frontiers | Frontiers in Bioinformatics

TYPE Brief Research Report
PUBLISHED 19 June 2025
DOI 10.3389/fbinf.2025.1604418

OPEN ACCESS

EDITED BY
Keith A. Crandall,
George Washington University, United States

REVIEWED BY
Youtao Lu,
University of Pennsylvania, United States
Guanjue Xiang,
Dana–Farber Cancer Institute, United States
Bo-Wei Zhao,
Zhejiang University, China
Jianlei Gu,
Yale University, United States

*CORRESPONDENCE
Juan I. Fuxman Bass,
fuxman@bu.edu

RECEIVED 01 April 2025
ACCEPTED 05 June 2025
PUBLISHED 19 June 2025

CITATION
Li Z and Fuxman Bass JI (2025) ICARus: a pipeline to extract robust gene expression signatures from transcriptome datasets. *Front. Bioinform.* 5:1604418. doi: 10.3389/fbinf.2025.1604418

ICARus: a pipeline to extract robust gene expression signatures from transcriptome datasets

Zhaorong Li[1] and Juan I. Fuxman Bass[1,2,3,4]*

[1]Bioinformatics Program, Boston University, Boston, MA, United States, [2]Department of Biology, Boston University, Boston, MA, United States, [3]Program in Molecular Biology, Cell Biology and Biochemistry, Boston University, Boston, MA, United States, [4]Biological Design Center, Boston University, Boston, MA, United States

Gene signature extraction from transcriptomics datasets has been instrumental to identify sets of co-regulated genes, identify associations with prognosis, and for biomarker discovery. Independent component analysis (ICA) is a powerful tool to extract such signatures to uncover hidden patterns in complex data and identify coherent gene sets. The ICARus package offers a robust pipeline to perform ICA on transcriptome datasets. While other packages perform ICA using one value of the main parameter (i.e., the number of signatures), ICARus identifies a range of near-optimal parameter values, iterates through these values, and assesses the robustness and reproducibility of the signature components identified. To test the performance of ICARus, we analyzed transcriptome datasets obtained from COVID-19 patients with different outcomes and from lung adenocarcinoma. We identified several reproducible gene expression signatures significantly associated with prognosis, temporal patterns, and cell type composition. The GSEA of these signatures matched findings from previous clinical studies and revealed potentially new biological mechanisms. ICARus with a vignette is available on Github https://github.com/Zha0rong/ICArus.

KEYWORDS

independent component analysis, transcriptomics, signatures, machine learning, robustness

Introduction

Transcriptomic data plays a crucial role in understanding the variation in gene expression patterns across diverse biological conditions and phenotypes. Common approaches to analyze such data involve conducting a differential expression and gene expression pattern analyses, which evaluate changes in expression across groups (Conesa et al., 2016). However, challenges arise when analyzing large transcriptomic datasets from sources like Genotype-Tissue Expression (GTEx) (Lonsdale et al., 2013) and the Cancer Genome Atlas (TCGA), since these data can often be classified according to multiple known (e.g., tissue, sex, age, tumor type, tumor stage, *etc.*) as well as unknown variables. This complicates identifying the contribution of the different variables to the differences in expression observed across

samples. To address these issues and enable the analysis of such large datasets, unsupervised algorithms like principal component analysis (PCA), Weighted Gene Co-Expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008), non-negative matrix factorization (NMF) (Jia et al., 2015), and independent component analyses (ICA) (Anglada-Girotto et al., 2022) have been developed. Unlike methods that compare gene expression between groups, these unsupervised algorithms identify gene expression modules or signatures associated with the phenotype labels of the samples.

ICA has been widely used to identify gene expression signatures in large transcriptomic datasets, including cancer, development, and exposure to treatments (Biton et al., 2014). ICA separates a multivariate signal, in this case gene expression, into additive subcomponents or signatures which are positive and negative contributions of each gene in the dataset. One key parameter in ICA is determining the optimal number of signatures to extract in a dataset as there is no ground truth for the actual number of independent contributing variables. Most pipelines, such as RobustICA (Anglada-Girotto et al., 2022) and BIODICA (Kairov et al., 2012) select this optimal parameter based on the number of components needed in PCA to explain a percentage of variance in the dataset. These studies often increase robustness by iterating the analysis using the same parameter; however, signatures often vary widely across parameter values. This can lead to the identification of low-confidence, non-reproducible signatures.

Here, we introduce the R package ICARus (Figure 1A), designed to streamline the application of ICA and extraction of high-confidence expression signatures that are robust across iterations and reproducible across parameter values. ICARus leverages the proportion of variance explained obtained from PCA to provide a range of near-optimal parameters for the ICA algorithm. Subsequently, for each parameter the ICA algorithm is applied, and the results are clustered and evaluated using the stability index proposed by Icasso to identify robust signatures for each parameter (Himberg and Hyvarinen, 2003). ICARus then clusters the robust signatures obtained to identify reproducible signatures across parameters. Finally, the gene expression signatures with the highest reproducibility scores are combined into meta-signatures and subjected to further analysis through Gene Set Enrichment Analysis (GSEA) or Fisher's Exact test to functionally interpret the signatures identified.

Materials and methods

Input data format for ICARus

The input data for ICARus is a normalized transcriptome dataset in matrix format, with rows being gene names and columns being samples (Figure 1B). Normalization methods such as Counts-per-Million (CPM) (Chen et al., 2025) and Ratio of median (Anders and Huber, 2010) are recommended. Different normalization method strategies will introduce differences in the final results of ICARus; however, most of the signatures are reproducible in the results obtained from different methods (Supplementary Figure S1).

Prefiltering of sparsely expressed genes in the input data is recommended as these genes introduce noise in the analysis, but since the filtering strategy varies between different datasets, it is not included in the pipeline.

Estimating the set of near-optimal parameters for the ICA algorithm

To estimate the set of near-optimal parameters, ICARus first performs PCA for the input dataset (Figure 1C). Prior work has used an optimal parameter N as the number of top principal components that collectively account for 99% of the variance observed in the dataset (Sastry et al., 2019). ICARus, also relies on the variance explained by PCA, but identifies the range of near-optimal values for n. After performing PCA, users can select whether to use: 1) the ranked distribution of standard deviations of each principal component, or 2) the cumulative proportion of variance explained by a certain number of principal components to determine the lower bound for the parameter set. In the first option, the standard deviation of each principal component is plotted against the ranked order of the principal components which takes the form of an elbow plot; whereas in the second option, the cumulative proportion of variance explained against the order of principal components takes the form of a knee plot. The elbow-point in the first plot and the knee-point in the second plot indicates the top n principal components that explain a large fraction of the variance in the data (Figure 1D), i.e., including more principal components does not lead to a marked increase in variance explained.

To pinpoint this critical elbow/knee point, the Kneedle Algorithm (Satopaa et al., 2011) is used, and this identified point is designated as the minimum number n for the near-optimal parameter set for subsequent ICA analysis. The Kneedle algorithm was implemented in an R package[1]. This set of parameters is then selected as every integer (n, n + k) where k can be user defined and is set as default to be 10.

Generating reproducible gene signatures

Following the identification of the near-optimal parameter set, ICARus initiates the generation of reproducible gene signatures employing two sequential strategies: intra-parameter iterations and inter-parameter iterations. For the intra-parameter iterations, ICARus conducts the ICA algorithm 100 times for each n value. Subsequently, the resulting signatures undergo sign correction suggested by a previous study (Anglada-Girotto et al., 2022) and hierarchical clustering to identify sets of robust signatures for each specific n. Within each cluster, the medoid is extracted and employed as the representative signature, while the stability of the signature cluster is assessed using the stability index proposed by Icasso (Figure 1E) (Himberg and Hyvarinen, 2003). To calculate the stability index, the similarities between signatures from different runs are calculated using the absolute value of the Pearson correlation coefficient $\sigma_{i,j}$. Then the stability index

1 https://github.com/etam4260/kneedle

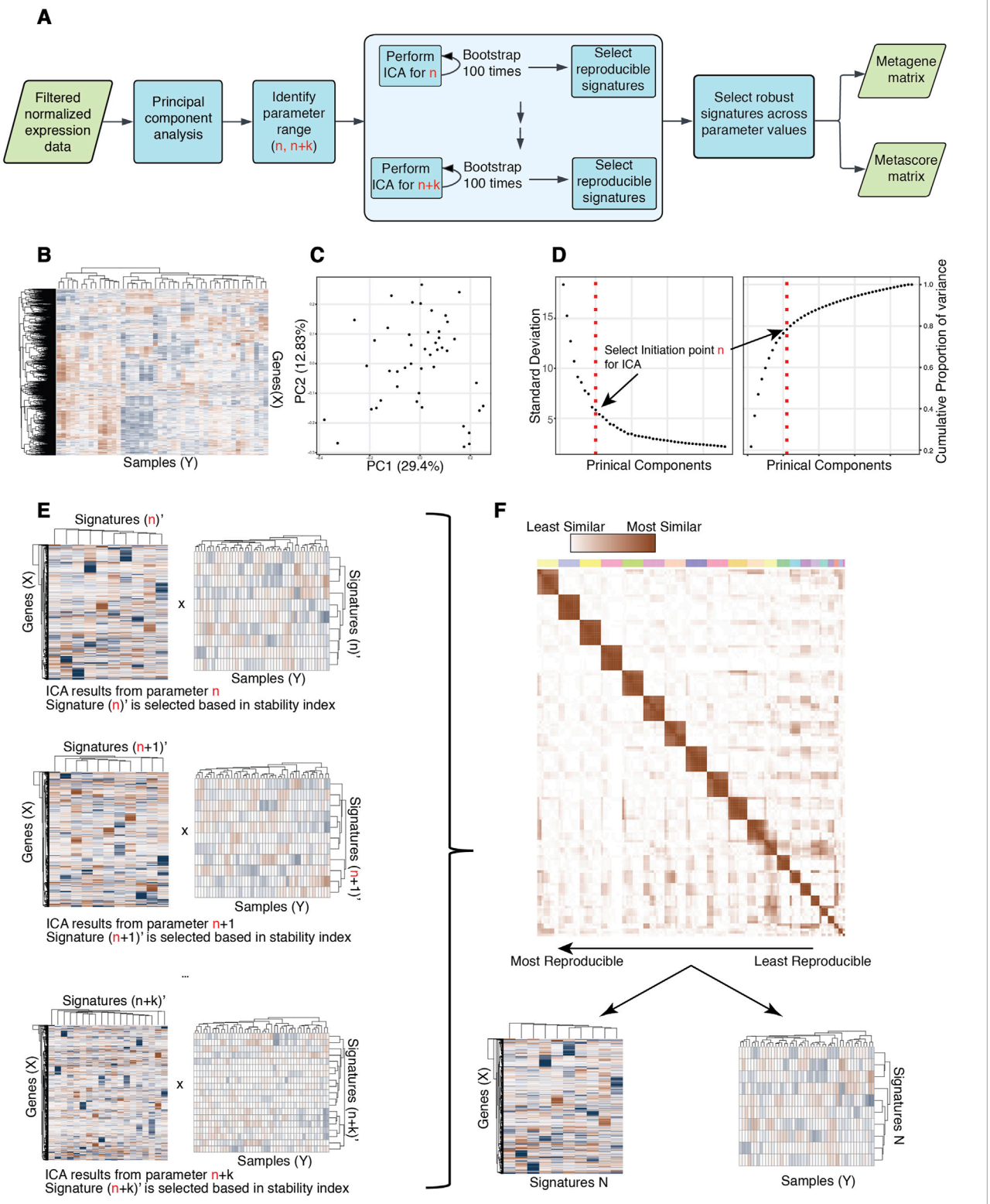

FIGURE 1
Overview of ICARus pipeline. **(A)** Pipeline diagram overview of ICARus. **(B)** The input for ICARus is a (Genes x Samples) normalized gene expression matrix where the rows are gene symbols (IDs) and columns are samples. **(C)** PCA plot of samples based on normalized gene expression. **(D)** Left = Standard deviation of each principal component sorted by principal component rank order. Right = Cumulative proportion of variance explained across principal components. The elbow and knee-points in these plots are used to identify the initiation point n for ICA. **(E)** For each parameter between n and n + k (k is defined by user) ICA is performed 100 times, and Icasso quality index is used to assess the robustness of independent components. **(F)** The independent components that pass the user defined robustness threshold for each tested parameter value are clustered. The sizes of clusters indicate the reproducibility of signatures across different parameter values. Signatures that pass the user defined reproducibility scores are output as genes x signatures and signatures x samples matrices.

of given cluster M is calculated using the following function in Equation 1:

$$\frac{1}{|C_M|^2}\sum_{i,j\in C_M}\sigma_{i,j} - \frac{1}{|C_M||C_{-M}|}\sum_{i\in C_M}\sum_{j\in C_{-M}}\sigma_{i,j} \quad (1)$$

where $|C_M|$ and $|C_{-M}|$ are the size of cluster M and the number of signatures not in cluster M (Himberg and Hyvarinen, 2003). The stability index calculated by this function ranges from 0 to 1, from least to most stable. The signatures with stability indices >0.75 are evaluated for reproducibility across values of n (Figure 1F). These robust signatures are subjected to hierarchical clustering. A signature obtained with one value of the parameter is considered reproducible if it clusters together with signatures obtained across multiple other n values within the near-optimal set. The user can specify whether to only keep the reproducible signatures originated from the starting point n, or to also keep the reproducible signatures originated from a higher parameter within the near-optimal set, as long as they can be reproduced in more than half of the remaining tested parameters.

ICARus outputs the number of near-optimal values that contribute a signature to the cluster and the average distance between these signatures for each cluster. These values can then be used to select reproducible signatures across many parameter values (Figure 1F).

Output signatures and downstream analysis of the gene signatures

The reproducible signatures extracted by ICARus consist of two parts: 1) a matrix of genes by signatures, where each value indicates the contribution of the gene to the signature (the distribution in scores of the signatures follows the normal distribution, with the mean of 0); and 2) a matrix of signatures by samples where each value indicates the contribution of the signature to the expression profile of the sample (Figure 1F). The gene scores of a particular signature can be used to perform Gene Set Enrichment Analysis (Subramanian et al., 2005) to identify pathways or gene sets associated with the signature for further biological interpretation. The signatures scores across samples can be used to associate signature values with sample phenotypes or temporal patterns.

Implementation

Steps that are described above were implemented in R with parallel backend computation as package *ICARus* and provided as pseudocode in Supplementary Material S1. The package and a vignette is available on Github https://github.com/Zha0rong/ICArus.

Test datasets

Peripheral leukocyte samples from COVID-19 patients

To illustrate the efficacy of ICARus in identifying relevant signatures, we applied it to a publicly available RNA-Seq dataset featuring 46 peripheral blood leukocyte samples collected from 11 COVID-19 patients infected with SARS-CoV-2, with varying clinical outcomes (fast recovery, prolonged recovery, and fatal) at different time points (Figure 2A) (Lam et al., 2023). Fast recovery patients had a median hospitalization time of 7 days, prolonged recovery patients had a median hospitalization time of 25 days, and fatal patients were patients that passed away due to complications of the infection. The count matrix was downloaded from the GEO repository (GSE221066), which included 26,475 genes and 55 samples. To prevent genes with sparse expression introducing noise in the analysis, only genes with non-zero expression in at least one-fourth of the samples were included in the analysis. This strategy filtered out 8,918 genes and retained 17,557 genes for the analysis. The count matrix was normalized using the Counts-Per-Million method (Chen et al., 2025).

Primary tumor samples of lung adenocarcinoma (LUAD)

To test the performance of ICARus on a large RNA-seq dataset with complex clinical phenotypes, we processed 539 lung adenocarcinoma primary tumor RNA-seq samples from TCGA database (Cancer Genome Atlas Research Network, 2014), which were downloaded through the TCGA-biolinks portal (Colaprico et al., 2016). The count matrix included 19,938 protein coding genes and 539 samples. To filter out genes with sparse and low expression in the dataset, only genes with non-zero expression in at least one-fourth of the samples were included in the analysis. This strategy filtered out 1,417 genes and retained 18,091 genes and 539 samples for the analysis. The count matrix was normalized using the Counts-Per-Million method (Chen et al., 2025).

Results

Identification of reproducible signatures in a COVID-19 expression dataset

To identify signatures associated with COVID-19 outcomes, we used a dataset of 46 samples derived from 11 patients with different outcomes (fast recovery, prolonged recovery, and fatal) at different time points (Figure 2A) (Lam et al., 2023). First, we determined the near-optimal parameter set in the COVID-19 expression dataset. We then selected the critical elbow-point in the PCA option provided by ICARus. This corresponded to 10 principal components; therefore, the nominated range for the ICA parameter was 10–19 independent components. We used 100 iterations for each of these parameter values, then identified the medoid signature, followed by clustering of signatures across the parameter values. This resulted in 10 signatures that were reproducible across more than half of the tested parameter values (Figures 2B,C). By comparing the signature scores between samples from patients with different clinical outcomes, we identified two signatures (signatures 4 and 10) that monotonically increase with outcome severity (Figure 2D). Next, we aimed to determine the biological processes associated with these signatures.

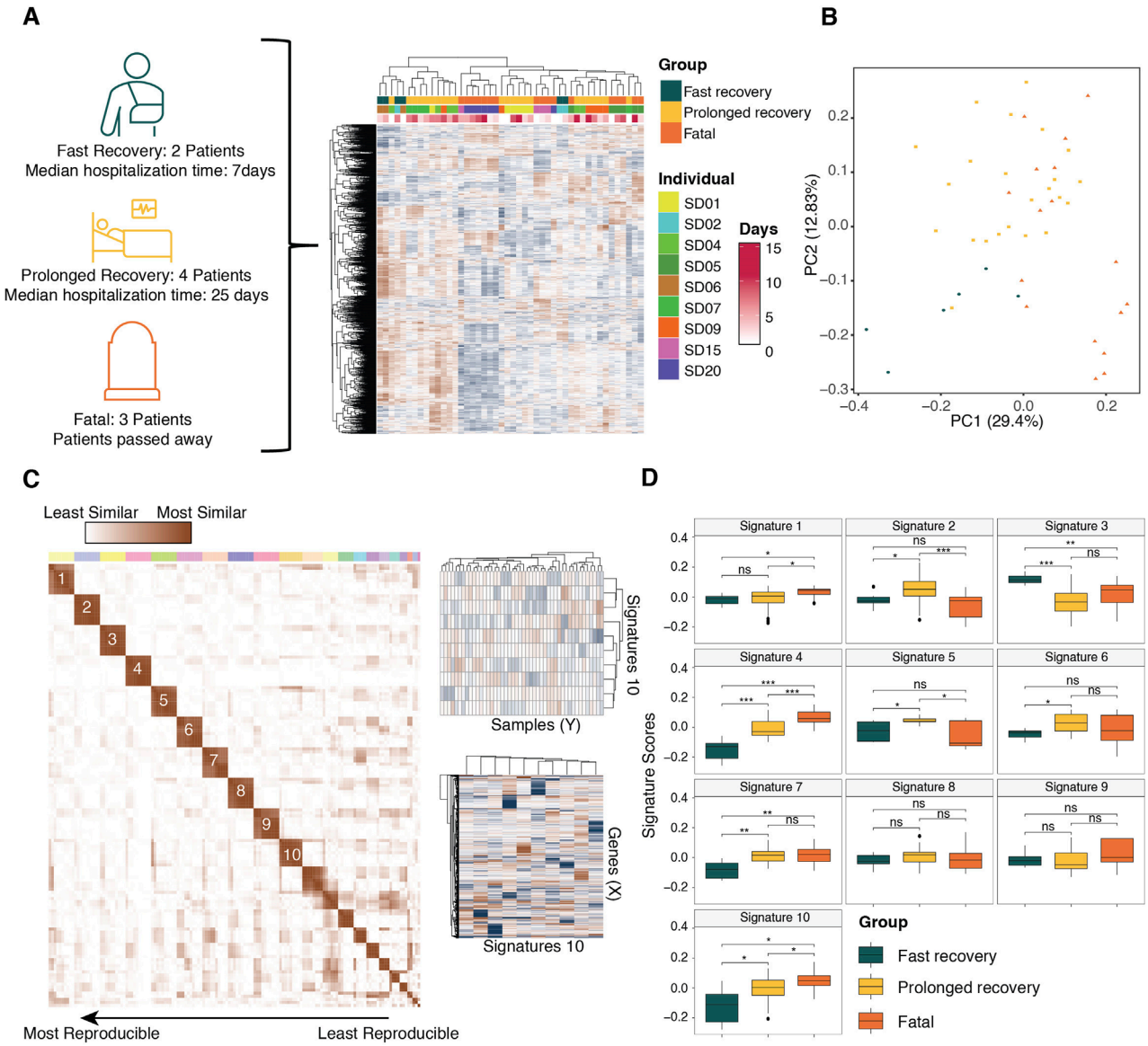

FIGURE 2
Application of ICARus to a COVID-19 transcriptomic dataset. **(A)** The test dataset consists of 46 samples of blood-derived leukocytes obtained from COVID-19 patients with different clinical outcomes at different time points during infection. After filtering genes with no expression in more than half of the dataset, 17,557 genes were kept for downstream analysis. **(B)** The PCA plot illustrates the separation of samples from different clinical outcomes. **(C)** The initiation parameter identified by ICARus for this dataset was 10 (n), and ICARus determined robust signatures using parameter values 10–19. Signatures across parameter values were clustered and only the signatures that were reproducible across more than 5 values were considered for downstream analysis. ICARus identified 10 robust and reproducible gene expression signatures from the test dataset. **(D)** The box plots showed the signature score distributions in different clinical outcomes. Statistical significance determined by Wilcoxon-ranked sum test. *p < 0.05, **p < 0.01, ***p < 0.005.

Signature 4 is associated with poor prognosis and fatal outcomes

Signature 4 exhibited a significant correlation with patient outcomes, with samples from fatal outcome patients having the highest scores and those from fast-recovery patients having the lowest scores (Figure 3A). By plotting signature scores across time points and clinical outcomes, we observed that signature 4 scores were higher in samples from fatal outcome patients, and lower in fast-recovery patients at every time point (Figure 3A). This observation is important as it rules out the possibilities of association driven by the bias at one or more time points and suggests that the biological functions associated with signature 4 can be used to differentiate fast recovery patients at any time point.

The GSEA analysis of signature 4 revealed a depletion of T and B cell activation and MHC class II antigen processing and presentation, and an enrichment of inflammation pathways and MHC Class I antigen presentation (Figure 3B). To identify which

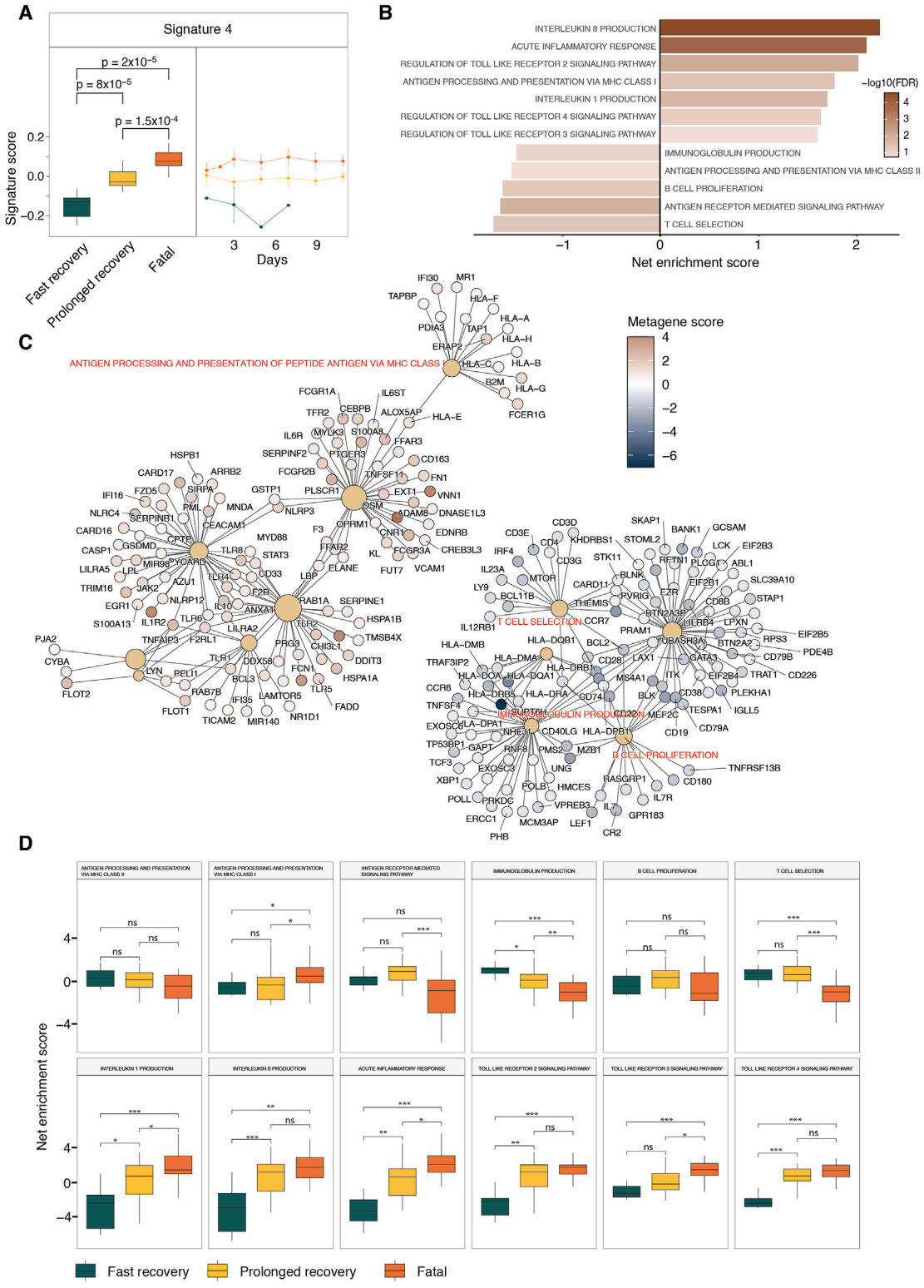

FIGURE 3
GSEA of clinical outcome-associated Signature 4. **(A)** The box plots show the signature 4 score distributions in different clinical outcomes. The line plot shows the temporal pattern of Signature 4 for different patient outcomes. Statistical significance determined by Wilcoxon-ranked sum test. **(B)** Bar graph displays the top enriched and depleted pathways from GSEA analysis results of Signature 4. Net enrichment scores are shown. **(C)** Network representation of the top enriched pathways and the driver genes associated with the enrichment results. **(D)** ssGSEA results of the top enriched pathways from GSEA analysis results of Signature 4.

genes were driving the observed enrichments, a net plot was generated (Figure 3C). In this plot, large brown nodes represent enrichment terms, while smaller red or blue nodes represent individual genes colored according to their scores in signature 4. Edges were drawn between nodes when a gene belonged to the core enrichment set of a given term. In the network graph, elevated scores of immune genes, such as IL1R2, MYD88, NRLP3, CASP1, TLR4, were shown to drive the enrichment of interleukin 1 and interleukin 8 producing signaling pathways (Figure 3C). This is consistent with previous studies linking the elevated expression of IL-1 and IL-8 with poor prognosis (Li et al., 2021; Cavalli et al., 2021). Further clinical studies also showed that the blocking of IL-1 in COVID-19 patients led to better prognosis (Cavalli et al., 2020). The network also showed that toll-like receptor genes such as TLR1, TLR2 and TLR4, which had elevated metagene scores, were driving the enrichment of toll-like receptor signaling pathways (Figure 3C). The TLR2 signaling pathway, elevated in signature 4, can also be associated with poor prognosis, consistent with several clinical studies showing elevated TLR2 expression was associated with poor prognosis in COVID-19 infection (Taniguchi-Ponciano et al., 2021; Xu Q. et al., 2022). Signature 4 has also a negative association with MHC-Class II antigen presenting pathways, which is driven by the suppression of MHC-Class II such as HLA-DMA, HLA-DMB and HLA-DRA (Figure 3C), suggesting a negative association with poor prognosis. This is consistent with previous studies showing that monocytes in COVID-19 patients have lower levels of MHC class II proteins (Xu Q. et al., 2022; Laing et al., 2020). These results were confirmed using ssGSEA (Reich et al., 2006) that calculate net enrichment scores of individual pathways in each sample (Figure 3D).

ICARus identified signature 10 as associated with a temporal phenotype

Signature 10 not only displayed an association with clinical outcomes (lowest in fast recovery, highest in fatal), but also showed a temporal phenotype (Supplementary Figure S2A). Prolonged recovery patients and fatal outcome patients had similar signature 10 scores in the beginning time point, but fatal outcome patients had consistent higher signature 10 scores in the later time points (Supplementary Figure S2A). GSEA analysis of signature 10 revealed positive associations with regulation of neutrophils chemotaxis/mediated immunity, actin filaments assembly/organization and extracellular matrix (Supplementary Figure S2B). A net plot was generated for the genes and the enriched terms to visualize the genes that drive the enrichment of given pathways (Supplementary Figure S2C). For example, matrix metalloproteinase genes such as MMP2 and MMP8 drive the enrichment of extracellular matrix disassembly and galectin genes such as LGALS1, LGALS3 and LGALS9 drive the enrichment of neutrophil mediated immune pathways. Previous studies (Schulte-Schrepping et al., 2020) have shown that elevated neutrophil counts are associated with a poor prognosis in COVID-19 patients, with clinical publications attributing the poor prognosis to the formation of neutrophil extracellular traps (NETs) (Zuo et al., 2021). NETs, composed of cell-free DNA, histones, and cytosolic proteins released by neutrophils, require the rearrangement of the actin cytoskeleton for their formation (Sprenkeler et al., 2022). NETs have been implicated in thrombosis and tissue damage (Papayannopoulos, 2018; Zuo et al., 2020), contributing to the poor prognosis of COVID-19 patients.

GSEA analysis results also revealed a negative association between signature 10 and regulation of T cell activation and T cell mediated immunity, driven by suppression of killer cell lectin-like receptors such as KLRC2/3/4, KLRD1 and KLRK1 (Supplementary Figure S2C). Previous studies have also shown decreasing T cell counts in COVID-19 patients with severe symptoms compared to the ones with non-severe symptoms (Liu et al., 2020). Further, another study reported an elevated number of neutrophils and decreasing number of T cells in COVID-19 patients with severe symptoms compared with COVID-19 patients with mild symptoms (Xu J. et al., 2022).

Identification of reproducible signatures in a TCGA-LUAD expression dataset

To demonstrate the application of ICARus on another larger dataset, we selected the TCGA-LUAD lung adenocarcinoma expression dataset (Figure 4A) (Cancer Genome Atlas Research Network, 2014). To identify the near-optimal parameter set, we selected the critical elbow-point in the PCA option provided by ICARus. This corresponded to 48 principal components, and therefore, the nominated range for the ICA parameter was 48–57 independent components. We performed 100 iterations for each of these parameter values, identified the medoid signature, then clustered signatures across the parameter values. This resulted in 22 signatures that were reproducible across more than half of the parameter values tested (Figure 4B).

Identification of gene signatures associated with disequilibrium of cell type proportion and adverse prognosis

To identify signatures associated with adverse prognosis, samples were stratified by the median score of each gene signature into two groups: samples with higher given gene signature scores and samples with lower given gene signature scores. The Cox proportional hazards model was used to perform the survival analysis and the likelihood ratio test was used to regress out co-variables such as age of diagnosis, gender, location of tumor origin, and cell type proportion. Four signatures were significantly associated with adverse prognosis (Figure 4C). We performed GSEA for signatures 9 and 10 to determine the pathways associated with adverse prognosis. GSEA of signature 10 showed an enrichment of keratinization related processes and depletion of metabolic, immune-related, and cell division related pathways (Figure 4D). Previous studies have shown that keratin gene expression activates the epithelial-mesenchymal transition in tumor cells and leads to poor prognosis (Li et al., 2024). GSEA of signature 9 showed an enrichment of appendage development pathways and depletion of macrophages and immune related pathways (Figure 4E). Previous studies have shown that the enrichment of appendage development pathways was associated with poor prognosis (Yu et al., 2024).

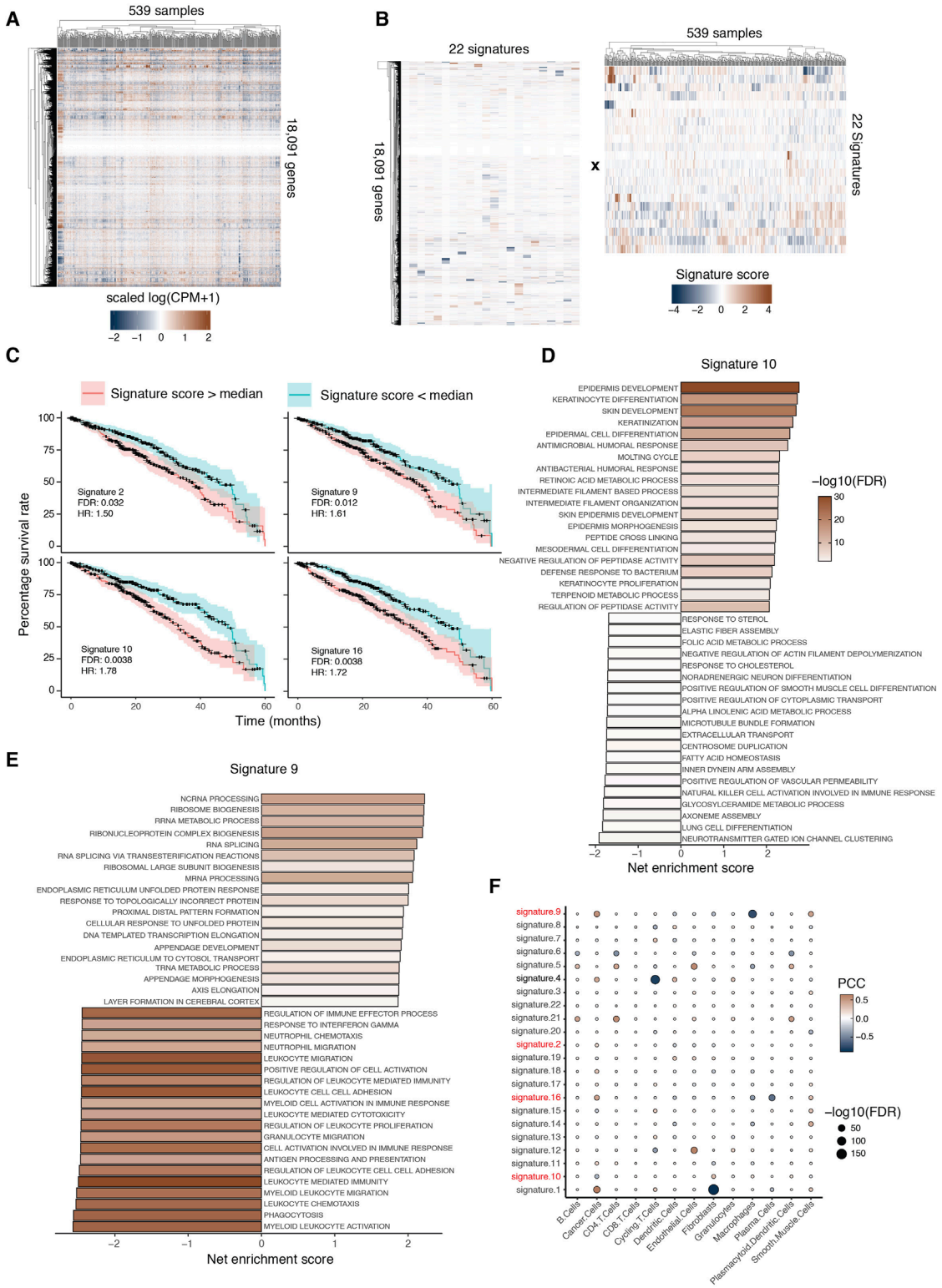

FIGURE 4
ICARUS extracts prognosis-related signatures from TCGA-LUAD database. **(A)** 539 primary tumor RNA-Seq samples were downloaded from the TCGA database, and 18,091 genes were kept in the analysis. **(B)** ICARUS identified 22 reproducible signatures. **(C)** Kaplan Meier plots of 4 signatures significantly associated with adversary prognosis. The adjusted p-values obtained from the Cox-proportional hazard ratio test, covariate factors such as gender, age, tissue of origin were regressed out using likelihood ratio test. HR = hazard ratio. **(D and E)** The bar graph displays the top enriched and depleted pathways from GSEA results of Signature 10 **(D)** and Signature 9 **(E)**. Net enrichment scores are shown. **(F)** Dot plot of Pearson correlation coefficient (PCC) and adjusted p-value of correlation tests between signature scores and cell type proportion. The color of the dot showed correlation coefficients and the size of the dots showed $-\log_{10}$ (adjusted p-value).

To determine whether signature 9 is indeed associated with a depletion of macrophages in the corresponding samples, we used BayesPrism (Chu et al., 2022) to deconvolve the 539 bulk RNA-Seq samples and predict the proportion of each cell type in the tumor microenvironment of each sample. Then, we performed correlation tests between the proportion of each cell type and score of each signature. The results were visualized using the dot plot where the rows are signatures and the columns are cell type proportions, the sizes of the dot are the negative log10 transformed FDR adjusted p-values of the correlation tests and the colors of the dots are the correlation coefficients (Figure 4F). We found several signatures associated with cell type proportions. In particular, signature 9 was significantly associated with a low proportion of macrophages, consistent with our GSEA results.

Discussion

We developed ICARus, an R package designed to assist researchers in identifying robust and reproducible gene signatures using ICA across multiple parameter values. This pipeline is highly versatile enabling users to select analysis parameters and stringency. First, ICARus enables the user to manually or automatically select near-optimal parameter sets using elbow or knee points of PCA results. Next, ICARus allows users to select the reproducibility criteria. Although our analyses focused on signatures identified in more than half of all parameters tested, the pipeline can also output signatures present in more than half of parameters from their first instance. This allows the identification of signatures specific to higher parameter values.

To show that the gene expression signatures extracted by ICARus are meaningful, we tested the package on two RNA-Seq datasets. The first dataset consisted of leukocytes samples which were obtained from COVID-19 patients with different clinical outcomes, and the second dataset consisted of primary tumor samples obtained from lung adenocarcinoma patients. The analysis of COVID-19 patient samples showed that ICARus identified biologically meaningful signatures that were associated with patient prognosis and the pathways that drive these associations.

Analysis of the primary tumor samples showed that ICARus identified gene signatures associated with prognosis and cell type proportion in the tumor microenvironment. The reproducible signatures identified by ICARus were associated with clinical phenotypes and temporal patterns consistent with previous studies. Furthermore, the network analyses of the signatures domonstrated that the signatures will provide biologically meaningful genes driving the enrichment of relevant biological functions. These genes can be used as input for some of the recently published algorithms that employ deep learning algorithms to study gene interactions networks and drug response (Zhao et al., 2024; Zhao et al., 2025; Zhao et al., 2022). In principle, ICARus can also be used to extract signatures from single cell RNA-seq datasets; however, the method may need adaptation to account for noise and missing values.

In summary, ICARus has demonstrated the ability to produce biologically meaningful and reproducible signatures which can be extended to other expression datasets.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

ZL: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. JIFB: Funding acquisition, Resources, Supervision, Visualization, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was funded by the National Institutes of Health grants R35 GM128625 awarded to JIFB.

Acknowledgments

We want to thank Devlin Moyer for testing and providing feedback on the ICARus package.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2025.1604418/full#supplementary-material

References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11 (10), R106. doi:10.1186/gb-2010-11-10-r106

Anglada-Girotto, M., Miravet-Verde, S., Serrano, L., and Head, S. A. (2022). Robustica: customizable robust independent component analysis. *BMC Bioinforma.* 23 (1), 519–9. doi:10.1186/s12859-022-05043-9

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., et al. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* 9 (4), 1235–1245. doi:10.1016/j.celrep.2014.10.035

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511 (7511), 543–550. doi:10.1038/nature13385

Cavalli, G., De Luca, G., Campochiaro, C., Della-Torre, E., Ripa, M., Canetti, D., et al. (2020). Interleukin-1 blockade with high-dose anakinra in patients with COVID-19, acute respiratory distress syndrome, and hyperinflammation: a retrospective cohort study. *Lancet Rheumatology* 2 (6), e325–e331. doi:10.1016/s2665-9913(20)30127-2

Cavalli, G., Larcher, A., Tomelleri, A., Campochiaro, C., Della-Torre, E., De Luca, G., et al. (2021). Interleukin-1 and interleukin-6 inhibition compared with standard management in patients with COVID-19 and hyperinflammation: a cohort study. *Lancet Rheumatol.* 3 (4), e253–e261. doi:10.1016/s2665-9913(21)00012-6

Chen, Y., Chen, L., Lun, A. T., Baldoni, P. L., and Smyth, G. K. (2025). edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Res.* 53 (2), gkaf018. doi:10.1093/nar/gkaf018

Chu, T., Wang, Z., Pe'er, D., and Danko, C. G. (2022). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. cancer* 3 (4), 505–517. doi:10.1038/s43018-022-00356-3

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44 (8), e71. doi:10.1093/nar/gkv1507

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17, 13–19. doi:10.1186/s13059-016-0881-8

Himberg, J., and Hyvarinen, A. (2003). "Icasso: software for investigating the reliability of ICA estimates by clustering and visualization," in *2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), Toulouse, France*, 259–268. doi:10.1109/NNSP.2003.1318025

Jia, Z., Zhang, X., Guan, N., Bo, X., Barnes, M. R., and Luo, Z. (2015). Gene ranking of RNAseq data via discriminant non-negative matrix factorization. *PloS One* 10 (9), e0137782. doi:10.1371/journal.pone.0137782

Kairov, U., Karpenyuk, T., Ramanculov, E., and Zinovyev, A. (2012). Network analysis of gene lists for finding reproducible prognostic breast cancer gene signatures. *Bioinformation* 8 (16), 773–776. doi:10.6026/97320630008773

Laing, A. G., Lorenc, A., Del Molino Del Barrio, I., Das, A., Fish, M., Monin, L., et al. (2020). A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nat. Med.* 26 (10), 1623–1635. doi:10.1038/s41591-020-1038-6

Lam, M. T. Y., Duttke, S. H., Odish, M. F., Le, H. D., Hansen, E. A., Nguyen, C. T., et al. (2023). Dynamic activity in cis-regulatory elements of leukocytes identifies transcription factor activation and stratifies COVID-19 severity in ICU patients. *Cell Rep. Med.* 4 (2), 100935. doi:10.1016/j.xcrm.2023.100935

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9 (1), 559–13. doi:10.1186/1471-2105-9-559

Li, G., Guo, J., Mou, Y., Luo, Q., Wang, X., Xue, W., et al. (2024). Keratin gene signature expression drives epithelial-mesenchymal transition through enhanced TGF-β signaling pathway activation and correlates with adverse prognosis in lung adenocarcinoma. *Heliyon* 10 (3), e24549. doi:10.1016/j.heliyon.2024.e24549

Li, L., Li, J., Gao, M., Fan, H., Wang, Y., Xu, X., et al. (2021). Interleukin-8 as a biomarker for disease prognosis of coronavirus disease-2019 patients. *Front. Immunol.* 11, 602395. doi:10.3389/fimmu.2020.602395

Liu, L., Xu, L., and Lin, C. (2020). T cell response in patients with COVID-19. *Blood Sci.* 2 (03), 76–78. doi:10.1097/bs9.0000000000000050

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45 (6), 580–585. doi:10.1038/ng.2653

Papayannopoulos, V. (2018). Neutrophil extracellular traps in immunity and disease. *Nat. Rev. Immunol.* 18 (2), 134–147. doi:10.1038/nri.2017.105

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J. P. (2006). GenePattern 2.0. *Nat. Genet.* 38 (5), 500–501. doi:10.1038/ng0506-500

Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., et al. (2019). The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat. Commun.* 10 (1), 5536. doi:10.1038/s41467-019-13483-w

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). "Finding a "kneedle" in a haystack: detecting knee points in system behavior," in *2011 31st international conference on distributed computing systems workshops* (IEEE), 166–171.

Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., et al. (2020). Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* 182 (6), 1419–1440.e23. doi:10.1016/j.cell.2020.08.001

Sprenkeler, E. G., Tool, A. T., Henriet, S. S., van Bruggen, R., and Kuijpers, T. W. (2022). Formation of neutrophil extracellular traps requires actin cytoskeleton rearrangements. *Blood, J. Am. Soc. Hematol.* 139 (21), 3166–3180. doi:10.1182/blood.2021013565

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Taniguchi-Ponciano, K., Vadillo, E., Mayani, H., Gonzalez-Bonilla, C. R., Torres, J., Majluf, A., et al. (2021). Increased expression of hypoxia-induced factor 1α mRNA and its related genes in myeloid blood cells from critically ill COVID-19 patients. *Ann. Med.* 53 (1), 197–207. doi:10.1080/07853890.2020.1858234

Xu, J., He, B., Carver, K., Vanheyningen, D., Parkin, B., Garmire, L. X., et al. (2022b). Heterogeneity of neutrophils and inflammatory responses in patients with COVID-19 and healthy controls. *Front. Immunol.* 13, 970287. doi:10.3389/fimmu.2022.970287

Xu, Q., Yang, Y., Zhang, X., and Cai, J. J. (2022a). Association of pyroptosis and severeness of COVID-19 as revealed by integrated single-cell transcriptome data analysis. *ImmunoInformatics* 6, 100013. doi:10.1016/j.immuno.2022.100013

Yu, X., Zheng, L., Xia, Z., Xu, Y., Shen, X., Huang, Y., et al. (2024). Comprehensive proteomic profiling of lung adenocarcinoma: development and validation of an innovative prognostic model. *Transl. Cancer Res.* 13 (5), 2187–2207. doi:10.21037/tcr-23-1940

Zhao, B. W., Su, X. R., Hu, P. W., Ma, Y. P., Zhou, X., and Hu, L. (2022). A geometric deep learning framework for drug repositioning over heterogeneous information networks. *Briefings Bioinforma.* 23 (6), bbac384. doi:10.1093/bib/bbac384

Zhao, B. W., Su, X. R., Yang, Y., Li, D. X., Li, G. D., Hu, P. W., et al. (2024). A heterogeneous information network learning model with neighborhood-level structural representation for predicting lncRNA-miRNA interactions. *Comput. Struct. Biotechnol. J.* 23, 2924–2933. doi:10.1016/j.csbj.2024.06.032

Zhao, B. W., Su, X. R., Yang, Y., Li, D. X., Li, G. D., Hu, P. W., et al. (2025). Regulation-aware graph learning for drug repositioning over heterogeneous biological network. *Inf. Sci.* 686, 121360. doi:10.1016/j.ins.2024.121360

Zuo, Y., Yalavarthi, S., Shi, H., Gockman, K., Zuo, M., Madison, J. A., et al. (2020). Neutrophil extracellular traps in COVID-19. *JCI insight* 5 (11), e138999. doi:10.1172/jci.insight.138999

Zuo, Y., Zuo, M., Yalavarthi, S., Gockman, K., Madison, J. A., Shi, H., et al. (2021). Neutrophil extracellular traps and thrombosis in COVID-19. *J. Thrombosis Thrombolysis* 51, 446–453. doi:10.1007/s11239-020-02324-z