Computational and Structural Biotechnology Journal 19 (2021) 4426–4434

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

ARMT: An automatic RNA-seq data mining tool based on comprehensive and integrative analysis in cancer research

Guanda Huang, Haibo Zhang, Yimo Qu, Kaitang Huang, Xiaocheng Gong, Jinfen Wei *, Hongli Du *

School of Biology and Biological Engineering, South China University of Technology, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 29 April 2021
Received in revised form 19 July 2021
Accepted 6 August 2021
Available online 10 August 2021

Keywords:
RNA-seq
Downstream analysis
Integration R package
GSVA

ABSTRACT

The comprehensive and integrative analysis of RNA-seq data, in different molecular layers from diverse samples, holds promise to address the full-scale complexity of biological systems. Recent advances in gene set variant analysis (GSVA) are providing exciting opportunities for revealing the specific biological processes of cancer samples. However, it is still urgently needed to develop a tool, which combines GSVA and different molecular characteristic analysis, as well as prognostic characteristics of cancer patients to reveal the biological processes of disease comprehensively. Here, we develop ARMT, an automatic tool for RNA-Seq data analysis. ARMT is an efficient and integrative tool with user-friendly interface to analyze related molecular characters of single gene and gene set comprehensively based on transcriptome and genomic data, which builds the bridge for deeper information between genes and pathways, to further accelerate scientific findings. ARMT can be installed easily from https://github.com/Dulab2020/ARMT.

1. Introduction

RNA-seq related applications have been greatly developed and become one of the most important methods in the field of life science research, especially in cancer researches [1–3]. With the rapid development of the application of RNA-seq technology, a large number of gene expression profile data sets have been accumulated in GEO [4], ArrayExpress [5] and other databases. Based on these gene expression profile data sets, researchers [6,7] have used a large number of statistical studies including differential analysis, enrichment analysis, survival analysis, correlation analysis, to study gene function, analyze gene expression regulation, and analyze in-depth the transcriptome gene map of cancer occurrence and development, which greatly promoted the development of cancer research [8,9].

The upstream analysis processes of transcripts, such as quality control (QC), mapping, quantification, have been standardized [10]. However, downstream functional analysis like differential analysis, and enrichment analysis, varies greatly due to different experimental designs and research purposes [11–13].

Screening for differentially expressed genes is common in the statistical analysis of RNA-seq data [14–16]. Many tools and algorithms for differentially expressed genes of RNA-seq have been developed, such as R packages: 'edgeR' [17], 'DESeq2′ [18], 'limma' [19], 'SAMseq' [20], 'Cuffdiff/Cuffdiff2′ [21,22], 'baySeq' [23], 'sleuth' [24] and other new tools[25–27]. Based on different statistical principles, different tools may lead to different results [28,29]. Since 2017, the most frequently cited methods for differential analysis include 'edgeR', 'limma', 'Cuffdiff/Cuffdiff2', and 'DEGseq2' (Fig. 1 A) and the generally acknowledged packages in these R packages include 'edgeR', 'limma' and 'DESeq2'. Previous study showed that the results calculated by 'edgeR' and 'limma' were similar [30]. The calculation speed of 'DESeq2' is significantly slower than that of 'edgeR' and 'limma' (Fig. 1 B).

Survival analysis is an important way to study disease [31], among which the Kaplan-Meier method is most widely used in the case of single factor and two variables. COVID-19 is widely spread nowadays and causes many deaths, many researchers use the survival analysis model and Cox model to look for the factors related to the prognosis of disease caused by this virus [32].

A single gene is effective to reveal biological significance to a certain extent, but the biological reaction process, such as metabolism, transcriptional regulation, stress response, is co-regulated by multiple genes. Therefore, it is essential to analyze the gene sets or signal pathways, which makes it easier to achieve the interpretability of biological regulation [33–35]. Enrichment analysis can be used to search for different genes with certain commonalities and identify metabolic pathways or signal pathways associated with phenotypes or diseases. At present, there are mainly four

* Corresponding author.
E-mail addresses: weijinfen@scut.edu.cn (J. Wei), hldu@scut.edu.cn (H. Du).

https://doi.org/10.1016/j.csbj.2021.08.009

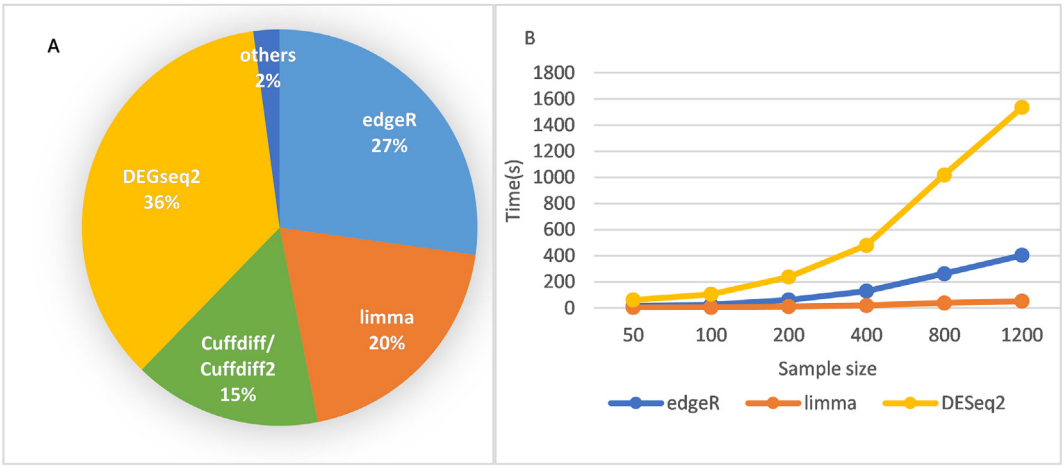

Fig. 1. Comparation of differential analysis tools A. The cite rate of tools for differential expression genes analysis **B.** The run time of differential analysis tools.

methods for gene enrichment analysis: Over - representation Analysis(ORA) [36], Functional Class Scoring (FCS) [37], Pathway Topology (PT) [38], Network Topology(NT) [39,40]. The most commonly used method is ORA like Gene Ontology (GO) [41] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [42] analysis based on hypergeometric distribution hypothesis test, which can only perform enrichment analysis based on the pathways defined in the databases with differentially expressed genes. While FCS can not only work with custom gene sets, but also consider the genes expression profile, which has higher calculation accuracy. The most famous of the FCS methods is gene set enrichment analysis (GSEA) [43], a supervised method based on population. While GSEA is thought to lack consideration of correlations between genes, resulting in an increase in the number of false positive gene sets [44]. Compared with GSEA, gene set variation analysis(GSVA) is an unsupervised and nonparameterized method for gene set analysis [45], which can estimate changes in pathway activity in the sample population. The GSVA also uses density estimates to assess sample enrichment, and allows for more extensive downstream analysis by ignoring phenotypic information, which is widely used in many studies recently [46–50].

The methods and tools, such as 'metascape' [51] for enrichment analysis and 'maftools' [52] for mutant mapping, are to achieve a certain type of function. There are also some tools for multiple functions, such as TCGAbiolinks [53], but most of them are inflexible in integrated analysis, too few of which can further analyze GSVA score. The downstream function analysis of transcriptome needs a more systematic and flexible analysis process to reveal functional mechanism more deeply and comprehensive with gene set analysis.

Base on some cancer studies by using TCGA multi-omics data and multiple analysis methods [12,13,45,54–58], we developed a tool--ARMT, for comprehensive and in-depth downstream statistical analysis of RNA-seq data. Given the widespread use of GSVA, integrating multiple functions of GSVA, ARMT can carry out further analysis on GSVA score. In addition, by incorporating some visualization functions of maftools [52], ARMT has advantages in integrated analysis and data mining between genome and transcriptome. Besides, a GUI interface made by 'shiny' package makes it easy for users to browse the visualized results. ARMT is more multifunctional in downstream analysis than other R packages [53,59–61] for RNA-seq data (Table 1). The workflow and function of ARMT would be elaborated in the following content, and we would take a case to illustrate its flexible and convenient use.

2. Method

2.1. Workflow overview

ARMT is an open-source R package on GitHub, which has comprehensive function with a GUI interface (Supplementary figure 1 and Supplementary figure 2) made by 'shiny' package, allowing users to operate efficiently and browse the visualization results conveniently. In ARMT, we integrate originally independent analysis methods, including GSVA, survival analysis, differential analysis, correlation analysis, enrichment analysis and mutant mapping, and provide enough adjustable parameters. The framework of ARMT is showed in Fig. 2 and its function can be described

Table 1
The comparation of R packages for RNA-seq analysis.

	KnowSeq	RNAseqR	RTCGA Toolbox	TCGAbiolinksGUI	ARMT
Create gene set file					√
Normalization for counts data	√	√			√
Survival Analysis			√	√	√
Cox proportional hazards model					√
Detect DEG	√	√	√	√	√
GO & KEGG	√	√		√	√
Differential Analysis for mutation and GSVA score					√
Correlation Analysis			√		√
Mutant mapping				√	√
GSVA					√
GUI				√	√

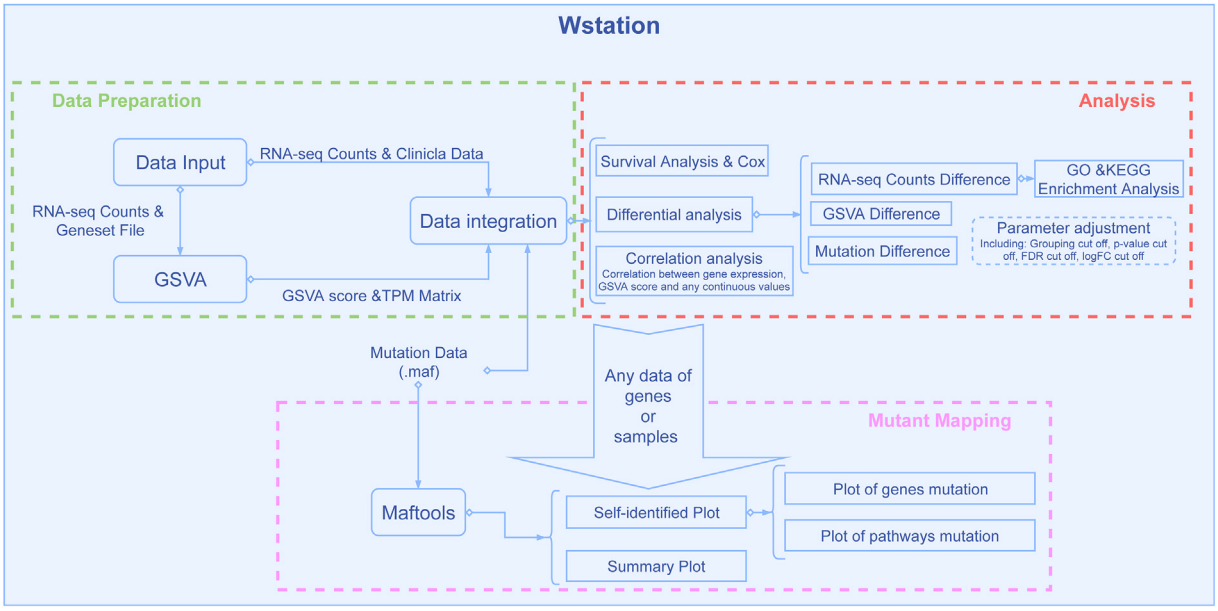

Fig. 2. The function & structure of ARMT. The data flow is illustrated and the framework can be divided to three parts (data preparation, analysis, mutant mapping).

as three parts: data preparation, analysis and mutant mapping. The user manual is included in the Supplementary data 1 and uploaded it to the GitHub repository (https://github.com/Dulab2020/ARMT).

2.2. Data preparation

We built TCGA clinical data into ARMT, which can be obtained directly by users in data preparation. The gene expression should be normalized with gene length when calculating GSVA score, and TPM is considered more suitable for with-in samples process [62] and more comparable between samples of different origins [10]. Therefore, before performing GSVA, the counts matrix should be normalized to TPM matrix and transformed to log2(TPM + 1) according to the official documentation, and the .gmt file of arbitrary gene sets can be built by ARMT. Then, clinical data, GSVA score, gene expression profile data (TPM) and mutation profile data can be merged through common samples, and this integrated data is used to next comprehensive analysis.

2.3. Analysis

After data preparation and integration, ARMT can carry out survival analysis, Cox proportional hazards regressive analysis, differential analysis, enrichment analysis and correlation analysis.

Survival analysis and Cox proportional-hazards regressive analysis can be carried out by various grouping of integrated data including mutation information, GSVA score, gene expression level and clinical information. Survival analysis is conducted for classification variables, ARMT divides samples into two groups for survival analysis through the selected factors in integration data. If the factor is a continuous variable, the samples will be grouped according to the level of values (top and bottom 50%, respectively). The result of survival analysis is demonstrated in Kaplan-Meier curve. The Cox proportional hazards model is also available in ARMT to analyze the combined effect of multiple factors, and the result is demonstrated with forest plot. The survival analysis and cox proportional-hazards model are carried out by R package 'survival'.

In differential analysis, the samples can be grouped by any factor in integration data. We provide enough adjustable parameters, including the threshold for the result (logFC, p-value and FDR) and the top and bottom threshold of the continuous factors (high and low). We employ 'edgeR' [17] packages to detect differentially expressed genes (DEGs) for high efficiency. To address the bias from normalization of counts matrix, four methods are provided, including 'TMM', 'TMMwsp', 'RLE', 'upperquartile'. As the GSVA score is decimal and 'edgeR' requires integer input, the differential analysis of GSVA score is carried out by 'limma' [19]. As mutant mapping by ARMT needs collaboration with 'maftools' [52], the differential analysis of mutation is also provided.

In enrichment analysis, the DEGs obtained by differential analysis or any gene list input by users can be enriched to GO or KEGG pathway by 'clusterprofiler' [63]. The result of enrichment can be screen by adjustable p-value and q-value.

The Spearman correlation coefficient and Pearson correlation coefficient could also be calculated by ARMT between any continuous variables, such as GSVA score and TPM value.

The above analysis method can be applied to multiple sets of data (such as pan-cancer) automatically in ARMT.

2.4. Mutant mapping

Here, the mutation information of genes can be plot out in summary, but only some specific genes or pathways are focused on by researchers. ARMT incorporates the 'maftools' [52] package's ability to read MAF mutation files and produce the mutant mapping (Supplementary figure 3), plotting out the distribution of specific genes mutations and the interaction heatmap between them (co-occurrence or mutually-exclusive), with an oncoplot of specific types of mutations. In the oncoplot, gene mutations can be displayed together with sample data and gene data from the transcriptome, facilitating the mining of deeper relationships between the genome and the transcriptome.

2.5. Visualization

ARMT's analysis results can be automatically visualized in the user interface with heat plots, volcano plots, forest plots and other methods; meanwhile, a searchable and sortable chart is also pro-

vided to display the calculation results (Supplementary figure 4, Supplementary figure 5 and Supplementary figure 6).

3. Result

3.1. Data

We downloaded the transcriptome data (counts matrix), mutation data, and clinical data of lung adenocarcinoma (LUAD) samples from TCGA. Then the transcriptome data has been standardized to obtain the TPM matrix of the cancer sample by ARMT. The data used in this paper is referenced to Wei JF [12] and has been agreed by the author.

According to Wei JF [12], the 14 gene signatures (*ALDOA*, *MIF*, *TUBB6*, *P4HA1*, *SLC2A1*, *PGAM1*, *ENO1*, *LDHA*, *CDKN3*, *TPI1*, *NDRG1*, *VEGFA*, *ACOT7* and *ADM*) and the 22 gene signatures (*SLC2A1*, *HK1*, *HK2*, *HK3*, *GPI*, *PFKL*, *PFKM*, *PFKP*, *ALDOA*, *ALDOB*, *ALDOC*, *TPI1*, *GAPDH*, *PGK1*, *PGAM1*, *PGAM4*, *ENO1*, *ENO2*, *ENO3*, *PKLR*, *PKM* and *LDHA*) had been verified to represent hypoxia and glycolysis gene set, respectively. These two gene sets were used to calculated GSVA score by ARMT in the following study.

3.2. Case

As previous studies have proposed hypoxia is associated with malignant progression, treatment resistance and poor clinical prognosis in various cancer types [64–67], thus, in this study, we focus on hypoxia for exploring hypoxia associated molecular characteristics to demonstrate the function and potential application of ARMT. In this paper, we will use an example to demonstrate the function of ARMT and to illustrate the use and highlight the analysis on GSVA score, we take a case of five steps for various analytical methods.

3.2.1. Step1: Differential analysis

In order to comprehensively reveal the molecular characteristics related to hypoxia, we first assessed the hypoxia level in each LUAD sample based on the GSVA score of hypoxia gene sets mentioned above, then grouped the samples into hypoxia-high and -low groups, respectively, according to hypoxia GSVA score (top and bottom 30%) to perform DEGs analysis. We obtained 6684 DEGs ($p < 0.05$, -log10(FDR) > 2, logFC > 1) (Fig. 3 A) and then filtered 1174 up-regulated genes with high criterion

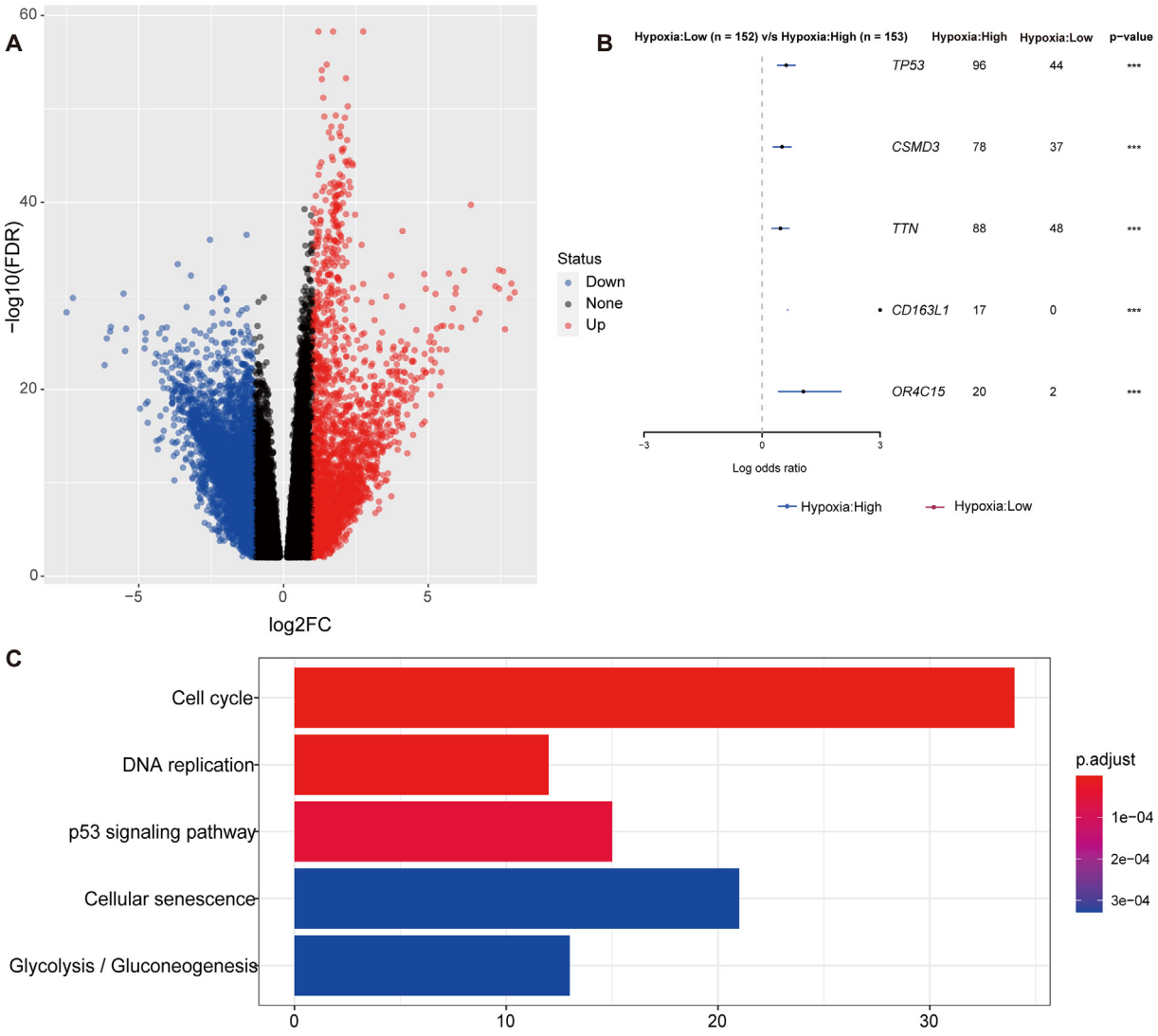

Fig. 3. Differential analysis for TCGA-LUAD samples according to hypoxia GSVA score. A. Volcano plot of differential expression genes **B.** Forest plot of differential mutation genes **C.** Bar plot of GO enrichment analysis for differential expression genes.

($p < 0.05$, -log10(FDR) > 20, logFC > 1) to do KEGG enrichment analysis, and found they were enriched in cell cycle and glycolysis-related pathways (Fig. 3 C). We then verified the difference of glycolysis by differential analysis on GSVA score (adj.p = $2.45*10^{-128}$) (Supplementary data 2). In order to identify genomic changes that associated with hypoxia, we conducted a differential analysis of the mutation data of LUAD sample between hypoxia-high and -low samples (Fig. 3 B). As Fig. 3 B showed five genes with SNV including *TP53*, *CSMD3*, *TTN*, *CD163L1*, *OR4C15* were significantly mutated in hypoxia-high groups ($p < 0.05$, FDR < 0.01), it is suggested that these genes were associated with hypoxia in tumor microenvironment.

3.2.2. Step2: Survival analysis & Cox Proportional-hazards model

In order to evaluate whether there was an effect of hypoxia and glycolysis on the survival of clinical patients, we grouped the clinical data of LUAD samples into GSVA-high and -low groups according to the level of glycolysis and hypoxia GSVA score for survival analysis. We found that both higher hypoxia and glycolysis score were significantly associated with poorer survival probability in LUAD patients (Fig. 4 A). The univariate and multivariate Cox proportional hazards regression analysis was also performed on glycolysis and hypoxia with other clinical features including gender, stage, and age (Fig. 4 B & Fig. 4 C), and the result indicated that hypoxia was an independent factor, but glycolysis was not independent.

3.2.3. Step3: Correlation analysis

The correlation between two variables is an important index of scientific research. ARMT can flexibly calculate correlation between any continuous variables, such as TPM value and GSVA score. Here, focusing on hypoxia and glycolysis, we used ARMT to calculate the Spearman correlation coefficient between their GSVA score and the TPM value of specific genes. The correlation heatmap is showed in Fig. 5, showing that glycolysis-related genes were correlated with hypoxia score and hypoxia-related genes were correlated with glycolysis score, which further verified the close correlation between glycolysis and hypoxia.

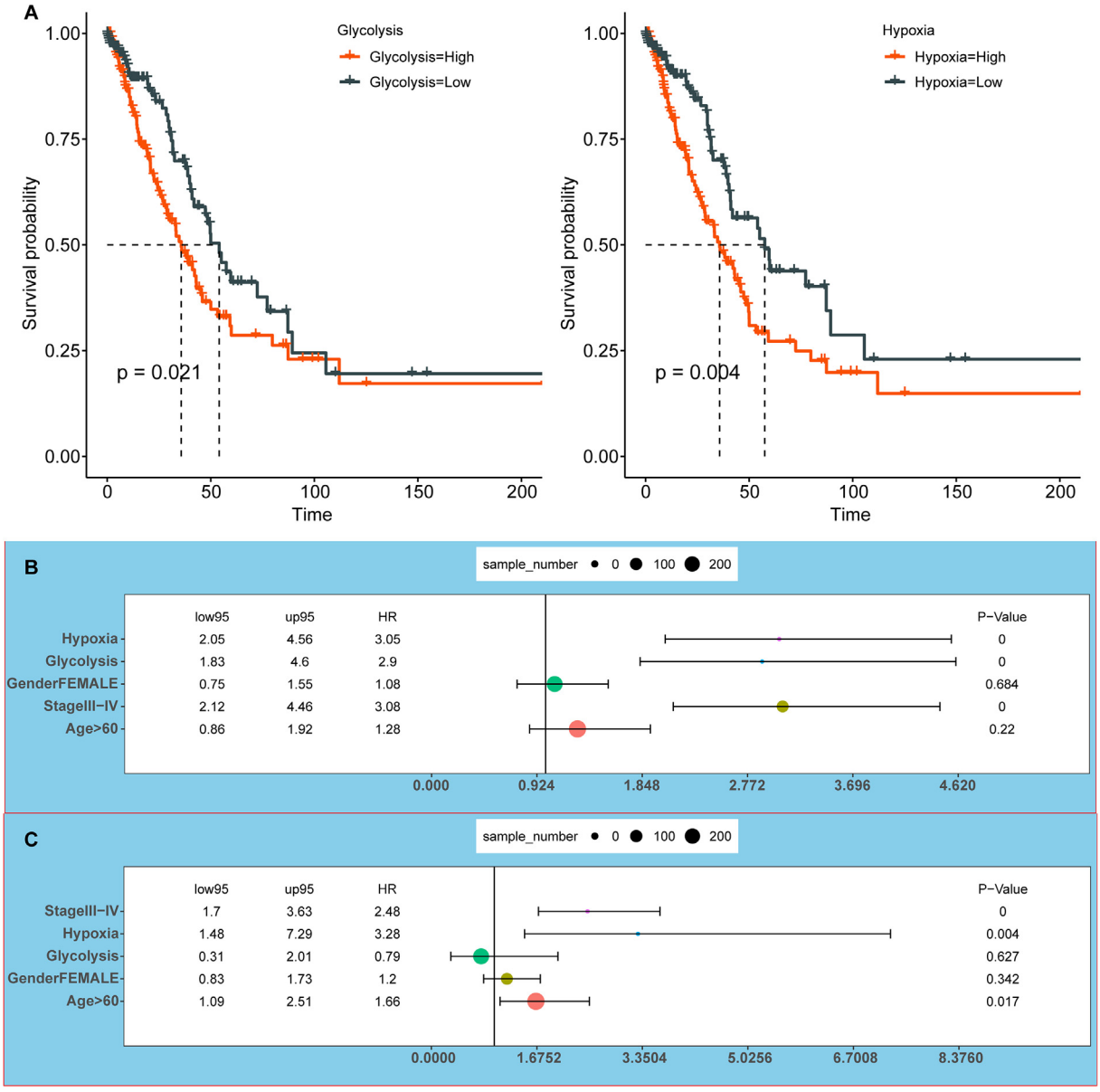

Fig. 4. Survival analysis and cox proportional hazards model. A. Kaplan-Meier curve of glycolysis and hypoxia level **B.** Forest plot for univariate cox proportional hazards regression analysis **C.** Forest plot for multivariate cox proportional hazards regression analysis.

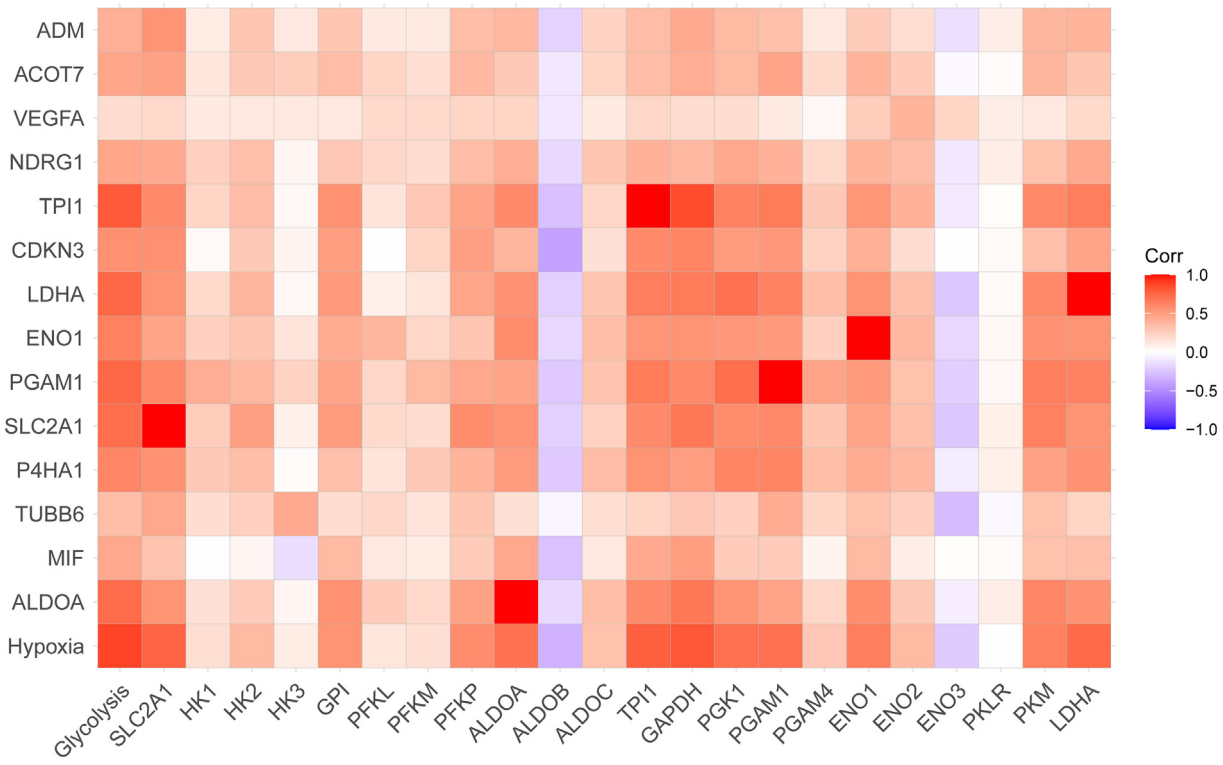

Fig. 5. Heatmap of Spearman correlation coefficient. The horizontal axis represents the glycolysis and glycolysis-related genes; the vertical axis represents the hypoxia and hypoxia-related genes.

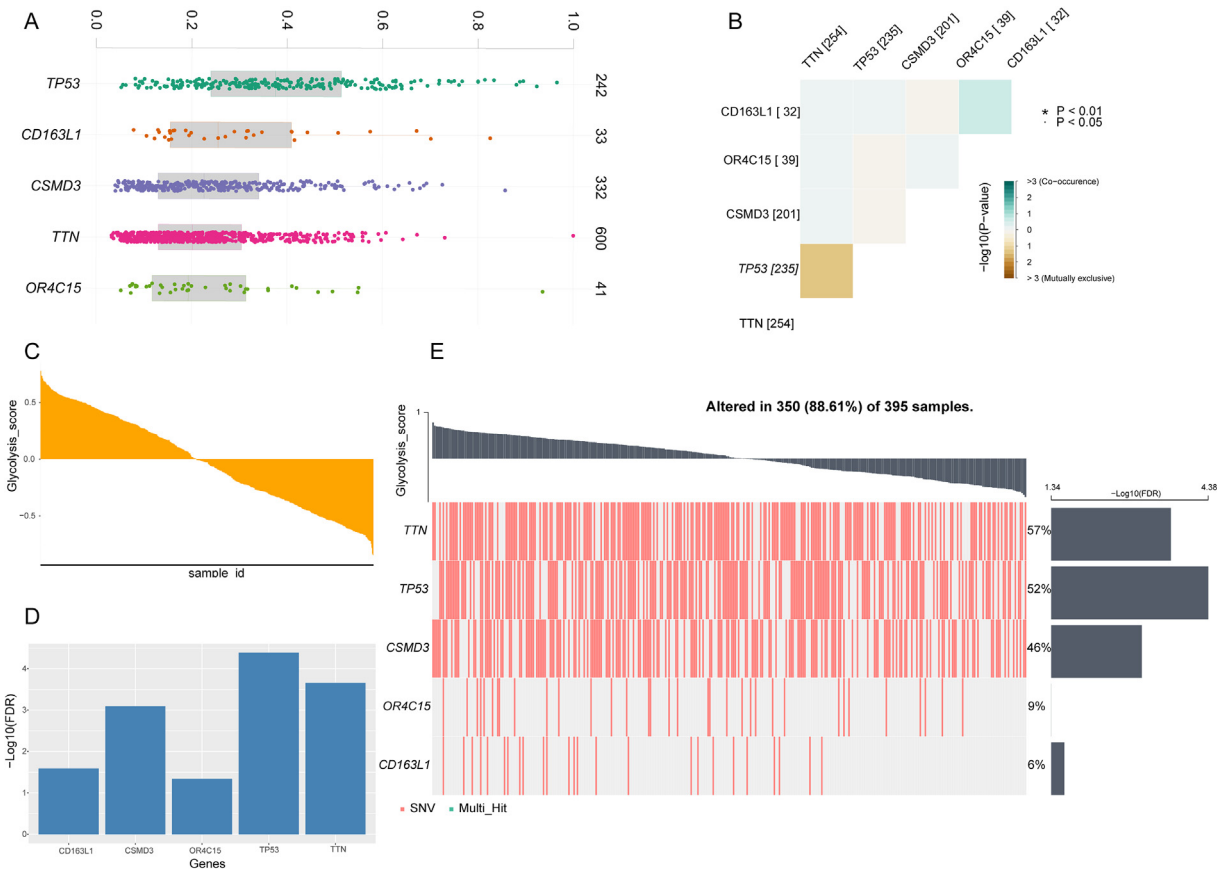

Fig. 6. The mutant map of TCGA-LUAD samples A. Variant allele frequency of differential mutation genes **B.** Mutation correlation heat map **C.** The GSVA score of glycolysis **D.** Differential analysis result of GSVA score **E.** Oncoplot of SNV mutation with data in C&D.

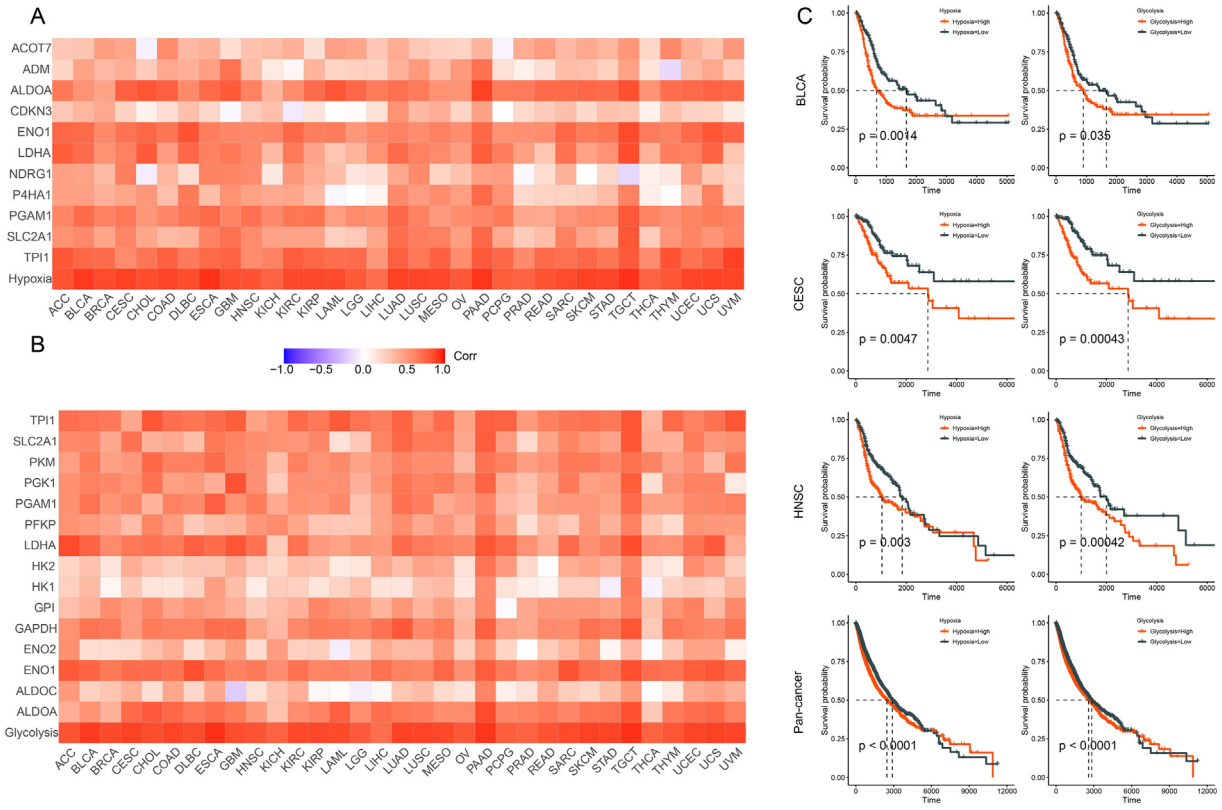

Fig. 7. Correlation analysis and survival analysis in pan-cancer. A. Heatmap of correlation coefficient between the GSVA score of glycolysis and hypoxia with hypoxia related-genes **B.** Heatmap of correlation coefficient between the GSVA score of hypoxia and glycolysis with glycolysis-related genes **C.** Kaplan-Meier curve of glycolysis and hypoxia level.

3.2.4. Step4: Mutant mapping

ARMT can associate GSVA and gene mutation information with the oncoplot so that we can set up a bridge between the transcriptome and genomes, which can further explore the potential relationship between the key genetic mutation and gene expression pattern. As the results shown above, *TP53*, *CD163L1*, *CSMD3*, *TTN*, *OR4C15* were the mutated genes in hypoxia score high samples (Fig. 3 B). Here, we visualized the variant allele frequency (VAF) of these five genes (Fig. 6 A), and presented their mutation correlation heat map to reveal that their mutation is independent. (Fig. 6 B).

Then, the differential analysis based on glycolysis GSVA score was preformed between samples with specific hypoxia-related gene mutation or neutral samples, to obtain their relationship. As shown in Fig. 6 C and Fig. 6 D, the samples were arranged by glycolysis GSVA score of all samples and the -log10(FDR) was shown based on differential analysis of glycolysis GSVA score between samples with *TP53*, *CD163L1*, *CSMD3*, *TTN*, *OR4C15* mutation or not (Fig. 6 D). Then, we used ARMT to display these two results data from GSVA score with oncoplot of mutation data together, which can reveal the potential relationship between hypoxia and glycolytic activation at mutation level (Fig. 6 E). The results showed that hypoxia score high mutated genes including *TP53*, *CSMD3* and *TTN* were also mutated in glycolysis score high samples, which suggested they may be the bridge molecular connecting hypoxia and glycolysis.

3.2.5. Step 5: Analysis of multiple sets of data (For Pan-Cancer et.al)

To make the analysis process more convenient, multiple sets of data can be automatically separated to analyze in one batch, which means ARMT has the ability of pan-cancer analysis.

Including all cancer samples in TCGA according to the cancer types, correlation analysis (between glycolysis score and expression of hypoxia-related genes or hypoxia score and glycolysis-related genes) and survival analysis (grouping based on hypoxia and glycolysis GSVA value) for each cancer were carried out in step 5. In Fig. 7 A, we found a high correlation between glycolysis score and hypoxia genes. In Fig. 7 B, there was a high coefficient between hypoxia score and glycolysis genes across 33 cancer types.

The survival analysis showed that high glycolysis and hypoxia score were associated with poorer survival probability in specific cancer types including BLCA, CESC, HNSC and pan-cancer patterns, which demonstrated these two factors may be the significant prognostic factors in these cancer types (Fig. 7 C).

4. Discussion

With the exponential growth of omics data, it is of great significance for life science and medicine research to explore valuable molecular mechanisms by comprehensive and integrated analysis. In the early stage, our group had conducted several Pan-cancer studies by using TCGA multi-omics data and multiple analysis methods [12,13,54–58]. According to the experience of previous studies, we believe that a comprehensive approach to integrated omics analysis is particularly predominant in revealing molecular mechanisms, and multiple valuable analysis approaches can be integrated and standardized. At present, there are many web tools for integration analysis of TCGA data [68], which promote our research on cancer through public databases. However, more and more studies show that complex diseases or phenotypes are difficult to explain with a single gene, so it is particularly important to use gene-set characterizing a specific biological process to reveal

the deeper biological mechanism. And there are few tools to integrate the analysis of GSVA score into RNA-seq downstream analysis.

Based on our previous research, we provided a more flexible used tool with GUI and have developed ARMT. In this study, we proposed a relatively standardized analysis process based on transcriptome and genomic data and developed them into an automated tool used without programming requirement, which can analyze any public or local relevant data on GUI according to the researcher's interest. From the case in this paper, ARMT can carry out grouping analysis based on the integrated data according to interest of researchers, providing a wide variety of grouping methods. ARMT not only integrates most known RNA-seq downstream analysis steps, but also provides the analysis function for GSVA score. Besides, it can carry out joint analysis with the mutation information of genome, which greatly reduces the time cost of programming and provides a relatively standard comprehensive analysis platform for researchers. We believe that this tool can promote the discovery of life science and medical mechanism.

ARMT provides a way in the integrated and comprehensive downstream functional analysis of RNA-seq, which is convenient and efficient, accelerates the scientific research, saving a lot of time and labor cost especially for users without programming skills.

Authors contribution

HD conceived the study. GH wrote the manuscript and developed ARMT. HZ assisted development and attended the introduction. YQ revised the manuscript. KH assisted development. XG tested ARMT and collected the data. JW revised the manuscript and attended the discussion. HD supervised the whole work and revised the manuscript.

Availability

The developed ARMT software has been shared on https://github.com/Dulab2020/ARMT and can be available freely.

Funding

This work was supported by the National Key R&D Program of China (2018YFC0910201) and the Key R&D Program of Guangdong Province (2019B020226001).

Data availability statement

All data relevant to the study are included in the article or uploaded as supplementary information.

Ethical approval

Not required.

Declaration of Competing Interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.08.009.

References

[1] Li JR, Sun CH, Li WY, Chao RF, Huang CC, et al. Cancer RNA-Seq Nexus: a database of phenotype-specific transcriptome profiling in cancer cells. Nucleic Acids Res 2016;44(D1):D944–51.
[2] Ren SC, Peng ZY, Mao JH, Yu YW, Yin CJ, et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res 2012;22(5):806–21.
[3] Wang JL, Dean DC, Hornicek FJ, Shi HR, Duan ZF. RNA sequencing (RNA-Seq) and its application in ovarian cancer. Gynecol Oncol 2019;152(1):194–201.
[4] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 2013;41(D1):D991–5.
[5] Athar A, Fullgrabe A, George N, Iqbal H, Huerta L, et al. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res 2019;47(D1):D711–5.
[6] Fakih M, Ouyang C, Wang CK, Tu TY, Gozo MC, et al. Immune overdrive signature in colorectal tumor subset predicts poor clinical outcome. J Clin Invest 2019;129(10):4464–76.
[7] Li Y, Jiang T, Zhou W, Li J, Li X, et al. Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. Nat Commun 2020;11(1).
[8] Calabrese C, Davidson NR, Demircioglu D, Fonseca NA, He Y, et al. Genomic basis for RNA alterations in cancer. Nature 2020;578(7793):129.
[9] Chen SJ, Huang V, Xu X, Livingstone J, Soares F, et al. Widespread and functional RNA circularization in localized prostate cancer. Cell 2019;176(4):831.
[10] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17.
[11] Uhlen M, Karlsson MJ, Zhong W, Tebani A, Pou C, et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science 2019;366(6472):1471.
[12] Wei JF, Huang KT, Chen ZX, Hu ML, Bai YM, et al. Characterization of glycolysis-associated molecules in the tumor microenvironment revealed by pan-cancer tissues and lung cancer single cell data. Cancers 2020;12(7).
[13] Liang JS, Cui Y, Meng YH, Li XS, Wang XP, et al. Integrated analysis of transcription factors and targets co-expression profiles reveals reduced correlation between transcription factors and target genes in cancer. Funct Integr Genomic 2019;19(1):191–204.
[14] Alfarsi LH, El Ansari R, Masisi BK, Parks R, Mohammed OJ, et al. Integrated analysis of key differentially expressed genes identifies DBN1 as a predictive marker of response to endocrine therapy in luminal breast cancer. Cancers 2020;12(6).
[15] Huang J, Zhang LR, Li ZM, Lu XJ. Screening and identification of key biomarkers for retinoblastoma Evidence from bioinformatics analysis. Medicine 2020;99(20).
[16] Shin W, Mun S, Choi S, Han K. Application of NanoString technologies in angioimmunoblastic T cell lymphoma. Genes Genom 2020;42(4):485–94.
[17] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139–40.
[18] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12).
[19] Ritchie ME, Phipson B, Wu D, Hu YF, Law CW, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43(7).
[20] Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 2013;22(5):519–36.
[21] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7(3):562–78.
[22] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 2013;31(1):46.
[23] Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. Bmc. Bioinformatics 2010;11.
[24] Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods 2017;14(7):687.
[25] Harris GM, Abbas S, Miles MF. GCSscore: an R package for differential gene expression analysis in Affymetrix/Thermo-Fisher whole transcriptome microarrays. Bmc. Genomics 2021;22(1).
[26] Marini F, Linke J, Binder H. ideal: an R/Bioconductor package for interactive differential expression analysis. Bmc. Bioinformatics 2020;21(1).
[27] Li ZZ, Zuo YM, Xu CH, Varghese RS, Ressom HW. INDEED: R package for network based differential expression analysis. Ieee Int C Bioinform 2018:2709–12.
[28] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 2015;16(1):59–70.
[29] Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. Bmc. Bioinformatics 2017;18.

[30] Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. Plos Comput Biol 2019;15(3).
[31] Sun J, Zhang ZC, Bao SQ, Yan CC, Hou P, et al. Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. J Immunother Cancer 2020;8(1).
[32] Grasselli G, Greco M, Zanella A, Albano G, Antonelli M, et al. Risk factors associated with mortality among patients with COVID-19 in intensive care units in Lombardy, Italy. Jama Intern Med 2020;180(10):1345–55.
[33] Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, et al. Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. Cell Rep 2018;23(1):255–69 e4.
[34] Frankish A, Uszczynska B, Ritchie GRS, Gonzalez JM, Pervouchine D, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics 2015;16.
[35] Archbold HC, Yang YX, Chen L, Cadigan KM. How do they do Wnt they do: regulation of transcription by the Wnt/ss-catenin pathway. Acta Physiol 2012;204(1):74–109.
[36] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. Plos Comput Biol 2012;8(2).
[37] Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007;23(8):980–7.
[38] Ihnatova I, Budinska E. ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data. Bmc. Bioinformatics 2015;16.
[39] Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformatics 2012;13.
[40] Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics 2012;28(18):I451–7.
[41] Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32:D258–61.
[42] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014;42(D1):D199–205.
[43] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA 2005;102(43):15545–50.
[44] 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(1):472–87.
[45] Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 2013;14.
[46] Han K, Pierce SE, Li A, Spees K, Anderson GR, et al. CRISPR screens in cancer spheroids identify 3D growth-specific vulnerabilities. Nature 2020;580(7801):136.
[47] Jin X, Demere Z, Nair K, Ali A, Ferraro GB, et al. A metastasis map of human cancer cell lines. Nature 2020;588(7837):331.
[48] Klemm F, Maas RR, Bowman RL, Kornete M, Soukup K, et al. Interrogation of the microenvironmental landscape in brain tumors reveals disease-specific alterations of immune cells. Cell 2020;181(7):1643.
[49] Shen JZ, Qiu ZX, Wu QL, Finlay D, Garcia G, et al. FBXO44 promotes DNA replication-coupled repetitive element silencing in cancer cells. Cell 2021;184(2):352.
[50] Zeng QQ, Michael IP, Zhang P, Saghafinia S, Knott G, et al. Synaptic proximity enables NMDAR signalling to promote brain metastasis. Nature 2019;573(7775):526.
[51] Zhou YY, Zhou B, Pache L, Chang M, Khodabakhshi AH, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10.
[52] Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 2018;28(11):1747–56.
[53] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016;44(8).
[54] He YT, Li XS, Meng YH, Fu SY, Cui Y, et al. A prognostic 11 long noncoding RNA expression signature for breast invasive carcinoma. J Cell Biochem 2019;120(10):16692–702.
[55] Li XS, Yu XK, He YT, Meng YH, Liang JS, et al. Integrated Analysis of MicroRNA (miRNA) and mRNA Profiles reveals reduced correlation between microrna and target gene in cancer. Biomed Res Int 2018;2018.
[56] Shi Y, Wei J, Chen Z, Yuan Y, Li X, et al. Integrative analysis reveals comprehensive altered metabolic genes linking with tumor epigenetics modification in pan-cancer. Biomed Res Int 2019;2019:6706354.
[57] Wang JJ, Li SD, Lin SD, Fu SY, Qiu L, et al. B-cell lymphoma 2 family genes show a molecular pattern of spatiotemporal heterogeneity in gynaecologic and breast cancer. Cell Proliferat 2020;53(6).
[58] Yu XK, Liang JS, Xu JR, Li XS, Xing S, et al. Identification and validation of circulating MicroRNA signatures for breast cancer early detection based on large scale tissue-derived data. J Breast Cancer 2018;21(4):363.
[59] Castillo-Secilla D, Galvez JMG, Carrillo-Perez F, Verona-Almeida M, Redondo-Sanchez D, et al. KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge. Comput Biol Med 2021;133.
[60] Chen LY, Wei KC, Huang ACY, Wang K, Huang CY, et al. RNASEQR-a streamlined and accurate RNA-seq sequence analysis program. Nucleic Acids Res 2012;40(6).
[61] Samur MK. RTCGAToolbox: A new tool for exporting TCGA firehose data. PLoS ONE 2014;9(9).
[62] Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol 2010;11(12).
[63] Yu GC, Wang LG, Han YY, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 2012;16(5):284–7.
[64] Bhandari V, Hoey C, Liu LY, Lalonde E, Ray J, et al. Molecular landmarks of tumor hypoxia across cancer types. Nat Genet 2019;51(2):308.
[65] Black JC, Atabakhsh E, Kim J, Biette KM, Van Rechem C, et al. Hypoxia drives transient site-specific copy gain and drug-resistant gene expression. Gene Dev 2015;29(10):1018–31.
[66] Fischer GM, Gopal YNV, McQuade JL, Peng W, DeBerardinis RJ, et al. Metabolic strategies of melanoma cells: Mechanisms, interactions with the tumor microenvironment, and therapeutic implications. Pigm Cell Melanoma R 2018;31(1):11–30.
[67] Schito L, Rey S. Cell-autonomous metabolic reprogramming in hypoxia. Trends Cell Biol 2018;28(2):128–42.
[68] Zhang Z, Li H, Jiang S, Li RJ, Li WY, et al. A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data. Brief Bioinform 2019;20(4):1524–41.