20. Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res., 22, 1184–1195
21. Shao, W., Zhao, Q. Y., Wang, X. Y., Xu, X. Y., Tang, Q., Li, M., Li, X. and Xu, Y. Z. (2012) Alternative splicing and trans-splicing events revealed by analysis of the *Bombyx mori* transcriptome. RNA, 18, 1395–1407
22. Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. Science, 338, 1587–1593
23. Xu, P., Kong, Y., Song, D., Huang, C., Li, X. and Li, L. (2014) Conservation and functional influence of alternative splicing in wood formation of *Populus* and *Eucalyptus*. BMC Genomics, 15, 780
24. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol., 31, 46–53
25. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252, 1651–1656
26. Aaronson, J. S., Eckman, B., Blevins, R. A., Borkowski, J. A., Myerson, J., Imran, S. and Elliston, K. O. (1996) Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. Genome Res., 6, 829–845
27. Kan, Z. Y., Rouchka, E. C., Gish, W. R. and States, D. J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. Genome Res., 11, 889–900
28. Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res., 29, 2850–2859
29. Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) Serial analysis of gene-expression. Science, 270, 484–487
30. Alvarez, H., Corvalan, A., Roa, J. C., Argani, P., Murillo, F., Edwards, J., Beaty, R., Feldmann, G., Hong, S. M., Mullendore, M., *et al.* (2008) Serial analysis of gene expression identifies connective tissue growth factor expression as a prognostic biomarker in gallbladder cancer. Clin. Cancer Res., 14, 2631–2638
31. Horan, M. P. (2009) Application of serial analysis of gene expression to the study of human genetic disease. Hum. Genet., 126, 605–614
32. Honda, H., Barrueto, F. F., Gogusev, J., Im, D. D. and Morin, P. J. (2008) Serial analysis of gene expression reveals differential expression between endometriosis and normal endometrium. Possible roles for *AXL* and *SHC1* in the pathogenesis of endometriosis. Reprod. Biol. Endocrinol., 6–59
33. Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., *et al.* (2006) CAGE: cap analysis of gene expression. Nat. Methods, 3, 211–222
34. Harbers, M. and Carninci, P. (2005) Tag-based approaches for transcriptome research and genome annotation. Nat. Methods, 2,

495–502
35. Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA, 100, 15776–15781
36. Maekawa, S., Matsumoto, A., Takenaka, Y. and Matsuda, H. (2007) Tissue-specific functions based on information content of gene ontology using cap analysis gene expression. Med. Biol. Eng. Comput., 45, 1029–1036
37. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S. J., McCurdy, S., Foy, M., Ewan, M., *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol., 18, 630–634
38. Ozsolak, F. and Milos, P. M. (2011) RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet., 12, 87–98
39. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 10, 57–63
40. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995) Quantitative monitoring of gene-expression patterns with a complementary-dna microarray. Science, 270, 467–470
41. Okoniewski, M. J. and Miller, C. J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics, 7, 276
42. Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan, L., Sandelin, A., Rinn, J. L., Regev, A., *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res., 22, 577–591
43. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature, 456, 470–476
44. Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., Wong, W. K. and Mockler, T. C. (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res., 20, 45–58
45. Keren, H., Lev-Maor, G. and Ast, G. (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet., 11, 345–355
46. Mamanova, L., Andrews, R. M., James, K. D., Sheridan, E. M., Ellis, P. D., Langford, C. F., Ost, T. W. B., Collins, J. E. and Turner, D. J. (2010) FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat. Methods, 7, 130–132
47. Faghihi, M. A. and Wahlestedt, C. (2009) Regulatory roles of natural antisense transcripts. Nat. Rev. Mol. Cell Biol., 10, 637–643
48. Yamashita, R., Sathira, N. P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., Hashimoto, S. i., Sugano, S., Nakai, K. and Suzuki, Y. (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res., 21, 775–789
49. Zhang, S. J., Liu, C. J., Yu, P., Zhong, X., Chen, J. Y., Yang, X., Peng, J., Yan, S., Wang, C., Zhu, X., *et al.* (2014) Evolutionary interrogation of human biology in well-annotated genomic framework of Rhesus Macaque. Mol. Biol. Evol., 31, 1309–1324
50. Derti, A., Garrett-Engele, P., MacIsaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M. and Babak, T. (2012) A quantitative atlas of polyadenylation in five mammals. Genome Res., 22, 1173–1183
51. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 5, 621–628
52. Jia, G., Huang, X., Zhi, H., Zhao, Y., Zhao, Q., Li, W., Chai, Y., Yang, L., Liu, K., Lu, H., *et al.* (2013) A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). Nat. Genet., 45, 957–961
53. Kumar, S., Banks, T. W. and Cloutier, S. (2012) SNP discovery through next-generation sequencing and its applications. Int. J. Plant Genomics, 2012, 1–15
54. Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P., O'Connell, M. A. and Li, J. B. (2013) Identifying RNA editing sites using RNA sequencing data alone. Nat. Methods, 10, 128–132
55. Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C. and Li, J. B. (2012) Accurate identification of human Alu and non-Alu RNA editing sites. Nat. Methods, 9, 579–581
56. Ward, J. A., Ponnala, L. and Weber, C. A. (2012) Strategies for transcriptome analysis in nonmodel plants. Am. J. Bot., 99, 267–276
57. Duan, J. L., Xia, C., Zhao, G. Y., Jia, J. Z. and Kong, X. Y. (2012) Optimizing *de novo* common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics, 13, 392
58. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet., 40, 1413–1415
59. Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., Zhuang, R., Lu, Z., He, Z., Fang, X., *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res., 20, 646–654
60. Allen, M. A., Hillier, L. W., Waterston, R. H. and Blumenthal, T. (2011) A global analysis of *C. elegans* trans-splicing. Genome Res., 21, 255–264
61. McManus, C. J., Duff, M. O., Eipper-Mains, J. and Graveley, B. R. (2010) Global analysis of trans-splicing in *Drosophila*. Proc. Natl. Acad. Sci. USA, 107, 12975–12979
62. Kong, Y., Zhou, H., Yu, Y., Chen, L., Hao, P. and Li, X. (2015) The evolutionary landscape of intergenic trans-splicing events in insects. Nat. Commun., 6, 8734
63. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res., 22, 1775–1789
64. Nacu, S., Yuan, W., Kan, Z., Bhatt, D., Rivers, C., Stinson, J., Peters, B. A., Modrusan, Z., Jung, K., Seshagiri, S., *et al.* (2011) Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med. Genomics, 4, 11
65. Rung, J. and Brazma, A. (2012) Reuse of public genome-wide gene expression data. Nat. Rev. Genet., 14, 89–99
66. Schliesky, S., Gowik, U., Weber, A. P. M. and Braeutigam, A. (2012) RNA-seq assembly — are we there yet? Front. Plant Sci., 3, 220
67. He, W., You, M., Vasseur, L., Yang, G., Xie, M., Cui, K., Bai, J., Liu, C., Li, X., Xu, X., *et al.* (2012) Developmental and insecticide-resistant insights from the *de novo* assembled transcriptome of the diamondback moth, *Plutella xylostella*. Genomics, 99, 169–177
68. Zhan, S., Merlin, C., Boore, J. L. and Reppert, S. M. (2011) The monarch butterfly genome yields insights into long-distance migra-

tion. Cell, 147, 1171–1185

69. Akbari, O. S., Antoshechkin, I., Amrhein, H., Williams, B., Diloreto, R., Sandler, J. and Hay, B. A. (2013) The developmental transcriptome of the mosquito *Aedes aegypti,* an invasive species and major arbovirus vector. G3, 3, 1493–1509
70. Merkin, J., Russell, C., Chen, P. and Burge, C. B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599
71. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol., 29, 644–652
72. Pevzner, P. A., Tang, H. X. and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA, 98, 9748–9753
73. Batzoglou, S. (2004). Algorithmic challenges in mammalian whole-genome assembly. In Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. John Wiley & Sons, Ltd
74. Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P. and Batzoglou, S. (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS One, 2, e484
75. Warren, R. L., Sutton, G. G., Jones, S. J. M. and Holt, R. A. (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics, 23, 500–501
76. Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. Genome Res., 17, 1697–1706
77. Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L. and Jones, C. D. (2007) Extending assembly of short DNA sequences to handle error. Bioinformatics, 23, 2942–2944
78. Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using *de Bruijn* graphs. Genome Res., 18, 821–829
79. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. Genome Res., 19, 1117–1123
80. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., *et al.* (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res., 20, 265–272
81. Surget-Groba, Y. and Montoya-Burgos, J. I. (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Res., 20, 1432–1440
82. Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., *et al.* (2009) *De novo* transcriptome assembly with ABySS. Bioinformatics, 25, 2872–2877
83. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., *et al.* (2010) *De novo* assembly and analysis of RNA-seq data. Nat. Methods, 7, 909–912
84. Schulz, M. H., Zerbino, D. R., Vingron, M. and Birney, E. (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics, 28, 1086–1092
85. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol., 29, 644–652
86. Schuster, S. C. (2008) Next-generation sequencing transforms today's biology. Nat. Methods, 5, 16–18
87. Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X. and Hao, P. (2011) Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics, 12, S2
88. Braeutigam, A., Kajala, K., Wullenweber, J., Sommer, M., Gagneul, D., Weber, K. L., Carr, K. M., Gowik, U., Mass, J., Lercher, M. J., *et al.* (2011) An mRNA blueprint for C-4 photosynthesis derived from comparative transcriptomics of closely related C-3 and C-4 species. Plant Physiol., 155, 142–156
89. Gowik, U., Brautigam, A., Weber, K. L., Weber, A. P. M. and Westhoff, P. (2011) Evolution of C-4 photosynthesis in the genus *Flaveria*: how many and which genes does it take to make C-4? Plant Cell, 23, 2087–2105
90. Wang, Y., Yu, Y., Pan, B., Hao, P., Li, Y., Shao, Z., Xu, X. and Li, X. (2012) Optimizing hybrid assembly of next-generation sequence data from *Enterococcus faecium*: a microbe with highly divergent genome. BMC Syst. Biol., 6(Suppl 3), S21
91. Falgueras, J., Lara, A. J., Fernandez-Pozo, N., Canton, F. R., Perez-Trabado, G. and Claros, M. G. (2010) SeqTrim: a high-throughput pipeline for preprocessing any type of sequence reads. BMC Bioinformatics, 11, 38
92. Lassmann, T., Hayashizaki, Y. and Daub, C. O. (2009) TagDust-a program to eliminate artifacts from next generation sequencing data. Bioinformatics, 25, 2839–2840
93. Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M. and Wang, Z. (2010) Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics, 11,663
94. Shi, H., Schmidt, B., Liu, W. and Mueller-Wittig, W. (2010) A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware. J. Comput. Biol., 17, 603–615
95. Kelley, D. R., Schatz, M. C. and Salzberg, S. L. (2010) Quake: quality-aware detection and correction of sequencing errors. Genome Biol., 11, R116
96. Yang, X., Chockalingam, S. P. and Aluru, S. (2013) A survey of error-correction methods for next-generation sequencing. Brief. Bioinform., 14, 56–66
97. Liu, B., Yuan, J., Yiu, S.-M., Li, Z., Xie, Y., Chen, Y., Shi, Y., Zhang, H., Li, Y., Lam, T.-W., *et al.*(2012) COPE: an accurate *k*-mer-based pair-end reads connection tool to facilitate genome assembly. Bioinformatics, 28, 2870–2874
98. Conway, T. C. and Bromage, A. J. (2011) Succinct data structures for assembling large genomes. Bioinformatics, 27, 479–486
99. Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M. and Brown, C. T. (2012) Scaling metagenome sequence assembly with probabilistic *de Bruijn* graphs. Proc. Natl. Acad. Sci. USA, 109, 13272–13277
100. HannonLab. (2009) FASTX TOOLKIT. http://hannonlab.cshl.edu/fastx_toolkit/
101. Joshi, N. A. and Fass, J. N. (2011) Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle
102. Andrews, S. (2010). FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
103. Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M. and Usadel, B. (2012) RobiNA: a user-friendly, integrated software

solution for RNA-Seq-based transcriptomics. Nucleic Acids Res., 40, W622–W627

104. Hansen, M. A., Oey, H., Fernandez-Valverde, S., Jung, C.-H. and Mattick, J. S. (2008). Biopieces: a bioinformatics toolset and framework. In 19th International Conference on Genome Informatics
105. Modolo, L. and Lerat, E. (2015) UrQt: an efficient software for the unsupervised quality trimming of NGS data. BMC Bioinformatics, 16, 137
106. Riesgo, A., Perez-Porro, A. R., Carmona, S., Leys, S. P. and Giribet, G. (2012) Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. Mol. Ecol. Resour., 12, 312–322
107. Looso, M., Preussner, J., Sousounis, K., Bruckskotten, M., Michel, C. S., Lignelli, E., Reinhardt, R., Hoeffner, S., Krueger, M., Tsonis, P. A., *et al.*(2013) A *de novo* assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. Genome Biol., 14, R16
108. MacManes, M. D. (2014) On the optimal trimming of high-throughput mRNA sequence data. Front. Genet., 5, 13
109. MacManes, M. D. and Eisen, M. B. (2013) Improving transcriptome assembly through error correction of high-throughput sequence reads. PeerJ, 1, e113
110. Mbandi, S. K., Hesse, U., Rees, D. J. G. and Christoffels, A. (2014) A glance at quality score: implication for *de novo* transcriptome reconstruction of Illumina reads. Front. Genet., 5, 17
111. Compeau, P. E. C., Pevzner, P. A. and Tesler, G. (2011) How to apply *de Bruijn* graphs to genome assembly. Nat. Biotechnol., 29, 987–991
112. Blumenthal, T. (1998) Gene clusters and polycistronic transcription in eukaryotes. BioEssays, 20, 480–487
113. Kazan, K. (2003) Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged. Trends Plant Sci., 8, 468–471
114. Leff, S. E. and Rosenfeld, M. G. (1986) Complex transcriptional units: diversity in gene-expression by alternative RNA processing. Annu. Rev. Biochem., 55, 1091–1117
115. Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P. and Rokas, A. (2009) Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol. Biol. Evol., 26, 2731–2744
116. Gruenheit, N., Deusch, O., Esser, C., Becker, M., Voelckel, C. and Lockhart, P. (2012) Cutoffs and *k*-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics, 13, 92
117. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., *et al.* (2013) *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat. Protoc., 8, 1494–1512
118. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol., 28, 511–515
119. Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111
120. Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C., Pugh, T. J., *et al.* (2010) Alternative expression analysis by RNA sequencing. Nat. Methods, 7, 843–847
121. Melicher, D., Torson, A., Dworkin, I. and Bowsher, J. (2014) A pipeline for the *de novo* assembly of the *Themira biloba* (Sepsidae: Diptera) transcriptome using a multiple *k*-mer length approach. BMC Genomics, 15, 188
122. Francis, W. R., Christianson, L. M., Kiko, R., Powers, M. L., Shaner, N. C. and Haddock, S. H. D. (2013) A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly. BMC Genomics, 14, 167
123. Kumar, S. and Blaxter, M. (2010) Comparing *de novo* assemblers for 454 transcriptome data. BMC Genomics, 11, 571
124. Ren, X., Liu, T., Dong, J., Sun, L., Yang, J., Zhu, Y. and Jin, Q. (2012) Evaluating *de bruijn* graph assemblers on 454 transcriptomic data. PLoS One, 7, e51188
125. O'Neil, S. and Emrich, S. (2013) Assessing *de novo* transcriptome assembly metrics for consistency and utility. BMC Genomics, 14, 465
126. Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., *et al.* (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res., 22, 557–567
127. Mundry, M., Bornberg-Bauer, E., Sammeth, M. and Feulner, P. G. D. (2012) Evaluating characteristics of *de novo* assembly software on 454 transcriptome data: a simulation approach. PLoS One, 7, e31410
128. Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R. and Dewey, C. N. (2014) Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. Genome Biol., 15, 553
129. Clark, S. C., Egan, R., Frazier, P. I. and Wang, Z. (2013) ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics, 29, 435–443
130. Henschel, R., Lieber, M., Wu, L.-S., Nista, P. M., Haas, B. J. and LeDuc, R. D. (2012). Trinity RNA-Seq assembler performance optimization. In Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond. 1–8
131. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., *et al.*(2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics, 30, 1660–1666
132. Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C. L. and Huang, X. (2015) Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data. Genome Biol., 16, 30
133. Li, Y., Hu, Y., Bolund, L. and Wang, J. (2010) State of the art *de novo* assembly of human genomes from massively parallel sequencing data. Hum. Genomics, 4, 271–277
134. Zhou, S., Liao, R. and Guan, J. (2013) When cloud computing meets bioinformatics: a review. J. Bioinform. Comput. Biol., 11, 1330002
135. Taylor, R. (2010) An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics, 11, S1
136. Check Hayden, E. (2009) Genome sequencing: the third generation. Nature, 457, 768–769
137. Schadt, E. E., Turner, S. and Kasarskis, A. (2010) A window into third-generation sequencing. Hum. Mol. Genet., 19, R227–R240
138. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.*(2009) Real-time DNA sequencing from single polymerase molecules. Science, 323, 133–138

139. Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., *et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat. Biotechnol., 30, 693–700
140. English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., *et al.* (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One, 7, e47768
141. Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A., *et al.* (2013) An evaluation of the PacBio RS platform for sequencing and *de novo* assembly of a chloroplast genome. BMC Genomics, 14, 1–12

