frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 13 March 2023
DOI 10.3389/fgene.2023.997383

OPEN ACCESS

EDITED BY
Panwen Wang,
Mayo Clinic Arizona, United States

REVIEWED BY
Teresa Colombo,
National Research Council (CNR), Italy
Tiziana Castrignanò,
University of Tuscia, Italy

*CORRESPONDENCE
Serghei Mangul,
serghei.mangul@gmail.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 18 July 2022
ACCEPTED 24 February 2023
PUBLISHED 13 March 2023

CITATION
Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, Muszyńska A, Munteanu V, Yang H, Rotman J, Tao L, Balliu B, Tseng E, Eskin E, Zhao F, Mohammadi P, P. Łabaj P and Mangul S (2023), RNA-seq data science: From raw data to effective interpretation. *Front. Genet.* 14:997383. doi: 10.3389/fgene.2023.997383

RNA-seq data science: From raw data to effective interpretation

Dhrithi Deshpande[1†], Karishma Chhugani[1†], Yutong Chang[1], Aaron Karlsberg[2], Caitlin Loeffler[3], Jinyang Zhang[4], Agata Muszyńska[5,6], Viorel Munteanu[7], Harry Yang[8], Jeremy Rotman[2], Laura Tao[9], Brunilda Balliu[9], Elizabeth Tseng[10], Eleazar Eskin[3,9,11], Fangqing Zhao[4,12], Pejman Mohammadi[13], Paweł P. Łabaj[5,14] and Serghei Mangul[2,15]*

[1]Department of Pharmacology and Pharmaceutical Sciences, USC Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, Los Angeles, CA, United States, [2]Department of Clinical Pharmacy, USC Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, Los Angeles, CA, United States, [3]Department of Computer Science, University of California, Los Angeles, CA, United States, [4]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China, [5]Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland, [6]Institute of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland, [7]Department of Computers, Informatics and Microelectronics, Technical University of Moldova, Chisinau, Moldova, [8]Department of Microbiology, Immunology and Molecular Genetics, University of California Los Angeles, Los Angeles, CA, United States, [9]Department of Computational Medicine, David Geffen School of Medicine at UCLA, CHS, Los Angeles, CA, United States, [10]Pacific Biosciences, Menlo Park, CA, United States, [11]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, United States, [12]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China, [13]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States, [14]Department of Biotechnology, Boku University Vienna, Vienna, Austria, [15]Department of Quantitative and Computational Biology, USC Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, United States

RNA sequencing (RNA-seq) has become an exemplary technology in modern biology and clinical science. Its immense popularity is due in large part to the continuous efforts of the bioinformatics community to develop accurate and scalable computational tools to analyze the enormous amounts of transcriptomic data that it produces. RNA-seq analysis enables genes and their corresponding *transcripts* to be probed for a variety of purposes, such as detecting novel exons or whole transcripts, assessing expression of genes and alternative transcripts, and studying alternative splicing structure. It can be a challenge, however, to obtain meaningful biological signals from raw RNA-seq data because of the enormous scale of the data as well as the inherent limitations of different sequencing technologies, such as *amplification bias* or *biases of library preparation*. The need to overcome these technical challenges has pushed the rapid development of novel computational tools, which have evolved and diversified in accordance with technological advancements, leading to the current myriad of RNA-seq tools. These tools, combined with the diverse computational skill sets of biomedical researchers, help to unlock the full potential of RNA-seq. The purpose of this review is to explain basic concepts in the computational analysis of RNA-seq data and define discipline-specific jargon.

KEYWORDS

RNA sequencing, transcriptome quantification, differential gene expression, high throughput sequencing, read alignment, bioinformatics

1 Introduction

High-throughput DNA sequencing technologies, including *next-generation sequencing* and the newly emerging *third-generation sequencing*, enable the gene sequences of living organisms to be probed in a cost-effective manner (Shendure and Ji, 2008). These sequencing technologies have also been adapted for RNA sequencing (RNA-seq), which enables the expression of various RNA populations, including mRNA and total RNA, to be detected and quantified. RNA-seq has reshaped biomedical research by expanding researchers' ability to analyze a vast range of biological data (Kukurba and Montgomery, 2015). To derive biological insights from RNA-seq data, researchers need to understand the steps involved in RNA-seq analysis and select appropriate tools to answer their research question.

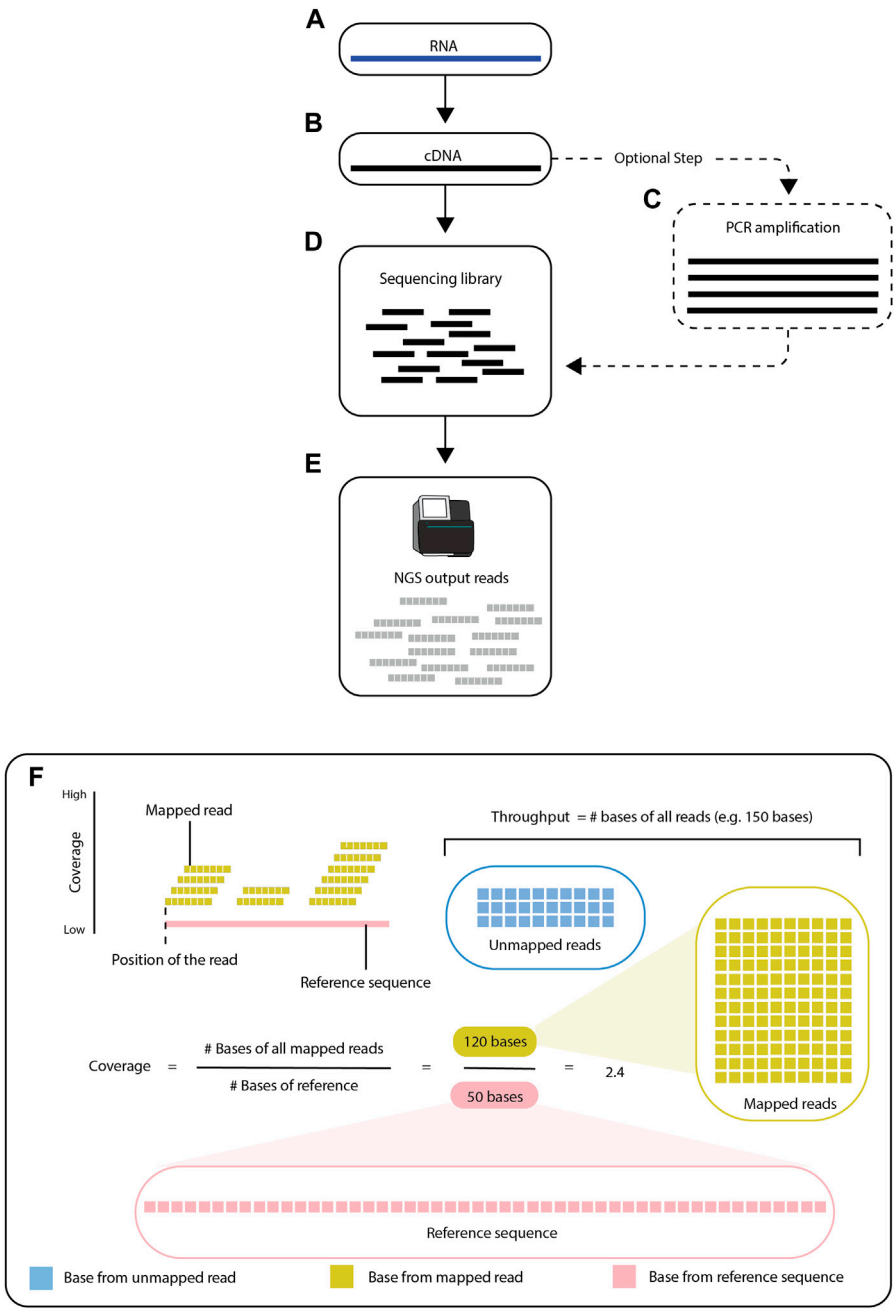


FIGURE 1
Overview of RNA-seq. RNA-seq is a process of creating short sequencing reads from RNA molecules. The steps consist of first converting the RNA **(A)** into cDNA **(B)**, then (optionally) amplifying the cDNA by PCR **(C)**, and finally fragmenting the cDNA into short pieces (known as fragments). After the sequencing library **(D)** is prepared, the fragments are used as input for next-generation sequencing **(E)**. The resulting sequence reads contained in FASTQ files are then aligned to a reference sequence **(F)**. Modern high-throughput sequencing machines can generate up to 150 million reads per run. The reference sequence, shown as a pink line, is known. The goal of the alignment is to find the *locus* in the reference sequence with the greatest match to each read. Reads are shown to align to the specific positions/locations and these mapped locations are recorded.

Biomedical researchers are often tasked with using computational methods for RNA-seq analysis, which are typically available wrapped as software tools and packages. In this review, we provide an overview of diverse methodologies for RNA-seq analyses that can be used to detect novel exons and transcripts, quantify gene expression and alternative splicing, and study alternative splicing structure. We discuss the steps from the generation of raw data using sequencing technologies to the effective interpretation and visualization of RNA-seq data using mapping and quantification techniques. By summarizing the biological and computational foundations of RNA-seq data generation, analysis, and software development, we hope this review will lead to a more deliberate use of existing computational tools.

2 RNA sequencing

RNA-seq uses high-throughput sequencing of nucleic acids to determine the nucleotide sequence of RNA molecules as well as the quantities of specific RNA species within populations of RNA molecules. RNA-seq analysis requires specialized computational tools that can account for the shortcomings of sequencing technologies, including the generation of *sequencing errors* (Le et al., 2013), *length biases* (Oshlack and Wakefield, 2009), and *fragmentation* (Tuerk et al., 2017). Computational analysis of RNA-seq data has led to many scientific advances, including novel therapeutic discoveries, detailed understanding of genetic regulatory regions, and identification of biomarkers and pathogenic mutations (Han et al., 2015).

Preparation of an RNA-seq library starts with extraction and isolation of RNA from a biological sample, such as a cell line or a frozen tissue sample. For RNA-seq performed with short-read sequencing (see Section 2.1), the isolated RNA is reverse-transcribed and converted into *cDNA*, which is then amplified by *polymerase chain reaction (PCR)* and fragmented into short sequences (either before or after PCR) (Prakash and Haeseler, 2017) (Figure 1). After the RNA molecules are processed, the RNA-seq library becomes the input for a sequencing platform (Kukurba and Montgomery, 2015), which generates reads (i.e., the sequenced fragments from the RNA-seq library).

2.1 High-throughput RNA-seq technologies

High-throughput sequencing techniques can derive millions of nucleotide sequences from an individual *transcriptome* (Stark et al., 2019). These nucleotide sequences provide multifold coverage of the whole transcriptome. High-resolution RNA-seq can identify which genes are actively transcribed in a sample and quantify the levels at which alternative transcripts of a gene are transcribed (Gerstein et al., 2007). The reads generated by different sequencing technologies have lengths ranging from hundreds of base pairs (usually referred to as short reads) to thousands of base pairs (referred to as long reads) (Shendure and Ji, 2008; Haas and Zody, 2010; Pollard et al., 2018). Illumina, Nanopore, and PacBio are among the most commonly used high-throughput sequencing platforms (Ye et al., 2015).

Illumina sequencing, considered a next-generation sequencing technology, is based on sequencing-by-synthesis chemistry and was first commercialized in 2006 (Shen and Shen, 2019). For Illumina RNA-seq, isolated RNAs are reverse-transcribed into single-stranded cDNA, which is then ligated to synthetic adapters, immobilized on a solid surface, and amplified by PCR. Then, a reaction mixture is added containing primers, DNA polymerase, and modified nucleotides. The modified *nucleotides have a fluorescent label* that serves as both a reversible terminator of DNA synthesis and an indicator of which nitrogenous base the nucleotide contains. As a new strand of DNA is synthesized using the immobilized cDNA as a template, each incorporated nucleotide is detected with a *charge-coupled device (CCD)* camera and identified by the color of the fluorescent label. The fluorescent label is then removed, and the next nucleotide is added in a new round of DNA synthesis. This cycle is repeated until each base in the cDNA is identified. The sequences of more than 10 million cDNA fragments can be simultaneously determined in parallel using the Illumina platform, giving rise to higher sequencing throughput compared with other sequencing platforms (Morganti et al., 2019; Workflows for RNA Sequencing, 2023).

Nanopore sequencing, which serves as the basis for the MinION, GridION, and PromethION platforms, was first introduced in 2014 by Oxford Nanopore Technologies. Nanopore sequencing can produce short or long reads from native DNA and RNA fragments of any length. Nanopores are very small holes in a membrane that can be created by pore-forming proteins or by non-biological means. The Nanopore sequencing method simultaneously sends an ionic current and a single strand of DNA or RNA through a nanopore. As the ionic current passes through each nucleotide that successively occupies the nanopore, it undergoes disruptions that are unique to the nitrogenous base. The patterns of disruption in the current can be interpreted to identify each base in the DNA or RNA strand that passes through the nanopore. Whereas short-read sequencing technologies such as Illumina require chemical modification or PCR amplification, Nanopore technology is capable of sequencing DNA or RNA without these additional steps, making it a third-generation sequencing technology (Bharagava et al., 2019).

PacBio sequencing, also known as SMRT (single-molecule, real-time) sequencing, was introduced in 2010 and generates full-length cDNA sequences (i.e., long reads) that characterize transcripts of targeted genes or across entire transcriptomes. Long reads generated by PacBio are accurate at the scale of a single molecule because they are generated by a process of circular consensus sequencing, in which the same cDNA is effectively read many times (Eid et al., 2009; Vierra et al., 2021). The comparatively high sensitivity of PacBio can be limited by external factors. For example, PacBio can produce full-length cDNA during the library preparation step; however, it can only generate high-quality reads if the target cDNA is short enough to be sequenced in multiple passes.

Each sequencing technology has inherent advantages and limitations, so no technology is best suited for all types of RNA-seq analysis (Box 1). Short-read technologies can generate data with a lower error rate and higher throughput than long-read technologies; however, the short-read length makes reconstruction and quantification of the transcriptome challenging (Korf, 2013; The RGASP Consortium et al., 2013a; The RGASP Consortium et al., 2013b; Amarasinghe et al., 2020). Long-read sequencing improves the accuracy of assembly (concatenation of individual reads to reassemble the transcriptome), or can even eliminate the need for assembly, as each read can cover an entire transcript. Long-read sequencing can also be used to produce complete, unambiguous information about

10.3389/fgene.2023.997383

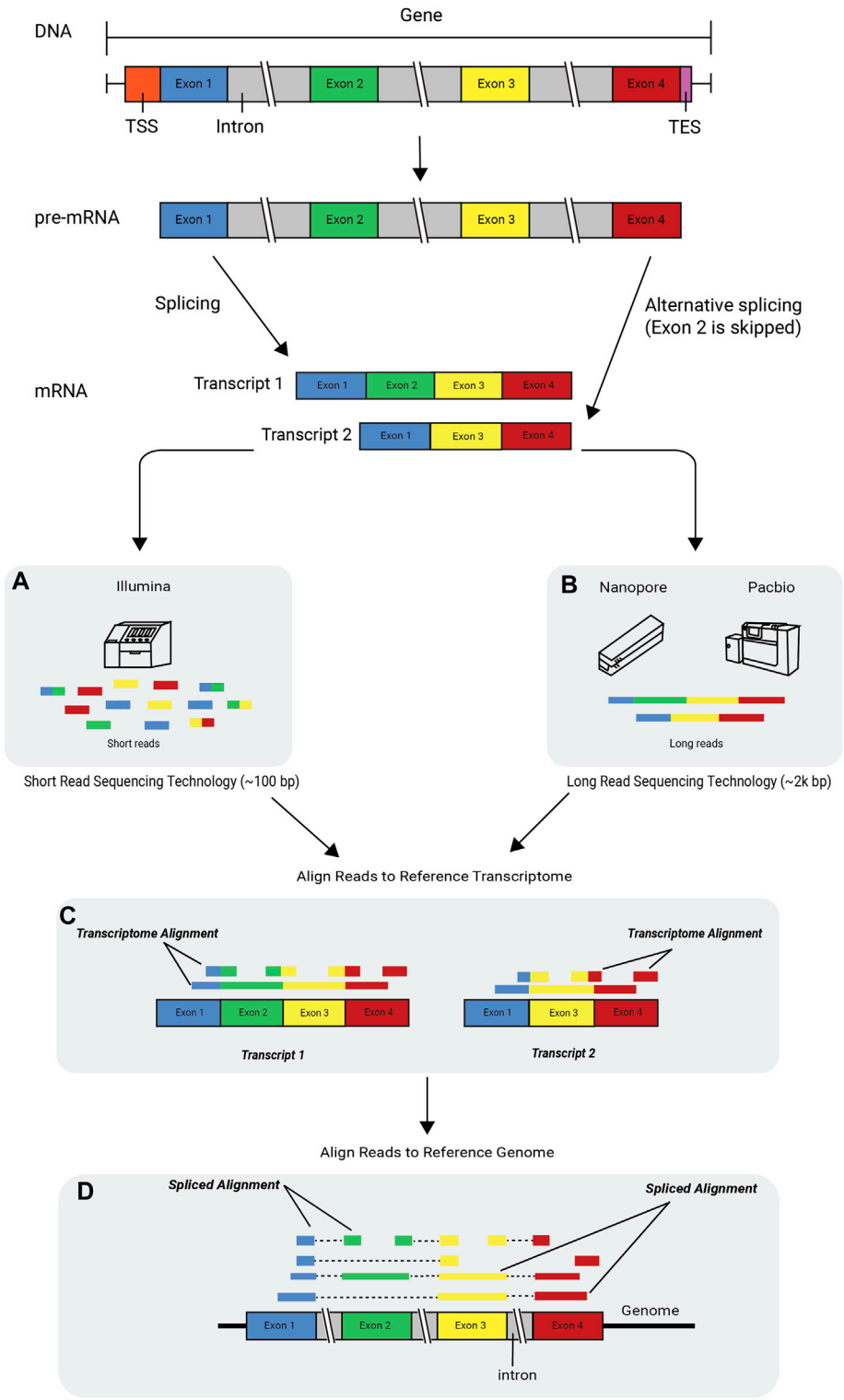


FIGURE 2
Alternative splicing and RNA-seq technologies. The flow of genetic information begins with DNA, which consists of introns and exons. DNA is transcribed into pre-mRNA and then further processed into mature mRNA by splicing out the introns and leaving the exons glued together. The mRNA is then translated into a protein. Transcripts with different arrangements of exons can be formed in a process called alternative splicing or exon skipping. An RNA-seq read is a short sequence sampled from a transcript. Reads are generated using sequencing technologies such as **(A)** the Illumina platform, which produces short reads, and the **(B)** Nanopore and PacBio platforms, which produce long reads. The figure depicts two scenarios in which uniquely mapped reads are aligned to a reference transcriptome **(C)** and a reference genome **(D)**, respectively. A few of the reads are multicolored, indicating that when aligned, they span across an exon-exon junction. Some of the shorter reads (single-colored) are aligned only to a single exon and do not span across the junction. *TSS*, transcription start site; *TES*, transcription end site.

Box 1 | Advantages and limitations of short and long reads

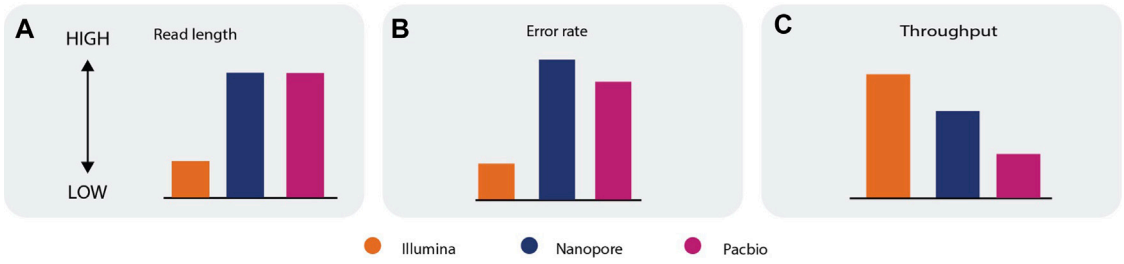
i. **Error rate**—Short read sequencing technologies have a lower error rate when compared to long read sequencing technologies **(a, b)**.

ii. **Throughput**—The throughput of long read sequencing technologies is typically lower than the throughput of short read sequencing technologies **(c)**.

iii. **Alignment**—Short reads suffer from multi-mapping issues, whereas longer reads, by nature of having more information, can be more accurately mapped to its origin. Due to a high error rate, pairwise alignment between the read, the reference transcriptome, and/or genome is more challenging for long reads compared to short reads.

iv. **Assemble novel transcripts**—Longer reads are preferred for *de novo* assembly, because they make the assembly step efficient. Most short reads do not span the shared region or shared exon junction, making the assembly step ambiguous. Full-length transcript sequencing eliminates the need for assembly.

v. **Estimate transcripts and gene expression**—Shorter reads are preferred for quantification of transcripts due to their higher throughput. However, assigning short reads to the transcripts requires more advanced probabilistic and statistical approaches. Longer reads have lower throughput, but they can usually cover the entire transcript and make determination of the transcript for each read a straightforward process.



alternative splicing, gene structure, regulatory elements, and coding regions. Long-read sequencing currently has a higher error rate and lower throughput compared with short-read sequencing, however (Figure 2) (Sedlazeck et al., 2018; De Maio et al., 2019; Mahmoud et al., 2019). Hybrid approaches that combine long reads and short reads can eliminate the limitations of each separate approach and can be used to accurately quantify and assemble known and novel transcripts (De Maio et al., 2019; Amarasinghe et al., 2020; Berbers et al., 2020), but they also have higher costs and more material requirements. Data gathered using Illumina, Nanopore, and PacBio sequencing technologies can be used to address a wide range of research areas, including transcriptome analysis, population-scale analysis, and clinical research (Wang et al., 2021).

3 RNA-seq data science: From raw data to effective interpretation

RNA-seq is multifaceted and can be used to uncover and expound new insights on, for example, a dysregulated gene or defective protein that has a downstream effect leading to a disease state (Costa et al., 2010). Computational analysis of RNA-seq data is central to decoding the biological complexities in the transcriptomes of living organisms, including humans (Costa et al., 2010). Here, we describe the major steps of computational analysis of RNA-seq data, beginning from the processing of raw data to the uncovering of biological insights.

3.1 Quality control of raw data

During the sequencing process, errors are introduced into reads that can bias the results of downstream analyses. Read trimming and data quality control to filter and assess the quality of raw reads (Yang et al., 2013) are therefore essential after the reads have been generated. Read trimming removes adapter sequences and portions of reads with low accuracy, as indicated by a low *PHRED quality score* (Martin, 2011; Dodt et al., 2012; Bolger et al., 2014). In addition, computational error correction can be applied to reduce the number of sequencing errors (Lima et al., 2020; Mitchell et al., 2020).

4 Read alignment

Read alignment is an essential step in RNA-seq downstream analysis. RNA-seq data typically lack information about the order and origin of the reads, including the specific part, homolog, or strand of the *genome* from which they originate. Computational alignment of the reads to an annotated reference transcriptome can establish where on the genome the reads originated (Figure 1) (Brown, 2002). Alignment of the reads to a reference sequence also reveals how many reads overlap each position on the reference sequence, which is known as the coverage. There are several bioinformatics tools (e.g., GenomeScope (Vurture et al., 2017), Smudgeplot (Ranallo-Benavidez et al., 2020), and Merqury (Rhie et al., 2020)) that can estimate the coverage without mapping the reads to a reference sequence (Ranallo-Benavidez et al., 2020; Rhie et al., 2020), as most of the overlap between reads is preserved with or without the reference sequence (Vurture et al., 2017) (Figure 1).

Alignment of RNA-seq reads to a complementary reference sequence can help determine which transcripts are expressed and the degree to which they are expressed, but the alignment approach is ill-equipped to discover transcripts that are missing from the reference sequence. Furthermore, even the human reference transcriptome remains incomplete (Nellore et al., 2016). Novel transcripts can be discovered by performing *de novo* assembly of RNA-seq reads to generate an

entire transcriptome without alignment to a reference sequence; however, this can be challenging and requires large amounts of computational time and resources (Grabherr et al., 2011) As an alternative, RNA-seq reads can be aligned to curated databases of known transcripts such as RefSeq (Pruitt et al., 2007), UCSC genome browser, Ensembl, GENCODE (GENCODE, 2022), and AceView (Larsson et al., 2005), and reads that fail to align to known transcripts can then be aligned to a reference genome to identify novel transcripts.

One computational challenge in aligning RNA-seq reads to a reference genome is the handling of spliced junctions, where one part of the read maps to the end of one exon and the rest of the read maps to another exon, which may be located thousands of base pairs away from the first exon. Spliced junctions are the result of the removal of non-coding parts of a gene, called introns, and the splicing together of the coding parts of the gene, called exons. Genes can generate multiple mRNA transcripts through alternative splicing. As a result, exons are combined or skipped in different ways and have alternative start/end sites. These varying combinations create different transcripts, known as *isoforms*, from the same gene. As a biological process, alternative splicing is evolutionarily advantageous, because it enables the production of different protein variants from the same genetic information (Figure 2). When genome annotations are available, existing exon structures can be used to map reads across known splice junctions; however, this knowledge-guided approach may be biased towards mapping only known junctions while failing to discover novel ones.

In cases where reads align to multiple transcripts, it might not be possible to discern from which transcript the reads originate. Splice alignment software packages (Wang et al., 2010; Dobin et al., 2013; Kim et al., 2019) are designed to minimize multi-mapping by correctly aligning reads across the exon–intron junctions of the reference genome (Figure 2). This can be a crucial first step of reference-guided assembly, wherein transcripts that are present in the sample but not annotated in the reference are assembled using the spliced read alignments to the reference.

In some instances, reads do not perfectly align with the reference sequence but instead contain mismatches, which can be caused either by sequencing errors or by biological variation such as mutations (Mitchell et al., 2020). RNA-seq alignment tools are typically equipped with a customizable threshold for tolerating mismatches in the alignment; however, it is important to distinguish between sequencing errors and real variation between the transcripts and the reference sequence. Specialized computational tools (Abate et al., 2014; Fernandez-Cuesta et al., 2015) can identify and classify genes using strategies such as *de novo* assembly (assembly of reads without alignment to a reference sequence), identification of reads that span fusion junctions, and filtering of gene fusion candidates based on various criteria.

5 Quantitative analysis of gene expression

RNA-seq enables quantitative analysis of gene expression at the level of alternative transcripts. The sequence fragments derived from mRNA can reveal which genes are expressed and how strongly they are expressed. Additionally, differential expression (DE) analysis can show how expression levels change under different conditions or between different populations.

5.1 Estimation of transcript and gene expression

Computational methods can estimate expression levels of genes and transcripts by counting the number of reads that match individual reference transcripts. Tools like HT-Seq-count, Rcount, and featureCounts (Liao et al., 2014; Anders et al., 2015; Schmid and Grossniklaus, 2015) are highly robust and widely used for such analyses; however, counting-based tools are ill-equipped to estimate the expression levels of different isoforms of expressed genes using short reads, as the majority of isoforms share a large percentage of exons and cannot be uniquely assigned to individual transcripts (Figure 2). The shorter the reads, the greater the probability that they will match multiple transcripts. A conservative approach to tackle this challenge is to consider only the reads that uniquely map to a single transcript (e.g., reads that map to transcript-specific splicing junctions or exons) (Conesa et al., 2016). An alternative approach that utilizes a larger fraction of the RNA-seq reads is to probabilistically assign reads to the isoforms from which they likely originated (Li and Dewey, 2011; Nicolae et al., 2011; Trapnell et al., 2012; Pertea et al., 2015).

A number of approaches quantify gene expression using complete read alignment, which requires large amounts of computational power and time to compare each read to reference sequences base-by-base. Pseudoalignment methods have been developed as an alternative approach that has a much smaller computational burden. These methods forgo the base-by-base accuracy of alignment and determine an approximate alignment of the reads on the genome, which is still sufficiently accurate to quantify gene expression. Pseudoalignment algorithms leverage a pre-compiled library of unique k-mers (exact substrings of length k) contained in known transcripts and assign reads to transcripts by counting the k-mer occurrences in the reads, thus achieving up to 100 times faster quantification compared with alignment-based methods (Bray et al., 2016). Sailfish (the pioneer of pseudoalignment) (Patro et al., 2014), Salmon (Patro et al., 2017), and Kallisto (Bray et al., 2016) each utilize pseudo-alignment-based algorithms to quantify the isoforms of expressed transcripts (Alser et al., 2020), each providing comparable accuracy in expression quantification. A more detailed explanation of these tools can be found in Supplementary Material S2.

5.2 Differential gene expression analysis

After gene and transcript expression levels are estimated, statistical approaches are employed to detect differences in expression levels across experimental groups (e.g., different sexes or cohorts exposed to different environmental

conditions) (Conesa et al., 2016). Expression levels measured for the same gene under different conditions cannot be directly compared, as each experiment represents a statistical sample, giving only the relative mRNA levels in comparison to the other mRNAs present in the sample. In addition, mRNA levels change over time, and reads can align to multiple places, making exact quantitation difficult. The purpose of statistical testing is to ensure that an observed change in mRNA levels is due to an actual difference in expression between experimental conditions.

To test whether the expression of a given gene is different between two groups, measurements are repeated in multiple replicates of the same experiments, and then a statistical test is applied. Through this process, the variation in expression between different conditions can be compared to the variation within replicates of the same condition. Each statistical test is based on a null hypothesis that the gene expression is the same between groups, which is usually true for the majority of genes. The value that indicates whether there is likely to be a true difference between groups is called the *p*-value, which gives the probability of observing a particular difference, or a more extreme difference, assuming that the null hypothesis is true. Small *p*-values give strong evidence against the null hypothesis. Genes with low *p*-values are considered to be differentially expressed, and the null hypothesis is rejected for those genes. The typical threshold for rejection of a null hypothesis is a *p*-value less than 0.05, but this cutoff is arbitrary and might need to be altered depending on how noisy the data are (Liu et al., 2006; Glaus et al., 2012; Shastry et al., 2020).

There are two types of error associated with statistical tests: Type I error and Type II error. A Type I error occurs if a test rejects a true null hypothesis. A Type II error occurs if a test accepts a false null hypothesis. The *p*-value indicates the probability of making a Type I error in a given test. For example, if the *p*-value threshold is set at 0.05 (i.e., 5%), and 20,000 genes are being tested, then 1,000 genes (5% · 20,000) will be wrongly considered to be differentially expressed because of Type I errors. There are two approaches to control Type I errors, also referred to as false positives. One approach is to control the family-wise error or the probability that there is at least one Type I error among all the rejected null hypotheses. The other approach is to control the false discovery rate, or the proportion of Type I errors among all the rejected null hypotheses. Both approaches involve calculation of an adjusted *p*-value (p-adj) for each gene, which can then be used for further analysis (Jafari and Ansari-Pour, 2018).

It is important to account for *noise* which includes sources of variation that are unrelated to the experimental variable of interest, when performing differential expression analysis. For example, *batch effects*, or confounding factors arising from samples being tested on different days, by different laboratory technicians, or in different laboratories (technical batch effects), can result in unwanted differences in measured values. In addition, variation due to intrinsic factors such as high GC content or gene body coverage evenness (biological batch effects) can affect the quantification of technical replicates of a sample. Existing statistical methods can effectively detect and adjust for hidden confounding factors (Li et al., 2014).

Other approaches to differential expression analysis that can produce more accurate results than conventional p-adj values use different metrics such as the minimum significant difference or the generalized linear model (GLM) framework (McCarthy et al., 2012), where a combination of *p*-values and log fold changes is applied to identify the genes or transcripts with the most significant differences in expression. Another alternative approach is the probability of positive log ratio (PPLR) (Liu et al., 2006), which was initially developed for microarray analysis and subsequently adjusted for RNA-seq data (Glaus et al., 2012). The PPLR uses a Bayesian hierarchical model to express the probability that the ratio of expression levels between two conditions is positive (i.e., the expression is upregulated in the second condition relative to the first). A PPLR value close to 1 means there is a very high probability that a given transcript is upregulated in the second condition (Liu et al., 2006). When the PPLR value is close to 0, there is a very low probability of upregulation, and consequently a high probability of downregulation, in the second condition relative to the first. There is no direct relation between PPLRs and *p*-values, as they look at the problem from different perspectives (i.e., in the probabilistic approach an uncertainty propagation between successive stages of analysis is possible and desired). Both approaches are capable of identifying large numbers of differentially expressed genomic features. If the number of differentially expressed features is too large, a more stringent cutoff for statistical significance can be applied to make the analysis more manageable.

Depending on the type of *normalization* performed on RNA-seq data, machine-learning approaches can be used to identify differentially expressed genes with classification models based on discrete or continuous distributions. Machine learning approaches have been used to manage, model, and categorize biological data, enabling high-impact discoveries in the field of biomedicine (Shastry et al., 2020). RNA-seq data are discrete in nature. The two most common ways to normalize RNA-seq data for machine learning-based differential expression analysis are to model the data as a Poisson or negative binomial distribution or transform the data to be similar to a distribution of microarray data. The Bioconductor MLSeq (Goksuluk et al., 2019) package is a comprehensive source of combinations of different normalization and machine-learning methods for RNA-seq analysis. After the data are normalized, genes or alternative transcripts (features) can be ranked, or standard sample classification can be performed, and the features that make the strongest contributions to the assignment of samples to particular groups can be extracted (Goksuluk et al., 2019). With a deep learning approach, it is also possible to predict differences in gene expression from histone modification signals (Sekhon et al., 2018).

Differential expression analysis can be complemented by *expression quantitative trait loci (eQTL)* analysis, which formally compares the expression levels of a given gene between groups with different copy numbers (0, 1, or 2) of the minor allele. Each read alignment technique produces different results, which may impact which genes are identified as differentially expressed (Castel et al., 2015). The power to detect differentially expressed genes and eQTLs depends on the sequencing depth of the sample, the minor allele frequency of the gene being tested, the expression level of the

gene, and the length of the gene (McKenna et al., 2010). The magnitude of the eQTL can be quantified by the log allelic fold change (Hu et al., 2015), and its significance is tested using a binomial distribution or over-dispersed generalizations (Kumasaka et al., 2016; Knowles et al., 2017; Mohammadi et al., 2019; Zou et al., 2019; Wang et al., 2020). Some of the popular approaches to detect eQTLs use transformation and linear regression models (Shabalin, 2012; Ongen et al., 2016; Taylor-Weiner et al., 2019).

The results of differential expression analyses can be validated using independent techniques such as *quantitative PCR (qPCR)*, which is statistically assessable (Skelly et al., 2011). Measurements of gene expression obtained by qPCR are relatively similar to measurements obtained by RNA-seq analysis, where a value can be calculated for the concentration of a target region in a given sample (Harvey et al., 2015; Romanel et al., 2015; Xie et al., 2019). Additional information about quantification of RNA splicing and splicing QTL (sQTL) analyses can be found in Supplementary Material S3.

6 Measurement of allele-specific expression

RNA-seq can measure allele-specific expression (ASE or allelic expression) to uncover the cis-regulatory effects of genetic variants (McKenna et al., 2010; Castel et al., 2015; Raghupathy et al., 2018). ASE represents gene expression measured independently for the paternal and maternal alleles of a gene. In a typical RNA-seq experiment, ASE can be measured only in genes that contain a heterozygous *single-nucleotide polymorphism (SNP)* within the transcribed region. This SNP, referred to as the aseSNP, can be used as a tag to identify reads that originate from each copy of the gene (Figure 3).

Allelic imbalance—the ratio between paternal and maternal allele expression—identifies genetic cis-regulatory differences between two haplotypes. The log allelic fold change can also be calculated to quantify the magnitude of allelic imbalance (Hu et al., 2015). An aseSNP is not itself a regulatory variant and should not induce an imbalanced ASE signal. However, there can be a bias in ASE data that falsely suggests that the haplotype carrying the reference allele for the aseSNP has slightly higher expression across all genes. This issue, known as allelic bias or reference bias, can be mitigated in two ways: by aligning the RNA-seq reads to a personalized reference genome that excludes likely biased sites (Dobin et al., 2013; van de Geijn et al., 2015; Gao and Zhao, 2018; Kristensen et al., 2019; Ferraro et al., 2020), or by aggregating the ASE signal from multiple aseSNPs in each gene (Chen et al., 2021). ASE data can also be used to improve statistical power for identifying eQTLs (Gao and Zhao, 2018; Kristensen et al., 2019; Zou et al., 2019; Ferraro et al., 2020) and to map the causal regulatory variants in eQTL data (Kim and Salzberg, 2011; Gao et al., 2018; Haas et al., 2019). Furthermore, ASE data are inherently robust to noise, so they are useful for identifying gene-by-environment interaction effects (Li, 2013) or the effects of rare genetic variants on gene expression to improve diagnostic accuracy for Mendelian diseases (Hoffmann et al., 2014; Ji et al., 2019).

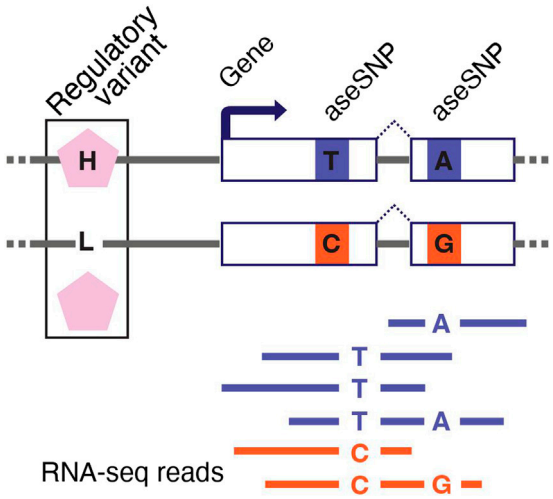


FIGURE 3
Measuring allele-specific expression with RNA-seq. RNA-seq can be used to generate allele-specific expression (ASE) data for genes with a heterozygous single-nucleotide polymorphism in the transcribed region (aseSNP). The aseSNP enables sequencing reads to be mapped to the haplotype from which they originate. Imbalance in ASE data is a functional indicator of a cis-regulatory difference between the two haplotypes that is driven by heterozygous regulatory variants. Data from multiple aseSNPs can be aggregated to improve ASE data quality. The non-coding regulatory variant depicted here has two alleles inducing higher (H) and lower (L) expression of the target gene.

7 Profiling circular RNA with RNA-seq

Circular RNA (circRNA) is a large class of RNA molecules with a covalently closed circular structure that plays important roles in various biological processes and metabolic mechanisms (Wu et al., 2020). In recent years, a variety of computational tools have been developed for circRNA study (Gao and Zhao, 2018; Chen et al., 2021). Identification of circRNAs is based on detection of reads spanning the circle junction, termed the back-splice junction (BSJ). Most tools (Cheng et al., 2016; Zhang et al., 2016; Gao et al., 2018) employ aligners (Humphreys et al., 2019; Wu et al., 2019; Zheng et al., 2019) to detect putative back-splicing events from fusion reads or split alignment results, whereas other splice-aware aligners (Wang et al., 2010; Zheng and Zhao, 2020) can align circular reads and detect BSJs directly.

Considering that most circRNAs are derived from exonic regions (Ji et al., 2019; Wu et al., 2020) where computational methods cannot accurately distinguish linear and circular reads, the BSJ read count is the most reliable measurement of circRNA expression levels. The BSJ read count is inferred from alignment results, and different filters and statistical strategies have been employed to improve its accuracy and sensitivity (Mangul et al., 2019; Zhang et al., 2020). Alternative approaches using pseudoalignment-based tools for circRNA quantification (Li et al., 2017) can substantially increase the computational efficiency compared with regular alignment-based methods. To compare the expression levels of circRNAs and their host

genes, the junction ratio, defined as the ratio of BSJ reads and linear reads mapped to the BSJ site, is often used for comparative analysis. Several computational methods have been developed for accurate estimation of junction ratios (Reimers and Carey, 2006; The Comprehensive R Archive Network, 2022). In addition, circRNAs exhibit alternative splicing patterns, and a number of specific tools have been developed for circular transcript assembly (Gao et al., 2016; Zhang et al., 2016; Wu et al., 2019; Zheng et al., 2019), internal structure visualization (Li et al., 2016; Mose et al., 2016), and differential expression analysis (Zhang et al., 2020; The Comprehensive R Archive Network, 2022). Several comprehensive databases have been constructed for circRNA annotation and prioritization analysis (Dong et al., 2018; Xia et al., 2018; Wu et al., 2020).

8 Discussion

As technology advances, RNA-seq methods have become increasingly popular and have revolutionized modern biology and clinical applications, driven by continuous efforts of the bioinformatics community to develop accurate and scalable computational tools. In addition, advancements in sequencing technologies have provided an unprecedented ability to analyze a wide range of biological data, enabling new explorations of novel and existing biological problems. To increase access to RNA-seq methods among new users and young scientists, we provided an overview of the fundamentals of RNA-seq and its associated computational methods and discussed the advantages and limitations of various applications.

Computational analysis of RNA-seq data can be used to tackle important biological problems such as estimating gene expression profiles across various phenotypes and conditions or detecting novel alternative splicing on specific exons. Specialized analyses of RNA-seq data can also help to detect changes in the concentration, function, or localization of transcription factors that affect splicing and can cause the onset of neurodegenerative diseases and cancers (Ozsolak and Milos, 2011; Szabo and Salzman, 2016). Some recently developed computational tools (Xu et al., 2014; Bolotin et al., 2015; Li et al., 2016; Mose et al., 2016; Mandric et al., 2020) are even capable of repurposing RNA-seq data to characterize the individual adaptive immune repertoire and microbiome (Varadhan and Roland, 2008). Additionally, computational deconvolution can be applied to RNA-seq data to study cell-type compositions in tissue samples (Melsted et al., 2017; Kang et al., 2019).

The interdisciplinary nature of RNA-seq applications and related analytic methods and software development introduces a host of terms that can challenge researchers in the wider scientific and medical research communities. The literature on RNA-seq methods has traditionally assumed that readers are familiar with the fundamental concepts of RNA-seq and related bioinformatics analyses (Nariai et al., 2013; Srivastava et al., 2016; Zakeri et al., 2017; Green et al., 2018; Li et al., 2018; Vaquero-Garcia et al., 2018). These methods may require diverse computational skills to be used effectively. A lack of computational skills can therefore limit the ability of biomedical researchers to unlock the full potential of RNA-seq, highlighting the need for a review that explains basic RNA-seq concepts and defines discipline-specific jargon.

Author contributions

SM conceived of the idea presented and supervised the project. DD, KC and SM led the project. DD, KC, YC, AK, CL, JZ, AM, VM, HY, JR, LT, BB, ET, EE, FZ, PM, PL and SM contributed to the writing ofthe manuscript. DD, KC, YC, and PM produced figures in the main text. KC and DD created Supplementary Materials and the box. All authors discussed the text and commented on the manuscript.All authors read and approved the final manuscript.

Funding

Serghei Mangul's work has been supported by the National Science Foundation Grants (2041984 and 2135954). P. Mohammadi's work has been supported by the National Institute of General Medical Sciences (R01GM140287). Fangqing Zhao's work has been supported by the National Natural Science Foundation of China (31722031, 91640117). Agata M's work has been co-funded by the European Union through the European Social Fund grant (POWR.03.02.00-00-I029).

Acknowledgments

We thank Dr. Patro for his valuable comments and suggestions on the manuscript.

Conflict of interest

ET was employed by the company Pacific Biosciences (United States).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.997383/full#supplementary-material

References

Abate, F., Zairis, S., Ficarra, E., Acquaviva, A., Wiggins, C. H., Frattini, V., et al. (2014). Pegasus: A comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst. Biol.* 8, 97. doi:10.1186/s12918-014-0097-z

Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P. I., et al. (2020). Technology dictates algorithms: Recent developments in read alignment. *Genome Biol.* 22, 249. doi:10.1186/s13059-021-02443-7

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21, 30. doi:10.1186/s13059-020-1935-5

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638

Berbers, B., Saltykova, A., Garcia-Graells, C., Philipp, P., Arella, F., Marchal, K., et al. (2020). Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified Bacillus. *Sci. Rep.* 10, 4310. doi:10.1038/s41598-020-61158-0

Bharagava, R. N., Purchase, D., Saxena, G., and Mulla, S. I. (2019). Applications of metagenomics in microbial bioremediation of pollutants. *Microb. Divers. Genomic Era* 2019, 459–477. doi:10.1016/B978-0-12-814849-5.00026-5

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., et al. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* 12, 380–381. doi:10.1038/nmeth.3364

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi:10.1038/nbt.3519

Brown, T. A. (2002). *Understanding a genome sequence. Genomes*. 2nd edition. Hoboken, NJ, USA: Wiley-Liss.

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 16, 195. doi:10.1186/s13059-015-0762-6

Chen, L., Wang, C., Sun, H., Wang, J., Liang, Y., Wang, Y., et al. (2021). The bioinformatics toolbox for circRNA discovery and analysis. *Brief. Bioinform.* 22, 1706–1728. doi:10.1093/bib/bbaa001

Cheng, J., Metge, F., and Dieterich, C. (2016). Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics* 32, 1094–1096. doi:10.1093/bioinformatics/btv656

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17, 13. doi:10.1186/s13059-016-0881-8

Costa, V., Angelini, C., De Feis, I., and Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-seq. *Biomed. Res. Int.* 2010, e853916. doi:10.1155/2010/853916

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics* 5, e000294. doi:10.1099/mgen.0.000294

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). Star: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Dodt, M., Roehr, J., Ahmed, R., and Dieterich, C. (2012). FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biology* 1, 895–905. doi:10.3390/biology1030895

Dong, R., Ma, X. K., Li, G. W., and Yang, L. (2018). CIRCpedia v2: An updated database for comprehensive circular RNA annotation and expression comparison. *Genomics Proteomics Bioinforma.* 16, 226–233. doi:10.1016/j.gpb.2018.08.001

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138. doi:10.1126/science.1162986

Fernandez-Cuesta, L., Sun, R., Menon, R., George, J., Lorenz, S., Meza-Zepeda, L. A., et al. (2015). Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol.* 16, 7. doi:10.1186/s13059-014-0558-0

Ferraro, N. M., Strober, B. J., Einson, J., Abell, N. S., Aguet, F., Barbeira, A. N., et al. (2020). Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* 369, eaaz5900. doi:10.1126/science.aaz5900

Gao, Y., Wang, J., Zheng, Y., Zhang, J., Chen, S., and Zhao, F. (2016). Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nat. Commun.* 7, 12060. doi:10.1038/ncomms12060

Gao, Y., Zhang, J., and Zhao, F. (2018). Circular RNA identification based on multiple seed matching. *Brief. Bioinform.* 19, 803–810. doi:10.1093/bib/bbx014

Gao, Y., and Zhao, F. (2018). Computational strategies for exploring circular RNAs. *Trends Genet.* 34, 389–400. doi:10.1016/j.tig.2017.12.016

GENCODE (2022). *GENCODE - home page*. Available at: https://www.gencodegenes.org/.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., et al. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681. doi:10.1101/gr.6339607

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28, 1721–1728. doi:10.1093/bioinformatics/bts260

Goksuluk, D., Zararsiz, G., Korkmaz, S., Eldem, V., Zararsiz, G. E., Ozcetin, E., et al. (2019). MLSeq: Machine learning interface for RNA-sequencing data. *Comput. Methods Programs Biomed.* 175, 223–231. doi:10.1016/j.cmpb.2019.04.007

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi:10.1038/nbt.1883

Green, C. J., Gazzara, M. R., and Barash, Y. (2018). MAJIQ-SPEL: Web-tool to interrogate classical and complex splicing variations from RNA-seq data. *Bioinformatics* 34, 300–302. doi:10.1093/bioinformatics/btx565

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* 20, 213. doi:10.1186/s13059-019-1842-9

Haas, B. J., and Zody, M. C. (2010). Advancing RNA-Seq analysis. *Nat. Biotechnol.* 28, 421–423. doi:10.1038/nbt0510-421

Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinforma. Biol. Insights* 9, BBI.S28991. doi:10.4137/bbi.s28991

Harvey, C. T., Moyerbrailean, G. A., Davis, G. O., Wen, X., Luca, F., and Pique-Regi, R. (2015). QuASAR: Quantitative allele-specific analysis of reads. *Bioinformatics* 31, 1235–1242. doi:10.1093/bioinformatics/btu802

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., et al. (2014). A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.* 15, R34. doi:10.1186/gb-2014-15-2-r34

Hu, Y. J., Sun, W., Tzeng, J. Y., and Perou, C. M. (2015). Proper use of allele-specific expression improves statistical power for cis-eQTL mapping with RNA-seq data. *J. Am. Stat. Assoc.* 110, 962–974. doi:10.1080/01621459.2015.1038449

Humphreys, D. T., Fossat, N., Demuth, M., Tam, P. P. L., and HoUlarcirc, J. W. K. (2019). Ularcirc: Visualization and enhanced analysis of circular RNAs via back and canonical forward splicing. *Nucleic Acids Res.* 47, e123. doi:10.1093/nar/gkz718

Jafari, M., and Ansari-Pour, N. (2018). Why, when and how to adjust your P values? *Cell. J. Yakhteh* 20, 604–607. doi:10.22074/cellj.2019.5992

Ji, P., Wu, W., Chen, S., Zheng, Y., Zhou, L., Zhang, J., et al. (2019). Expanded expression landscape and prioritization of circular RNAs in mammals. *Cell. Rep.* 26, 3444–3460. doi:10.1016/j.celrep.2019.02.078

Kang, K., Meng, Q., Shats, I., Umbach, D. M., Li, M., Li, Y., et al. (2019). CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Comput. Biol.* 15, e1007510. doi:10.1371/journal.pcbi.1007510

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi:10.1038/s41587-019-0201-4

Kim, D., and Salzberg, S. L. (2011). TopHat-fusion: An algorithm for discovery of novel fusion transcripts. *Genome Biol.* 12, R72. doi:10.1186/gb-2011-12-8-r72

Knowles, D. A., Davis, J. R., Edgington, H., Raj, A., Fave, M. J., Zhu, X., et al. (2017). Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* 14, 699–702. doi:10.1038/nmeth.4298

Korf, I. (2013). Genomics: The state of the art in RNA-seq analysis. *Nat. Methods* 10, 1165–1166. doi:10.1038/nmeth.2735

Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B., and Kjems, J. (2019). The biogenesis, biology and characterization of circular RNAs. *Nat. Rev. Genet.* 20, 675–691. doi:10.1038/s41576-019-0158-7

Kukurba, K. R., and Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harb. Protoc.* 2015, 951–969. doi:10.1101/pdb.top084970

Kumasaka, N., Knights, A. J., and Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* 48, 206–213. doi:10.1038/ng.3467

Larsson, T. P., Murray, C. G., Hill, T., Fredriksson, R., and Schiöth, H. B. (2005). Comparison of the current RefSeq, Ensembl and EST databases for counting

genes and gene discovery. *FEBS Lett.* 579, 690–698. doi:10.1016/j.febslet.2004.12.046

Le, H. S., Schulz, M. H., McCauley, B. M., Hinman, V. F., and Bar-Joseph, Z. (2013). Probabilistic error correction for RNA sequencing. *Nucleic Acids Res.* 41, e109. doi:10.1093/nar/gkt215

Li, B., and Dewey, C. N. (2011). Rsem: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinforma.* 12, 323. doi:10.1186/1471-2105-12-323

Li, B., Li, T., Pignon, J. C., Wang, B., Wang, J., Shukla, S. A., et al. (2016). Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* 48, 725–732. doi:10.1038/ng.3581

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Genomics* 1303. doi:10.48550/ARXIV.1303.3997

Li, M., Xie, X., Zhou, J., Sheng, M., Yin, X., Ko, E. A., et al. (2017). Quantifying circular RNA expression from RNA-seq data using model-based framework. *Bioinformatics* 33, 2131–2139. doi:10.1093/bioinformatics/btx129

Li, S., Labaj, P. P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., et al. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* 32, 888–895. doi:10.1038/nbt.3000

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., et al. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50, 151–158. doi:10.1038/s41588-017-0004-9

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656

Lima, L., Marchet, C., Caboche, S., Da Silva, C., Istace, B., Aury, J. M., et al. (2020). Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. *Brief. Bioinform.* 21, 1164–1181. doi:10.1093/bib/bbz058

Liu, X., Milo, M., Lawrence, N. D., and Rattray, M. (2006). Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics* 22, 2107–2113. doi:10.1093/bioinformatics/btl361

Mahmoud, M., Gobet, N., Cruz-Davalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biol.* 20, 246. doi:10.1186/s13059-019-1828-7

Mandric, I., Rotman, J., Yang, H. T., Strauli, N., Montoya, D. J., Van Der Wey, W., et al. (2020). Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. *Nat. Commun.* 11, 3126. doi:10.1038/s41467-020-16857-7

Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., et al. (2019). Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLOS Biol.* 17, e3000333. doi:10.1371/journal.pbio.3000333

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* 17, 10–12. doi:10.14806/ej.17.1.200

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi:10.1093/nar/gks042

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Melsted, P., Hateley, S., Joseph, I. C., Pimentel, H., Bray, N., and Pachter, L. (2017). *Fusion detection and quantification by pseudoalignment*. 166322 Preprint. doi:10.1101/166322

Mitchell, K., Brito, J. J., Mandric, I., Wu, Q., Knyazev, S., Chang, S., et al. (2020). Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biol.* 21, 71. doi:10.1186/s13059-020-01988-3

Mohammadi, P., Castel, S. E., Cummings, B. B., Einson, J., Sousa, C., Hoffman, P., et al. (2019). Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* 366, 351–356. doi:10.1126/science.aay0256

Monlong, J., Calvo, M., Ferreira, P. G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* 5, 4698. doi:10.1038/ncomms5698

Morganti, S., Tarantino, P., Ferraro, E., D'Amico, P., Duso, B. A., and Curigliano, G. (2019). Next generation sequencing (ngs): A revolutionary technology in pharmacogenomics and personalized medicine in cancer. *Adv. Exp. Med. Biol.* 1168, 9–30. doi:10.1007/978-3-030-24100-1_2

Mose, L. E., Selitsky, S. R., Bixby, L. M., Marron, D. L., Iglesia, M. D., Serody, J. S., et al. (2016). Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics* 32, 3729–3734. doi:10.1093/bioinformatics/btw526

Nariai, N., Hirose, O., Kojima, K., and Nagasaki, M. (2013). Tigar: Transcript isoform abundance estimation method with gapped alignment of RNA-seq data by variational bayesian inference. *Bioinformatics* 29, 2292–2299. doi:10.1093/bioinformatics/btt381

Nellore, A., Collado-Torres, L., Jaffe, A. E., Alquicira-Hernandez, J., Wilks, C., Pritt, J., et al. (2016). Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. *Bioinformatics* 33, 4033–4040. doi:10.1093/bioinformatics/btw575

Nicolae, M., Mangul, S., Măndoiu, I. I., and Zelikovsky, A. (2011). Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.* 6, 9. doi:10.1186/1748-7188-6-9

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32, 1479–1485. doi:10.1093/bioinformatics/btv722

Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4, 14. doi:10.1186/1745-6150-4-14

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi:10.1038/nrg2934

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi:10.1038/nmeth.4197

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32, 462–464. doi:10.1038/nbt.2862

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi:10.1038/nbt.3122

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., and Sandhu, M. S. (2018). Long reads: Their purpose and place. *Hum. Mol. Genet.* 27, R234–R241. doi:10.1093/hmg/ddy177

Prakash, C., and Haeseler, A. V. (2017). An enumerative combinatorics model for fragmentation patterns in RNA sequencing provides insights into nonuniformity of the expected fragment starting-point and coverage profile. *J. Comput. Biol.* 24, 200–212. doi:10.1089/cmb.2016.0096

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences. (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65. doi:10.1093/nar/gkl842

Raghupathy, N., Choi, K., Vincent, M. J., Beane, G. L., Sheppard, K. S., Munger, S. C., et al. (2018). Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* 34, 2177–2184. doi:10.1093/bioinformatics/bty078

Ranallo-Benavidez, T. R., Jaron, K. S., Schatz, M. C., and GenomeScope, 2. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi:10.1038/s41467-020-14998-3

Reimers, M., and Carey, V. J. (2006). Bioconductor: An open source framework for bioinformatics and computational biology. *Methods Enzymol.* 411, 119–134. doi:10.1016/S0076-6879(06)11008-3

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21, 245. doi:10.1186/s13059-020-02134-9

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi:10.1038/nbt.2931

Romanel, A., Lago, S., Prandi, D., Sboner, A., and Demichelis, F. (2015). Aseq: Fast allele-specific studies from next-generation sequencing data. *BMC Med. Genomics* 8, 9. doi:10.1186/s12920-015-0084-2

Schmid, M. W., and Grossniklaus, U. (2015). Rcount: Simple and flexible RNA-seq read counting. *Bioinformatics* 31, 436–437. doi:10.1093/bioinformatics/btu680

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346. doi:10.1038/s41576-018-0003-4

Sekhon, A., Singh, R., and Qi, Y. (2018). DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics* 34, i891–i900. doi:10.1093/bioinformatics/bty612

Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358. doi:10.1093/bioinformatics/bts163

Shastry, K. A., and Sanjay, H. A. (2020). "Machine learning for bioinformatics," in *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*. Editors K. G. Srinivasa, G. M. Siddesh, and S. R. Manisekhar (Berlin, Germany: Springer), 25–39. doi:10.1007/978-981-15-2445-5_3

Shen, C.-H. (2019). "Chapter 11 - techniques in sequencing," in *Diagnostic molecular biology*. Editor C. H. Shen (Cambridge, MA, USA: Academic Press), 277–302. doi:10.1016/B978-0-12-802823-0.00011-0

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145. doi:10.1038/nbt1486

Simoneau, J., Gosselin, R., and Scott, M. S. (2020). Factorial study of the RNA-seq computational workflow identifies biases as technical gene signatures. *Nar. Genomics Bioinforma.* 2, lqaa043. doi:10.1093/nargab/lqaa043

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* 21, 1728–1737. doi:10.1101/gr.119784.110

Srivastava, A., Sarkar, H., Gupta, N., and Patro, R. (2016). RapMap: A rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinforma. Oxf. Engl.* 32, i192–i200. doi:10.1093/bioinformatics/btw277

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: The teenage years. *Nat. Rev. Genet.* 20, 631–656. doi:10.1038/s41576-019-0150-2

Szabo, L., and Salzman, J. (2016). Detecting circular RNAs: Bioinformatic and experimental challenges. *Nat. Rev. Genet.* 17, 679–692. doi:10.1038/nrg.2016.114

Taylor-Weiner, A., Aguet, F., Haradhvala, N. J., Gosai, S., Anand, S., Kim, J., et al. (2019). Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* 20, 228. doi:10.1186/s13059-019-1836-7

The Comprehensive R Archive Network (2022). *The comprehensive R archive Network*. Available at: https://cran.r-project.org/.

The RGASP ConsortiumAbril, J. F., Engstrom, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., et al. (2013a). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi:10.1038/nmeth.2714

The RGASP ConsortiumSteijger, T., Sipos, B., Grant, G. R., Kahles, A., Ratsch, G., et al. (2013b). Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10, 1185–1191. doi:10.1038/nmeth.2722

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi:10.1038/nprot.2012.016

Tuerk, A., Wiktorin, G., and Güler, S. (2017). Mixture models reveal multiple positional bias types in RNA-Seq data and lead to accurate transcript concentration estimates. *PLOS Comput. Biol.* 13, e1005515. doi:10.1371/journal.pcbi.1005515

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015). Wasp: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12, 1061–1063. doi:10.1038/nmeth.3582

van Ijzendoorn, D. G. P., Szuhai, K., Briaire-de Bruijn, I. H., Kostine, M., Kuijjer, M. L., and Bovee, J. V. M. G. (2019). Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLOS Comput. Biol.* 15, e1006826. doi:10.1371/journal.pcbi.1006826

Vaquero-Garcia, J., Norton, S., and Barash, Y. (2018). *LeafCutter vs. MAJIQ and comparing software in the fast moving field of genomics*. 463927 Preprint. doi:10.1101/463927

Varadhan, R., and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.* 35, 335–353. doi:10.1111/j.1467-9469.2007.00585.x

Vierra, M., Kingan, S., Tseng, E., Hon, T., Rowell, W., Mountcastle, J., et al. (2021). *From RNA to full-length transcripts: The PacBio Iso-Seq method for transcriptome analysis and genome annotation - PacBio*. Available at: https://www.pacb.com/proceedings/from-rna-to-full-length-transcripts-the-pacbio-iso-seq-method-for-transcriptome-analysis-and-genome-annotation/.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi:10.1093/bioinformatics/btx153

Wang, A. T., Shetty, A., O'Connor, E., Bell, C., Pomerantz, M. M., Freedman, M. L., et al. (2020). Allele-specific QTL fine mapping with PLASMA. *Am. J. Hum. Genet.* 106, 170–187. doi:10.1016/j.ajhg.2019.12.011

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., et al. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 38, e178. doi:10.1093/nar/gkq622

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365. doi:10.1038/s41587-021-01108-x

Workflows for RNA Sequencing (2023). *A guide to Illumina solutions for next-generation RNA sequencing applications*.

Wu, J., Wang, C., Cui, Y., Xu, T., Wang, C., Wang, X., et al. (2019). CircAST: Full-length assembly and quantification of alternatively spliced isoforms in circular RNAs. *Genomics Proteomics Bioinforma.* 17, 522–534. doi:10.1016/j.gpb.2019.03.004

Wu, W., Ji, P., and Zhao, F. (2020). CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol.* 21, 101. doi:10.1186/s13059-020-02018-y

Xia, S., Feng, J., Chen, K., Ma, Y., Gong, J., Cai, F., et al. (2018). Cscd: A database for cancer-specific circular RNAs. *Nucleic Acids Res.* 46, D925–D929. doi:10.1093/nar/gkx863

Xie, J., Ji, T., Ferreira, M. A. R., Li, Y., Patel, B. N., and Rivera, R. M. (2019). Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. *BMC Bioinforma.* 20, 530. doi:10.1186/s12859-019-3141-6

Xu, G., Strong, M. J., Lacey, M. R., Baribault, C., Flemington, E. K., and Taylor, C. M. (2014). RNA CoMPASS: A dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. *PLOS ONE* 9, e89445. doi:10.1371/journal.pone.0089445

Yang, Q., Hu, Y., Li, J., and Zhang, X. (2017). ulfasQTL: an ultra-fast method of composite splicing QTL analysis. *BMC Genomics* 18, 963. doi:10.1186/s12864-016-3258-1

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., et al. (2013). HTQC: A fast quality control toolkit for Illumina sequencing data. *BMC Bioinforma.* 14, 33. doi:10.1186/1471-2105-14-33

Ye, H., Meehan, J., Tong, W., and Hong, H. (2015). Alignment of short reads: A crucial step for application of next-generation sequencing data in precision medicine. *Pharmaceutics* 7, 523–541. doi:10.3390/pharmaceutics7040523

Zakeri, M., Srivastava, A., Almodaresi, F., and Patro, R. (2017). Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics* 33, i142–i151. doi:10.1093/bioinformatics/btx262

Zhang, J., Chen, S., Yang, J., and Zhao, F. (2020). Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nat. Commun.* 11, 90. doi:10.1038/s41467-019-13840-9

Zhang, X.-O., Dong, R., Zhang, Y., Zhang, J. L., Luo, Z., Zhang, J., et al. (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* 26, 1277–1287. doi:10.1101/gr.202895.115

Zheng, Y., Ji, P., Chen, S., Hou, L., and Zhao, F. (2019). Reconstruction of full-length circular RNAs enables isoform-level quantification. *Genome Med.* 11, 2. doi:10.1186/s13073-019-0614-1

Zheng, Y., and Zhao, F. (2020). Visualization of circular RNAs and their internal splicing events from transcriptomic data. *Bioinforma. Oxf. Engl.* 36, 2934–2935. doi:10.1093/bioinformatics/btaa033

Zou, J., Hormozdiari, F., Jew, B., Castel, S. E., Lappalainen, T., Ernst, J., et al. (2019). Leveraging allelic imbalance to refine fine-mapping for eQTL studies. *PLOS Genet.* 15, e1008481. doi:10.1371/journal.pgen.1008481