SOFTWARE **Open Access**

Smccnet 2.0: a comprehensive tool for multi-omics network inference with shiny visualization

Weixuan Liu[1*], Thao Vu[1], Iain R. Konigsberg[2], Katherine A. Pratte[3], Yonghua Zhuang[4] and Katerina J. Kechris[1]

*Correspondence:
weixuan.liu@cuanschutz.edu

[1] Department of Biostatistics and Informatics, School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[2] Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[3] Department of Biostatistics, National Jewish Health, Denver 80206, CO, USA
[4] Department of Pediatrics, University of Colorado Anschutz Medical Campus, Aurora 80045, CO, USA

Abstract

***Summary*:** Sparse multiple canonical correlation network analysis (SmCCNet) is a machine learning technique for integrating omics data along with a variable of interest (e.g., phenotype of complex disease), and reconstructing multi-omics networks that are specific to this variable. We present the second-generation SmCCNet (SmCCNet 2.0) that adeptly integrates single or multiple omics data types along with a quantitative or binary phenotype of interest. In addition, this new package offers a streamlined setup process that can be configured manually or automatically, ensuring a flexible and user-friendly experience.

***Availability*:** This package is available in both CRAN: https://cran.r-project.org/web/packages/SmCCNet/index.html and Github: https://github.com/KechrisLab/SmCCNet under the MIT license. The network visualization tool is available at https://smccnet.shinyapps.io/smccnetnetwork/.

Keywords: Multi-omics integration, Network analysis, Automated pipeline

Background

Advances in sequencing and mass spectrometry technologies have allowed access to extensive -omics data sets such as transcriptomics, proteomics, and metabolomics, which allows the integration of different omics data to gain biological insights into complex diseases [1]. In recent years, multi-omics integration methods such as multi-omics factor analysis (MOFA and MOFA +) have been proposed to integrate multiple different layers of molecular profiles and capture biological-relevant information using latent factors [2, 3]. Another strategy is multi-omics network inference for integrating multiple -omics data sets to infer molecular interactions with respect to the trait(s) of complex disease and gain insights into associated biological processes [4]. Recent multi-omics network inference methods include knowledge-guided multi-omics network inference (KiMONo) [5] and Biomarker discovery using Latent Variable Approaches for Omics Studies (DIABLO) [6].

Canonical Correlation Analysis (CCA) [7], which seeks to find the linear combination (canonical weight) that maximizes the correlation between two sets of data [7]. When there are more than two datasets, Sparse multiple Canonical Correlation Analysis (SmCCA) is used to maximize all pairwise relationships between multipel datasets [8] and has been widely used for multi-omics integration [9]. Although there are many existing multi-omics integration methods based on sparse multiple canonical correlation analysis, many methods focus on predictive tasks [10, 11], and only a few studies have focused on reconstructing multi-omics interaction networks with respect to the trait of complex disease [6]. Sparse multiple Canonical Correlation Network Analysis (SmCCNet) is such a canonical correlation-based integration method that reconstructs phenotype-specific multi-omics networks, and simulation studies have shown that it outperforms other methods in detecting the correct features [12]. SmCCNet starts by using SmCCA to construct a global multi-omics interaction network with respect to a trait, then implements the hierarchical clustering algorithm to partition the global networks into multiple subnetwork modules. Each subnetwork module represents a specific subset of potential biological pathways/processes. SmCCNet has been applied to different multi-omics integration tasks such as extracting protein-metabolite networks [13], mRNA-miRNA networks [14], and microbiome-proteomics network [15] associated with disease phenotypes.

Despite successful applications, SmCCNet in its current version has several limitations. In the SmCCA modeling step: (1) the first version of SmCCNet (SmCCNet 1.0) only analyzes two omics data types with a quantitative phenotype, and it does not consider a single-omics data or more than two -omics data; (2) SmCCNet 1.0 can only treat a binary phenotype as quantitative phenotype, similar to other methods like DIABLO and (3) SmCCNet 1.0 cannot select scaling factors that prioritize correlation structures of interest (e.g., omics-phenotype correlation over omics-omics correlation). Besides these modeling steps, SmCCNet 1.0 has other drawbacks in downstream network steps: (1) after clustering molecular features into different modules, sometimes the subnetworks will contain molecular features that contribute less to the network, and SmCCNet 1.0 is not able to eliminate those features; (2) SmCCNet 1.0 uses principal component analysis to summarize each network, but it fails to consider the network topology (i.e., how each molecular feature interacts with other network features); (3) SmCCNet 1.0 requires more than 1000 lines of code with hard-coded setup to run through the pipeline, which usually takes more than 24 h to run through even after parallel processing, making it computationally inefficient; and (4) there is a lack of visualization options in SmCCNet 1.0, especially for multi-omics network visualization.

To enhance the practical utility of SmCCNet with improved or novel methods and functionalities, we have rewritten and upgraded the software (SmCCNet 2.0) to flexibly accommodate one or more omics data types, as well as a binary phenotype. We also created an automated end-to-end pipeline that obtains the final network result with just a single line of code to substantially enhance the usability of the pipeline. Regarding network analysis steps, we proposed a new network pruning algorithm after the network clustering step to reduce the subnetworks so that the most informative network structure can be obtained. To summarize each final subnetwork, we implemented the NetSHy network summarization algorithm to take network topology into account. Furthermore, the other improvements from SmCCNet 2.0 include a data preprocessing pipeline, enhanced computational efficiency, an online RShiny application for network visualization, and the storage of accessible and reproducible network analysis results in a user-specified directory.

Methods and implementation

Overall, we showcased that we were able to improve the following functions within SmCCNet:

- Multi-omics SmCCNet with quantitative phenotype allows integration of more than two -omics data (improved functionality, Section Multi-omics SmCCNet with Quantitative Phenotype).
- Novel hybrid SmCCNet algorithm with binary phenotype (novel functionality and algorithm, section "Multi-omics SmCCNet with binary phenotype").
- Single-omics implementation of SmCCNet algorithm with either quantitative or binary phenotype (novel functionality and algorithm, section "Single-omics SmCCNet with quantitative/binary phenotype").
- Novel model-wise optimal scaling factors selection algorithm for multi-omics SmCCNet (novel functionality and algorithm, section "Scaling factor determination").
- Implementation of NetSHy network summarization method to summarize network based on network topology (novel functionality, section "Network clustering and pruning").
- Subnetwork pruning algorithm to reduce the size of multi-omics network (novel functionality and algorithm, section "Network Clustering and pruning").
- New subnetwork visualization RShiny application for multi-omics interaction network visualization (novel functionality, section "Network visualization").
- Fast Automated SmCCNet conducts end-to-end pipeline with a single line of code and further improves the algorithm speed (novel functionality, section "Automated SmCCNet").
- New -omics data preprocessing pipeline to filter out features with low variability, regress out clinical covariates, and center/scale (novel functionality, a general preprocessing step before running SmCCNet).
- Simpler coding setup to run SmCCNet manually and boost the algorithm speed by 100–1000x (improved functionality, described in section "Methods and implementation").

Before running SmCCNet, the user can apply a streamlined function to preprocess the omics data, including filtering features with low Coefficient of Variation (CoV), centering and scaling each molecular feature, and regressing out effects from covariates. The data preprocessing pipeline can be implemented by using the *dataPreprocess()* function. The end-to-end pipeline of SmCCNet takes in any number of molecular profiles (omics data) and either a quantitative or binary phenotype, and outputs the single/multi-omics subnetwork modules that are associated with the phenotype. The general workflow of SmCCNet is shown in Fig. 1.

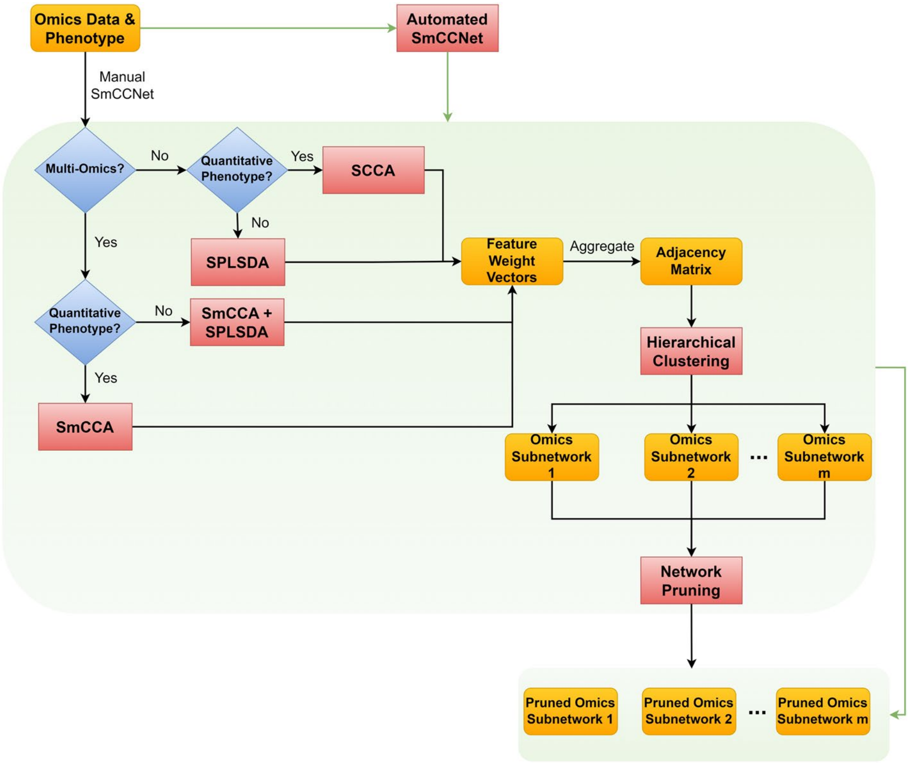


Fig. 1 SmCCNet Workflow. The workflow of the second-generation SmCCNet, which takes in single-/multi-omics data with quantitative/binary phenotype and outputs multiple pruned omics subnetwork modules. Steps in the box represent the stepwise process of executing the SmCCNet, and if automated SmCCNet is used, all these steps within the box will be executed automatically

Number of omics data and phenotype modality

In general, SmCCNet consists of the following steps (See Figs. 2, 3, and 4):

- Step I: Determination of SmCCA/SPLSDA sparsity penalty parameters. The user can select the penalties for feature selection based on prior knowledge. Alternatively, the user can pick sparsity penalties based on a K-fold cross-validation (CV) procedure that minimizes the total prediction error (e.g., Fig. 2). The K-fold CV procedure enhances the robustness of selected penalties when generalizing to similar independent omics data sets.
- Step II: Subsampliing algorithm that randomly subsample omics features, apply SmCCA/SPLSDA with chosen penalties and compute a canonical weight vector for each subsample. Repeat the process many times.
- Step III: Feature similarity matrix computation based on canonical weight matrix.
- Step IV: Hierarchical tree clustering to the similarity matrix to simultaneously identify multiple subnetworks.
- Step V: Network pruning algorithm to prune each subnetwork obtained from Step IV and visualize the -omics network with an RShiny application (https://smccnet.shinyapps.io/smccnetnetwork/) or Cytoscape.

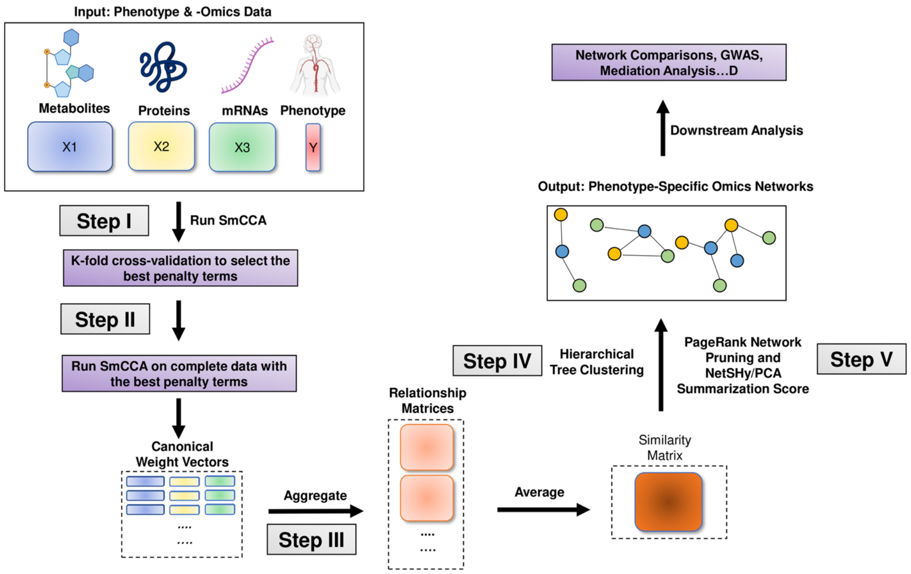

Input: Phenotype & -Omics Data
Metabolites
Proteins
mRNAs
Phenotype
X1
X2
X3
Y
Step I
Run SmCCA
K-fold cross-validation to select the best penalty terms
Step II
Run SmCCA on complete data with the best penalty terms
Canonical Weight Vectors
Aggregate
Step III
Relationship Matrices
Average
Similarity Matrix
Step IV
Hierarchical Tree Clustering
PageRank Network Pruning and NetSHy/PCA Summarization Score
Step V
Output: Phenotype-Specific Omics Networks
Downstream Analysis
Network Comparisons, GWAS, Mediation Analysis…D

Fig. 2 Multi-Omics SmCCNet with Quantitative Phenotype Workflow. The workflow of the second-generation SmCCNet with multi-omics data and quantitative phenotype

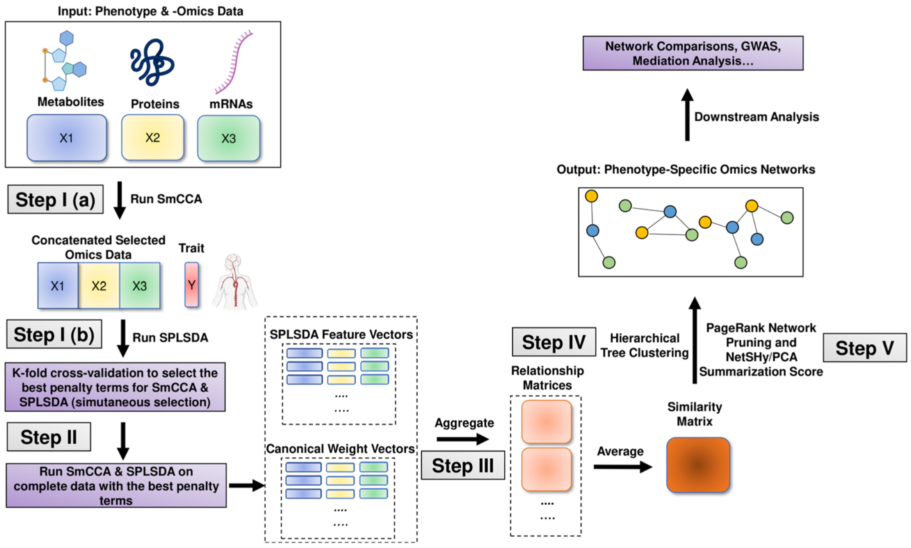

Input: Phenotype & -Omics Data
Metabolites
Proteins
mRNAs
X1
X2
X3
Step I (a)
Run SmCCA
Concatenated Selected Omics Data
Trait
X1
X2
X3
Y
Step I (b)
Run SPLSDA
K-fold cross-validation to select the best penalty terms for SmCCA & SPLSDA (simutaneous selection)
Step II
Run SmCCA & SPLSDA on complete data with the best penalty terms
SPLSDA Feature Vectors
Canonical Weight Vectors
Aggregate
Step III
Relationship Matrices
Average
Similarity Matrix
Step IV
Hierarchical Tree Clustering
PageRank Network Pruning and NetSHy/PCA Summarization Score
Step V
Output: Phenotype-Specific Omics Networks
Downstream Analysis
Network Comparisons, GWAS, Mediation Analysis…

Fig. 3 Multi-Omics SmCCNet with Binary Phenotype Workflow. The workflow of the second-generation SmCCNet with multi-omics data and binary phenotype

Steps III to V remain consistent across all scenarios, regardless of the number of omics data types used and the phenotype modality involved. However, Steps I and II differ depending on the specific scenario. Below, we provide detailed descriptions for Steps I and II for each scenario:

Multi-omics SmCCNet with quantitative phenotype

If multi-omics data is used with quantitative phenotype, same as SmCCNet 1.0, we implement the SmCCA algorithm (Eq. 1) for feature selection and network construction, which is achieved by using *getRobustWeightsMulti()*. For T omics

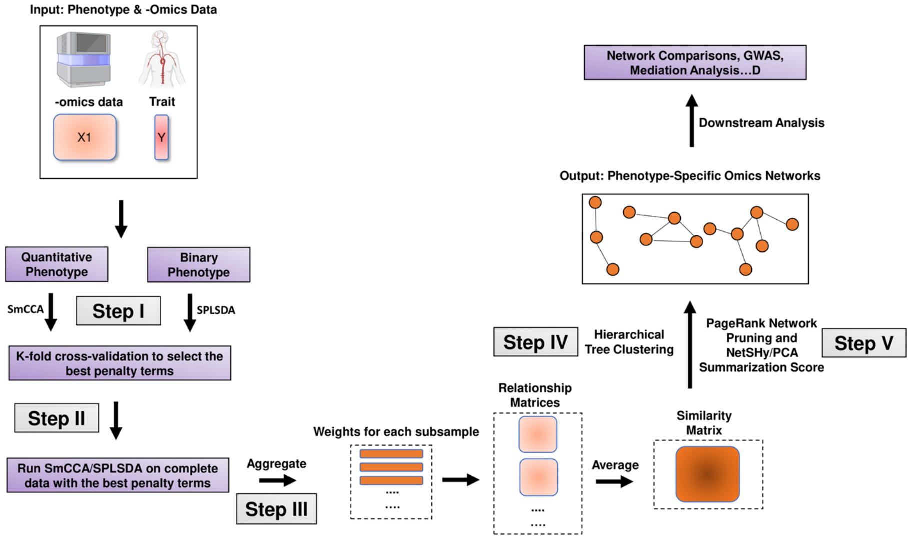

Input: Phenotype & -Omics Data
-omics data
Trait
X1
Y
Quantitative Phenotype
Binary Phenotype
SmCCA
Step I
SPLSDA
K-fold cross-validation to select the best penalty terms
Step II
Run SmCCA/SPLSDA on complete data with the best penalty terms
Aggregate
Step III
Weights for each subsample
Relationship Matrices
Average
Similarity Matrix
Step IV
Hierarchical Tree Clustering
PageRank Network Pruning and NetSHy/PCA Summarization Score
Step V
Output: Phenotype-Specific Omics Networks
Downstream Analysis
Network Comparisons, GWAS, Mediation Analysis...D

Fig. 4 Single-Omics SmCCNet with Quantitative/Binary Phenotype workflow The workflow of the second-generation SmCCNet with single-omics data and quantitative/binary phenotype

data $X_1, X_2, ...X_T$ and a quantitative phenotype Y measured in the same subjects. SmCCA finds the canonical weights $w_1, w_2, ..., w_T$ that maximize the (weighted or unweighted) sum of pairwise canonical correlations between $X_1, X_2, ..., X_T$ and Y, under sparsity constraints in Equ 1. In SmCCNet, the sparsity constraint functions $P_t(\cdot), t = 1, 2, ..., T$, are the least absolute shrinkage and selection operators (LASSO) [16]. The weighted version corresponds to $a_{i,j}, b_i$ (also called scaling factors), which are not all equal; the unweighted version corresponds to $a_{i,j} = b_i = 1$ for all $i,j = 1, 2, ..., T$, where $a_{i,j}$ are for between -omics relationships, while b_i is for the single omics and phenotype relationship.

$$(w_1, w_2, ..., w_T) = \arg\max_{\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_T} \left(\sum_{\substack{i<j \\ i,j=1,2,...,T}} a_{i,j} \tilde{w}_i^T X_i^T X_j \tilde{w}_j + \sum_{i=1}^{T} b_i \tilde{w}_i^T X_i^T Y \right),$$
$$\text{subject to} \left\| \tilde{w}_t \right\|^2 = 1, \quad P_t(\tilde{w}_t) \le c_t, \quad t = 1, 2, ..., T. \tag{1}$$

The sparsity penalties c_t influence how many features will be included in each subnetwork. With pre-selected sparsity penalties, the SmCCNet algorithm creates a network similarity matrix based on SmCCA canonical weights from repeatedly subsampled omics data and the phenotype and then finds multi-omics modules that are relevant to the phenotype. The subsampling scheme improves network robustness by analyzing subsets of omics features multiple times and forms a final similarity matrix by aggregating results from each subsampling step.

In step I, we use *k*-fold cross-validation to determine the optimal penalty parameters based on the loss function. In SmCCNet 1.0, the loss function is defined to be the prediction error, which is defined as follows:

$$PredErr = |trainCC - testCC|, \tag{2}$$

where *trainCC* and *testCC* are defined as the training canonical correlation and testing canonical correlation respectively. In SmCCNet 2.0, the loss function is defined to be the scaled prediction error:

$$scaledPredErr = \frac{|trainCC - testCC|}{|testCC|}. \tag{3}$$

Compared to prediction error, the scaled prediction error aims not only to minimize the discrepancy between the training and testing canonical correlations but also to maximize the testing canonical correlation. This approach effectively prevents the selection of penalty parameters that could result in extremely low testing canonical correlations.

In Step II, to enhance the robustness of the multi-omics network, we employ a subsampling algorithm.[1] This algorithm selects only a fraction of the molecular features of each molecular profile during each iteration of subsampling. The complete workflow of multi-omics SmCCNet with quantitative phenotype is shown in Fig. 2. The detailed implementation can be found in the package vignette https://cran.r-project.org/web/packages/SmCCNet/vignettes/SmCCNet_Vignette_MultiOmics.pdf.

The setup of parameters substantially impacts the results and robustness of the SmCCNet. Specifically, the choice of scaling factors influences which molecular features are incorporated into the final networks. If scaling factors are set to emphasize the omics-phenotype relationship, more molecular features with strong correlations to phenotypes are selected. Conversely, prioritizing between-omics associations leads to the inclusion of molecular features with robust inter-omics connections. To optimize these scaling factors, cross-validation can be used (details can be found in Sect.), and we recommend setting the cross-validation tuning grid to favor the omics-phenotype relationship. This approach aims to generate subnetworks with a stronger association to phenotypes. Additionally, the number of subsampling iterations affects the robustness of the networks; a greater number of iterations typically results in more robust subnetwork outcomes. Given these considerations, depending on the computational resources available, we suggest performing subsampling between 100 and 1000 times to enhance the robustness of the results.

Multi-omics SmCCNet with binary phenotype

We developed the hybrid algorithm between Sparse Partial Least Squared Discriminant Analysis (SPLSDA) [17] and SmCCA for feature selection and network construction, which is achieved by using *getRobustWeightsMultiBinary()*. SPLSDA is a two-step approach to implement the partial least squared algorithm with the binary outcome, which is a combination of partial least squared and logistic regression.

First, SmCCA (Sparse Multiple Canonical Correlation Analysis) is applied without involving phenotype data to filter molecular features and identify those that are interconnected. Next, SPLSDA (Sparse Partial Least Squares Discriminant Analysis) is employed

[1] We drop subsampling from step I, which substantially increases the computational speed.

on the chosen features across all molecular profiles to determine which features are associated with the phenotype. Finally, the canonical weights obtained from SmCCA and the feature importance weights from SPLSDA are combined into a weighted average, providing a consolidated measure of each feature's relevance.

To select the optimal penalty parameters, k-fold cross-validation is implemented on the complete hybrid algorithm to evaluate penalty terms on SmCCA and SPLSDA simultaneously. In SmCCNet 2.0, various metrics can be used to evaluate each set of penalty parameters, which include prediction accuracy, AUC score, precision, recall, and F1 score.

After the optimal penalty terms are selected, the hybrid method is run on the complete dataset, and, the same as regular SmCCNet, we use the subsampling scheme to ensure the robustness of the multi-omics network.

The complete workflow of multi-omics SmCCNet with quantitative phenotype is shown in Fig. 3. Consider $X_1, X_2, ..., X_T$ as T omics datasets, and Y as the phenotype data. The hybrid SmCCNet algorithm with binary phenotype is defined as follows (Step II in Fig. 3 run through Stage 1–3):

- **Stage 1: Weighted/Unweighted Sparse Multiple Canonical Correlation Analysis (SmCCA, Step I(a) in Fig.** 3): The is performed on $X_1, X_2, ..., X_T$ (excluding phenotype data). The output is canonical weight vectors (with nonzero entries, zero entries are filtered) $\tilde{W}_t \in \mathbb{R}^{p_t^{(\text{sub})} \times 1}, t = 1, 2, ..., T$, which represent the omics-omics connections. In this step, we filter out features that have no connections with other features, which helps reduce dimensionality. Note that we tend to set relaxed penalty terms for this step to include as many omics features as possible to increase the performance of the classifier in the next step.
- **Stage 2: Subset Omics Data (Step I(a) in Fig.** 3): Each dataset $X_1, X_2, ..., X_T$ is subsetted to include only omics features selected in Step 1, calling subsetted data $X_t^{(\text{sub})} \in \mathbb{R}^{n \times p_t^{(\text{sub})}}$.
- **Stage 3: Multi-omics Data Concatenation and Sparse Partial Least Squared Discriminant Analysis Implementation (SPLSDA, Step I(b) in Fig.** 3): The subsetted datasets $X_1^{(\text{sub})}, X_2^{(\text{sub})}, ..., X_T^{(\text{sub})}$ are concatenated into $X^{(\text{sub})} = [X_1^{(\text{sub})}, X_2^{(\text{sub})}, ..., X_T^{(\text{sub})}] \in \mathbb{R}^{n \times p^{(sub)}}, p^{(sub)} = \sum_{i=1}^{T} p_i$. The SPLSDA algorithm is then run to extract R latent factors and a projection matrix, by default, R is set to 3. The projection matrix is defined as $Z \in \mathbb{R}^{p^{(\text{sub})} \times R}$. Latent factors are defined as $L = [r_1, r_2, ..., r_R] = X^{(\text{sub})} \cdot Z \in \mathbb{R}^{n \times R}$.
- **Stage 4: Latent Factors Aggregation (After Step II and Before Step III in Fig.** 3): The R latent factors are aggregated into one using logistic regression, defined by $\text{logit}(Y) = \alpha_1 r_1 + \alpha_2 r_2 + ... + \alpha_R r_R$. Feature weights are given by aggregation of the projection matrix from Sparse PLSDA $W_t^* = Z_t \cdot \alpha \in \mathbb{R}^{p_t^{(\text{sub})} \times 1}, t = 1, 2, ..., T, \alpha = [\alpha_1, \alpha_2, ..., \alpha_R] \in \mathbb{R}^{R \times 1}$, where Z_t is the subset of projection matrix Z such that it only includes features from the tth omics data.
- **Stage 5: Final Canonical Weight Normalization and Calculation for Global Network Construction (Step III in Fig.** 3): The feature weights $W_1^*, W_2^*, ..., W_T^*$ based on SPLSDA are normalized to have an L2 norm of 1. Let γ_1 and γ_2 be two scalars

representing the strength of omics-omics and omics-phenotype connections, respectively. The final canonical weight is obtained by combining the canonical weight from step 1 and the feature weight from the classifier from step 4: $W_t = \frac{\gamma_1}{\gamma_1+\gamma_2}\tilde{W}_t + \frac{\gamma_2}{\gamma_1+\gamma_2}W_t^*, t = 1, 2, ..., T$.

The configuration of parameters in SmCCNet influences both the results and the robustness of the inferred networks. One key parameter is the between-shrinkage factor, which determines the molecular features included in the final networks. Setting this factor to emphasize the omics-phenotype relationship (higher values) leads to the selection of molecular features with strong correlations to phenotypes. Conversely, a focus on between-omics associations results in the inclusion of features with strong inter-omics connections. Generally, it is advisable to set the between-shrinkage factor to favor the omics-phenotype relationship to generate subnetworks with a stronger association to phenotypes. Additionally, the robustness of the networks can be influenced by the number of subsampling iterations, as discussed in Sect. , and the number of latent factors in SPLSDA. A higher number of latent factors typically leads to more robust network outcomes. Depending on the computational resources available and the data's dimensionality, we recommend setting the number of latent factors between 3 and 10. Furthermore, the choice of evaluation metric can also impact the final network results. While the default evaluation method is the AUC score, other metrics such as accuracy, precision, recall, and F1 score may be considered based on specific analytical needs or study goals. This flexibility allows researchers to tailor the evaluation to better reflect the focus and nuances of their specific data and research objectives.

Single-omics SmCCNet with quantitative/binary phenotype

In multi-omics SmCCNet, the between-omics interaction is taken into account. However, in the single-omics setting, this is no longer considered. Therefore, we developed two functions separately to tackle single-omics analysis (Fig. 4). If a quantitative phenotype is used, then sparse canonical correlation analysis (SCCA) is used to construct the global network by using the function *getRobustWeightsSingle()*; if a binary phenotype is used, then both stage 3 and 4 of the hybrid algorithm above are used with SPLSDA algorithm by using the function *getRobustWeightsSingleBinary()*. For more information about the single-omics SmCCNet pipeline setup, runnable examples are provided in the package vignette. In addition, this pipeline has been applied to the proteomics network analysis to identify the single-omics networks associated with pulmonary function and smoking behavior [18].

Scaling factor determination

The scaling factors ($a_{i,j}$ and b_j) in Eq. 1 can be supplied to prioritize the correlation structure of interest in Steps I and II of the SmCCNet Pipeline. Users can choose to supply their own choice of scaling factors or select them with the model-based approaches. We provide three different methods for selecting the scaling factors.

Prompt to define scaling factors

If a user is able to supply the scaling factors for the model based on prior knowledge, an interactive function *scalingFactorInput()* can be used to enter scaling factors manually for each pairwise correlation. For instance, when entering *scalingFactorInput(c(’mRNA’,’miRNA’, ’phenotype’))*, three sequential prompts will appear, requesting the scaling factors for mRNA-miRNA, mRNA-phenotype, and miRNA-phenotype relationships, respectively.

Pairwise correlation to select scaling factors with automated SmCCNet

As an alternative, the pairwise correlation between each pair of omics data can be used to set the scaling factors. For this option, SCCA is run with a stringent penalty pair. The resulting canonical correlation will be treated as the between-omics scaling factor, while a scaling factor of 1 will be used for the omics-phenotype relationship. In addition, we introduce another parameter called the shrinkage factor to prioritize either the omics-omics relationship or the omics-phenotype relationship. For example, in a multi-omics analysis with two omics data, if the omics-omics correlation is 0.8 by SCCA, and the shrinkage parameter is 2, then the final scaling factors are set to $(a, b_1, b_2) = c(0.4, 1, 1)$, where a are the between-omics relationship and b’s are the omics-phenotype relationships. This method is currently implemented in the automated SmCCNet approach.

Cross-validation to select scaling factors

The approach employs cross-validation to identify optimal scaling factors, illustrated using two omics types as an example. Initially, candidate sets of scaling factors are generated with all omics-omics scaling factors set to 1, and omics-phenotype scaling factors adjusted so their sum equals 1 for comparability. For instance, scaling factors $(a_{1,2}, b_1, b_2)$ must fulfill the condition $a_{1,2} + b_1 + b_2 = 1$. A nested grid search strategy is then applied to simultaneously optimize the scaling factors and penalty parameters. Within this framework, as different sets of scaling factors are evaluated, the optimal penalty parameters are selected. For each candidate set of scaling factors, the optimal sparse penalty parameters (denoted as $l1$, $l2$) are identified via k-fold cross-validation. The evaluation metric’s value corresponding to these parameters is recorded, which is associated with the optimal penalty parameters for each candidate set. This process is repeated across all potential combinations of scaling factors. The set of scaling factors yielding the best performance, according to the chosen evaluation metric, is selected as the optimal set, together with its associated optimal penalty parameters. Given the exponential increase in possible scaling factor combinations with more than three types of -omics data, the use of the automated SmCCNet algorithm is recommended for selecting optimal scaling factors in analyses involving larger numbers of -omics data types.

Network clustering and pruning

The adjacency matrix is formed by taking the outer product of the canonical weights. After obtaining the adjacency matrix, hierarchical clustering [19] is implemented to

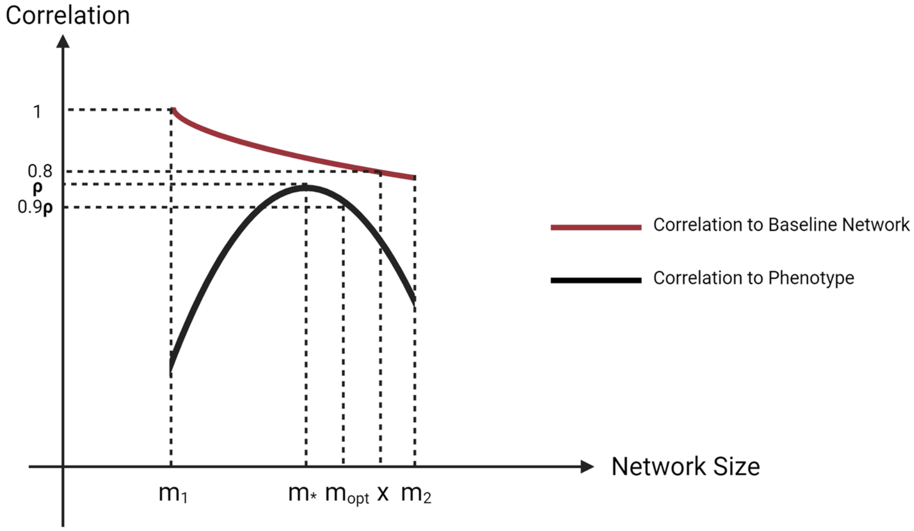
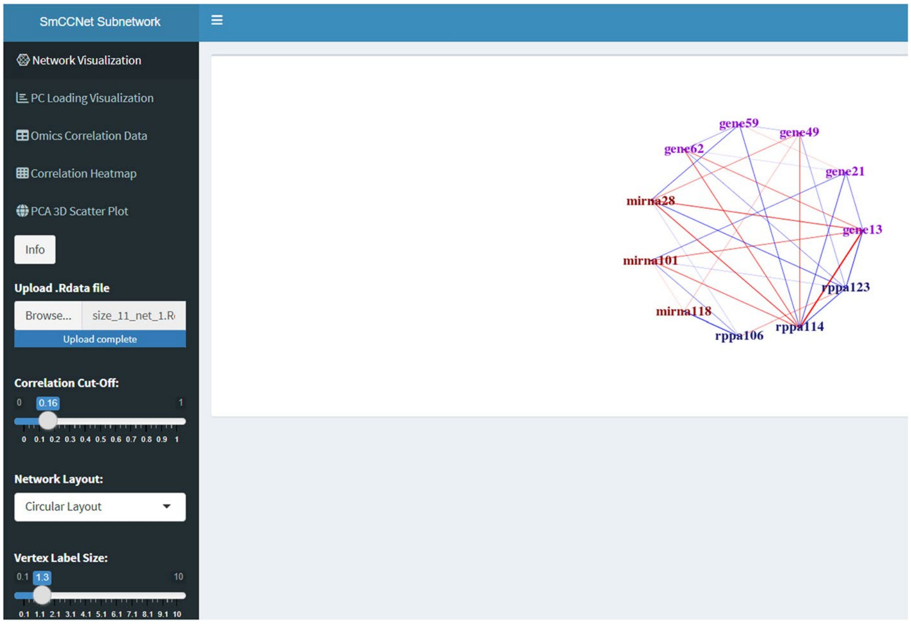


Fig. 5 Network Pruning Algorithm Conceptual Figure. Conceptual figure of network pruning algorithm with the y-axis to be NetSHy/PCA summarization score's correlation to phenotype (black) or baseline network at m_1 (red). m_* is the network size with the highest correlation to phenotype. x ($x > m_*$) is the maximum network size that has a least 0.8 correlation to the baseline network at m_1. m_{opt} corresponds to the optimal network size

partition molecular features into different network modules, and this is achieved by using the function *getAbar()*.

The objective of Step V is to prune the network by removing features (nodes) that have no/little contribution to the subnetwork using a network summarization score of Principal Component Analysis (PCA) [20] or network summarization via a hybrid approach leveraging topological properties (NetSHy) [21] to produce a densely connected pruned subnetwork that maintains a high summarization correlation with respect to the phenotype (Fig. 5). Initially, the network features are ranked based on their PageRank scores [22]. Beginning with a user-defined minimum baseline network size, the method iteratively includes additional features, evaluating the summarization correlation with respect to both the phenotype and the baseline network at each step until reaching the optimal subnetwork size. The network pruning step is achieved by implementing the function *networkPruning()*, and the step-by-step description is given as follows:

- Calculate PageRank score for all molecular features in the unpruned network and rank them according to PageRank score.
- Start from minimally possible network size m_1, iterate the following steps until reaching the maximally possible network size m_2 (defined by users):
  - Add one more molecular feature into the network based on node ranking, then calculate NetSHy/PCA summarization score (PC1, PC2, PC3) for this updated network.
  - Calculate the correlation between this network summarization score and phenotype for the current network size $i \in [m_1, m_2]$, and only use the PC with the highest correlation (determined by absolute value) w.r.t. phenotype, define this correlation as $\rho_{(i,pheno)}$.



Upload .Rdata file: size_11_net_1.R

Correlation Cut-Off: 0.16

Network Layout: Circular Layout

Vertex Label Size: 1.3

Fig. 6 Example R Shiny Interface for Network Information Visualization. The example interface for network visualization. The users can upload network *.Rdata* file to the application, and tune the visualization parameters to obtain the optimal visual representation of the -omics network. Other network-relevant information (PC loading bar graph, correlation heatmap, 3-D subject plot, feature-phenotype correlation table) can also be visualized using this application. Example figures can be found in Sect.

- Identify network size m_* ($m_* \in [m_1, m_2]$) with $\rho_{(m_*, pheno)}$ being the maximally possible summarization score correlation w.r.t. phenotype (determined by absolute value).
- Treat m_* as the new baseline network size, let $\rho_{(m_*, i)}$ be the correlation of summarization score between network with size m_* and network with size i. Define x to be the network size ($x \in [m_*, m_2]$), such that $x = \max\{i|(i \in [m_*, m_2])\&(|\rho_{(m_*, i)}| > 0.8)\}$.
- Between network size of m and x, the optimal network size m_{opt} is defined to be the maximum network size such that $|\rho_{m_{(opt,pheno)}}| \geq 0.9 \cdot |\rho_{(m,pheno)}|$.

Network visualization

The SmCCNet pipeline saves the final subnetwork information in a `.Rdata` file, which does not include data for network visualization. To enable the translation of this `.Rdata` file into a visual representation of the network, we have developed an R Shiny application, accessible at https://smccnet.shinyapps.io/smccnetnetwork/ (Fig. 6). This application provides a user-friendly platform for visualizing single or multi-omics networks, utilizing subnetworks created and stored by SmCCNet. Users can obtain visualizations simply by uploading a.Rdata file with the naming convention 'size_a_net_b.Rdata', where 'a' represents the pruned network size, and 'b' indicates the network module index following hierarchical clustering.

To refine the network visualization, the application offers several adjustable parameters. The Correlation Cut-Off slider allows users to filter network edges based on correlation values, enabling a focus on stronger connections. The Network Layout drop-down menu presents different layout options, which facilitates the selection of the preferred visual arrangement for the network. Users can also adjust the sizes of vertex labels and vertices through the respective Vertex Label Size and Vertex Size sliders. Moreover, the Edge Intensity slider provides control over the color intensity and width of the edges. After adjusting these parameters to their satisfaction, users can generate the network visualization by clicking the ’Plot Network’ button. The ’Download Plot’ button enables the download of the network visualization as a PDF.

Additionally, this application also enables the demonstration of (1) the correlation matrix heatmap between network features; (2) the visualization of PC loadings for the first 3 PCs; (3) The 3-D graph visualizing the distribution of subjects with respect to the first 3 PCs, which serves as a quality-check method to provide some inferences on network-phenotype association; and (4) the feature-phenotype correlation table can also be shown in the application (see Fig. 6).

The application is optimally designed for visualizing final subnetworks of a relatively small size (e.g., $<$ 100 nodes). For larger networks, manual adjustments, such as moving nodes to prevent label overlap, are often necessary. In these instances, we recommend users employ Cytoscape [23] for network visualization. Communication between R and Cytoscape is facilitated by the `RCy3` package [24].

Automated SmCCNet

In this version of the SmCCNet package, we introduce a pipeline known as Automated SmCCNet, which can be implemented with *fastAutoSmCCNet()*. This method streamlines the SmCCNet code and substantially reduces computation time. Users are simply required to input a list of omics data and a phenotype variable. The program then automatically determines whether it is dealing with a single-omics or multi-omics problem, and whether to use CCA or PLS for quantitative or binary phenotypes respectively. For details of how each method is established and how parameters and coefficients are set, we recommend the user to refer to the multi-omics and single-omics vignettes.

Specifically, for multi-omics SmCCNet, if CCA is employed, the program can automatically select the scaling factors (importance of the pair-wise omics or omics-phenotype correlations to the objective function). This is achieved by calculating the pairwise canonical correlation between each pair of omics under the most stringent penalty parameters. The scaling factor for the omics data A, B pair in SmCCA is set to the absolute value of the pairwise canonical correlation between omics A and B divided by the between-omics correlation shrinkage parameter. By default, all scaling factors linked to the phenotype-specific correlation structure are set to 1. In Automated SmCCNet, users only need to provide a BetweenShrinkage parameter, a positive real number that helps reduce the significance of the omics-omics correlation component. The larger this number, the more the between-omics correlation is shrunk.

Moreover, for multi-omics SmCCNet with a binary phenotype, the scaling factor is not implemented. However, the user needs to provide values for γ_1 (omics-omics connection importance) and γ_2 (omics-phenotype connection importance, see multi-omics vignette section 5 for detail). The automated SmCCNet program offers a method to calculate γ_1 while setting the value of γ_2 to 1. This is generally done by averaging all the pairwise omics-omics canonical correlations in the multi-omics dataset.

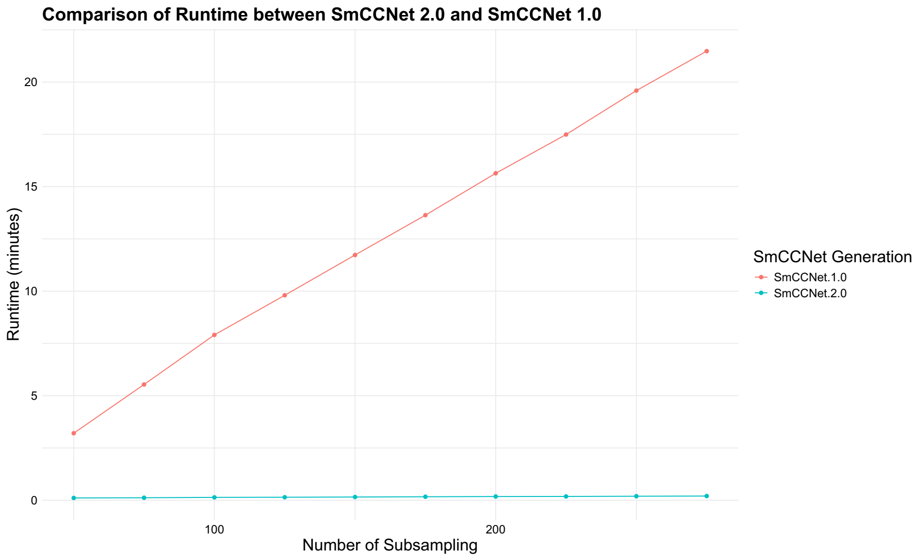


Fig. 7 SmCCNet 2.0 Runtime Performance Compared to SmCCNet 1.0. The graphs shows the x-axis is the number of subsamples and the y-axis represents the time in minutes, which demonstrates SmCCNet 2.0 runtime (in minutes) compared to SmCCNet 1.0 with respect to different number of subsampling iterations. 4 simulated datasets are used: 3 omics data with 50 subjects and 100 features with a quantitative phenotype. At each iteration of subsampling, 70% of the features are randomly subsampled. 5-fold cross-validation is performed on both methods

The program can also automatically select the percentage of features subsampled. If the number of features from an omics data is less than 300, then the percentage of feature subsampled is set to 0.9, otherwise, it's set to 0.7. The candidate penalty terms range from 0.1 to 0.5 with a step size of 0.1 for single/multi-omics SmCCA, and from 0.5 to 0.9 with a step size of 0.1 for single/multi-omics SPLSDA[2] (for both omics-omics SmCCA step and omics-phenotype classifier, see section 5 in the multi-omics vignette for details).

The automated version of SmCCNet typically offers a computational speed advantage over the standard manual SmCCNet, primarily due to the heuristic selection of scaling factors and the parallelization of the cross-validation step. This parallelization is achieved through the use of a parallelized map function in *furrr* package [25], substantially improving the computational speed.

Computational runtime analysis

SmCCNet 2.0 substantially improves the computational time compared to SmCCNet 1.0, which we demonstrate using simulated three omics datasets, each with 50 subjects and 100 features, and one quantitative phenotype. During the random subsampling phase, 70% of the features are selected. We also evaluate different number of subsampling iterations from 50 to 500 with the step size of 50. SmCCNet runs consistently

[2] Penalty terms in SmCCA is in the opposite direction of the SPLSDA, in SmCCA, a higher value of penalty term implies a less stringent sparsity penalty.

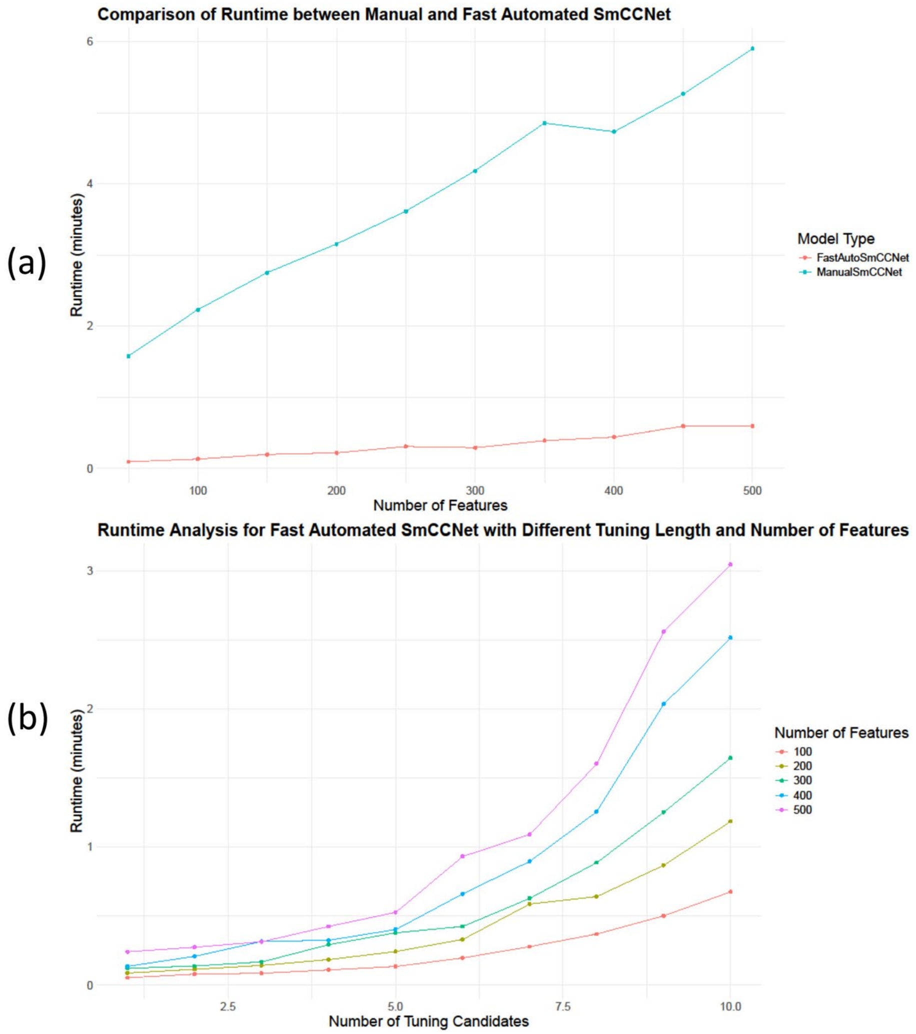


Fig. 8 SmCCNet 2.0 Runtime Analysis. **a** Runtime comparison between Automated SmCCNet and Manual SmCCNet (in minutes) with respect to different number of features in each dataset. 4 simulated datasets are used: 3 omics data with 50 subjects with a quantitative phenotype. For both methods, we use 125 tuning grids for sparsity penalty parameters, and 27 tuning grids for scaling factor parameters. 5-fold cross-validaation is performed on both methods. **b** Runtime analysis of SmCCNet 2.0

faster than SmCCNet 1.0, and the runtime difference increases as the number of subsampling iterations increases (Fig. 7).

Additionally, we also conduct a runtime analysis to compare between automated SmCCNet and manual SmCCNet. Automated SmCCNet runs consistently faster than manual SmCCNet and the runtime difference increases as the number of features in each omics data increases (Fig. 8a). Furthermore, to examine the runtime of SmCCNet under different number of features and sparsity penalty tuning grids, we perform the runtime analysis to evaluate the runtime of automated SmCCNet. As the number of tuning parameter candidates for each omics data increases, the runtime also increases (Fig. 8b). As the number of features in each omics data increases, the runtime increases as well, and the runtime increment is relatively consistent (Fig. 8b).

Example

We demonstrate the second-generation SmCCNet utilizing multi-omics data sourced from the Cancer Genome Atlas Program's (TCGA) project [26] on breast invasive carcinoma (Firehose Legacy). The dataset contains RNA sequencing data with normalized counts, microRNA (miRNA) expression data, and log-ratio normalized reverse phase protein arrays (RPPA) protein data, all procured from tumor samples. After data preprocessing, there are 107 subjects in our final data with 5039 genes, 823 miRNAs, and 175 RPPAs. Furthermore, we regress out age, race, and radiation therapy status from each molecular feature. We provided 2 different examples of using fast automated SmCCNet for multi-omics analysis. Example of a more flexible multi-omics SmCCNet pipeline can be found in package vignette https://cran.r-project.org/web/packages/SmCCNet/vignettes/SmCCNet_Vignette_MultiOmics.pdf. We use survival time as the quantitative phenotype, and survival status as the binary phenotype.

Multi-omics with quantitative phenotype

In the TCGA breast cancer example with a quantitative phenotype (survival time), the analysis can be run with the following code (assuming all X (-omics data list) and Y (survival time) are preprocessed):

```
X <- list(Gene, miRNA, RPPA)
result <- fastAutoSmCCNet(X = X, Y = Y,
                          K = 5, subSampNum = 100,
                          DataType = c(’Gene’, ’miRNA’, ’RPPA’),
                          CutHeight = 0.995,
                          summarization = ’NetSHy’,
                          BetweenShrinkage = 5,
                          seed = 123456)
```

In the first phase of the SmCCNet algorithm, 5-fold cross-validation is used to optimize the sparsity penalty for each molecular profile and determine the best scaling factors. We consider the SmCCA penalty parameter for each molecular profile ranging from 0.1 to 0.5, with increments of 0.1, resulting in a total of 125 combinations. The preliminary CCA canonical correlation is 0.960 (gene-miRNA), 0.689 (gene-protein), 0.632 (protein-miRNA), combined with the between-omics shrinkage factor of 5, resulting in the scaling factor of 0.192 (gene-miRNA), 0.138 (gene-protein), 0.126 (protein-miRNA). After the 5-fold cross-validation, the optimal penalty parameters for molecular profiles are determined to be 0.1 (gene), 0.2 (miRNA), and 0.5 (protein), yielding a total test canonical correlation of 0.799 (normalized scaling factors such that they sum up to 1), and the scaled prediction error of 0.521.

Following this, the complete SmCCNet algorithm is applied with the identified parameters. A subsampling scheme is utilized, selecting 70% of features per iteration for genes and miRNAs and 90% for proteins with 100 iterations to construct the global similarity matrix. Hierarchical clustering with a cut height of 0.995 and a network pruning algorithm set to retain networks between 10 and 100 nodes in size are used to extract the

Table 1 Summary of final subnetwork information for survival time, with information of network index, network size, highest NetSHy score correlation to survival time, number of genes, number of miRNAs, and number of proteins

Network Index	Network Size	PC Correlation to Phenotype	Number of Gene	Number of miRNA	Number of Protein
1	21	0.23509	14	7	0
2	98	0.24307	32	26	40
3	19	0.30108	5	11	3
4	60	0.16220	27	20	13
5	12	0.18218	5	3	4

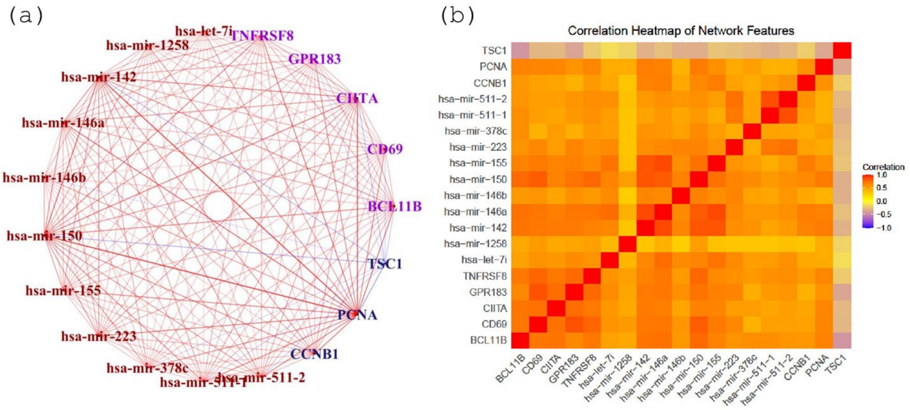


Fig. 9 Multi-omics SmCCNet Result for Survival Time. Multi-omics SmCCNet subnetwork result for TCGA breast cancer data with respect to patient's survival time (subnetwork 3). (**a**): Multi-omics network with respect to survival time. Purple nodes are genes, brown nodes are miRNAs, and dark blue nodes are proteins. Red edges represent positive association between two nodes, and negative edges represent negative association between two nodes. The color depth and edge width represent the strength of association between two nodes (edges are filtered based on a Pearson's correlation threshold of 0.3). (**b**): the correlation heatmap between all subnetwork molecular features

final network modules. The robustness and relevance of the networks are summarized using the NetSHy network summarization score.

After executing the SmCCNet algorithm, we identified five final multi-omics subnetworks (Table 1). Among these, network module 3 demonstrated the strongest association with survival time. Network analysis aims to uncover potential mechanistic insights into the biology of omics data and interpret their relationships with specific phenotypes. Furthermore, it seeks to identify master regulators, which could serve as potential therapeutic targets. SmCCNet plays a pivotal role in achieving these objectives by generating subnetwork results that provide various forms of output. Specifically, SmCCNet formulates hypotheses based on the omics data provided, which can then be validated through existing literature or explored in future research. As an example interpretation, visualization of network module 3 (Fig. 9a) through the Shiny application revealed a hub structure centered on the protein *PCNA*, which has strong connections to the miRNAs *miR-150*. *PCNA* has been studied as a potential biomarker for breast cancer [27], while *miR-150* is known to promote breast cancer growth by targeting the pro-apoptotic purinergic receptor [28]. Despite this, the interaction between *PCNA* and *miR-150* in breast cancer has been minimally studied, leading to a potential area for future research.

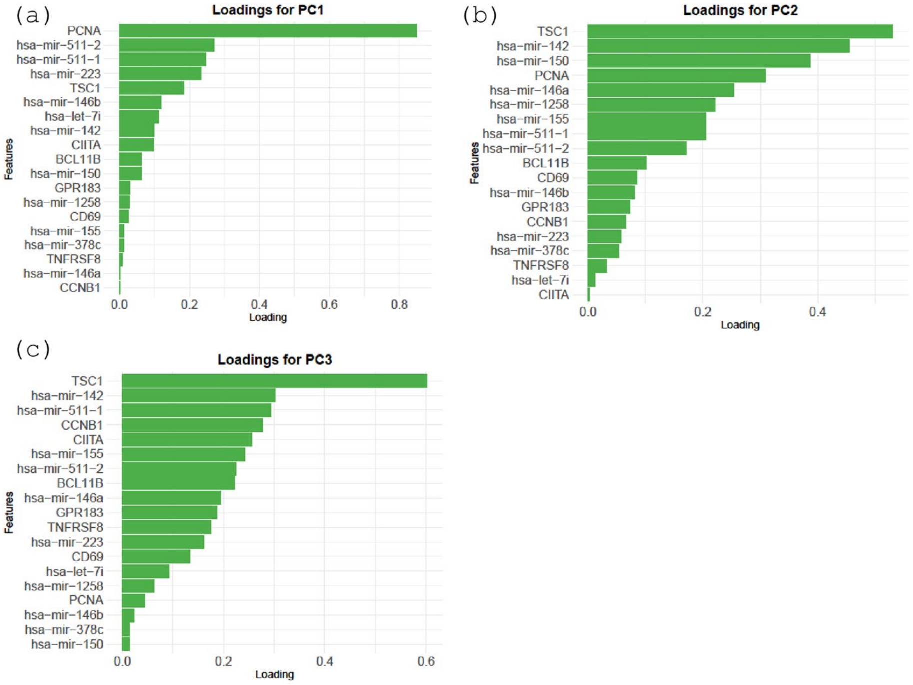


Fig. 10 Final Subnetwork NetSHy Loadings for Survival Time. The NetSHy sumamrization loadings of all the final subnetwork features based on subnetwork 3. with panel **a**, **b**, and **c** represent PC1, PC2, and PC3 respectively

Additionally, the correlation heatmap (Fig. 9b) indicates strong correlations among molecular features in network module 3, which indicates the efficacy of our hierarchical clustering algorithm in grouping highly correlated molecular features that are significantly associated with the phenotype of interest. Notably, the heatmap reveals an almost perfect correlation among *miR-150*, *miR-142*, and *miR-146a*, which are closely connected to *PCNA*. This connection hints at a possible immune-related pathway involving *miR-150* and *miR-146a*, particularly their time-dependent relationship with T-cell differentiation [29].

The NetSHy loading plots (Fig. 10a-c) reveal that network connections oriented around *PCNA* predominantly influence the first principal component (PC), while *TSC1*-oriented connections play a major role in both the second and third PCs. Notably, the third PC (PC3) exhibits the highest correlation with survival time, with a correlation coefficient (ρ) of -0.301. Intriguingly, *TSC1* by itself shows a relatively modest correlation with survival time ($\rho = -0.119$). However, its network connections with other molecular features enhance this association threefold. This implies the importance of further investigating the interactions between *TSC1* and other molecular features within the network, such as *miR-142*, to better understand breast cancer's biological mechanism.

Multi-omics with binary phenotype

In the TCGA breast cancer example with a binary phenotype (survival status), the analysis can be run with the following code (assuming all X (-omics data list) and Y (survival status) are preprocessed):

```
X <- list(Gene, miRNA, RPPA)
result <- fastAutoSmCCNet(X = X, Y = as.factor(Y),
                          K = 5, subSampNum = 100,
                          DataType = c('Gene', 'miRNA', 'RPPA'),
                          CutHeight = 0.995,
                          summarization = 'NetSHy',
                          BetweenShrinkage = 5,
                          seed = 123456,
                          EvalMethod = 'auc',
                          )
```

In the first phase of the SmCCNet algorithm, 5-fold cross-validation is used to optimize the sparsity penalty for each molecular profile and determine the best scaling factors. We consider the SmCCA penalty parameter for each molecular profile ranging from 0.5 to 0.9, with increments of 0.1, and the SPLSDA penalty parameter ranging from 0.5 to 0.9, with increments of 0.1 as well, resulting in 625 combinations. AUC score is used to identify the optimal penalty parameters. Same as before, the preliminary CCA canonical correlation between -omics data is 0.960 (gene-miRNA), 0.689 (gene-protein), 0.632 (protein-miRNA), combined with the between-omics shrinkage factor of 5, resulting in the scaling factor of 0.192 (gene-miRNA), 0.138 (gene-protein), 0.126 (protein-miRNA) for the SmCCA step (exclude phenotype). The relative importance of the between-omics relationship and the omics-phenotype relationship is set to 5, meaning that the omics-phenotype relationship in the model is 5 times as important as the between-omics relationship in the network construction step. After the 5-fold cross-validation, the optimal penalty parameters for SmCCA are determined to be 0.5 (gene), 0.7 (miRNA), and 0.5 (protein); and the optimal penalty parameter for SPLSDA is 0.9, yielding validation AUC score of 0.709. SmCCNet is a network inference pipeline that balances the trade-off between omics-phenotype association and omics-omics association. As a result, the AUC score serves as a quality check for the final networks. If the importance of the omics-phenotype association is emphasized (increase between-shrinkage factor in *fastAutoSmCCNet()*), the AUC score will increase; conversely, if the omics-omics association is prioritized, the AUC score will decrease, but a stronger association between molecular features will be observed. An AUC score of 0.709 indicates a good predictive performance, while still effectively capturing potential biological interactions with respect to breast cancer survival status in the resulting networks.

Similarly, as in the quantitative phenotype example, the complete SmCCNet algorithm is applied with the optimal parameters. A subsampling scheme is utilized, selecting 70% of features per iteration for 100 iterations to construct the global similarity matrix. Hierarchical clustering with a cut height of 0.995 and a network pruning algorithm set to retain networks between 10 and 100 nodes in size are used to extract the final network modules. The robustness and relevance of the networks are summarized using the NetSHy network summarization score.

After executing the SmCCNet algorithm, we identified four final multi-omics subnetworks (Table 2). Among these, network module 3 exhibited the strongest association with survival time. Network analysis aims to uncover potential mechanistic

Table 2 Summary of final subnetwork information for survival status, with information of network index, network size, highest NetSHy score correlation to survival time, number of genes, number of miRNAs, and number of proteins

Network Index	Network Size	PC Correlation to Phenotype	Number of Gene	Number of miRNA	Number of Protein
1	13	0.33536	11	2	0
2	17	0.42283	9	7	1
3	13	0.43232	9	4	0
4	14	0.38436	10	4	0

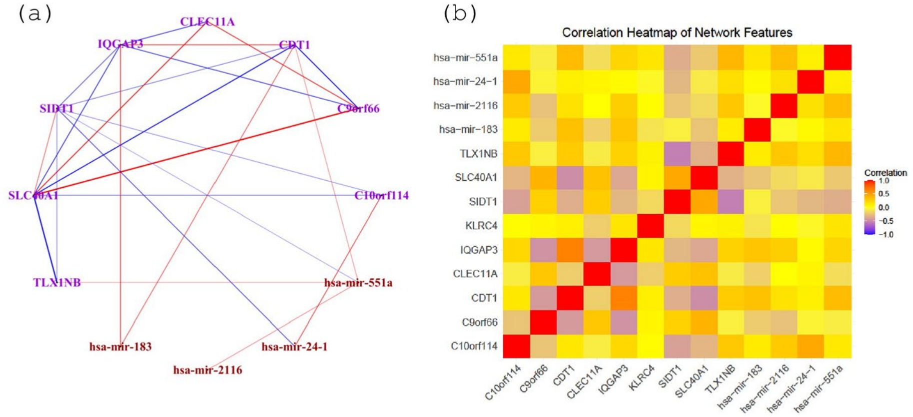


Fig. 11 Multi-omics SmCCNet Result for Survival Status. Multi-omics SmCCNet subnetwork results for TCGA breast cancer data with respect to patient's survival status (subnetwork 3). **a**: Multi-omics network with respect to survival time. Purple nodes are genes, brown nodes are miRNAs, and dark blue nodes are proteins. Red edges represent positive association between two nodes, and negative edges represent negative association between two nodes. The color depth and edge width represent the strength of association between two nodes (edges are filtered based on a Pearson's correlation threshold of 0.3, after edge filtering, not all network nodes are presented in the network); **b**: the correlation heatmap between all subnetwork molecular features

insights into the biology of omics data and interpret their relationships with specific phenotypes. Furthermore, it seeks to identify master regulators, which could serve as potential therapeutic targets. SmCCNet plays a pivotal role in achieving these objectives by generating subnetwork results that provide various forms of output. Specifically, SmCCNet formulates hypotheses based on the omics data provided, which can then be validated through existing literature or explored in future research. As an example interpretation, visualization of network module 3 (Fig. 11a) through our Shiny application shows that there is less network connectivity after edge filtering based on Pearson's correlation (threshold $= 0.3$). After edge filtering, *SLC40A1* has a relatively higher network connectivity. A study has shown that malignant breast cancer cells modulate their iron metabolism by downregulating the iron exporter gene *SLC40A1* to accommodate their high demand for iron [30]. Additionally, the correlation heatmap (Fig. 11b) has a weaker signal than the survival time network, but still demonstrates some high correlation between network molecular features. For instance, there is a strong negative correlation between *TLX1NB* and *SIDT1*. While there is no established study confirming the biological association between *TLX1NB*

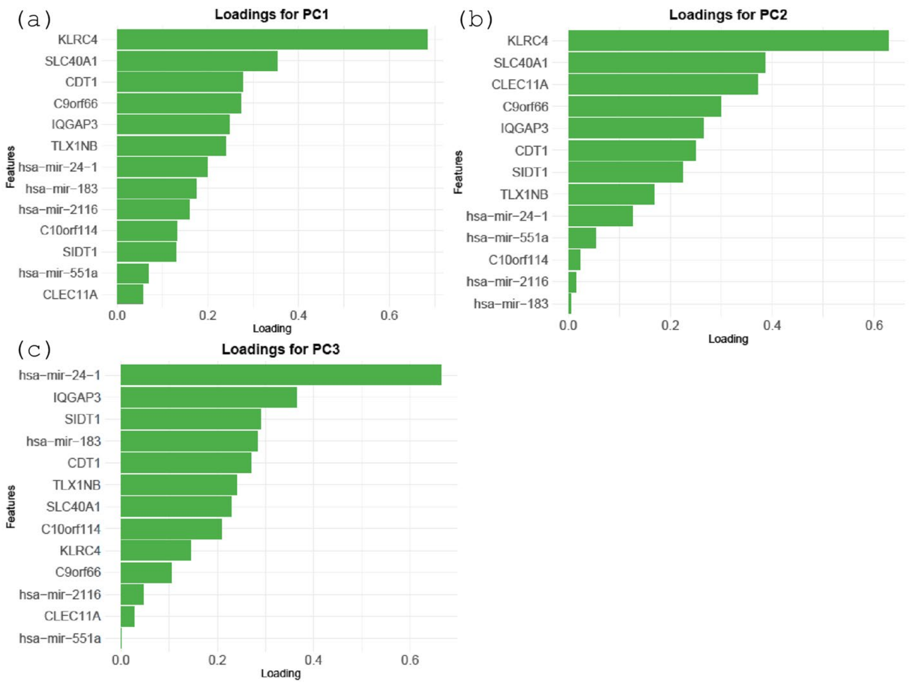


Fig. 12 Final Subnetwork NetSHy Loadings for Survival Status. The NetSHy sumamrization loadings of all the final subnetwork features based on subnetwork 3. with panel **a**, **b**, and **c** represent PC1, PC2, and PC3 respectively

and *SIDT1* in the context of breast cancer, future studies can be conducted to validate their association.

The NetSHy loading plots (Fig. 12a–c) reveal that network connections oriented around *KLRC4* predominantly influence the first and the second principal component (PC), while *miR-24–1*-oriented connections play a major role in both the third PCs. Interestingly, *KLRC4* is filtered out after the stringent Pearson's correlation edge filtering as shown in Fig. 11a, suggesting that it does not have strong interactions with other network molecular features, but it is still influencing the network in other ways. *KLRC4* is associated with a stronger immune response in breast cancer, which correlates with good breast cancer prognosis, highlighting its potential importance in breast cancer targeted immunotherapy treatments [31]. Additionally, the first PC (PC1) exhibits the highest correlation with survival status, with a biserial correlation coefficient (ρ) of −0.432. Interestingly, The highest individual feature-phenotype correlation is only 0.299, which suggests that adding network interactions improves the association with patients' survival status compared to individual molecular features.

Conclusion

The second-generation SmCCNet is a powerful and comprehensive tool for multi-omics network inference with respect to a quantitative or binary variable (e.g., an exposure or phenotype for a complex disease). This upgraded tool incorporates numerous new features including generalization to single or multi-omics data, a novel algorithm for single/multi-omics data with binary phenotype, an automated pipeline to streamline the algorithm with a single line of code, a network pruning algorithm, a topology-based network summarization method, a new network visualization tool, and much more. Additionally, compared to the first-generation SmCCNet, this new version substantially reduces the computational time, and the end-to-end pipeline can be set up easily with either a manual form for more specific parameter control, or through the new automated version. In the TCGA breast cancer data example, we demonstrated how final subnetworks can be obtained using SmCCNet and the Shiny application. We also provide examples of how to interpret the final subnetworks. Depending on the type of multi-omics data supplied, additional interpretation methods such as enrichment / overrepresentation analysis, network mediation analysis, or genome-wide association study (GWAS) can be conducted, and some recent multi-omics study with SmCCNet have demonstrated how SmCCNet subnetwork results can be interpreted in these ways [13, 14, 18]. In the future, more features such as time-to-event data and longitudinal data will be incorporated into the pipeline.

Acknowledgements
We express our sincere gratitude to Wen (Jenny) Shi and Laura M Saba for their contributions to the development of the first-generation of SmCCNet.

Author Contributions
W.L. created the software, wrote the main manuscript text, and prepared all figures and tables; T.V. provided the network summarization algorithm of the software and provided feedback; I.K., K.P., and Y.Z. tested and validated the software and provided feedback; K.K. supervised the project provided guidelines and ideas to the software and manuscript and edited the software and manuscript. All authors reviewed the manuscript.

Funding
This work is supported in part by funds from the National Heart, Lung, and Blood Institute, National Instites of Health (R01 HL152735, TransOmics for Precision Medicines Fellowship: https://topmed.nhlbi.nih.gov/awards/15744).

Availability of data and materials
The TCGA breast cancer data used in example section is available at: http://linkedomics.org/data_download/TCGA-BRCA/. Data are extracted and available along with all subnetwork results, figures and R scripts at: https://github.com/liux4283/SmCCNet2.0Result. All results are obtained based on SmCCNet version 2.0.1.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Conflict of interest
No Conflict of interest is declared.

Received: 7 April 2024 Accepted: 14 August 2024
Published online: 24 August 2024

References

1. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:1–15.
2. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):8124.
3. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21:1–17.
4. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data. Front Genet. 2019;10:535.
5. Henao JD, Lauber M, Azevedo M, Grekova A, Theis F, List M, Ogris C, Schubert B. Multi-omics regulatory network inference in the presence of missing data. Brief Bioinform. 2023;24(5):309.
6. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, Lê Cao K-A. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055–62.
7. Hotelling H. Relations between two sets of variates. In: Breakthroughs in Statistics: Methodology and Distribution, pp. 162–190. Springer, 1992.
8. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology 2009;8(1).

9. Jiang MZ, Aguet F, Ardlie K, Chen J, Cornell E, Cruz D, Durda P, Gabriel SB, Gerszten RE, Guo X, et al. Canonical correlation analysis for multi-omics: application to cross-cohort analysis. PLoS Genet. 2023;19(5):1010517.
10. Rodosthenous T, Shahrezaei V, Evangelou M. Integrating multi-omics data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. Bioinformatics. 2020;36(17):4616–25.
11. Moon S, Hwang J, Lee H. Sdgcca: supervised deep generalized canonical correlation analysis for multi-omics integration. J Comput Biol. 2022;29(8):892–907.
12. Shi WJ, Zhuang Y, Russell PH, Hobbs BD, Parker MM, Castaldi PJ, Rudra P, Vestal B, Hersh CP, Saba LM, et al. Unsupervised discovery of phenotype-specific multi-omics networks. Bioinformatics. 2019;35(21):4336–43.
13. Mastej E, Gillenwater L, Zhuang Y, Pratte KA, Bowler RP, Kechris K. Identifying protein-metabolite networks associated with copd phenotypes. Metabolites. 2020;10(4):124.
14. Zhuang Y, Hobbs BD, Hersh CP, Kechris K. Identifying miRNA-mRNA networks associated with COPD phenotypes. Front Genet. 2021;12: 748356.
15. Graham BI, Harris JK, Zemanick ET, Wagner BD. Integrating airway microbiome and blood proteomics data to identify multi-omic networks associated with response to pulmonary infection. The microbe. 2023;1: 100023.
16. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol. 1996;58(1):267–88.
17. Chung D, Keles S. Sparse partial least squares classification for high dimensional data. Stat Appl Gene Mol Biol 2010;9(1).
18. Konigsberg IR, Vu T, Liu W, Litkowski EM, Pratte KA, Vargas LB, Gilmore N, Abdel-Hafiz M, Manichaikul AW, Cho M, et al. Proteomic networks and related genetic variants associated with smoking and chronic obstructive pulmonary disease. medRxiv, 2024–02 2024.
19. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(1):86–97.
20. Abdi H, Williams LJ. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. 2010;2(4):433–59.
21. Vu T, Litkowski EM, Liu W, Pratte KA, Lange L, Bowler RP, Banaei-Kashani F, Kechris KJ. Netshy: network summarization via a hybrid approach leveraging topological properties. Bioinformatics. 2023;39(1):818.
22. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: bring order to the web. Technical report, Technical report, stanford University 1998.
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
24. Gustavsen JA, Pai S, Isserlin R, Demchak B, Pico AR. Rcy3: Network biology using cytoscape from within r. F1000Research 2019;8.
25. Vaughan D, Dancho M. Furrr: apply mapping functions in parallel using futures. R package version 0.1. 0 2018.
26. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):71–71.
27. Malkas LH, Herbert BS, Abdel-Aziz W, Dobrolecki LE, Liu Y, Agarwal B, Hoelz D, Badve S, Schnaper L, Arnold RJ, et al. A cancer-associated PCNA expressed in breast cancer has implications as a potential biomarker. Proc Natl Acad Sci. 2006;103(51):19472–7.
28. Huang S, Chen Y, Wu W, Ouyang N, Chen J, Li H, Liu X, Su F, Lin L, Yao Y. miR-150 promotes human breast cancer growth and malignant behavior by targeting the pro-apoptotic purinergic p2x7 receptor. PLoS ONE. 2013;8(12):80707.
29. Gan L, Sun T, Li B, Tian J, Zhang J, Chen X, Zhong J, Yang X, Li Q. Serum miR-146a and miR-150 as potential new biomarkers for hip fracture-induced acute lung injury. Mediators Inflamm. 2018;2018(1):8101359.
30. Jiang XP, Elliott RL, Head JF. Manipulation of iron transporter genes results in the suppression of human and mouse mammary adenocarcinomas. Anticancer Res. 2010;30(3):759–65.
31. Tan W, Liu M, Wang L, Guo Y, Wei C, Zhang S, Luo C, Liu N. Novel immune-related genes in the tumor microenvironment with prognostic value in breast cancer. BMC Cancer. 2021;21:1–16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.