Frontiers in Zoology

REVIEW

Open Access

De novo assembly of transcriptomes and differential gene expression analysis using short-read data from emerging model organisms – a brief guide

Daniel J. Jackson[1*†], Nicolas Cerveau[1†] and Nico Posnien[2*†]

Abstract

Many questions in biology benefit greatly from the use of a variety of model systems. High-throughput sequencing methods have been a triumph in the democratization of diverse model systems. They allow for the economical sequencing of an entire genome or transcriptome of interest, and with technical variations can even provide insight into genome organization and the expression and regulation of genes. The analysis and biological interpretation of such large datasets can present significant challenges that depend on the 'scientific status' of the model system. While high-quality genome and transcriptome references are readily available for well-established model systems, the establishment of such references for an emerging model system often requires extensive resources such as finances, expertise and computation capabilities. The de novo assembly of a transcriptome represents an excellent entry point for genetic and molecular studies in emerging model systems as it can efficiently assess gene content while also serving as a reference for differential gene expression studies. However, the process of de novo transcriptome assembly is non-trivial, and as a rule must be empirically optimized for every dataset. For the researcher working with an emerging model system, and with little to no experience with assembling and quantifying short-read data from the Illumina platform, these processes can be daunting. In this guide we outline the major challenges faced when establishing a reference transcriptome de novo and we provide advice on how to approach such an endeavor. We describe the major experimental and bioinformatic steps, provide some broad recommendations and cautions for the newcomer to de novo transcriptome assembly and differential gene expression analyses. Moreover, we provide an initial selection of tools that can assist in the journey from raw short-read data to assembled transcriptome and lists of differentially expressed genes.

Keywords Transcriptome assembly, De novo assembly, RNA-seq, Short reads, Emerging model system, Genome, Annotation, Differential gene expression

[†]Daniel J. Jackson, Nicolas Cerveau and Nico Posnien contributed equally to this work.

*Correspondence:
Daniel J. Jackson
djackso@gwdg.de
Nico Posnien
nposnie@gwdg.de
Full list of author information is available at the end of the article

Introduction

A major goal in Biology is to understand the processes that underlie the phenotypic variation observed in nature. Because information about organismal phenotypes, such as appearance and function, are stored in the genome, holistic approaches have greatly benefitted from advances in sequencing technologies which provide the opportunity to meaningfully integrate molecular, genetic and morphological data. Accordingly, genomic databases have grown rapidly since the completion of the first animal genome in 1998 [1]. These genomic resources have revolutionized multiple biological disciplines and have generated unprecedented insights into disparate phenomena. For instance, the reconstruction of phylogenetic relationships employing genomic information (i.e. phylogenomics) has resolved previously undetermined relationships and radically changed our view of the tree of life [2]. Cost-efficient sequencing of organismal communities directly from environmental samples (i.e. metagenomics) have provided novel opportunities for the description and monitoring of biodiversity (e.g. [3]), and the analyses of population dynamics, patterns of adaptation and the demographic history of organisms profit from extensive genome re-sequencing (e.g. [4]) (Fig. 1Ai). As a genome contains the information for all protein coding genes, as well as non-coding regulatory sequences (Fig. 1Ai), studies focused on chromatin accessibility (e.g. ATAC-seq, [5]), transcription factor binding sites, histone modifications (e.g. ChIP-seq, [6]) and epigenetic DNA modifications (e.g. bisulfite sequencing for methylation studies, [7]) require a well-assembled genome sequence. Moreover, studies assessing mobile element insertions (i.e. transposable elements; [8]) and the identification of neutral sites for population genomics inferences (e.g. [9]) often require reference genomes.

Despite the exciting opportunities that genome-scale approaches provide, the generation of a complete genome sequence requires considerable financial and computational resources, as well as different sets of expertise to assemble, annotate and analyze the final dataset.

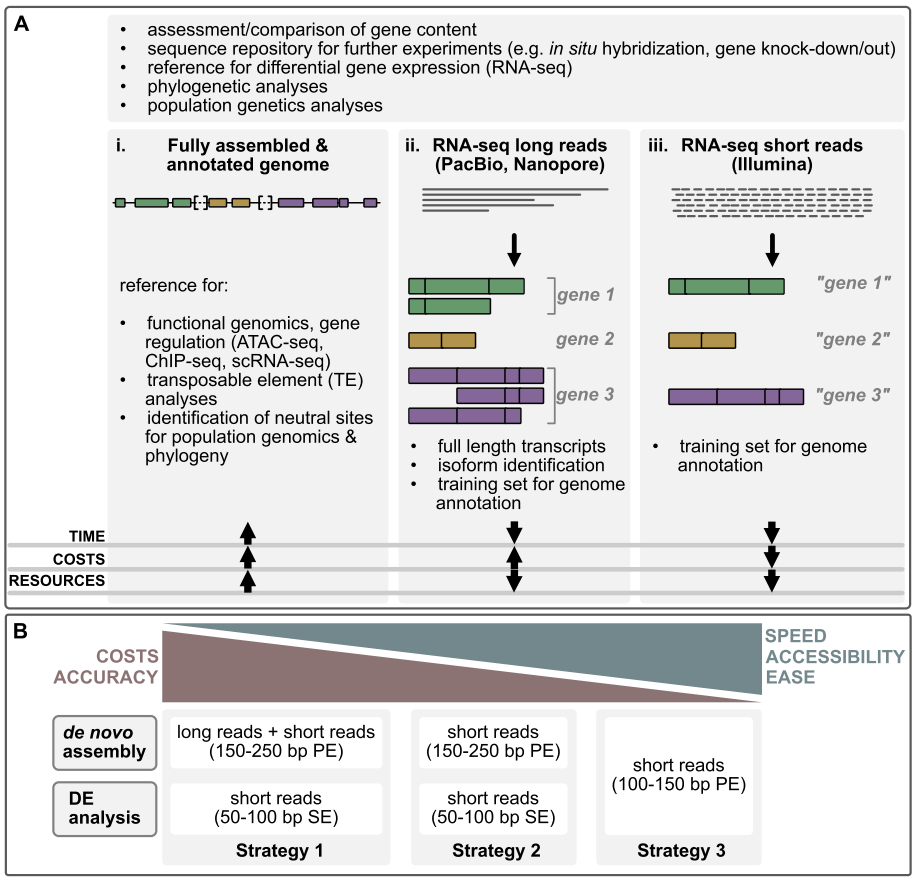

Fig. 1 Overview of applications for genome and transcriptome resources. **A** Overview of major applications facilitated by "omics" data. The upper box summarizes applications provided by all methods outlined in Ai-iii. (Ai) Chromosome-level genome with high-quality gene annotation. (Aii) Transcriptome assembly based on long reads (e.g. PacBio Iso-Seq). (Aiii) Transcriptome assembly based on short reads (Illumina, 100–250 bp). **B** Recommendations for experimental design decisions

Genome assemblies may be challenged by large genome size [10] and high content of repetitive sequences (for example typical for molluscan genomes; [11]). Accordingly, genome assembly approaches often rely on the integration of Illumina short reads and long-read data for the generation of continuous fragments (contigs), which are combined into larger fragments (i.e. scaffolds), preferentially entire chromosomes, by optical mapping approaches or genome wide three-dimensional genomic contact information (i.e. HiC) [12–16]. The success of a genome annotation effort can also strongly depend on the structure of the genome, such as the ratio of coding to non-coding sequences and the availability of gene expression data to train gene model prediction algorithms [17]. Therefore, assembling and annotating a genome de novo is often not achievable for an individual research group, but requires input from different labs usually in larger consortia. Moreover, despite ever-growing genomic databases it remains challenging to establish causative links between a genome sequence and the organismal phenotype it produces [18]. This is in part because the sequence information stored in the genome is transferred to other molecules (e.g. proteins or regulatory RNAs) which only then affect cell and organ morphology and function. For example, the ultimate action of a functional protein is achieved and regulated by multiple molecular processes, such as transcription, translation and post-translational modifications [19]. While sequencing-based methods to study and quantify these regulatory steps have been established in recent years [20], many such functional genomic approaches remain restricted to well-established model organisms. However, as many basic aspects can be well analyzed without a full genomic sequence, high-throughput sequencing of transcribed genes (i.e. transcriptome sequencing) by RNA-seq is broadly used to assess the content and abundance of transcripts in any organism, tissue and more recently in individual cells [21–24] (Fig. 1Aii—iii). Differential gene expression (DGE) studies based on RNA-seq data can be used to quantify differences in transcript abundance across multiple natural (e.g. between species, populations or between developmental stages or habitats) and experimental (e.g. between treatments or genetic modifications) conditions. Therefore, RNA-seq has become a standard approach in all domains of biology to better understand how genetic information defines organismal phenotypes (e.g. [25, 26]).

Gene expression studies based on RNA-seq are easily performed for model organisms with a reference genome and a high-quality annotation (Fig. 2A). The annotation of a well assembled genome is often based on an initial round of automated annotation, followed by iterative rounds of manual curation by a dedicated community [27–30]. Despite the efforts of large consortia, such as the Darwin Tree of Life [31] and the European Reference Genome Atlas (ERGA) projects [32], to establish genomic resources for non-model systems, the annotation of these genomes often does not reach the high quality of model organism genomes. Accordingly, DGE analyses derived from poorly annotated genomes in these non-model systems may not be the ideal approach; an incorrect or incomplete genome annotation has a major

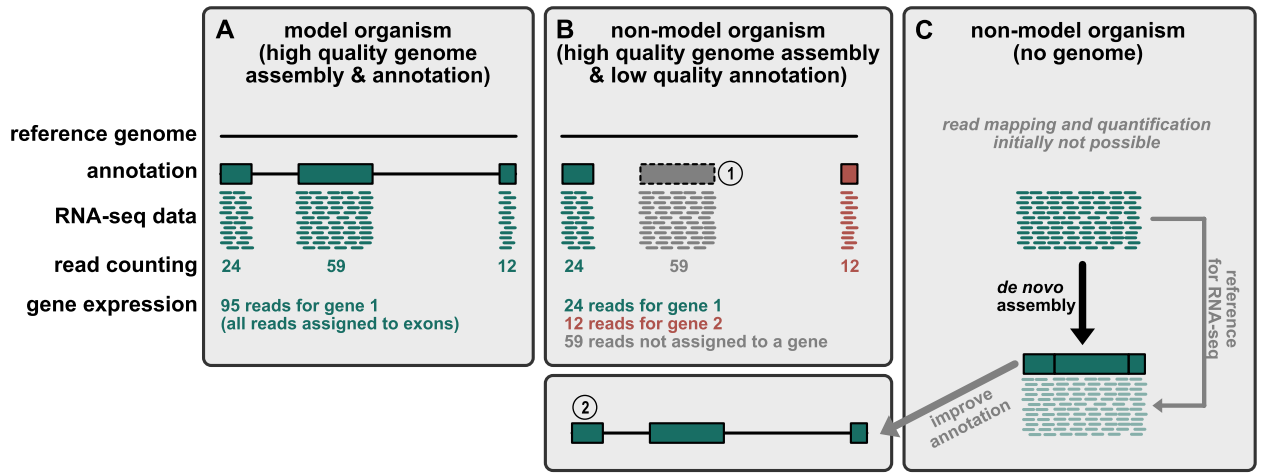

Fig. 2 Overview of current challenges during RNA-seq analysis. **A** RNA-seq analysis in a model organism with a high-quality genome annotation. **B** RNA-seq analysis of a non-model organism with a genome reference, but incomplete annotation. 1 – After mapping RNA-seq reads to the reference genome, the additional exon can be annotated, while no information about the connection to the other two exons is available. 2 – Mapping the transcript obtained by de novo assembly of the RNA-seq data onto the genome allows annotating the full gene model. **C** If no genome reference is available, RNA-seq data cannot be readily used to quantify gene expression. A de novo assembly of the RNA-seq is required to reconstruct transcripts, which can serve as mapping references for RNA-seq data and to improve existing genome annotations

impact on the analysis of gene expression data (Fig. 2B) ([33], and unpublished observation). Moreover, for many emerging model organisms a well assembled and annotated reference genome often does not exist (Fig. 2C). As RNA-seq captures the transcribed regions of the genome, such as protein coding genes and regulatory RNA molecules [21, 34], the establishment of de novo transcriptome resources (i.e. a de novo transcriptome assembly) has become a powerful and time and resource efficient tool to study the gene content of an organism. Transcriptomic resources have been successfully employed to estimate genetic divergence in population genetic studies [35, 36], to conduct meta-transcriptomic surveys of environmental samples [37, 38] and to reconstruct phylogenetic relationships [39–44]. They can also facilitate the efficient identification and molecular isolation of specific gene sequences for further studies, such as in situ hybridization or loss of function experiments using RNA interference (RNAi) [45] (Fig. 1A). Moreover, de novo transcriptome assemblies can serve as a reference for DGE studies (Fig. 2C), and such resources can also be employed to improve genome annotations (Fig. 2B, 1&2). As many emerging model organisms are studied by small research communities, the economic cost and technical challenges of generating a high-quality genome assembly and annotation are not realistically achievable within the timeframe of a typical project (3–4 years), and therefore many of these methods are restricted to well-established and/or well-funded model systems. Therefore, the de novo assembly of transcriptomes remains a valuable approach to study the genetic and molecular underpinnings of phenotypic traits and to improve genome annotation.

Researchers that wish to add a genetic component to their research questions may be overwhelmed by the many decisions that need to be taken to obtain the best possible data. For instance, 'which next generation sequencing technology should I choose, short reads or long reads? How many replicates do I need for a DGE study? What assembly software should I use and how do I assess the quality of any output?'. In this article we provide an overview of typical workflows for de novo transcriptome assemblies and subsequent DGE analyses. We also highlight some of the challenges associated with study design, sample preparation, de novo assembly, DGE analysis and evaluating the outputs of these exercises. In general, we refrain from making dogmatic recommendations concerning the use of particular software packages and encourage the user to explore the impact that different approaches and tools have on their own data. To this end, we recommend adopting a systematic and organized approach to exploring any RNA-seq dataset so that objective comparisons can readily be made and interpreted. We also refrain from making any cost comparisons or explicit budget recommendations because the available technologies and their corresponding shortcomings, strengths and 'cost-per-base' prices are evolving extremely rapidly. Rather we encourage the researcher to inform themselves of current market prices and to explicitly weigh the technical advantages and disadvantages of each technology according to their research needs. Importantly, we are primarily concerned with emerging model organisms for which limited, or no genome or transcriptome resources exist, and we assume this category includes organisms for which there also exists little functional genetic information, such as spatial gene expression or gene function data.

Generating and using RNA-seq data

The generation of RNA-seq data is a cost and time efficient entry point for genetic and molecular analyses in an emerging model system [26]. A typical RNA-seq experiment starts with the isolation of total RNA from the organism, tissue or developmental stage of interest. The RNA molecules are fragmented, reverse transcribed into complementary DNA (cDNA) and sequencing adapters are incorporated. By selecting polyadenylated molecules during library preparation, the fragments can be enriched for messenger RNAs (mRNAs). This depletes many non-coding RNA molecules and the majority of ribosomal RNAs from subsequent analyses. If regulatory RNAs are intended to be studied, it is important to retain the full complement of RNA molecules using random priming during the library preparation (Fig. 3) as many of these molecules are not polyadenylated. The cDNA library is amplified by PCR, and these libraries are then subjected to Illumina short-read sequencing resulting in 50–250 bp single- (SE) or paired-end (PE) reads. The quality of these short reads is assessed, and high-quality reads are assembled to reconstruct the original transcripts (i.e. de novo transcriptome assembly) (Fig. 3). This data can then be used to assess the complement of transcripts expressed in a certain tissue or developmental stage. If RNA extracted from whole bodies or multiple organs at different life stages is used for the de novo assembly, the transcriptome will ideally represent a comprehensive resource for the identification of all transcribed gene sequences, and it may serve as a reference to quantitatively estimate transcript abundance (i.e. DGE) across biological or experimental conditions.

Readers will likely be aware that long-read sequencing technologies, such as PacBio and Oxford Nanopore are an alternative to assembling Illumina-derived short reads [46] to generate a de novo transcriptome. As these long-read technologies sequence the entire transcript from the five-prime untranslated region (5'-UTR)

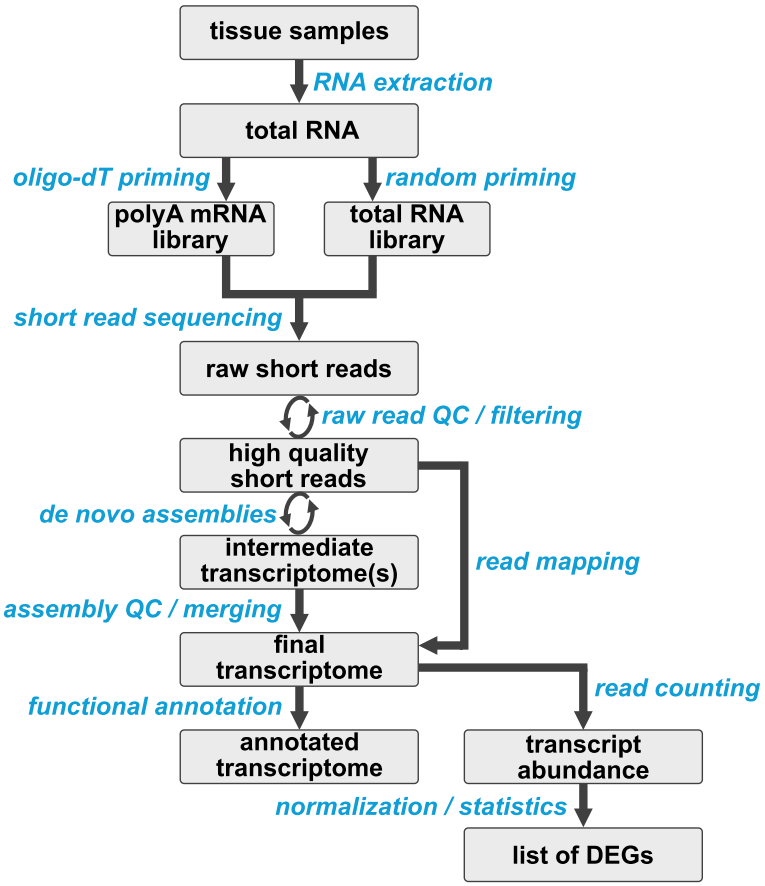

Fig. 3 General outline of major steps for de novo assembly and DGE analyses

to the polyadenylation tail, they are indispensable when the accurate identification and discrimination of splice variants [47] is the goal (Fig. 1Aii). Time-consuming and computationally intense de novo assembly steps are often not required when working with long reads. However, this data is commonly "polished" with short-read data to improve sequence accuracy. It is also important to note that unfragmented, high-quality RNA is needed for efficient long-read sequencing, which may not always be feasible if field samples or preserved material are all that is available. In addition, estimates of gene expression (whether absolute or relative) are almost exclusively based on high-depth short-read data [48, 49]. This is in part because every transcript should be represented by an individual long read and current long-read technologies generally do not generate this sequence depth to accurately infer global levels of gene expression [50, 51]. Gene expression quantification is further hampered because the proper assignment of long reads to particular isoforms tends to be complicated for emerging model systems without reference genomes [50]. Therefore, current long-read based transcriptomics require higher financial and computational resources than short reads and the choice of data to be generated should thus be aligned with the project goals and funds available (Fig. 1Aii, B).

If funding, computational resources, expertise and fresh material are not limiting, we recommend generating a transcriptomic reference based on a combination of long and short reads from the entire organism and/or multiple tissues and from different life stages to obtain a comprehensive collection of highly accurate, full-length transcripts. Subsequent gene expression analyses should then be derived from short-read data (50 – 100 bp reads, SE mode). If resources are limited, a transcriptomic reference can be generated by the de novo assembly of short reads exclusively. Preferentially, longer short reads (150 – 250 bp) in PE mode are used for the assembly of the reference, and shorter reads (50 – 100 bp) in SE mode are generated for subsequent DGE analyses. The most cost-efficient approach is to use the same short reads for the de novo assembly and for subsequent gene expression analysis. In this case, we recommend generating 100–250 bp reads in PE mode (Fig. 1B). We focus here on typical RNA-seq workflows that aim at generating de

novo transcriptome references from short reads which can subsequently serve as reference for differential gene expression (DGE) analyses.

Generating a transcriptome assembly de novo

Planning, sample preparation and short-read sequencing

As gene expression is highly context dependent, the completeness of a transcriptome strongly depends on the biological input material. The planned downstream applications must therefore be carefully considered at this stage of the project. For instance, if the transcriptome should serve as reference for subsequent analyses of embryonic gene expression, the de novo assembly should be based on RNA extracted from preferentially multiple embryonic tissues. If an analysis of total gene content is the goal, for example to support a genome annotation, one should obtain RNA-seq data from as many tissues and life stages as possible. In principle, to generate a de novo transcriptome assembly experimental replication of the tissues/stages is not necessary. If resources are available, we would rather recommend investing in sequencing more samples from different tissues and/or life stages to capture more transcript diversity. However, if the same RNA-seq reads used to assemble a transcriptome de novo will serve for DGE studies, proper biological replication (see below) is a prerequisite. It should also be kept in mind that biological replicates derived from different individuals (especially from a highly heterozygous population) may generate assembly artefacts and significantly increase computing time [52]. If the degree of heterozygosity is high or unknown, we recommend to sequence from as few individuals as possible or try to use individuals derived from inbred lines.

Despite the availability of impressively low-input RNA-seq options, we strongly encourage the isolation of adequate amounts of high-quality total RNA (a minimum of 0.3 µg total RNA at a concentration of 30—100 ng/µl). In our experience there is no advantage for the researcher to prepare the sequencing libraries themselves prior to submitting them to an experienced sequencing center because critical steps can be performed for a better cost and usually to a higher standard by dedicated technicians or pipetting robots. Moreover, most sequencing centers will routinely perform quality control (QC) on all RNA samples submitted by clients and will identify problems associated with RNA extracted from unusual tissues, such as tissues with high enzymatic activity or inhibitors of ligases and/or polymerases. For instance, highly fragmented RNA can be easily identified by assessing the fragment size distribution (e.g. fragment analyzer). It is important to note that many sequencing centers tend to handle RNA extracted from predominantly mammalian or well-established model organisms. Accordingly, a typical first check for RNA integrity that employs estimates of the size and relative abundance of 28 s and 18 s ribosomal RNA (rRNA) fragments [53] 'fails' for many emerging model systems as the rRNA fragments tend to be of different sizes in different organisms. If the RNA-seq data will be used in downstream DGE analyses (Figs. 2C, and 3), close consultation with the sequencing center will ensure the sensible pooling of barcoded samples to avoid any flow-cell specific technical biases [54]. In general, a clear line of communication and regular check-ins with the sequencing center are therefore essential to avoid preventable missteps.

Once a high-quality RNA-seq library is prepared, several important decisions concerning sequencing, such as read length, type (i.e. PE vs. SE) and the number of reads (i.e. sequencing depth) need to be made. Generally, longer PE reads (up to 500 bp comprised of 2 × 250 bp PE reads) give more complete and less fragmented de novo assembly results [55]. Deciding on the number of reads to sequence should be informed predominantly by the complexity of the transcriptome (i.e. the diversity of genes expressed) and by the desired sensitivity of the analysis, as greater depth is required to detect rare transcripts. As this information is usually not available for emerging model systems, one can only make educated estimates. A good starting point is to consult RNA-seq experiments performed in close relatives of the species of interest. It may be tempting to follow the "a lot helps a lot" rule, but simulation studies have demonstrated negative effects on the quality of the de novo assembly (i.e. the identification of new transcripts) due to saturation. For instance, in the fruit fly *Drosophila melanogaster* (genome size: ~ 180 Mbp), saturation effects have been observed when about 2 million reads were used for the assembly [55]. Similarly, a de novo transcriptome assembly for the common house spider *Parasteatoda tepidariorum* (genome size: ~ 1,411 Mbp) showed that transcript discovery saturated when 99 to 132 million reads of a full set of 330 million PE reads were used. The addition of more reads increased the number of assembled transcripts, but these tended to be short and not informative [56]. Therefore, instead of sequencing deeper (i.e. more reads from the same sample), we recommend investing resources in either sequencing from more diverse samples (for example more tissues or more developmental stages) to generate a more comprehensive reference assembly, or more biological replicates for DGE analyses. Generally, the number of reads depends on the transcriptome complexity and can range between 2–5 million reads for insects with small genomes (e.g. *Drosophila melanogaster*) to 100 to 120 million reads for chelicerates with rather large genomes (e.g. *Parasteatoda tepidariorum*).

Pre-processing and quality control of raw read data

Once a dataset of short reads in FASTQ format have been acquired several quality checks must be performed. While a variety of approaches exist for the targeted sequencing of different molecules (e.g. small RNAs, 16S and other PCR amplicons), we assume here that the desired short-read data represents mRNA transcripts, and therefore that the sequencing center has performed a polyadenylation selection step to reduce rRNA abundance during library preparation. As typically multiple samples are pooled for each Illumina run, all reads must be assigned back to their parent sample via the unique barcodes present in the sequencing adapters (i.e. de-multiplexing). The sequencing center will remove these barcodes, indices and other non-native sequences prior to providing them to the researcher in a FASTQ file. While in theory this means the researcher should be able to proceed directly to performing a de novo assembly, we strongly recommend assessing the raw data for read/base quality, read length, read number, GC composition and adapter contamination (see Table 1 for a set of tools that can perform these functions). Moreover, we strongly recommend to check for contamination either from unwanted organisms [57] or by rRNA [58, 59]. Contamination of mRNA-seq data by rRNA is often not investigated, however we have experience with samples from otherwise equivalently RNA-extracted replicates sequenced in the same run suffering from extremely high (85%) proportions of rRNA. Based on the results of these raw-read quality metrics, reads should be trimmed and low-quality reads should be removed (Table 1). It may take multiple iterations of trimming and filtering before a set of parameters is identified appropriate for the data being processed. Once the raw data has passed these quality controls, a de novo assembly can be performed.

The issue of read normalization prior to performing a de novo assembly should also be mentioned at this stage. It has been reported that reducing the redundancy of reads originating from the same original transcript prior to assembly not only reduces the computational complexity of the problem (i.e. assembly time), but also the quality of the assembly can benefit from such measures [104]. Indeed, some packages include a read normalization step by default prior to any assembly action (for example Trinity; [105]) and different conceptional bases for read normalization have been used [104, 106]. As highlighted below for other parameters, we would encourage to empirically assess the effect of including/omitting read normalization on the final assembly (see below). Importantly, this normalization should not be confused with the mandatory normalization of data within and across samples during DGE analyses (see below).

Assembly

Given the abundance of options, the novice (and expert) may have difficulty in deciding which assembly package to employ and which will give the "best" assembly. While there will rarely be a clear answer to this question (especially a priori), the QC steps that assess the quality of a de novo assembly (see below) will help guide the selection of the appropriate package. In general, we recommend using popular and highly cited short-read assemblers that have a respectable half-life in the literature and remain available and supported (Table 1) [107].

These short-read assemblers typically employ methods based on the construction of de Bruijn graphs which represent pathways of sequence-overlap through the raw data. Each graph represents a transcript or a group of similar transcripts [66]. During de Bruijn graph construction and the resulting assembly each read is divided into smaller fragments of size k, which must be shorter than the read length. These substrings of the reads, so called k-mers, are aligned against the collection of k-mers that exist in the raw data with an overlap of the length k-1. This iterative process therefore extends the aligned region one nucleotide at a time and continues until no additional overlap is found. Therefore, k-mer length has a strong influence on the final assembly and must be carefully chosen [108–110]. Excessively short k-mers will generate highly fragmented and duplicated assemblies but will detect rare transcripts, while long k-mers can miss rare transcripts but will resolve repetitive or error-prone regions. Transcriptome assemblers can be distinguished based on whether they employ a single k-mer value during assembly, or multiple k-mer values (Table 1).

A study based on 10 assemblers and 9 different datasets concluded that it is difficult to identify the 'best assembler' as their performance is intimately linked to the input data supplied [111]. The literature is replete with systematic comparisons of assemblers applied to the same input data that produce divergent outputs [68, 78, 112, 113], and our own experience reflects this. Therefore, we encourage researchers new to assembling transcriptome data to adopt a philosophy of performing multiple assemblies with a diversity of tools and settings, and systematically comparing their outputs (see below). In line with this approach, the 'next generation' of de novo transcriptome assemblers are aggregate tools that take the output from a variety of independent assembly packages, and either assess their individual completeness using a variety of metrics, or merge and then de-duplicate them according to similarity thresholds (Table 1).

Table 1 A list of software packages available for the various steps required to assemble a transcriptome de novo

de novo transcriptome assembly		
Preprocessing and quality control of raw read data		
FastQC	[60]	Read quality check
RNA-QC-chain	[61]	Sequencing quality and contamination trimming
RSeqQC	[62]	Read quality and distribution statistics
MultiQC	[63]	Summarization and visualization tool
HTQC	[64]	Read filtering, QC and visualizing
SortMeRNA	[58]	Filtering of rRNAs from meta-transcriptomic data
BBDuk (BBtools)	[59]	Read trimming, rRNA removal, filtering, error correction and much more
Trimmomatic	[65]	Read trimming, rRNA removal, filtering
Assembly* de novo *single k-mer		
Trinity	[66]	Each single k-mer assembler has its unique set of features and we encourage the user to systematically compare the outputs of different packages
SOAPdenov2	[67]	
IDBA-tran	[68]	
Trans-Abyss	[69]	
Assembly* de novo *multiple k-mer		
rnaSPADES	[70]	Each multiple k-mer assembler has its unique set of features and we encourage the user to systematically compare the outputs of different packages
SKESA	[71]	
Velvet	[72]	
Transcriptome aggregation tools		
TransPi	[73]	The principle of merging/aggregating the outputs of multiple de novo assemblies and then reducing redundancy lies at the core of these tools. The details of how they achieve this can generate divergent outputs which should be systematically compared
Cerveau and Jackson	[74]	
Nakasugi et al	[75]	
Mikado	[76]	
ConSemble	[77]	
Quality assessment of* de novo *transcriptome assembly		
Detonate	[78]	Model based score to evaluate transcriptome quality
TransRate	[79]	Quality assessment detecting chimeras, structural and sequencing errors
rnaQUAST	[80]	Based on reference genome and database
BUSCO	[81]	Based on single orthologue database
DOGMA	[82]	Based on protein domain database
Bellerophon	[83]	Result concatenation tool
Functional annotation		
Transdecoder	[84]	predicts CDS for each transcript
esl-translate	[85]	part of the HMMER package; reports all potential ORFs for each transcript
Trinotate	[86, 87]	Comprehensive functional annotation pipeline, generates easily accessible database with all annotation results
dammit	[88, 89]	Comprehensive functional annotation pipeline using reciprocal homology assignment
EnTAP	[90]	Integration of multiple functional annotations
Differential gene expression analyses based on transcriptome references		
Read mapping to reference transcriptome		
STAR	[91, 92]	Read mapper (traditional read alignment)
HISAT2	[93]	Read mapper (traditional read alignment)
Kallisto	[94]	Pseudo-aligner
Salmon	[95]	Quasi-mapper
RapMap	[96]	Quasi-mapper
Transcript and mapping result grouping		
Corset	[97]	Tools to reduce the diversity of reference transcripts inevitably generated from a de novo assembly – employed when performing read mapping for DGE
Grouper	[98]	
Compacta	[99]	

Table 1 (continued)

de novo transcriptome assembly		
Statistical analyses/differential gene expression		
DEseq2	[100, 101]	Two of the most employed statistical tools for DGE analyses. They each employ different data normalization concepts and can therefore generate divergent results
edgeR	[102, 103]	

Quality assessment of de novo transcriptome assemblies

Evaluating the quality of a de novo assembled transcriptome is a key step that should precede its use in any downstream application. While the quality of an assembly can be best assessed based on prior genome or transcriptome data from the target organism (i.e. reference-based quality assessment), we focus our attention on reference-free tools because we are primarily concerned with scenarios focused on emerging model organisms. One of the first metrics of an assembly that should receive attention is the total number of transcripts (also often referred to as contigs). This value will be related to the total number of genes present in the genome of the organism and can therefore be predicted to be constrained to a reasonable minimum. For instance, metazoan models with well annotated genomes contain 10-22 k protein coding genes [114–116]. However, this guideline cannot account for mechanisms such as gene duplication and the presence of splice variants which will significantly inflate these numbers. Therefore, in addition to the number of transcripts, their length distribution should be considered as any transcriptome assembly will invariably include spurious short contigs that should be excluded from further analyses. In some reports a majority of transcripts are short [117] and can be removed without affecting downstream analyses. While it would be unwise to recommend a concrete threshold, many researchers ignore transcripts with lengths shorter than 200 bp (default threshold minimum transcript size for Trinity) and even up to 400 bp [118]. Any length threshold should be carefully chosen, for instance by taking the original read length and insert size (i.e. the distance between PE reads) into account. Generally, contigs shorter than the average read length should be excluded from the reference. It is important to note that frequently employed contiguity metrics, such as the N50 metric, which were developed for assessing the contiguity of genome assemblies [119] are uninformative in the context of transcriptomes because transcripts vary greatly in their lengths.

Another important measure of assembly quality is the average and/or total number of reads mapping to a contig as this metric allows identifying transcripts which are poorly supported by the original RNA-seq reads [80, 120]. By mapping reads originally used for the assembly back to the assembled transcriptome one can evaluate the proportion of reads used to generate the assembly and high-confidence transcripts supported by correctly paired reads can be identified and retained [121]. In addition, the distribution of reads across transcripts can be used to detect chimeras and other assembly artefacts [122]. Tools such as Detonate and TransRate (Table 1) can be used to map reads and generate a score that reflects the overall quality of the transcriptome by detecting chimeras, assembly, and sequencing errors. The ExN50 metric [123] represents a combination of transcript abundance estimation and transcript length to assess and compare the quality of transcriptome assemblies.

Another informative, and highly cited tool to assess assembly quality is BUSCO (benchmarking universal single-copy orthologs) [81]. BUSCO compares a transcriptome (or genome) against a curated database of single-copy orthologous genes. The concept is that if 90% of the BUSCO genes are in an assembly, then that assembly by extension is likely to be 90% complete for all genes. The tool provides various databases on different phylogenetic levels and the user selects the most appropriate lineage, such as all eukaryotes, plants, bacteria, or specific groups, such as spiders or molluscs. A BUSCO output will therefore return a meaningful estimate of the completeness, duplication, fragmentation, and lack of transcripts of genes that should be present in the assembly. Similarly, DOGMA assesses the completeness of a transcriptome by surveying the assembly for conserved protein domains [82] rather than complete genes.

With the growing number of tools and approaches available to assess the quality of transcriptome assemblies, aggregate packages such as Bellerophon [83] have been developed and aim to incorporate multiple lines of evidence to identify an optimal assembly. In general, we recommend to systematically apply the outlined quality checks for multiple assemblies to facilitate relative comparisons.

Functional annotation of a de novo transcriptome assembly

Once an optimal de novo assembly has been generated the next step is to predict putative functions of the proteins coded by the assembled transcripts. Such a functional annotation is commonly achieved by

identifying similar (ideally homologous) sequences in protein or nucleotide databases. Any functional information assigned to the sequence in that database can be inferred to apply to the transcript in the de novo transcriptome assembly.

Typical protein or nucleotide databases are the UniProt (many proteins, non-curated)/SwissProt (fewer proteins, highly curated) [124] protein databases, as well as the most comprehensive, but non-curated NCBI databases for protein (NR) and nucleotide (NT) sequences [125]. The search for homologous sequences in such databases is often a time-consuming computational task because it mostly relies on tools employing the Basic local alignment search tool (BLAST) logic [126]. As protein sequences tend to be more conserved than nucleotide sequences, it is generally computationally less demanding to perform the homology searches in protein databases using translated sequences as input. Therefore, a typical first step of a functional annotation pipeline is to identify the coding sequence (CDS) and/or the longest open reading frames (ORFs) for each transcript in the de novo assembly (Table 1). Once sequence homologs are identified, the functional information associated to them is assessed employing functional databases. For instance, the InterPro database contains information about protein domains and families [127] and the gene ontology (GO) knowledgebase is a highly curated collection of functional information (e.g. biological function, molecular functions) for genes and gene products derived predominantly from model organisms [128]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database allows placing de novo transcripts into the context of biological networks and pathways [129, 130]. To streamline the major computational steps required for a functional transcriptome annotation, comprehensive pipelines, such as Trinotate and dammit (Table 1) have been established. The output of such pipelines are easily accessible files, such as Excel spreadsheets or a Sqlite database, as well as more advanced annotation files for further bioinformatic analyses. Potential problems during functional annotation caused by the typically high number of transcripts are for instance addressed by the Eukaryotic Non-Model Transcriptome Annotation Pipeline (EnTAP) (Table 1), which filters transcripts based on expression levels and coding potential and subsequently combines multiple functional annotations (i.e. based on annotation pipelines) into one high-confidence annotation.

The quality of the functional annotation strongly depends on the ability to identify clear homologous sequences, as well as on the quality of the functional information stored in these databases. If the de novo transcriptome is derived from an organism that is related to a model organism, sequence homology can often be easily established, and putative functions can be assigned for many transcripts. However, if relatives of the study organism are not well represented in typical databases, it may be challenging to functionally annotate the majority of transcripts. Hence, depending on the phylogenetic relation to organisms for which high-quality functional information is available in databases, the functional annotation should follow slightly different routes. For instance, while BLAST searches in highly curated protein databases are most time- and resources-efficient, it may be advantageous to base the homology assignment on larger protein databases or even more comprehensive nucleotide databases if many transcripts remain un-annotated after a first annotation round. It is also important to note that it may not be possible to assign putative functions to lineage-restricted or novel, as well as fast-evolving genes in the de novo assembly as they tend to lack homologous sequences in the respective databases. In these cases, putative functions may only be deduced from the annotation of short protein domains rather than comprehensive functional information. A final consideration is related to the purpose of the de novo assembly. If the assembly is intended to provide a broad overview of the transcripts expressed in an organism, a full functional annotation is desirable. However, if the major goal is to identify expression differences between conditions, it could be sufficient to functionally annotate only those transcripts that show significant differences in gene expression. It must be kept in mind that various scenarios of protein evolution, for example domain shuffling within and between proteins, may complicate the interpretation of superficial annotation efforts.

DGE analyses based on de novo transcriptome assemblies

To establish causal relationships between the genome and biological phenomena, gene expression studies most of the time represent an easy entry point to identify lists of candidate genes that could be responsible for a certain phenotypic outcome. The analysis of differential gene expression from RNA-seq data is therefore one of the most employed methods in molecular biology nowadays, and many best-practice guidelines are available [21, 22, 131–134]. Here, we briefly summarize the major analysis steps, standard recommendations and we emphasize special requirements when de novo assemblies are used as the reference.

Experimental design, sequencing and pre-processing of RNA-seq reads

Careful planning is a critical phase of any DGE experiment and will save significant time and financial costs in the later phases of a project. The most common first decision relates to the number of replicates to generate. A

study specifically aimed at answering this question in the well-established unicellular yeast model system *Saccharomyces cerevisiae* found that 20 high quality replicates in a simple 2-way comparison were required to detect > 85% of the significantly differentially expressed genes (regardless of fold-change). With a commonly employed experimental design of 3 replicates only 20–40% of all significantly differentially expressed genes could be detected [135]. From such studies a general recommendation of 6 biological replicates for all experimental conditions emerged, while 12 replicates are recommended when the identification of the majority of differentially expressed genes is required [135].

Beyond replication, factors related to the Illumina sequencing platform and the generated reads should be considered at the project planning stage. A general consensus emerges from the literature that recommends stranded sequencing in order to maximize the disambiguation of potential read placement (i.e. caused by overlapping gene bodies on different strands) [136], read lengths of 50–100 bp [137] and 25–100 million reads per sample/replicate (depending on the genome/transcriptome size) [138]. Unless splice-variant quantification is paramount [137], SE reads suffer no disadvantage to PE reads [139]. On the contrary, SE reads usually provide more sequencing depth for the same cost and they often generate higher mapping rates. However, improvements of the library preparation and sequencing protocols, diminishes the cost differences between PE and SE reads. A dedicated study showed that 40 bp PE reads resulted in more consistent gene expression estimates compared to 75 bp SE reads [140], while the same costs apply for both approaches. We consider the number of biological replicates, followed by sequencing depth, to be the two primary variables that should be considered in any DGE experiment, and we strongly recommend discussing these main parameters closely with the sequencing center.

Raw short reads provided by the sequencing center should be subjected to proper quality assessment, trimming and filtering as described above to obtain high-quality short reads to proceed with (Fig. 3). Specifically, the identification of samples with high percentage of rRNA reads will avoid artefacts associated with inaccurate estimates of library sequencing depth and therefore transcript coverage during DGE analysis [141–143].

In case the same RNA-seq reads will be used for de novo assemblies and subsequent DGE analyses (Figs. 1B and 2C), it is important to note that the read normalization step that may be done prior to the assembly must be omitted prior to the DGE analysis as this step will eliminate all differences in expression. Moreover, proper replication is required, even though the assembly per se does not require replicates. The type of sequencing as recommended above (Fig. 1B) represents an acceptable compromise between short SE reads for quantification and long PE reads for the assembly.

Read mapping, transcript quantification and statistical analyses

Mapping RNA-seq reads to a reference (whether a de novo transcriptome assembly or an annotated genome) lies at the core of any DGE analysis. A multitude of freely available read-mapping tools now exist (Table 1 for a selection) and the mapping quality can be software dependent [144]. Accordingly, there is no shortage of studies that compare the performances of different tools [48, 49, 145–147]. More traditional read-mapping algorithms (e.g. STAR and HISAT2, Table 1) are splice-aware, and have been reported to be more accurate with regards to miRNAs and lowly expressed genes [148] than pseudo-aligners (e.g. Kallisto, Table 1) and quasi-mappers (e.g. Salmon and RapMap, Table 1). Conversely, these more recent read-mapping tools have been reported to be more accurate than splice-aware methods for the quantification of long non-coding RNAs [149]. Overall, we advise the reader to appraise the strengths and weaknesses of the selected mapping tool given the primary goal of the read-mapping exercise, such as mRNA quantification, miRNA quantification or isoform quantification.

One issue that all read-mapping tools must deal with is how reads that mapped with equal fit to more than one location in the reference are handled. Such ambiguous read-mapping rates can be as high as 37% [150] and they could for instance be the result of splice-variants, duplicated genes, pseudo-genes and low complexity genes in the reference. This issue is particularly important when de novo assemblies serve as the mapping reference because each gene is often represented by many transcripts (whether true isoforms or spurious contigs). Several strategies exist to deal with multi-mapping reads including eliminating these reads from the DGE analysis, splitting them equally between their possible true origins, assigning them to a location based on a probability distribution constructed from unambiguously mapped reads, or collapsing the potential targets into gene groups and providing expression level estimates for these groups rather than the individual genes. While there appears to be no consensus on this issue [reviewed in 109], and new solutions are being actively developed [151, 152] we recommend that any DGE analysis should compare the impact of ignoring multi-mapping reads (i.e. only considering uniquely mapping reads) vs. assigning them to a location using at least one explicit strategy. If a de novo transcriptome assembly serves as the mapping reference, multiple tools have been established

which cluster transcripts based on the likelihood of having the same reads mapped to it (e.g. Corset, Grouper, Compacta; Table 1). This allows gene expression to be estimated on the level of clusters, rather than individual transcripts. Another option to reduce the complexity of the transcriptome reference is to only keep the longest isoform for each putative gene locus. This approach will only result in reliable results if the isoform-clustering of transcripts is performed correctly.

Importantly, some RNA-seq methods rely on sequencing only the 3-prime ends of RNA fragments (e.g. QuantSeq, [153]) or focus on transient RNA molecules (e.g. TT-seq, [154]) and they profit from or require a well-assembled and annotated reference genome to unambiguously assign RNA-seq reads to specific genes [155, 156]. Also commonly employed droplet-based single-cell RNA-seq technologies (such as the 10X Genomics platform) rely on 3' end sequencing [157], and 25% of the reads generated in single-nucleus RNA-seq data typically represent intronic sequences [158]. Therefore, a reference genome is often a prerequisite when gene expression is being assessed at single-cell/single-nucleus resolution. These special expression quantification methods are therefore not advisable if only de novo assembled transcriptome references are available.

After read mapping, a count matrix summarizing the number of reads mapped to each gene/transcript in each biological condition and replicate is generated and used as the input for the statistical assessment of expression differences. In summary, gene expression data is typically modelled based on a negative binomial distribution to estimate the mean gene/transcript expression, as well as the variance among replicates and to normalize the read counts for differences in the library size (i.e. total number of mapped reads) (e.g. DEseq2, edgeR; Table 1). A statistical comparison of the mean expression between conditions results in an adjusted p-value (after accounting for multiple testing), which can be used to identify a list of significantly differentially expressed genes (i.e. genes whose adjusted p-value falls below 0.05 or 0.01 for more conservative estimates). It has been demonstrated that the mainstream DGE software tools differ in their abilities to identify differentially expressed genes and in their rates of false positives [159]. To minimize this source of technical bias, tools have been developed by the community to take the consensus of multiple DGE analysis outputs and this has been shown to increase the robustness of DGE predictions [160]. Moreover, filtering out genes/transcripts supported by low read counts (for example less than 10 reads) is a common practice in RNA-seq data analysis, and can increase the number of differentially expressed genes detected and improve the sensitivity of DGE analyses [161, 162].

It is important to distinguish between the two major applications of transcript abundance estimation. The procedure outlined above assumes that raw read counts for each transcript are compared between biological/experimental conditions. Hence, the read counts to be compared are mapped against the same reference transcript/gene and accordingly the transcript length does not need to be considered. In contrast, comparisons of expression levels of different transcripts/genes within the same sample are regularly employed to generate heatmaps and principal component analyses (PCA) plots to globally describe the expression data. Moreover, abundance estimates across transcripts are typically used during the quality assessment of de novo assemblies to identify low-quality transcripts (see above). Statistical tools used for such cross-sample comparisons should not be used in this case as they do not account for differences in transcript/gene length. Instead, normalization procedures which account for library size and transcript/gene length such as RPKM (reads per kilobase of exon per million reads mapped) and its derivates FPKM (fragments per kilobase of exon per million fragments mapped) and TPM (transcripts per kilobase million) should be employed [163–165]. Special caution must be also applied if gene expression between different populations/species with different mapping references should be compared. A study in different *Drosophila* species showed that despite library size/transcript length normalization prior to the DGE analysis, false-positive significantly differentially expressed genes were identified [33]. To our knowledge, the problem of interspecific expression comparison is not yet conclusively solved.

Conclusion and outlook

Generating a de novo transcriptome for an emerging model organism represents a significant challenge for beginners, with many different scientific and technical aspects to consider. Data quality benefits from a clear communication with the sequencing center generating the raw data. Moreover, the quality and selection of open-source software tools that exist today, along with an extensive scientific literature and advice from colleagues can greatly support the journey from the initial scientific question to the discovery of the genes associated with a trait of interest. While comprehensive de novo assembled transcriptomes are excellent starting points to study and compare the gene content of an organism, DGE studies promise exciting new insights for subsets of biological processes or tissues. However, without tempering the excitement, it is important to keep in mind that any DGE analysis typically results in long lists of differentially expressed genes. These initial lists should always be regarded as candidates for

further investigation, rather than the final answer to a question. Often the real work begins with these lists and will require a range of additional computational analyses and wet lab experiments that should verify and test any resulting interpretations. While beyond the scope of this review, computational approaches to narrow down these often daunting lists of candidate genes may include assigning GO terms [166–168], or pathway enrichment analyses [169] which allow identifying specific sets of genes or perhaps an entire signaling pathway that is significantly up-regulated in the tissue/condition of interest. As these enrichment analyses are based on homology assignments, the downstream analysis pipeline should be flexible enough to accommodate unexpected outcomes, such as novel, fast-evolving, or lineage-restricted genes for which no functional data may be available in existing databases. Following the rational that genes with similar expression profiles are likely to be co-regulated, one can also reconstruct co-expression networks [170], which place individual candidate genes into a systemic regulatory context. Such networks may be very powerful to link uncharacterized candidate genes to genes with known functions. Eventually, independent verification of any DGE or network analysis should be performed. This can be achieved quantitatively (for example via qPCR), spatially (e.g. in situ hybridization) or functionally using gain- and loss-of-function methods (RNAi or CRISPR). The selection of candidate genes for such further studies could be based on their homology with genes known to be associated with similar traits in other model systems.

As new sequencing technologies are continuously and rapidly being deployed it is difficult to predict how the analysis of genomes and transcriptomes will evolve. What is certain is the identification of genes associated with biological traits of interest will continue to fascinate scientists from a variety of disciplines, from medicine to evolution to agriculture. By using the methods and techniques we have surveyed here, these questions can be readily addressed in organisms that do not have the historical pedigree that traditional models enjoy.

Abbreviations

bp	Base pairs
BLAST	Basic local alignment search tool
BUSCO	Benchmarking universal single-copy orthologs
cDNA	Complementary DNA
DGE	Differential gene expression
mRNA	Messenger RNA
PE	Paired-end
QC	Quality control
qPCR	Quantitative polymerase chain reaction
RNAi	RNA interference
rRNA	Ribosomal RNA
SE	Single-end

Acknowledgements
We are thankful for the valuable suggestions of the three reviewers of our work as they helped improving the manuscript. We acknowledge support by the Open Access Publication Funds of the Göttingen University.
This article is supported by the publication fund "NiedersachsenOPEN" (supported by "zukunftniedersachen").

Authors' contributions
Conceptualization: DJJ, NC, NP; Writing—Original Draft: DJJ, NC, NP; Writing—Review & Editing: DJJ, NC, NP; Visualization: NP, DJJ; Project administration: DJJ, NP.

Funding
Open Access funding enabled and organized by Projekt DEAL. DJJ is supported by a grant of the German Research Foundation (DFG) (528314512). NP is supported by grants of the German Research Foundation (DFG) (PO 1648/8–1, PO 1648/7–1, PO 1648/6–1).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]University of Göttingen, Department of Geobiology, Goldschmidtstr.3, Göttingen 37077, Germany. [2]University of Göttingen, Department of Developmental Biology, GZMB, Justus-Von-Liebig-Weg 11, Göttingen 37077, Germany.

Received: 5 March 2024 Accepted: 12 June 2024
Published online: 20 June 2024

References

1. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282:2012–8. Available from: https://doi.org/10.1126/science.282.5396.2012.
2. Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS. Ancient gene linkages support ctenophores as sister to other animals. Nature. 2023;618:110–7. Available from: https://doi.org/10.1038/s41586-023-05936-6.
3. Yan Z-G, Zhu X-M, Zhang S-W, Jiang H, Wang S-P, Wei C, et al. Environmental DNA sequencing reveals the regional difference in diversity and community assembly mechanisms of eukaryotic plankton in coastal waters. Front Microbiol. 2023;14:1132925. Available from: https://doi.org/10.3389/fmicb.2023.1132925.
4. Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, et al. Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. Mol Biol Evol. 2020;37:2661–78. Available from: https://doi.org/10.1093/molbev/msaa120.
5. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8. Available from: https://doi.org/10.1038/nmeth.2688
6. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet [Internet]. 2012;13:840–52. Available from: https://doi.org/10.1038/nrg3306.

7. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008;452:215–9. Available from: https://doi.org/10.1038/nature06745.
8. Rodriguez F, Arkhipova IR. An Overview of Best Practices for Transposable Element Identification, Classification, and Annotation in Eukaryotic Genomes. In: Branco MR, de Mendoza Soler A, editors. Transposable Elements: Methods and Protocols. New York, NY: Springer US; 2023. p. 1–23. Available from: https://doi.org/10.1007/978-1-0716-2883-6_1.
9. Kapun M, Nunez JCB, Bogaerts-Márquez M, Murga-Moreno J, Paris M, Outten J, et al. Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource. Mol Biol Evol. 2021;38:5782–805. Available from: https://doi.org/10.1093/molbev/msab259.
10. Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, et al. Eukaryotic genome size databases. Nucleic Acids Res. 2007;35:D332–8. Available from: https://doi.org/10.1093/nar/gkl828
11. Sigwart JD, Lindberg DR, Chen C, Sun J. Molluscan phylogenomics requires strategically selected genomes. Philos Trans R Soc Lond B Biol Sci. 2021;376:20200161. Available from: https://doi.org/10.1098/rstb.2020.0161
12. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5. Available from: https://doi.org/10.1126/science.aal3327.
13. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46. Available from: https://doi.org/10.1038/s41586-021-03451-0.
14. Yuan Y, Chung CY-L, Chan T-F. Advances in optical mapping for genomic research. Comput Struct Biotechnol J. 2020;18:2051–62. Available from: https://doi.org/10.1016/j.csbj.2020.07.018.
15. Leinonen M, Salmela L. Optical map guided genome assembly. BMC Bioinformatics. 2020;21. Available from: https://doi.org/10.1186/s12859-020-03623-1.
16. Luo J, Wei Y, Lyu M, Wu Z, Liu X, Luo H, et al. A comprehensive review of scaffolding methods in genome assembly. Brief Bioinform. 2021;22. Available from: https://doi.org/10.1093/bib/bbab033.
17. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv. 2023. Available from: https://doi.org/10.1101/2023.06.10.544449.
18. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;169:1177–86. Available from: https://doi.org/10.1016/j.cell.2017.05.038.
19. Buchberger E, Reis M, Lu T-H, Posnien N. Cloudy with a Chance of Insights: Context Dependent Gene Regulation and Implications for Evolutionary Studies. Genes. 2019;10. Available from: https://doi.org/10.3390/genes10070492.
20. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:83. Available from: https://doi.org/10.1186/s13059-017-1215-1.
21. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631–56. Available from: https://doi.org/10.1038/s41576-019-0150-2.
22. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI, et al. RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. Annu Rev Biomed Data Sci. 2019;2:139–73. Available from: https://doi.org/10.1146/annurev-biodatasci-072018-021255.
23. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6:377–82. Available from: https://doi.org/10.1038/nmeth.1315.
24. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58:610–20. Available from: https://doi.org/10.1016/j.molcel.2015.04.005.
25. Buchberger E, Bilen A, Ayaz S, Salamanca D, Matas de Las Heras C, Niksic A, et al. Variation in Pleiotropic Hub Gene Expression Is Associated with Interspecific Differences in Head Shape and Eye Size in Drosophila. Mol Biol Evol. 2021;38:1924–42. Available from: https://doi.org/10.1093/molbev/msaa335
26. Oppenheim SJ, Baker RH, Simon S, DeSalle R. We can't all be supermodels: the value of comparative transcriptomics to the study of non-model insects. Insect Mol Biol. 2015;24:139–54. Available from: https://doi.org/10.1111/imb.12154.
27. Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, et al. FlyBase: updates to the *Drosophila* genes and genomes database. Genetics. 2024; Available from: https://doi.org/10.1093/genetics/iyad211.
28. Drysdale RA, Crosby MA, FlyBase Consortium. FlyBase: genes and gene models. Nucleic Acids Res. 2005;33:D390-5. Available from: https://doi.org/10.1093/nar/gki046.
29. Tribolium Genome Sequencing Consortium, Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, et al. The genome of the model beetle and pest Tribolium castaneum. Nature. 2008;452:949–55. Available from: https://doi.org/10.1038/nature06784.
30. Herndon N, Shelton J, Gerischer L, Ioannidis P, Ninova M, Dönitz J, et al. Enhanced genome assembly and a new official gene set for Tribolium castaneum. BMC Genomics. 2020;21:47. Available from: https://doi.org/10.1186/s12864-019-6394-6.
31. Darwin Tree of Life Project Consortium. Sequence locally, think globally: The Darwin Tree of Life Project. Proc Natl Acad Sci U S A. 2022;119. Available from: https://doi.org/10.1073/pnas.2115642118.
32. Mazzoni CJ, Ciofi C, Waterhouse RM. Biodiversity: an atlas of European reference genomes. Nature. 2023;619:252. Available from: https://doi.org/10.1038/d41586-023-02229-w.
33. Torres-Oliva M, Almudi I, McGregor AP, Posnien N. A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. BMC Genomics. 2016;17:392. Available from: https://doi.org/10.1186/s12864-016-2646-x.
34. Sun Y-M, Chen Y-Q. Principles and innovative technologies for decrypting noncoding RNAs: from discovery and functional prediction to clinical application. J Hematol Oncol. 2020;13:109. Available from: https://doi.org/10.1186/s13045-020-00945-8.
35. Fachrul M, Karkey A, Shakya M, Judd LM, Harshegyi T, Sim KS, et al. Direct inference and control of genetic population structure from RNA sequencing data. Commun Biol. 2023;6:804. Available from: https://doi.org/10.1038/s42003-023-05171-9.
36. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. Am J Hum Genet. 2013;93:641–51. Available from: https://doi.org/10.1016/j.ajhg.2013.08.008.
37. Hewson I, Eggleston EM, Doherty M, Lee DY, Owens M, Shapleigh JP, et al. Metatranscriptomic analyses of plankton communities inhabiting surface and subpycnocline waters of the Chesapeake Bay during oxic-anoxic-oxic transitions. Appl Environ Microbiol. 2014;80:328–38. Available from: https://doi.org/10.1128/aem.02680-13.
38. Shakya M, Lo C-C, Chain PSG. Advances and Challenges in Metatranscriptomic Analysis. Front Genet. 2019;10. Available from: https://doi.org/10.3389/fgene.2019.00904.
39. González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, et al. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proc Biol Sci. 2015;282:20142332. Available from: https://doi.org/10.1098/rspb.2014.2332.
40. Peters RS, Meusemann K, Petersen M, Mayer C, Wilbrandt J, Ziesmann T, et al. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. BMC Evol Biol. 2014;14:52. Available from: https://doi.org/10.1186/1471-2148-14-52.
41. Bucek A, Šobotník J, He S, Shi M, McMahon DP, Holmes EC, et al. Evolution of Termite Symbiosis Informed by Transcriptome-Based Phylogenies. Curr Biol. 2019;29:3728–3734.e4. Available from: https://doi.org/10.1016/j.cub.2019.08.076.
42. Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T. A transcriptome approach to ecdysozoan phylogeny. Mol Phylogenet Evol. 2014;80:79–87. Available from: https://doi.org/10.1016/j.ympev.2014.08.001.
43. Zhao L, Wang S, Lou F, Gao T, Han Z. Phylogenomics based on transcriptome data provides evidence for the internal phylogenetic relationships and potential terrestrial evolutionary genes of lungfish. Front Mar Sci. 2021;8. Available from: https://doi.org/10.3389/fmars.2021.724977.
44. Dylus D, Altenhoff A, Majidian S, Sedlazeck FJ, Dessimoz C. Inference of phylogenetic trees directly from raw sequencing reads using

Read2Tree. Nat Biotechnol. 2024;42:139–47. Available from: https://doi.org/10.1038/s41587-023-01753-4.
45. Mehlhorn S, Hunnekuhl VS, Geibel S, Nauen R, Bucher G. Establishing RNAi for basic research and pest control and identification of the most efficient target genes for pest control: a brief guide. Front Zool. 2021;18:60. Available from: https://doi.org/10.1186/s12983-021-00444-7.
46. Hook PW, Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. Nat Rev Genet [Internet]. 2023;24:627–41. Available from: https://doi.org/10.1038/s41576-023-00600-1
47. Guizard S, Miedzinska K, Smith J, Smith J, Kuo RI, Davey M, et al. nf-core/isoseq: simple gene and isoform annotation with PacBio Iso-Seq long-read sequencing. Bioinformatics. 2023;39. Available from: https://doi.org/10.1093/bioinformatics/btad150.
48. Corchete LA, Rojas EA, Alonso-López D, De Las Rivas J, Gutiérrez NC, Burguillo FJ. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci Rep. 2020;10:19737. Available from: https://doi.org/10.1038/s41598-020-76881-x.
49. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13. Available from: https://doi.org/10.1186/s13059-016-0881-8.
50. Gleeson J, Leger A, Prawer YDJ, Lane TA, Harrison PJ, Haerty W, et al. Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Res [Internet]. 2022;50:e19–e19. Available from: https://doi.org/10.1093/nar/gkab1129.
51. Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J. Methodologies for transcript profiling using long-read technologies. Front Genet. 2020;11. Available from: https://doi.org/10.3389/fgene.2020.00606.
52. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24:1384–95. Available from: https://doi.org/10.1101/gr.170720.113.
53. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;7:3. Available from: https://doi.org/10.1186/1471-2199-7-3.
54. Takele Assefa A, Vandesompele J, Thas O. On the utility of RNA sample pooling to optimize cost and statistical power in RNA sequencing experiments. BMC Genomics. 2020;21:312. Available from: https://doi.org/10.1186/s12864-020-6721-y.
55. O'Neil ST, Emrich SJ. Assessing De Novo transcriptome assembly metrics for consistency and utility. BMC Genomics. 2013;14:465. Available from: https://doi.org/10.1186/1471-2164-14-465.
56. Posnien N, Zeng V, Schwager EE, Pechmann M, Hilbrant M, Keefe JD, et al. A comprehensive reference transcriptome resource for the common house spider Parasteatoda tepidariorum. PLoS One. 2014;9:e104885. Available from: https://doi.org/10.1371/journal.pone.0104885.
57. Alvarez RV, Landsman D. GTax: improving de novo transcriptome assembly by removing foreign RNA contamination. Genome Biol. 2024;25. Available from: https://doi.org/10.1186/s13059-023-03141-2.
58. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7. Available from: https://doi.org/10.1093/bioinformatics/bts611.
59. Bushnell B, Rood J, Singer E. BBMerge – Accurate paired shotgun read merging via overlap. PLoS One. 2017;12:e0185056. Available from: https://doi.org/10.1371/journal.pone.0185056.
60. FastQC. updated March 1 2023 [cited 2024 Mar 5]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
61. Zhou Q, Su X, Jing G, Chen S, Ning K. RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC Genomics. 2018;19:144. Available from: https://doi.org/10.1186/s12864-018-4503-6.
62. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5. Available from: https://doi.org/10.1093/bioinformatics/bts356.
63. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8. Available from: https://doi.org/10.1093/bioinformatics/btw354.
64. Yang X, Liu D, Liu F, Wu J, Zou J, Xiao X, et al. HTQC: a fast quality control toolkit for Illumina sequencing data. BMC Bioinformatics. 2013;14:33. Available from: https://doi.org/10.1186/1471-2105-14-33.
65. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. Available from: https://doi.org/10.1093/bioinformatics/btu170.
66. 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. Nat Biotechnol. 2011;29:644–52. Available from: https://doi.org/10.1038/nbt.1883.
67. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1:18. Available from: https://doi.org/10.1186/2047-217X-1-18.
68. Peng Y, Leung HCM, Yiu S-M, Lv M-J, Zhu X-G, Chin FYL. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics. 2013;29:i326–34. Available from: https://doi.org/10.1093/bioinformatics/btt219.
69. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7:909–12. Available from: https://doi.org/10.1038/nmeth.1517.
70. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience. 2019;8. Available from: https://doi.org/10.1093/gigascience/giz100.
71. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018;19:153. Available from: https://doi.org/10.1186/s13059-018-1540-z.
72. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9. Available from: https://doi.org/10.1101/gr.074492.107.
73. Rivera-Vicéns RE, Garcia-Escudero CA, Conci N, Eitel M, Wörheide G. TransPi-a comprehensive TRanscriptome ANalysiS PIpeline for de novo transcriptome assembly. Mol Ecol Resour. 2022;22:2070–86. Available from: https://doi.org/10.1111/1755-0998.13593.
74. Cerveau N, Jackson DJ. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. BMC Bioinformatics. 2016;17:525. Available from: https://doi.org/10.1186/s12859-016-1406-x.
75. Nakasugi K, Crowhurst R, Bally J, Waterhouse P. Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana. PLoS One. 2014;9:e91776. Available from: https://doi.org/10.1371/journal.pone.0091776.
76. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience. 2018;7. Available from: https://doi.org/10.1093/gigascience/giy093.
77. Voshall A, Behera S, Li X, Yu X-H, Kapil K, Deogun JS, et al. A consensus-based ensemble approach to improve transcriptome assembly. BMC Bioinformatics. 2021;22:513. Available from: https://doi.org/10.1186/s12859-021-04434-8.
78. 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 Biol. 2014;15:553. Available from: https://doi.org/10.1186/s13059-014-0553-5.
79. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016;26:1134–44. Available from: https://doi.org/10.1101/gr.196469.115.
80. Bushmanova E, Antipov D, Lapidus A, Suvorov V, Prjibelski AD. rnaQUAST: a quality assessment tool for *de novo* transcriptome assemblies. Bioinformatics. 2016;32:2210–2. Available from: https://doi.org/10.1093/bioinformatics/btw218.
81. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54. Available from: https://doi.org/10.1093/molbev/msab199.

82. Dohmen E, Kremer LPM, Bornberg-Bauer E, Kemena C. DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics. 2016;32:2577–81. Available from: https://doi.org/10.1093/bioinformatics/btw231.
83. Kerkvliet J, de Fouchier A, van Wijk M, Groot AT. The Bellerophon pipeline, improving de novo transcriptomes and removing chimeras. Ecol Evol [Internet]. 2019;9:10513–21. Available from: https://doi.org/10.1002/ece3.5571
84. Haas B. TransDecoder Github. updated July 16 2023 [cited 2024 Mar 5]. Available from: https://github.com/TransDecoder/TransDecoder.
85. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7:e1002195. Available from: https://doi.org/10.1371/journal.pcbi.1002195.
86. Haas B. Trinotate Github. updated September 8 2023 [cited 2024 Mar 5]. Available from: https://github.com/Trinotate/Trinotate.
87. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep. 2017;18:762–76. Available from: https://doi.org/10.1016/j.celrep.2016.12.063.
88. Scott C, Ward TP. dammit Github. updated December 10 2019 [cited 2024 Mar 5]. Available from: https://github.com/dib-lab/dammit.
89. Scott C. dammit Documentation. 2020 [cited 2024 Mar 5]. Available from: https://dib-lab.github.io/dammit/.
90. Hart AJ, Ginzburg S, Xu MS, Fisher CR, Rahmatpour N, Mitton JB, et al. EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. Mol Ecol Resour. 2020;20:591–604. Available from: https://doi.org/10.1111/1755-0998.13106.
91. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. Available from: https://doi.org/10.1093/bioinformatics/bts635.
92. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51:11.14.1–11.14.19. Available from: https://doi.org/10.1002/0471250953.bi1114s51.
93. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15. Available from: https://doi.org/10.1038/s41587-019-0201-4.
94. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7. Available from: https://doi.org/10.1038/nbt.3519.
95. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9. Available from: https://doi.org/10.1038/nmeth.4197.
96. Srivastava A, Sarkar H, Gupta N, Patro R. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. Bioinformatics. 2016;32:i192–200. Available from: https://doi.org/10.1093/bioinformatics/btw277.
97. Davidson NM, Oshlack A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol. 2014;15:410. Available from: https://doi.org/10.1186/s13059-014-0410-6.
98. Malik L, Almodaresi F, Patro R. Grouper: graph-based clustering and annotation for improved *de novo* transcriptome analysis. Bioinformatics. 2018;34:3265–72. Available from: https://doi.org/10.1093/bioinformatics/bty378.
99. Razo-Mendivil FG, Martínez O, Hayano-Kanashiro C. Compacta: a fast contig clustering tool for de novo assembled transcriptomes. BMC Genomics. 2020;21:148. Available from: https://doi.org/10.1186/s12864-020-6528-x.
100. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550. Available from: https://doi.org/10.1186/s13059-014-0550-8.
101. Love MI. DESeq2 Github. updated February 2024 [cited 2024 Mar 5]. Available from: https://github.com/thelovelab/DESeq2.
102. Yunshun Chen <yuchen@wehi.edu.au>, Aaron Lun<alun@wehi.edu.au>, Davis McCarthy <dmccarthy@wehi.edu.au>, Xiaobei Zhou <xiaobei.zhou@uzh.ch>, Mark Robinson<mark.robinson@imls.uzh.ch>, Gordon Smyth <smyth@wehi.edu.au>. edgeR. Bioconductor; 2017. Available from: https://doi.org/10.18129/B9.BIOC.EDGER.
103. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. Available from: https://doi.org/10.1093/bioinformatics/btp616.
104. Durai DA, Schulz MH. Improving in-silico normalization using read weights. Sci Rep. 2019;9. Available from: https://doi.org/10.1038/s41598-019-41502-9.
105. 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. Nat Protoc. 2013;8:1494–512. Available from: https://doi.org/10.1038/nprot.2013.084.
106. Wedemeyer A, Kliemann L, Srivastav A, Schielke C, Reusch TB, Rosenstiel P. An improved filtering algorithm for big read datasets and its application to single-cell assembly. BMC Bioinformatics. 2017;18:324. Available from: https://doi.org/10.1186/s12859-017-1724-7.
107. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to de novo transcriptome assembly and annotation. Brief Bioinform. 2022;23. Available from: https://doi.org/10.1093/bib/bbab563.
108. Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, et al. Separating homeologs by phasing in the tetraploid wheat transcriptome. Genome Biol. 2013;14:R66. Available from: https://doi.org/10.1186/gb-2013-14-6-r66.
109. Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhart P. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics [Internet]. 2012;13:92. Available from: https://doi.org/10.1186/1471-2164-13-92.
110. Durai DA, Schulz MH. Informed *k*mer selection for *de novo* transcriptome assembly. Bioinformatics. 2016;32:1670–7. Available from: https://doi.org/10.1093/bioinformatics/btw217.
111. Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience. 2019;8. Available from: https://doi.org/10.1093/gigascience/giz039.
112. Jänes J, Hu F, Lewin A, Turro E. A comparative study of RNA-seq analysis strategies. Brief Bioinform. 2015;16:932–40. Available from: https://doi.org/10.1093/bib/bbv007.
113. 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. Nat Biotechnol [Internet]. 2010;28:511–5. Available from: https://doi.org/10.1038/nbt.1621
114. Sarov M, Barz C, Jambor H, Hein MY, Schmied C, Suchold D, et al. A genome-wide resource for the analysis of protein localisation in Drosophila. Elife. 2016;5:e12068. Available from: https://doi.org/10.7554/eLife.12068.
115. Kim Y, Park Y, Hwang J, Kwack K. Comparative genomic analysis of the human and nematode*Caenorhabditis elegans*uncovers potential reproductive genes and disease associations in humans. Physiol Genomics. 2018;50:1002–14. Available from: https://doi.org/10.1152/physiolgenomics.00063.2018.
116. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse. Nat Rev Genet. 2017;18:425–40. Available from: https://doi.org/10.1038/nrg.2017.19.
117. Tao X, Gu Y-H, Wang H-Y, Zheng W, Li X, Zhao C-W, et al. Digital gene expression analysis based on integrated de novo transcriptome assembly of sweet potato [Ipomoea batatas (L.) Lam]. PLoS One. 2012;7:e36234. Available from: https://doi.org/10.1371/journal.pone.0036234.
118. Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, de Souza Gonçalves P, et al. De novo assembly and transcriptome analysis of the rubber tree (Hevea brasiliensis) and SNP markers development for rubber biosynthesis pathways. PLoS One. 2014;9:e102665. Available from: https://doi.org/10.1371/journal.pone.0102665.
119. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. Available from: https://doi.org/10.1038/35057062.
120. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8:1765–86. Available from: https://doi.org/10.1038/nprot.2013.099.
121. Haas B. Assessing the Read Content of the Transcriptome Assembly. updated January 29 2022 [cited 2024 Mar 5]. Available from: https://

github.com/trinityrnaseq/trinityrnaseq/wiki/RNA-Seq-Read-Representation-by-Trinity-Assembly.
122. Ma C, Kingsford C. Detecting, categorizing, and correcting coverage anomalies of RNA-seq quantification. Cell Syst. 2019;9:589–599.e7. Available from: https://doi.org/10.1016/j.cels.2019.10.005.
123. Haas B. Trinity Transcriptome Contig Nx and ExN50 Statistics. updated February 5 2023 [cited 2024 Mar 5]. Available from: https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcript ome-Contig-Nx-and-ExN50-stats.
124. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9. Available from: https://doi.org/10.1093/nar/gkaa1100.
125. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50:D20–6. Available from: https://doi.org/10.1093/nar/gkab1112.
126. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. Available from: https://doi.org/10.1016/S0022-2836(05)80360-2.
127. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023;51:D418–27. Available from: https://doi.org/10.1093/nar/gkac993.
128. Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224. Available from: https://doi.org/10.1093/genetics/iyad031.
129. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30. Available from: https://doi.org/10.1093/nar/28.1.27.
130. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51:D587–92. Available from: https://doi.org/10.1093/nar/gkac963.
131. Chowdhury HA, Bhattacharyya DK, Kalita JK. Differential expression analysis of RNA-seq reads: Overview, taxonomy and tools. IEEE/ACM Trans Comput Biol Bioinform. 2018;1–1. Available from: https://doi.org/10.1109/tcbb.2018.2873010.
132. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11:220. Available from: https://doi.org/10.1186/gb-2010-11-12-220.
133. Chalifa-Caspi V. RNA-Seq in Nonmodel Organisms. In: Shomron N, editor. Deep Sequencing Data Analysis. New York, NY: Springer US; 2021. p. 143–67. Available from: https://doi.org/10.1007/978-1-0716-1103-6_8.
134. Cheng H, Wang Y, Sun M-A. Comparison of Gene Expression Profiles in Nonmodel Eukaryotic Organisms with RNA-Seq. In: Wang Y, Sun M-A, editors. Transcriptome Data Analysis: Methods and Protocols. New York, NY: Springer New York; 2018. p. 3–16. Available from: https://doi.org/10.1007/978-1-4939-7710-9_1.
135. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22:839–51. Available from: https://doi.org/10.1261/rna.053959.115.
136. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, et al. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics. 2015;16:675. Available from: https://doi.org/10.1186/s12864-015-1876-7.
137. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol. 2015;16:131. Available from: https://doi.org/10.1186/s13059-015-0697-y.
138. Lamarre S, Frasse P, Zouine M, Labourdette D, Sainderichin E, Hu G, et al. Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size. Front Plant Sci. 2018;9:108. Available from: https://doi.org/10.3389/fpls.2018.00108.
139. Corley SM, MacKenzie KL, Beverdam A, Roddam LF, Wilkins MR. Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. BMC Genomics. 2017;18. Available from: https://doi.org/10.1186/s12864-017-3797-0.
140. Freedman AH, Gaspar JM, Sackton TB. Short paired-end reads trump long single-end reads for expression analysis. BMC Bioinformatics. 2020;21. Available from: https://doi.org/10.1186/s12859-020-3484-z.
141. Deyneko IV, Mustafaev ON, Tyurin AA, Zhukova KV, Varzari A, Goldenkova-Pavlova IV. Modeling and cleaning RNA-seq data significantly improve detection of differentially expressed genes. BMC Bioinformatics. 2022;23:488. Available from: https://doi.org/10.1186/s12859-022-05023-z.
142. Sheng Q, Vickers K, Zhao S, Wang J, Samuels DC, Koues O, et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. Brief Funct Genomics. 2017;16:194–204. Available from: https://doi.org/10.1093/bfgp/elw035.
143. Kumar G, Ertel A, Feldman G, Kupper J, Fortina P. iSeqQC: a tool for expression-based quality control in RNA sequencing. BMC Bioinformatics. 2020;21:56. Available from: https://doi.org/10.1186/s12859-020-3399-8.
144. Srivastava A, Malik L, Sarkar H, Zakeri M, Almodaresi F, Soneson C, et al. Alignment and mapping methodology influence transcript abundance estimation. Genome Biol. 2020;21:239. Available from: https://doi.org/10.1186/s13059-020-02151-8.
145. Schaarschmidt S, Fischer A, Zuther E, Hincha DK. Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana. Int J Mol Sci. 2020;21:1720. Available from: https://doi.org/10.3390/ijms21051720.
146. Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J. 2020;18:1569–76. Available from: https://doi.org/10.1016/j.csbj.2020.06.014.
147. Donato L, Scimone C, Rinaldi C, D'Angelo R, Sidoti A. New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies. Neural Comput Appl. 2021;33:15669–92. Available from: https://doi.org/10.1007/s00521-021-06188-z.
148. Wu DC, Yao J, Ho KS, Lambowitz AM, Wilke CO. Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics. 2018;19:510. Available from: https://doi.org/10.1186/s12864-018-4869-5.
149. Zheng H, Brennan K, Hernaez M, Gevaert O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. Gigascience. 2019;8. Available from: https://doi.org/10.1093/gigascience/giz145.
150. McDermaid A, Chen X, Zhang Y, Wang C, Gu S, Xie J, et al. A New Machine Learning-Based Framework for Mapping Uncertainty Analysis in RNA-Seq Read Alignment and Gene Expression Estimation. Front Genet. 2018;9:313. Available from: https://doi.org/10.3389/fgene.2018.00313.
151. Hita A, Brocart G, Fernandez A, Rehmsmeier M, Alemany A, Schvartzman S. MGcount: a total RNA-seq quantification tool to address multi-mapping and multi-overlapping alignments ambiguity in non-coding transcripts. BMC Bioinformatics. 2022;23:39. Available from: https://doi.org/10.1186/s12859-021-04544-3.
152. Deschamps-Francoeur G, Boivin V, Abou Elela S, Scott MS. CoCo: RNA-seq read assignment correction for nested genes and multimapped reads. Bioinformatics. 2019;35:5039–47. Available from: https://doi.org/10.1093/bioinformatics/btz433.
153. Moll P, Ante M, Seitz A, Reda T. QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods. 2014 [cited 2024 Mar 5];11:i–iii. Available from: https://www.nature.com/articles/nmeth.f.376.
154. Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, et al. TT-seq maps the human transient transcriptome. Science. 2016;352:1225–8. Available from: https://doi.org/10.1126/science.aad9841.
155. Ma F, Fuqua BK, Hasin Y, Yukhtman C, Vulpe CD, Lusis AJ, et al. A comparison between whole transcript and 3′ RNA sequencing methods using Kapa and Lexogen library preparation methods. BMC Genomics. 2019;20:9. Available from: https://doi.org/10.1186/s12864-018-5393-3.
156. Tandonnet S, Torres TT. Traditional versus 3′ RNA-seq in a non-model species. Genom Data. 2017;11:9–16. Available from: https://doi.org/10.1016/j.gdata.2016.11.002.

157. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65:631–643.e4. Available from: https://doi.org/10.1016/j.molcel.2017.01.023.
158. Grindberg RV, Yee-Greenbaum JL, McConnell MJ, Novotny M, O'Shaughnessy AL, Lambert GM, et al. RNA-sequencing from single nuclei. Proc Natl Acad Sci U S A. 2013;110:19802–7. Available from: https://doi.org/10.1073/pnas.1319700110.
159. Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, et al. Robustness of differential gene expression analysis of RNA-seq. Comput Struct Biotechnol J. 2021;19:3470–81. Available from: https://doi.org/10.1016/j.csbj.2021.05.040.
160. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One. 2017;12:e0190152. Available from: https://doi.org/10.1371/journal.pone.0190152.
161. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, et al. RNA-seq analysis is easy as 1–2–3 with limma, Glimma and edgeR. F1000Res. 2018;5:1408. Available from: https://doi.org/10.12688/f1000research.9005.3.
162. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015. p. 6461–4. Available from: https://doi.org/10.1109/EMBC.2015.7319872.
163. Abrams ZB, Johnson TS, Huang K, Payne PRO, Coombes K. A protocol to evaluate RNA sequencing normalization methods. BMC Bioinformatics. 2019;20:679. Available from: https://doi.org/10.1186/s12859-019-3247-x.
164. Zhao S, Ye Z, Stanton R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA. 2020;26:903–9. Available from: https://doi.org/10.1261/rna.074922.120.
165. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94. Available from: https://doi.org/10.1186/1471-2105-11-94.
166. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008;9:509–15. Available from: https://doi.org/10.1038/nrg2363.
167. Tipney H, Hunter L. An introduction to effective use of enrichment analysis software. Hum Genomics. 2010;4:202–6. Available from: https://doi.org/10.1186/1479-7364-4-3-202.
168. Gene Ontology. 1999–2024 [cited 2024 Mar 5]. Available from: https://geneontology.org/.
169. Chicco D, Agapito G. Nine quick tips for pathway enrichment analysis. PLoS Comput Biol. 2022;18:e1010348. Available from: https://doi.org/10.1371/journal.pcbi.1010348.
170. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform. 2018;19:575–92. Available from: https://doi.org/10.1093/bib/bbw139.

Publisher's Note