About

The software is divided into four modules, including “Data Preparation”, “GET Module”, “COMP Module”, and “RUN Module”. Four modules connect with each other, and have a sequence relationship. Additionally, in order to facilitate users to view the generated results in real time, the software is divided into three parts. The upper part of the software contains title and module switching options. The lower part of the software is further divided into two parts, with parameters setting on the left part and results display on the right part.

“Data Preparation” is divided into TCGA dataset preparation and external validation dataset preparation. For the preparation of TCGA dataset, the software plans to provide several ways, including direct acquisition from built-in downloaded datasets, automatic downloading from websites, downloading from specified website links, and uploading locally. The software will preprocess and integrate the obtained data according to certain rules. In addition, users can also preprocess and integrate the data themselves, then upload the integrated dataset to the software directly. Users can choose different data preparation ways based on actual situation. If there is no need to make too many personalized adjustments on data, it is most convenient to download the data automatically from websites, and then preprocess and integrate them by

software. On the contrary, it is most appropriate for users to preprocess and integrate the data locally first, and then directly upload the integrated dataset to the software. For the preparation of external validation dataset, two data preparation ways of obtaining from the built-in downloaded datasets as well as automatically downloading from websites are no longer provided due to the wide range of sources for external validation dataset. In terms of omics data types, the software plans to add radiomics data on the basis of five original data types from MOVICS R package, containing mRNA, lncRNA, DNA methylation, copy number alterations, and somatic mutation.

"GET Module" is divided into five sub-modules, including "Get Elites", "Get Optimal Clustering Number", "Consensus Clustering", "Silhouette" and "Multi-omics Heatmaps". "Get Elites" processes the obtained omics data sequentially and performs dimensionality reduction according to certain rules, which will be used for cluster analysis later. "Get Optimal Clustering Number" combines two statistics of CPI and Gaps-statistics to plot and give the optimal clustering number. "Consensus Clustering" provides 10 clustering algorithms, including iClusterBayes, SNF, PINSPlus, NEMO, COCA, LRAcluster, ConsensusClustering, IntNMF, CIMLR, and MoCluster, from which users can choose one or more for clustering. If you want to run consensus clustering, you should choose at least two algorithms to get the consensus clustering diagram. "Silhouette"

calculates and visualizes the similarity between samples in each subtype derived from clustering results using Silhouette Coefficient, which can be used to evaluate the clustering results. “Multi-omics Heatmaps” combines multi-omics data and clustering results to generate multi-omics heatmaps, which can be utilized to evaluate the clustering results based on the expression differences of different subtypes in specific omics features.

“COMP Module” compares characteristics of different subtypes obtained from clustering results, which is divided into seven sub-modules. “Compare survival outcome” generates KAPLAN-MEIER curves to show the significance of survival differences among different subtypes. “Compare clinical features” generates a table to screen out clinical variables which are significantly associated with subtypes. “Compare mutational frequency” utilizes a table to display the mutation frequency of genes that meet certain conditions, and then draws a waterfall chart to display the genes whose mutation frequency is significantly different in each subtype. “Compare total mutation burden” compares the total mutation burden (TMB) among subtypes by drawing a box-violin plot, and then uses a table to show the TMB of each sample. “Compare fraction genome altered” compares the fraction of genome altered by copy number gain or loss among subtypes through bar charts, and tabulates according to the specifics of each sample at the same time. Fraction genome altered (FGA) represents the fraction of the genome altered by

copy number gain or loss. Specifically, FGA can be divided into FGG and FGL, which represent the fraction of genome gained and the fraction of genome lost respectively caused by copy number gain or loss. “Compare drug sensitivity” uses IC_{50} to compare the responses to drugs in GDSC database among subtypes, using box-violin plots for visualization. Besides, we listed estimated IC_{50} for each sample in the form of a table. “Compare agreement with other subtypes” compares the consistency of the clustering results with the current classification results (e.g., PAM50, pstage), which can evaluate the clustering results. A bar chart of four evaluation indicators, containing Rand Index (RI), Adjusted Mutual Information (AMI), Jaccard Index (JI), and Fowlkes-Mallows (FM) as well as an alluvial diagram are utilized for visualization. In addition, the software also lists the values of four evaluation indicators above in the form of a table.

“RUN Module” performs downstream analyses, which is also divided into seven sub-modules. “Run differential expression analysis” provides three types of algorithms including DESeq2, edgeR, and limma to find out differentially expressed genes for each subtype, which are displayed in the form of a table. “Run biomarker identification procedure” sets conditions to further screen out up-regulated and down-regulated marker genes separately for each subtype, which are displayed in the form of heatmaps and tables. “Run gene set enrichment analysis” screens out up-regulated

and down-regulated pathways respectively for each subtype based on the given gene set background files, and then shows the information of these pathways through a table. Additionally, the software also calculates enrichment scores of the screened pathways in each subtype, which are displayed in the form of heatmaps and tables. Analogously, “Run gene set variation analysis” calculates enrichment scores of each sample in each subtype according to the given gene set background files, and then displays the results using tables and a corresponding heatmap. Nearest Template Prediction (NTP) is a model-free method, and “Run nearest template prediction” utilizes NTP to predict the subtype of each sample in external validation dataset based on marker genes for each subtype obtained from TCGA dataset, which are displayed through a table. Then, a heatmap is drawn to show the consistency between prediction results and clustering results. Similarly, Partition around Medoids (PAM) is also a model-free prediction method. “Run partition around medoids classifier” uses PAM to predict the subtype of each sample in external validation dataset and evaluates the consistency between prediction results and clustering results through a similarity and reproducibility indicator named IGP. The predicted subtype of each sample and IGP value of each subtype are displayed in the form of tables respectively. The prediction results of external validation dataset from both NTP and PAM methods can be used to carry out analyses in “COMP Module”, which can validate the clustering

results. “Run consistency evaluation using Kappa statistics” calculates Kappa statistics, and then generates heatmaps to evaluate the consistency between clustering results and prediction results or the consistency between prediction results derived from NTP and PAM.

In addition to the four main modules above, since the software involves many parameters, and certain steps should be followed during the use, we specially set up a module called “Users Guide”. In order to help users better understand the software, a module called “About” was also set up.

References:

[1]. Lu Xiaofan, Meng Jialin, Zhou Yujie, et al.MOVICS: an R package for multi-omics integration and visualization in cancer subtyping[J].Bioinformatics,2021, 36 (22-23): 5539-5541.

[2]. Lu Xiaofan, Meng Jialin, Su Liwen, et al.Multi-omics consensus ensemble refines the classification of muscle-invasive bladder cancer with stratified prognosis, tumour microenvironment and distinct sensitivity to frontline therapies[J].Clinical and Translational Medicine,2021, 11 (12): e601.

[3]. Meng Jialin, Lu Xiaofan, Jin Chen, et al.Integrated multi-omics data reveals the molecular subtypes and guides the androgen receptor signalling inhibitor treatment of prostate cancer[J].Clinical and

Translational Medicine,2021, 11 (12): e655.

[4]. Ruan Xinjia, Ye Yuqing, Cheng Wenxuan, et al.Multi-omics integrative analysis of lung adenocarcinoma: an in silico profiling for precise medicine[J].Front Med (Lausanne),2022, 9: 894338.

[5]. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A (2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun, 9(1):4453.

[6]. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics, 19(1):71-86.

[7]. Meng C, Helm D, Frejno M, Kuster B (2016). moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. J Proteome Res, 15(3):755-765.

[8]. Hoadley KA, Yau C, Wolf DM, et al (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158(4):929-944.

[9]. Monti S, Tamayo P, Mesirov J, et al (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Mach Learn, 52:91-118.

[10]. Chalise P, Fridley BL (2017). Integrative clustering of multi-level omic data based on non-negative matrix factorization algorithm. PLoS One, 12(5):e0176278.

[11]. Wu D, Wang D, Zhang MQ, Gu J (2015). Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics, 16(1):1022.

[12]. Rappoport N, Shamir R (2019). NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics, 35(18):3348-3356.

[13]. Nguyen H, Shrestha S, Draghici S, Nguyen T (2019). PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics, 35(16):2843-2846.

[14]. Wang B, Mezlini AM, Demir F, et al (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat Methods, 11(3):333-337.

[15]. Gu Z, Eils R, Schlesner M (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18):2847–2849.

[16]. Mayakonda A, Lin D C, Assenov Y, et al. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res, 28(11):1747-1756.

[17]. Shyr C, Tarailo-Graovac M, Gottlieb M, Lee JJ, van Karnebeek C, Wasserman WW. (2014). FLAGS, frequently mutated genes in public exomes. BMC Med Genomics, 7(1): 1-14.

[18]. Cerami E, Gao J, Dogrusoz U, et al. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer

Genomics Data. Cancer Discov, 2(5):401-404.

[19]. Gao J, Aksoy B A, Dogrusoz U, et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal, 6(269):pl1-pl1.

[20]. Geeleher P, Cox N, Huang R S (2014). pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. PLoS One, 9(9):e107468.

[21]. Geeleher P, Cox N J, Huang R S (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol, 15(3):1-12.

[22]. Robinson MD, McCarthy DJ, Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139-140.

[23]. McCarthy DJ, Chen Y, Smyth GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40(10):4288-4297.

[24]. Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15(12):550-558.

[25]. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 43(7):e47.

[26]. Yu G, Wang L, Han Y, He Q (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS, 16(5):284-287.

[27]. Subramanian A, Tamayo P, Mootha V K, et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci, 102(43):15545-15550.

[28]. ssgsea: Barbie, D.A. et al (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature, 462(5):108-112.

[29]. Hänzelmann, S., Castelo, R. and Guinney, J. (2013). GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics, 14(1):7.

[30]. Lee, E. et al (2008). Inferring pathway activity toward precise disease classification. PLoS Comp Biol, 4(11):e1000217.

[31]. Tomfohr, J. et al (2005). Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics, 6(1):1-11.

[32]. Eide P W, Bruun J, Lothe R A, et al. (2017). CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci Rep, 7(1):1-8.

[33]. Hoshida, Y. (2010). Nearest Template Prediction: A Single-Sample-Based Flexible Class Prediction with Confidence Assessment. PLoS One, 5(11):e15543.

[34]. Tibshirani R, Hastie T, Narasimhan B and Chu G (2002). Diagnosis of

multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci, 99,6567–6572.

[35]. Kapp A V, Tibshirani R. (2007). Are clusters found in one dataset present in another dataset? Biostatistics, 8(1):9-31.

[36]. Pierre-Jean M, Deleuze J F, Le Floch E, et al (2019). Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. Brief Bioinformatics.

[37]. Rappoport N, Shamir R (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res, 46(20):10546-10562.

[38]. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature, 2012, 490(7418):61-78.

[39]. Yau C, Esserman L, Moore D H, et al. (2010). A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. Breast Cancer Res, 5(12):1-15.

[40]. Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. J R Stat Soc Series B Stat Methodol, 63(2):411-423.

[41]. Strehl, Alexander; Ghosh, Joydeep. (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. J Mach Learn Res, 3(Dec):583–617.

[42]. Le DT, Uram JN, Wang H, et al. (2015). PD-1 Blockade in tumors with

mismatch-repair deficiency. N Engl J Med, 372(26):2509–2520.

[43]. Davoli T, Uno H, Wooten E C, et al. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science, 355(6322):261-U75.

[44]. Rand W M. (1971). Objective criteria for the evaluation of clustering methods. J Am Stat Assoc, 66(336):846-850.

[45]. Vinh, N. X., Epps, J., Bailey, J. (2009). Information theoretic measures for clusterings comparison. Proceedings of the 26th Annual International Conference on Machine Learning - ICML ’09. p. 1.

[46]. Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient). Dictionary of Bioinformatics and Computational Biology.

[47]. Fowlkes E B, Mallows C L. (1983). A method for comparing two hierarchical clusterings. J Am Stat Assoc, 78(383):553-569.

[48]. Tibshirani R, Hastie T, Narasimhan B and Chu G (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci, 99(10):6567–6572.

[49]. Kapp A V, Tibshirani R. (2007). Are clusters found in one dataset present in another dataset?, Biostatistics, 8(1):9-31.

To cite this software:

(APA Style)

Yan, F.R., Lu, X.F., Meng, J. L., Zhu, J.K., Cheng, W.X., Wang, X., Yang, J.L., Wang, W.X. (2022). MOVICS: Multi-Omics integration and VIsualization in

Cancer Subtyping (V1.0) [Computer software]. Retrieved from http://www.movics-cpu.com:3838/.

(MLA Style)

Zhu, Junkai, et al. "MOVICS: Multi-Omics integration and VIsualization in Cancer Subtyping." A Shiny Software to Perform Multi-omics Analysis, V1.0, China Pharmaceutical University Research Center of Biostatistics and Computational Pharmacy, Nov. 2022, http://www.movics-cpu.com:3838/.

Contact us:

If you find some errors during the use of this software or if you have some advice on this software, we welcome you to write letters to us:

Lu, Xiaofan: ***xlu.cpu@foxmail.com***

Thanks for your using and we are looking forward to your valuable feedback!