Computational and Structural Biotechnology Journal 21 (2023) 86–98

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Temporal progress of gene expression analysis with RNA-Seq data: A review on the relationship between computational methods

Juliana Costa-Silva [a], Douglas S. Domingues [b], David Menotti [a], Mariangela Hungria [c], Fabrício Martins Lopes [d,*]

[a] Department of Informatics – Federal University of Paraná, Rua Coronel Francisco Heráclito dos Santos, 100, 81531-990 Curitiba, Paraná, Brazil
[b] Department of Genetics, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Av. Pádua Dias, 11, 13418-900 Piracicaba, São Paulo, Brazil
[c] Department of Soil Biotecnology - Embrapa Soybean, Cx. Postal 231, 86000-970 Londrina, Paraná, Brazil
[d] Department of Computer Science, Universidade Tecnológica Federal do Paraná – UTFPR, Av. Alberto Carazzai, 1640, 86300-000, Cornélio Procópio, Paraná, Brazil

ARTICLE INFO

Article history:
Received 15 August 2022
Received in revised form 25 November 2022
Accepted 25 November 2022
Available online 1 December 2022

Keywords:
RNA-Seq
Differential expression analysis
Gene expression
Bioinformatics

ABSTRACT

Analysis of differential gene expression from RNA-seq data has become a standard for several research areas. The steps for the computational analysis include many data types and file formats, and a wide variety of computational tools that can be applied alone or together as pipelines. This paper presents a review of the differential expression analysis pipeline, addressing its steps and the respective objectives, the principal methods available in each step, and their properties, therefore introducing an organized overview to this context. This review aims to address mainly the aspects involved in the differentially expressed gene (DEG) analysis from RNA sequencing data (RNA-seq), considering the computational methods. In addition, a timeline of the computational methods for DEG is shown and discussed, and the relationships existing between the most important computational tools are presented by an interaction network. A discussion on the challenges and gaps in DEG analysis is also highlighted in this review. This paper will serve as a tutorial for new entrants into the field and help established users update their analysis pipelines.

Contents

* Corresponding author.
E-mail addresses: julianacostasilvati@gmail.com (J. Costa-Silva), fabricio@utfpr.edu.br (F.M. Lopes).

https://doi.org/10.1016/j.csbj.2022.11.051

1. Introduction

RNA-seq is the standard sequencing technique to reveal the presence of, and quantify, transcripts in a biological sample. As a result, it allows differential expression analysis from RNA sequencing data (RNA-seq) sequences. Therefore, several methods have been proposed with different approaches and improvements in order to perform the differentially expressed gene (DEG) analysis from RNA-seq [1–3].

Quantification of gene expression, identification of novel transcripts and detection of fusion transcripts are the major applications of RNA-Seq. Fusion transcripts detection are described more detailed in [4].

Especially during the last decade, RNA-seq has become an indispensable tool for DEG analysis [5]. In this context, considering the diversity of methods proposed in the literature, there are some interesting questions that deserve attention: which methodologies are available for DEG analysis? Which are the different approaches used by the methods for the definition of DEGs? What are the differences between the analyses performed by each method? Finally, how do we choose a method for DEG analysis?.

Some common goals are usually present in the methodologies for DEG such as (i) improving the accuracy of the results; (ii) to removing biases from the analysis [6], and (iii) filling in some gaps in the existing analyses [7].

However, although most methods share these goals, the computational tools developed for DEG analysis present different approaches for evaluating their results. Therefore, defining which method to use to obtain greater precision in DEG analysis is not a trivial question, mainly because of the large number of variables involved in this decision.

Expression analysis can be performed in coding genes, non-coding genes (small RNAs, lncRNAs) and transposable elements (coding and non-coding parts). Each one of these "types" of sequences has specificities regarding the expression analysis, and how the computational tools perform them is not the focus of the present manuscript. On the contrary, this review focuses on how the computational methods infer the genes (or reads) with differential expression.

When considering methods and their implementations (software) developed since the popularization of RNA-seq, each experiment is tagged by one or more tools in the search for increasingly efficient DEG analysis. In this sense, some studies address this context and present a review of existing approaches to the analysis, their characteristics and applications [8–10].

Given the challenge of determining whether the count of a transcript or exon is significantly different between experimental conditions, the pioneer edgeR tool for DEG analysis [11], using parametric analysis has become widely used since its creation. Another challenge is the simultaneous transcript discovery and abundance estimation without requiring by prior gene annotations. Cufflinks method was developed to address these analytic issues [7].

Despite the great popularity of computational tools initially developed [12,7,6], many advances are still needed. In this context, some tools have sought to improve aspects, such as the impact of the depth of sequencing in DEG identification [13], the quantification of genes and isoforms with or without reference genomes [14], distinct experimental conditions and their influence on these analyses [14], among others.

Considering the number of existing methods and software dedicated to DEG analysis, it is possible to observe that the choice for parametric methods [1] is recurrent and, in general, presents adequate results. Moreover, it is also notable the series of steps to identify DEG needs to be computationally simplified, given the volume of pipelines developed after the RNA-seq popularization [15–18,1].

However, the variety of particularities found in DEG analysis generate specific studies that address each of the details of existing methods and lead to alternative approaches. Some studies have recently reviewed aspects that affect the outcome of DEG analysis, considering different applications such as phylogeny, single-cell and bulk RNA [19–22]. In particular, the adopted approach can significantly affect the outcome of the DEG analysis, as well as the conclusions that a single tool is unlikely to be optimal in all circumstances. Besides, the number of replicates and the heterogeneity of the samples should be considered when selecting the pipeline [19].

The statistical approaches for DEG analysis, such as the Negative Binomial distribution and likelihood (quasi-likelihood) are presented and discussed regarding the problem of over-scattering in RNA-seq samples and phylogeny [20]. Analysis of expression data integrated with public data and meta-data, can also be an option, however, it is a more complex analysis, more information about integrated data analysis can be found in [23]. Systematically comparing approaches for DEG analysis coupled with gene annotations is an approach that has been of interest [24].

Regarding single-cell data, three different scenarios were considered in order to assess the DEGs analyses showing significant differences in the results when considering different methods, such as the number of DEGs and their sensitivity and specificity [21].

The popularization of methods that integrate the steps of the DEG analysis (pipeline) partly because analysis has several steps, and for each step has a specific type of file. Another challenge for the understanding and applying DEG methods relies on the various options for tools in each step. It is common to find studies on related themes that use totally different analytical methods [22,25,26]. Interpreting results in DEG analysis is a relevant aspect to be considered, such as the different data visualisations that help the understanding of the results [22].

The methods for DEG analysis have some characteristics that allow their grouping. One of them, adopted in this review, is how to treat the expression distribution. Approaches that consider a specific distribution for the data analysis, i.e., that data will have a certain statistical distribution, are known as parametric approaches. In contrast, some methods do not consider data distribution (data with unknown distribution) and are known as non-parametric approaches. An additional option is to consider both or more approaches to indicate the DEGs, and these are known as hybrid approaches.

In this context, we review the principal methods of DEG analysis and describes the evolution of computational methods, their properties and relationships. Moreover, a historical context is presented with the main methodologies implemented since the increased use of the RNA-seq, seeking to identify the main alternatives used to perform DEG analysis and clarify issues about this context, such as the questions raised early in this review.

To support the understanding of biological systems, expression

[1] Which state that the data follow a certain distribution.

analysis can also be performed considering only cells of interest. This is possible through single-cell RNA sequencing (scRNA-seq). Although this review focuses on differential expression analysis techniques for bulk RNA-seq data, some considerations about expression analysis using single-cell data and its methods are also briefly presented.

The steps commonly used in the differential expression analysis from RNA-seq data will be presented in the Differential Expression Analysis section, followed by a brief history about the methodologies of differential expression analysis, discussing their main properties, similarities, differences and applications in the Methods for the section of differential gene expression analysis. In the end, a discussion about the convergent and divergent points between the analyzed methodologies is presented in the Discussion section, along with some conclusions and guidelines that can help in the choice of methodologies and definition of experiments for DEG analysis.

2. Differential expression analysis (pipeline)

Differential expression analysis consists of several steps, which are presented in this section to provide an overview and show the challenges involved. Naturally, the first step is to explore the key steps and the respective methods available. The second step is to choose the composition of an analysis protocol for differential expression, commonly referred to as a differential expression analysis pipeline.

In RNA-Seq experiments the number of replicates must also be taken into consideration; in this context, the reader can rely on studies that evaluate the impact of the number of replicates on the final result of the analysis, like [27].

Fig. 1 presents the steps commonly adopted in differential expression analysis, such as quality control and trimming, alignment, counting, normalization and expression analysis (to define the genes/sequences with differential expression).

Considering the importance of choosing methodologies for each of these steps, some studies sought to establish a protocol for the analysis [28,29]. Instead, it is also possible to find studies that present divergences in the choices of methods in some steps of differential expression analysis [25,26,22].

This review aims to address mainly the aspects involved in DEG analysis and contextualize the related methods although preparing sequencing libraries represents an important issue regarding its impact on the results, as reported in previous studies [30–32]. Some important issues regarding its impact on the results. Some important issues must be considered in a differential expression analysis pipeline, such as (i) access to the reference genome or transcriptome, (ii) quality of annotations, and (iii) number of samples. The following sections present some important considerations regarding these issues and the key steps involved in differential expression analysis.

2.1. Quality assessment and trimming

Quality assessment and sequence trimming are the first step of the analysis, which is also common to other analyses involving sequencing data, such as genome and transcript assemblies [33,34].

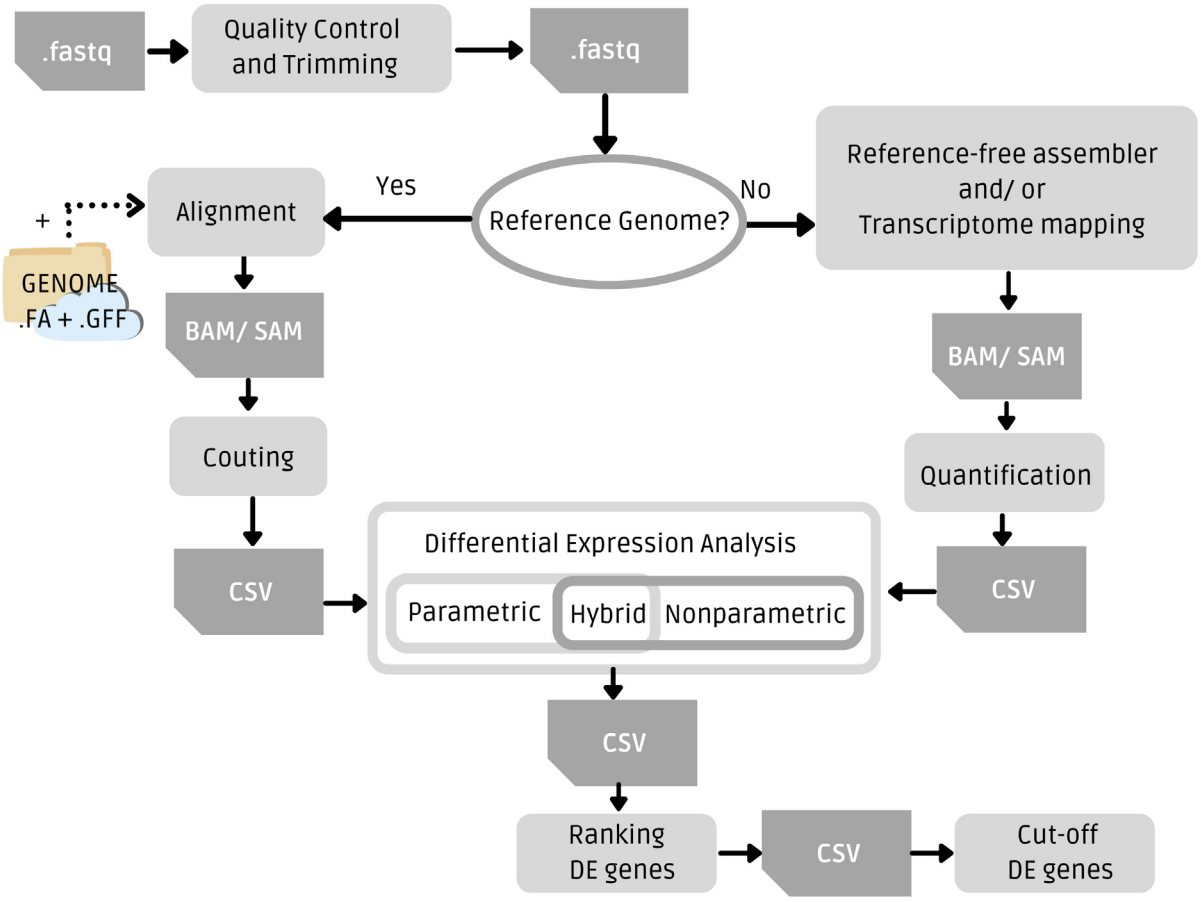

Fig. 1. Overview of the main steps in differential expression analysis from bulk RNA-seq data and file type generated in each step.

Quality assessment aims to identify and remove sequences identified as having low quality [35]. More specifically, trimming removes sequences from adapters used in sequencing. However, it is very common to find the term "cleaning" or "trimming" to refer to both steps, i.e., removal of both adapters and low-quality sequences. Table 1 shows the main methodologies and implementations (software) available in the literature for this task.

The sequence trimming process evaluates the quality of the reads[2] obtained in sequencing. Two characteristics are evaluated: the presence of adapter sequences (used in sample preparation) and the quality of the reads by the sequencer. As a result, reads that get a specific score within the defined scale [41] for the sequencing technique are selected. In this process, each base pair is evaluated from the quality score informed by the sequencer. It is possible to choose the cut-off score, besides defining the reads that should be kept or discarded.

Considering identification of the quality of the bases, the quality information in the FASTQ [42] file is used. In contrast, the removal of sequences identified as adapters is performed by similarity search [38].

A tool that can also be used in sequence quality analysis is FASTQC [43]. Despite its popularity in quality analysis, FASTQC is a tool that generates quality reports and does not perform the removal of low-quality sequences or adapters. It does, however, provide information to guide these cleanings. Table 1 provides information on tools that can be used as both functionalities: removal of adapters and low-quality sequences.

An additional point to observe in quality assessment are batch effects. Batch effects arise from differences between samples that are not rooted in the experimental design and can have various sources, spanning from different handlers or experiment locations to different batches of reagents and even biological artifacts such as growth location [44]. Various methods have been developed to detect and/or remove batch effects in genomics data, particularly RNA-seq data. For example, svaseq [45], Combat-Seq [46].

Table 1
Main methodologies for the removal of adapters and low-quality sequences. The methodologies are ordered chronologically considering the year of publication year.

Name	Year	Reference
Btrim	2011	[36]
CutAdapt	2011	[37]
Trimmomatic	2014	[38]
AdapterRemoval v2	2016	[27]
Atropos	2017	[39]
fastp	2018	[40]

Table 2
The principal sequence mapping methods. The methodologies are ordered chronologically considering the publication year.

Name	Algorithm	Year	Reference
BWA	Burrows–Wheeler Transform	2009	[48]
RUM	Burrows–Wheeler Transform	2011	[49]
Bowtie2	Burrows–Wheeler Transform	2012	[50]
BWBBLE	Burrows–Wheeler Transform	2013	[51]
STAR	Maximal Mappable Prefix (MMP)	2013	[52]
Tophat2	Burrows–Wheeler Transform	2013	[53]
HISAT2	Graph Based	2019	[54,55]

2.2. Alignment

After trimming the sequences, the mapping and counting of mapped reads occur. In this process, the aim is to identify how many reads are aligned to a genome region. The result is a read count table aligned to each gene. Among the difficulties in the mapping process are processing time and the computational capacity required, which have major challenges.

Table 2 shows the main mapping methodologies, where it can be observed that the methodologies recurrently apply Burrows-Wheeler Transformation algorithm [47].

The principal sequence mapping methodologies are briefly described bellow.

- **BWA** [48]: is based on the backward search associated with the Burrows-Wheeler transform. This method efficiently aligns short reads to large reference sequences, such as the human genome. BWA allows gaps and mismatches. In terms of memory optimization and search technique used, it performs a similar strategy adopted by the Bowtie method [56];
- **RUM** [49]: is based on an aggregation of methods, in which reads are mapped against the genome and transcriptome using the Bowtie tool. The reads not mapped by Bowtie are aligned to the genome with the BLAT tool [57]. The result of the mappings is presented in SAM format;
- **Bowtie 2** [50]: describes the application of the Burrows-Wheeler transform. Thus, for each read, the method performs four key steps (i) the extraction of "seeds" (sequence snippets) from the read and their reverse complement; (ii) the alignment of the seeds to the genome (reference), producing Burrows-Wheeler alignment bands; (iii) the selection of the bands randomly and repeatedly (weighted by priority), applying the displacement of each selected lane on the reference genome, using a method to compress the suffix matrix, and still effectively support the search for arbitrary patterns, called FM index [58], and applying the "walk-left" strategy of the FM index; and (iv) the resolution of similar alignments, observing the edges.
- **BWBBLE** [51]: based on mapping using multiple genomes as a reference, proposing the concept of a linear reference multi-genome. This concept incorporates the catalog of all known gene variants with a reference genome (e.g. SNPs, insertions, deletions and inversions), and uses a read alignment algorithm based on the Burrows-Wheeler transform;
- **STAR** [52]: proposed to specifically address many of the challenges of mapping RNA-seq data, such as junction detection and characterization and mapping sequences derived from non-contiguous genomic regions. In addition, it uses a novel strategy for junction alignments. The alignment comprises two major steps: seed search and clustering. The main idea behind the STAR seed search step is the sequential search for a Maximum Mappable Prefix (MMP). In the clustering step, STAR builds alignments of the entire read sequence by joining all the seeds that were aligned to the genome in the first step.
- **Tophat 2** [53]: directs attention to the problem of multiple alignments in junction reads. It uses the Bowtie 2 tool as a dependency. In the mapping step, reads aligned to over one exon are treated as unmapped. These reads are fragmented and aligned to the genome. Tophat 2 considers that the alignment distance between fragments may indicate possible splice regions. The genomic sequences around these junction sites are concatenated, and the resulting spliced sequences are treated as a set of potential transcription fragments. Any reads not mapped in the previous stages (or poorly mapped) are then realigned with Bowtie2 against this new transcript.
- **HISAT2** [55]: has a graph-based search strategy as its fundamental characteristic. HISAT2 starts the alignment process by generating a linear graph of the reference genome and then it

[2] In next-generation sequencing, a read refers to the DNA sequence from one fragment (a small section of DNA).

adds mutations, insertions and deletions as alternative paths of the graph. The authors claim that graph representations are more efficient in terms of memory utilization and/or alignment speed compared to the linear reference representation of genomes and alleles.

Considering the summarized presentation of the alignment methodologies, it is possible to highlight that each method presents a key strategy and is concerned with some specific alignment problems, which must be considered during the selection of the differential expression analysis pipeline.

The choice of mapping method is directly related to the objectives of a study, like those aiming to identify new transcripts. The mapping method should consider these findings or focus on this type of identification.

For analyses without a reference genome there are a few options: i) It is possible to generate a transcriptome assembly using the reads from the experiment itself and map it to the generated assembly, using tools such as tools such as RSEM [14], Cufflinks [28]; ii) Map against the transcriptome, using HISAT2 [55], TopHat [53] and/or kallisto [59]; iii) Map considering several genomes, with tools such as HISAT2.

As mapping methods evolve, we note the trend of using multiple genomes as a reference, as seen in the HISAT2 tool [55], presented as the successor to TopHat. More detailed information on mapping techniques can be obtained in the literature [60].

As a result of mapping, the methods produce mapping files in formats such as SAM (Sequence Alignment/Map format) and/or BAM (Binary Alignment/Map format) [48]. The resulting file contains all the mappings of a read and information such as alignment position and alignment score identified as MAPQ (acronym for MAPping Quality), which is presented in Phred (Phred-scaled) scale. Phred (Q) scale [61] is a quality indicator based on the probability of error in the alignment of a read at a reference position.

The approach in considering these mappings allows variations. It is possible to consider only mappings with Phred values above a threshold or reads that obtained unique mapping (in only one region of the reference genome). To support this task, counting tools are used, which process files in SAM and/or BAM format and produce a table with genes and the number of mapped reads. Some counting methods and their fundamental properties are presented in Section 2.4.

2.3. Pseudoalignment

Besides the Burrows-Wheeler transform and graph-based alignments, some strategies prioritize the balance between computational performance and the results produced. In this context, the methods that use the pseudoalignment strategy are applied.

It is important to note that most of the mapping methods described earlier could be applied to analyses in which the reference genome and its respective annotation are available. When an assembled genome is unavailable, there is a need for mapping without reference (Fig. 1). The transcriptome is assembled, and expression is estimated based on this assembly. In general, there are two types of approaches to transcriptome assembly: (i) genome-guided (or genome-based) assembly; and (ii) *de novo* assembly.

Some methods have been developed to generate transcript identification, such as Trinity [62] and Oases [63]. There are also strategies to estimate expression levels from data without a reference genome, including RSEM [14] and eXpress [64]. Several strategies perform both steps - transcript identification and estimation of expression levels - such as Scripture [65], Cufflinks [7] and StringTie [66], Salmon [67] and Kallisto [59].

One important method that uses pseudoalignment and transcript quantification is Salmon [67]. Salmon uses the quasi-mapping strategy [68], which requires a set of transcripts as a reference (which can be a reference assembly or *de novo*) or only the reads, to quantify the transcripts. This strategy comprises three steps (i) a simplified mapping model, (ii) a phase that estimates initial expression levels and model parameters, and (iii) a phase that refines the expression estimates. This inference procedure allows Salmon to build a probabilistic model based on the sequencing data, which includes more information, and then improves the conditional probability that a fragment is part of a transcript.

In pseudoalignment strategies, it is important to point out that the methods for transcript identification and quantification will report normalized quantification values as output and not counts of mapped reads. Therefore, the choice of the method for differential expression analysis should consider the type of data input expected (count or normalized values).

2.4. Counting

The count of mapped reads is the step to identify how many reads were mapped in each genomic region (reference). This step does not define which genes are differentially expressed; however, it represents the basis for the following analytical steps because, for each sequencing file presented to the mapper, a count of reads mapped to a particular gene will be produced. Consequently, there is a need for an annotation file of the reference genome. Tools like RSubread [69] and QuasR [70] are good options for this task.

The annotation file is usually a GFF (General Feature Format), which consisting of one row per feature: each row presents nine data columns. The file columns are separated by tabs and arranged in the following order: < seqname >, < source>, < feature >, < start >, < end >, < score >, < strand >, < frame > [*attributes*] [*comments*] (<> mandatory fields and [ ] optional fields) [71]. The GTF (General Transfer Format) is identical to the GFF in its version 2. Files have the same format and fields, but have different extensions.

GFF is a widely used text file format for storing genome annotations, describing sequence-based annotations. In addition, GFF files present genome features in a tab-delimited, single-feature-per-line table, making it ideal for use with multiple [72] data analysis pipelines. More details on sequencing data format are available in [71].

Before performing the count, it is necessary to consider the alignments options, including (i) read fully aligned to a gene; (ii) read partially aligned to a gene; (iii) read aligned to a junction (intron and exon); (iv) read aligned to a junction of exons (no alignment with intron); (v) read partially aligned to two genes, and (vi) read aligned to two genes. For this task some methods can be used together or alone, such as the HTSeq-count [73] which is part of the HTSeq framework, the BEDTools [74] toolkit and the featureCounts [75]. Implemented in R language, the Rsubread [69] method has functionality for alignment, quantification and analysis of RNA-seq data and can also be a counting option.

The choice of method and how to consider mappings in the count should be made based on the dataset and its properties. For eventual situations, where little prior knowledge is available, a comparison is recommended between the count in the most restrictive mode and most permissive mode of each software tool to define an adequate parameterization.

2.5. Normalization

This step of differential expression analysis aims to define which variations in the mapping count will be considered as differ-

ential expressions. If a gene has more mapped reads, it does not mean that it is differentially expressed because as each gene has an extension of base pairs, a smaller sequence in base pair sizes may have a proportionally smaller quantity of aligned reads. In this context, normalization is one of the basic steps for differential expression analysis.

The Reads per Kilobase per Million (RPKM) method was proposed in 2008 to generate accurate quantification of gene expression from RNA-seq [8] data. This method normalizes the expression of RNA-seq data using as a basis the total transcript size and the number of reads sequenced. In this way, RPKM allows small genes or transcripts not to be penalized compared to larger sequences.

The Fragments Per Kilobase Million (FPKM) normalization approach is analogous to RPKM, but supports one, two or more sequences from the same molecular source [7]. FPKM considers fragments and reads. In paired-end experiments, forward and reverse reads of the same sequence are considered as a fragment.

In 2012, the Transcripts Per Million (TPM) normalization was presented as a modification of the RPKM approach and aimed to remove RPKM bias [76].

The Trimmed Mean of M-values (TMM) [77] method aims to ensure that a gene with equal expression level in two samples is not detected as differentially expressed. To accurately estimate expression levels, it is necessary to quantify total RNA production, which cannot be estimated directly. However, the relative RNA production between two samples can be more easily determined by calculating the overall fold-change. The TMM method has been proposed as a robust and straightforward way of estimating RNA production. The TMM method is used by the edgeR package [11], in practice very similar to the method used by the DESeq package [6]. The results of these two methods are also similar in some points [78,79].

Although normalization methods like TMM, RPKM are classic and therefore widely used, important to note that recent methods also provide good results. The method: deviation based on the number of conditions, or cdev, to quantify the success of normalization, cdev measures how much one expression array differs from another. More information about this method can be found at [80]. Other options that can be considered are [81,82].

With normalized count data, it is possible to identify the expression variation evaluated under different conditions coherently. For these analyses, several methods are used to identify expression variations, assuming that these data follow a particular statistical distribution or not. A review of the principal methods, their characteristics and applications are presented in the next section, the focus of this review.

3. Methods for differentially expressed gene analysis

While the overview of differential expression and its steps is presented, this review focuses on presenting and discussing differential expression identification/inference methods and their properties.

In the identification of DEG, the aim is to infer which genes have decreased or increased transcriptional activity in certain experimental assays. Thus, the methods for identifying DEGs consider the quantification of RNA transcription. There are several ways to quantify transcripts (as presented in the Pseudoalignment Section 2.3) and define differential expression.

Contextualizing the differential expression analysis requires going back to 1991, when the identification of the transcriptional profile in mammals was proposed [83] using the Expressed Sequence Tag (EST) technique, based on the partial sequencing of cloned cDNAs to evaluate expression. Some years later, the Serial Analysis of Gene Expression (SAGE) method [84] was proposed and, in parallel, publications using the Microarray technique emerged [85], which became for years the most popular choice among transcriptional pattern studies.

In 2006 the first study with RNA-seq (high-throughput platform mRNA sequencing) data [86] was published, using Roche's 454 technology. This sequencing technique generates large numbers of short reads called high-throughput sequencers, in platforms that perform this type of sequencing.

Initially aiming to present an overview and the relations between the main methodologies and the computational tools available in the literature, we identified those most cited the early RNA-seq popularization in 2009. These methodologies were organized in temporal form, generating a timeline, which is presented in Fig. 2.

Confirmation of the popularity of the RNA-seq technology occurred in 2008, with the encouragement of a trio of papers [8,10,9], bringing novel approaches to the analyses. These studies did not specifically generate computational tools. However, they paved the way for the expression analysis from RNA-seq data. For this reason, these studies are identified in Fig. 2 in gray as a group of seminal studies.

Initial studies sought to build a consensus on the methodologies for DEG analysis. In this context, the study by [8] proposed the Enhanced Read Analysis of Gene Expression (ERANGE) software, while in the study by [10] the expression data were only normalized. The study by [9] performed a correlation with hybridization data.

Among the most popular technologies for generating RNA-seq data are Illumina Genome Analyzer and HiSeq [87], which enable the production of single or paired-end reads. In this way, RNA-seq also produces quality mappings, accurate identification of alternative splicing, and transcript reconstruction, among other types of studies. Regarding its predecessor, the Microarray, RNA-seq allows the study of new transcripts, offers higher resolution, better detection range and less technical variability. These factors have led to a major expansion of RNA-seq, and it has become the first choice in transcriptome analysis for many research groups [88].

With the popularization of the RNA-seq technique, computational tools (software) and proposals of new methodologies for DEG analysis emerged [121]. Since the proposal of the ERANGE tool [8], several other tools have been and continue to be proposed in the literature, as presented in Fig. 2. To evaluate gene expression data generated with RNA-seq, the starting point, after mapping and counting mapped reads, is usually the decision on the type of tool to be used to identify DEGs.

Expression analysis aims to identify genes whose mapped read count between two conditions has a significant difference. If the difference is greater than a random variation, the gene is identified as differentially expressed [89]. Differential expression analysis methods need to validate the statistical significance of this variation [90], which is done by adopting statistical tests such as Wald Test, Fischer's exact test, F-test, T-test, Likelihood ratio test, p-values and q-values, to cite just a few. The cut-off (threshold) adopted to indicate DEGs and non-DEGs, is decided from evaluating the distribution of the mapped reads.

In this review, methodologies were considered regarding the statistical distribution applied in the differential expression analysis, proposing a division into four groups: parametric, non-parametric, parametric or non-parametric, and hybrid. Only computational tools for DEG analysis, implemented and made available in software, were considered, regardless of the programming language or form of access. Methodologies without computational tools were not considered since this is an essential criterion for application in real problems. The adopted criteria to select the computational tools are describe in supplementary material.

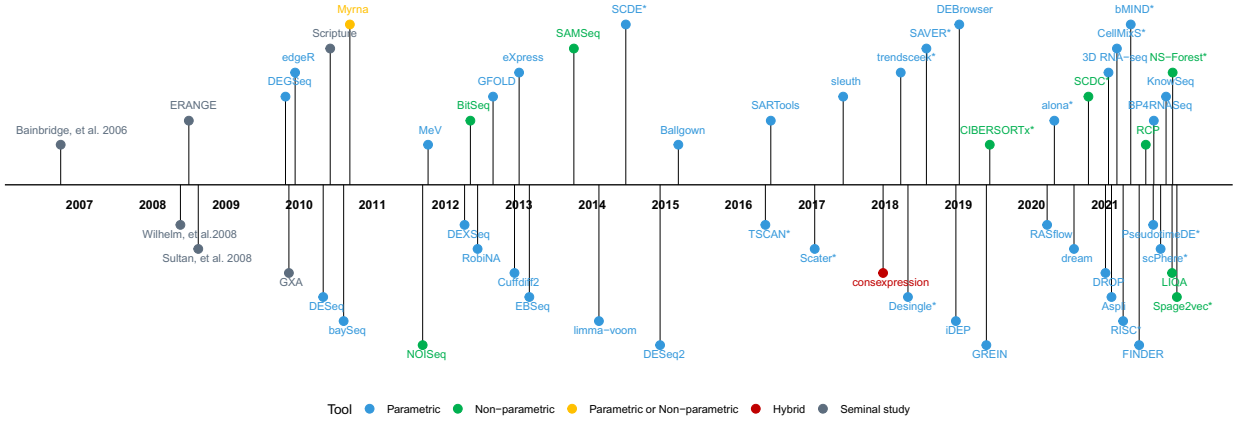

Fig. 2. Timeline with the main methodologies and computational tools for DEG analysis. In blue are shown the computational tools that use parametric methods to indicate differentially expressed genes. In green the non-parametric tools and in yellow the tools that allow the use of parametric or non-parametric methods. In red are the tools considered hybrid, for using parametric and non-parametric methods together in the indication of DEG. The items that are identified in grey are pioneer publications and/or publications that boosted the analysis. The distribution of methods in the timeline considers the month and year of publication. The items that contain * in the name indicate methods developed in the context of single-cell sequencing analysis.

All computational tools that adopt or describe the use of some parametric statistical distribution for the inference of DEGs, as well as tools that make partial or total use of this class of statistical distribution were considered parametric. The tools that do not use parametric distributions in their analysis and/or do not present any a priori statement about the data distribution for DEG inference were considered non-parametric. In this review, computational tools described as hybrid use parametric methods associated with non-parametric methods for DEG inference.

The explanation considering parametric, non-parametric and hybrid was adopted in this review to present the main methodologies in the literature in order to organize and contextualize them. Considering this scenario, the following sections will describe the respective particularities of parametric, non-parametric and hybrid methods.

3.1. Parametric methods

Parametric methodologies are those that start from the premise that data present a certain distribution. When using these tools, input data are considered to be distributed according to the statistical distribution adopted by the method, such as Negative Binomial, Poisson or Gaussian. This strategy is adopted by the first computational tools developed for DEG analysis [12,11,6].

Poisson distribution is adopted with some frequency to represent of RNA-seq data by computational tools [91–93]. Methods that use parametric analysis represent most of the tools developed and made available for use since the popularization of RNA-seq, as presented in Fig. 2.

Some considerations are essential when choosing parametric methods for DEG analysis. The parametric methods assume that the expression data is distributed according to a statistical distribution. Therefore, identification of DEGs in this context can be defined as the genes at the extremities of the distribution chosen, according to the experiment and sampling using a statistical significance value as the *p*-value.

Among the most widely used distributions for DEG analysis are the Negative Binomial [30,12,11,6], Poisson [94,95,91–93] and Gaussian [96–98] distributions.

The Poisson distribution is characterized by its suitability in application to technical replicate data [94]. On the other hand, data from biological replicates have higher variance and, for this reason, are best represented by a Negative Binomial distribution [99,15,16,100,101]. Gaussian distribution or normal distribution is a bell-shaped curve, and it is assumed that during any measurement, values will follow a normal distribution with an equal number of measurements above and below the mean value [102]. The Gaussian or normal distribution is used in simulation data [98], parameters estimation [97] and in expression analysis with Microarray data [96].

These and many other characteristics of parametric methods made them a suitable model to be followed, which generated a large volume of tools that use parametric methods of analysis, as visualized in Fig. 2, in which many of the initially proposed methods were the basis for most of the recently proposed methods.

3.2. Non-parametric methods

Non-parametric methods for DEG analysis arise in a context of innovation, with the need for solutions to the analysis of experiments with few replicates, in which the estimation of variance with precision becomes difficult. By observing the distribution of the groups of methods in Fig. 2, it is possible to notice that the main non-parametric computational tools for DEG analysis were presented between 2010 and 2013 [13,14,103,99].

Non-parametric methods include inference, non-parametric descriptive statistics, statistical models, and statistical tests. These methods do not determine a data distribution model a priori. The structure of the models is defined based on the distribution of the data, commonly known as data-driven. The non-parametric expression, associated with a tool for DEG analysis, shows that the number and nature of the parameters are adjusted according to the distribution of the data [104].

Among the principal tools identified in this review, NOIseq is a method that assesses differential gene expression between groups through the relationship between expression change and absolute expression differences [13,105]. NOIseq uses mapped, corrected, and normalized count reads and, models the noise distribution by contrasting the logarithm of expression change and absolute expression differences between groups. NOIseq defines a gene as differentially expressed between groups if the corresponding logarithm of expression change and absolute expression difference values have a high probability of being higher than the noise values [106].

Another tool for non-parametric DEG analysis is SAMseq [99]. For comparisons between groups, SAMseq uses the Wilcoxon

two-sample classification statistic. On the other hand, SAMseq considers the different depths through a re-sampling process in differential data analysis. In the Wilcoxon and FDR classification statistics, the null distribution is estimated using the mutation method [106].

3.3. Hybrids

For this review, hybrid methods were considered to be those approaches that associate parametric and non-parametric methodologies for the identify DEG.

The approach identified as hybrid in this review was the consexpression [18]. Consexpression approach is a pipeline for gene expression analysis that adopts the identification of DEGs from the joint analysis of nine tools. Among them, two are non-parametric: NOISeq [13] and SAMSeq [99], and seven are parametric [12,6,11,100,107,97,98]. Genes so indicated by the consensus of five or more tools are considered differentially expressed. In addition, the tools for the option of parametric or non-parametric analyses shown in Fig. 2 use one method according to the user's choice in an isolated manner.

3.4. Single-cell methods

Analysis of gene expression at the single-cell level provides insight into the oscillatory or non-linear behavior in asynchronous cells and reveals the cell-to-cell variability due to gene expression's stochastic nature [108]. Despite research possessing interest in the stochastic nature of gene expression and its implications for many years, the techniques available to quantify gene expression were limited in some early single-cell experiments [109,110].

Developing of a collection of fluorescent proteins, with unique biochemical characteristics, has enabled single-cell experiments [111]. The raw data generated through sequencing is processed to obtain molecular count arrays (count arrays) or alternatively read arrays (read arrays), depending on whether unique molecular identifiers (UMIs) have been incorporated into the single-cell library construction protocol.

The generated read or count matrices have the dimensions of barcodes X number of transcripts. At this point, the term "barcode" is used because all the reads associated with the same barcode may not correspond to the read of the same cell. A barcode may mislabel multiple cells (doublet) or may not mark any cells (empty droplet/well) [111].

Before evaluating DEG data, it is necessary to ensure that all barcodes correspond to cells; this is quality control in single-cell analysis. Three covariates are commonly used: count depth, genes per barcode and the fraction of mitochondrial gene counts per barcode [112].

As with bulk RNA-seq analyses, it is necessary to normalize single-cell data. The most commonly used normalization is counts per million or CPM normalization, which comes from bulk expression analysis and normalizes the count data using a size factor proportional to the count depth per cell. The normalization process in single-cell data is detailed in [113].

After normalizing the matrices, the data is typically transformed with $log(x+1)$. This transformation make the distance between transformed expression values represent log-fold changes, as many expression analysis tools created for bulk analysis [97,98,114] assume that data are normally distributed.

Some tools for expression analysis with single-cell data use as a basis (totally or partially) bulk analysis tools, such as the tools presented in Table 3.

The dependence of single-cell methods on bulk methods for expression analysis indicates the particular applicability of previously created methods for expression analysis with single-cell data. On the other hand, some tools were also identified that propose new methods for expression analysis with no dependency on previous method, such as SCDE* [116].

Table 3
Single-cell expression analysis methods. The methods are ordered chronologically considering the publication year.

Method	Based on previous methods	Year
trendsceek* [115]	DESeq2 [107]	2014
alona* [114]	limma-voom [97]	2020

4. Discussion

This review presents a temporal overview of the computational tools developed for differential gene expression analysis. This analysis is processed in several steps, as briefly described in the Methods for DEG analysis section. Then, based on the principal methods identified in the literature and addressed in this review, their properties and applications are presented.

In the context of the trimming step, the study by [41], the authors stated that by applying more aggressive parameters at the sequence trimming stage, over ten percent (10%) of the genes had significant changes in estimated expression levels. Still regarding quality, several studies report that by applying more aggressive and commonly used parameters such as Phred quality rate > 20 and read size $>$ 50bp, no significant differences in results are found [88,117], suggesting that a soft cut or even no cut at all results in the most biologically accurate gene expression estimates.

One study also indicate that most expression changes could be mitigated by imposing a minimum length filter after the cut-off, suggesting that differential gene expression may be driven primarily by spurious mapping of short reads [41].

Regarding the methods for differential expression analysis, this review identified that many of the available computational tools dedicated to DEG analysis are derived from previous methods or used them as part of their solutions. Therefore, a study was performed to recover the relationships between computational tools from the current literature. As a result, a network of interactions between the computational tools for DEG analysis was produced (Fig. 3).

The edges of the network were defined through the dependency between the tools. When a tool has an edge directed to another tool, it indicates that it uses the tool pointed at in the DEG identification. The relationship was identified based on the methodology described or the dependencies of the software.

It is possible to notice that few methodologies can be considered totally original (diamond-shaped node in Fig. 3), which shows the development of various computational methodologies for expression analysis are based on some specific methodologies. It is possible to notice that edgeR [11], DESeq2 [107] and limma-voom [97] methods are highlighted in this criterion, which are used by other methods as a basis for differential gene expression analysis. More specifically, considering only the principal methods available in the literature, edgeR is adopted by 14 methods other, DESeq2 by 12 methods, and, limma-voom by 10 methods.

Another interesting aspect that deserves to be highlighted is that the methodologies that adopt parametric statistical distributions in their analyses (in blue) are more abundant in the literature than non-parametric (in green) and hybrid methodologies (in red). The discrepancy between the number of parametric and non-parametric methodologies shows a scenario that needs improvements in the methods, since the tools used as the basis of their predecessors bring some improvement or functionality to the DEG analysis.

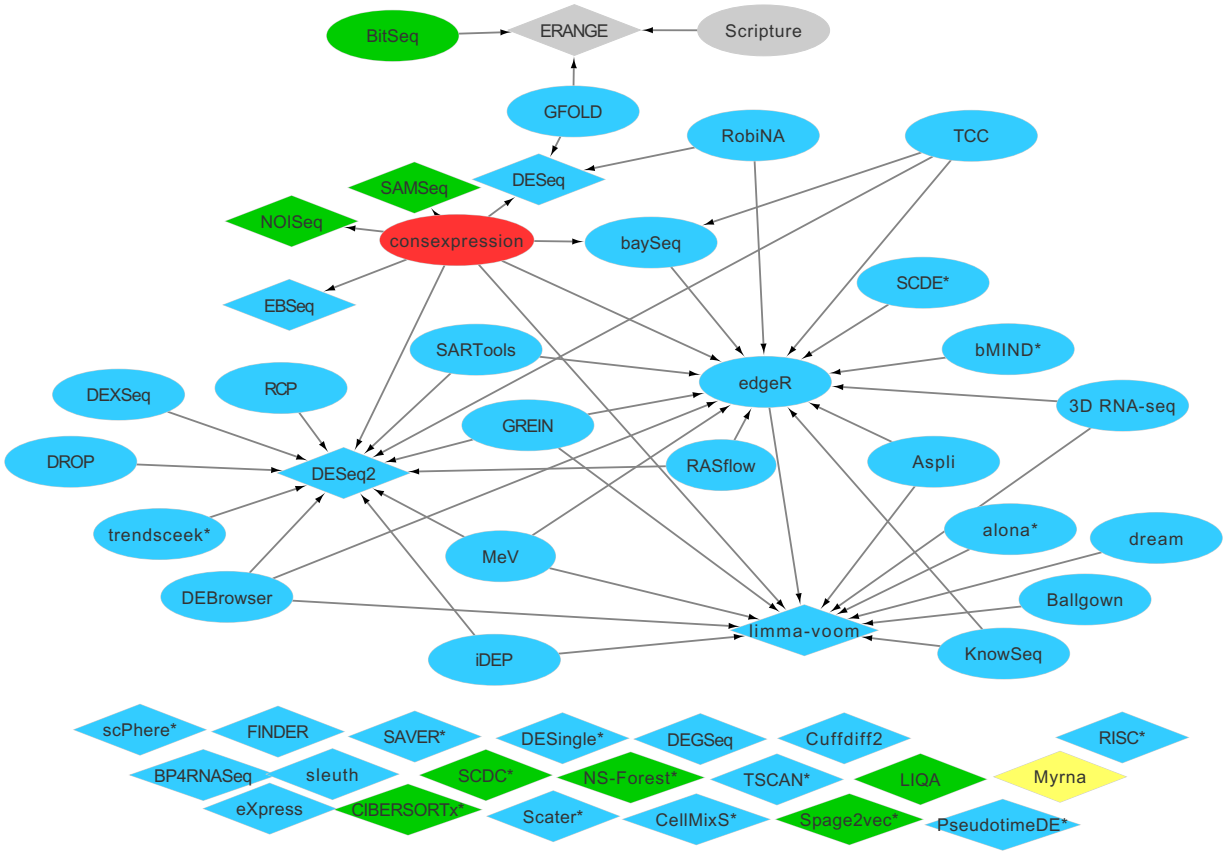

Fig. 3. Network of interactions between methodologies for the analysis of differential gene expression. The edges represent a tool that partially or totally uses another tool as a base, as an inheritance of the base method. The colors of the nodes show their method: blue (parametric), green (non-parametric), red (hybrid), yellow (parametric or non-parametric) and gray (seminal studies). The diamond-shaped nodes show tools that do not depend on any other. The items that contain * in the name indicate methods developed in the context of single-cell sequencing analysis.

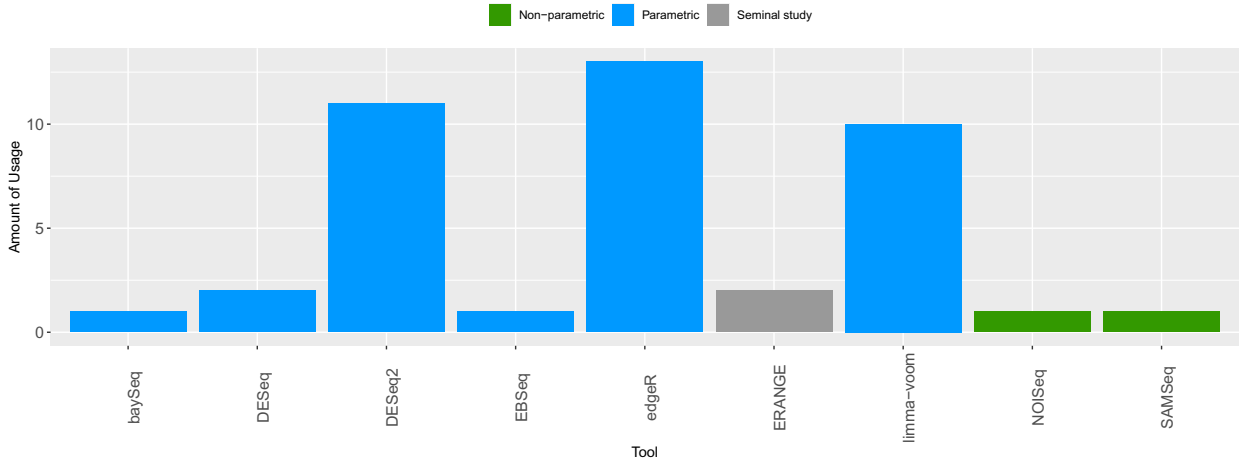

Fig. 4. Histogram of the use of the methods: The X-axis presents only the tools used as a base (dependency) for some other tool. The bars are colored according to the category of the tool, following the identification colors of the timeline (Fig. 2) and the interaction network (Fig. 3), where blue shows parametric methods, green non-parametric and gray "seminal studies". The items that contain * in the name indicate tools developed in the context of single-cell sequencing analysis.

Only one hybrid tools was identified by this review: consexpression [18]. This tool present the ability to combine the advantages of parametric and non-parametric methods.

It is also possible to observe that some methodologies have not been adopted by other methods and appear at the bottom without edges in this network. The most recent methodologies containing * in their name indicate methods developed in the context of single-cell sequencing analysis.

An issue that arises from the analysis of these interactions is that computational tools that implement non-parametric methods may represent a necessary solution, because, in the interaction network, non-parametric methodologies are used as a basis for other parametric and hybrid methods. This may indicate a direction in the development of more adequate analysis methods, and in this scenario, there is a certain convergence among parametric methods. Another point that draws attention is the opportunities to

explore the development of non-parametric (data-driven) and hybrid methodologies.

The observation regarding the number of methods among of the tools considered in this review is also relevant to point out the most used. The Fig. 4 depicts this analysis, indicating preference as edgeR, DESeq2 and limma-voom. The preference can be associated with maintenance, ease of use and the vast documentation made available by developers and community regarding these tools.

Among the 56 tools analyzed by this review, only 9 are used as a base for other studies, indicating that there are many tools, but most of them (24 tools) use a base method previously created. The data also show that most of the computational solutions for DEG analysis are based on the same analysis method.

Among the analyzed tools, few were identified that do not use other methods as a base (called original in this review). Among them, approximately half are not used as a basis for the developing other methods, some because they were developed to be used in a defined pipeline, such as Cuffdiff2 [16] and sleuth [98]. The complete list of tools analyzed in this review is available in supplementary material, including selected tools and references [121–149].

Single-cell RNA expression analysis (scRNA-seq) is revolutionizing organismal science, allowing unbiased identification of previously uncharacterized molecular heterogeneity at the cellular level [118]. Single-cell sequencing has become popular [119] and, in this review, some tools for the expression analysis using single-cell data are pointed out with an "*" in the name. The methods proposed in the literature for single-cell data analysis are based on methods proposed for the analysis of other more general RNA-Seq experiments (Fig. 4).

In this context, a comparison considering different methods for DEG analysis among scRNA-Seq populations is presented in [21,120]. Three different scenarios were considered, showing significant differences between the methods in terms of the number of genes identified with differential expression. Besides, it is reported that DEG methods specific for scRNA-Seq did not perform systematically better than non-specific methods (DESeq [6] and limma-voom [97]). These findings reinforce that scRNA-Seq specific DEG analysis tools use as a basis the tools developed for bulk RNA-Seq analysis.

The alona* [114] tool uses limma-voom [97], while the trendsceek* [115] uses DESeq2 [107].

Although single-cell expression analysis is an important topic, it is still in its infancy and has its own characteristics. This topic has also been supported by previous studies such as [120]. Clearly, there is a need that future studies can and should include analytical techniques for single-cell data as an emerging topic, and that deserves a dedicated review.

5. Conclusion

A review of computational methods for DEG analysis since the popularization of RNA-seq in 2009 was presented, reporting and discussing the most important tools in the current literature, contributing to the understanding of the steps involved, and the methods available, along with their particularities and applications.

The development of expression analysis pipelines by including new functionalities has become a trend. Avoiding the need for many replicates, required by RNA-seq sequencing, and yet maintaining satisfactory results, is a challenge that deserves attention in the developing DEG analysis methodologies.

The review presents fundamental concepts and computational tools for expression analysis, in which it is possible to identify the tendency to reuse methodologies in the development of computational tools (software) and also, to incorporate new functionalities to the existing software.

Therefore, it was possible to verify that the context of parametric methodologies presents a more stable scenario, revealing convergence with the methods available in the literature. In contrast, this review points out a challenge in developing non-parametric (data-driven) and hybrid methodologies for DEG analysis.

In conclusion, this review brings a discussion about different methodologies applied in differential expression analysis. In addition, it contributes with notes and directions to the community to clarify some aspects of the analysis and serves as support to beginners data analysts in bioinformatics at beginning of their careers.

Author's contributions

Costa-Silva J., Domingues D. S., Menotti D., Hungria M. and Lopes F. M. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors have read and approved the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

INCT - Plant-Growth Promoting Microorganisms for Agricultural Sustainability and Environmental Responsibility (CNPq 465133/2014-4, Fundação Araucária-STI); Fundação Araucária e do Governo do Estado do Paraná SETI (Grant 035/2019, 138/2021). Fundação Araucária - NAPI de Bioinformática (grant PDI 66/2021); CNPq (312823/2019-3, 308879/2020-1 and 440412/2022-6); FAPESP (2016/10896-0, 2018/08042-8 and 2019/15477-3).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.11.051.

References

[1] Yingying Cao, Simo Kitanovski, and Daniel Hoffmann. intepareto: an r package for integrative analyses of rna-seq and chip-seq data. BMC Genom, 21:802, 12 2020.
[2] Wenbin Guo, Nikoleta A. Tzioutziou, Gordon Stephen, Iain Milne, Cristiane P. G. Calixto, Robbie Waugh, John W.S. Brown, and Runxuan Zhang. 3d rna-seq: a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of rna-seq data for biologists. RNA Biol, pages 1–14, 12 2020.
[3] Jiménez-Jacinto Verónica, Sanchez-Flores Alejandro, Vega-Alvarado Leticia. Integrative differential expression analysis for multiple experiments (ideamex): A web server tool for integrated rna-seq data analysis. Front Genet 2019;10(279):3.
[4] Kumar Shailesh, Vo Angie Duy, Qin Fujun, Li Hui. Comparative assessment of methods for the fusion transcripts detection from rna-seq data. Sci Rep 2016;6(1–10):2.
[5] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. Nature Rev Genet, 20:631–656, 11 2019.
[6] Anders Simon, Huber Wolfgang. Differential expression analysis for sequence count data. Nature Proc 2010.
[7] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol, 28:511–515, 2010.
[8] Mortazavi Ali, Williams Brian A, McCue Kenneth, Schaeffer Lorian, Wold Barbara. Mapping and quantifying mammalian transcriptomes by rna-seq. Nat. Methods 2008;5(621–628):7.

[9] Brian T Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature, 453:1239–1243, 2008.
[10] M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 321:956–960, 8 2008.
[11] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26:139–140, 1 2010.
[12] Hardcastle Thomas J, Kelly Krystyna A. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinform 2010;11:422.
[13] Sonia Tarazona, F. Garcia-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in rna-seq: A matter of depth. Genome Res, 21:2213–2223, 12 2011.
[14] Li Bo, Dewey Colin N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinform 2011;12(323):1.
[15] Eleanor A. Howe, Raktim Sinha, Daniel Schlauch, and John Quackenbush. Rna-seq analysis in mev. Bioinformatics, 27:3209–3210, 11 2012.
[16] Cole Trapnell, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. Differential analysis of gene regulation at transcript resolution with rna-seq. Nature Biotechnology, 31:46–53, 1 2013.
[17] Frazee Alyssa C, Pertea Geo, Jaffe Andrew E, Langmead Ben, Salzberg Steven L, Leek Jeffrey T. Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat Biotechnol 2015;33(243–246):3.
[18] Costa-Silva Juliana, Domingues Douglas, Lopes Fabricio Martins. Rna-seq differential expression analysis: An extended review and a software tool. PLOS ONE 2017;12(e0190152):12.
[19] Seyednasrollah Fatemeh, Laiho Asta, Elo Laura L. Comparison of software packages for detecting differential expression in rna-seq studies. Briefings Bioinform 2015;16(59–70):1.
[20] Xun Gu. Statistical detection of differentially expressed genes based on rna-seq: from biological to phylogenetic replicates. Briefings Bioinform 2016;17 (243–248):3.
[21] Jaakkola Maria K, Seyednasrollah Fatemeh, Mehmood Arfa, Elo Laura L. Comparison of methods to detect differentially expressed genes between single-cell populations. Briefings Bioinform 2017;18(735–743):9.
[22] Adam McDermaid, Brandon Monier, Jing Zhao, Bingqiang Liu, and Qin Ma. Interpretation of differential gene expression results of rna-seq data: Review and integration. Briefings Bioinform, 20:2044–2054, 11 2019.
[23] Daniel Toro-Domínguez, Juan Antonio Villatoro-Garciá, Jordi Martorell-Marugán, Yolanda Román-Montoya, Marta E. Alarcón-Riquelme, and Pedro Carmona-Saéz. A survey of gene expression meta-analysis: methods and applications. Briefings Bioinform, 22:1694–1705, 3 2021.
[24] Jürgen Jänes, Fengyuan Hu, Alexandra Lewin, and Ernest Turro. A comparative study of rna-seq analysis strategies. Briefings Bioinform, 16:932–940, 11 2015.
[25] Shancheng Ren, Zhiyu Peng, Jian Hua Mao, Yongwei Yu, Changjun Yin, Xin Gao, Zilian Cui, Jibin Zhang, Kang Yi, Weidong Xu, Chao Chen, Fubo Wang, Xinwu Guo, Ji Lu, Jun Yang, Min Wei, Zhijian Tian, Yinghui Guan, Liang Tang, Chuanliang Xu, Linhui Wang, Xu Gao, Wei Tian, Jian Wang, Huanming Yang, Jun Wang, and Yinghao Sun. Rna-seq analysis of prostate cancer in the chinese population identifies recurrent gene fusions, cancer-associated long noncoding rnas and aberrant alternative splicings. Cell Res, 22:806–821, 5 2012.
[26] Weirong Cui, Yulan Qian, Xiaoke Zhou, Yuxin Lin, Junfeng Jiang, Jiajia Chen, Zhongming Zhao, and Bairong Shen. Discovery and characterization of long intergenic non-coding rnas (lincrna) module biomarkers in prostate cancer: An integrative analysis of rna-seq data. BMC Genom, 16:1–10, 6 2015.
[27] Schubert Mikkel, Lindgreen Stinus, Orlando Ludovic. Adapterremoval v2: Rapid adapter trimming, identification, and read merging. BMC Res Notes 2016;9(88):2.
[28] Trapnell Cole, Roberts Adam, Goff Loyal, Pertea Geo, Kim Daehwan, Kelley David R, Pimentel Harold, Salzberg Steven L, Rinn John L, Pachter Lior. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nat Protoc 2012;7:562–78.
[29] Zong Hong Zhang, Dhanisha J Jhaveri, Vikki M Marshall, Denis C Bauer, Janette Edson, Ramesh K Narayanan, Gregory J Robinson, Andreas E Lundberg, Perry F Bartlett, Naomi R Wray, et al. A comparative study of techniques for differential expression analysis on rna-seq data. PloS one, 9:e103207, 2014.
[30] Robinson Mark D, Smyth Gordon K. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007;23:2881–7.
[31] Lauren M. McIntyre, Kenneth K Lopiano, Alison M Morse, Victor Amin, Ann L Oberg, Linda J Young, and Sergey V Nuzhdin. Rna-seq: Technical variability and sampling. BMC Genom, 12, 2011.
[32] Hansen Kasper D, Brenner Steven E, Dudoit Sandrine. Biases in illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 2010;38:e131.
[33] Wang Zhong, Gerstein Mark, Snyder Michael. Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10(57–63):1.
[34] MacManes Matthew D. On the optimal trimming of high-throughput mrna sequence data. Front Genet 2014;5.
[35] Li Xing, Nair Asha, Wang Shengqin, Wang Liguo. Quality control of rna-seq experiments. Methods Mol Biol 2015;1269:137–46.
[36] Kong Yong. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics 2011;98(152–3):8.
[37] Martin Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 2011;17:10–2.
[38] Bolger Anthony M, Lohse Marc, Usadel Bjoern. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 2014;30(2114–2120):8.
[39] Didion John P, Martin Marcel, Collins Francis S. Atropos: Specific, sensitive, and speedy trimming of sequencing reads. PeerJ 2017;2017.
[40] Chen Shifu, Zhou Yanqing, Chen Yaru, Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics (Oxford, England) 2018;34:i884–90.
[41] Williams Claire R, Baccarella Alyssa, Parrish Jay Z, Kim Charles C. Trimming of sequence reads alters rna-seq gene expression estimates. BMC Bioinform 2016;17:103.
[42] Peter J.A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. Nucleic Acids Res, 38:1767–1771, 12 2009.
[43] Simon Andrews. Babraham bioinformatics – fastqc a quality control tool for high throughput sequence data, 2010.
[44] Sprang Maximilian, Andrade-Navarro Miguel A, Fontaine Jean Fred. Batch effect detection and correction in rna-seq data using machine-learning-based automated assessment of quality. BMC Bioinform 2022;23(1–15):7.
[45] Leek Jeffrey T. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res 2014;42:161.
[46] Zhang Xiaokang, Jonassen Inge. Rasflow: An rna-seq analysis workflow with snakemake. BMC Bioinform 2020;21(1–9):3.
[47] Burrows Michael, Wheeler David. A block-sorting lossless data compression algorithm. Citeseer: In Digital SRC Research Report; 1994.
[48] Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009;25(1754–1760):7.
[49] Gregory R. Grant, Michael H. Farkas, Angel D. Pizarro, Nicholas F. Lahens, Jonathan Schug, Brian P. Brunk, Christian J. Stoeckert, John B. Hogenesch, and Eric A. Pierce. Comparative analysis of rna-seq alignment algorithms and the rna-seq unified mapper (rum). Bioinformatics, 27:2518–2528, 9 2011.
[50] Langmead Ben, Salzberg Steven L. Fast gapped-read alignment with bowtie 2. Nature Methods 2012;9:357–9.
[51] Huang Lin, Popic Victoria, Batzoglou Serafim. Short read alignment with populations of genomes. Bioinformatics 2013;29:i361–70.
[52] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: Ultrafast universal rna-seq aligner. Bioinformatics, 29:15–21, 1 2013.
[53] Kim Daehwan, Pertea Geo, Trapnell Cole, Pimentel Harold, Kelley Ryan, Salzberg Steven L. Tophat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;14 (R36):4.
[54] Kim Daehwan, Langmead Ben, Salzberg Steven L. Hisat: a fast spliced aligner with low memory requirements. Nat Methods 2015;12(357–360):4.
[55] Kim Daehwan, Paggi Joseph M, Park Chanhee, Bennett Christopher, Salzberg Steven L. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nature Biotechnol 2019;37(907–915):8.
[56] Langmead Ben, Trapnell Cole, Pop Mihai, Salzberg Steven L, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol 2009;10:R25.
[57] Kent WJ. Blat—the blast-like alignment tool. Genome Res 2002;12(656–664):3.
[58] Ferragina P, Manzini G. Opportunistic data structures with applications. In: Proceedings 41st Annual Symposium on Foundations of Computer Science. p. 390–8.
[59] Bray Nicolas L, Pimentel Harold. Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. Nat Biotechnol 2016;34(525–527):5.
[60] Stefan Canzar and Steven L. Salzberg. Short read mapping: An algorithmic tour. Proc IEEE, 105:436–458, 3 2017.
[61] Ewing Brent, Green Phil. Base-calling of automated sequencer traces using phred. ii. error probabilities. Genome Res 1998;8(186–194):3.
[62] Grabherr Manfred G, Haas Brian J, Yassour Moran, Levin Joshua Z, Dawn A, et al. Thompson,Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Full-length transcriptome assembly from rna-seq data without a reference genome. Nature Biotechnol 2011;29:644–52.
[63] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. Bioinformatics, 28:1086–1092, 4 2012.
[64] Roberts Adam, Pachter Lior. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 2013;10(71–73):1.
[65] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. Nature Biotechnol, 28:503–510, 5 2010.
[66] Pertea Mihaela, Kim Daehwan, Pertea Geo M, Leek Jeffrey T, Salzberg Steven L. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. Nat Protoc 2016;11(1650–1667):9.
[67] Patro Rob, Duggal Geet, Love Michael I, Irizarry Rafael A, Kingsford Carl. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 2017;14(417–419):4.
[68] Srivastava Avi, Sarkar Hirak, Gupta Nitish, Patro Rob. Rapmap: A rapid, sensitive and accurate tool for mapping rna-seq reads to transcriptomes. Bioinformatics 2016;32:i192–200.

[69] Yang Liao, Gordon K Smyth, and Wei Shi. The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads. Nucl Acids Res, 47, 2019.
[70] Dimos Gaidatzis, Anita Lerch, Florian Hahne, and Michael B. Stadler. Quasr: quantification and annotation of short reads in r. Bioinformatics (Oxford, England), 31:1130–1132, 4 2015.
[71] Zhang Hongen. Overview of sequence data formats. Methods Mol Biol 2016;1418:3–17.
[72] Rastogi Achal, Gupta Dinesh. Gff-ex: a genome feature extraction package. BMC Res Notes 2014;7:315.
[73] Anders S, Pyl PT, Huber W. Htseq–a python framework to work with high-throughput sequencing data. Bioinformatics 2015;31(166–169):1.
[74] Quinlan Aaron R, Hall Ira M. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(841–842):3.
[75] Liao Y, Smyth GK, Shi W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30 (923–930):4.
[76] Wagner Günter P, Kin Koryu, Lynch Vincent J. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. Theory Biosci 2012;131:281–5.
[77] Robinson Mark D, Oshlack Alicia, et al. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol 2010;11:R25.
[78] Li Peipei, Piao Yongjun, Shon Ho S, Ryu Keun H. Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. BMC Bioinform 2015;16:347.
[79] Soneson Charlotte, Delorenzi Mauro. A comparison of methods for differential expression analysis of rna-seq data. BMC Bioinform 2013;14:91.
[80] Tran Diem-Trang, Might Matthew. cdev: a ground-truth based measure to evaluate rna-seq normalization performance. PeerJ 2021;9(e12233):10.
[81] Yance Feng and Lei M. Li. Muren: a robust and multi-reference approach of rna-seq transcript normalization. BMC Bioinform, 22:386, 12 2021.
[82] Yan Zhou, Bin Yang, Junhui Wang, Jiadi Zhu, and Guoliang Tian. A scaling-free minimum enclosing ball method to detect differentially expressed genes for rna-seq data. BMC Genom, 22:479, 12 2021.
[83] Mark D Adams, Jenny M Kelley, Jeannine D Gocayne, Mark Dubnick, Mihael H Polymeropoulos, Hong Xiao, Carl R Merril, Andrew Wu, Bjorn Olde, Ruben F Moreno, et al. Complementary dna sequencing: expressed sequence tags and human genome project. Science, 252:1651–1656, 1991.
[84] Velculescu Victor E, Zhang Lin, Vogelstein Bert, Kinzler Kenneth W. Serial analysis of gene expression. Science 1995;270:484–7.
[85] Schena Mark, Shalon Dari, Davis Ronald W, Brown Patrick O. Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science 1995;270:467–70.
[86] Matthew N Bainbridge, René L Warren, Martin Hirst, Tammy Romanuik, Thomas Zeng, Anne Go, Allen Delaney, Malachi Griffith, Matthew Hickenbotham, Vincent Magrini, et al. Analysis of the prostate cancer cell line lncap transcriptome using a sequencing-by-synthesis approach. BMC Genom, 7:246, 2006.
[87] McGettigan Paul A. Transcriptomics in the rna-seq era. Curr Opin Chem Biol 2013;17:4–11.
[88] Luis A. Corchete, Elizabeta A. Rojas, Diego Alonso-López, Javier De Las Rivas, Norma C. Gutiérrez, and Francisco J. Burguillo. Systematic comparison and assessment of rna-seq procedures for gene expression quantitative analysis. Sci Rep, 10:19737, 12 2020.
[89] Joshi Reema, Sarmah Rosy. Survey of methods used for differential expression analysis on rna seq data. In: International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making. Springer; 2019. p. 226–39.
[90] Li Chung-I, Samuels David C, Zhao Ying-Yong, Shyr Yu, Guo Yan. Power and sample size calculations for high-throughput sequencing-based experiments. Briefings Bioinform 2017;19(6):1247–55.
[91] Wang Likun, Feng Zhixing, Wang Xi, Wang Xiaowo, Zhang Xuegong. Degseq: an r package for identifying differentially expressed genes from rna-seq data. Bioinformatics 2010;26(136–138):1.
[92] Langmead Ben, Hansen Kasper D, Leek Jeffrey T. Cloud-scale rna-sequencing differential expression analysis with myrna. Genome Biol 2010;11(1–11):8.
[93] Jianxing Feng, Clifford A. Meyer, Qian Wang, Jun S. Liu, X. Shirley Liu, and Yong Zhang. Gfold: a generalized fold change for ranking differentially expressed genes from rna-seq data. Bioinformatics, 28:2782–2788, 11 2012.
[94] Marioni John C, Mason Christopher E, Mane Shrikant M, Stephens Matthew, Gilad Yoav. Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008;18(1509–1517):7.
[95] Bullard James H, Purdom Elizabeth, Hansen Kasper D, Dudoit Sandrine. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. BMC Bioinform 2010;11:94.
[96] Bloom Joshua S, Khan Zia, Kruglyak Leonid, Singh Mona, Caudy Amy A. Measuring differential gene expression by short read sequencing: Quantitative comparison to 2-channel gene expression microarrays. BMC Genom 2009;10(221):5.
[97] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res, page gkv007, 2015.
[98] Pimentel Harold, Bray Nicolas L, Puente Suzette, Melsted Páll, Pachter Lior. Differential analysis of rna-seq incorporating quantification uncertainty. Nat Methods 2017;14(687–690):6.
[99] Li Jun, Tibshirani Robert. Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. Stat Methods Med Res 2013;22:519–36.
[100] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart M G Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. Bioinformatics, 29:1035–1043, 2013.
[101] Varet Hugo, Brillet-Guéguen Loraine, Coppée Jean-Yves, Dillies Marie-Agnès. Sartools: A deseq2- and edger-based r pipeline for comprehensive differential analysis of rna-seq data. PLOS ONE 2016;11(e0157022):6.
[102] Dasgupta Amitava, Wahed Amer. Chapter 4 - laboratory statistics and quality control. In: Dasgupta Amitava, Wahed Amer, editors. Clin Chem. Immunology and Laboratory Quality Control. San Diego: Elsevier; 2014. p. 47–66.
[103] Glaus Peter, Honkela Antti, Rattray Magnus. Identifying differentially expressed transcripts from rna-seq data with biological variation. Bioinformatics 2012;28(1721–1728):7.
[104] Penfold Christopher A, Buchanan-Wollaston Vicky, Denby Katherine J, Wild David L. Nonparametric bayesian inference for perturbed and orthologous gene regulatory networks. Bioinformatics 2012;28:i233–41.
[105] Sonia Tarazona, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. Data quality aware analysis of differential expression in rna-seq with noiseq r/bioc package. Nucl Acids Res, page gkv711, 2015.
[106] Dongmei Li. Statistical methods for rna sequencing data analysis. Computat Biol, pages 85–99, 11 2019.
[107] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol, 15:550, 12 2014.
[108] Longo Diane, Hasty Jeff. Dynamics of single–cell gene expression. Mol Syst Biol 2006;2(64):1.
[109] Maloney Peter C, Rotman Boris. Distribution of suboptimally induced β-d-galactosidase in escherichia coli. J Mol Biol 1973;73:1.
[110] Spudich John L, Koshland DE. Non-genetic individuality: chance in the single cell. Nature 1976;262(467–471):8.
[111] Luecken Malte D, Theis Fabian J. Current best practices in single–cell rna–seq analysis: a tutorial. Mol Syst Biol 2019;15(e8746):6.
[112] Jonathan A Griffiths, Antonio Scialdone, and John C Marioni. Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol, 14(4):e8046, 2018.
[113] Peng Jiajie, Han Lu, Shang Xuequn. A novel method for predicting cell abundance based on single-cell rna-seq data. BMC Bioinform 2021;22(1–15):8.
[114] Franzén Oscar, Björkegren Johan LM. alona: a web server for single-cell rna-seq analysis. Bioinformatics 2020;36(3910–3912):6.
[115] Edsgärd Daniel, Johnsson Per, Sandberg Rickard. Identification of spatial expression trends in single-cell gene expression data. Nat Methods 2018;15 (339–342):4.
[116] Kharchenko Peter V, Silberstein Lev, Scadden David T. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014;11(740–742):5.
[117] Liao Yang, Shi Wei. Read trimming is not required for mapping and quantification of rna-seq reads at the gene level. NAR Genom Bioinform 2020;2:9.
[118] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol, 16:241, 11 2015.
[119] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of gene expression states from single-cell rna-seq data. Nature Biotechnol, pages 1–9, 4 2021.
[120] Kadota Koji, Shimizu Kentaro. Commentary: A systematic evaluation of single cell rna-seq analysis pipelines. Front Genet 2020;11(941):9.
[121] Misha Kapushesky, Ibrahim Emam, Ele Holloway, Pavel Kurnosov, Andrey Zorin, James Malone, Gabriella Rustici, Eleanor Williams, Helen Parkinson, and Alvis Brazma. Gene expression atlas at the european bioinformatics institute. Nucl Acids Res, 38:D690–D698, 11 2009.
[122] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. Nature Proc, pages 1–1, 4 2012.
[123] Marc Lohse, Anthony M. Bolger, Axel Nagel, Alisdair R. Fernie, John E. Lunn, Mark Stitt, and Björn Usadel. Robina: A user-friendly, integrated software solution for rna-seq-based transcriptomics. Nucleic Acids Res, 40:W622–W627, 7 2012.
[124] Sun Jianqiang, Nishiyama Tomoaki, Shimizu Kentaro, Kadota Koji. Tcc: An r package for comparing tag count data with robust normalization strategies. BMC Bioinform 2013;14(1–14):7.
[125] Law Charity W, Chen Yunshun, Shi Wei, Smyth Gordon K. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. Genome Biol 2014;15:1.
[126] Ji Zhicheng, Ji Hongkai. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. Nucleic Acids Res 2016;44(e117):7.
[127] Davis J McCarthy, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. Bioinformatics, 33:1179–1186, 4 2017.
[128] Miao Zhun, Deng Ke, Wang Xiaowo, Zhang Xuegong. Desingle for detecting three types of differential expression in single-cell rna-seq data. Bioinformatics 2018;34(3223–3224):9.
[129] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I. Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang.

Saver: gene expression recovery for single-cell rna sequencing. Nature Methods, 15:539–542, 7 2018.
[130] Steven Xijin Ge, Eun Wo Son, and Runan Yao. idep: An integrated web application for differential expression and pathway analysis of rna-seq data. BMC Bioinformatics, 19:1–24, 12 2018.
[131] Alper Kucukural, Onur Yukselen, Deniz M. Ozata, Melissa J. Moore, and Manuel Garber. Debrowser: Interactive differential expression analysis and visualization tool for count data 06 biological sciences 0604 genetics 08 information and computing sciences 0806 information systems. BMC Genom, 20:6, 1 2019.
[132] Aaron M. Newman, Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust, Mohammad S. Esfahani, Bogdan A. Luca, David Steiner, Maximilian Diehn, and Ash A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnol, 37:773–782, 5 2019. Using Seurat28, clusters were identified by (supl.1).
[133] Naim Al Mahi, Mehdi Fazel Najafabadi, Marcin Pilarczyk, Michal Kouril, and Mario Medvedovic. Grein: An interactive web platform for re-analyzing geo rna-seq data. Sci Rep, 9:1–9, 12 2019.
[134] Hoffman Gabriel E, Roussos Panos. Dream: powerful differential expression analysis for repeated measures designs. Bioinformatics 7 2020..
[135] Partel Gabriele, Wählby Carolina. Spage2vec: Unsupervised representation of localized spatial gene expression signatures. FEBS J 2021;288(1859–1870):3.
[136] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M. Perou, Fei Zou, and Yuchao Jiang. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. Briefings Bioinform, 22:416–427, 1 2021.
[137] Vicente A. Yépez, Christian Mertes, Michaela F. Müller, Daniela Klaproth-Andrade, Leonhard Wachutka, Laure Frésard, Mirjana Gusic, Ines F. Scheller, Patricia F. Goldberg, Holger Prokisch, and Julien Gagneur. Detection of aberrant gene expression events in rna sequencing data. Nature Protocols, 16:1276–1296, 2 2021.
[138] Estefania Mancini, Andres Rabinovich, Javier Iserte, Marcelo Yanovsky, and Ariel Chernomoretz. Aspli: integrative analysis of splicing landscapes through rna-seq assays. Bioinformatics, 37:2609–2616, 9 2021. Originial Method.
[139] Almut Lütge, Joanna Zyprych-Walczak, Urszula Brykczynska Kunzmann, Helena L. Crowell, Daniela Calini, Dheeraj Malhotra, Charlotte Soneson, and Mark D. Robinson. Cellmixs: quantifying and visualizing batch effects in single-cell rna-seq data. Life Sci Alliance, 4, 6 2021.
[140] Liu Yang, Wang Tao, Zhou Bin, Zheng Deyou. Robust integration of multiple single-cell rna sequencing datasets using a single reference space. Nat Biotechnol 2021;39(877–884):3. Pag 11 fim da coluna 1 ferramentas utilizadas.
[141] Jiebiao Wang, Kathryn Roeder, and Bernie Devlin. Bayesian estimation of cell type–specific gene expression with prior derived from single-cell data. Genome Res, 31:1807–1818, 10 2021.
[142] Sagnik Banerjee, Priyanka Bhandary, Margaret Woodhouse, Taner Z. Sen, Roger P. Wise, and Carson M. Andorf. Finder: an automated software package to annotate eukaryotic genes from rna-seq data and associated protein sequences. BMC Bioinformatics, 22:1–26, 4 2021. STAR In addition to constructing genes from expression data, FINDER uses BRAKER2 [65] to predict genes de novo.
[143] Eliah G. Overbey, Amanda M. Saravia-Butler, Zhe Zhang, Komal S. Rathi, Homer Fogle, Willian A. da Silveira, Richard J. Barker, Joseph J. Bass, Afshin Beheshti, Daniel C. Berrios, Elizabeth A. Blaber, Egle Cekanaviciute, Helio A. Costa, Laurence B. Davin, Kathleen M. Fisch, Samrawit G. Gebre, Matthew Geniza, Rachel Gilbert, Simon Gilroy, Gary Hardiman, Raúl Herranz, Yared H. Kidane, Colin P.S. Kruse, Michael D. Lee, Ted Liefeld, Norman G. Lewis, J. Tyson McDonald, Robert Meller, Tejaswini Mishra, Imara Y. Perera, Shayoni Ray, Sigrid S. Reinsch, Sara Brin Rosenthal, Michael Strong, Nathaniel J. Szewczyk, Candice G.T. Tahimic, Deanne M. Taylor, Joshua P. Vandenbrink, Alicia Villacampa, Silvio Weging, Chris Wolverton, Sarah E. Wyatt, Luis Zea, Sylvain V. Costes, and Jonathan M. Galazka. Nasa genelab rna-seq consensus pipeline: Standardized processing of short-read rna-seq data. iScience, 24:102361, 4 2021.
[144] Dongyuan Song and Jingyi Jessica Li. Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. Genome Biol, 22:1–25, 12 2021.
[145] Sun Shanwen, Lei Xu, Zou Quan, Wang Guohua. Bp4rnaseq: a babysitter package for retrospective and newly generated rna-seq data analyses using both alignment-based and alignment-free quantification method. Bioinformatics 2021;37(1319–1321):6.
[146] Jiarui Ding and Aviv Regev. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. Nature Commun, 12:2554, 12 2021.
[147] Daniel Castillo-Secilla, Juan Manuel Gálvez, Francisco Carrillo-Perez, Marta Verona-Almeida, Daniel Redondo-Sánchez, Francisco Manuel Ortuno, Luis Javier Herrera, and Ignacio Rojas. Knowseq r-bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. Comput Biol Med, 133:104387, 6 2021.
[148] Brian Aevermann, Yun Zhang, Mark Novotny, Mohamed Keshk, Trygve Bakken, Jeremy Miller, Rebecca Hodge, Boudewijn Lelieveldt, Ed Lein, and Richard H. Scheuermann. A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell rna sequencing. Genome Res, 31:1767–1780, 10 2021.
[149] Yu Hu, Li Fang, Xuelian Chen, Jiang F. Zhong, Mingyao Li, and Kai Wang. Liqa: long-read isoform quantification and analysis. Genome Biology, 22:1–21, 12 2021. LIQA to quantify isoform expression and detect differential alternative splicing (DAS).