Review Article

A Comprehensive Methodological Review of Major Developments in Bioinformatics Pipelines for Transcriptomic Data Analysis

Awais Ali[1*], Syed Luqman Ali[1], Abdelkarem Omneya[2]

[1]Department of Biochemistry, Abdul Wali Khan University Mardan, Madan 23200- Pakistan
[2]Department of Chemical Pathology, Medical Research Institute, Alexandria University, Egypt

Received: 21 June, 2024; Accepted: 07 September, 2024

DOI: 10.22037/nbm.v13i1.45616

Abstract

Background: Recent advances in transcriptomics have provided new insights to analyze a wide range of biological data. RNA sequencing (RNA-Seq) is a common method used to study the complete set of RNA molecules (the transcriptome) in different cell types, genetic backgrounds, and environments. While many computational tools exist for analyzing large RNA-Seq datasets, there is still a need to thoroughly compare methods used to separate mixed cell populations (deconvolution).

Materials and Methods: This review highlights recent software and database improvements for processing RNA-Seq data, including steps like matching sequences to the genome, reconstructing RNA molecules, and measuring RNA abundance.

Results: We examine how well different methods work under various experimental conditions and discuss important factors such as data quality, sequence alignment, data visualization, identifying gene expression differences, and data standardization. A novel approach is also introduced: an ensemble learning-based deconvolution method combining multiple techniques to improve accuracy, mitigate data contamination, and reduce errors. Our findings provide valuable guidance for using omics tools effectively and developing better analysis methods. This review offers detailed instructions for planning and evaluating Illumina sequencing experiments.

Conclusion: We cover basic concepts, RNA-Seq analysis steps, computational workflows, and potential difficulties.

Keywords: Transcriptomic, Bioinformatics pipelines, Contamination, Illumine, RNA-Seq analysis

***Corresponding Author:** Awais Ali, Department of Biochemistry, Abdul Wali Khan University Mardan, Mradan 23200- Pakistan. Email: awaisalibio@gmail.com . Intuitional email : awaisali@awkum.edu.pk

Please cite this article as: Ali A, Ali SL, Omneya A. A Comprehensive Methodological Review of Major Developments in Bioinformatics Pipelines for Transcriptomic Data Analysis. Novel Biomed. 2025;13(1):46-60.

Introduction

Recent technological advances have made it possible to measure different biological molecules quickly and accurately. This has led to extensive studies of genes, allowing scientists to compare how molecules vary under different conditions. RNA sequencing (RNA-Seq) is a powerful method that accurately measures the RNA transcript level. It combines finding new RNA molecules with measuring their levels efficiently. However, variation in the results is a significant challenge due to different protocols. Therefore, we need to develop and follow detailed guidelines to help obtain reliable results.

New sequencing technologies focus on studying

individual cells in genomics, transcriptomics, and epigenomics. For example, single-cell RNA sequencing (scRNA-Seq) can identify rare and complex cell types.

Recent technological advances have made it possible to measure different biological molecules quickly and accurately through high-throughput methods**.** RNA sequencing (RNA-Seq) is a powerful technique for measuring RNA transcript levels, combining finding new RNA molecules and measuring their levels efficiently. It has become invaluable in biological research[1,2]. However, variation in the results is a major challenge due to different protocols. Therefore, we need to develop and follow detailed protocols and guidelines to help obtain reliable results[3]. In genomics, transcriptomics, and epigenomics, Next-Generation Sequencing (NGS) technologies are increasingly focused on characterizing individual cells, with single-cell RNA sequencing (scRNA-Seq) being particularly valuable for identifying rare and complex cell populations.

RNA sequencing (RNA-Seq) outperforms microarrays for studying gene activity because it can analyze all RNA molecules simultaneously (complete transcriptome sequencing) and offers a wide range of computational tools for mapping, assembling, and quantifying RNA sequences. In contrast, microarrays can only look at a limited set of RNA molecules, which gives limited profiling through hybridization. RNA-Seq is essential for detecting differentially modulated transcripts, splice variants, and non-coding RNAs (e.g., miRNA, lncRNA, pseudogenes). The large amount of information from RNA-Seq is valuable for many scientific domains, including toxicity prediction, mechanistic investigations, and biomarker discovery, ushering in a new era of molecular exploration[4–6].

Single-cell RNA sequencing (scRNA-Seq) is a revolutionary technique that explores tissue complexity and cell characteristics at the molecular level. Scientists can learn more about cell types, how cells respond to different conditions, and how genes are regulated by analyzing the transcriptome of single cells. This technology has become essential for answering many biological questions, including identifying distinct cell types, understanding cellular diversity in responses, revealing gene expression variability, and deducing gene regulation networks within cell populations. There are pipelines for computational and library-based scRNA-Seq data analysis (Figure 1)[7–10]. These innovations will pave the way for a new era of genetic research and medical breakthroughs[11].

Properly planning an RNA-Seq experiment is important but often overlooked. A well-designed experiment is essential for answering biological questions accurately. Critical factors in RNA-Seq experiments include the choice of library type (single end/paired-end), sequencing depth (number of reads), and the inclusion of an adequate number of replicates for each condition. Transcriptome deconvolution is a technique to determine the types of cells in a tissue sample (RNA sample) based on the genes turned on or off in that sample (gene expression). This helps correct differences between samples. Scientists have developed computer programs to estimate the number of different cell types in a tissue sample using RNA-Seq data. This allows researchers to learn more from large studies like The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium[12,13].

Sample collection: In the fast-paced world of scientific research, collecting and analyzing samples are crucial for understanding complicated phenomena, generating creative solutions, and pushing improvements across numerous disciplines. Traditional sample-collecting procedures have frequently been tedious, time-consuming, and prone to human error. However, the introduction of computer analysis and strong computational tools has transformed how we approach sample gathering.

By exploiting the potential of computational tools, researchers may simplify and optimize the collecting process, resulting in more accurate, efficient, and insightful studies [14,15].

Advantages of computational technologies in sample collection

Efficiency, Accuracy, and enhanced sample diversity: Computational technologies enable unparalleled speed and precision in sample collection, dramatically lowering human error and enhancing overall efficiency. Automated robotic systems with modern sensors and machine learning algorithms can accurately gather and handle samples with minimum human interaction. These technologies provide uniform

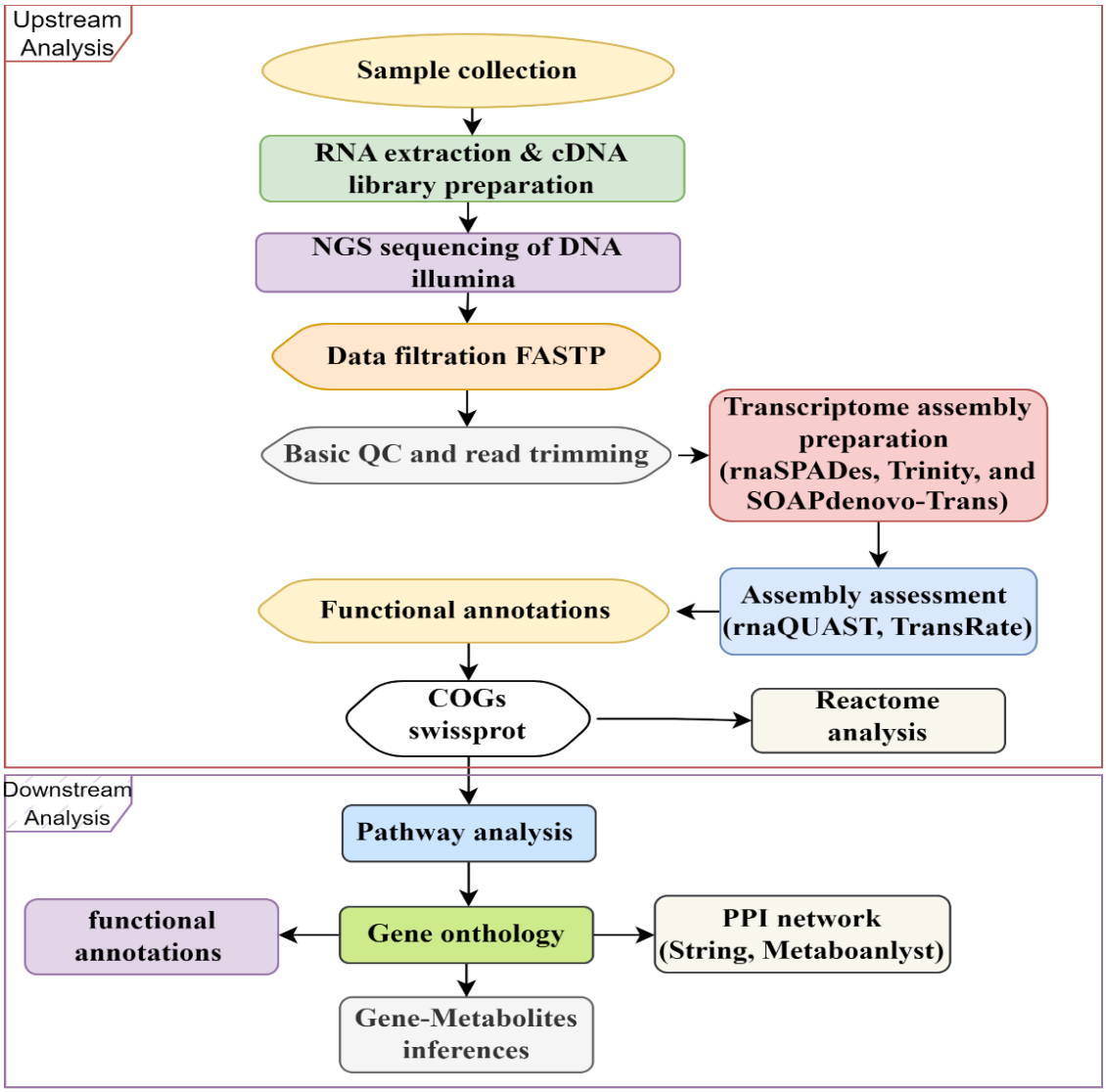

Figure 1. Systematic flow chart of transcriptomics analysis, including Upstream and downstream analysis.

and repeatable data capture, avoiding the discrepancies commonly associated with manual collection [16–18]. Furthermore, real-time data processing capabilities enable researchers to make instant tweaks or revisions throughout the gathering process, resulting in higher-quality data and more reliable analysis[8,1920].

Customized Sampling Techniques: Computational analysis offers a significant advantage in sample collection through the development of tailored sampling strategies. Advanced algorithms enable the optimization of sample acquisition based on specific research objectives and constraints. Researchers can employ computational tools to identify optimal sampling parameters within budgetary and temporal limitations21 by considering sample size, location, and frequency. This personalized approach ensures the collection of highly relevant and representative samples, enhancing the precision and impact of subsequent analyses[22].

Data-driven Insights: Computational analysis significantly accelerates the sample collection while enhancing data analysis capabilities. Advanced analytical techniques like machine learning and data mining can uncover latent patterns, correlations, and trends within collected samples. These derived insights facilitate predictive modeling, hypothesis formulation, and the generation of novel research directions[23].

Researchers can maximize data utility and deepen their understanding of complex phenomena by integrating computational methods with traditional scientific inquiry.

RNA extraction and cDNA library preparation: Continuous advancements in technology and methodology characterize the dynamic landscape of genetic research. Central to the elucidation of gene expression and regulation are RNA extraction and cDNA library construction. Computational tools have revolutionized these foundational processes, enhancing efficiency, accuracy, and reproducibility in genetic studies[24].

RNA extraction, a critical initial step in gene expression analysis, was traditionally laborious and error-prone. However, computational approaches have streamlined this workflow. Researchers can assess sample quality, detect contaminants, and optimize extraction protocols25 by analyzing raw sequencing data. This enables the consistent and rapid production of high-quality RNA, a prerequisite for subsequent downstream analyses[26, 27].

A) Optimizing cDNA Library Preparation: Constructing complementary DNA (cDNA) libraries is vital for investigating gene expression profiles. Historically, cDNA library preparation has been a challenging process that has often yielded biased and low-quality results. However, computational advancements have significantly improved cDNA library construction[28]. Sophisticated algorithms facilitate the optimization of primer selection, reverse transcription conditions, and experimental design, thereby mitigating biases and enhancing the representativeness of the final cDNA library. This technological progress empowers researchers to capture the full complexity of gene expression patterns accurately.

B) Data Integration and Analysis: Computational tools are instrumental in streamlining RNA extraction, cDNA library preparation, and subsequent data integration and analysis stages. The vast quantities of genetic data generated necessitate sophisticated computational approaches to extract meaningful insights. By employing advanced bioinformatics methodologies, researchers can identify differential gene expression patterns, uncover complex regulatory networks, and correlate findings across multiple datasets[29]. Furthermore, integrating RNA-Seq data with other omics data types, such as proteomics and epigenomics, provides a holistic perspective on gene expression dynamics.

C) Driving Scientific Discoveries: The synergistic interplay between computational tools and RNA extraction/cDNA library preparation has catalyzed groundbreaking discoveries across diverse scientific domains. Researchers are empowered to elucidate disease etiologies, investigate developmental processes, and identify potential therapeutic targets with unprecedented precision. Integrating computational methodologies with experimental approaches has accelerated research timelines, fostered reproducibility, and facilitated collaborative knowledge sharing within the scientific community[30].

Library preparation procedure care to reduce contamination for Transcriptome analysis on a molecular level

The RNA-Seq sample preparation process is susceptible to biases, including mRNA enrichment and RNA fragmentation, which can compromise data accuracy. Adherence to rigorous laboratory protocols, encompassing sterile equipment and high-quality reagents, is essential to mitigate contamination and enhance data quality. Moreover, the complex composition of biological samples, such as tick organs, necessitates additional considerations, including applying complementary techniques like fluorescence in situ hybridization (FISH) to overcome interference and address potential contamination issues[31].

Recent advancements in computational capabilities and reduced sequencing costs have facilitated extensive genomic reconstruction from diverse sample types. However, the rapid accumulation of draft genomes has outpaced the capacity for comprehensive annotation regarding the recovery of microbial genomes from complex environmental samples. Ensuring the accuracy of these reconstructed genomes is essential for downstream analyses. CheckM2 is a state-of-the-art machine learning tool designed to assess the quality of microbial genomes, including those derived from isolates, single cells, and metagenomic-assembled genomes (MAGs)[32]. Distinguishing itself from its predecessor, CheckM2 accurately predicts genome completeness and contamination without relying on taxonomic information, especially excelling in

Table 1: Major tools for Reducing contamination from RNA-seq data and genome data.

Database	Function-Based	Contamination Reduction	References
Cleanseq	Python-Based Pipeline	98%	35
BLAST	Python, Linux, And Windows	Depending Upon Raw/Short Sequences	36
Burrows-Wheeler Aligner (BWA)	Python/ Galaxy/ Cloudbiolinux	86.4%	37
Sequence Alignment/Map (Samtools)	Mac, Linux, Or Windows	60-80%	38
The Genome Analysis Toolkit (GATK)	Linux, Macos X, Java 1.8	97- 99%	39
Checkm	Python	90%	32
Kraken	Linux	98-99.1%	40
Confindr	Python-Based Pipeline	Approximately 95.5%	41
Fastq Screen	Perl, Linux-Based	Approximately 89.4%	42
Mutscan	R Package, C++	Detection And Remove Duplication	43
Decontaminer	Perl	90%	44
Microdecon	R Package	98.1%	45
Haplocheck	Windows	97%	46
Deconseq	Perl, Linux/Unix, Or Mac O	64-72%	47

evaluating genomes from understudied lineages characterized by compact genome sizes.

Contaminant data tools are crucial for maintaining data integrity in scientific research. Implementing extraction and no-template controls, coupled with advanced bioinformatic techniques, facilitates identifying and removing contaminant sequences. These rigorous measures ensure data reliability and enhance the overall quality of research outcomes[33].

Researchers should incorporate additional strategies to minimize contamination in RNA-Seq experiments, including control sequences, careful sample size estimation, and meticulous attention to batch effects. These precautionary steps collectively contribute to the generation of high-quality and reproducible data. Additionally, specialized bioinformatics tools and databases have been developed for detecting, removing, and verifying contaminants within whole genome sequencing (WGS) data (Table 1).

Additionally, to mitigate potential contamination, sequences exhibiting significant homology to both the horse and human genomes were excluded from the analysis unless designated explicitly for human contamination filtering. All human matches, including potential paralogs, were retained as a conservative measure for Illumina sequencing data. Alternatively, putative horse sequences were filtered based on edit distance, excluding those more similar to the human genome (hg19) than the horse genome (EquCab2) to safeguard data integrity[34]. HTS of cDNA enables the identification of many RNA molecules in a cell or tissue at a particular temporal point.

Next-generation sequencing: NGS has become the predominant technology for the sequencing and analysis of mRNA and small RNA due to its cost-effectiveness and efficiency. Prominent NGS platforms include the Roche 454, Illumina Solexa, and ABI SOLiD systems[48].

NGS has revolutionized genomic research by enabling rapid and precise genome analysis. However, the massive volume of data generated presents significant computational challenges. Typically, NGS produces tens of millions of short sequence reads (100-200 base pairs) derived from longer DNA fragments (200-400 base pairs)[49].

A primary advantage of computational techniques in next-generation sequencing is their capacity for managing large-scale datasets. The advent of high-throughput sequencing technologies has enabled the generation of terabytes of sequencing data from a single experiment in a timely and cost-effective way[50]. Sophisticated computational tools are essential to analyze and interpret these immense datasets effectively. By leveraging parallel processing,

advanced algorithms, and optimized data structures, these tools facilitate extracting meaningful biological insights from raw sequencing data[51].

Additionally, computational tools exhibit remarkable versatility in processing next-generation sequencing data. Their functional repertoire encompasses various analyses, ranging from basic quality control and read alignment to complex tasks such as variant calling, de novo assembly, and transcriptome analysis. This flexibility allows researchers to tailor computational pipelines to specific research objectives[39]. Moreover, the seamless integration of these tools with existing bioinformatics infrastructure facilitates data management and interoperability.

Moreover, computational methods are necessary to address the inherent challenges of NGS data, including sequencing errors, biases, and complex genetic variants. Advanced algorithms and statistical models facilitate identifying and correcting sequencing inaccuracies, mitigating library preparation biases, and precisely detecting genetic variations such as single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variants. Moreover, these computational approaches enable the functional characterization of genetic variants by assessing their impact on protein structure, gene expression, and disease phenotypes[52, 53].

Interestingly, computational technologies have significantly contributed to the democratization of genomic research. User-friendly interfaces, comprehensive documentation, and active online communities have facilitated accessibility to NGS technologies and complex data analysis for researchers with diverse backgrounds[54]. This democratization fosters collaboration, knowledge sharing, and accelerated scientific discovery, ultimately advancing our understanding of the genome and driving the development of personalized medicine. These tools are crucial collaborators in the era of next-generation sequencing. Their capacity to manage vast datasets, their versatility in data processing, and their proficiency in addressing NGS-specific challenges position them as invaluable assets in genomic research. By harnessing the power of computational tools, researchers can fully exploit the potential of NGS data, driving groundbreaking discoveries and transformative applications across diverse fields[39].

The rapid advancement of sequencing technology has led to the development of multiple next-generation sequencing platforms. A standard workflow for data analysis and processing, often utilizing publicly available bioinformatics tools, is depicted in Figure 1. These platforms generate vast quantities of short-sequence reads at high speed. For instance, the Illumina-Solexa system can produce approximately 50 million short reads (30-50 nucleotides) in approximately three days[55]. The ABI-SOLiD system offers comparable throughput using a different technology[56]., while the Roche-454 system generates fewer but longer reads. The sequencing throughput continues to increase rapidly, highlighting the dynamic nature of this field[57].

Raw Data Processing and Quality Assessment: The transcriptomic RNA-seq data processing can be divided into three steps.

Quality control (QC): Data excellence is crucial in bulk and single-cell RNA sequencing analyses [58]. Bulk RNA sequencing entails evaluating RNA quality, with subpar samples excluded from sequencing. The quality of initial sequencing data is assessed through parameters like read count and read quality scores. An essential metric is the mapping quality score, indicating the likelihood of post-read alignment misplacement and providing an overall data quality measure. This assessment covers per-base and per-sequence quality, duplicate identification, GC content, and the detection of overrepresented sequences or adapters. Notable tools for swift diagnostics include FastQC (for Illumina) and NGSQC (platform-independent). Subsequently, the removal of low-quality reads and bases, as well as adapter trimming, can be efficiently executed using tools such as FASTX-Toolkit and Trimmomatic[59].

Mapping: In RNA-seq studies, data can be aligned to the genome or a pre-existing transcriptome. Genome alignment is chosen when the aim is to discover new isoforms, enhancing our understanding of gene expression diversity. For a typical high-quality human sample library, >70% of the mapping rate to the human genome is expected. TopHat2 and STAR are popular algorithms (spliced aware mapper, i.e., accounting for splice junctions during mapping) for genome mapping[60]. Meanwhile, transcriptome-based mapping can be performed using bowtie2/bwa[61,62]vav.

Transcriptome-based mapping yields a lower percentage of reads mapped due to a lack of sequences of unannotated transcripts. When a reference genome is unavailable (for non-model organisms) or incomplete, de novo transcriptome assembly should be performed. RapMap gives rapid results; more ever, it is a sensitive and accurate tool for mapping RNA-seq reads to transcriptomes[63]. Trinity[64], Oases[65], and SOAP denovo-Trans[66] are popular algorithms for transcriptome assembly. Longer reads or paired-end reads will aid in the de novo assembly. All the reads from multiple samples should be combined, and then the combined assembly is done to generate the reference set of contigs. These reference contigs can be used for mapping, expression quantification, and comparison across different groups.

Quantification*:* The important step in RNA-seq data analysis is to estimate gene and transcript abundance. HTSeq 2.0 or featureCounts[67,68], quantify the abundance by aggregating the number of mapped reads in each exon (determined by Gene Transfer Format [GTF] file), as desired. In addition to aggregating the raw counts, various per-sample normalization methods have been developed to account for feature-length and library sizes. RPKM (reads per kilobase of exon model per million reads), FPKM (Fragment per kilobase of exon model per million reads), and TPM (Transcript Per Million) are common measures of expression provided by various analysis tools: cufflinks[69,] RSEM[70], eXpress[71] and Kallisto. FPKM is used for paired-end read sequencing, and RPKM is used for single-end reads.

Meanwhile, unlike RPKM, TPM skips normalization for gene length after sequencing depth normalization. It makes the sum of all TPMs in all samples the same and aids in comparing expression profiles among different samples. Transcript abundance is computed as an average of a normalized abundance of exons in the transcript.

Besides, RNA-Seq by Expectation-Maximization (RSEM) software is precise and easy to use for calculating transcript abundances from RNA-Seq data[70]. Additionally, RSEM has provided invaluable advice for the economical design of quantification experiments using the currently pricey RNA-Seq technology. With de novo transcriptome assemblies, it is especially helpful for quantification because it does not rely on the existence of a reference genome.

Transcriptome analysis and list of the most commonly used bioinformatics pipelines/resources:

RNA-seq analysis can be grouped into two types: those using a reference genome and those not relying on one. Regardless of the approach, the RNA-seq workflow consists of four main stages: aligning and assembling the data, quantifying it, normalizing, and performing differential expression (DE) analysis, which helps identify the genes showing differing expression between two sets of samples (Figure 2)[73]. The software and input files needed vary for each phase depending on the specific analytical method chosen. Noteworthy pipelines and their alternatives can be found in Table 2.

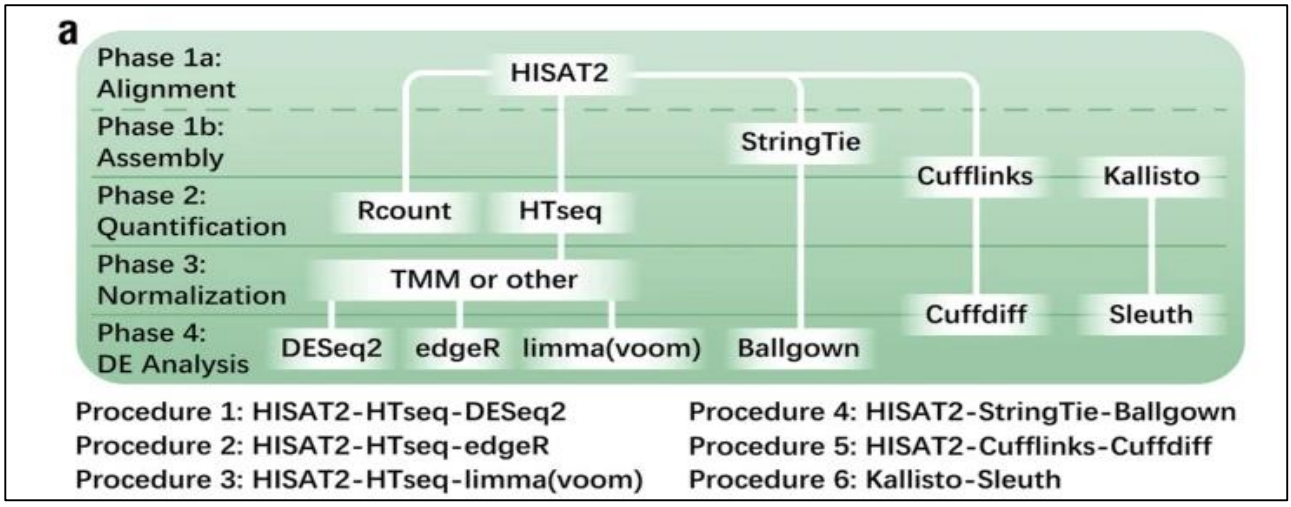

Figure 2. A systematic flow chart representation of the six procedures for RNA-seq analysis was compared.

Table 2: List of bioinformatics pipelines/resources for transcriptome analyses, mapping sequences, reconstruction and quantification.

Method	Tools	Data input format	References
Mapping Sequences	STAR	FASTA or FASTQ	84
	Stringtie	Gene Transfer Format (GTF)	75
	ZOOM	FASTA, FASTQ, CFASTA	83
	MicroRazerS	FASTA	100
	TopHat	FASTA, FASTQ, GFF	76
	SpliceMap	FASTA, FASTQ	102
	RNA-Mate	CFASTA	103
	mrsFAST	FASTA, FASTQ	104
	VASA-seq	single cell	99
Reconstruction	Cuffl inks	Galaxy (function)	69
	eXpress	FASTA	105
	StringTie	multi-FASTA	106
	VASA-seq	single cell	99
Quantifications	RSEM	FASTA, FASTQ	70
	Sailfish	FASTA	107
	Express	FASTA	105
	Alevin-fry	sc/snRNA	108
	VASA-seq	single cell	99

Phase 1, the alignment and assembly phase, begins with data files in the FASTQ format containing raw sequenced reads[74]. This involves aligning these reads to a reference genome and linking them to transcripts. The software tool HISAT[75], an advanced version of TopHat[76], is widely used due to its efficiency and lower computing requirements than STAR. These programs use unique algorithms to accurately align the reads to neighboring exons in the reference genome, enhancing the mapping success. They can even work with incomplete reference genome annotations by filling the gaps and assembling exon-intron structures. Given the massive data generated, specialized algorithms are needed for mapping short sequences to the genome, where tools like BLAST and BLAT fall short[36]. Several recent programs utilize techniques like hashing and short key indexing, with MAQ standing out as the top choice due to its time-saving, contamination reduction, and low error rate[77]. Check Table 3 for a rundown of these essential tools.

Phase 2, in the first step, gene expression levels are assessed by analyzing how sequencing reads align with the reference genome. This crucial process provides insights into the activity of genes and their transcripts, forming a foundational step in genomic research. Commonly used quantification tools include Rcount, HTseq[68], StringTie[75], and Cufflinks[69]. Table 2 Gene expression analysis tools can be categorized into two groups, hinging on the assessment criteria of either counts or FPKM values. Tools like Rcount, HTseq, and Kallisto operate based on counts, while StringTie and Cufflinks rely on FPKM values. Both HTseq and Rcount exclusively tally reads uniquely mapped to a single gene, discarding those aligned to multiple positions or overlapping with more than one gene. This dichotomy in evaluation methods enhances precision and diversity in gene expression studies, providing scientists with powerful options for their research endeavors[68], while Rcount assigns weights to each alignment of a multiread[75]. Therefore, Rcount is better at counting multireads and gene-overlapping regions.

Phase 3, normalization step in which a matrix of expression values can be modeled to determine which gene or transcript features are likely to have altered expression levels. This step may have a major impact on the results of the DE analysis.

In phase 4, the primary task in RNA-seq data analysis is to identify shifts in gene expression.

Table 3: Tools and databases used for illumine seq data preparation and assessment.

Data BASE	Link	Function-based	Reference
RMAP	http://rulai.cshl.edu/rmap/	Python	78
MAQ	http://maq.sourceforge.net	C++, Perl	79
CloudBurst	http://cloudburst-bio.sourceforge.net/	Python	80
SeqMap	http://biogibbs.stanford.edu/~jiangh/SeqMap/	Linux	81
SHRiMP	http://compbio.cs.toronto.edu/shrimp/	Linux 2.6	82
ZOOM	http://www.bioinfor.com/zoom	Desktop application	83
SOAPv1	https://github.com/aquaskyline/SOAPdenovo-Trans	command-driven program	66
PASS-	http://pass.cribi.unipd.it/	Command-line tool	84
MOM	http://mom.csbc.vcu.edu/	Command-line tool	85
NovoAlign	http://www.novocraft.com/	Command-line tool/Linux-64, OSX-64	86
ProbeMatch	http://www.cs.wisc.edu/~jignesh/probematch/	Command-line tool	87
BFAST	http://genome.ucla.edu/bfast	Command-line tool/ GPL-3.0	88
NovaSeq 6000	https://www.illumina.com/systems/sequencing-platforms/novaseq.html	command line instrument	89
SOAPv2	http://soap.genomics.org.cn/	Command-line tool	66
Trans-Abbys	https://www.bcgsc.ca/resources/software/trans-abyss	Galaxy/ CloudBio Linux	90
BUSCO	https://busco.ezlab.org/	Command-line tool	91
DOGMA	https://domainworld-services.uni-muenster.de/dogma/	Web application	92
TransRate	https://github.com/blahah/transrate	Command-line tool	93
CD-HIT	http://weizhongli-lab.org/cd-hit/	Linux, C++	94

Once gene and transcript expression levels are approximated, statistical methods are applied to pinpoint variations in expression across different experimental groups, helping to distinguish genuine gene expression changes from random fluctuations. These tests evaluate the significance of changes based on specific assumptions about random errors' distribution, distinct from errors linked to experimental variables, and these cannot be rectified through normalization. Key software tools for this analysis stage include DESeq2[95], edgeR[96], limma, Ballgown[61], Cuffdiff[97], and Sleuth[98].

VASA-seq is a cutting-edge technique for single-cell transcriptome analysis. Its exceptional sensitivity, complete RNA coverage, and high throughput make it a standout method. VASA-seq's efficient RNA processing and compatibility with various experimental setups, including plate-based and droplet microfluidic, enhance its applicability. Notably, it has provided valuable insights into alternative splicing during mammalian development, shedding light on critical processes like blood development and heart morphogenesis[99]. VASA-seq is compatible with plate-based and droplet microfluidic systems, making it highly versatile. Importantly, it provides comprehensive insights into alternative splicing, particularly in critical biological processes such as blood and heart development[99]. The specialized tool MicroRazerS offers a fast and flexible solution for mapping small RNA sequences to reference genomes, outperforming traditional methods like MegaBLAST in speed and efficiency[100].

For small RNA analysis, MicroRazerS[100] is a specialized tool designed for mapping sequences onto reference genomes. It offers a significant speed advantage over traditional methods like Mega BLAST while maintaining comparable accuracy. MicroRazerS's enhanced sensitivity, manageability,

and adjustability contribute to its effectiveness in small RNA analysis.

Overcoming the Challenge of Duplicated Genomic Regions in RNA Sequencing: Analyzing large RNA sequencing datasets presents a challenge with duplicated genomic regions, which can complicate the accurate alignment of sequencing reads due to multiple suitable alignment sites. This ambiguity can lead to inaccurate quantification of genes or transcripts. As a result, various methods have been developed to address the presence of multireads, each requiring distinct strategies for precise quantification(summarized in Table 2) [101].

Downstream analysis

A) Analyzing the Functional Relevance of Differentially Expressed Genes and Gene Modules: Understanding the functional significance of differentially expressed genes (DEGs) and gene modules is essential for interpreting RNA-Seq data. By analyzing pathways (like KEGG)[109], Gene Ontology terms, and existing gene signatures such as MSigDB, GeneSigDB, and DSigDB, researchers can identify the biological processes, molecular functions, and cellular components associated with these genes. Statistical tests like Fisher's exact and KS tests are commonly used to determine if a particular module is enriched for genes related to a specific function[110]. For DEGs from RNA-Seq data, GO-seq is a recommended tool that addresses gene length bias, providing more accurate identification of differentially expressed genes.

B) Analyzing gene modules and interactions and Biological Systems: Genes do not function in isolation; other genes influence their activities within the cell. Transcription factors, for example, regulate the expression of downstream genes. Effective gene collaboration is essential for various cellular processes, including DNA replication, growth, and cell death. To understand these interactions, researchers systematically analyze biological systems. By identifying sets of genes associated with specific biological conditions using techniques like clustering, biclustering, and differential network analysis, they can uncover functional relationships and gain insights into the complex interplay of genes within cells[111].

C) Clustering and Network Analysis in Bioinformatics: Clustering is an unsupervised machine-learning technique to group similar data points. Bioinformatics commonly identifies patterns within gene expression data, such as grouping genes with similar expression profiles or clustering samples based on their shared characteristics. Hierarchical clustering and K-means clustering are two widely used algorithms for this purpose[112].

Biclustering is a dynamic approach different from traditional clustering in that it identifies groups of genes and samples that are co-expressed or co-regulated. Unlike traditional clustering, biclustering considers gene and sample dimensions simultaneously, allowing for the discovery of hidden patterns and relationships within large datasets. The Biclustering package[113] provides researchers with various biclustering algorithms to explore different types of biclusters, offering valuable insights into complex biological systems[114].

Differential network analysis is a supervised machine learning technique used to identify changes in gene coexpression patterns between different conditions or time points. Analyzing gene expression data can reveal new and evolving relationships between genes, such as discovering changes in musical harmonies over time. Differential network analysis provides valuable insights into the dynamic nature of biological systems and helps researchers understand how gene interactions change in response to various stimuli or conditions[115,116]

(D) functional annotations and pathways inferences: Metabolic pathways can be enriched with insights from leading databases and web servers. KEGG (https://www.genome.jp/kegg/) offers a pathway analysis tool that aids researchers in interpreting gene lists within a biological context. This platform provides valuable insights into the roles of genes and proteins within intricate cellular networks[109], METACyc (https://metacyc.org/) contains metabolic pathways and enzymes spanning many organisms. It is invaluable for studying metabolic processes and uncovering metabolic pathway associations within different biological systems[117]. MetaboAnalyst (https://www.metaboanalyst.ca/) is an advanced web application that enables metabolomics data analysis. It assists researchers in comprehending complex metabolic profiles, identifying biomarkers, and extracting meaningful insights from large datasets[118]. Metascape (https://metascape.org/gp/index.html) is a

web server for gene annotation and functional enrichment analysis. It aids in identifying enriched biological functions, pathways, and processes within gene sets, allowing researchers to glean deeper insights into the functional implications of their data[119]. PlantCyc (https://plantcyc.org/) is a curated database on plant metabolic pathways and enzymes. It offers a comprehensive resource for studying plant-specific metabolic processes, contributing to a better understanding of plant biology and biochemistry; Reactome (https://reactome.org/) for human's knowledgebase of biological pathways and reactions, catering to diverse organisms. Its pathway analysis tool aids researchers in unraveling the interconnectedness of genes and proteins, shedding light on complex cellular mechanisms, plant reactome for plants[120], MonaGO (https://bio.tools/monago) identifies significant Gene Ontology terms, highlighting essential biological functions, WikiPathways (https://www.wikipathways.org/)[121] similarly Gene Set Enrichment Analysis (GSEA) available at https://www.gsea-msigdb.org/gsea/index.jsp[122], Cytoscape 2.0 (https://cytoscape.org/) Cytoscape visualizes molecular networks, clarifying complex interactions in pathways[9,123], similary genome.jp (https://www.genome.jp/), Pathway Commons (pathwaycommons.org) aggregates pathway data, simplifying exploration of intricate networks, BioCyc (https://biocyc.org/) encompasses diverse organisms' pathways, offering comprehensive insights[17,124–126], Human Protein Reference Database (https://www.hprd.org/) HPRD offers comprehensive information on human proteins, including their interactions, post-translational modifications, and associated diseases. It contributes to our understanding of protein functions in health and disease[127]. HumanCyc (https://humancyc.org/) focuses on metabolic pathways and enzyme-catalyzed reactions in humans, and It aids researchers in exploring the intricate metabolic processes occurring within the human body[128], and the NextProt platform (https://www.nextprot.org/) [7,15,27,129]. These resources provide comprehensive information through enrichment analysis, offering valuable insights into specific pathways. Among the databases and web servers listed, KEGG is a leading resource for pathway enrichment analysis, offering valuable insights into the biological pathways of various organisms. Reactome, in contrast, is influential in delivering detailed information on gene expression, focusing on human biology. Both databases are essential for understanding complex biological processes, but each specializes in distinct research areas.

Gene ontology and pathway analysis: The Gene Ontology (GO) knowledgebase is a vast hub of gene function information vital for humans and machines. It is crucial for analyzing extensive molecular biology and genetics experiments in biomedical research (http://geneontology.org/) [130]. similarly, The *Generic GO Term Finder* (https://go.princeton.edu/GOTermFinder/GOTermFinder_help.shtml) is a user-friendly tool that quickly identifies significant GO terms shared among a list of genes from any organism. With the modern *LAGO* implementation and the *GO Term Mapper*, users can gain further insights into functional relationships between genes and bin genes into broad categories similar to *BUSCO (https://busco.ezlab.org/)* and *OrthoFinder (https://github.com/davidemms/OrthoFinder)*[131]. These tools are essential to any researcher or bioinformatician's toolkit for analyzing gene function and relationships. However, the most common use is PantherDB (https://pantherdb.org/)[132]. Nevertheless, KEGG, METACyc, Reactome, WikiPathways, Pathway Commons, and BioCyc provide comprehensive insights into molecular interactions, cellular processes, and organismal pathways, aiding researchers in understanding gene and protein roles within complex biological networks.

Conclusion

Transcriptomic studies continue to be a cornerstone of biomedical research. This review emphasizes the significance of robust analytical approaches for interpreting transcriptomic data. By highlighting the importance of careful experimental design and the use of advanced data analysis techniques in addition to summarization of different deconvolution methods and ensemble learning approaches, this review contributes to improving the accuracy and reliability of transcriptomic analysis. The insights offered in this

paper will be instrumental in advancing our understanding of gene expression and its role in health and disease.

Acknowledgment

I sincerely thank the mentors at my institution whose unwavering support was instrumental in completing this dissertation. Their guidance enriched my academic knowledge and provided invaluable direction during critical junctures. Special appreciation goes to my colleagues, whose steadfast encouragement propelled me to give my utmost to see this endeavor to its fruition.

Conflict of interest

The authors further declare that they have no conflict of interest.

References

1. Nie L, Wu G, Culley DE, Scholten JC, Zhang WJCrib. Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. 2007;27(2):63-75.
2. Alami MM, Ouyang Z, Zhang Y, et al. The Current Developments in Medicinal Plant Genomics Enabled the Diversification of Secondary Metabolites' Biosynthesis. Int J Mol Sci. 2022;23:15932.
3. Lim WK, Mathuru ASJCZ. Design, challenges, and the potential of transcriptomics to understand social behavior. 2020;66(3):321-30.
4. Pathak RK, Kim J-MJFiVS. Vetinformatics from functional genomics to drug discovery: insights into decoding complex molecular mechanisms of livestock systems in veterinary science. 2022;9:1008728.
5. Yan X, Hu Z, Feng Y, et al. Comprehensive genomic characterization of long non-coding RNAs across human cancers. Cancer Cell. 2015;28:529–40.
6. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. 2010;38(18):e178-e.
7. ALi A, Manzoor U, Ali S, Marsool M, Parida P, Marsool A, et al. Currently trending and futuristic biological modalities in the management of different types of diabetes: A comprehensive review. 2023;30:2948-70.
8. Ali SL, Ali A, Alamri A, et al. Genomic annotation for vaccine target identification and immunoinformatics-guided multi-epitope-based vaccine design against Songling virus through screening its whole genome encoded proteins. Front Immunol. 2023;14:1284366.
9. Dinh P, Tran C, Dinh T, et al. Hsa_circRNA_0000284 acts as a ceRNA to participate in coronary heart disease progression by sponging miRNA-338-3p via regulating the expression of ETS1. J Biomol Struct Dyn. 2023;1–14.
10. Peng H, Kaplan N, Lavker RMJIO, Science V. Single cell RNA seq (scRNA-seq) defines the early and late corneal epithelial transit amplifying (TA) cells. 2019;60(9):1370.
11. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50:1–14.
12. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. 2020;21:1-15.
13. Rondina MT, Voora D, Simon LM, Schwertz H, Harper JF, Lee O, et al. Longitudinal RNA-Seq analysis of the repeatability of gene expression and splicing in human platelets identifies a platelet SELP splice QTL. 2020;126(4):501-16.
14. Mitra S, Drautz-Moses DI, Alhede M, Maw MT, Liu Y, Purbojati RW, et al. In silico analyses of metagenomes from human atherosclerotic plaque samples. 2015;3:1-14.
15. Saleem Naz Babari I, Islam M, Saeed H, Nadeem H, Imtiaz F, Ali A, et al. Design, synthesis, in-vitro biological profiling and molecular docking of some novel oxazolones and imidazolones exhibiting good inhibitory potential against acetylcholine esterase. 2024:1-18.
16. Wang L, Xi Y, Sung S, Qiao HJBg. RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. 2018;19:1-13.
17. Nwanna E, Ojo R, Shafiq N, et al. An In Silico In Vitro and In Vivo Study on the Influence of an Eggplant Fruit (Solanum anguivi Lam) Diet on Metabolic Dysfunction in the Sucrose-Induced Diabetic-like Fruit Fly (Drosophila melanogaster). Foods. 2024;13:559.
18. Shafiq N, Arshad M, Ali A, et al. Integrated computational modeling and in-silico validation of flavonoids-Alliuocide G and Alliuocide A as therapeutic agents for their multi-target potential: Combination of molecular docking, MM-GBSA, ADMET and DFT analysis. South African J Bot. 2024;169:276–300.
19. Hajar A, Swathi NL, Ali A. Immunological Insights Into Nutritional Deficiency Disorders. In: Causes and Management of Nutritional Deficiency Disorders. IGI Global. 2024;60–83.
20. Cristancho MA ,Botero-Rozo DO, Giraldo W, Tabima J, Riaño-Pachón DM, Escobar C, et al. Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales. 2014;5:594.
21. Slyper M, Porter CB, Ashenberg O, Waldman J, Drokhlyansky E, Wakiro I, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. 2020;26(5):792-802.
22. Brown CL, Scott H, Mulik C, Freund AS, Opyr MP, Metz GA, et al. Fecal 1H-NMR metabolomics: A comparison of sample preparation methods for NMR and novel in silico baseline correction. 2022;12(2):148.
23. Sarkar H, Srivastava A, Bravo HC, Love MI, Patro RJB. Terminus enables the discovery of data-driven, robust transcript groups from RNA-seq data. 2020;36(Supplement_1):i102-i10.
24. Lu C, Meyers BC, Green PJ. Construction of small RNA cDNA libraries for deep sequencing. Methods. 2007;43:110–7.
25. Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, et al. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. 2008;44(1):3-12.

26. Inglis PW, Pappas MdCR, Resende LV, Grattapaglia DJPo. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. 2018;13(10):e0206085.
27. Manzoor U, Ali A, Ali SL, Abdelkarem O, Kanwal S, Alotaibi SS, et al. Mutational screening of GDAP1 in dysphonia associated with Charcot-Marie-Tooth disease: clinical insights and phenotypic effects. 2023;21(1):119.
28. Olivares D, Perez-Hernandez J, Perez-Gil D, Chaves FJ, Redon J, Cortes RJJoTM. Optimization of small RNA library preparation protocol from human urinary exosomes. 2020;18:1-9.
29. Williams Z, Ben-Dov IZ, Elias R. Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. Proc Natl Acad Sci. 2013;110:4255–60.
30. Anastasakis DG, Jacob A, Konstantinidou P, Meguro K, Claypool D, Cekan P, et al. A non-radioactive, improved PAR-CLIP and small RNA cDNA library preparation protocol. 2021;49(8):e45-e.
31. Greay TL, Gofton AW, Paparini A. Recent insights into the tick microbiome gained through next-generation sequencing. Parasit Vectors. 2018;11:1–14.
32. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GWJGr. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. 2015;25(7):1043-55.
33. Djebali S, Wucher V, Foissac S, Hitte C, Corre E, Derrien TJERM, et al. Bioinformatics pipeline for transcriptome sequencing analysis. 2017:201-19.
34. Schubert M, Ginolhac A, Lindgreen S, Thompson JF, Al-Rasheid KA, Willerslev E, et al. Improving ancient DNA read mapping against modern reference genomes. 2012;13:1-15.
35. Wang C, Xia Y, Liu Y, Kang C, Lu N, Tian D, et al. CleanSeq: A pipeline for contamination detection, cleanup, and mutation verifications from microbial genome sequencing data. 2022;12(12):6209.
36. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32:W20–W25.
37. Abuín JM, Pichel JC, Pena TF, Amigo JJB. BigBWA: approaching the Burrows–Wheeler aligner to Big Data technologies. 2015;31(24):4003-5.
38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. 2009;25(16):2078-9.
39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. 2010;20(9):1297-303.
40. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:1–12.
41. Low AJ, Koziol AG, Manninger PA, Blais B, Carrillo CDJP. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. 2019;7:e6995.
42. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research; 7.
43. Chen S, Huang T, Wen T, Li H, Xu M, Gu JJBb. MutScan: fast detection and visualization of target mutations by scanning FASTQ data. 2018;19:1-11.
44. Granata I, Sangiovanni M, Guarracino M. DecontaMiner: a pipeline for the detection and analysis of contaminating sequences in human NGS sequencing data. In: Dynamics of mathematical models in biology. Springer. 2016;137–48.
45. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KRJED. microDecon: A highly accurate read-subtraction tool for the post- sequencing removal of contamination in metabarcoding studies. 2019;1(1):14-25.
46. Weissensteiner H, Forer L, Fendt L, Kheirkhah A, Salas A, Kronenberg F, et al. Haplocheck: Phylogeny-based contamination detection in mitochondrial and whole-genome sequencing studies. 2020:2020.05. 06.080952.
47. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6:17288.
48. Ondov BD, Varadarajan A, Passalacqua KD, Bergman NHJB. Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. 2008;24(23):2776-7.
49. Kchouk M, Gibrat J-F, Elloumi M. Generations of sequencing technologies: from first to next generation. Biol Med; 9.
50. Goktas TM. K-mer-based data structures and pipelines for sequence mapping and analysis: University of British Columbia; 2023.
51. Freire B, Ladra S, Paramá JR. Memory-efficient assembly using Flye. IEEE/ACM Trans Comput Biol Bioinforma. 2021;19:3564-77.
52. Ansorge WJ. Next generation DNA sequencing techniques and applications. N Biotechnol 2010; 27: S3.
53. Anderson MW, Schrijver I. Next generation DNA sequencing and the future of genomic medicine. Genes (Basel). 2010;1:38–69.
54. Padhan JK, Kumar P, Sood H, Chauhan RSJMP-IJoP, Industries R. Prospecting NGS-transcriptomes to assess regulation of miRNA-mediated secondary metabolites biosynthesis in Swertia chirayita, a medicinal herb of the North-Western Himalayas. 2016;8(3):219-28.
55. Hu Z, Cheng L, Wang H. The Illumina-solexa sequencing protocol for bacterial genomes. In: Bacterial Pangenomics. Springer. 2015;91–7.
56. Mazza T, Castellana S. Multi-Sided compression performance assessment of ABI SOLiD WES data. Algorithms. 2013;6:309–18.
57. Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol. 2009;25:195–203.
58. Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics. 2021;37:963–7.
59. Panwar MBS. Quality Control and Differential Analysis Tools for Sequencing Data with User-Friendly GUI Implementation Based on PySimpleGUI. 2023.
60. Bianchi A, Di Marco A, Pellegrini C. Comparing HISAT and STAR-based pipelines for RNA-Seq Data Analysis: a real experience. In: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). IEEE. 2023;218–24.
61. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek

JTJNb. Ballgown bridges the gap between transcriptome assembly and expression analysis. 2015;33(3):243-6.
62. Guan Y, Li M, Qiu Z, Xu J, Zhang Y, Hu N, et al. Comprehensive analysis of DOK family genes expression, immune characteristics, and drug sensitivity in human tumors. 2022;36:73-87.
63. Srivastava A, Sarkar H, Gupta N, Patro RJB. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. 2016;32(12):i192-i200.
64. Sewe SO, Silva G, Sicat P, et al. Trimming and validation of illumina short reads using Trimmomatic, Trinity assembly, and assessment of RNA-seq data. In: Plant bioinformatics: Methods and protocols. Springer. 2022;211–32.
65. Schulz MH, Zerbino DR, Vingron M, Birney EJB. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. 2012;28(8):1086-92.
66. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. 2014;30(12):1660-6.
67. Putri GH, Anders S, Pyl PT. Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics. 2022;38:2943–5.
68. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166-9.
69. Ghosh S, Chan C-KK. Analysis of RNA-Seq data using TopHat and Cufflinks. In: Plant bioinformatics. Springer. 2016;339-61.
70. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:1–16.
71. Roberts A, Feng H, Pachter L. Fragment assignment in the cloud with eXpress-D. BMC Bioinformatics. 2013;14:1–9.
72. ALi A, Khan UK, Ullah W, Aslam M. Advancement in Medicinal plants transcriptomic approaches for deciphering bioactive secondary metabolites. 2022.
73. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PMJNar. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. 2010;38(6):1767-71.
74. Chen S, Zhou Y, Chen Y, Gu JJB. fastp: an ultra-fast all-in-one FASTQ preprocessor. 2018;34(17):i884-i90.
75. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SLJNp. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. 2016;11(9):1650-67.
76. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.
77. Arafah A, Rehman MU, Mir TM, Wali AF, Ali R, Qamar W, et al. Multi-therapeutic potential of naringenin (4′, 5, 7-trihydroxyflavonone): experimental evidence and mechanisms. 2020;9(12):1784.
78. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics. 2008;9:1–8.
79. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18:1851–8.
80. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics. 2009;25:1363–9.
81. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008;24:2395-6.
82. Diwan A, Harke S, Panche A. Application of proteomics in shrimp and shrimp aquaculture. Comp Biochem Physiol Part D Genomics Proteomics. 2022;101015.
83. Lin H, Zhang Z, Zhang MQ, Ma B, Li MJB. ZOOM! Zillions of oligos mapped. 2008;24(21):2431-7.
84. Weyl EG, Fabinger M. Pass-through as an economic tool: Principles of incidence under imperfect competition. J Polit Econ. 2013;121:528–83.
85. Eaves HL, Gao Y. MOM: maximum oligonucleotide mapping. Bioinformatics. 2009;25:969–70.
86. Wenya W, Erli PJJoBNU. Comparison of BWA-MEM and NovoAlign using human whole-genome next-generation sequencing reads. 2021;57(3):337-44.
87. Kim YJ, Teletia N, Ruotti V, Maher CA, Chinnaiyan AM, Stewart R, et al. ProbeMatch: rapid alignment of oligonucleotides to genome allowing both gaps and mismatches. 2009;25(11):1424.
88. Almeida AE, Menini N, Verbesselt J, Torres RdS, editors. BFAST explorer: An effective tool for time series analysis. IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium; 2018: IEEE.
89. Modi A, Vai S, Caramelli D, Lari M. The Illumina sequencing protocol and the NovaSeq 6000 system. Bacterial Pangenomics: methods and protocols: Springer; 2021. p. 15-42.
90. Lopez de Heredia U, Vázquez-Poletti JL. RNA-seq analysis in forest tree species: bioinformatic problems and solutions. Tree Genet Genomes. 2016;12:1–17.
91. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. In: Gene prediction. Springer. 2019;227–45.
92. Leenheer P De, Debruyne C. DOGMA-MESS: A tool for fact-oriented collaborative ontology evolution. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer. 2008;797–806.
93. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly SJGr. TransRate: reference-free quality assessment of de novo transcriptome assemblies. 2016;26(8):1134-44.
94. Fu L, Niu B, Zhu Z, Wu S, Li WJB. CD-HIT: accelerated for clustering the next-generation sequencing data. 2012;28(23):3150-2.
95. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
96. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:1–9.
97. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. 2012;7(3):562-78.
98. Pimentel H, Bray NL, Puente S, Melsted P, Pachter LJNm. Differential analysis of RNA-seq incorporating quantification uncertainty. 2017;14(7):687-90.
99. Salmen F, De Jonghe J, Kaminski TS, Alemany A, Parada GE, Verity-Legg J, et al. High-throughput total RNA sequencing in single cells using VASA-seq. 2022;40(12):1780-93.
100. Emde A-K, Grunert M, Weese D. MicroRazerS: rapid

alignment of small RNA reads. Bioinformatics. 2010;26:123-4.
101. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13:36-46.
102. Au KF, Jiang H, Lin L. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010;38:4570-8.
103. Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DT, Kolle G, et al. RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. 2009;25(19):2615-6.
104. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, et al. mrsFAST: a cache-oblivious algorithm for short-read mapping. 2010;7(8):576-7.
105. Arnold F, Podehl G. Best of both worlds–a mapping from EXPRESS-G to UML. In: The Unified Modeling Language.«UML»'98: Beyond the Notation: First International Workshop, Mulhouse, France, June 3-4, 1998. Selected Papers 1. Springer. 1999;49-63.
106. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SLJNb. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. 2015;33(3):290-5.
107. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32:462-4.
108. He D, Zakeri M, Sarkar H, Soneson C, Srivastava A, Patro RJNM. Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. 2022;19(3):316-22.
109. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
110. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. 2005;102(43):15545-50.
111. Korir PK, Roberts L, Ramesar R, Seoighe CJBrn. A mutation in a splicing factor that causes retinitis pigmentosa has a transcriptome-wide effect on mRNA splicing. 2014;7:1-11.
112. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. 2000;406(6797):747-52.
113. Dolnicar S, Kaiser S, Lazarevski K, Leisch FJJoTR. Biclustering: Overcoming data dimensionality problems in market segmentation. 2012;51(1):41-9.
114. Chia BKH, Karuturi RKM. Differential coexpression framework to quantify goodness of biclusters and compare biclustering algorithms. Algorithms Mol Biol. 2010;5:1–12.
115. Gillis J, Pavlidis P. A methodology for the analysis of differential coexpression across the human lifespan. BMC Bioinformatics. 2009;10:1–15.
116. Li H, Karuturi RKMJIjodm, bioinformatics. Significance analysis and improved discovery of disease-specific Differentially Co-expressed Gene Sets in microarray data. 2010;4(6):617-38.
117. Karp PD, Riley M, Paley SM, Pellegrini-Toole AJNar. The metacyc database. 2002;30(1):59-61.
118. Pang Z, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. 2021;49(W1):W388-W96.
119. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. 2019;10(1):1523.
120. Naithani S, Gupta P, Preece J, D'Eustachio P, Elser JL, Garg P, et al. Plant Reactome: a knowledgebase and resource for comparative pathway analysis. 2020;48(D1):D1093-D103.
121. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. 2021;49(D1):D613-D21.
123. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker TJB. Cytoscape 2.8: new features for data integration and network visualization. 2011;27(3):431-2.
124. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc collection of microbial genomes and metabolic pathways. 2019;20(4):1085-93.
125. Shoukat W, Hussain M, Ali A. Design, Synthesis, characterization and biological screening of novel thiosemicarbazones and their derivatives with Potent Antibacterial and Antidiabetic Activities. J Mol Struct. 2024;139614.
126. Riaz R, Parveen S, Shafiq N. Virtual screening, ADME prediction, drug-likeness, and molecular docking analysis of Fagonia indica chemical constituents against antidiabetic targets. Mol Divers. 2024;1–22.
127. Goel R, Harsha H, Pandey A, Prasad TKJMB. Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. 2012;8(2):453-63.
128. Romero P. The humancyc pathway-genome database and pathway tools software as tools for imaging and analyzing metabolomics data. Handb Metabolomics. 2012;419–38.
129. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, et al. neXtProt: a knowledge platform for human proteins. 2012;40(D1):D76-D83.
130. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. 2004;20(18):3710-5.
131. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:1–14.
132. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. In: Protein networks and pathway analysis. Springer. 2009;123–40.