SmCCNet 2.0: an Upgraded R package for Multi-omics Network Inference with Automated Pipeline, 2022, pp. 1–5

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

SmCCNet 2.0: an Upgraded R package for Multi-omics Network Inference

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

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

[*]Corresponding author. weixuan.liu@cuanschutz.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

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 in `https://smccnet.shinyapps.io/smccnetnetwork/`.

Key words: multi-omics integration, network analysis, automated pipeline

Introduction

The goal of multi-omics network inference is to integrate multiple omics data sets (e.g., transcriptomic, proteomic) to infer molecular associations and gain insights on associated biological processes (Hawe et al., 2019). Recent methods include knowledge-guided multi-omics network inference (KiMONo) (Henao et al., 2023) and Biomarker discovery using Latent Variable Approaches for Omics Studies (DIABLO) (Singh et al., 2019).

Sparse multiple Canonical Correlation Network Analysis (SmCCNet) is a canonical correlation-based integration method that reconstructs phenotype-specific multi-omics networks (Shi et al., 2019). It has been applied to different multi-omics integration tasks such as extracting protein-metabolite networks (Mastej et al., 2020) and mRNA-miRNA networks (Zhuang et al., 2021) associated with disease phenotypes. The algorithm is based on a sparse multiple canonical analysis (SmCCA) (Witten and Tibshirani, 2009) for T omics 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) (Tibshirani, 1996). 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 $i, j = 1, 2, ..., T$.

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

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.

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

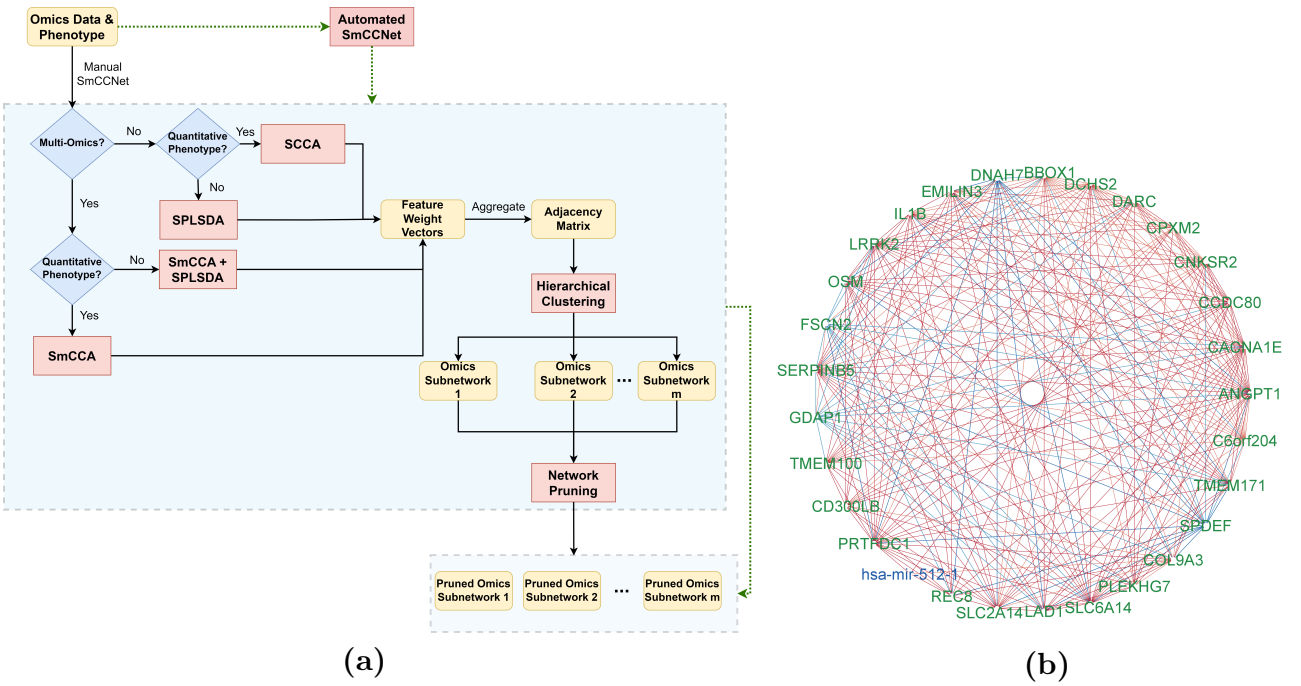

Fig. 1. SmCCNet workflow (a) the workflow of the second-generation SmCCNet, which takes in single-/multi-omics data with quantitative/binary phenotype and outputs multiple pruned omics subnetwork modules. Steps in the box represent the stepwise process of executing the SmCCNet, and if automated SmCCNet is used, all these steps will be executed automatically. (b) an example of a multi-omics network of TCGA breast cancer data visualized in Cytoscape with the green node representing genes, and the blue node representing miRNA. The strength of the node connections is indicated by the thickness of the edges. Red edges and blue edges are for negative and positive connections, respectively. If more than two types of omics features are presented, there would be more than two colors for nodes

The first version of SmCCNet could only analyze two omics data types with a quantitative phenotype and required a complicated manual setup. To enhance the practical utility of SmCCNet, we have upgraded the software 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. Other improvements include a data preprocessing pipeline, improved computational efficiency, an online RShiny application for network visualization, and the storage of accessible and reproducible network analysis results in a user-specified directory.

Method 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 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 multi-omics subnetwork modules that are associated with the phenotype. The general workflow of SmCCNet is as follows (for a quantitative phenotype):

- Step I: Determine SmCCA sparsity penalties c_t. 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 (Figure 1a). The K-fold CV procedure enhances the robustness of selected penalties when generalizing to similar independent omics data sets.
- Step II: Randomly subsample omics features, apply SmCCA with chosen penalties and compute a feature relationship matrix for each subsample. Repeat the process many times and define the similarity matrix to be the average of all feature relationship matrices.
- Step III: Apply hierarchical tree cutting to the similarity matrix to simultaneously identify multiple subnetworks.
- Step IV: Apply network pruning algorithm to prune each subnetwork obtained from Step III.
- Step V: Visualize the omics network with an RShiny application (`https://smccnet.shinyapps.io/smccnetnetwork/`) or Cytoscape.

Number of Omics Data and Phenotype Modality

While the stepwise workflow above uses multi-omics with quantitative phenotype as an illustration, it can be adapted to a single-omics setting or binary phenotype setting. When these settings are presented, the only change in the workflow is the choice of model in Steps I and II: (1) when single-omics data is used with a quantitative phenotype, then sparse canonical correlation analysis (SCCA) is used to construct the global network by using the function *getRobustWeightsSingle()*; (2) when single-omics data is used with a binary phenotype, then sparse partial least squared discriminant analysis (SPLSDA) (Chung and Keles, 2010) is implemented to construct the global network by using the function *getRobustWeightsSingleBinary()*; (3) when multi-omics data is used with a quantitative phenotype, then regular SmCCA is used to construct the global network by using the

function *getRobustWeightsMulti()*; (4) when multi-omics data is used with a binary phenotype, then a hybrid of SmCCA and SPLSDA is used to construct the global network by using the function *getRobustWeightsMultiBinary()*.

In particular, a novel stepwise hybrid approach is developed for multi-omics data with a binary phenotype by (1) filtering molecular features using SmCCA (no phenotype involved) to identify interconnected molecular features, then (2) implementing SPLSDA on selected features from all molecular profiles to find molecular features that are associated with the phenotype, and (3) the canonical weight from SmCCA and feature importance weight are aggregated by taking the weighted average. Multiple binary classification metrics (accuracy, ROC-AUC score, precision, recall, and F1 score) can be used to evaluate the model and select the optimal penalty parameters for both SmCCA and SPLSDA (see Section 5 of the package vignette [1].

Scaling Factor Determination

The scaling factors ($a_{i,j}$ and b_j) in Eqn 1 can be supplied to prioritize the correlation structure of interest in Steps I and II. We now 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 $(a, b_1, b_2) = c(0.4, 1, 1)$. This method is currently implemented in the automated SmCCNet approach.

Cross-Validation to Select Scaling Factors

The last approach uses cross-validation to find the scaling factors. We use two -omics as an example. Initially, candidate scaling factors are created by setting all omics-omics scaling factors to 1 and varying the omics-phenotype scaling factors, ensuring their sum is 1 for comparability, e.g., the scaling factor $(a_{1,2}, b_1, b_2)$ should satisfy $a_{1,2} + b_1 + b_2 = 1$. A nested grid search is employed to simultaneously optimize scaling factors and penalty parameters. In this framework, while assessing different scaling factors, the optimal penalty parameters are selected. For each distinct set of scaling factors, the sparse penalty parameters ($l1$, $l2$) are determined through K-fold cross-validation, aiming to minimize the scaled prediction error. The minimal scaled prediction error associated with the optimal penalty parameters is then recorded for each set. This procedure is reiterated for all possible sets of scaling factors. The set of scaling factors yielding the lowest scaled prediction error is ultimately selected as the optimal set. As the number of omics data increases, the number of possible combinations for scaling factors increases exponentially. Therefore, for analyses with more omics data (> 3), we recommend using the automated SmCCNet algorithm to select the optimal scaling factors.

Network Clustering and Pruning

As mentioned in Step II, the adjacency matrix is formed by taking the outer product of the canonical weights. After obtaining the adjacency matrix, hierarchical clustering (Murtagh and Contreras, 2012) is implemented to partition molecular features into different network modules, and this is achieved by using the function *getAbar()*.

The objective of Step IV 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) (Abdi and Williams, 2010) or network summarization via a hybrid approach leveraging topological properties (NetSHy) (Vu et al., 2023) to produce a densely connected pruned subnetwork that maintains a high summarization correlation with respect to the phenotype. Initially, the network features are ranked based on their PageRank scores (Page et al., 1998). 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()* (see Section 2.6 of the package vignette [1]).

Automated SmCCNet

We introduce a new Automated SmCCNet approach using the function *fastAutoSmCCNet()*. This method streamlines the SmCCNet code and significantly 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. Specifically, for multi-omics SmCCNet, if CCA is employed, the program can automatically select the scaling factors based on pairwise correlation; if PLS is employed, the scaling factors are determined by setting the trade-off parameters that indicate the relative importance of the between-omics relationship and the omics-phenotype relationship (similar to using pairwise correlation). Details about parameter setup can be found in the automated SmCCNet vignette[2]. This automated version of SmCCNet is faster than the standard SmCCNet due to the heuristic selection of the scaling factors and the parallelization of the

[1] `https://cran.r-project.org/web/packages/SmCCNet/vignettes/SmCCNet_Vignette_MultiOmics.pdf`

[2] `https://cran.r-project.org/web/packages/SmCCNet/vignettes/SmCCNet_Vignette_AutoSmCCNet.pdf`

cross-validation step, resulting in a substantial increase in computational speed.

Example

We demonstrate the second-generation SmCCNet utilizing multi-omics data sourced from the Cancer Genome Atlas Program's (TCGA) project (Colaprico et al., 2016) 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. Tumor purity is used as the quantitative phenotype, which is defined as the percentage of cancer cells within a tumor tissue sample. After data pre-processing, there are 105 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 ran SmCCNet for multi-omics analysis of a quantitative phenotype with 5-fold cross-validation to select the optimal penalty parameters as well as scaling factors. See **Supplementary material** for more details on the parameters and the optimal scaling and penalty factors selected. We visualize network module 4 as an example using Cytoscape (in Figure 1(b)). In this particular network example, there is only one miRNA that connects 28 genes, with each of the red edges representing a positive association between two molecular features and the blue edges representing a negative association. Specifically, our result indicates the association between *hsa-mir-512-1* and *REC8* with a correlation of 0.309. Previous work has suggested that *REC8* overexpression markedly decreased the proliferation, migration, and invasion capabilities of breast cancer cells (He et al., 2022).

Conclusion

The second-generation SmCCNet is a powerful 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 by up to 1000 times, 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.

Acknowledgements

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

Competing interests

No competing interest is declared.

Funding

This work is supported in part by funds from the National Heart, Lung, and Blood Institute, National Institues of Health (R01 HL152735, TransOmics for Precision Medicines Fellowship).

Data Availability

The TCGA breast cancer data used in example section is available at: `http://linkedomics.org/data_download/TCGA-BRCA/`.

References

H. Abdi and L. J. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2 (4):433–459, 2010.

D. Chung and S. Keles. Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016.

J. S. Hawe, F. J. Theis, and M. Heinig. Inferring interaction networks from multi-omics data. *Frontiers in genetics*, 10: 535, 2019.

S. He, D. Liu, and Z. Chen. Rec8 inhibits proliferation, migration and invasion of breast cancer cells by targeting cdc20. *Molecular Medicine Reports*, 26(1):1–11, 2022.

J. D. Henao, M. Lauber, M. Azevedo, A. Grekova, F. Theis, M. List, C. Ogris, and B. Schubert. Multi-omics regulatory network inference in the presence of missing data. *Briefings in Bioinformatics*, 24(5):bbad309, 2023.

E. Mastej, L. Gillenwater, Y. Zhuang, K. A. Pratte, R. P. Bowler, and K. Kechris. Identifying protein–metabolite networks associated with copd phenotypes. *Metabolites*, 10 (4):124, 2020.

F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bring order to the web. Technical report, Technical report, stanford University, 1998.

W. J. Shi, Y. Zhuang, P. H. Russell, B. D. Hobbs, M. M. Parker, P. J. Castaldi, P. Rudra, B. Vestal, C. P. Hersh, L. M. Saba, et al. Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics*, 35(21):4336–4343, 2019.

A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

T. Vu, E. M. Litkowski, W. Liu, K. A. Pratte, L. Lange, R. P. Bowler, F. Banaei-Kashani, and K. J. Kechris. Netshy: network summarization via a hybrid approach leveraging topological properties. *Bioinformatics*, 39(1):btac818, 2023.

D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1), 2009.

Y. Zhuang, B. D. Hobbs, C. P. Hersh, and K. Kechris. Identifying mirna-mrna networks associated with copd phenotypes. *Frontiers in genetics*, 12:748356, 2021.