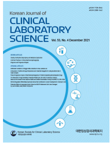
Korean J Clin Lab Sci. 2023;55(4):235-243

https://doi.org/10.15324/kjcls.2023.55.4.235

Korean Society for
Clinical Laboratory Science

Korean Journal of
CLINICAL LABORATORY SCIENCE



REVIEW ARTICLE

Big Data Analytics in RNA-sequencing

Sung-Hun WOO[1], Byung Chul JUNG[2]

[1]Department of Biomedical Laboratory Science, College of Software and Digital Healthcare Convergence, Yonsei University, Wonju, Korea
[2]Department of Nutritional Sciences and Toxicology, University of California, Berkeley, California, USA

RNA 시퀀싱 기법으로 생성된 빅데이터 분석

우성훈[1], 정병출[2]

[1]연세대학교 소프트웨어디지털헬스케어융합대학 임상병리학과, [2]캘리포니아대학교 버클리캠퍼스 영양과학 및 독성학과

ARTICLE INFO

Received October 31, 2023
Revised 1st November 10, 2023
Revised 2nd November 13, 2023
Accepted November 13, 2023

Key words
Big data
Computational biology
Sequence analysis, RNA

ABSTRACT

As next-generation sequencing has been developed and used widely, RNA-sequencing (RNA-seq) has rapidly emerged as the first choice of tools to validate global transcriptome profiling. With the significant advances in RNA-seq, various types of RNA-seq have evolved in conjunction with the progress in bioinformatic tools. On the other hand, it is difficult to interpret the complex data underlying the biological meaning without a general understanding of the types of RNA-seq and bioinformatic approaches. In this regard, this paper discusses the two main sections of RNA-seq. First, two major variants of RNA-seq are described and compared with the standard RNA-seq. This provides insights into which RNA-seq method is most appropriate for their research. Second, the most widely used RNA-seq data analyses are discussed: (1) exploratory data analysis and (2) pathway enrichment analysis. This paper introduces the most widely used exploratory data analysis for RNA-seq, such as principal component analysis, heatmap, and volcano plot, which can provide the overall trends in the dataset. The pathway enrichment analysis section introduces three generations of pathway enrichment analysis and how they generate enriched pathways with the RNA-seq dataset.

INTRODUCTION

Investigating the transcriptomic differences between physiological and pathological conditions helps us gain insights into the mechanisms underlying diseases and the development of therapeutic strategies. The traditional approach relied on low-throughput techniques such as reverse transcription polymerase chain reaction and quantitative polymerase chain reaction, which are limited to analyzing single or a few transcripts of interest [1]. However, alteration of particular gene expression may not always directly lead to the phenotype of interest, but expression change of multiple gene sets can be involved in the consequential biological phenotypes [2]. With rapid technological advancements, researchers are able to analyze global transcriptome profiling. The first transcriptome study was conducted using complementary DNA microarray to monitor the expression of 45 Arabidopsis genes with a single reaction [3]. This study has opened new avenues for investigating transcriptomes on a genome-wide scale, going beyond single-gene analysis. Although current DNA microarrays can offer comprehensive coverage of the genome, this technique is limited to pre-defined

Corresponding author: Byung Chul JUNG
Department of Nutritional Sciences and Toxicology, University of California, Berkeley, California 94720, USA
E-mail: sandbag9@berkeley.edu
ORCID: https://orcid.org/0000-0003-0732-0122

pISSN 1738-3544 eISSN 2288-1662

transcripts due to the requirement of hybridization with pre-fixed probes on the DNA chip. In addition, the other disadvantages of microarrays are relatively (1) high cost, (2) low specificity, and (3) low reproducibility. After next-generation sequencing (NGS) became available, RNA-sequencing (RNA-seq) has been gradually overtaking DNA microarray as the tool of choice for studying global transcriptome to overcome the disadvantages of microarray [4]. In contrast to hybridization-based microarrays, RNA-seq does not require predesigned probes and provides more sensitive and accurate data at a lower cost [5]. However, RNA-seq presents numerous challenges that need to be overcome for accurate data interpretation. Our goals in this review are to describe distinct RNA-Seq to select the most appropriate assay and how to interpret the RNA-Seq data to gain insights into relevant biological meaning.

MAIN ISSUE

1. Types of RNA-sequencing

RNA-seq is one of the most popular high-throughput technologies that uses NGS to reveal patterns and quantify cellular transcriptomes. Since the development of RNA-seq, nearly 100 distinct methods have evolved from the standard RNA-seq protocol [6]. Nevertheless, the majority of RNA-seq data in public repositories have been generated using Illumina sequencing technology, and most of the steps have not changed substantially [6]. Therefore, this review focuses on the pros and cons of each RNA-seq methods between standard RNA-seq and other two popular variants of RNA-seq, rather than describing the detailed workflows of each RNA-seq.

1) Standard RNA-sequencing

The standard RNA-seq procedure consists of several steps including RNA isolation, converting to complementary DNA (cDNA), adaptor ligation, constructing a sequencing library, sequencing by synthesis, and analysis. Standard RNA-seq provides transcriptome information on gene expression profiling, splice variant analysis and single nucleotide polymorphism (SNP) discovery (Table 1) [7]. However, standard RNA-seq requires a relatively higher amount of RNA and a higher RNA integrity number (RIN), which can be a significant obstacle for analyzing less abundant cell populations in tissue and low-quality RNA from forensic and certain clinical samples [8, 9].

2) 3' Tag RNA-sequencing

One popular variation of RNA-seq is 3' Tag RNA-seq, which uses oligo-dT priming for cDNA conversion, resulting in constructed libraries that are enriched near the 3' end of polyadenylated messenger RNAs (mRNAs) [10]. Because the 3' end of mRNA in mammals is more stable than other mRNA regions, 3' Tag RNA-seq is less sensitive to RNA degradation [11]. In addition, 3' Tag RNA-seq requires much fewer sequencing reads to identify differentially expressed genes (DEGs) between the samples, leading to substantial cost savings as well

Table 1. Comparison of distinct RNA-seq described in this review

	Standard RNA-seq	3' Tag RNA-seq	De novo transcript assembly
Purpose	Discovery of DEGs Splice variant analysis SNP discovery	Discovery of DEGs	Working with unavailable model organism and/or high amount of genomic alteration sample
Applicable species	Prokaryote, eukaryote	Eukaryote	Prokaryote, eukaryote
Recommended read depth	10~30 million reads	<5 million reads	100~200 million reads
Cost	Average	Low	High
Recommended input RNA (ng)	>500	>25	>500
Recommended RIN	>8.0	>5.0	>5.3 (10.1111/1755-0998.12485)
Reference genome	Required	Required	Not required

Abbreviations: RNA-seq, RNA-sequencing; DEGs, differentially expressed genes; SNP, single nucleotide polymorphism; RIN, RNA integrity number.

as providing low-noise gene expression profiles. However, 3' Tag RNA-seq does have certain limitations. For instance, it is only suitable for eukaryotic total RNA samples due to the requirement of a poly-A tail. In addition, it is not suitable for identifying splice variant analysis and SNP discovery. Therefore, 3' Tag RNA-seq can be the best sequencing option in case the primary goal of the analysis is to identify DEGs between eukaryotic samples that have a lower quantity and lower RIN.

3) De Novo Transcript Assembly

The resulting sequenced reads generated from both standard RNA-seq and 3' Tag RNA-seq are required reference genomes for mapping and assembling them to reveal the transcript (Table 2). In contrast, de novo transcript assembly involves the process of directly joining overlapping reads into longer contiguous sequences and don't need reference genome for analysis. Therefore, de novo transcript assembly becomes a useful approach for analyzing cellular transcriptomes with an unavailable reference genome sequence. Moreover, de novo transcript assembly successfully generates transcripts even in cases where a reference-guided assembly may fail to reconstruct them correctly due to gaps, high fragmentation, or significant alterations in the genomic sequence, as is often the case in cancer cells [12]. Nonetheless, it is essential to acknowledge certain limitations associated with de novo transcript assembly. de novo transcript assembly requires a high amount of sequencing read counts, which causes a much higher cost. Moreover, most genomes contain lots of repetitive regions, which make it difficult to achieve high-quality transcript assembly and often cause errors leading to misarrangements in the assembly results [13].

In summary, in order to select a suitable RNA-seq method, several factors should be considered such as experimental objectives, RNA sample quality and availability of a reference transcriptome.

2. RNA-sequencing Data Analysis

1) Exploratory Data Analysis

Exploratory data analysis refers to the approach of investigating and summarizing the main characteristics of the data sets to facilitate a better understanding of the data [14]. Several statistical data visualization methods, such as principal component analysis (PCA), heatmap, and volcano plot, are categorized under exploratory data analysis. With these analyses, we can quickly reveal the overall trends in the dataset, which can guide us in determining the most suitable analytical approach.

(1) Principal Component Analysis (PCA)

Since RNA-seq is involved in complex multiple steps, extreme deviation of intrasample called an outlier is often generated [15, 16]. Finding and removing outliers in RNA-seq datasets is a prerequisite for improving quality and preventing misinterpretation of the biological meaning derived from RNA-seq data. Given that data generated from RNA-seq is high-dimensional, considering a global overview of the data is necessary to determine intrasample outliers. In this regard, PCA is

Table 2. Most widely studied species of latest reference genome assembly available in the UCSC Browser (http://genome.ucsc.edu.)

Taxonomic name	UCSC version	Genome assembly name	Release date
Homo sapiens	hs1	T2T CHM13v2.0	24 Jan. 2022
Mus musculus	mm39	GRCm39	Jun. 2020
Rattus norvegicus	rn7	mRatBN7.2	Nov. 2020
Danio rerio	danRer11	GRCz11	May 2017
Drosophila melanogaster	dm6	BDGP Release 6+ISO1 MT	Aug. 2014
Caenorhabditis elegans	ce11	WBcel235	Feb. 2013
Saccharomyces cerevisiae S288C	sacCer3	R64	Apr. 2011

Abbreviation: UCSC, University of California Santa Cruz.

the most widely used way to distinguish outliers. PCA is a widely used exploratory data analysis to reduce the dimensionality of the dataset, which can allow for visualization of the dominant patterns of the data [17]. Principal components are linear combinations of the genes that collectively explain the variation across the samples [18]. As depicted in Figure 1, the PCA plot provides the overall similarity of the dataset. Through the PCA plot in Figure 1 (right), we can successfully identify an outlier that deviates significantly from the overall distribution pattern of the control group.

(2) Heatmap

Heatmap is a graphical representation of the data using a color gradient for easier interpretation and visualization of RNA-seq data [19]. Genes with higher expression levels are typically colored red, while those with lower expression levels are colored blue, thus providing a simultaneous illustration of gene expression patterns within large datasets across all the samples. In most cases, rows representing each gene and columns representing each sample are reordered by certain clustering algorithms. This reordering ensures the data matrices with similar patterns are placed closely on the heatmap. In biology, a hierarchical clustering algorithm is the most widely applied algorithm for generating a heatmap [20, 21]. The hierarchical clustering algorithm is a type of unsupervised machine learning algorithm, which pairs objects based on the degree of similarity [22]. Therefore, a heatmap combined with a hierarchical clustering algorithm provides better visualization of patterns, relationships, and similarities across all the samples (Figure 2) [21, 23]. In ideal RNA-seq data, hierarchical clustering algorithms would pair the same groups

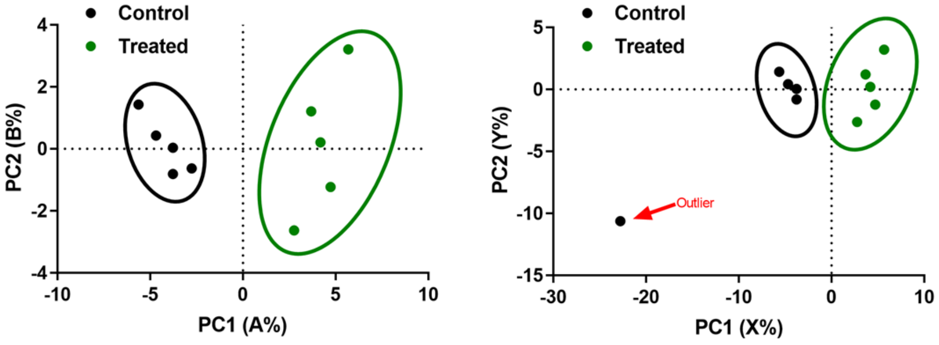


Figure 1. Comparing PCA plot of simulated dataset with or without outlier. PC1 (A or X%) and PC2 (B or Y%) describes the most and second most variation within the data, which accounts for (A or X%) or (B or Y%) of the variance respectively (A>B; X>Y). Abbreviations: PCA, principal component analysis; PC1, principal component1; PC2, principal component2.

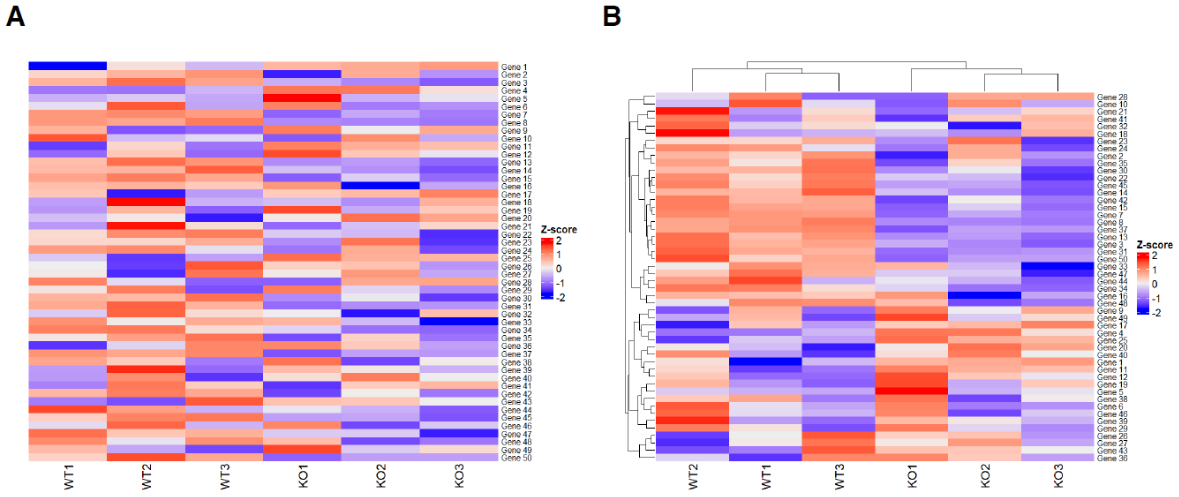


Figure 2. Comparison of heatmap generated from the same simulated dataset without hierarchical clustering algorithm (A) and with hierarchical clustering algorithm (B). Heatmaps are generated by ComplexHeatmap R package [21, 23].

together, as the similarity within the same groups is greater than the similarity between different groups.

(3) Volcano Plot

A volcano plot is a type of scatter plot to explore the most interesting genes within large datasets. Typically, the x-axis of a volcano plot represents $\log_2$ of the fold change (FC), and the y-axis represents the $-\log_{10}$ of the adjusted *P*-value, which is called as "double filtering" criterion [24, 25]. As depicted in Figure 3 [26], interesting genes (DEGs) meet the two criteria: (1) $|\log_2 FC|>1$ and (2) adjusted *P*-value<0.05. Double filtering criteria can be beneficial to exclude (1) genes with large expression differences that are caused by large variations in the dataset (outlier) and (2) genes that show statistical significance but have low expression differences, which could be false positives caused by batch effect or low expression level [24, 27]. In a volcano plot, the points representing each gene that is located in the upper-right and upper-left corners are considered the most promising findings, which can be candidate biomarkers and therapeutic targets. However, double filtering criteria itself relies on arbitrary cutoffs and there is no standard rule for setting up a cutoff threshold, which results in filtering out potentially valuable genes that are close to the cutoff threshold.

Therefore, researchers should be aware that while exploratory data analysis is useful for an initial overview, it does not provide sufficient information to derive meaningful conclusions.

2) Pathway Enrichment Analysis

With the advent of high-throughput technologies such as RNA-seq, the main hurdle has shifted from acquiring gene expression profiles to the correct interpretation of the transcriptomic data to gain insights into relevant biological meaning. Typically, RNA-seq data generates the gene list of hundreds to thousands of DEGs, making it nearly impossible to manually search the literature and interpret the biological nuances. A standard strategy to overcome this issue is a pathway enrichment analysis which identifies a smaller list of interpretable biological pathways from overwhelmingly large gene lists [28]. Biological pathways are achieved with statistical testing whether provided gene lists are enriched in particular pathways from a variety of databases (Table 3). Pathway enrichment analysis can be categorized into over representation analysis (ORA), functional class scoring (FCS), and topology-based pathway analysis (TPA) (Figure 4) [29, 30].

(1) Over Representation Analysis

ORA is the first generation of pathway enrichment analysis that explores the list of DEGs that are enriched in certain biological pathways from public repositories

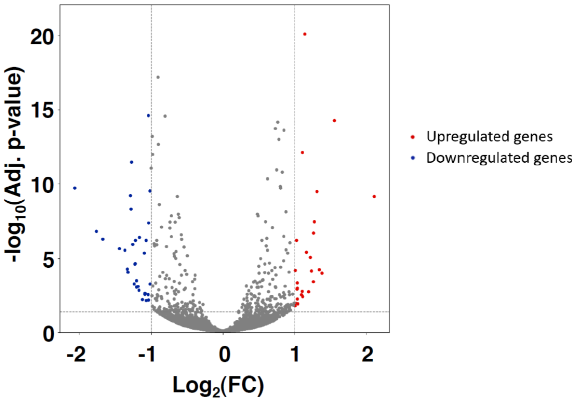


Figure 3. Typical volcano plot generated from the simulated dataset. Bioinfokit python package was used to illustrate the volcano plot [26]. Abbreviation: FC, fold change.

Table 3. Popular public repositories commonly used in pathway enrichment analysis

Database	Website	TPA availability	Reference (DOI)
KEGG	https://www.kegg.jp/	Yes	10.1093/nar/28.1.27
GO	https://geneontology.org/	No	10.1002/pro.4218
REACTOME	https://reactome.org/	Yes	10.1093/nar/gkab1028
WikiPathways	https://www.wikipathways.org/	Yes	10.1093/nar/gkaa1024
MsigDB	https://www.gsea-msigdb.org/gsea/msigdb	No	10.1016/j.cels.2015.12.004

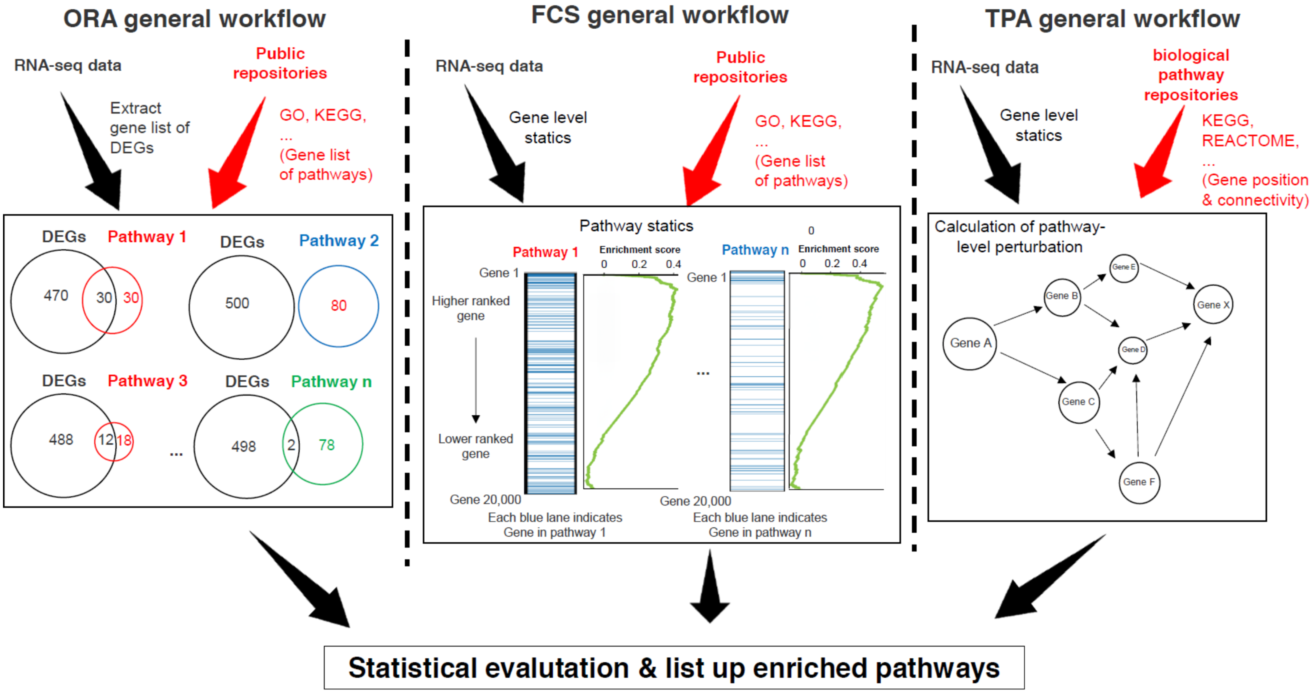


Figure 4. Schematic image of pathway enrichment analysis.
Abbreviations: ORA, over representation analysis; FCS, functional class scoring; TPA, topology-based pathway analysis; RNA-seq, RNA-sequencing; DEGs, differentially expressed genes.

(Table 3). The first step of ORA is identifying DEGs from RNA-seq data using certain criteria such as false discovery rate (FDR) and/or FC of gene expression. The next step involves counting the number of selected DEGs within each pathway. This counting process is performed for each pathway individually. Subsequently, the statistical evaluation of each pathway is carried out using a Fisher's exact test based on the hypergeometric distribution [28, 31]. Since hundreds of pathways (hypothesis) are statistically evaluated simultaneously, each statistical evaluation has a false positive error probability (Type I error) [32, 33]. Therefore, multiple-testing correction is required to correct this error. Benjamini-Hochberg FDR procedure is the most commonly applied method to correct *P*-value as the adjusted *P*-value (Q-value) [34]. However, there are still limitations to ORA. Since the selection of DEGs relies on arbitrary cutoffs, such as FDR and/or FC of gene expression, standardization can be challenging. Additionally, ORA is that it tends to identify DEGs associated with substantial expression changes by arbitrary cutoffs. This tendency can exclude sets of functionally related genes with milder expression changes, which could coordinately exert as much influence as a single gene with a large expression change. Moreover, once DEGs are chosen, ORA considers the entire list of genes for analysis. This approach results in each gene within the DEGs list having an equal impact on pathway enrichment, regardless of differences in their FDR and gene expression levels.

(2) Functional Class Scoring

FCS is the second generation of pathway enrichment analysis to overcome the limitations of ORA. In contrast to ORA, FCS does not filter out with a particular threshold to isolate the list of DEGs. Instead, FCS calculates and assigns the gene score to each gene and analyzes pathway enrichment based on these assigned gene scores, ensuring that all genes are considered in the analysis. The most widely used FCS tool is gene set enrichment analysis (GSEA) [2, 35]. GSEA computes gene scores with several methods such as signal-tonoise ratio, t-test and gene expression between two phenotypes [36]. Then, it evaluates the distribution of a set of genes from each pathway of the database repository to assign an enrichment score through a weighted Kolmogorov-Smirnov-like statistic [2]. The statistical significance of the enrichment score is evaluated by an empirical phenotype-based permutation test for larger replicates or a gene set for smaller replicates (below 7 replicates) [36]. Lastly, the enrichment score is adjusted by multiple hypotheses testing to reduce a false positive finding. Since the FCS method uses all available information from RNA-seq data, it has a better resolution to detect the pathways associated with weak but coordinated gene sets. However, FCS also comes with certain limitations. FCS does not account for the relationships within the gene sets of the pathway, often referred to as the 'gene independence assumption' [37], which is far removed from the interconnected nature of biology. Similarly, another limitation of FCS is that it does not consider the interplay between pathways. Given that biological pathways are not isolated entities, and one pathway can affect the activity of others, the approach of FCS neglects actual biological processes.

(3) Topology-based Pathway Analysis

To overcome the limitations of ORA and FCS, TPA was developed. Unlike the first two-generation analysis, TPA takes into consideration not only the lists of genes and gene ranks but also the integration of topological information from particular biological pathway repositories such as KEGG, REACTOME, or WikiPathway [38]. There are several publicly available TPA based packages such as SPIA and SEMgsa [39, 40]. The algorithm of TPA is fundamentally similar to that of FCS. However, the main difference lies in the fact that TPA considers topological features of the genes such as the position of genes within a pathway and its interaction with other genes [41]. Essentially, TPA computes a pathway-level perturbation using both expression and the topology of the pathway, which enables a better assessment of relevant pathway derived from the RNA-seq data [42-44]. Although TPA is the latest generation, this method also has its own limitations, which may be addressed in future methods. First, it is not feasible to consider activation and inactivation time and spatial distribution for the pathway components in the model. Second, the genuine pathway underlying the RNA-seq data can be different from the pathways of the database. Lastly, the limited database is available due to the cell and tissue specificity of the pathway.

CONCLUSION

As technology advances, RNA-seq has become the first choice for interpreting cellular transcriptomes both in research and clinical applications. To provide general considerations for choosing appropriate types of RNA-seq, we introduce 3' Tag RNA-seq and de novo transcript assembly as well as standard RNA-seq in this review (Table 1). For example, if the primary objectives of the experiment are to isolate DEGs from a small amount or low-quality eukaryotic RNA samples, then 3' Tag RNA-seq would be a better option compared to standard RNA-seq. Moreover, we also summarize three generations of pathway enrichment analysis, which has become one of the foremost tools of RNA-seq data. Basically, all the pathway enrichment analyses aim to simplify the complex transcriptomic profiling, identifying a smaller number of significantly enriched pathways. As described above, all three have their own advantages and limitations. It is important for researchers to understand that none of the existing approaches is flawless. Therefore, these tools should be used to generate more refined hypotheses for uncovering biological meanings, rather than making definitive conclusions.

요 약

차세대 염기서열 분석이 개발되고 널리 사용됨에 따라 RNA-시퀀싱(RNA-sequencing, RNA-seq)이 글로벌 전사체 프로파일링을 검증하기 위한 도구의 첫번째 선택으로 급부상하게 되

었다. RNA-seq의 상당한 발전으로 다양한 유형의 RNA-seq가 생물정보학(bioinformatics) 발전과 함께 진화했으나, 다양한 RNA-seq 기법 및 생물정보학에 대한 전반적인 이해 없이는 RNA-seq의 복잡한 데이터를 해석하여 생물학적 의미를 도출하기는 어렵다. 이와 관련하여 본 리뷰에서는 RNA-seq의 두 가지 주요 섹션을 논의하고 있다. 첫째, Standard RNA-seq과 주요하게 자주 사용되는 두 가지 RNA-seq variant method를 비교하였다. 이 비교는 어떤 RNA-seq 방법이 연구 목적에 가장 적절한지에 대한 시사점을 제공한다. 둘째, 가장 널리 사용되는 RNA-seq에서 생성된 데이터 분석; (1) 탐색적 자료 분석 및 (2) enriched pathway 분석에 대해 논의하였다. 데이터 세트의 전반적인 추세를 제공할 수 있는 주 성분 분석, Heatmap 및 Volcano plot과 같이 RNA-seq에 대해 가장 널리 사용되는 탐색적 자료 분석을 소개하였다. Enriched pathway 분석 섹션에서는 3가지 세대의 enriched pathway 분석에 대해 소개하고 각 세대가 어떤 식으로 RNA-seq 데이터 세트로부터 enriched pathway를 도출하는지를 소개하였다.

Funding: None

Acknowledgements: None

Conflict of interest: None

Author's information (Position): Woo SH[1], Graduate student; Jung BC[2], Researcher.

Author Contributions

- Writing - original draft: Woo SH, Jung BC.
- Writing - review & editing: Woo SH, Jung BC.

Ethics approval

This article does not require IRB/IACUC approval because there are no human and animal participants.

ORCID

Sung-Hun WOO	https://orcid.org/0000-0002-1642-9341
Byung Chul JUNG	https://orcid.org/0000-0003-0732-0122

REFERENCES

1. Jung BC, Kang S. Epigenetic regulation of inflammatory factors in adipose tissue. Biochim Biophys Acta Mol Cell Biol Lipids. 2021;1866:159019. https://doi.org/10.1016/j.bbalip.2021.159019
2. 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. Proc Natl Acad Sci U S A. 2005;102:15545-15550. https://doi.org/10.1073/pnas.0506580102
3. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270:467-470. https://doi.org/10.1126/science.270.5235.467
4. Mutz KO, Heilkenbrinker A, Lönne M, Walter JG, Stahl F. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol. 2013;24:22-30. https://doi.org/10.1016/j.copbio.2012.09.004
5. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9:e78644. https://doi.org/10.1371/journal.pone.0078644
6. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631-656. https://doi.org/10.1038/s41576-019-0150-2
7. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57-63. https://doi.org/10.1038/nrg2484
8. Sigurgeirsson B, Emanuelsson O, Lundeberg J. Sequencing degraded RNA addressed by 3' tag counting. PLoS One. 2014;9:e91851. https://doi.org/10.1371/journal.pone.0091851
9. Jung BC, You D, Lee I, Li D, Schill RL, Ma K, et al. TET3 plays a critical role in white adipose development and diet-induced remodeling. Cell Rep. 2023;42:113196. https://doi.org/10.1016/j.celrep.2023.113196
10. Weng X, Juenger TE. A high-throughput 3'-Tag RNA sequencing for large-scale time-series transcriptome studies. Methods Mol Biol. 2022;2398:151-172. https://doi.org/10.1007/978-1-0716-1912-4_13
11. Wu X, Bartel DP. Widespread influence of 3'-end structures on mammalian mRNA processing and stability. Cell. 2017;169:905-917.e11. https://doi.org/10.1016/j.cell.2017.04.036
12. Raghavan V, Kraft L, Mesny F, Rigerte L. A simple guide to de novo transcriptome assembly and annotation. Brief Bioinform. 2022;23:bbab563. https://doi.org/10.1093/bib/bbab563
13. Liao X, Li M, Zou Y, Wu FX, Yi P, Wang J. Current challenges and solutions of de novo assembly. Quant Biol. 2019;7:90-109. https://doi.org/10.1007/s40484-019-0166-9
14. Teo YY. Exploratory data analysis in large-scale genetic studies. Biostatistics. 2010;11:70-81. https://doi.org/10.1093/biostatistics/kxp038
15. Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, et al. A beginner's guide to analysis of RNA sequencing data. Am J Respir Cell Mol Biol. 2018;59:145-157. https://doi.org/10.1165/rcmb.2017-0430tr
16. Chen X, Zhang B, Wang T, Bonni A, Zhao G. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinformatics. 2020;21:269. https://doi.org/10.1186/s12859-020-03608-0
17. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26:303-304. https://doi.org/10.1038/nbt0308-303
18. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374:20150202. https://doi.org/10.1098/rsta.2015.0202
19. Khomtchouk BB, Van Booven DJ, Wahlestedt C. HeatmapGenerator: high performance RNAseq and microarray visualization software suite to examine differential gene expression levels using an R

and C++ hybrid computational pipeline. Source Code Biol Med. 2014;9:30. https://doi.org/10.1186/s13029-014-0030-2

20. Engle S, Whalen S, Joshi A, Pollard KS. Unboxing cluster heatmaps. BMC Bioinformatics. 2017;18(Suppl 2):63. https://doi.org/10.1186/s12859-016-1442-6
21. Gu Z. Complex heatmap visualization. iMeta. 2022;1:e43. https://doi.org/10.1002/imt2.43
22. El Bouchefry K, de Souza RS. Learning in big data: introduction to machine learning. In: Škoda P, Adam F, editors. Knowledge discovery in big data from astronomy and earth observation: AstroGeoInformatics. Elsevier: 2020. p. 225-249.
23. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847-2849. https://doi.org/10.1093/bioinformatics/btw313
24. Li W. Volcano plots in analyzing differential expressions with mRNA microarrays. J Bioinform Comput Biol. 2012;10:1231003. https://doi.org/10.1142/s0219720012310038
25. Ebrahimpoor M, Goeman JJ. Inflated false discovery rate due to volcano plots: problem and solutions. Brief Bioinform. 2021;22:bbab053. https://doi.org/10.1093/bib/bbab053
26. Bedre R. reneshbedre/bioinfokit: bioinformatics data analysis and visualization toolkit [Internet]. Zenodo [cited 2022 Sep 4]. Available from: https://doi.org/10.5281/zenodo.3698145
27. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733-739. https://doi.org/10.1038/nrg2825
28. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14:482-517. https://doi.org/10.1038/s41596-018-0103-9
29. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. Front Physiol. 2015;6:383. https://doi.org/10.3389/fphys.2015.00383
30. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8:e1002375. https://doi.org/10.1371/journal.pcbi.1002375
31. Jung SH. Stratified Fisher's exact test and its sample size calculation. Biom J. 2014;56:129-140. https://doi.org/10.1002/bimj.201300048
32. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci. 2003;18:71-103. https://doi.org/10.1214/ss/1056397487
33. Camargo A, Azuaje F, Wang H, Zheng H. Permutation - based statistical tests for multiple hypotheses. Source Code Biol Med. 2008;3:15. https://doi.org/10.1186/1751-0473-3-15
34. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990;9:811-818. https://doi.org/10.1002/sim.4780090710
35. Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. BMC Bioinformatics. 2021;22:191. https://doi.org/10.1186/s12859-021-04124-5
36. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023;39:btac757. https://doi.org/10.1093/bioinformatics/btac757
37. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res. 2016;25:472-487. https://doi.org/10.1177/0962280212460441
38. Wang Y, Li J, Huang D, Hao Y, Li B, Wang K, et al. Comparing Bayesian-based reconstruction strategies in topology-based pathway enrichment analysis. Biomolecules. 2022;12:906. https://doi.org/10.3390/biom12070906
39. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009;25:75-82. https://doi.org/10.1093/bioinformatics/btn577
40. Grassi M, Tarantino B. SEMgsa: topology-based pathway enrichment analysis with structural equation models. BMC Bioinformatics. 2022;23:344. https://doi.org/10.1186/s12859-022-04884-8
41. Ma J, Shojaie A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. BMC Bioinformatics. 2019;20:546. https://doi.org/10.1186/s12859-019-3146-1
42. Ibrahim MA, Jassim S, Cawthorne MA, Langlands K. A topology-based score for pathway enrichment. J Comput Biol. 2012;19:563-573. https://doi.org/10.1089/cmb.2011.0182
43. Zhao K, Rhee SY. Interpreting omics data with pathway enrichment analysis. Trends Genet. 2023;39:308-319. https://doi.org/10.1016/j.tig.2023.01.003
44. Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019;20:203. https://doi.org/10.1186/s13059-019-1790-4