iModMix: Integrative Module Analysis for Multi-omics Data

Isis Narváez-Bandera[1] (Isis.Narvaez-Bandera@moffitt.org), Ashley Lui[2,3] (Ashley.Lui@moffitt.org), Yonatan Ayalew Mekonnen[2] (YonatanAyalew.Mekonnen@moffitt.org), Vanessa Rubio[2] (Vanessa.Rubio@moffitt.org), Noah Sulman[4] (Noah.Sulman@epi.usf.edu), Christopher Wilson[1] (cmwilson0109@gmail.com), Hayley D. Ackerman[2,3] (Hayley.Ackerman@moffitt.org), Oscar E. Ospina[1] (Oscar.Ospina@moffitt.org), Guillermo Gonzalez-Calderon[1] (Guillermo.Gonzalez-Calderon@moffitt.org), Elsa Flores[2,3] (Elsa.Flores@moffitt.org), Qian Li[5] (qian.li@stjude.org), Ann Chen[6] (Ann.Chen@hci.utah.edu), Brooke Fridley[7]* (blfridley@cmh.edu), Paul Stewart[6,8]* (paul.stewart@hci.utah.edu).

[1]Department of Biostatistics and Bioinformatics, [2]Department of Molecular Oncology, [3]Cancer Biology and Evolution Program, Moffitt Cancer Center; [4]Health Informatics Institute, University of South Florida; [5]Department of Biostatistics, St. Jude Children's Research Hospital; [6]Huntsman Cancer Institute, University of Utah; [7]Division of Health Services and Outcome Research, Children's Mercy Research Institute, Kansas City; [8]Department of Nutrition and Integrative Physiology, University of Utah.
*Corresponding Authors

Abstract

Summary: The integration of metabolomics with other omics ("multi-omics") offers complementary insights into disease biology. However, this integration remains challenging due to the fragmented landscape of current methodologies, which often require programming experience or bioinformatics expertise. Moreover, existing approaches are limited in their ability to accommodate unidentified metabolites, resulting in the exclusion of a significant portion of data from untargeted metabolomics experiments. Here, we introduce *iModMix - Integrative Module Analysis for Multi-omics Data*, a novel approach that uses a graphical lasso to construct network modules for integration and analysis of multi-omics data. *iModMix* uses a horizontal integration strategy, allowing metabolomics data to be analyzed alongside proteomics or transcriptomics to explore complex molecular associations within biological systems. Importantly, it can incorporate both identified and unidentified metabolites, addressing a key limitation of existing methodologies. *iModMix* is available as a user-friendly R Shiny application that requires no programming experience (https://imodmix.moffitt.org), and it includes example data from several publicly available multi-omic studies for exploration. An R package is available for advanced users (https://github.com/biodatalab/iModMix).

Availability and implementation: Shiny application: https://imodmix.moffitt.org. The R package and source code: https://github.com/biodatalab/iModMix.

Introduction

Integration of metabolomics with other omics modalities, such as proteomics and transcriptomics, is essential for a comprehensive understanding of complex biological systems and disease mechanisms. The scale and complexity of these so-called "multi-omics" data necessitate computational approaches for effective integration and analysis. However, multi-omics analysis presents several challenges. These include the need for computational tools that are adaptable to diverse experimental setups and data types, as well as user-friendly enough to be used without extensive programming experience (Jendoubi 2021). Moreover, many approaches are unable to incorporate unidentified metabolites, resulting in the exclusion of large portions of data from untargeted metabolomics experiments, thereby limiting the potential for novel discoveries. Current pathway tools often restrict the inclusion of metabolite identifications due to naming mismatches and inconsistent use of standard identifiers like KEGG or HMDB IDs. This results in exclusion of some identified metabolites which cannot be fully utilized for analysis and further narrowing and hindering comprehensive insights (Jendoubi 2021).

In general, approaches for integrating metabolomics with other omics can be categorized into vertical or horizontal integration. Vertical integration takes multiple omics datasets as input and produces a single output, such as clustering assignments or a merged dataset, which no longer reflects the individual features of the original data. iClusterPlus (Qianxing et al. 2013) is an example of vertical integration: it takes multiple omics datasets from the same samples as input, extracts latent factors across omics datasets, clusters samples in latent space, and outputs the clustering assignments by sample. In contrast, horizontal integration utilizes multiple omics datasets in parallel while maintaining

the context of the original inputs. PIUmet (Benedetti et al. 2023) is an example of horizontal integration: it takes metabolomics and proteomics data as input, and the resulting network output contains both metabolites and proteins.

iModMix is a horizontal integration framework that constructs network modules from two input omics datasets (*e.g.*, metabolomics and proteomics, metabolomics and transcriptomics). It first uses graphical lasso to estimate a sparse Gaussian Graphical Model (GGM) (Friedman, Hastie, and Tibshirani 2008) for each input omics dataset. GGMs capture direct associations within the input omics datasets, which is an improvement over Weighted Gene Correlation Network Analysis (WGCNA) (Langfelder et al. 2008) module creation that includes both direct and indirect associations. Similar to WGCNA, a Topological Overlap Matrix (TOM) is next calculated (Yip et al. 2007), which quantifies the extent to which pairs of features share common neighbors. Hierarchical clustering is then performed on TOM dissimilarity (1 – TOM), using a dynamic tree cutoff to group related features into modules. *iModMix* next takes the first principal component of the abundances of the features in the module, called an eigenfeature (*e.g.*, eigenmetabolites for metabolomics, eigenproteins for proteomics). These eigenfeatures are representative of the contents of the module, and they can be used in place of normal metabolite abundance or protein expression for testing for differential expression or association with experimental conditions. Finally, integration between omics types is achieved by correlating the eigenfeatures from different omics experiments.

A major advantage of iModMix is its ability to generate network modules and their associated eigenfeatures independently of feature annotation. This allows iModMix to incorporate identified and unidentified metabolites into its results. Since this method does not rely on existing pathway databases like KEGG, it can uncover new associations among features. *iModMix* is available as an R Shiny application (https://imodmix.moffitt.org) that provides an easy to use graphical interface and detailed documentation for users without programming expertise, and it is available as an R package (https://github.com/biodatalab/iModMix). *iModMix* can run on standard desktop computers with at least 8GB of RAM and a multi-core processor. The running time depends on the size and complexity of the dataset, but typical data sizes of 20,000 variables, the analysis can be completed in under an hour.

Implementation

Data Upload

The *iModMix* workflow accepts two omics datasets at a time (Fig. 1A). If available, users can provide feature-level annotation corresponding to the input omics dataset. A separate metadata file containing sample IDs and at least one sample grouping column (*e.g.*, treatment, control) are necessary for conducting PCA, heatmaps, and boxplots. Once uploaded, the tool scales the data by centering each feature to have a mean of zero and a standard deviation of one. Missing values are imputed using the k-nearest neighbors' algorithm, or users can also provide their own imputed data if a different method is preferred. We provide ten datasets encompassing gene expression and metabolomics data to illustrate *iModMix's* capabilities. These datasets, from (Benedetti et al. 2023), have been formatted to meet *iModMix* specifications, with variable columns labeled as "Feature_ID" and sample columns labeled as "Samples". These datasets are designed to facilitate user engagement with *iModMix* and are readily accessible for download on Zenodo (https://doi.org/10.5281/zenodo.13988161). We also provide an in-house dataset of 20 lung adenocarcinoma mouse samples (10 wild type, 10 knockout), with 7353 metabolomics features and 7928 protein groups, to highlight *iModMix's* functionality in handling unidentified metabolites.

Module Construction

iModMix first estimates partial correlations using the glassoFast R package (Friedman, Hastie, and Tibshirani 2008) (Fig. 1B) to build a GGM. Next, the TOM is computed and TOM dissimilarity (1 – TOM) is used for hierarchical clustering (Yip et al. 2007). Modules are created using the dynamic tree cut method, specifically the 'hybrid' method, which refines assignments derived from a static cut height by analyzing the shape of each branch (Langfelder, Zhang, and Horvath 2008). Finally, *iModMix* calculates eigenfeatures for each module using the moduleEigengenes function from the WGCNA R package (Langfelder et al. 2008) (Fig. 1B) that are used for downstream omics analyses.

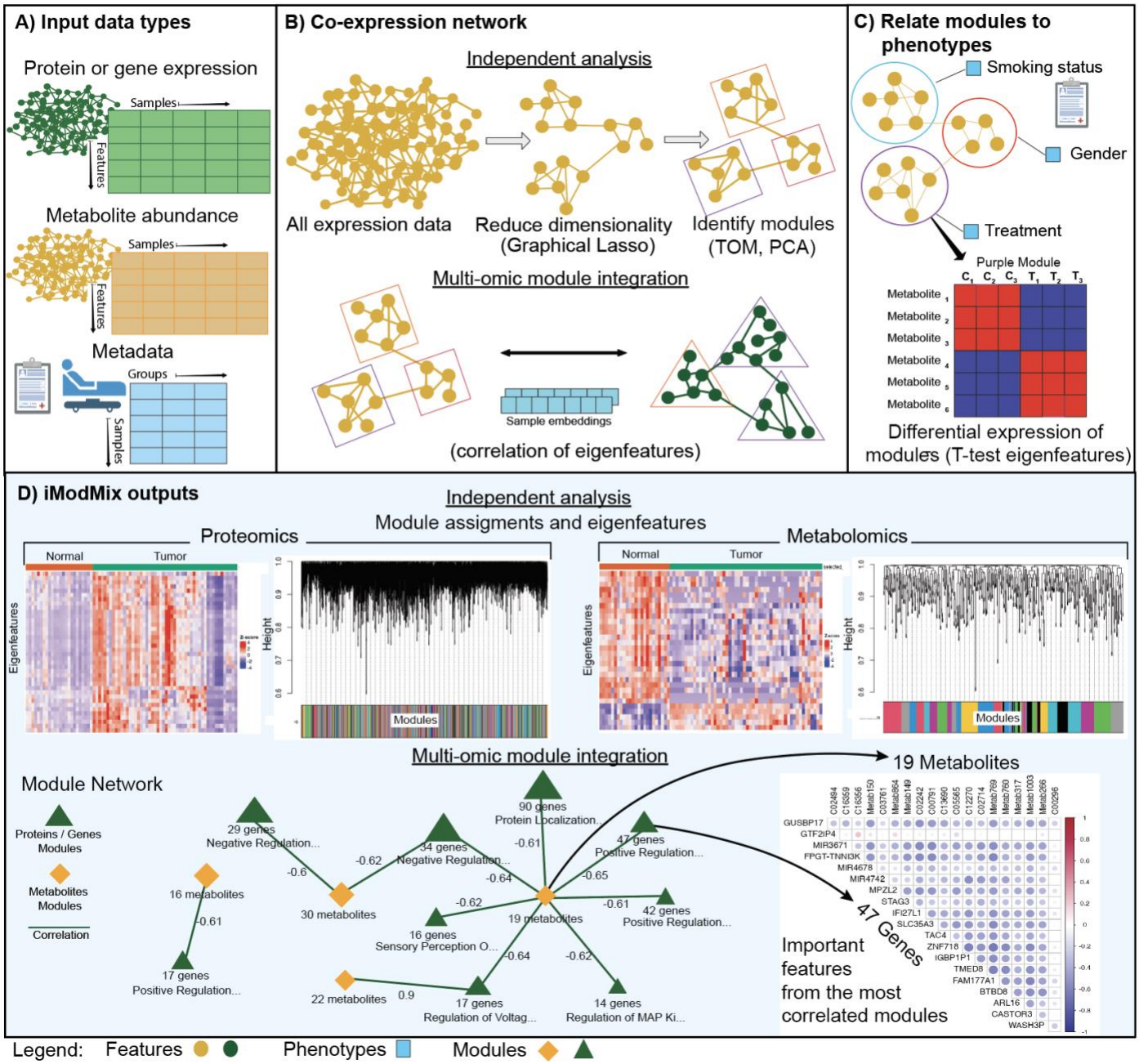

Figure 1: Overview of *iModMix*. (A) *iModMix* integrates transcriptomics, proteomics, metabolomics, and clinical data to understand biological systems comprehensively. (B) Co-expression network: The tool performs independent analyses of each omics dataset to identify biologically relevant modules, and it then integrates the datasets by correlating eigenfeatures. (C) *iModMix* evaluates associations between phenotype-related modules across different data types. (D) Results of the integrated analysis of transcriptomics and metabolomics data from normal and tumor samples using *iModMix*. The network interaction and most correlated modules are also visualized.

Module Exploration

The phenotype analysis section provides a framework for classifying phenotypes based on eigenfeatures. Users can upload metadata, select phenotypes of interest, and set significance thresholds for statistical tests. The module performs classification using Student's t-test, comparing selected phenotypes and generating results that include t-statistics and p-values. Significant eigenfeatures are visualized through boxplots. Additionally, users can explore specific modules and view associated features, PCA loading, and heatmap plots to view feature behavior across different phenotypes (Fig. 1C). *iModMix* supports pathway enrichment analysis of genes or proteins by utilizing all available libraries in Enrichr (Kuleshov et al. 2016). This allows users to choose their preferred library, such as KEGG, GO, Human Gene Atlas, or WikiPathways.

Multi-omics analysis

iModMix integrates modules across omics datasets by calculating the Spearman correlation between eigenfeatures. *iModMix* then constructs an interactive network, where protein/gene modules are represented as green triangles and metabolite modules as yellow diamonds, with correlation coefficients shown on connecting arrows. Users can select the number of top multi-omics module correlations to visualize and explore detailed correlations within these modules through corrplots and tables. Lists of metabolites and proteins/genes within highly correlated modules are provided, facilitating further pathway analysis and offering insights into the relationships between different omics layers. Additionally, classification between features of each layer and phenotypes using t-tests and boxplots is provided, enabling users to visualize and identify significant differences. (Fig. 1B).

Case-Study: ccRCC Dataset with Identified Metabolites

We used 24 normal and 52 tumor clear cell renal cell carcinoma (ccRCC) samples (Golkaram et al. 2022; Tang et al. 2023; Benedetti et al. 2023) as a case study. It contained 23001 genes from RNA-seq and 904 identified metabolites from untargeted metabolomics. Applying *iModMix* identified 751 gene modules and 34 metabolite modules. Differential expression analysis of the modules through t-test confirmed changes in metabolite abundance between groups recapitulated those from prior work (Benedetti et al. 2023) highlighting reduced levels of gamma-glutamyltyrosine (module ME#BA3241, P-value: 0.0003), creatinine/C00791 (module ME#C06162 P-value: 4.2681E-15), xanthosine/C01762 (module ME#66628D P-value: 0.0015), docosahexaenoate (DHA; 22:6n3)/C06429 (module ME#904A67 P-value: 4.6906E-12), and 1-methyladenosine/C02494 (module ME#E1C62F P-value: 8.1847E-14) in tumors compared to normal tissue. Conversely, metabolites with increased abundance in tumors compared to normal tissues included proline/C00148, glutamine/C00064 (module ME#6C856F, P-value: 0.006), maltose/C00208 and 1-methylnicotinamide/C02918 (module ME#904A67 P-value: 4.6906E-12) (Benedetti et al. 2023). The top five most correlated metabolomic and transcriptomic module pairs included pathways such as Inflammatory Response (P-value: 0.0001), Protein Polyubiquitination (P-value: 0.0001), tRNA Aminoacylation For Mitochondrial Protein Translation (P-value: 0.0084), and Sphingosine-1-Phosphate Receptor Signaling Pathway (P-value: 0.0068). The inflammatory response plays a crucial role in the development and prognosis of ccRCC. Research indicates that inflammation is involved at all stages of the disease, influencing tumor progression and treatment responses (Zhong et al. 2023).

Case-Study: Lung Adenocarcinoma Dataset with Unidentified Metabolites

Matched proteomics and metabolomics dataset was generated using two mouse models for lung adenocarcinoma (LUAD) (10 wild type, 10 knockout), with 7353 metabolomics features (identified and unidentified) and 7928 protein groups. Applying *iModMix* generated 412 gene modules, and 287 metabolite modules. Strong correlations were observed between these modules, with correlations as high as 0.93. The most correlated pair (protein module ME#FFDA24 with metabolite module ME#5E985E; correlation of 0.93), included 16 genes (Hsd11b1, Cavin2, Ehd2, Myh10, Clic5, Msn, Myo1c, Epb41l2, Pakap, Sptan1, Akap12, Cyp2b10, Cav1, Cav3, Sptbn1, Ace, Ace3, Cavin1) and 26 metabolites, one of which was identified as gamma-linolenic acid/C06426. Pathway analysis of these 16 genes and the identified metabolite using MetaboAnalyst (Pang et al. 2024) revealed significant enrichment in Proteoglycans in cancer (P-value: 7.9492E-4) and Renin-angiotensin system (P-value: 7.9861E-4). Similarly, the third most correlated modules (protein module ME#FFDD25 and metabolite module ME#FFB214; correlation of 0.93) comprised 16 genes (Myl6, Tmod1, Vsnl1, Ppp1r14a, Aldh1a1, Myl12b, Dlc1, Tgfb1i1, Plscr4, Macf1, Pcdh18, Limch1, Lims1, Ilk, Specc1l, Vcl, Rasip1) and 21 metabolites, seven of which are identified metabolites (Phenylpyruvic acid/C00166, Indoleacetic acid/C00954, Homovanillic/C05582, Hydroxyphenyllactic acid/C03672, 4-Hydroxyphenylpyruvic/C01179, Deoxyuridine/C00526, Thymine/C00178). Pathway analysis using MetaboAnalyst highlighted significant enrichment in Phenylalanine, tyrosine and tryptophan biosynthesis (P-value: 8.0031E-4) and Focal adhesion (P-value: 8.9650E-4). Proteoglycans are critical in LUAD, influencing tumor progression, metastasis, and microenvironment. Their dysregulation impacts essential cellular processes such as epithelial-to-mesenchymal transition and cancer stemness, which are key to cancer cell plasticity and aggressiveness (Karagiorgou et al. 2022). Besides, focal adhesions play a significant role in LUAD progression by mediating cell adhesion, migration, and signaling pathways essential for

tumorigenesis. Studies highlight the involvement of focal adhesion kinase (FAK) in LUAD, suggesting potential therapeutic targets and biomarkers for this cancer type (Zhou et al. 2018).

Benchmarking

We benchmark *iModMix* to identify the optimal parameters for gene network construction using the *iModMix* algorithm. The process considers three input parameters: num_genes, num_fold, and lambda. Expression data from RC20 and lung adenocarcinoma mouse samples were used, and genes with the highest variance were selected, constrained by num_genes. To identify the best sparsity parameter (lambda), we applied five-fold cross-validation using the Graphical Lasso (Glasso) algorithm, which calculates a sparse Gaussian Graphical Model (GGM) for each input omics dataset. The Glasso model was fitted using the glassoFast function with lambda, and metrics such as the number of non-zero partial correlations, log-likelihood, and execution time were evaluated. Hierarchical clustering analysis was then used to identify gene modules. Stability was achieved in both datasets with an alpha value of 0.25 (Fig. S1A, Fig. S1B), allowing for efficient and accurate model evaluation across various configurations. This benchmarking process demonstrates that *IModMix* offers a refined approach to network construction by focusing on direct associations, making it a robust alternative to WGCNA for complex omics data.

Summary:

The iModMix pipeline provides an empirical framework for uncovering novel associations between multi-omics data, with the unique capability to incorporate both identified and unidentified metabolites. By generating de novo network modules independent of pathway databases, iModMix enables the discovery of previously unrecognized interactions between features. Unlike methods that rely on predefined biological functions or annotations, iModMix leverages feature matrices alone, making it highly adaptable to a wide range of omics combinations, including those with limited annotation or poorly characterized interactions. This flexibility highlights its potential for expanding multi-omics integration to novel and challenging datasets, advancing our understanding of complex biological systems.

Acknowledgments

Funding: This work was supported by the Biostatistics and Bioinformatics Shared Resource and the Proteomics and Metabolomics Shared Resource at the H. Lee Moffitt Cancer Center & Research Institute, and NCI-designated Comprehensive Cancer Center (P30-CA076292); the Cancer Research Institute (Stewart); P01CA250984, “Identifying Metabolic Vulnerabilities in Lung Cancer”; T32 CA233399/CA/NCI NIH HHS/United States; and the Anna D. Valentine and Charles L. Oehler Award (Li/Fridley/Chen), which funded prior work.

The research reported in this publication was also supported by the Huntsman Cancer Foundation. Research reported in this publication utilized the Cancer Bioinformatics Shared Resource at Huntsman Cancer Institute at the University of Utah and was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA042014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Benedetti, E., EM. Liu, C. Tang, F. Kuo, M. Buyukozkan, T. Park, J.. Park, F. Correa, AA. Hakimi, AM. Intlekofer, J. Krumsiek, and E. Reznik. 2023. 'A multimodal atlas of tumour metabolism reveals the architecture of gene-metabolite covariation', *Nature metabolism*, 5.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics*, 9.

Golkaram, Mahdi, Fengshen Kuo, Sounak Gupta, Maria I. Carlo, Michael L. Salmans, Raakhee Vijayaraghavan, Cerise Tang, Vlad Makarov, Phillip Rappold, Kyle A. Blum, Chen Zhao, Rami Mehio, Shile Zhang, Jim Godsey, Traci Pawlowski, Renzo G. DiNatale, Luc G. T. Morris, Jeremy Durack, Paul Russo, Ritesh R. Kotecha, Jonathan Coleman, Ying-Bei Chen, Victor E. Reuter, Robert J. Motzer, Martin H. Voss, Li Liu, Ed Reznik, Timothy A.

Chan, A. Ari Hakimi, Mahdi Golkaram, Fengshen Kuo, Sounak Gupta, Maria I. Carlo, Michael L. Salmans, Raakhee Vijayaraghavan, Cerise Tang, Vlad Makarov, Phillip Rappold, Kyle A. Blum, Chen Zhao, Rami Mehio, Shile Zhang, Jim Godsey, Traci Pawlowski, Renzo G. DiNatale, Luc G. T. Morris, Jeremy Durack, Paul Russo, Ritesh R. Kotecha, Jonathan Coleman, Ying-Bei Chen, Victor E. Reuter, Robert J. Motzer, Martin H. Voss, Li Liu, Ed Reznik, Timothy A. Chan, and A. Ari Hakimi. 2022. 'Spatiotemporal evolution of the clear cell renal cell carcinoma microenvironment links intra-tumoral heterogeneity to immune escape', *Genome Medicine 2022 14:1*, 14.

Jendoubi, Takoua. 2021. 'Approaches to Integrating Metabolomics and Multi-Omics Data: A Primer', *Metabolites 2021, Vol. 11, Page 184*, 11.

Karagiorgou, Zoi, Panagiotis N. Fountas, Dimitra Manou, Erik Knutsen, Achilleas D. Theocharis, Zoi Karagiorgou, Panagiotis N. Fountas, Dimitra Manou, Erik Knutsen, and Achilleas D. Theocharis. 2022. 'Proteoglycans Determine the Dynamic Landscape of EMT and Cancer Cell Stemness', *Cancers 2022, Vol. 14, Page 5328*, 14.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. 2016. 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update', *Nucleic Acids Research*, 44.

Langfelder, Peter, Steve Horvath, Peter Langfelder, and Steve Horvath. 2008. 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics 2008 9:1*, 9.

Langfelder, Peter, Bin Zhang, and Steve Horvath. 2008. 'Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R', *Bioinformatics*, 24.

Pang, Zhiqiang, Lei Xu, Charles Viau, Yao Lu, Reza Salavati, Niladri Basu, Jianguo Xia, Zhiqiang Pang, Lei Xu, Charles Viau, Yao Lu, Reza Salavati, Niladri Basu, and Jianguo Xia. 2024. 'MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics', *Nature Communications 2024 15:1*, 15.

Qianxing, Mo, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, and Shen R. 2013. 'Pattern discovery and cancer gene identification in integrated cancer genomic data - PubMed', *Proceedings of the National Academy of Sciences of the United States of America*, 110.

Tang, Cerise, Amy X Xie, Minwei Liu, Fengshen Kuo, Minsoo Kim, Renzo Di Natale, Mahdi Golkaram, Yingbei Chen, Sounak Gupta, Robert Motzer, Paul Russo, Jonathan Coleman, Maria I Carlo, Martin H Voss, Ritesh R Kotecha, Chung Han Lee, Wesley Tansey, Nikolaus Schultz, A Ari Hakimi, and Ed Reznik. 2023. 'Immunometabolic coevolution defines unique microenvironmental niches in ccRCC', *Cell metabolism*, 35.

Yip, Andy M, Steve Horvath, Andy M Yip, and Steve Horvath. 2007. 'Gene network interconnectedness and the generalized topological overlap measure', *BMC Bioinformatics 2007 8:1*, 8.

Zhong, Weimin, Huijing Chen, Jiayi Yang, Chaoqun Huang, Yao Lin, Jiyi Huang, Weimin Zhong, Huijing Chen, Jiayi Yang, Chaoqun Huang, Yao Lin, and Jiyi Huang. 2023. 'Inflammatory response-based prognostication and personalized therapy decisions in clear cell renal cell cancer to aid precision oncology', *BMC Medical Genomics 2023 16:1*, 16.

Zhou, Bo, Gui-Zhen Wang, Zhe-Sheng Wen, Yong-Chun Zhou, Yun-Chao Huang, Ying Chen, and Guang-Biao Zhou. 2018. 'Somatic Mutations and Splicing Variants of Focal Adhesion Kinase in Non–Small Cell Lung Cancer', *JNCI: Journal of the National Cancer Institute*, 110.