Bioinformatics, 2025, **41(3)**, btaf114
https://doi.org/10.1093/bioinformatics/btaf114
Advance Access Publication Date: 17 March 2025
Applications Note

OXFORD

Gene expression

mastR: an R package for automated identification of tissue-specific gene signatures in multi-group differential expression analysis

Jinjin Chen[1,2], Ahmed Mohamed[1,2], Dharmesh D. Bhuva[1,2,3], Melissa J. Davis[1,2,3,4,5,*,†], Chin Wee Tan[1,2,4,*,†]

[1]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC 3052, Australia
[2]Department of Medical Biology, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, VIC 3010, Australia
[3]South Australian immunoGENomics Cancer Institute (SAiGENCI), Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, SA 5005, Australia
[4]Frazer Institute, Faculty of Medicine, The University of Queensland, Brisbane, QLD 4102, Australia
[5]Department of Clinical Pathology, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, VIC 3010, Australia

*Corresponding authors. Chin Wee Tan, Walter and Eliza Hall Institute, 1G Royal Parade, Victoria 3052, Australia. Email: cwtan@wehi.edu.au; Melissa Davis, Walter and Eliza Hall Institute, 1G Royal Parade, Victoria 3052, Australia. Email: melissajdavis@isomorphiclabs.com

†equal contribution, co-senior authors.

Associate Editor: Can Alkan

Abstract

Motivation: Biomarker discovery is important and offers insight into potential underlying mechanisms of disease. While existing biomarker identification methods primarily focus on single cell RNA sequencing (scRNA-seq) data, there remains a need for automated methods designed for labeled bulk RNA-seq data from sorted cell populations or experiments. Current methods require curation of results or statistical thresholds and may not account for tissue background expression. Here we bridge these limitations with an automated marker identification method for labeled bulk RNA-seq data that explicitly considers background expressions.

Results: We developed *mastR*, a novel tool for accurate marker identification using transcriptomic data. It leverages robust statistical pipelines like *edgeR* and *limma* to perform pairwise comparisons between groups, and aggregates results using rank-product-based permutation test. A signal-to-noise ratio approach is implemented to minimize background signals. We assessed the performance of *mastR*-derived NK cell signatures against published curated signatures and found that the *mastR*-derived signature performs as well, if not better than the published signatures. We further demonstrated the utility of *mastR* on simulated scRNA-seq data and in comparison with *Seurat* in terms of marker selection performance.

Availability and implementation: *mastR* is freely available from https://bioconductor.org/packages/release/bioc/html/mastR.html. A vignette and guide are available at https://davislaboratory.github.io/mastR. All statistical analyses were carried out using R (version ≥4.3.0) and Bioconductor (version ≥3.17).

1 Introduction

Biomarkers are biological features that infer the states of cells, tissues, or individuals, either diseased or healthy. Biomarkers may include molecular features like genes, and proteins which can be used in research and clinical settings to provide insights into disease diagnosis, prognosis, and treatment. In recent years, biomarkers have been identified through various -omics approaches, including transcriptomics, proteomics, and metabolomics, providing an improved overview of the molecular landscape of the system being studied (Lawlor *et al.* 2009, Wang and Yu 2013, Rodrigues *et al.* 2016). This influx of omics data has advanced the development of computational and bioinformatics methods to identify biomarkers, providing opportunities to accelerate biomarker discovery and thereby facilitating diagnostic and therapeutic developments for various diseases and cancers (Kaur *et al.* 2021, Lee and Kim 2021, Vlachavas *et al.* 2021). However, separating the background signal of the tissue microenvironment from the target marker's signal remains a complex problem.

While many recent marker identification methods focus on emerging data types such as single-cell RNA sequencing (scRNA-seq) or spatial omics data, where disease signals are more easily separated from background, bulk RNA-seq remains a valuable resource due to its widespread availability, cost and already established pool of datasets. Existing tools for bulk RNA-seq marker identification, such as *edgeR* (Robinson *et al.* 2010), *limma* (Ritchie *et al.* 2015), or *DESeq2* (Love *et al.* 2014), require manual curation of differential expression (DE) results, a process that is labor-intensive and prone to inconsistencies or biases. In the context of marker identification, the package *MarkerPen* (Qiu *et al.* 2021) made some progress toward automated marker identification. However, it is semi-supervised and relies on predefined marker lists, thereby limiting their applicability to new datasets.

Received: 22 October 2024; **Revised:** 7 March 2025; **Editorial Decision:** 7 March 2025; **Accepted:** 13 March 2025

Although some scRNA-seq methods can technically be applied to labeled bulk RNA-seq data, due to the small sample sizes in bulk RNA-seq data, machine learning-based approaches are generally unsuitable. While for DE-based scRNA-seq methods, studies indicate they generally do not perform as well as DE-based methods specifically designed for bulk RNA-seq data (Squair *et al.* 2021, Heumos *et al.* 2023), underscoring the need for more tailored approaches. Moreover, many of the current tools' workflows often fail to fully utilize the statistical information (e.g. *P* value or fold change) across multiple comparisons or datasets. These tools apply a direct intersection of the results, and therefore did not account for the effects of tissue-specific background expression, leading to nonspecific marker identification.

To address these challenges, we developed an R/Bioconductor package *mastR* (*M*arkers *A*utomated *S*creening *T*ool in *R*), offering integration of the following key features as a comprehensive framework: (i) an automated workflow that integrates statistical information across multiple DE comparisons and datasets, (ii) explicit consideration of tissue-specific background expression to enhance marker specificity.

mastR builds upon established DE analysis tools [*edgeR* (Robinson *et al.* 2010) and *limma* (Ritchie *et al.* 2015)] by implementing a rank-product-based scoring approach to integrate statistical information across multiple DE comparisons. This approach is particularly effective in scenarios where standard one-vs-all comparisons face limitations, such as when the target group shares high similarity with specific subgroup(s). While standard workflows might not identify markers that distinguish the target group from dissimilar groups due to the dominance of the similar subgroup in the analysis, *mastR*'s rank-product scoring method assigns balanced weights across all group comparisons. This design choice helps mitigate bias introduced by group size differences and enables the identification of markers that might be overlooked in standard approaches. Furthermore, *mastR* incorporates tissue-specific background expression through a signal-to-noise ratio (SNR) metric implemented in its marker selection algorithm, enhancing the specificity and reliability of identified markers. Through validation on both simulated and public datasets, we demonstrate that *mastR* achieves high accuracy and computational efficiency while maintaining robustness across diverse experimental contexts. These features make *mastR* valuable for both research applications and clinical marker identification where tissue context and complex group relationships need to be considered.

2 Methods

mastR's workflow involves four steps (Fig. 1): (i) build a markers pool; (ii) identify the signature of the target group; (iii) refine the signature by removing the background signal of the sample microenvironment; and (iv) visualizing the resulting signature.

2.1 Build a marker pool

The *mastR* pipeline begins by generating a pool of candidate markers. This pool can be compiled either using the functions in *mastR* or by custom curation and selection of marker genes from databases or publications. For the former, the R/Bioconductor package *mastR* allows extraction of marker genes specific to immune cell types, relevant pathways, and/or gene sets from existing data sources, which can be retrieved via *get_lm_sig/get_panglao_sig/get_gsc_sig* functions in *mastR*. This includes leukocyte gene signature matrices from CIBERSORT [LM7 (Tosolini *et al.* 2017) and LM22 (Newman *et al.* 2015), immune cell signature matrices defining 7 and 22 immune cell types, respectively], PanglaoDB (scRNA-seq experiments from mouse and human that includes marker genes for 25 different immune cell types.) (Franzen *et al.* 2019), and MSigDB (Molecular Signatures Database, a collection of annotated gene sets) (Subramanian *et al.* 2005, Liberzon *et al.* 2015), respectively.

mastR provides *gsc_plot* function to help visualize the overlap of sets of markers. These sets of marker genes can be seamlessly merged as the original pool of markers using the *merge_markers* function in *mastR*, with all marker gene sources saved in the *longDescription* attribute. This merged pool will be used in subsequent analyses. When the markers pool is used in the downstream filtering step, all the marker genes in the pool will be preserved as these are determined to be of biological significance.

2.2 Identify signature of the target group

To identify group-specific signatures, *mastR* uses three main steps: (i) differential expression analysis (DEA), (ii) feature selection to select highly differentially expressed genes (DEGs) based on their rank-product score; and (iii) constraining selected genes within the markers pool (Fig. 1B).

Firstly, DEA is performed using *edgeR* (Robinson *et al.* 2010) and *limma* (Ritchie *et al.* 2015) workflow (i.e. filtering, normalizing, sample weighting and linear modeling). Given the "Group" and "Batch" factors in the data, *mastR* automatically generates the appropriate design matrix to be used during data filtering, normalization, and batch effect correction. Here, batch factor is used as the fixed effect in linear modeling as it was found that the use of batch-corrected data rarely improves the analysis of sparse data, whereas batch covariate modeling improves the analysis for substantial batch effects (Nguyen *et al.* 2023). *mastR* allows either raw counts or log-normalized data as input with different processing pipelines conducted on different types of input. Raw count data is filtered by the *filterByExpr* function in *edgeR*, normalized using the trimmed mean of M-values (TMM) method and analyzed using the "*limma voom with treat*" pipeline. For log-normalized data, genes are filtered by user-defined thresholds and "*limma trend with treat*" method is used. In most cases in this study, the log-fold-change (logFC) equal to 0 was used to perform DE analysis, with the only exception being when generating NK cell signature using DICE [Database of Immune Cell Expression, Expression quantitative trait loci (eQTLs) and Epigenomics] project (Schmiedel *et al.* 2018) in which case a logFC of 1.5 was used.

Secondly, feature selection is conducted to select genes specific to the target group across multiple comparisons. The probability $score_g$ is computed by comparing the rank product (RP) score RP_g with permutated random score *rp* from bootstrap approach [Equations (1)–(3)]. The common DE genes from $n-1$ comparisons (where n is the total number of groups) are identified and ranked based on the given gene statistics (e.g. *P* value, adjusted *P*-value or log fold change) for each comparison. The ranks for each marker gene across all comparisons are log-transformed and summed, before a permutation test was applied to bootstrap the null distribution of the random RP. The resulting marker genes are ordered by

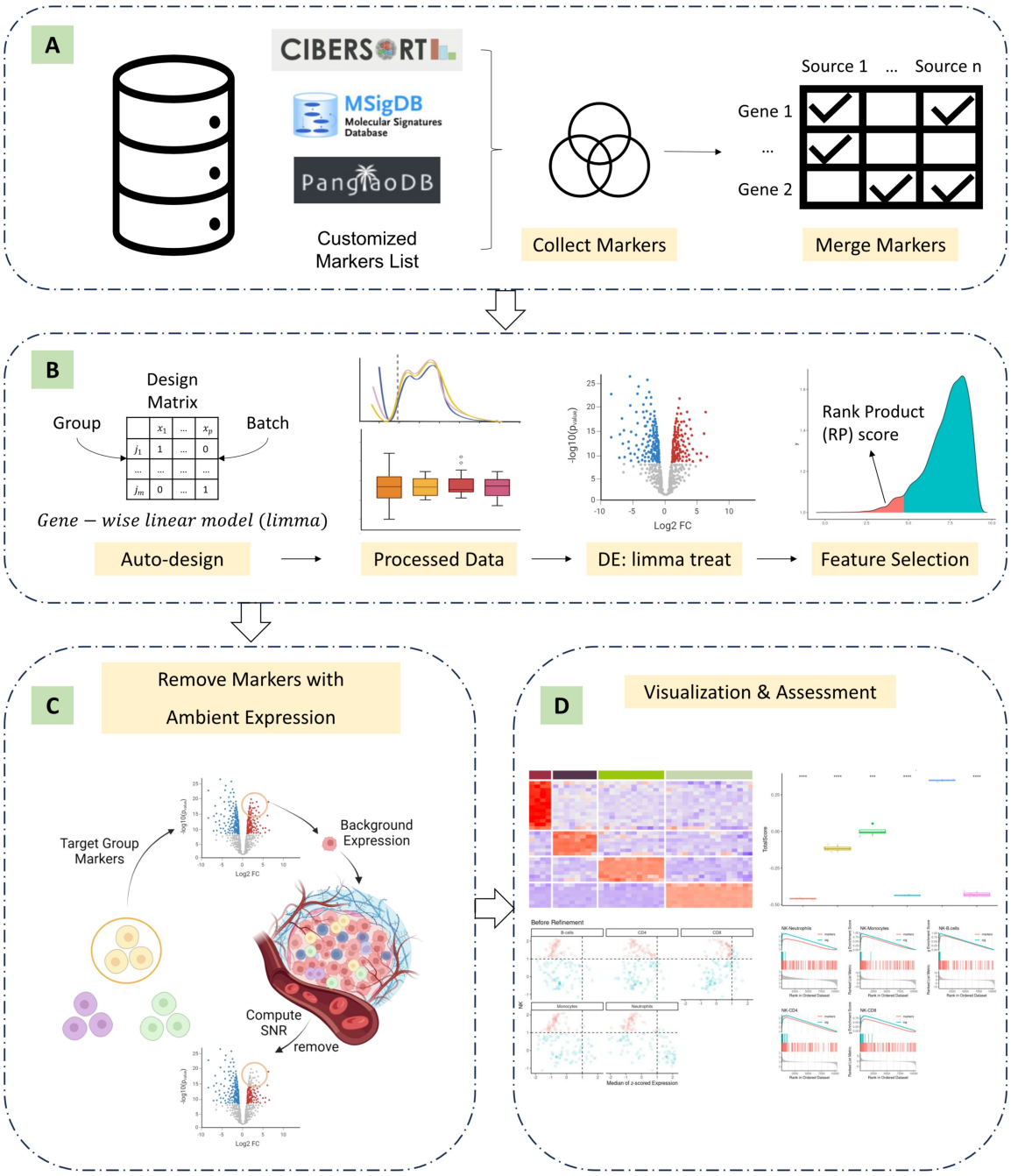

Figure 1. Schematic of the *mastR* workflow. The workflow of *mastR* can be divided into four main sections: (A) build markers pool; (B) identify signature markers; (C) refine signature by filtering based on background expression and (D) visualize and access signature performance. The *mastR* workflow recommends integrating markers from multiple sources (e.g. PanglaoDB, MSigDB) to form an initial set of markers. *mastR* then generates a design matrix based on the given "Group" and "Batch" factors to be used during data processing and DE analysis. The data processing includes an *edgeR* data filtering and normalization pipeline, and a *limma-voom-treat* based linear modeling DE approach to compare the target group with all other groups. *mastR* then computes the marker's RP score based on the ranked product across the DE comparisons and bootstrapped permutation null distribution for further feature selection across multiple comparisons. The selected features will be constrained by their intersection with the initial set of markers. *mastR* allows for filtering of genes based on the SNR with a background dataset to remove features with inherent expression in a specific context or disease. *mastR* then provides visualization functions to assess the performance of the signature.

$score_g$ (with the smaller the values being more significant) and filtered by a selected threshold (default as 0.05).

$$RP_g = \sum_{i=1}^{n-1} \ln(rank_{g,i}) \quad (1)$$

where $rank_{g,i}$ is the rank of gene g in ith list,

$$rp = \left\{ \sum_{i=1}^{n-1} \ln(rrank_{g,i}) \right\} \quad (2)$$

where $rrank_{g,i}$ is the shuffled rank of gene g in ith list,

$$score_g = P(RP_g > rp) \quad (3)$$

The computation can be summarized as below:

1) Rank the gene statistics in increasing order (decreasing order of ||logFC|| when statistics is logFC) ⇒ $rank_{g,i}$: rank of *g*th gene under *i*th comparison;
2) Sum log-rank for each gene across comparisons as RP_g: RP of *g*th gene;
3) Independently permute statistics value within each comparison relative to gene ID, repeat step (1)–(2) ⇒ $rp_g^{(k)}$: random RP of gth gene;
4) Repeat step (3) *K* times, form reference null distribution with $rp_g^{(k)}(k = 1, 2, \ldots, K)$;
5) Determine the probability associated with each gene ⇒ $score_g$.

In some marker identification studies, the presence of two or more closely related groups in the data pose challenges for the identified markers to be effective in distinguishable these groups (Burel *et al.* 2022). To accommodate this situation, threshold filtering based on RP can be omitted for the target comparison(s) in question by setting parameters "keep.top" and "keep.group," allowing for a greater number of DEGs in the targeted comparison(s).

Thirdly, the identified marker genes are limited to those in the markers pool (i.e. common genes are retained) as the resulting signature. This refinement approach enhances both the discriminative power and the precision of the resulting signature when there is prior knowledge. When the input involves multiple datasets, *mastR* aggregates the individual signature lists identified by each dataset using either a "Robust Rank Aggregation (RRA)," "union," or "intersect."

The aggregation method "RRA" detects marker genes that are consistently ranked higher than stochastically expected under the null hypothesis of uncorrelated inputs and assigns a significance score to each gene (Kolde *et al.* 2012). It is recommended for robust gene selection from large numbers of DEGs. The "union" method is recommended for small numbers of marker genes identified per dataset; and the "intersect" method, is best used in situations characterized by high levels of marker intersection.

mastR provides a series of step-by-step functions as well as an integrated wrapper function to implement the above analyses.

2.3 Refine signature by accounting for background expression

To avoid background microenvironment confounding effects, *mastR* can further refine the marker genes by filtering out genes with ubiquitous expression. *mastR* utilizes an approach which remove genes with low "SNR" based on Cohen's d (Knapp 1990), which have limited discriminative power between the group of interest and the "background" or "environment" [Equations (4) and (5)]. Considering situations where the background and signal expressions do not originate from the same batch, and that re-normalizing the entire data is time-consuming, in order to make the sample microenvironment and signal data comparable between batches, the relative expression of the genes within the samples are used to compute SNRs, making "signal" and "noise" comparable across datasets.

For DE analysis, we assume genes are not differentially expressed (null hypothesis) and the gene expression within each sample should follow a normal distribution denoted as $X \sim N(\mu, \sigma^2)$. The parameters, mean (μ) and standard deviation (σ), can be estimated through maximum likelihood estimation (MLE). The percentile (accumulated density) for each gene in each sample can then be obtained using the Gaussian cumulative distribution function (CDF) $F(x|\mu, \sigma^2)$, and the SNR computed as outlined in Equation (5).

$$\hat{x}_S = F^{-1}(0.5|\mu_S, {\sigma_s}^2), \ \hat{x}_B = F^{-1}(0.5|\mu_B, {\sigma_B}^2) \quad (4)$$

$$snr = \frac{\hat{x}_S - \hat{x}_B}{\sigma_B} \quad (5)$$

where F^{-1} is the inverse CDF function, $\hat{x}_S$ represents the 50th percentile (median) of a normal distribution fitted to the observed log-transformed gene expression in the signal dataset *S*, $\hat{x}_B$ represents the 50th percentile (median) of a normal distribution fitted to the observed log-transformed gene expression in the background dataset *B*, μ is the mean of the normal distribution, σ is the standard deviation of the normal distribution, *snr* is the SNR of each gene.

This crucial step removes the effect of sample purity for the identified signature markers. By excluding the marker genes with similar expression in the sample microenvironment, the SNR approach ensures only the marker genes with robust and specific expression patterns in the group of interest are retained, leading to a more refined and accurate signature marker list.

3 Results

In this study, we evaluated *mastR*'s performance on both simulated and biological datasets, with *mastR* exhibiting high accuracy and robustness (Supplementary Figs S1–S5 and Supplementary Tables S3 and S4). Briefly, a natural killer (NK) cell specific signature from DICE dataset was identified using *mastR* (Supplementary Fig. S1) and validated in an independent immune cell dataset (Supplementary Fig. S4), showing high specificity for NK cells. The resulting performance metric on the simulated data (Supplementary Table S3) suggest *mastR* is highly accurate, have low false discovery rates and computationally fast.

We then compared the performance of the *mastR*-derived NK cells signature with existing published NK signatures with *mastR* demonstrating comparable, if not better performance in identifying NK cells (Supplementary Fig. S6). Assessing the average expression of the unique markers for each signature across the cell types, *mastR* is able to identify novel and highly specific marker genes for NK cells (Supplementary Fig. S7).

While the focus of this study was to evaluate *mastR* for marker identification in bulk RNA-seq data, we also looked at the potential application of *mastR* for scRNA-seq datasets. Here we compared *mastR*'s performance with Seurat (Hao *et al.* 2021), one of the statistical packages designed for scRNA-seq data. Interestingly, *mastR* performed better than Seurat and requires significantly lower computation time (Supplementary Table S4).

Till this end, we have assessed *mastR* for both bulk and scRNA-seq data, however it can theoretically be applied to all multi-omics data. Moving forward, the aim for this work is to validate the performance of *mastR* using experimental data across diverse omics types to improve application and generalizability across a range of research contexts.

Acknowledgements

We would like to thank Ashley Weir for reviewing and providing valuable feedback for this manuscript.

Author contributions

Jinjin Chen (Conceptualization [equal], Formal analysis [lead], Methodology [lead], Project administration [equal], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), Ahmed Mohamed (Investigation [supporting], Software [supporting]), Dharmesh D. Bhuva (Formal analysis [supporting], Visualization [supporting], Writing—review & editing [supporting]), Melissa J. Davis (Conceptualization [equal], Funding acquisition [lead], Methodology [supporting], Project administration [supporting], Supervision [equal], Writing—original draft [supporting]), and Chin Wee Tan (Conceptualization [supporting], Methodology [supporting], Project administration [lead], Supervision [lead], Visualization [supporting], Writing—original draft [supporting], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

J.C. was supported by Melbourne Research Scholarships (MRS) from the University of Melbourne. C.W.T. was supported by the Australian Academy of Sciences (AAS): Regional Collaborations Programme COVID-19 Digital Grants scheme. D.D.B. and M.J.D. were supported by the Grant-in-Aid Scheme administered by Cancer Council Victoria and by a research grant from the Australian Lions Childhood Cancer Foundation. M.J.D. was funded by the Betty Smyth Centenary Fellowship in Bioinformatics, the Cure Brain Cancer Foundation and National Breast Cancer Foundation joint grant CBCNBCF-19-009, and the National Health and Medical Research Council grant APP2021286. WEHI acknowledges the support of the Operational Infrastructure Program of the Victorian Government. The South Australian immunoGENomics Cancer Institute (SAiGENCI) has received grant funding from the Australian Government.

Data availability

The source code for *mastR* is freely available from the Bioconductor website at https://bioconductor.org/packages/release/bioc/html/mastR.html. Datasets are freely available for download from the following public data repositories and URLs. DICE (Database of Immune Cell Expression, Expression quantitative trait loci (eQTLs) and Epigenomics) at https://dice-database.org; TCGA-COAD at https://portal.gdc.cancer.gov/projects/TCGA-COAD; CCLE (Cancer Cell Line Data Repository) at https://sites.broadinstitute.org/ccle; im_data_6 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi and can be accessed with GSE60424; pbmc3k.final at https://github.com/satijalab/seurat-data; NK Crinier at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6269138/; NK Cursons at https://aacrjournals.org/cancerimmunolres/article/7/7/1162/469488/A-Gene-Signature-Predicting-Natural-Killer-Cell and NK Shembrey at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9853446.

References

Burel JG, Chawla A, Greenbaum JA *et al.* Distinguishing cell–cell complexes from dual lineage cells using single-cell transcriptomics is not trivial. *Cytom Part A* 2022;**101**:547–51.

Franzen O, Gan LM, Bjorkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019;**2019**:baz046.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87 e3529.

Heumos L, Schaar AC, Lance C *et al.*; Single-cell Best Practices Consortium. Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023;**24**:550–72.

Kaur H, Kumar R, Lathwal A *et al.* Computational resources for identification of cancer biomarkers from omics data. *Brief Funct Genomics* 2021;**20**:213–22.

Knapp TR. Statistical power analysis for the behavioral-sciences, 2nd edition—Cohen, J. *Educ Psychol Meas* 1990;**50**:225–7.

Kolde R, Laur S, Adler P *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;**28**:573–80.

Lawlor K, Nazarian A, Lacomis L *et al.* Pathway-based biomarker search by high-throughput proteomics profiling of secretomes. *J Proteome Res* 2009;**8**:1489–503.

Lee SM, Kim HU. Development of computational models using omics data for the identification of effective cancer metabolic biomarkers. *Mol Omics* 2021;**17**:881–93.

Liberzon A, Birger C, Thorvaldsdóttir H *et al.* The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;**1**:417–25.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.

Newman AM, Liu CL, Green MR *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;**12**:453–7.

Nguyen HCT, Baik B, Yoon S *et al.* Benchmarking integration of single-cell differential expression. *Nat Commun* 2023;**14**:1570.

Qiu Y, Wang J, Lei J *et al.* Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. *Bioinformatics* 2021;**37**:3228–34.

Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.

Rodrigues D, Jerónimo C, Henrique R *et al.* Biomarkers in bladder cancer: a metabolomic approach using in vitro and ex vivo model systems. *Int J Cancer* 2016;**139**:256–68.

Schmiedel BJ, Singh D, Madrigal A *et al.* Impact of genetic polymorphisms on human immune cell gene expression. *Cell* 2018;**175**:1701–15 e1716.

Squair JW, Gautier M, Kathe C *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 2021;**12**:5692.

Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.

Tosolini M, Pont F, Poupot M *et al.* Assessment of tumor-infiltrating TCRVgamma9Vdelta2 gammadelta lymphocyte abundance by deconvolution of human cancers microarrays. *Oncoimmunology* 2017;**6**:e1284723.

Vlachavas EI *et al.* A detailed catalogue of multi-omics methodologies for identification of putative biomarkers and causal molecular networks in translational cancer research. *Int J Mol Sci* 2021;**22**:2822.

Wang J, Yu G. A systems biology approach to characterize biomarkers for blood stasis syndrome of unstable angina patients by integrating MicroRNA and messenger RNA expression profiling. *Evid Based Complement Alternat Med* 2013;**2013**:510208.

Bioinformatics, 2025, 41, 1–5
https://doi.org/10.1093/bioinformatics/btaf114
Applications Note