

聂斌^{1,2}, 李雪玲², 陈樱³

(1.安徽建筑大学电子与信息工程学院,安徽 合肥 230000;

2.中国科学院合肥物质科学研究院健康与医学技术研究所,安徽 合肥 230000;

3.安徽医科大学第一附属医院外科,安徽 合肥 230000)

摘要:从 bulk 转录组学和 DNA 甲基化等数据中推断单个细胞类型比例及基因表达的方法,称为细胞类型反卷积方法。细胞类型反卷积方法在细胞异质性和肿瘤免疫微环境研究中发挥了重要的作用,然而各类方法在不同数据上的表现并不稳定。为了克服单个反卷积方法结果的偏倚和功能限制,本文基于 R-shiny 框架构建了一个从混合组织中推测细胞比例及基因表达研究的在线应用(www.deconvolution.cn),用于转录组数据和 DNA 甲基化数据的细胞比例分析、基因表达、富集分析及单细胞数据处理等功能。将工具应用在 GSE119409 直肠癌数据上,得到了新辅助放疗应答与细胞比例和细胞水平基因表达的关系。本应用提供了一个细胞类型反卷积研究的统一交互平台,对研究癌症预后和治疗应答机制都具有重要的意义。

关键词: R-shiny; 细胞类型反卷积; 新辅助放疗; 基因表达

中图分类号: R197

文献标识码: B

DOI: 10.3969/j.issn.1006-1959.2025.10.009

文章编号: 1006-1959(2025)10-0056-06

Construction of an Online Tool for Inferring Cellular Proportions and Gene Expression from Heterogeneous Tissues

NIE Bin^{1,2}, LI Xueling², CHEN Ying³

(1.School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230000, Anhui, China;

2.Institute of Health and Medical Technology, Hefei Institute of Physical Sciences, Chinese Academy of Sciences, Hefei 230000, Anhui, China;

3.Department of Surgery, the First Affiliated Hospital of Anhui Medical University, Hefei 230000, Anhui, China)

Abstract: Methods for inferring the proportion of individual cell types and gene expression from bulk transcriptomic data or DNA methylation are collectively referred to as cell type deconvolution methods. Cell type deconvolution has played an important role in the study of cell heterogeneity and tumor immune microenvironment, but the performance of various methods on different data is not stable. In order to overcome the results bias and functional limitations of a single deconvolution method, an online application for prediction of cell proportion and gene expression from heterogeneous tissues based on the R-shiny framework is constructed (www.deconvolution.cn), for transcriptome data and DNA methylation data cell proportion analysis, gene expression, enrichment analysis and single cell data processing functions. The tool was applied to GSE119409 rectal cancer data, and the relationship between neoadjuvant radiotherapy response and cell proportion and cellular level gene expression was obtained. This application provides a unified interactive platform for the study of cell type deconvolution, which is of great significance for the study of cancer prognosis and therapeutic response mechanism.

Key words: R-shiny; Cell type deconvolution; Neoadjuvant radiotherapy; Gene expression

生物组织,尤其是肿瘤组织,通常由多种不同类型的细胞组成,这些细胞在形态、功能和基因表达上存在显著差异^[1,2]。在肿瘤免疫微环境中,相同基因在不同种类的细胞中的基因表达并不完全相同^[3,4]。由于异质混合物的 bulk 样本仅代表平均表达水平,而不是此类混合物中存在的不同细胞类型中每个基因的单次测量,因此许多相关分析(例如差异基因表达)通常被细胞类型比例的差异所混淆^[5,6]。而荧光激

活细胞分选(FACS)、流式细胞术(FCM)、单细胞测序(scRNA-seq)等生物技术具有多种局限性,从而限制了其大规模应用^[7,8]。由于这些原因,在过去十多年中,已经开发了许多从 bulk 转录组学数据或 DNA 甲基化或空间转录组学中推断单个细胞类型比例的方法,以及使用单细胞 RNA 测序(scRNA-seq)数据推断批量 RNA 测序或 DNA 甲基化样本中细胞比例的新方法,这些方法统称为细胞类型反卷积方法^[9]。这些技术通过数学模型和算法,从 bulk 转录组数据中推断出细胞类型的丰度和基因表达,从而在不进行单细胞分离的情况下,提供了一种研究细胞异质性的方法。尽管细胞类型反卷积计算方法层出不穷,但是每种反卷积方法都有自己的优势和

基金项目:安徽省医学重点专项(编号:202304295107020037)

作者简介:聂斌(1993.8-),男,安徽淮南人,硕士研究生,主要从事生物信息学研究

通讯作者:李雪玲(1976.8-),女,山东济南人,博士,研究员,主要从事生物信息学研究

局限性,使用单个方法可能会获得有偏差的反卷积结果,虽然很多反卷积方法解决了影响反卷积结果的不同因素,但一次只关注一个或两个单独的方面^[10,11],面对各有特点的反卷积方法,如何在不同数据集上选择最优的方法,是一个具有挑战的问题。为了克服单个反卷积方法结果的偏倚和功能限制,获得更准确的细胞构成比例及细胞水平的基因表达,并且更加方便生物或医学研究者的日常使用,本文构建了一个细胞异质性研究的在线应用。用户无需处理复杂的输入数据格式及代码实现过程,即可得到最精确结果,进行可视化分析或下载数据进行其他感兴趣的研究,帮助生物或医学研究者快速得到可验证的生物学假说,供下游生物实验验证。

1 在线应用的功能体系及架构搭建

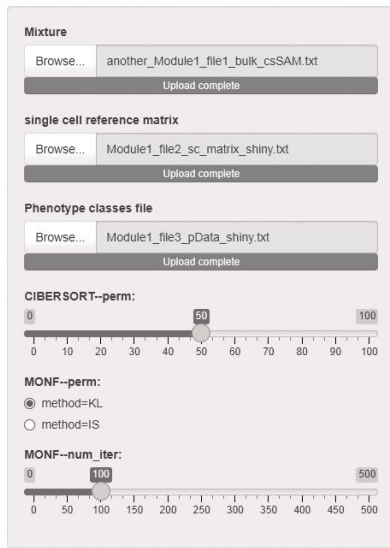
1.1 功能体系 本应用包含五大模块:RNA-seq 反卷积、DNA 甲基化反卷积、计算细胞水平基因表达谱、富集分析可视化和单细胞数据处理模块。功能模块一和二为基础功能,提供更精确的细胞分数;模块三和模块四基于模块一,扩展了基础反卷积功能;模块五作为模块一的辅助,提供更简便的方法处理单细胞数据作为反卷积的参考。RNA-seq 反卷积模块(图 1):此模块输入部分用户可以上传自己的 bulk 表达矩阵和单细胞参考矩阵,也可以选择提供的示例数据,并设置部分方法的参数(或者默认),输出部分包含参考矩阵和 7 种方法结果(Music^[12]、NNLS^[13]、SCDC^[14]、CIBERSORT^[15]、MONF^[16]、DWLS^[17]、Ensemble),以及推测出的最佳方法名称。DNA 甲基化反卷积模块:需要用户上传 DNA 甲基化数据及参考矩阵,方便的是工具提供了多种组织类型的参考标签,用户可直接选择,输出部分是三种 DNA 甲基化反卷积方法的结果(PRMeth^[18]、MethylResolver^[19]、Methyl-CIBERSORT^[20]),以及集成结果和推测的最优方法。计算细胞类型级别基因表达谱:此模块基于模块一结果,且需另外上传模块一中 Mixture 文件对应的样本分组信息。通过细胞类型特异性意义分析(csSAM)方法,鉴定出不同细胞类型的分组差异表达基因和基因表达谱。用户可以设置 T 检验 P 值和假阳性率(FDR)的阈值, P 值默认 0.05。富集分析可视化模块(图 2):基于模块三的单细胞类型差异基因,进行 GO(Gene Ontology)和 KEGG(Kyoto Encyclopedia of Genes and Genomes)富集分析。反卷积结果中的每种细胞类型都会生成多组富集分析结果,完整矢量图结果可以下载。单细胞数据处理模块(图 3):此模块可处理 10X 和 rds 两种常用单细胞

数据格式,自动整理,标准化,降维,聚类,采用 SingeR^[21]作为细胞注释方法,用户可以设置聚类分群的 resolution 和细胞注释参考数据集,提供可视化及完整数据下载。

1.2 架构搭建 本系统采用 B/S 架构。相比 C/S 架构,B/S 架构连接互联网后,通过浏览器即可访问系统,用户操作简便,也便于升级维护。开发语言使用 R-shiny 作为基础架构,运用 bootstrap 对其界面与功能进行完善^[22]。Shiny 是基于 R 语言的交互式 web 框架,具有成熟的机器学习和统计模型解决方案,便于集成多种基础反卷积方法。分别搭建应用的用户界面(UI)和后台功能(Server),相应代码封装并部署在服务器上。UI 包括设置 navbarPage、tabPanel、sidebarPanel、mainPanel、tabsetPanel 五层结构,提供了上传下载,方法的参数选择等交互式按钮。通过引入 bootstrap 对界面布局进行美化,提供清晰的操作展示界面。Server 端负责所有数据的计算处理。其中集成六种基础反卷积方法的核心算法思想如下公式所示:

$$\min_{F, \{\omega_m\}} \sum_{m=1}^M \omega_m \|F - F^{(m)}\|_2 + \lambda \sum_{m=1}^M \omega_m \log \omega_m$$
$$\text{subject to } \sum_{m=1}^M \omega_m = 1, \omega_m \geq 0.$$

以 RNA 测序数据为例,运行 6 个基础反卷积方法后,可以得到 6 个反卷积结果,其中 N 表示 bulk 样本数, K 表示细胞类型数, m 表示第 m 个基本反卷积方法。通过整合 6 种基础反卷积方法产生的结果来集成细胞类型反卷积结果 F ,并预测基础反卷积方法所占权重,权重最大的为最佳结果。因单细胞参考数据读取需占用较大内存,1GB rds 类型单细胞文件读取到 RAM 需占用约 16G 内存,因此服务器提供了 128G 内存,计算结束后自动释放单细胞对象。为了加快工具的响应速度,使用了 reactive 表达式控制程序是否需要更新运算^[23]。通过 parallel 包实现部分异步处理请求,提高服务器的并发处理能力。为了解决集中式 App 框架的内容承载量有限问题,将重复的 UI 和 Server 封装成模块(图 4),通过 source 引入,也使得程序便于移植和开发。另外,将不同反卷积方法封装成对应的函数,自动处理不同方法需要的数据类型自动构建多种对象,例如 SCDC 需要的 Expressionset 对象,Music 需要的 SingleCellExperiment 对象;服务器使用 Nginx 实现反向代理,提高应用的性能、可靠性和安全性。具体配置见表 1。



Signature Matrix

The best method: NNLS Music SCDC CIBERSORT MONF DWLS Ensemble

Output of slider:

Show 25 entries

Epithelial cells	Stromal cells	Myeloids	T cells	B cells
0.389997669113931	3.163314427257439	1.473324527763739	9.48994328177232	11.95992851949388
4.896637401097133	0	0	0	0
0.5633299664979003	66.03960530329229	3.899976691139309	1.169993007341793	1.516657602109731
0.2166653717299616	5.546633516287018	0.04333307434599232	0.04333307434599232	0.4333307434599232
4.073308988523279	0.7366622638818694	0.3466645947679386	1.386658379071754	2.03665449426164
4.636638955021178	0.9533276356118311	0.129999223037977	1.256659156033777	1.516657602109731
5.069969698481102	3.596645170717363	1.29999223037977	9.359944058734342	6.976624969704764
5.849965036708964	0.6933291895358772	0.129999223037977	1.949988345569655	1.213326081687785
7.366622638818694	2.8166498324895	1.1266599329958	32.32647346211028	1.776656048185685
22.87996325468394	5.719965813670986	0.1733322973839693	1.949988345569655	1.559990676455724
32.80313727991619	5.676632739324994	0.9099945612658389	7.236623415780719	5.97996425974694
7.756620307932625	1.1266599329958	0.04333307434599232	1.603323750801716	1.646656825147708
10.22660554565419	1.646656825147708	0.4766638178059156	1.343325304725762	1.993321419915647

图 1 RNA-seq 反卷积运行结果

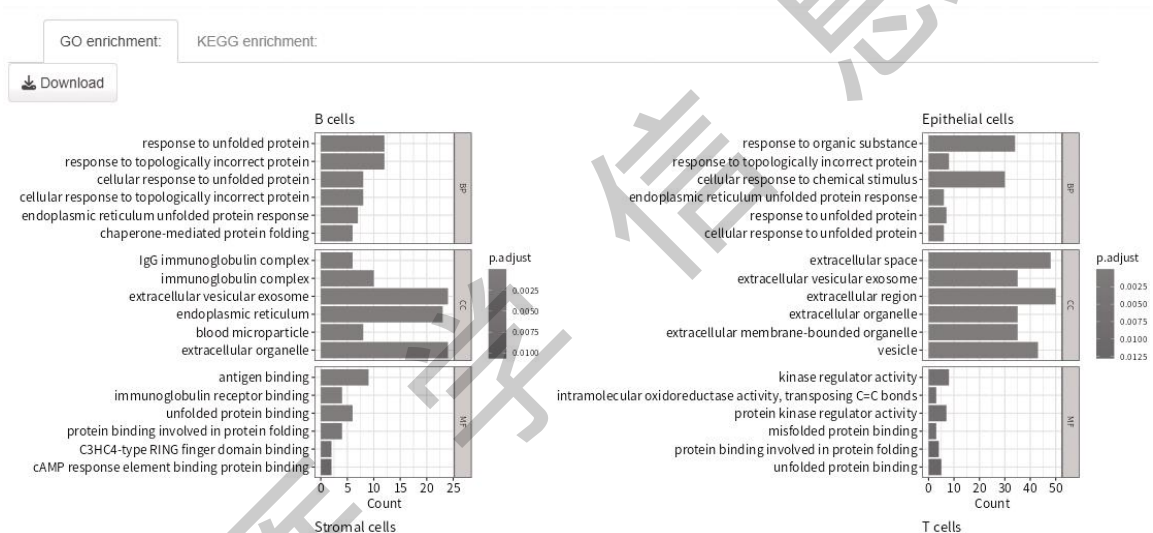


图 2 不同细胞类型富集分析结果

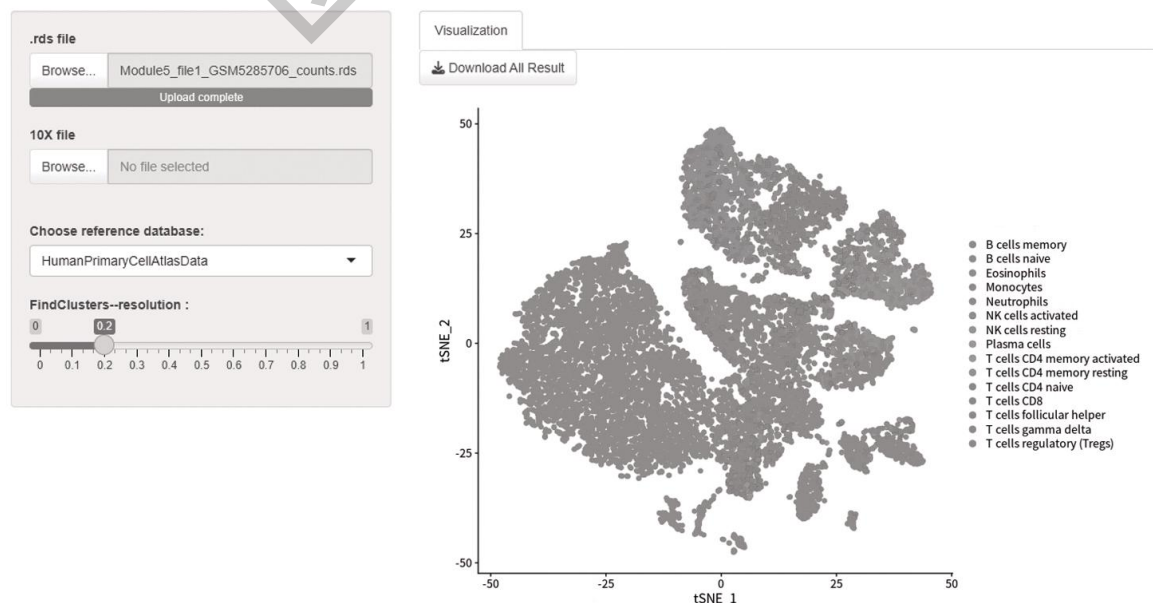


图 3 单细胞数据集分析结果

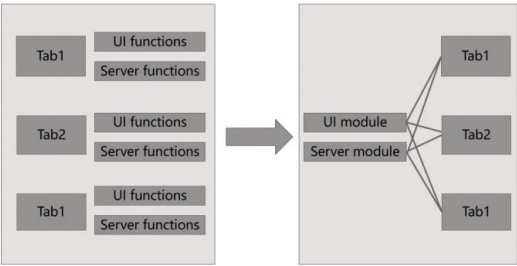


图 4 将 UI 和 Server 模块化图示

表 1 服务器环境配置

环境	配置
CPU	Intel(R) Xeon(R) Gold 6132
RAM	128G
操作系统	Ubuntu18.04.6
R	4.3
shiny	1.7.5
bootstrap	3.4

2 在 GSE119409 数据上的应用

2.1 数据来源 从 GEO 下载基于 Affymetrix Human Genome U133 Plus 2.0 Array(GPL570)的直肠癌微阵列数据集 GSE119409，由 66 个局部晚期直肠癌患者的活检样本组成，活检样本均在患者进行新辅助放疗(neoadjuvant Radiotherapy, nRT)和手术切除之前取得，随后患者接受中等剂量的新辅助放疗 (30 Gy/10 次)，并进行全直肠系膜切除术 (Total Mesorectal Excision, TME),在剔除未知反应样本后，筛选出了 41 例对新辅助放疗无应答的样本和 15 例应答的样本。

2.2 应用效果 将 GSE119409 数据集按照放疗应答与非应答分组，使用模块一的细胞分数结果比较两组病人免疫细胞比例差异，不同亚型的肿瘤样本表现出明显的免疫细胞浸润模式(图 5、图 6)。NK 细胞和 T 细胞中的 CD4 细胞获得了显著的差异($P<0.05$)。

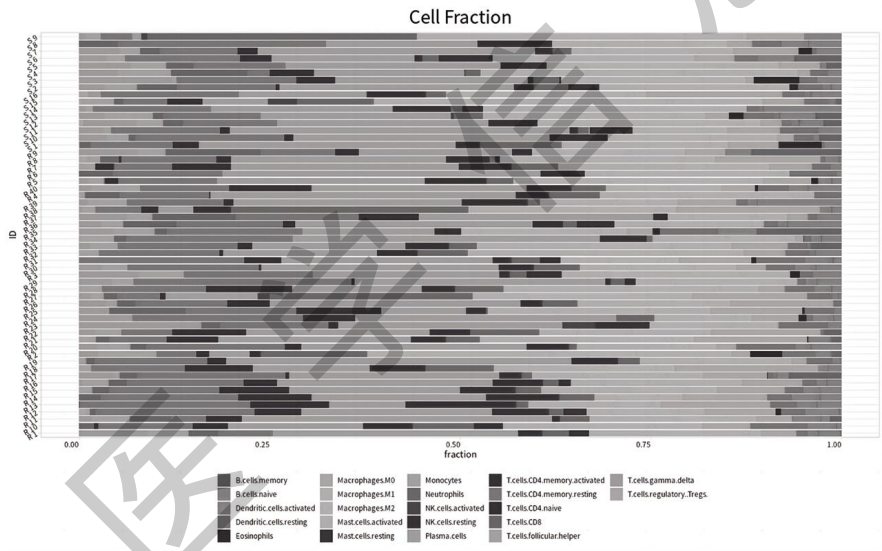


图 5 GSE119409 数据集中免疫细胞比例

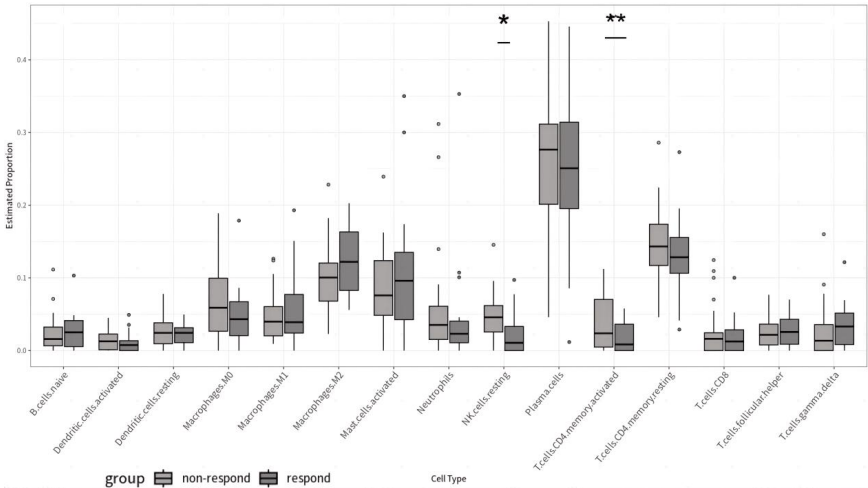


图 6 GSE119409 放疗应答与非应答分组免疫细胞比例差异

分析细胞水平的差异基因可以进一步得到与新辅助放疗应答相关的信息。首先使用模块五将结直肠癌数据集 GSE132465 制作成具有 6 种细胞类型、2659 个基因的签名矩阵(图 7)。接着运行模块三,可以得到不同细胞类型的组间差异基因及相应的 P 值、六种细胞的富集分析结果,主要功能富集在免疫、应激、血管生成、创伤反应方面(图 8)。B 细胞的细胞水平差异基因 GO 富集包括对外部刺激的反应、免疫反应、受损的 DNA 结合外切酶活性和对创伤的炎症反应等。T 细胞的细胞水平差异基因的 GO 富集包括细胞信号传导、细胞投影、中胚层迁移

与原肠形成、神经板形态发生和发育、受体激动剂活性等。髓系细胞的细胞水平差异基因 GO 富集包括巨噬细胞激活、超氧代谢过程、神经元投射和神经递质受体活性等。基质细胞的细胞水平差异基因 GO 富集包括血管形态发生、对创伤反应、细胞对内源性刺激的反应和 GABA 受体活性等。相较于传统转录组的组间差异分析,本应用可以获得细胞水平的特异性基因表达和组间差异基因,通过在 GSE119409 数据上的分析结果,发现了与直肠癌放射治疗反应密切相关的细胞水平的差异基因及富集模式。

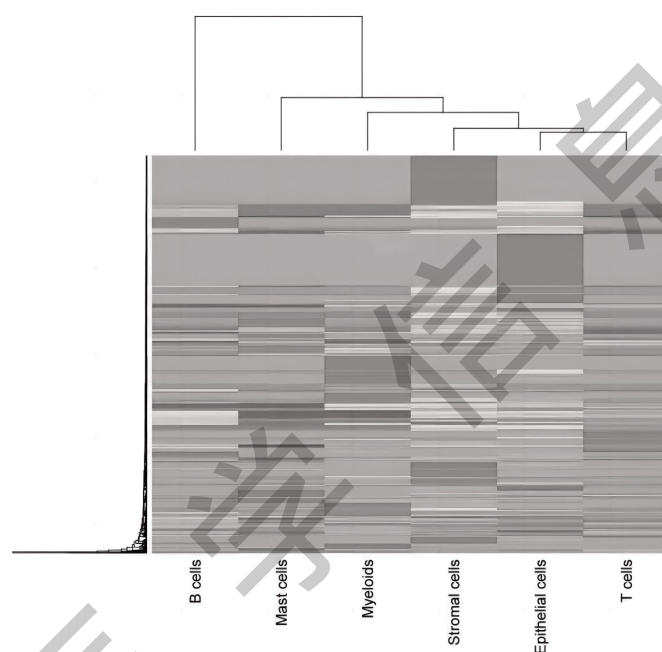


图 7 GSE132465 构建的六种细胞签名矩阵热图

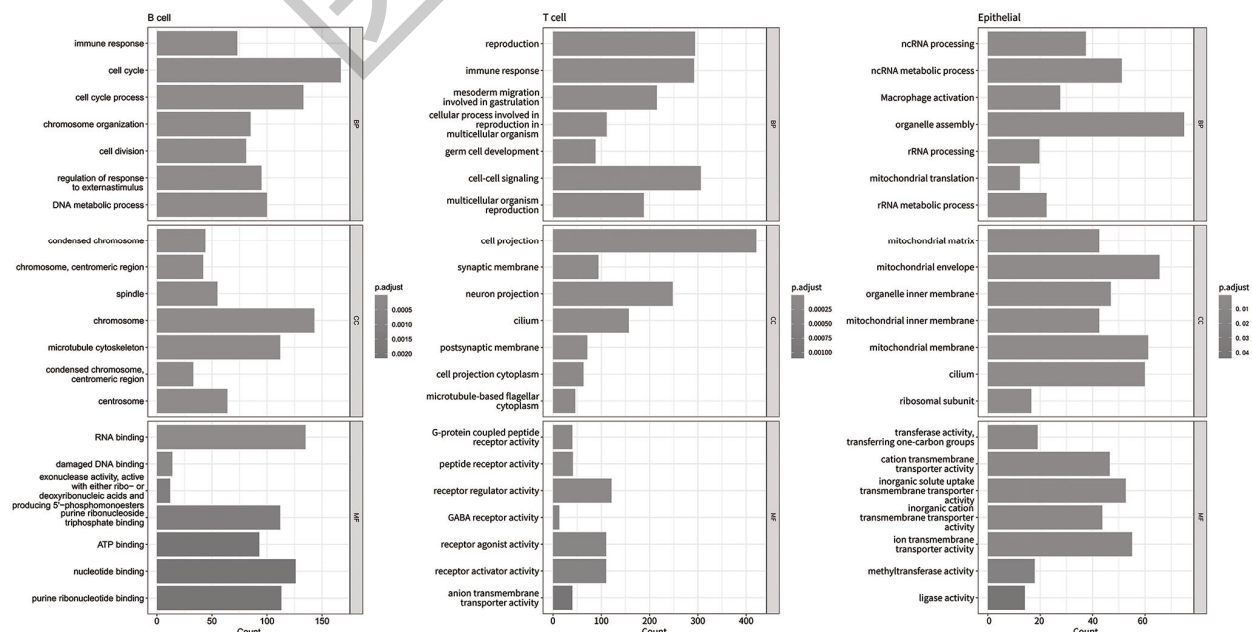


图 8 六种细胞类型的 GO 富集分析

3 总结

本文通过 R-shiny 框架构建了从混合组织中推测细胞比例及基因表达的在线应用,提供了标准的数据输入格式,自动输出不同反卷积方法的结果,并推断出最优的反卷积结果,进而实现组模式反卷积,得到各细胞类型的基因表达谱和差异基因、富集分析及单细胞数据处理,为用户提供了统一方便的研究平台,通过本应用推测细胞比例及基因表达,既可以充分挖掘现有的 bulk 转录组学和 DNA 甲基化数据,利用其在大规模样本分析中的优势,又可以为用户提供更精确的结果。通过公开数据集的应用,得到了传统 bulk 转录组差异分析无法得到的细胞水平的基因表达结果。并在完成系统功能模块设计同时,对各模块做了详细的解释说明,展示了核心功能的详细设计内容和实现细节。未来本应用将提升算法的准确度并不断挖掘和引入新的反卷积方法,拓展应用场景,为研究者提供有力的数据分析工具,为进一步研究疾病发展机制,制定疾病预防、诊断、治疗、预后和策略提供支持。

参考文献:

[1] Sharma A, Merritt E, Hu X, et al. Non-genetic intra-tumor heterogeneity is a major predictor of phenotypic heterogeneity and ongoing evolutionary dynamics in lung tumors[J]. Cell Rep, 2019, 29(8): 2164–2174.

[2] Hendry S, Salgado R, Gevaert T, et al. Assessing tumor infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group [J]. Adv Anat Pathol, 2017, 24(6): 311–335.

[3] Shen-Orr SS, Tibshirani R, Khatri P, et al. Cell type-specific gene expression differences in complex tissues [J]. Nature Methods, 2010, 7(4): 287–289.

[4] Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science [J]. Genome Biol, 2020, 21(1): 31.

[5] Li B, Li T, Liu JS, et al. Computational deconvolution of tumor-infiltrating immune components with bulk tumor gene expression data [J]. Methods Mol Biol, 2020, 2120: 249–262.

[6] Wang X, Park J, Susztak K, et al. Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference [J]. Nature Communications, 2019, 10(1): 380.

[7] Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles [J]. Nat Methods, 2015, 12(5): 453–457.

[8] Finotello F, Mayer C, Plattner C, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data [J]. Genome Med, 2019, 11(1): 34.

[9] Dong M, Thennavan A, Urrutia E, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references [J]. Briefings in Bioinformatics, 2020, 22(1): 416–427.

[10] Baron M, Veres A, Wolock SL, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structure [J]. Cell Systems, 2016, 3(4): 346–360.e4.

[11] Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression [J]. Genome Biol, 2016, 17(1): 218.

[12] Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape [J]. Genome Biol, 2017, 18(1): 220.

[13] Kang K, Huang C, Li Y, et al. CDSeqR: Fast Complete Deconvolution for Gene Expression Data from Bulk Tissues [J]. BMC Bioinformatics, 2021, 22(1): 262.

[14] Chu T, Wang Z, Pe'er D, et al. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology [J]. Nat Cancer, 2022, 3(4): 505–517.

[15] Menden K, Marouf M, Oller S, et al. Deep Learning-based Cell Composition Analysis from Tissue Expression Profiles [J]. Science Advances, 2020, 6(30): eaba2619.

[16] Chen Y, Wang Y, Chen Y, et al. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis [J]. Nat Commun, 2022, 13(1): 6735.

[17] Avila Cobos F, Alquicira-Hernandez J, Powell JE, et al. Benchmarking of cell type deconvolution pipelines for transcriptomics data [J]. Nat Commun, 2020, 11(1): 5650.

[18] He D, Chen M, Wang W, et al. Deconvolution of tumor composition using partially available DNA methylation data [J]. BMC Bioinformatics, 2022, 23(1): 355.

[19] Arneson D, Yang X, Wang K. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents [J]. Commun Biol, 2020, 3(1): 422.

[20] Chakravarthy A, Furness A, Joshi K, et al. Pan-cancer deconvolution of tumour composition using DNA methylation [J]. Nat Commun, 2018, 9(1): 3220.

[21] Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage [J]. Nat Immunol, 2019, 20(2): 163–172.

[22] Tu JJ, Li HS, Yan H, et al. EnDecon: cell type deconvolution of spatially resolved transcriptomics data via ensemble learning [J]. Bioinformatics, 2023, 39(1): btac825.

[23] 胡晓雯, 薛铭琰, 张枫, 等. 基于 R shiny 的法定传染病可视化分析系统的设计和初步应用 [J]. 南京医科大学学报(自然科学版), 2021, 41(3): 444–449, 459.

收稿日期: 2024-03-17; 修回日期: 2024-04-23

编辑/成森