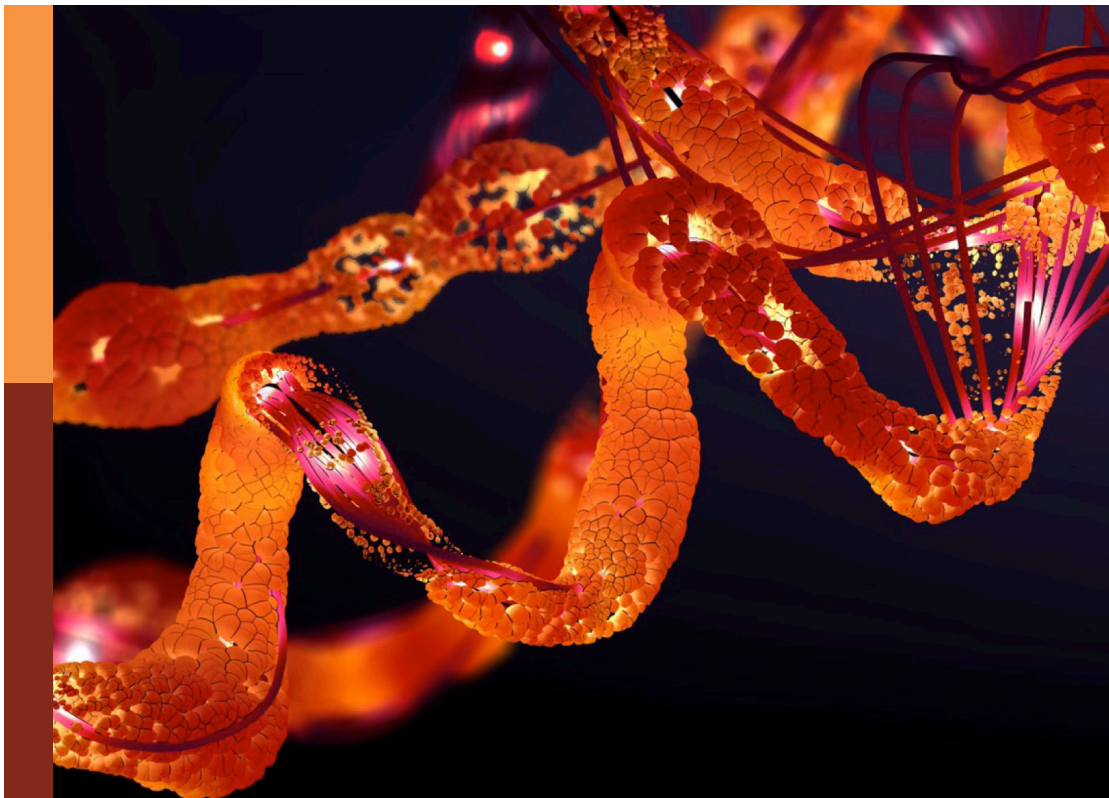
frontiers | Research Topics

Advances in methods and tools for multi-omics data analysis

Edited by
Ornella Cominetti, Sergio Oller Moreno and Sumeet Agarwal

Published in
Frontiers in Molecular Biosciences



ISSN 1664-8714
ISBN 978-2-8325-2342-1
DOI 10.3389/978-2-8325-2342-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Advances in methods and tools for multi-omics data analysis

Topic editors

Ornella Cominetti — Nestlé Research Center, Switzerland

Sergio Oller Moreno — Institute for Bioengineering of Catalonia (IBEC), Spain

Sumeet Agarwal — Indian Institute of Technology Delhi, India

Citation

Cominetti, O., Moreno, S. O., Agarwal, S., eds. (2023). *Advances in methods and tools for multi-omics data analysis*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-2342-1

Table of contents

frontiers | Frontiers in Molecular Biosciences

TYPE Editorial
PUBLISHED 24 April 2023
DOI 10.3389/fmolb.2023.1186822

OPEN ACCESS

EDITED AND REVIEWED BY
Wolfram Weckwerth,
University of Vienna, Austria

*CORRESPONDENCE
Ornella Cominetti,
Ornella.Cominetti@rd.nestle.com

RECEIVED 15 March 2023
ACCEPTED 14 April 2023
PUBLISHED 24 April 2023

CITATION
Cominetti O, Agarwal S and Oller-Moreno S (2023), Editorial: Advances in methods and tools for multi-omics data analysis.
Front. Mol. Biosci. 10:1186822.
doi: 10.3389/fmolb.2023.1186822

Editorial: Advances in methods and tools for multi-omics data analysis

Ornella Cominetti [1]*, Sumeet Agarwal [2] and Sergio Oller-Moreno [3]

[1]Nestlé Research Center, Lausanne, Switzerland, [2]Indian Institute of Technology Delhi, New Delhi, India, [3]Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Barcelona, Spain

KEYWORDS

multi-omics, personalized medicine, machine learning, integrative omics, deep generative models, protein protein interaction (PPI) network, EATRIS European infrastructure for translational medicine

Editorial on the Research Topic
Advances in methods and tools for multi-omics data analysis

Multi-omics data analysis is a rapidly growing field with significant potential in personalized medicine. Despite the many advances in this domain, there are still important challenges that need to be addressed, such as standardization of methods and limitations in interpreting results.

The Research Topic "*Advances in Methods and Tools for Multi-Omics Data Analysis*" showcases novel techniques and tools, including machine/deep learning tools, multi-factor analysis, Bayesian statistics, network-based models, and computational approaches for network-based integrative multi-omics analysis.

This article Research Topic comprises one perspective article, which addresses the translational challenges of multi-omics research in the realm of European personalized medicine, four review articles covering challenges in the path towards achieving precision medicine in cancer treatment and immuno-oncology, computational approaches for network-based integrative multi-omics analysis, deep generative models for learning joint embeddings of single-cell multi-omics data, and methods for characterization and visualization of protein–protein interaction networks in a multi-omics integration context.

Additionally, there are five original research articles ranging from the presentation of a new statistical framework for the analysis of longitudinal multi-omics data to the application of multi-omics methods to different diseases such as cancer, cutaneous melanoma, inflammatory diseases, myocardial dysfunction, and tuberous sclerosis complex-related angiomyolipoma.

Various machine learning (ML) techniques that can be used to integrate multi-omics data are discussed in this Research Topic. Network-based diffusion/propagation methods and multiview/multi-modal ML are two such techniques that can exploit information captured in each omics dataset and infer associations between different data types. Deep learning methods are an example of multiview/multi-modal learning, which can capture complex non-linear associations in a multi-layered manner.

Complex longitudinal omics datasets require new statistical approaches for their analysis, and ALASCA, a new package presented by Jarmund, Madssen and

Giskeødegård, shows a promising framework for tackling the longitudinal and multivariate nature of multi-omics studies, as well as covariate adjustment. In contrast, Magnusson et al. used a different approach for longitudinal omics data analysis: they applied linear mathematical mixed time-delayed splice variant models to predict protein abundances from mRNA expression.

The review article by Brombacher and colleagues provides a systematic overview of current deep generative models (DGM)-based approaches for learning joint embeddings from multi-omics data and illustrates how small sample sizes impact the amount of information that can be recovered from such datasets. Specifically, the review examines how the performance of popular DGM-based approaches to infer joint low-dimensional representations is influenced by varying numbers of cells, which is particularly relevant at the stage of designing an experiment.

Robin and colleagues discuss the role of protein-protein interactions (PPIs) in cellular mechanisms and the construction of PPI networks. PPIs are involved in physical and biochemical processes in structured environments and can be constructed using prediction methods and high-throughput experiments. Computational methods have emerged as a promising way to identify PPIs, and integration methods can be used to filter false interactions. Visualization is a key step in analyzing PPI networks, but the complexity of proteomes in different organisms presents a challenge.

The article by Oldoni and colleagues summarizes the outcome of the 2021 European Infrastructure for Translational Medicine (EATRIS)-Plus Multi-omics Stakeholder Group workshop. This multidisciplinary and cross-institutional working group aims to become a European reference group for implementing personalized medicine across Europe. This perspective article discusses the potential of precision medicine in healthcare and the challenges associated with it, such as moving beyond genomics to integrated multi-omics and multi-modal complex biomarker generation, new technologies and digital health, data standardization to enable multi-modal integration and AI-supported drug modeling, variability in omics data at source, data privacy and regulatory aspects, and economic implications.

Moreover, the Research Topic includes original research articles that demonstrate the application of multi-omics in various diseases. For example, Wang and colleagues present a study where ultra-performance liquid chromatography-mass spectrometer (UPLC-MS) was used to measure plasma proteins and metabolites in patients with renal cysts, sporadic angiomyolipoma, and tuberous sclerosis complex (TSC)-related angiomyolipoma before and after immunosuppressant treatment, with the aim of finding potential diagnostic and prognostic biomarkers as well as revealing the underlying mechanism of TSC tumorigenesis.

In another contribution, Wang et al. applied Least Absolute Shrinkage and Selection Operator (LASSO) and Support Vector Machine-Recursive Feature Elimination (SVM-RFE) algorithms to identify a cancer cell stemness feature, identifying 3 specific subtypes of melanoma with different survival outcomes.

One of the key challenges in multi-omics data analysis is the combination of different data types to identify composite biomarker signatures. This merging is complicated by the fact that multi-omics data often needs to be coupled with other data, such as imaging data, phenotypic data, and medical data (Electronic Health Records and patient-related outcomes). The integration of multi-omics and multi-modal data marks a significant step closer to personalized medicine, although many challenges remain before these biomarkers can be fully implemented in routine clinical care. The contributions in this Research Topic represent a small but solid base of step towards achieving these goals.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 16 February 2021
doi: 10.3389/fmolb.2020.598725

The Integrative Analysis Identifies Three Cancer Subtypes and Stemness Features in Cutaneous Melanoma

*Xiaoran Wang[1†], Qi Wan[1†], Lin Jin[2], Chengxiu Liu[3], Chang Liu[1], Yaqi Cheng[1] and Zhichong Wang[1]**

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-Sen University, Guangzhou, China, [2]The First Affiliated Hospital of Shandong First Medical University, Shandong, China, [3]Department of Ophthalmology, Affiliated Hospital of Qingdao University Medical College, Qingdao, China

Background: With the growing uncovering of drug resistance in melanoma treatment, personalized cancer therapy and cancer stem cells are potential therapeutic targets for this aggressive skin cancer.

Methods: Multi-omics data of cutaneous melanoma were obtained from The Cancer Genome Atlas (TCGA) database. Then, these melanoma patients were classified into different subgroups by performing "CancerSubtypes" method. The differences of stemness indices (mRNAsi and mDNAsi) and tumor microenvironment indices (immune score, stromal score, and tumor purity) among subtypes were investigated. Moreover, the Least Absolute Shrinkage and Selection Operator (LASSO) and Support Vector Machine-Recursive Feature Elimination (SVM-RFE) algorithms were performed to identify a cancer cell stemness feature, and the likelihood of immuno/chemotherapeutic response was further explored.

Results: Totally, 3 specific subtypes of melanoma with different survival outcomes were identified from TCGA. We found subtype 2 of melanoma with the higher immune score and stromal score and lower mRNAsi and tumor purity score, which has the best survival time than the other subtypes. By performing Kaplan–Meier survival analysis, we found that mRNAsi was significantly associated with the overall survival time of melanomas in subtype 2. Correlation analysis indicated surprising associations between stemness indices and subsets of tumor-infiltrating immune cells. Besides, we developed and validated a prognostic stemness-related genes feature that can divide melanoma patients into high- and low-risk subgroups by applying risk score system. The high-risk group has a significantly shorter survival time than the low-risk subgroup, which is more sensitive to CTLA-4 immune therapy. Finally, 16 compounds were screened out in the Connectivity Map database which may be potential therapeutic drugs for melanomas.

Conclusion: Thus, our finding provides a new framework for classification and finds some potential targets for the treatment of melanoma.

Keywords: cutaneous melanoma, classification, stemness feature, prognosis, canccer stem cell

OPEN ACCESS

Edited by:
Mahendra Pratap Kashyap, University of Alabama at Birmingham, United States

Reviewed by:
Rakesh Pathak, National Institutes of Health Clinical Center (NIH), United States
Sanjay Rathod, University of Pittsburgh, United States

****Correspondence:***
Zhichong Wang
wangzhichong@gzzoc.com

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *26 August 2020*
Accepted: *31 December 2020*
Published: *16 February 2021*

Citation:
Wang X, Wan Q, Jin L, Liu C, Liu C, Cheng Y and Wang Z (2021) The Integrative Analysis Identifies Three Cancer Subtypes and Stemness Features in Cutaneous Melanoma. Front. Mol. Biosci. 7:598725. doi: 10.3389/fmolb.2020.598725

INTRODUCTION

Melanoma is a quite lethal tumor once it has spread (metastasized). Melanoma arises from the precursor lesion with an accumulation of unrestrained mutations; orthotopic melanoma can be cured by resection in combination with continuously proven adjuvant therapy (Bray et al., 2018; Siegel et al., 2019).

Progression of melanoma can be characterized by the genetically distinct subpopulations which are related to a high occurrence of chemotherapy resistance. Given that about 90% of metastatic tumors develop resistance, a high incidence of melanoma in the reduced overall survival rate is due to the resistance to chemotherapies (Chow et al., 2011). At present, there are some validated adjuvant treatments for melanoma. Still, considering the side effects and different drug treatment responses of melanoma patients, the best choice and implementation of comprehensive melanoma therapy are unresolved. It is critical to find a more targeted selection for advanced melanoma patients.

Among tumor cells, the strong chemoresistance of tumor stem cells is closely related to high mortality after metastasis. Cancer stem cells are defined as the precursors by tumorigenesis, self-renewal, and pluripotency, namely, a subset of tumor-initiating cells (Abbaszadegan et al., 2017). To date, melanoma stem cells have been identified as a subpopulation of melanoma cells which can express cellular markers, like CD271, CD133, ABCB5, MDR1, etc. (Civenni et al., 2011; Keshet et al., 2008; Sharma at al., 2010). According to recent studies, melanoma stem cells can participate in related signal transduction pathways and play vital roles in escaping from immune surveillance and resistance to radiation therapy or chemotherapy (El-Khattouti et al., 2014; Mohme et al., 2017; Pak et al., 2004). Studies have found that molecules related to the expression of stem markers in tumors can enhance the resistance of tumors to chemotherapy, which is the basis for cancer stem cells to resist the toxic effects of chemotherapy drugs. The expression level of some stem cell-related markers is positively correlated with chemotherapy tolerance. The reason why cancer stem cells can escape from the cytotoxic effect of chemotherapeutic drugs includes their drug excretion mechanism, anti-apoptosis mechanism, and DNA damage repair mechanism (Meng et al., 2014; Schoning et al., 2017). Cancer stem cells also could express stronger stem cell-related potentials when they resist chemotherapy by activating specific pathways (Takeda et al., 2016). Therefore, the study of the characteristics of drug resistance mechanism of cancer stem cells has excellent application prospects and significance, and it is meaningful for complementary drug treatment programs to melanoma patients.

Current therapeutic strategies targeting tumor stem cells mainly include targeting specific surface markers or intracellular signal transduction pathways, inducing tumor stem cell differentiation, and changing the tumor stem cell microenvironment (Pei et al., 2020; Qin et al., 2020; Zhang et al., 2020). However, some studies have shown that tumor cells can be dedifferentiated into tumor stem cells to supplement depleted tumor stem cells under the influence of their surrounding environment. The ability of this new tumor stem cell to tolerate chemotherapy is still unknown. The heterogeneity of tumors and the complexity of the surrounding microenvironment make tumor treatment extremely complicated, so understanding the tumor heterogeneity and its external environment is vital (Lian et al., 2019). In particular, changes in the immune environment related to tumors will help us further to understand the melanoma therapeutic strategy.

MATERIALS AND METHODS

Data Collection and Cancer Subtype Identification

The transcriptome profile of RNA sequencing data and matched DNA methylation data of cutaneous melanoma as well as clinical information were obtained from the TCGA database. After data processing like distribution check, imputation, and normalization, three data types including gene expression, miRNA expression, and DNA methylation merged into a final dataset for integrative analysis. Next, these melanoma patients were divided into different subgroups by performing three clustering methods in R package ("CancerSubtypes").

Stemness Index Calculation

Stemness Index Workflow (https://bioinformaticsfmrp.github.io/PanCanStem_Web/) provides the steps and processes to regenerate our stemness indices (mRNAsi and mDNAsi), which train a stemness signature using normal stem cells and apply the one-class algorithm to define a stemness index for each tumor sample. The mRNA stemness index based on a gene set contains 11,774 genes, and the DNA stemness index calculated by a DNA methylation set contains 151 differentially methylated CpG sites. We first scored melanoma patients by applying Stemness Index Workflow and then scaled the stemness indices range from 0 to 1.

Tumor Microenvironment Estimation

The immune score, stromal score, and tumor purity were calculated from gene expression data by applying the ESTIMATE algorithm in R package ("ESTIMATE"). By running the ESTIMATE algorithm, immune score, stromal score, and tumor purity of each melanoma patient can be estimated. Then, we also scaled the value of immune score, stromal score, and tumor purity range from 0 to 1.

Evaluation of the Relationship Between Subtype and Clinical Variables

To clarify the clinicopathologic characteristics of the cancer subtypes, the subgroup analysis of clinical variables including mRNAsi, mDNAsi, immune score, stromal score, tumor purity, age, sex, and metastatic status was performed. Next, Kaplan–Meier plots were used to explore the prognostic value of stemness index (mRNAsi and mDNAsi) and found that only mRNAsi had a significant association with overall survival time in all melanoma patients. Hence, mRNAsi was screened out for

further analysis. Afterwards, each subtype of melanoma was divided into low and high mRNAsi groups by median cutoff of value, and Kaplan–Meier plots were drawn. The differences between low and high mRNAsi groups in subtypes were compared by log-rank tests. Eventually, Kaplan–Meier survival analysis showed that the mRNAsi was only significantly associated with overall survival in subtype 2. In addition, the cutaneous melanoma patients in subtype 2 were randomly divided into a 70% training dataset and a 30% validation dataset. In training datasets, samples were divided into high and low mRNAsi groups. "Limma" package in R software was applied to identify the differentially expressed genes (DEGs). The |log 2 fold change (FC)| ≥0.5 and p values <0.05 were considered as the cutoff criterion for DEGs. Then, univariate Cox regression analysis was used to screen the prognostic DEGs (p values <0.05). Next, for subsequently selecting the important mRNAsi-related features, the Least Absolute Shrinkage and Selection Operator (LASSO) and Support Vector Machine-Recursive Feature Elimination (SVM-RFE) algorithms were applied to reduce the prognostic DEGs.

Identification and Validation of Stemness Features

LASSO and SVM-RFE algorithms jointly determine the qualified seed of DEGs for the risk formula, and the risk score is generated as follows: risk score = $\sum_{i=1}^{N}(coef_i \times expr_i)$, in which N means the number of feature genes, $expr_i$ means the expression level of genes, and $coef_i$ means regression coefficient calculated by multivariate Cox regression analysis.

The risk score of each sample in training dataset was estimated, and the patients were accordingly classified into high- and low-risk group by the median cutoff. Univariate and multivariate Cox logistic analyses for OS were performed on the patient clinical characteristics (age, gender, stage, and metastasis) and the risk score of stemness features.

To compare the differences between high- and low-risk groups, we drew Kaplan–Meier survival curves and calculated the significance by log-rank tests. The area under the curve (AUC) of receiver operating characteristic curves (ROC) was used to evaluate the 5-year overall survival predictive accuracy of the model. Besides, to test the robustness of our results, stemness features were further verified in a validation dataset (GSE65904) which was downloaded from the GEO database.

Evaluation of the Association Between Stemness Indices and Immune Microenvironment

To explore the relationship between stemness indices and immune microenvironment in different melanoma subtype, single sample gene set enrichment analysis (ssGSEA) method in R package ("GSVA") was applied to specifically discriminate 24 human immune cells, including innate and adaptive immune cells. The innate immune cells contain natural killer (NK) cells, CD56bright NK cells, CD56dim NK cells, dendritic cells (DCs), activated DCs (aDCs), immature DCs (iDCs), plasmacytoid DCs (pDCs), neutrophils, macrophages, eosinophils, and mast cells, and the adaptive immune cells, including T cells, B cells, and cytotoxic cells. Moreover, the T cells consist of T effector memory (Tem), T central memory cells (Tcm), CD8 T cells, Tgd cells, regulatory T cells (Treg), T helper cells and T follicular helper cells (TFH), Th1, Th2, and Th17. Next, the correlation analysis between stemness indices (mRNAsi/mDNAsi) and 24 immune cells expression was performed.

Immuno/Chemotherapeutic Response Prediction

To explore the potential immuno/chemotherapeutic drugs, we predicted the candidate compounds response for each sample based on the Connectivity Map website (https://portals.broadinstitute.org/cmap/). The significant compounds were selected ($p < 0.05$). Additionally, immune checkpoint inhibitors have been approved as routine drugs for melanoma. Thus, we also predicted the potential response to immunotherapy by using the TIDE website tool (http://tide.dfci.harvard.edu/).

Statistical Analysis

All statistical analyses were conducted using the R package (v.3.5.2) and corresponding packages. Survival analysis was applied by using "survival" and "survivalROC" package. LASSO algorithm was conducted by "glmnet" package. SVM algorithm was calculated with the "e1017" package. The correlation coefficient was calculated by Spearman test. For comparisons of two groups and more than two groups, Kruskal–Wallis test and one-way analysis of variance were used as non-parametric and parametric methods, respectively. The association between subgroup and clinicopathological characteristics was analyzed with the chi-square test.

RESULTS

Data Collection and Cancer Subtype Identification

After combining multi-omics data into integrative analysis, 449 melanoma patient samples were obtained from the TCGA database. Then, according to prior studies, these patients were divided into three subtypes by three clustering methods including consensus clustering (CC), consensus non-negative matrix factorization (CNMF), and similarity network fusion with CC (SNFCC) (Lu et al., 2018). Although all the clustering methods can classify melanoma patients into 3 subtypes with different survival outcomes (CC: p value = 4.23e-10; NMF: p value = 2.62e-09; SNFCC: p value = 7.51e-09) (**Figure 1A**) and clear boundaries between different color areas (**Figure 1B**), combined with the value of average silhouette width (ASW) which works as a measure of cluster coherence to assess whether samples are more similar within subtypes. SNFCC showed more advantages than other methods and were selected for subsequent analysis (CC: ASW = 0.74; CNMF: ASW = 0.82; SNFCC: ASW = 0.89) (**Figure 1C**). In SNFCC, subtype 1 contains 205 samples, subtype 2 consists of 177 samples, and subtype 3

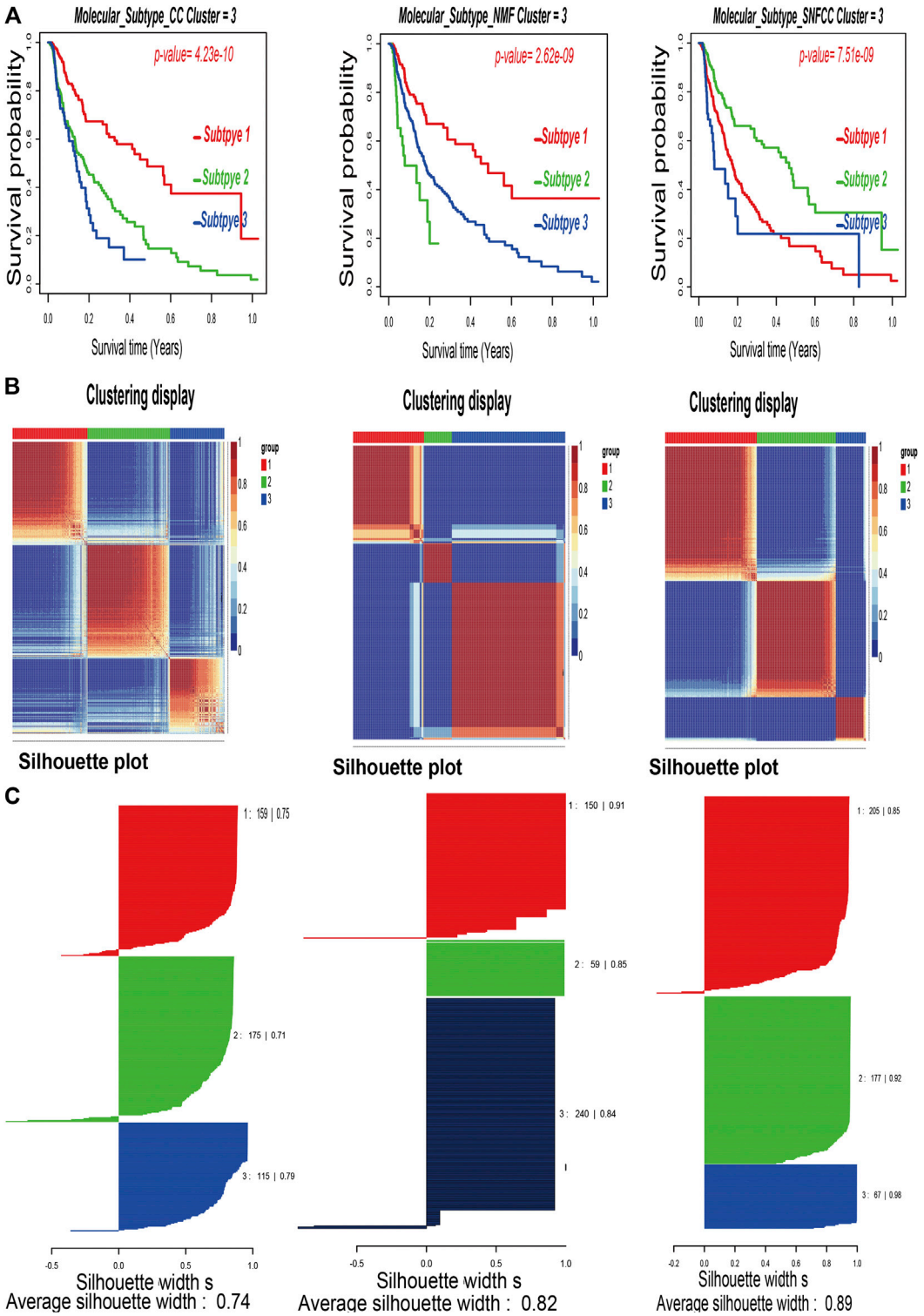


FIGURE 1 | Classification of melanoma patients by three clustering methods including consensus clustering (CC), consensus non-negative matrix factorization (CNMF), and similarity network fusion with CC (SNFCC). **(A)**: Kaplan–Meier survival analysis of three subtypes with log-rank test *p* value; **(B)**: clustering heatmap of three subtype samples; **(C)**: average silhouette width representing the coherence of clusters.

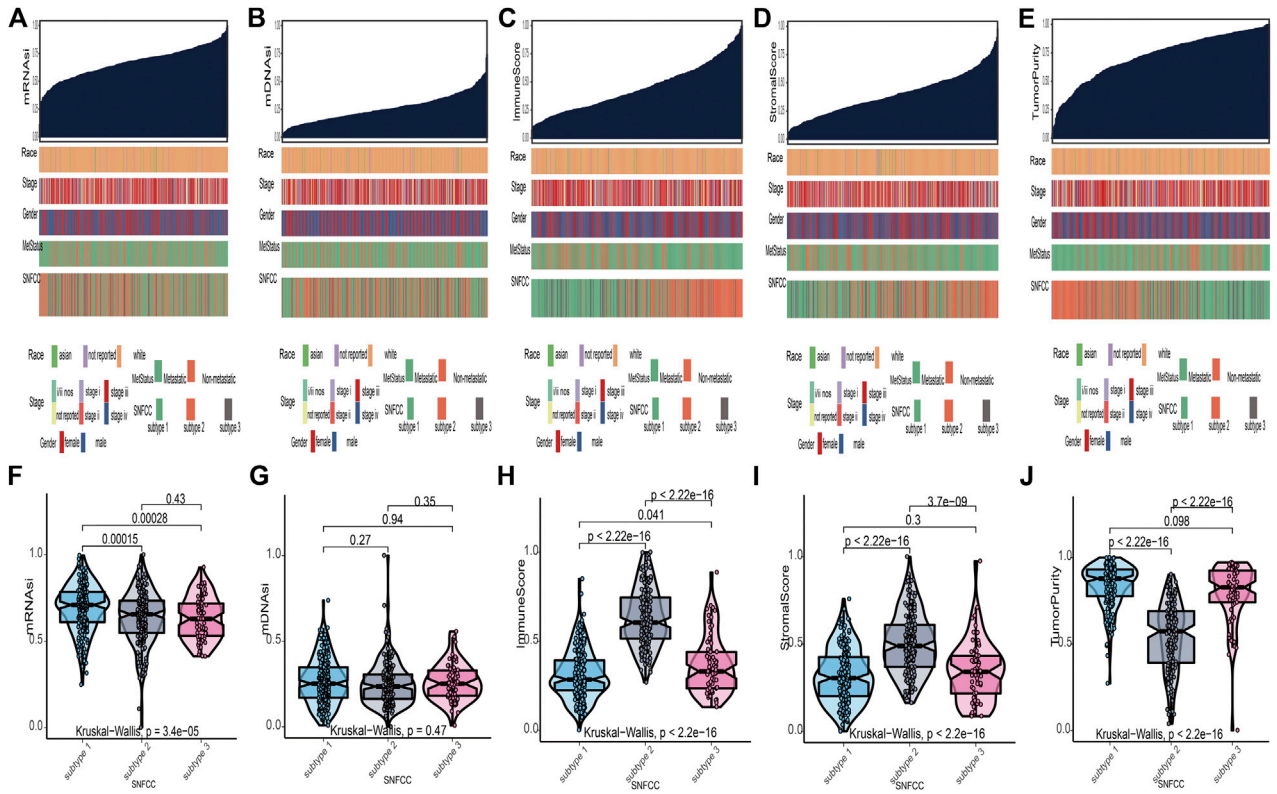


FIGURE 2 | Clinical variables associated with the stemness indices (mRNAsi and mDNAsi) and tumor microenvironment indices (immune score, stromal score, and tumor purity) in melanoma. **(A)**: the association between clinical variables (race, stage, gender, metastatic status, and subtype) and mRNAsi; **(B)**: the association between clinical variables and mDNAsi; **(C)**: the association between clinical variables and immune score; **(D)**: the association between clinical variables and stromal score; **(E)**: the association between clinical variables and tumor purity. Columns represent samples sorted by score of indices from low to high **(top row)**. Rows represent clinical variables. **(F)**: boxplots of mRNAsi in individual samples stratified by subtype; **(G)**: boxplots of mDNAsi in individual samples stratified by subtype; **(H)**: boxplots of immune score in individual samples stratified by subtype; **(I)**: boxplots of stromal score in individual samples stratified by subtype; **(J)**: boxplots of tumor purity in individual samples stratified by subtype.

includes 67 samples. Among three subtypes, subtype 2 has the longest survival time compared to others.

Clinicopathologic Characteristics of the Cancer Subtypes

According to the methods, we acquired stemness indices (mRNAsi and mDNAsi) and tumor microenvironment indices (immune score, stromal score, and tumor purity) of 449 melanoma patients. After excluding adjacent, duplicated, and incomplete samples, data of 427 patients were included for further subgroup analysis. Firstly, the melanoma patients were ordered by their values of stemness and tumor microenvironment indices (from low to high) to explore whether any clinical feature was associated with these calculated indices (**Figures 2A–E**). Remarkably, the patients in subtype 1 had higher value of mRNAsi (median value = 0.71) than subtype 2 (median value = 0.66) and subtype 3 (median value = 0.63) patients (**Table 1**). Boxplots of mRNAsi suggested that there is a significant difference among subtypes (**Figure 2F**). Similarly, subgroup analysis of tumor purity showed that patients in subtype 1 (median value = 0.88) had higher values than subtype 2 (median value = 0.57) and subtype 3 (median value = 0.83) (**Figure 2J** and **Table 1**). As for immune and stromal score, results manifested that subtype 2 samples had higher values (immune median value = 0.61; stromal median value = 0.49) than subtype 1(immune median value = 0.28; stromal median value = 0.30) and subtype 3 (immune median value = 0.33; stromal median value = 0.34) (**Figures 2H,I** and **Table 1**). However, there is no statistical difference among the three subtypes in mDNAsi index (**Figure 2G**). The median values of three subtypes were 0.25, 0.24, and 0.25, respectively (**Table 1**). Next, the subgroup analysis of other clinical variables like overall survival time, age, gender, race, metastatic status, and stages was also applied. The results showed that survival time, age, metastatic status, and stages were statistically different among melanoma subtypes (**Table 1**).

Relationship Between Stemness Indices and Tumor Microenvironment

Kaplan–Meier curves of mRNAsi and mDNAsi manifested that only mRNAsi was significantly associated with overall survival

TABLE 1 | Clinicopathological variables of subtypes in melanoma. IQR means interquartile range.

		Subtype 1	Subtype 2	Subtype 3	*p*	Test
n		195	167	65		
Survival time (median [IQR])		2.84 [1.30, 5.88]	4.44 [2.27, 9.48]	1.28 [1.01, 2.18]	0.000	Kruskal–Wallis test
Age (median [IQR])		60.00 [49.00, 70.00]	55.00 [45.00, 68.50]	63.00 [56.00, 76.00]	0.002	Kruskal–Wallis test
Gender (%)	Female	64 (32.8)	69 (41.3)	27 (41.5)	0.191	Chi-square test
	Male	131 (67.2)	98 (58.7)	38 (58.5)		
Race (%)	Asian	4 (2.1)	4 (2.4)	4 (6.2)	0.459	Chi-square test
	Not reported	5 (2.6)	3 (1.8)	2 (3.1)		
	White	186 (95.4)	160 (95.8)	59 (90.8)		
MetStatus (%)	Metastatic	159 (81.5)	155 (92.8)	15 (23.1)	0.000	Chi-square test
	Non-metastatic	36 (18.5)	12 (7.2)	50 (76.9)		
Stage (%)	I/II nos	4 (2.1)	5 (3.0)	1 (1.5)	0.000	Chi-square test
	Not reported	13 (6.7)	18 (10.8)	3 (4.6)		
	Stage I	30 (15.4)	40 (24.0)	3 (4.6)		
	Stage II	59 (30.3)	30 (18.0)	40 (61.5)		
	Stage III	77 (39.5)	67 (40.1)	15 (23.1)		
	Stage IV	12 (6.2)	7 (4.2)	3 (4.6)		
mRNAsi (median [IQR])		0.71 [0.61, 0.79]	0.66 [0.55, 0.73]	0.63 [0.53, 0.72]	0.000	Kruskal–Wallis test
mDNAsi (median [IQR])		0.25 [0.17, 0.35]	0.24 [0.16, 0.31]	0.25 [0.18, 0.33]	0.533	Kruskal–Wallis test
Stromal score (median [IQR])		0.30 [0.20, 0.43]	0.49 [0.37, 0.61]	0.34 [0.22, 0.43]	0.000	Kruskal–Wallis test
Immune score (median [IQR])		0.28 [0.22, 0.39]	0.61 [0.51, 0.75]	0.33 [0.24, 0.44]	0.000	Kruskal–Wallis test
Tumor purity (median [IQR])		0.88 [0.78, 0.93]	0.57 [0.39, 0.69]	0.83 [0.74, 0.93]	0.000	Kruskal–Wallis test

time in all melanoma patients, and low mRNAsi group had a longer survival time than high mRNAsi group (log-rank p = 0.009) (**Figure 3A**). Therefore, mRNAsi was selected out for the next analysis. Subgroup analysis of mRNAsi showed that subtype 2 was significantly correlated to overall survival time (log-rank p = 0.037), whereas Kaplan–Meier curves of subtype 1 and subtype 3 showed that there was no statistical difference (**Figure 3B**). In addition, correlation analysis revealed that mRNAsi was positively correlated with mDNAsi (r = 0.155, p = 0.001) and tumor purity (r = 0.370, p = 0.000), while immune and stromal score were negatively associated with mRNAsi (r = −0.220, p = 0.000; r = −0.590, p = 0.000) (**Figure 3C**).

Identification and Validation of Stemness Features

To identify stemness features, subtype 2 samples were randomly divided into a training dataset (n = 117) and a validation dataset (n = 50). The clinical characteristics of training and validation datasets are listed in **Table 2**, and statistical results indicated that they were balanced between two datasets. Firstly, based on the selection criteria, 364 DEGs were screened out in training dataset, in which 319 genes were significantly downregulated and 45 genes were significantly upregulated (**Figure 4A**). Next, the univariate analysis of 364 DEGs was conducted, and the results showed that 27 prognostic DEGs were significantly associated with overall survival time in the training dataset (**Figure 4B**). Finally, 11 mRNAsi-related genes were selected by performing LASSO and SVM-RFE algorithm, and these genes were further used to construct a risk score system (**Figures 4C–E**). By applying this risk model, a risk score for each sample in the training dataset will be generated. Then, melanoma patients were divided into a high-risk group (n = 58) and a low-risk group (n = 59) by using the median cutoff value of the risk scores. Kaplan–Meier curves showed that patients in the high-risk group have a shorter survival time than low-risk group with a log-rank test of $p < 0.001$. To estimate the prediction power of 11 mRNAsi-related genes' signature, the ROC curve was drawn, and five years of AUC was 0.944 (**Figure 5A**). Besides, in order to confirm the robustness of the result, a verification test was conducted in the validation dataset and GSE65904 dataset. The validation and GSE65904 datasets were classified into high-risk and low-risk groups according to the training dataset. Kaplan–Meier curves showed that there is a significant difference between high-risk and low-risk groups in both validation dataset (log-rank p < 0.001) and GSE65904 dataset (log-rank $p < 0.001$) (**Figure 5B** and **Figure 5C**). The five years of AUC were 0.846 and 0.680, respectively. What is more, to explore the prognostic value of risk score and other clinical features (age, race, gender, and metastatic status), univariate and multivariate logistic regression were applied. Based on the results, only the risk score was significantly associated with overall survival in both univariate and multivariate analysis (**Table 3**).

Association Between Stemness Indices and Immune Microenvironment

To evaluate the associations between stemness indices and immune microenvironment, correlations analysis between immune cell individuals and mRNAsi (**Figure 5D**) and mDNAsi (**Figure 5E**) was performed. In mRNAsi, most of the immune cells were negatively correlated with mRNAsi, in which iDC, macrophages, mast cells, NK cells, TFH, and Tgd were commonly negatively correlated with three subtypes, while only Th2 cell was commonly positively correlated with three subtypes. As for mDNAsi, less immune cells were associated with mDNAsi compared to mRNAsi and

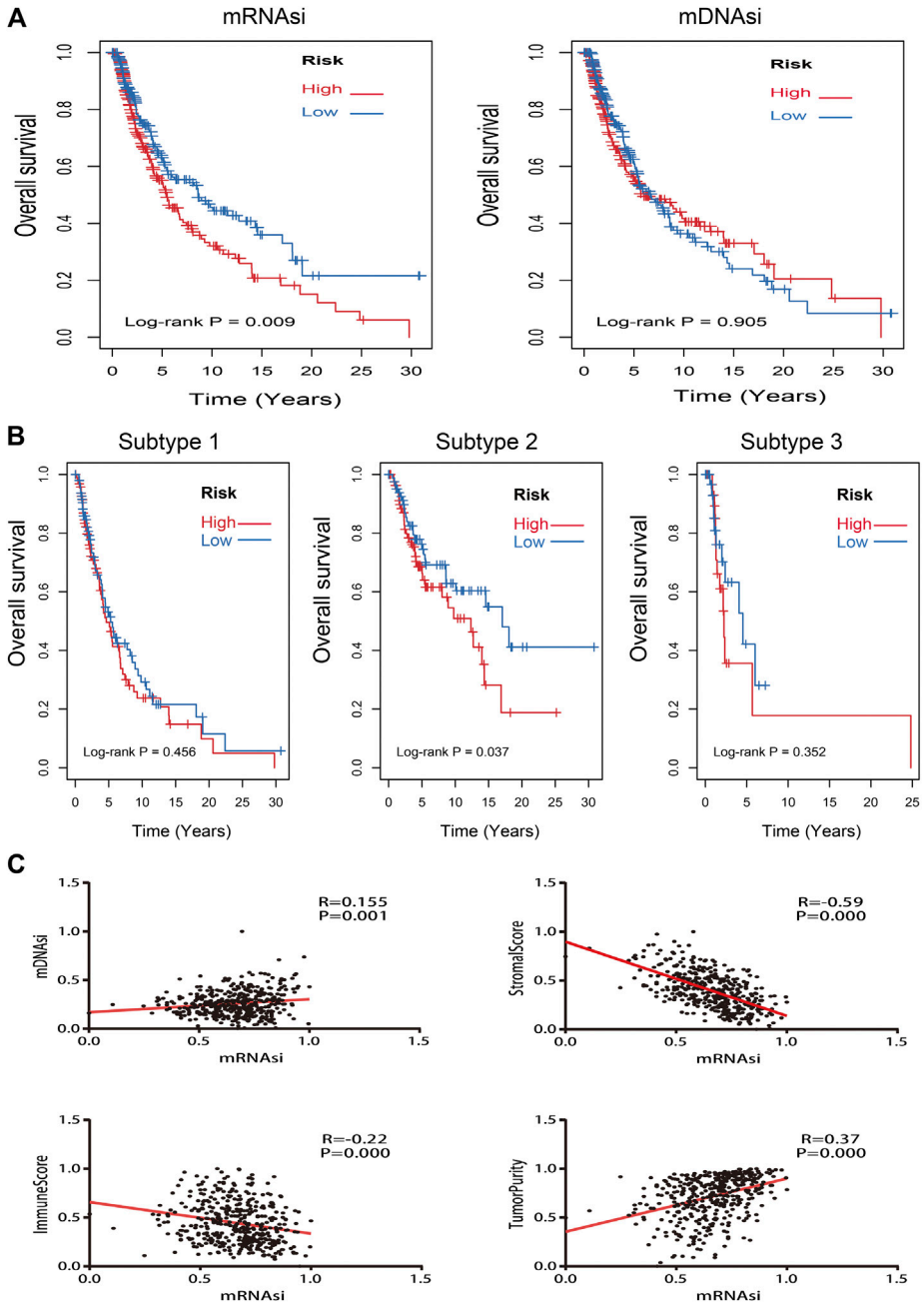


FIGURE 3 | Kaplan–Meier survival analysis and correlation analysis of stemness indices. **(A)**: Kaplan–Meier analysis of mRNAsi and mDNAsi in all melanoma samples; **(B)**: Kaplan–Meier analysis of each subtype of melanoma patients with high or low mRNAsi; **(C)**: the correlation analysis between mRNAsi and other indices (mDNAsi, immune score, stromal score, and tumor purity).

only CD8 T cell and cytotoxic cell were commonly negatively associated with three subtypes.

Immuno/Chemotherapeutic Response Prediction

Immunotherapy is regarded as an emerging therapy and widely used in melanoma. Therefore, we conducted the TIDE algorithm and subclass mapping to compare the expression profile of the two subgroups and another published dataset containing 47 patients with melanoma that responded to immune checkpoint inhibitors (CTLA-4 and PD-1). Interestingly, we found that the low-risk group in subtype 2 is more promising to respond to anti-CTLA-4 therapy (Bonferroni corrected $p = 0.007$) (**Figure 6B**). Then, we applied the same method to predict immune checkpoint inhibitors for other melanoma subtypes. We surprisingly found

TABLE 2 | Clinicopathological variables of training and validation dataset. IQR means interquartile range.

		Training samples	Validation samples	*p*	Test
n		117	50		
OS.time (median [IQR])		4.28 [2.28, 9.34]	4.50 [2.26, 9.44]	0.917	Kruskal–Wallis test
OS (median [IQR])		0.00 [0.00, 1.00]	0.00 [0.00, 1.00]	0.428	Kruskal–Wallis test
Age (median [IQR])		53.00 [44.00, 68.00]	57.50 [46.25, 69.50]	0.318	Kruskal–Wallis test
Gender (%)	Female	51 (43.6)	18 (36.0)	0.459	Chi-square test
	Male	66 (56.4)	32 (64.0)		
Race (%)	Asian	2 (1.7)	2 (4.0)	0.668	Chi-square test
	Not reported	2 (1.7)	1 (2.0)		
	White	113 (96.6)	47 (94.0)		
MetStatus (%)	Metastatic	109 (93.2)	46 (92.0)	1	Chi-square test
	Non-metastatic	8 (6.8)	4 (8.0)		
Stage (%)	I/II nos	4 (3.4)	1 (2.0)	0.197	Chi-square test
	Not reported	16 (13.7)	2 (4.0)		
	Stage I	24 (20.5)	16 (32.0)		
	Stage II	18 (15.4)	12 (24.0)		
	Stage III	50 (42.7)	17 (34.0)		
	Stage IV	5 (4.3)	2 (4.0)		
mRNAsi (median [IQR])		0.66 [0.53, 0.73]	0.67 [0.58, 0.74]	0.362	Kruskal–Wallis test
mDNAsi (median [IQR])		0.24 [0.17, 0.32]	0.20 [0.14, 0.29]	0.089	Kruskal–Wallis test
Stromal score (median [IQR])		0.52 [0.37, 0.63]	0.46 [0.38, 0.57]	0.434	Kruskal–Wallis test
Immune score (median [IQR])		0.61 [0.52, 0.74]	0.60 [0.49, 0.79]	0.969	Kruskal–Wallis test
Tumor purity (median [IQR])		0.57 [0.42, 0.67]	0.60 [0.35, 0.71]	0.737	Kruskal–Wallis test

that low-risk groups no matter in subtype 1 (**Figure 6A**) or subtype 3 (**Figure 6C**) significantly responded to anti-CTLA-4 therapy (Bonferroni corrected $p = 0.03$; Bonferroni corrected $p = 0.012$). Moreover, chemotherapy is a common treatment for melanoma. Therefore, the Connectivity Map database was also applied to predict potential compounds. Compounds significantly correlated with at least two cancer subtypes will be selected (**Figure 6D**). Eventually, 16 compounds were significantly enriched, including anisomycin, cephaeline, chenodeoxycholic acid, digitoxigenin, ellipticine, gossypol, helveticoside, hycanthone, lanatoside C, metixene, nitrofural, ouabain, oxedrine, prednisone, proscillaridin, and valinomycin.

DISCUSSION

Worldwide, cutaneous melanoma is known as a common type of malignancy with high morbidity and mortality, while the traditional classification lacks clinical benefits and strategies for treatment are still ineffective. Therefore, in this study, we tried to establish a more evaluable classification system to help figure out better treatment choices for advanced melanoma patients. Therapies without inclusive consideration of gene transcription characters would bring treatment indeterminacy (Hamid et al., 2018). Given that, we sought to take gene expression, miRNA expression, and DNA methylation into account to partition melanoma profile and compared three clustering models. We successfully categorized melanoma patients into 3 validated subtypes. Interestingly, significant difference in overall survival time was observed among these 3 subtypes, which suggests that there exist biological relevance and distinction among subgroups. In addition, it's generally accepted that melanoma tumors are composed of a mixture of different cell types such as cancer cells, cancer stem cells, and immune cells. We also defined the stemness indices (mRNAsi and mDNAsi) and tumor microenvironment indices (immune score, stromal score, and tumor purity) for different melanoma subtypes. The results manifested that subtype 2 with higher immune score and stromal score and lower mRNAsi and tumor purity score has the best survival time compared to other subtypes, which was consistent with our next findings that low risk of mRNAsi has longer survival than high risk. Correlation analysis also proved that intimate associations exist among these indices. Thus, our research provides a framework for exploring how the context of diverse cell types among subtypes may elucidate the observed diverse clinical outcomes and treatment effects.

Cancer cells are recently hypothesized to be derived from cancer stem cells which are closely correlated with relapse of malignant tumors, drug resistance, and metastasis. Recent studies have found that some stemness-related genes can not only initiate malignant neoplastic cascade and maintain the oncogenicity of stem cells, but also enhance the chemotherapy resistance of tumor stem cells (Chen et al., 2016; Chiou et al., 2017; Kharas and Lengner, 2017; Redmer et al., 2017). Therefore, therapeutic targeting genes associated with melanoma stem cells are urgently important. In this study, we developed and validated a robust stemness-related signature which contains 11 genes (LOC284837, OCLN, ABCC9, MEGF6, TSPYL5, RAB27B, TF, TNXB, KIAA0495, TCEA3, and KCNH5). Among these stemness-related genes, some have been identified to be associated with stem cells. For instance, TF (tissue factor) is a multifunctional membrane protein which correlates with various advanced cancers. The overexpression of TF can increase the activity of breast cancer stem cells in vitro (Shaker et al., 2017). The activated RAB27B expression will promote the secretion of colorectal cancer stem cell exosomes (Cheng et al., 2019). TSPYL5 is

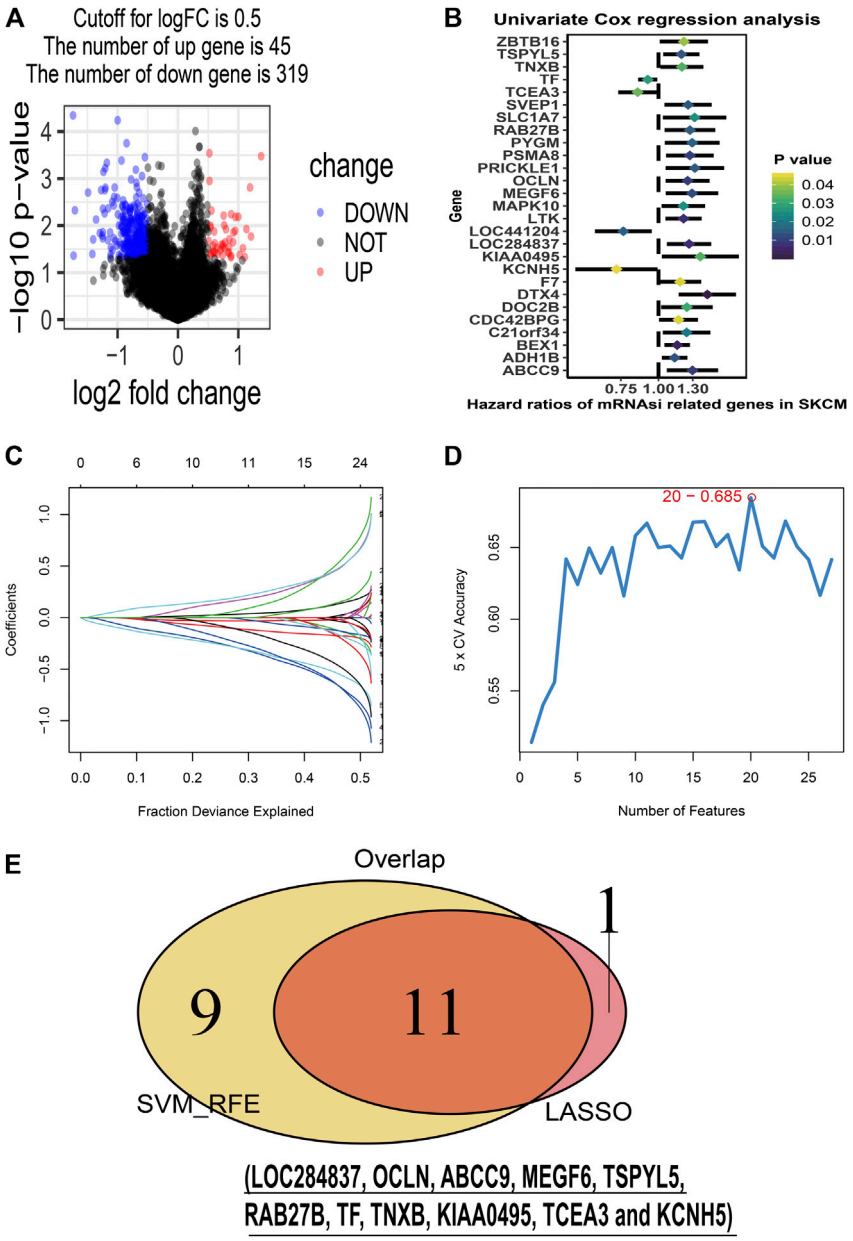


FIGURE 4 | Stemness-related genes feature selection. **(A)**: volcano plot of the differentially expressed stemness-related genes in training dataset; **(B)**: forest plots of the prognostic differentially expressed stemness-related genes; **(C)**: the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm coefficient profiles of the 12 genes that met the prognostic criteria initially; **(D)**: Support Vector Machine-Recursive Feature Elimination (SVM-RFE) algorithms. The point highlighted indicates the lowest error rate, and the corresponding genes at this point are the best signature selected by SVM-RFE. **(E)**: the Venn plot of overlap genes selected by LASSO and SVM-RFE algorithms.

highly expressed in human pluripotent stem cells, and the overexpression of TSPYL5 is proven to promote cell proliferation and migration (Na et al., 2019). Moreover, KCNH5 and TCEA3 are shown to have high concentrations in mesenchymal stem cells and mouse embryonic stem cells (Cha et al., 2013; Jeong et al., 2013). Besides, the univariate and multivariate regression analysis indicated that the risk score of stemness-related signature could be regarded as an independent prognostic model in melanoma. Hence, it seems reasonable to believe that our identified stemness-related signature can be regarded as a prognostic biomarker for further clinical research. Consistent with taking advantage of integrated stemness indices to classified melanoma in our study, mounting evidence suggests that the control of melanoma stem cell could be typically administrated to melanoma patients (Luo et al., 2012; Rappa et al., 2008; Santini et al., 2012).

In this study, we explored the different immune environment of melanoma with different stemness indices. In mRNAsi of this study,

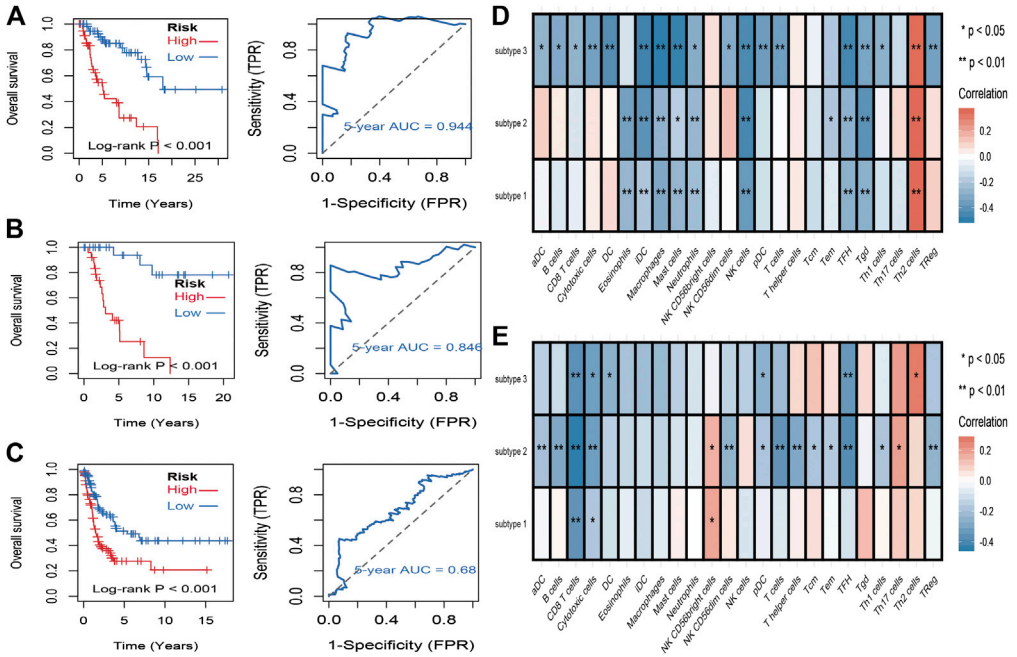


FIGURE 5 | Identification and validation of stemness-related genes feature for survival prediction. **(A)**: Kaplan–Meier analysis of 11 mRNAsi-related genes' signature and 5 years of the receiver operating characteristic (ROC) curve in training dataset. **(B)**: Kaplan–Meier analysis of 11 mRNAsi-related genes' signature and 5 years of the receiver operating characteristic (ROC) curve in validation dataset. **(C)**: Kaplan–Meier analysis of 11 mRNAsi-related genes' signature and 5 years of the receiver operating characteristic (ROC) curve in GSE65904 dataset. **(D)**: correlations between the mRNAsi and the subsets of tumor-infiltrating immune cells estimated by "ssGSEA" method. **(F)**: correlations between the mDNAsi and the subsets of tumor-infiltrating immune cells estimated by "ssGSEA" method.

TABLE 3 | Univariate and multivariate Cox regression analyses of 11 mRNAsi-related genes signature and clinical variables associated with overall survival in subtype 2 datasets.

	Univariate analysis				**Multivariate analysis**			
Marker	unicox_p	HR	lower .95	upper .95	mutlicox_p	exp(coef)	lower .95	upper .95
Age	0.003	1.026	1.009	1.044	0.051	1.019	1.000	1.039
Gender	0.487	1.197	0.722	1.985	0.716	1.111	0.629	1.962
Race	0.887	1.177	0.123	11.283	0.976	0.968	0.119	7.875
MetStatus	0.065	3.202	0.930	11.021	0.065	3.512	0.926	13.314
Stage	0.186	1.154	0.933	1.428	0.985	0.998	0.784	1.270
Risk score	0.000	1.225	1.154	1.300	0.000	1.270	1.176	1.370

we found that T helper 2 cell (Th2 cell) was the only commonly positively correlated with three subtypes of melanoma. Th2 cells are induced by interleukin 4, which can be secreted by basophils, eosinophils, mast cells, natural killer T cells, or differentiated Th2 cells (Lee et al., 2002). The main effect of Th2 cells is to activate B cell, and then humoral immunity would be stimulated by plasma cells. Nevertheless, tumor immunotherapy requires cellular immunity which is mainly activated by Th1. Both Th1 cells and Th2 cells can secrete cytokines to promote their proliferation and inhibit each other's proliferation (Saito et al., 1999). Under normal immune environment, Th1 cells and Th2 cells are in a relatively balanced state. Th2 bias signifies the imbalance of Th1/Th2. Th2 could strongly inhibit Th1 responses (Guenova et al., 2013). Th2 cells promote tumor growth and prevent tumor rejection. The bias of Th2 is regarded as one of the mechanisms of tumor immune escape. Previous research proved that the tumor microenvironment of advanced melanoma is composed of Th2-type polarization that facilitates disease progression. Studies have also shown that Th2 dominance could mediate chronic inflammation which could promote melanoma metastasis (Nevala et al., 2009). It has been reported that, in melanoma, plasmacytoid dendritic cells can break this kind of immune homeostasis by OX40L and ICOSL to support melanoma progression (Aspord et al., 2013). Reversing the imbalance of Th1/Th2 has been a concerned treatment for tumors and other diseases (Kidd, 2003). Our results further supported the importance of treatment to Th2 bias in melanoma. To date, immunotherapy is pivotal for the treatment of patients with advanced melanoma patients. Cytotoxic lymphocyte-associated antigen 4 (CTLA-4) can compete with CD28 receptor-binding antigen-presenting cell surface binding sites. CD28 receptors can

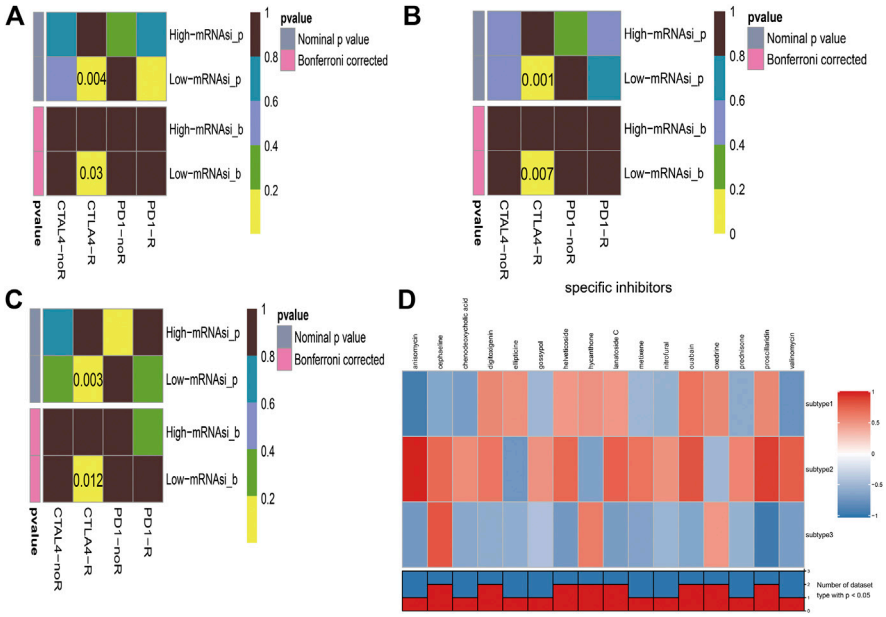


FIGURE 6 | Immunotherapeutic response and potential compounds identification. **(A)**: differential immunotherapeutic response targeting CTLA-4 and PD-1 between the high- and low-risk patients in subtype 1; **(B)**: differential immunotherapeutic response targeting CTLA-4 and PD-1 between the high- and low-risk patients in subtype 2; **(C)**: differential immunotherapeutic response targeting CTLA-4 and PD-1 between the high- and low-risk patients in subtype 3; **(D)**: heatmap of potential compounds and enrichment score (positive in red, negative in blue) obtained from the Connectivity Map database for each melanoma subtype. The bottom panel showed that the number of subtypes significantly enriched in compounds.

activate T cells. CTLA-4 is a highly homologous molecule with CD28 and binds to the B7 molecule (CD80/CD86), and the binding strength is higher than CD28. So once CTLA-4 is highly expressed and combined, it will be a loss of the co-stimulatory signal, and then CTLA-4 would inhibit lymphocyte activation and proliferation. CTLA-4 plays a key role in regulating the T-cell system and is often used as suppressive immune molecules in tumor therapy. Anti-CTLA-4 monoclonal antibodies can augment T-cell activation and proliferation and amplify immunity by blockading CTLA-4 pathways, which enhances the patient's ability to perform an antitumor immune response. The use of CTLA-4 monoclonal antibodies to block the CTLA-4 pathway in clinical immunotherapy of tumors also has been the current research hotspot (Carreno et al., 2000; Wells et al., 2001). However, in clinical data, the treatment has no survival benefits (Boasberg et al., 2010; Robert et al., 2011). Immune checkpoint inhibitors also have some severe side effects mostly because the blockade of the immune checkpoint pathway makes the immune responses of related organs and tissues amplified; it cannot be terminated in time, and autoimmune damage would occur. Which kind of patients is appropriate to a special treatment remains unclear. As of now, we still do not have sufficient evidence to guide clinical decisions. In this study, we comprehensively described the stemness and environmental characteristics of melanoma and found that low-risk mRNAsi groups are promising to respond to anti-CTLA-4 therapy which may provide effective measurement solutions to help the final clinical decision and hoped to help patients with advanced melanoma get the maximum remission rate.

Additionally, 16 potential compounds were identified to significantly correlate with at least two cancer subtypes. Few of these compounds have been used in melanoma researches in vitro or in vivo. For example, previous experiments proved that low doses of anisomycin can inhibit one-third of protein synthesis in melanoma cells and induce cancer cell apoptosis (Slipicevic et al., 2013). Gossypol was demonstrated to have more cytotoxic to melanoma cell lines than the conventional drugs like melphalan, cisplatin, and dacarbazine (Blackstaffe et al., 1997). What is more, nitrofural is known to act as pro-drugs, and the combination of olaparib and nitrofural will enhance the effect for the treatment of melanoma (McNeil et al., 2013). Although large part of compounds had not been reported for the treatment of melanoma, these undiscovered compounds may be regarded as the promising drug for the subsequent melanoma research.

Although our preliminary results have several implications for patients with melanoma, several limitations must be considered. Firstly, melanoma patients are recruited from public database, and findings in this research are carried out by bioinformatics methods. Secondly, the sample size in this study is small, and experimental verifications are lacking. Thus, additional fundamental researches are needed to explore the underlying mechanisms.

In conclusion, our studies provide a comprehensive cellular characterization for melanoma classification and additional subtypes that may benefit from stemness-related genes targeted therapies. Our studies also afford strategies to assess

more promising population for immunotherapy and identify several potential compounds that could supply more effective treatment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: TCGA-SKCM cohorts were downloaded from TCGA database (https://www.cancer.gov/tcga/). GSE65904 cohorts were downloaded from GEO database (https://www.ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

QW and JL were responsible for writing-original draft preparation; XW was responsible for writing-review and editing; CL, YC, and CL were responsible for data curation; ZW was responsible for project administration and funding acquisition. All the authors commented and approved the text

FUNDING

This work was supported by the National Key R&D program of China (2018YFC1106000).

REFERENCE

Abbaszadegan, M. R., Bagheri, V., Razavi, M. S., Momtazi, A. A., Sahebkar, A., and Gholamin, M. (2017). Isolation, identification, and characterization of cancer stem cells: A review. *J. Cell. Physiol.* 232 (8), 2008–2018. doi:10.1002/jcp.25759

Aspord, C., Leccia, M. T., Charles, J., and Plumas, J. (2013). Plasmacytoid dendritic cells support melanoma progression by promoting Th2 and regulatory immunity through OX40L and ICOSL. *Canc. Immunol. Res.* 1 (6), 402–415. doi:10.1158/2326-6066.CIR-13-0114-T

Blackstaffe, L., Shelley, M. D., and Fish, R. G. (1997). Cytotoxicity of gossypol enantiomers and its quinone metabolite gossypolone in melanoma cell lines. *Melanoma. Res.* 7 (5), 364–372. doi:10.1097/00008390-199710000-00002

Boasberg, P., Hamid, O., and O'Day, S. (2010). Ipilimumab: unleashing the power of the immune system through CTLA-4 blockade. *Semin. Oncol.* 37 (5), 440–449. doi:10.1053/j.seminoncol.2010.09.004

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Canc. J. Clin.* 68 (6), 394–424. doi:10.3322/caac.21492

Carreno, B. M., Bennett, F., Chau, T. A., Ling, V., Luxenberg, D., Jussif, J., et al. (2000). CTLA-4 (CD152) can inhibit T cell activation by two different mechanisms depending on its level of cell surface expression. *J. Immunol.* 165 (3), 1352–1356. doi:10.4049/jimmunol.165.3.1352

Cha, Y., Heo, S. H., Ahn, H. J., Yang, S. K., Song, J. H., Suh, W., et al. (2013). Tcea3 regulates the vascular differentiation potential of mouse embryonic stem cells. *Gene Expr.* 16 (1), 25–30. doi:10.3727/105221613x13776146743343

Chen, H. Y., Lin, L. T., Wang, M. L., Lee, S. H., Tsai, M. L., Tsai, C. C., et al. (2016). Musashi-1 regulates AKT-derived IL-6 autocrinal/paracrinal malignancy and chemoresistance in glioblastoma. *Oncotarget* 7 (27), 42485–42501. doi:10.18632/oncotarget.9890

Cheng, W. C., Liao, T. T., Lin, C. C., Yuan, L. E., Lan, H. Y., Lin, H. H., et al. (2019). RAB27B-activated secretion of stem-like tumor exosomes delivers the biomarker microRNA-146a-5p, which promotes tumorigenesis and associates with an immunosuppressive tumor microenvironment in colorectal cancer. *Int. J. Cancer* 145 (8), 2209–2224. doi:10.1002/ijc.32338

Chiou, G. Y., Yang, T. W., Huang, C. C., Tang, C. Y., Yen, J. Y., Tsai, M. C., et al. (2017). Musashi-1 promotes a cancer stem cell lineage and chemoresistance in colorectal cancer cells. *Sci. Rep.* 7 (1), 2172. doi:10.1038/s41598-017-02057-9

Chow, E. K., Zhang, X. Q., Chen, M., Lam, R., Robinson, E., Huang, H., et al. (2011). Nanodiamond therapeutic delivery agents mediate enhanced chemoresistant tumor treatment. *Sci. Transl. Med.* 3 (73), 73ra21. doi:10.1126/scitranslmed.3001713

Civenni, G, Walter, A, Kobert, N, Mihic-Probst, D, Zipser, M, Belloni, B, et al. (2011). Human CD271-positive melanoma stem cells associated with metastasis establish tumor heterogeneity and long-term growth. *Canc. Res.* 71 (8), 3098–109.doi:10.1158/0008-5472.CAN-10-3997

El-Khattouti, A., Selimovic, D., Haïkel, Y., Megahed, M., Gomez, C. R., and Hassan, M. (2014). Identification and analysis of CD133(+) melanoma stem-like cells conferring resistance to taxol: An insight into the mechanisms of their resistance and response. *Canc. Lett.* 343 (1), 123–133. doi:10.1016/j.canlet.2013.09.024

Guenova, E., Watanabe, R., Teague, J. E., Desimone, J. A., Jiang, Y., Dowlatshahi, M., et al. (2013). TH2 cytokines from malignant cells suppress TH1 responses and enforce a global TH2 bias in leukemic cutaneous T-cell lymphoma. *Clin. Canc. Res.* 19 (14), 3755–3763. doi:10.1158/1078-0432.CCR-12-3488

Hamid, O., Robert, C., Ribas, A., Hodi, F. S., Walpole, E., Daud, A., et al. (2018). Antitumour activity of pembrolizumab in advanced mucosal melanoma: a post-hoc analysis of KEYNOTE-001, 002, 006. *Br. J. Cancer* 119 (6), 670–674. doi:10.1038/s41416-018-0207-6

Jeong, S. G., Ohn, T., Kim, S. H., and Cho, G. W. (2013). Valproic acid promotes neuronal differentiation by induction of neuroprogenitors in human bone-marrow mesenchymal stromal cells. *Neurosci Lett.* 554, 22–27. doi:10.1016/j.neulet.2013.08.059

Keshet, G. I., Goldstein, I., Itzhaki, O., Cesarkas, K., Shenhav, L., Yakirevitch, A., et al. (2008). MDR1 expression identifies human melanoma stem cells. *Biochem. Biophys. Res. Commun.* 368 (4), 930–936. doi:10.1016/j.bbrc.2008.02.022

Kharas, M. G., and Lengner, C. J. (2017). Stem cells, cancer, and MUSASHI in blood and guts. *Trends Cancer* 3 (5), 347–356. doi:10.1016/j.trecan.2017.03.007

Kidd, P. (2003). Th1/Th2 balance: the hypothesis, its limitations, and implications for health and disease. *Altern. Med. Rev.* 8 (3), 223–246.

Lee, D. U., Agarwal, S., and Rao, A. (2002). Th2 lineage commitment and efficient IL-4 production involves extended demethylation of the IL-4 gene. *Immunity* 16 (5), 649–660. doi:10.1016/s1074-7613(02)00314-x

Lian, H., Han, Y. P., Zhang, Y. C., Zhao, Y., Yan, S., Li, Q. F., et al. (2019). Integrative analysis of gene expression and DNA methylation through one-class logistic regression machine learning identifies stemness features in medulloblastoma. *Mol. Oncol.* 13 (10), 2227–2245. doi:10.1002/1878-0261.12557

Lu, X., Zhang, Q., Wang, Y., Zhang, L., Zhao, H., Chen, C., et al. (2018). Molecular classification and subtype-specific characterization of skin cutaneous melanoma by aggregating multiple genomic platform data. *J. Canc. Res. Clin. Oncol.* 144 (9), 1635–1647. doi:10.1007/s00432-018-2684-7

Luo, Y., Dallaglio, K., Chen, Y., Robinson, W. A., Robinson, S. E., McCarter, M. D., et al. (2012). ALDH1A isozymes are markers of human melanoma stem cells and potential therapeutic targets. *Stem Cells* 30 (10), 2100–2113. doi:10.1002/stem.1193

McNeil, E. M., Ritchie, A. M., and Melton, D. W. (2013). The toxicity of nitrofuran compounds on melanoma and neuroblastoma cells is enhanced by Olaparib and ameliorated by melanin pigment. *DNA Repair* 12 (11), 1000–1006. doi:10.1016/j.dnarep.2013.08.017

Meng, E., Mitra, A., Tripathi, K., Finan, M. A., Scalici, J., McClellan, S, et al. (2014). ALDH1A1 maintains ovarian cancer stem cell-like properties by altered regulation of cell cycle checkpoint and DNA repair network signaling. *PLoS One* 9 (9), e107142. doi:10.1371/journal.pone.0107142

Mohme, M., Riethdorf, S., and Pantel, K. (2017). Circulating and disseminated tumour cells - mechanisms of immune surveillance and escape. *Nat. Rev. Clin. Oncol.* 14 (3), 155–167. doi:10.1038/nrclinonc.2016.144

Na, H. J., Yeum, C. E., Kim, H. S., Lee, J., Kim, J. Y., and Cho, Y. S. (2019). TSPYL5-mediated inhibition of p53 promotes human endothelial cell function. *Angiogenesis* 22 (2), 281–293. doi:10.1007/s10456-018-9656-z

Nevala, W. K., Vachon, C. M., Leontovich, A. A., Scott, C. G., Thompson, M A., Markovic, S. N., et al. (2009). Evidence of systemic Th2-driven chronic

inflammation in patients with metastatic melanoma. *Clin. Canc. Res.* 15 (6), 1931–1939. doi:10.1158/1078-0432.CCR-08-1980

Pak, B. J., Lee, J., Thai, B. L., Fuchs, S. Y., Shaked, Y., Ronai, Z., et al. (2004). Radiation resistance of human melanoma analysed by retroviral insertional mutagenesis reveals a possible role for dopachrome tautomerase. *Oncogene* 23 (1), 30–38. doi:10.1038/sj.onc.1207007

Pei, J., Wang, Y., and Li, Y. (2020). Identification of key genes controlling breast cancer stem cell characteristics via stemness indices analysis. *J. Transl. Med.* 18 (1), 74. doi:10.1186/s12967-020-02260-9

Qin, S., Long, X., Zhao, Q., and Zhao, W. (2020). Co-Expression network analysis identified genes associated with cancer stem cell characteristics in lung squamous cell carcinoma. *Canc. Invest.* 38 (1), 13–22. doi:10.1080/07357907.2019.1697281

Rappa, G., Fodstad, O., and Lorico, A. (2008). The stem cell-associated antigen CD133 (Prominin-1) is a molecular therapeutic target for metastatic melanoma. *Stem Cells* 26 (12), 3008–3017. doi:10.1634/stemcells.2008-0601

Redmer, T., Walz, I., Klinger, B., Khouja, S., Welte, Y., Schäfer, R., et al. (2017). The role of the cancer stem cell marker CD271 in DNA damage response and drug resistance of melanoma cells. *Oncogenesis* 6 (1), e291. doi:10.1038/oncsis.2016.88

Robert, C., Thomas, L., Bondarenko, I., O'Day, S., Weber, J., Garbe, C., et al. (2011). Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N. Engl. J. Med.* 364 (26), 2517–2526. doi:10.1056/NEJMoa1104621

Saito, S., Tsukaguchi, N., Hasegawa, T., Michimata, T., Tsuda, H., and Narita, N. (1999). Distribution of Th1, Th2, and Th0 and the Th1/Th2 cell ratios in human peripheral and endometrial T cells. *Am. J. Reprod. Immunol.* 42 (4), 240–245. doi:10.1111/j.1600-0897.1999.tb00097.x

Santini, R., Vinci, M. C., Pandolfi, S., Penachioni, J. Y., Montagnani, V., Olivito, B., et al. (2012). Hedgehog-GLI signaling drives self-renewal and tumorigenicity of human melanoma-initiating cells. *Stem Cells* 30 (9), 1808–1818. doi:10.1002/stem.1160

Schöning, J. P., Monteiro, M., and Gu, W. (2017). Drug resistance and cancer stem cells: the shared but distinct roles of hypoxia-inducible factors HIF1α and HIF2α. *Clin. Exp. Pharmacol. Physiol.* 44 (2), 153–161. doi:10.1111/1440-1681.12693

Shaker, H., Harrison, H., Clarke, R., Landberg, G., Bundred, N. J., Versteeg, H. H., et al. (2017). Tissue Factor promotes breast cancer stem cell activity in vitro. *Oncotarget* 8 (16), 25915–25927. doi:10.18632/oncotarget.13928

Sharma, B. K., Manglik, V., and Elias, E. G. (2010). Immuno-expression of human melanoma stem cell markers in tissues at different stages of the disease. *J. Surg. Res.* 163 (1), e11–e15. doi:10.1016/j.jss.2010.03.043

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA A Canc. J. Clin.* 69 (1), 7–34. doi:10.3322/caac.21551

Slipicevic, A., Øy, G. F., Rosnes, A. K., Stakkestad, Ø., Emilsen, E., Engesæter, B., et al. (2013). Low-dose anisomycin sensitize melanoma cells to TRAIL induced apoptosis. *Canc. Biol. Ther.* 14 (2), 146–154. doi:10.4161/cbt.22953

Takeda, H., Okada, M., Suzuki, S., Kuramoto, K., Sakaki, H., Watarai, H., et al. (2016). Rho-associated protein kinase (ROCK) inhibitors inhibit survivin expression and sensitize pancreatic cancer stem cells to gemcitabine. *Anticanc. Res.* 36 (12), 6311–6318. doi:10.21873/anticanres.11227

Wells, A. D., Walsh, M. C., Bluestone, J. A., and Turka, L. A. (2001). Signaling through CD28 and CTLA-4 controls two distinct forms of T cell anergy. *J. Clin. Invest.* 108 (6), 895–903. doi:10.1172/JCI13220

Zhang, Y., Tseng, J. T., Lien, I. C., Li, F., Wu, W., and Li, H. (2020). mRNAsi Index: machine learning in mining lung adenocarcinoma stem cell biomarkers. *Genes* 11 (3), 257. doi:10.3390/genes11030257

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 15 August 2022
DOI 10.3389/fmolb.2022.967397

OPEN ACCESS

EDITED BY
Liangcai Zhao,
Wenzhou Medical University, China

REVIEWED BY
Jiangjiang Zhu,
Southwest Minzu University, China
Chao Tong,
Henan Agricultural University, China

*CORRESPONDENCE
Kaijiang Yu,
drkaijiang@163.com
Changsong Wang,
changsongwangicu@163.com

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 16 June 2022
ACCEPTED 19 July 2022
PUBLISHED 15 August 2022

CITATION
Jia X, Peng Y, Ma X, Liu X, Yu K and Wang C (2022), Analysis of metabolic disturbances attributable to sepsis-induced myocardial dysfunction using metabolomics and transcriptomics techniques. *Front. Mol. Biosci.* 9:967397. doi: 10.3389/fmolb.2022.967397

Analysis of metabolic disturbances attributable to sepsis-induced myocardial dysfunction using metabolomics and transcriptomics techniques

Xiaonan Jia[1], Yahui Peng[1], Xiaohui Ma[1], Xiaowei Liu[2], Kaijiang Yu[1]* and Changsong Wang[3]*

[1]Departments of Critical Care Medicine, The First Affiliated Hospital of Harbin Medical University, Harbin Medical University, Harbin, China, [2]Departments of Critical Care Medicine, The Fourth Affiliated Hospital of Harbin Medical University, Harbin Medical University, Harbin, China, [3]Departments of Critical Care Medicine, Harbin Medical University Cancer Hospital, Harbin Medical University, Harbin, China

Background: Sepsis-induced myocardial dysfunction (SIMD) is the most common and severe sepsis-related organ dysfunction. We aimed to investigate the metabolic changes occurring in the hearts of patients suffering from SIMD.

Methods: An animal SIMD model was constructed by injecting lipopolysaccharide (LPS) into mice intraperitoneally. Metabolites and transcripts present in the cardiac tissues of mice in the experimental and control groups were extracted, and the samples were studied following the untargeted metabolomics–transcriptomics high-throughput sequencing method. SIMD-related metabolites were screened following univariate and multi-dimensional analyses methods. Additionally, differential analysis of gene expression was performed using the DESeq package. Finally, metabolites and their associated transcripts were mapped to the relevant metabolic pathways after extracting transcripts corresponding to relevant enzymes. The process was conducted based on the metabolite information present in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

Results: One hundred and eighteen significant differentially expressed metabolites (DEMs) (58 under the cationic mode and 60 under the anionic mode) were identified by studying the SIMD and control groups. Additionally, 3,081 significantly differentially expressed genes (DEGs) (1,364 were down-regulated and 1717 were up-regulated DEGs) were identified in the

Abbreviations: SIMD, Sepsis-induced myocardial dysfunction; LPS, lipopolysaccharide; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEMs, differentially expressed metabolites; AMPK, AMP-activated protein kinase; DEGs, differentially expressed genes; LVEF, left ventricular ejection fraction; LVEDd, left ventricular end-diastolic dimension; LVESd, left ventricular end-systolic dimension; LVFS, left ventricular fractional shortening; NGS, Next-Generation Sequencing; OPLS-DA, orthogonal partial least squares–discriminant analysis; PCA, principal component analysis; ALA, Alpha-linolenic acid.

transcriptomes. The comparison was made between the two groups. The metabolomics–transcriptomics combination analysis of metabolites and their associated transcripts helped identify five metabolites (D-mannose, D-glucosamine 6-phosphate, maltose, alpha-linolenic acid, and adenosine 5′-diphosphate). Moreover, irregular and unusual events were observed during the processes of mannose metabolism, amino sugar metabolism, starch metabolism, unsaturated fatty acid biosynthesis, platelet activation, and purine metabolism. The AMP-activated protein kinase (AMPK) signaling pathways were also accompanied by aberrant events.

Conclusion: Severe metabolic disturbances occur in the cardiac tissues of model mice with SIMD. This can potentially help in developing the SIMD treatment methods.

KEYWORDS

sepsis, metabolic, myocardial dysfunction, SIMD, transcriptomics

1 Introduction

Sepsis is defined as a life-threatening organ dysfunction that is caused by an overreaction of the body to infection (Singer et al., 2016). It is a serious global problem and the most common cause of in-hospital mortality (Rudd et al., 2020). Sepsis-induced myocardial dysfunction (SIMD) is the most common and severe sepsis-related organ dysfunction. SIMD induces or exacerbates dysfunction in other organs. The prognosis of patients with SIMD is poor, resulting in an extremely high mortality rate (70–90%) (Martin et al., 2019; Ravikumar et al., 2021). It is known that the mechanisms underlying SIMD involve the release of circulating myocardial inhibitory substances, the release of nitric oxide and reactive oxygen species, abnormalities in calcium handling, downregulation of adrenergic pathways, and mitochondrial dysfunction (Hollenberg and Singer, 2021; Yang and Zhang, 2021). However, these abnormalities fail to explain the mechanisms underlying the onset and progression of SIMD. Circulating troponin and NT-proBNP exhibit good specificity and sensitivity in the cases of myocardial ischemic disease and cardiac failure. However, similar roles are not observed in the case of SIMD (Hollenberg and Singer, 2021). At present, a viable biomarker for SIMD is yet to be identified.

Metabolomics allows the exploration of small molecule metabolites in blood or tissues. Results obtained by conducting qualitative and quantitative analyses revealed that the relationship between metabolites and physiological/pathological changed over time (Rochfort, 2005; Patti et al., 2012). Metabolites are the end products of the biochemical activities occurring in the body. Therefore, metabolomics is the omics study that is closest to phenotyping. A number of metabolomics-oriented studies have been conducted to understand the pathogenesis, progression, and patient prognosis of sepsis (Neugebauer et al., 2016; Ping et al., 2019; Ping et al., 2021). However, there is a lack of such studies on SIMD. Additionally, transcriptomics facilitates the investigation of gene function and gene structure at a global level to identify differentially expressed genes (DEGs) within cells, tissues, or individuals under different physiological or pathological states (Velculescu et al., 1995; Virlon et al., 1999; Morris, 2009).

However, metabolomics solely may lead to incomplete findings. Therefore, metabolomics–transcriptomics combination analysis can be performed to accurately identify key metabolites, hub genes, and metabolic pathways associated with the ‘cause’ and ‘result’ dimensions (Rochfort, 2005; Griffin, 2006).

Mannose metabolism, amino sugar metabolism, starch metabolism, unsaturated fatty acid biosynthesis, platelet activation, and purine metabolism have some studies in sepsis (Gao et al., 2015; Bakalov et al., 2016; Hwang et al., 2019; She et al., 2022). However, there are no relevant studies in Sepsis-induced myocardial dysfunction (SIMD). The AMPK signaling pathway has some studies in SIMD (Song et al., 2020; Wang et al., 2021).

We aimed to investigate metabolite changes occurring in the heart tissues of mice suffering from SIMD. The SIMD-related metabolites and metabolic pathways were identified and studied by conducting untargeted metabolomics–transcriptomics combination analysis. Overall, the findings of this study provide new insights into the processes associated with the pathogenesis, early diagnosis, and treatment of SIMD.

2 Methods

2.1 Animal model establishment

Male C57BL/6 mice (age: 6–8 weeks) were purchased from Charles River (Beijing, China). The mice under study had free access to food and water. The mice belonging to the experimental group were administered intraperitoneal injections of lipopolysaccharide (LPS) (20 mg/kg) once to induce SIMD.

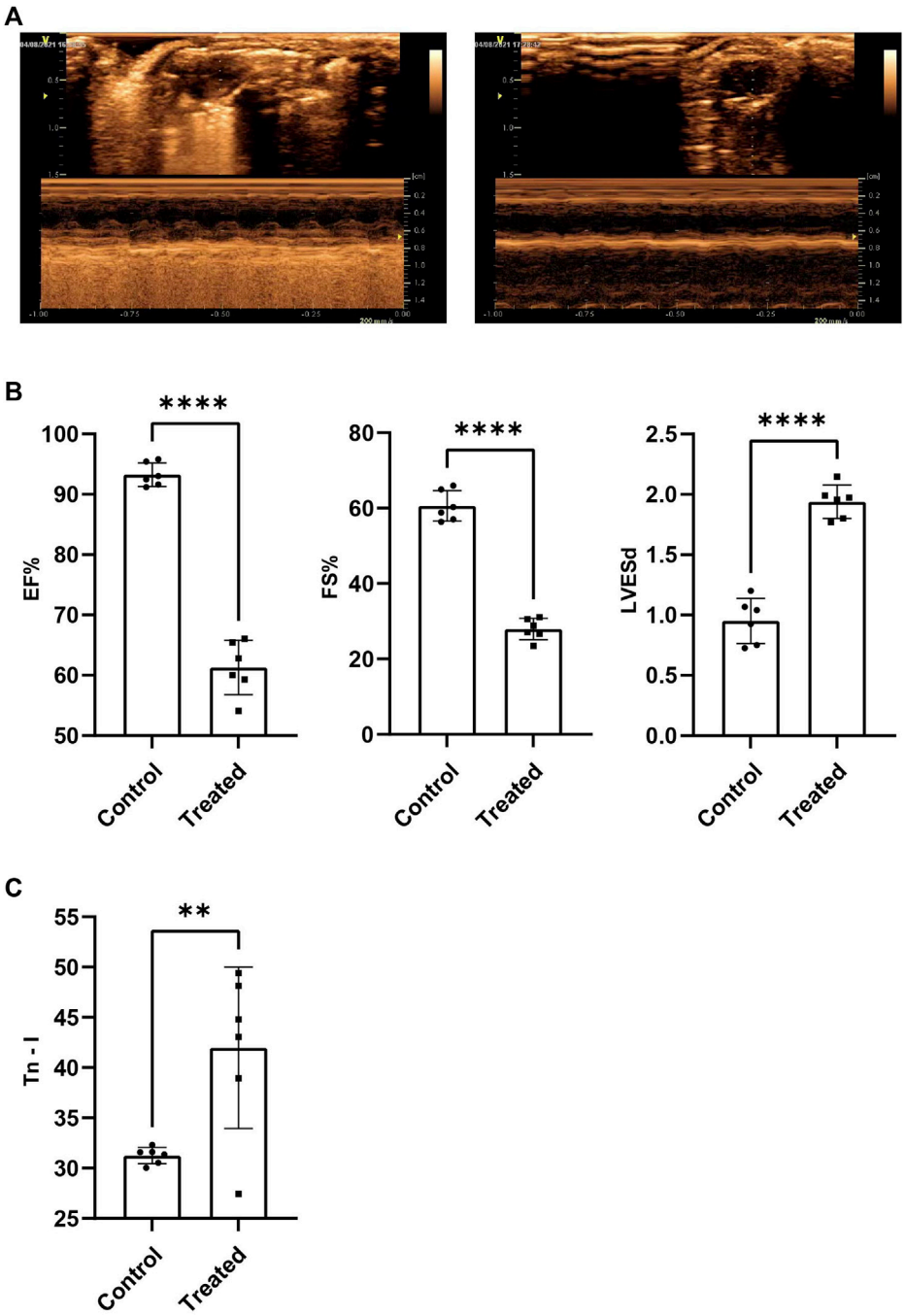


FIGURE 1
The cardiac function of SIMD mice decreased significantly. **(A)** Representative images of mice heart examined by echocardiography. **(B)**EF%, FS % and LVESd (n = 6, $p < 0.001$) **(C)**Tn-I in serum were measured by ELISA assays (n = 6, $p < 0.01$).

The volume of saline that was administered intraperitoneally to the mice belonging to the control group was the same as the volume of LPS injections. The mice were subjected to conditions of echocardiography after 6 hours of injection, and 2D and M-mode echocardiographic measurements were taken under these conditions. A high-resolution *in vivo* imaging system (VIVID E9, GE, United States) was used to record the data. The left ventricular ejection fraction (LVEF), left ventricular end-diastolic dimension (LVEDd), left ventricular end-systolic dimension (LVESd), and left ventricular fractional shortening

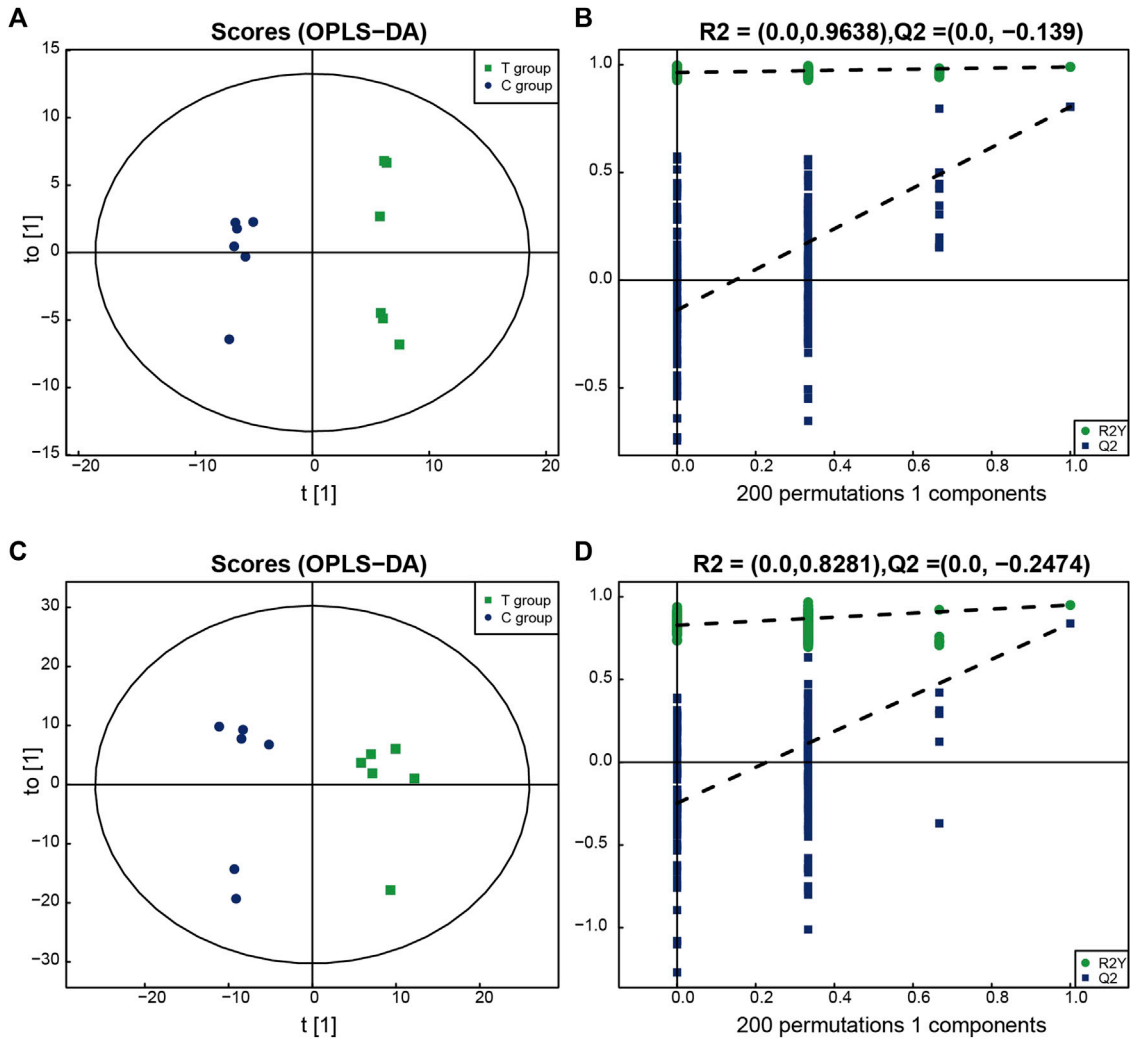


FIGURE 2
The multidimensional results in positive and negative ionization modes are shown in this figure. **(A,B)** OPLS-DA score plot and OPLS-DA validation plot intercepts in positive ionization modes: the Treated group vs. the Control group. R2Y = (0.0, 0.9638), Q2 = (0.0,−0.139). **(C,D)** OPLS-DA score plot and OPLS-DA validation plot intercepts in negative ionization modes: the Treated group vs. the Control group. R2Y = (0.0, 0.8281), Q2 = (0.0,−0.2474).

(LVFS) functioned as the measurement indicators. A short-axis view of the heart was obtained from the parasternal approach. The ejection fraction was also calculated. The formula for calculation is as follows:

$$\left(\mathrm{LVEDd}^3 - \mathrm{LVESd}^3\right)/\mathrm{LVEDd}^3 \times 100$$

The ejection fractional shortening was calculated as follows:

$$(\mathrm{LVEDd} - \mathrm{LVESDd})/\mathrm{LVEd} \times 100$$

Subsequently, the whole heart tissue samples were harvested following the execution of the mice. The blood in the heart chamber was rinsed with PBS. The samples were stored at a temperature of −80°C for subsequent use. All necessary permissions were obtained from the Ethics Committee of Harbin Medical University, and all procedures met the relevant regulatory standards.

2.2 Enzyme-linked immunosorbent assay

Commercially available enzyme-linked immunosorbent assay (ELISA) kits (Meimian Biotechnology, Jiangsu, China) were used to determine the levels of Tn-I. The instructions

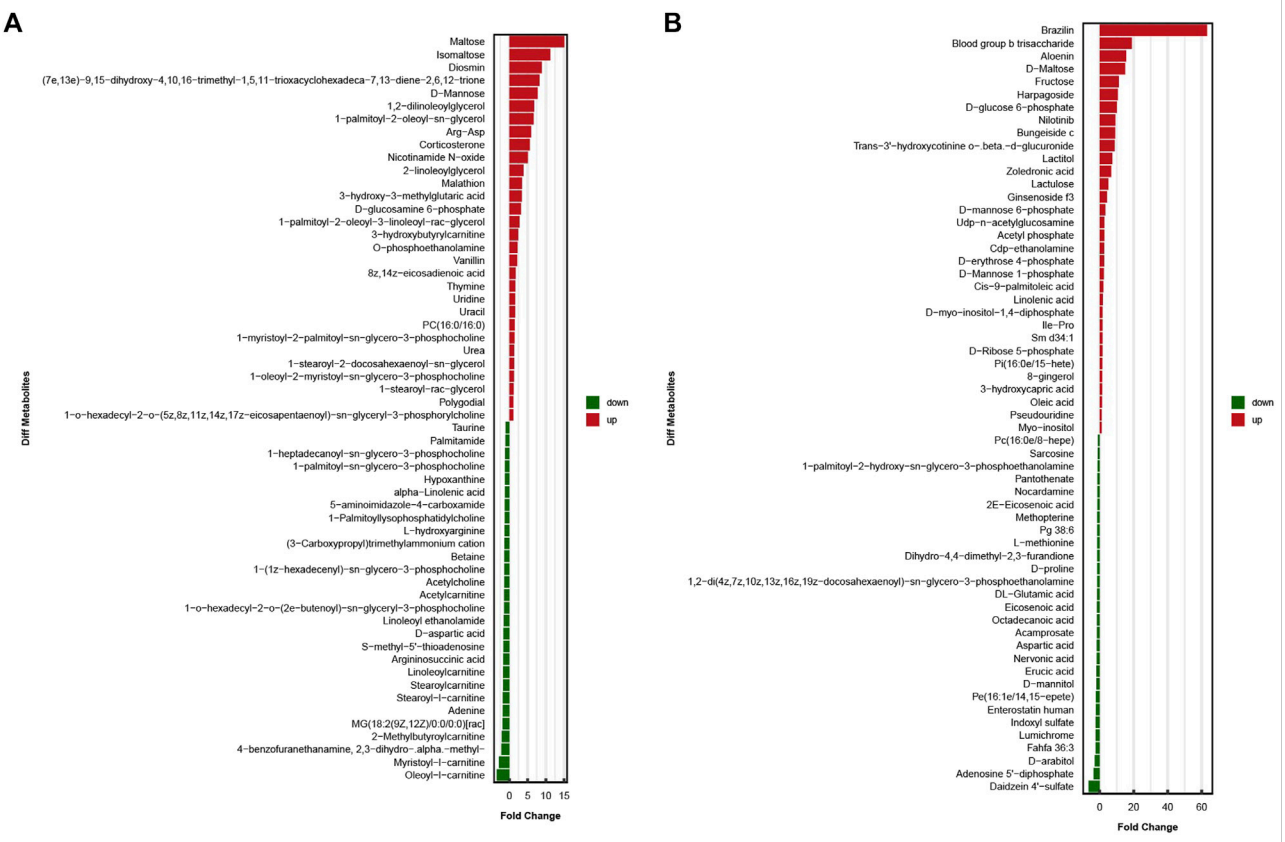


FIGURE 3
In two ionization modes, one-dimensional, and multi-dimensional analysis results of differential metabolites. **(A)**significant differences in metabolite expression in positive ionization modes. **(B)**significant differences in metabolite expression in negative ionization modes.

provided by the manufacturer were followed to conduct the studies.

2.3 Untargeted metabolomics studies

Heart tissue samples collected from mice belonging to the experimental group (n = 6) and the control group (n = 6) were slowly thawed at 4°C. Following this, the samples were treated with a pre-chilled solution consisting of methanol, water, and acetonitrile (methanol:acetonitrile:water = 2:2:1, v/v). The solution was vortexed, after which it was sonicated over a period of 30 min at a low temperature. Subsequently, the sample solution was allowed to stand at −20°C for 10 min, following which it was centrifuged at 14,000 g over a period of 20 min at 4°C. Subsequently, the supernatant was extracted, and it was dried under conditions of vacuum. The samples were analyzed using the mass spectrometry technique.

Analyses were performed using an UHPLC (1,290 Infinity LC, Agilent Technologies) coupled to a quadrupole time-of-flight (AB Sciex TripleTOF 6,600). For HILIC separation, samples were analyzed using a 2.1 mm × 100 mm ACQUIY UPLC BEH 1.7 μm column (waters, Ireland). In both ESI positive and negative modes, the mobile phase contained A = 25 mM ammonium acetate and 25 mM ammonium hydroxide in water and B = acetonitrile. The gradient was 85% B for 1 min and was linearly reduced to 65% in 11 min, and then was reduced to 40% in 0.1 min and kept for 4 min, and then increased to 85% in 0.1 min, with a 5 min re-equilibration period employed. For RPLC separation, a 2.1 mm × 100 mm ACQUIY UPLC HSS T3 1.8 μm column (waters, Ireland) was used. In ESI positive mode, the mobile phase contained A = water with 0.1% formic acid and B = acetonitrile with 0.1% formic acid; and in ESI negative mode, the mobile phase contained A = 0.5 mM ammonium fluoride in water and B = acetonitrile. The gradient was 1%B for 1.5 min and was linearly increased to 99% in 11.5 min and kept for 3.5 min. Then it was reduced to 1% in 0.1 min and a 3.4 min of re-equilibration period was employed. The gradients were at a flow rate of 0.3 ml/min, and the column

temperatures were kept constant at 25°C. A 2 μL aliquot of each sample was injected.

The ESI source conditions were set as follows: Ion Source Gas1 (Gas1) as 60, Ion Source Gas2 (Gas2) as 60, curtain gas (CUR) as 30, source temperature: 600°C, IonSpray Voltage Floating (ISVF) ± 5500 V. In MS only acquisition, the instrument was set to acquire over the m/z range 60–1,000 Da, and the accumulation time for TOF MS scan was set at 0.20 s/spectra. In auto MS/MS acquisition, the instrument was set to acquire over the m/z range 25–1,000 Da, and the accumulation time for product ion scan was set at 0.05 s/spectra. The product ion scan is acquired using information dependent acquisition (IDA) with high sensitivity mode selected. The parameters were set as follows: the collision energy (CE) was fixed at 35 V with ±15 eV; declustering potential (DP), 60 V (+) and −60 V (−); exclude isotopes within 4 Da, candidate ions to monitor per cycle: 10.

The extracted data were used for metabolite structure identification and subjected to data pre-processing techniques. Subsequently, the data quality was evaluated and analyzed.

2.4 Pre-processing of the metabolomics data

The MzXML files were generated from the raw MS data (wiff.scan files). ProteoWizard MSConvert was used for data conversion. Following this, the data were imported into the free XCMS software. The parameters for peak pick up were determined (centWave: m/z, 25 ppm; prefilter, c (10,100); peak width: c (10,60)). The parameters for peak grouping were also set (minfrac: 0.5; bw: 5; mzwid: 0.025). The isotopes and adducts were annotated using the Collection of Algorithms of Metabolite pRofile Annotation (CAMERA). The extracted ion features consisted of variables that were characterized by >50% of the non-zero measurements in at least one of the sets recorded. The accuracy of the m/z values (<25 ppm) and the mass spectroscopy–mass spectroscopy (MS/MS) spectral data were compared with those present in an internal database developed using authentic standards to analyze the metabolites.

2.5 Transcriptomics

The heart tissues of the mice belonging to the experimental and control groups were used for the extraction of total RNA. The process of sample extraction was performed using TRIzol. A bioanalyzer (Agilent 2,100) was used to determine the purity and concentration of the extracted RNA. The ribosomal RNA (rRNA) Removal Kit was used for ribosomal RNA removal. The rest of the total RNA samples were subjected to conditions of ionization to break down the samples into fragments that were 200–300 bp long. A random primer consisting of six bases and reverse transcriptase were used to synthesize the first complementary DNA (cDNA) strand. RNA was used as a template during the process. Subsequently, the second strand was generated using the first cDNA strand as the template. This process was followed to generate a specific library. The polymerase chain reaction (PCR) amplification process was used to increase the number of fragments in the library following the process of library construction. Subsequently, based on the library fragment size, the library selection process was conducted (library size: 450 bp). The quality of the libraries was determined using the Agilent 2,100 Bioanalyzer. This same system was also used to test the effective and total library concentrations. The amount of data required for the construction of the library and the effective concentration of the library were analyzed. The mixing of the libraries characterized by different index sequences was based on the results. The mixed libraries were diluted to 2 nM and deformed using alkali to form single-stranded libraries. The libraries were analyzed using the paired-end (PE) sequencing method (Next-Generation Sequencing (NGS); Illumina NovaSeq 6,000 sequencing platform) post the process of extraction and purification of RNA and library construction.

All raw data were filtered to obtain high-quality sequences. These sequences (Clean Data) were aligned with the reference genome. HISAT2 was used for sequence alignment, and this software could be accessed through http://ccb.jhu.edu/software/hisat2/index.shtml. The expression level of each gene was determined based on the alignment results. Subsequently, differential analysis of sample genes was performed using DESeq to identify DEGs satisfying the criteria of $|\log2FoldChange| > 1$ and $p < 0.05$. The ggplots2 package was used to plot the volcano plots for the DEGs.

2.6 Metabonomics–transcriptomics combination analysis

The differentially expressed metabolites (DEMs) and gens (DEGs) were extracted. The gens corresponding to the relevant enzymes were also extracted. The relevant data were obtained by analyzing the metabolite information presented in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. This database can be accessed through the website https://www.kegg.jp/dbget-bin/www_bfind?compound. Finally, DEMs and their associated DEGs were mapped with the corresponding metabolic pathways.

2.7 Statistical analysis

All statistical analyses were performed using SPSS 19.0, and the plots were generated using GraphPad Prism 8.0 (statistically significant results: $p < 0.05$). The metabolite-related data were analyzed using ropls (R package). Multiple algorithms were used

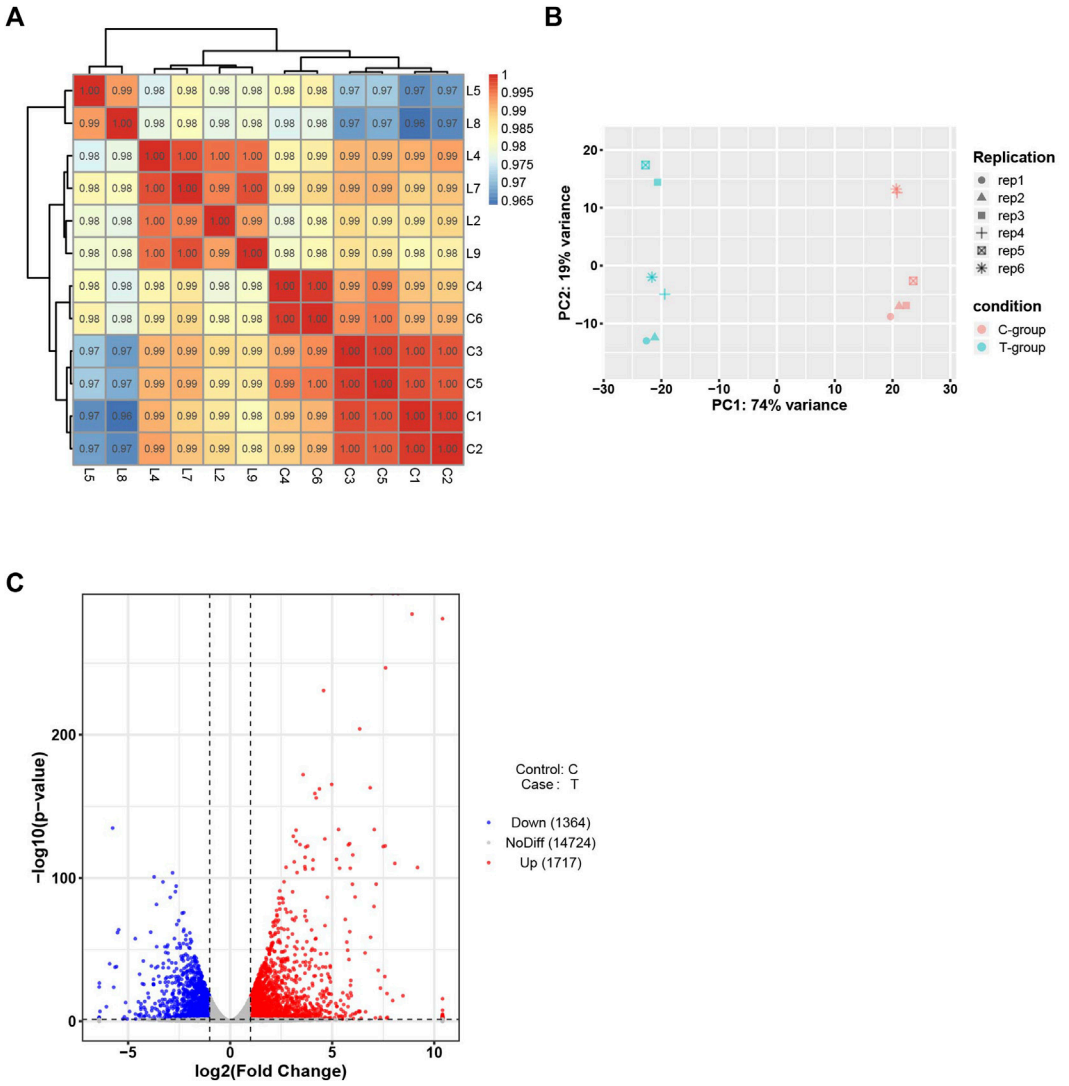


FIGURE 4
Bioinformatics analysis of RNA-seq data. **(A)** Sample correlation test. **(B)**PCA of mRNAs. **(C)** Volcano plot of mRNAs. C-group the control group (n = 6). T-group the SIMD group (n = 6).

to realize multivariate data analysis. The orthogonal partial least squares–discriminant analysis (OPLS–DA) and pareto-scaled principal component analysis (PCA) methods were used for data analysis. The 7-fold cross-validation method was used, and response permutation tests were conducted to determine the robustness of the model. For each variable associated with the OPLS–DA model, the variable importance in the projection (VIP) value was calculated. This helped determine the contribution of the variables toward the classification process. The student's t-test was conducted for all metabolites characterized with VIP values > 1. The significance of each metabolite was determined by conducting the tests at the univariate level.

3 Results

3.1 Sepsis-induced myocardial dysfunction model

Results obtained by conducting echocardiography tests suggested that the overall cardiac function of the members of the experimental group, and EF%, FS% recorded for the experimental group were significantly lower than those recorded for the control group ($p < 0.001$) (Figures 1A,B). LVESd recorded for the experimental group is significantly higher than the control group ($p < 0.001$) (Figure 1B). It was also observed that the circulating Tn-I level in the experimental

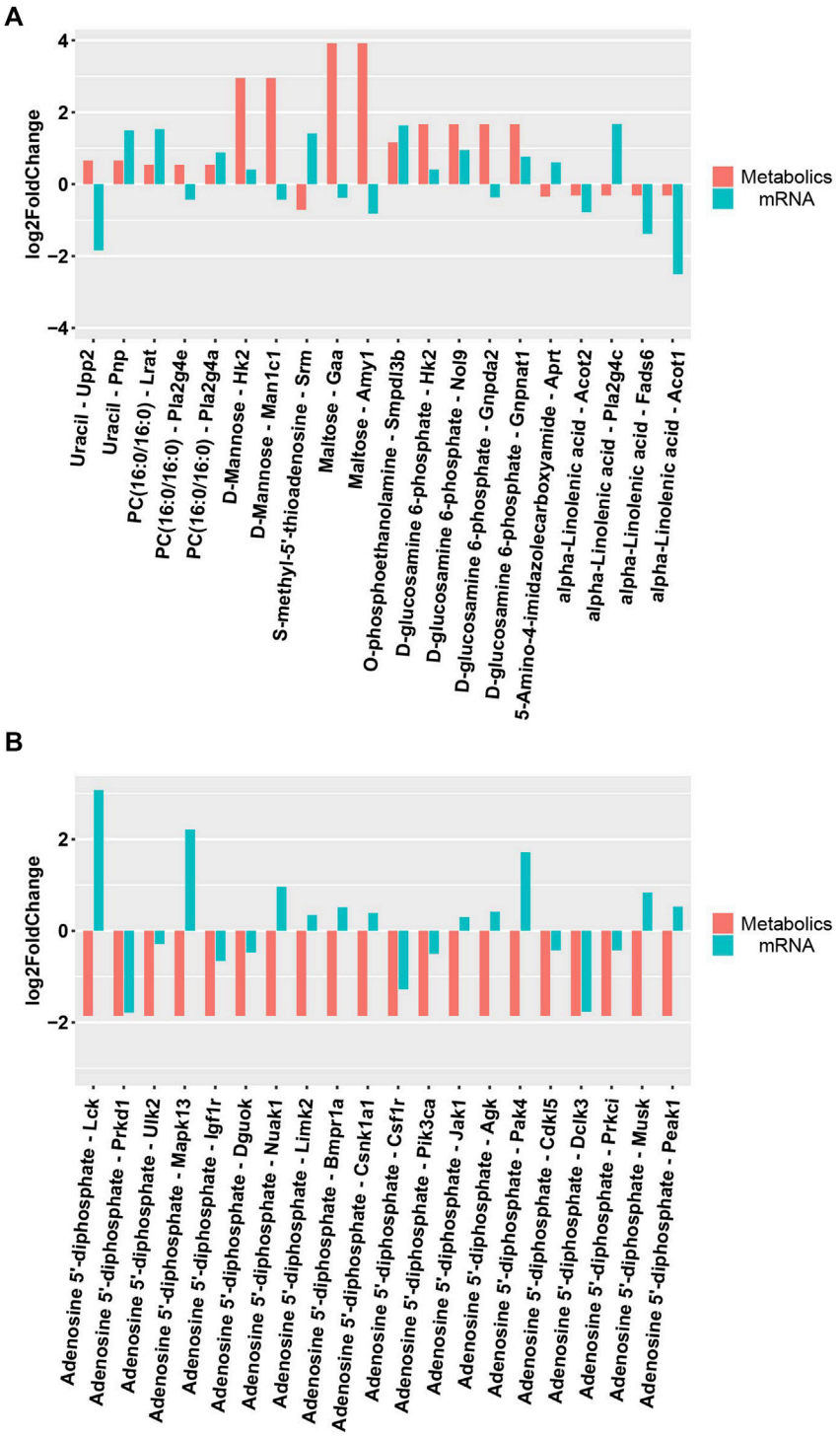


FIGURE 5
Differential expression of metabolites and related transcripts. **(A)**Differential expression of metabolites and related transcripts in positive ionization modes. **(B)** Differential expression of metabolites and related transcripts in negative ionization modes.

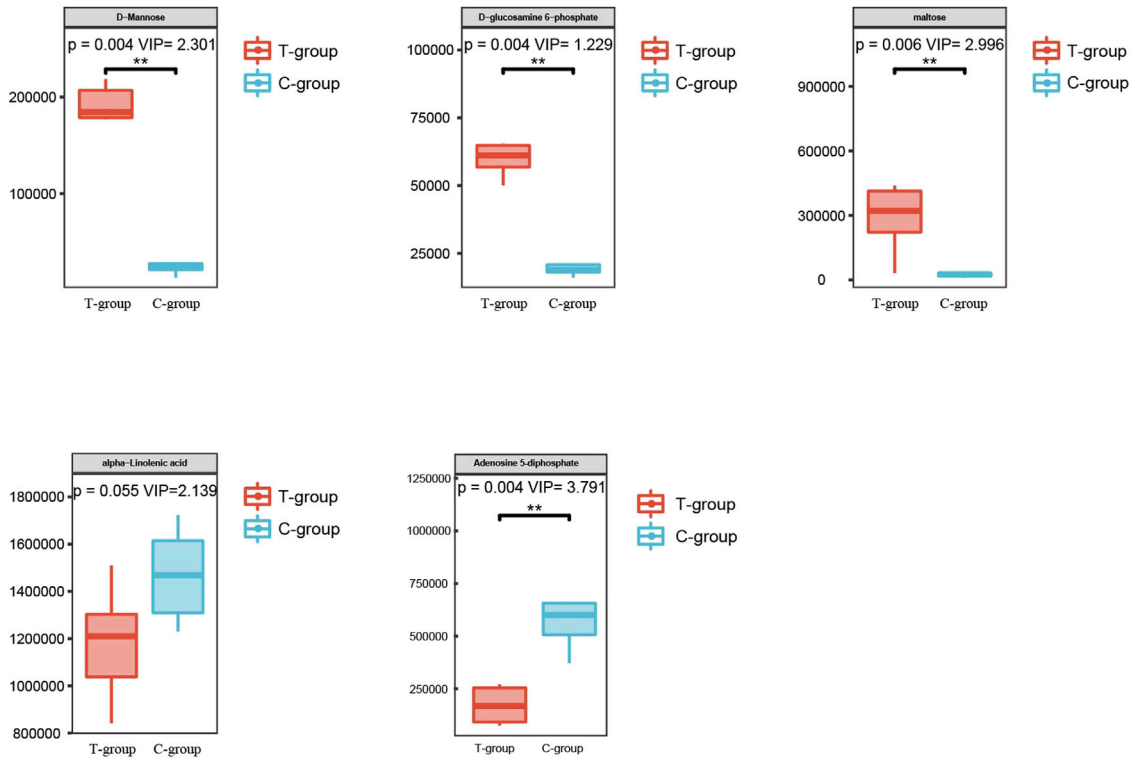


FIGURE 6
Metabolites with significant differences between the two groups.

group was significantly higher than the Tn-I level recorded for the control group ($p < 0.05$) (Figure 1C). These indicated the successful establishment of the SIMD model.

3.2 Metabolomics validation of the model

All the identified metabolites were analyzed using a multi-dimensional statistical analysis method. The OPLS–DA permutation test plot and the OPLS–DA score plot generated under both the positive and negative ion modes are shown in the Figure 2. It was observed that the model could be used to differentiate between one group of samples from the other, and overfitting could be avoided. This indicated the good robustness of the model.

3.3 Identification of differentially expressed metabolites

The results obtained under the positive and negative ion modes were combined, and a total of 1,027 metabolites were identified. Of all these samples, 390 metabolites were identified under the positive ion mode, and 637 metabolites were identified under the negative ion mode. Univariate and multi-dimensional analyses methods were used to screen 118 DEMs (criteria for OPLS–DA: VIP >1; $p < 0.05$). Among these, 58 significant DEMs were identified under the cationic mode, and 60 significant DEMs were identified under the anionic mode. The results obtained under these two modes are presented in the Figure 3.

3.4 Transcriptomics analysis of sepsis-induced my

ocardial dysfunction

High-throughput transcriptome analysis of the heart tissues in the experimental and control groups was performed. The correlation coefficients between the samples ranged from 0.8 to 1, indicating an extremely strong correlation. PCA results indicated high intra-group similarity between the samples in the experimental and control groups. Of the 3,081 DEGs recorded, down-regulation was observed for 1,364 DEGs, and up-regulation was observed for 1717 DEGs (Figure 4).

TABLE 1 Pathways of metabolites and related transcripts.

DEMs	DEGs	Pathway
D-Mannose	Hk2	Fructose and mannose metabolism
D-Glucosamine 6-phosphate	Gnpda2Hk2; Gnpnat1	Amino sugar and nucleotide sugar metabolism
Maltose	Gaa	Starch and sucrose metabolism
	Amy1	Carbohydrate digestion and absorption
Alpha-Linolenic acid	Acot1	Biosynthesis of unsaturated fatty acids
Adenosine 5′-diphosphate	Prkci	Platelet activation
	Igf1r; Pik3ca	AMPK signaling pathway
	Dguok	Purine metabolism

3.5 Metabolomics–transcriptomics combination analysis

DEMs obtained under the negative and positive ion modes and the transcriptome data were subjected to conditions of the metabolomics–transcriptomics combination analysis method. The change in the fold of the top 20 DEM–DEG pairs is shown in the Figure 5. Finally, multiple common metabolites were identified by analyzing the mice in both groups. The common metabolites were identified to be D-mannose, D-glucosamine 6-phosphate, maltose, alpha-linolenic acid, and adenosine 5′-diphosphate (Figure 6).

3.6 Differentially expressed metabolites and differentially expressed genes: Analysis of the kyoto encyclopedia of genes and genomes pathway

The DEMs and DEGs were mapped simultaneously to the KEGG pathway database to identify the common pathways associated with the DEMs and DEGs (Supplementary Tables S1–S9). The results are presented in the Table 1.

4 Discussion

Metabolomics allows for a more precise exploration of disease diagnosis and pathogenesis. The metabolomics–transcriptomics combination analysis method used helped us to identify significant DEMs between the two groups, including D-mannose, maltose, D-glucose 6-phosphate, alpha-linolenic acid, and adenosine 5′-diphosphate. Additionally, metabolite-related metabolic pathways were also investigated.

D-mannose, a common monosaccharide, is a digestive product of polysaccharides and glycoproteins. However, the amount of mannose present in the daily diet is significantly small. Hexokinase converts mannose to mannose-6-phosphate, which is then converted to fructose 6-phosphate by mannose phosphate isomerase. This eventually participates in the glycolytic pathway to produce lactic acid, glucose, and pentose (Wood and Cahill, 1963; Ganda et al., 1979). Elevated lactate levels indicate cellular dysfunction in patients with sepsis. Hyperlactataemia is a sign of severe sepsis and results in high mortality (Singer et al., 2016). Mannan-binding lectin (MBL) is a crucial complement component in the human body and is an important part of the processes associated with innate immunity. Infection caused by pathogenic microorganisms induces the secretion of MBL, which specifically recognizes and binds to mannose on the surface of microorganisms. This triggers complement activation and mediates the process of generation of inflammatory response (Fujita, 2002). It has been reported that in the sera of individuals with sepsis attributable to Gram-negative bacterial infections, MBL recognizes and binds to mannose on LPS to activate the complementary MBL pathway and initiate the body's innate immunity to participate in the inflammatory response (Fujita, 2002). This results in a significant reduction in the MBL levels. The results reported herein reveal that the mannose levels in the heart tissues of mice with LPS-induced SIMD were significantly higher than the mannose levels recorded for the control group. Additionally, the expression level of Hk2, a gene that mediates the process of D-mannose metabolism, was significantly high. A large amount of D-mannose was deposited in cardiac tissues, and this activated MBL to

trigger innate immune responses and induce an inflammatory response.

D-glucosamine 6-phosphate, a type of glucosamine, is an important energy source for many bacteria present in the body. It is also an important component of bacterial cell walls (Matsuura, 2013). Moreover, D-glucosamine 6-phosphate is also associated with the virulence of some bacteria (Kawada-Matsuo et al., 2016).

Maltose is a disaccharide that is produced in the body during starch catabolism. It can be metabolized to form two glucose molecules. Researchers have previously used magnetic resonance imaging-based metabolomics techniques to study conditions of sepsis. The results revealed that the maltose content in the metabolites of patients with sepsis was significantly lower than the maltose contents of patients not suffering from sepsis. However, no such changes were observed in the sham-operated and control groups (Bakalov et al., 2016). This suggested that the significant reduction in the maltose content was associated with the chronic depletion of the long-term inflammatory response. We used an early-state 6 h animal model to conduct the studies. The experimental results suggested a significant increase in the maltose content. However, whether the maltose content changes as sepsis progresses needs to be further investigated.

Alpha-linolenic acid (ALA) is a type of omega-3 essential fatty acid. It is a polyunsaturated fatty acid with three double bonds. It has been previously reported that ALA and its metabolites significantly inhibit the generation of LPS-induced inflammatory response, and their action results in a decrease in the rate of cellular reactive oxygen species (ROS) and NO production. These could also inhibit the expression of iNOS and TNF-α in cells and reduce the mortality in mice suffering from endotoxin-mediated septic shock (Kumar et al., 2016).

Mitochondrial dysfunction is an adverse mechanism associated with the cardiac dysfunction observed in patients with sepsis (Ravikumar et al., 2021). It results in the inability of the body to synthesize sufficient amounts of adenosine triphosphate (ATP) to provide energy for the heart (Wasyluk et al., 2021). Insufficient ATP synthesis also results in a reduction in the adenosine diphosphate (ADP) content in cardiac tissues. This result agrees with the results reported herein. It was also observed that the amount of adenosine 5′-diphosphate in the heart tissues of mice in the experimental group was significantly lower than the content of adenosine 5′-diphosphate in the heart tissues of mice belonging to the control group.

DEMs and DEGs were linked to mannose metabolism, aminoglycan metabolism, starch metabolism, unsaturated fatty acid biosynthesis, platelet activation, purine metabolism, and AMP-activated protein kinase (AMPK) signaling pathways. AMPK significantly affects the process of cellular energy homeostasis (Carling et al., 2011). Stressors such as hypoglycemia, hypoxia, and ischemia that remarkably deplete ATP can activate this pathway (Canto and Auwerx, 2010; Hardie, 2011; Mihaylova and Shaw, 2011), which positively regulates the signaling pathways that replenish cellular ATP supply.

There are some limitations to this study. Although LPS is an important myocardial inhibitory factor, the predisposing factors for cardiac dysfunction are not limited to Gram-negative bacteria-induced sepsis. Therefore, we will further explore the metabolic alterations and pathogenic mechanisms associated with Gram-positive bacteria-induced SIMD in the future.

5 Conclusion

In summary, significant changes in metabolites occur in the cardiac tissues of patients suffering from SIMD. These changes are primarily associated with mannose metabolism, aminoglycan metabolism, starch metabolism, unsaturated fatty acid biosynthesis, platelet activation, purine metabolism, and AMPK signaling pathways. The problems associated with the aberrant metabolic events can be addressed to help improve the prognoses of patients with SIMD and provide new insights into the processes associated with diagnosis and disease management.

Data availability statement

The authors acknowledge that the data presented in this study must be deposited and made publicly available in an acceptable repository, prior to publication. Frontiers cannot accept a manuscript that does not adhere to our open data policies.

Ethics statement

The animal study was reviewed and approved by Harbin Medical University Cancer Hospital ethics committee.

Author contributions

XJ, KY and CW participated in the design, interpretation of the studies and analysis of the data and review of the manuscript. XJ designed the research and wrote the manuscript. YP and XM performed the data analysis. XL reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This project was funded by the National Natural Science Foundation of China (Nos. 81770276).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.967397/full#supplementary-material

References

Bakalov, V., Amathieu, R., Triba, M. N., Clement, M. J., Reyes Uribe, L., Le Moyec, L., et al. (2016). Metabolomics with nuclear magnetic resonance spectroscopy in a *Drosophila melanogaster* model of surviving sepsis. *Metabolites* 6 (4), E47. doi:10.3390/metabo6040047

Canto, C., and Auwerx, J. (2010). AMP-activated protein kinase and its downstream transcriptional pathways. *Cell. Mol. Life Sci.* 67 (20), 3407–3423. doi:10.1007/s00018-010-0454-z

Carling, D., Mayer, F. V., Sanders, M. J., and Gamblin, S. J. (2011). AMP-Activated protein kinase: nature's energy sensor. *Nat. Chem. Biol.* 7 (8), 512–518. doi:10.1038/nchembio.610

Fujita, T. (2002). Evolution of the lectin-complement pathway and its role in innate immunity. *Nat. Rev. Immunol.* 2 (5), 346–353. doi:10.1038/nri800

Ganda, O. P., Soeldner, J. S., Gleason, R. E., Cleator, I. G., and Reynolds, C. (1979). Metabolic effects of glucose, mannose, galactose, and fructose in man. *J. Clin. Endocrinol. Metab.* 49 (4), 616–622. doi:10.1210/jcem-49-4-616

Gao, D. N., Zhang, Y., Ren, Y. B., Kang, J., Jiang, L., Feng, Z., et al. (2015). Relationship of serum mannose-binding lectin levels with the development of sepsis: A meta-analysis. *Inflammation* 38 (1), 338–347. doi:10.1007/s10753-014-0037-5

Griffin, J. L. (2006). The cinderella story of metabolic profiling: Does metabolomics get to go to the functional genomics ball? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361 (1465), 147–161. doi:10.1098/rstb.2005.1734

Hardie, D. G. (2011). AMP-Activated protein kinase: An energy sensor that regulates all aspects of cell function. *Genes Dev.* 25 (18), 1895–1908. doi:10.1101/gad.17420111

Hollenberg, S. M., and Singer, M. (2021). Pathophysiology of sepsis-induced cardiomyopathy. *Nat. Rev. Cardiol.* 18 (6), 424–434. doi:10.1038/s41569-020-00492-2

Hwang, J. S., Kim, K. H., Park, J., Kim, S. M., Cho, H., Lee, Y., et al. (2019). Glucosamine improves survival in a mouse model of sepsis and attenuates sepsis-induced lung injury and inflammation. *J. Biol. Chem.* 294 (2), 608–622. doi:10.1074/jbc.RA118.004638

Kawada-Matsuo, M., Oogai, Y., and Komatsuzawa, H. (2016). Sugar allocation to metabolic pathways is tightly regulated and affects the virulence of Streptococcus mutans. *Genes (Basel)* 8 (1), E11. doi:10.3390/genes8010011

Kumar, N., Gupta, G., Anilkumar, K., Fatima, N., Karnati, R., Reddy, G. V., et al. (2016). 15-Lipoxygenase metabolites of alpha-linolenic acid, [13-(S)-HPOTrE and 13-(S)-HOTrE], mediate anti-inflammatory effects by inactivating NLRP3 inflammasome. *Sci. Rep.* 6, 31649. doi:10.1038/srep31649

Martin, L., Derwall, M., Al Zoubi, S., Zechendorf, E., Reuter, D. A., Thiemermann, C., et al. (2019). The septic heart: Current understanding of molecular mechanisms and clinical implications. *Chest* 155 (2), 427–437. doi:10.1016/j.chest.2018.08.1037

Matsuura, M. (2013). Structural modifications of bacterial lipopolysaccharide that facilitate gram-negative bacteria evasion of host innate immunity. *Front. Immunol.* 4, 109. doi:10.3389/fimmu.2013.00109

Mihaylova, M. M., and Shaw, R. J. (2011). The AMPK signalling pathway coordinates cell growth, autophagy and metabolism. *Nat. Cell Biol.* 13 (9), 1016–1023. doi:10.1038/ncb2329

Morris, D. R. (2009). Ribosomal footprints on a transcriptome landscape. *Genome Biol.* 10 (4), 215. doi:10.1186/gb-2009-10-4-215

Neugebauer, S., Giamarellos-Bourboulis, E. J., Pelekanou, A., Marioli, A., Baziaka, F., Tsangaris, I., et al. (2016). Metabolite profiles in sepsis: Developing prognostic tools based on the type of infection. *Crit. Care Med.* 44 (9), 1649–1662. doi:10.1097/CCM.0000000000001740

Patti, G. J., Yanes, O., and Siuzdak, G. (2012). Innovation: Metabolomics: The apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 13 (4), 263–269. doi:10.1038/nrm3314

Ping, F., Guo, Y., Cao, Y., Shang, J., Yao, S., Zhang, J., et al. (2019). Metabolomics analysis of the renal cortex in rats with acute kidney injury induced by sepsis. *Front. Mol. Biosci.* 6, 152. doi:10.3389/fmolb.2019.00152

Ping, F., Li, Y., Cao, Y., Shang, J., Zhang, Z., Yuan, Z., et al. (2021). Metabolomics analysis of the development of sepsis and potential biomarkers of sepsis-induced acute kidney injury. *Oxid. Med. Cell. Longev.* 2021, 6628847. doi:10.1155/2021/6628847

Ravikumar, N., Sayed, M. A., Poonsuph, C. J., Sehgal, R., Shirke, M. M., and Harky, A. (2021). Septic cardiomyopathy: From basics to management choices. *Curr. Probl. Cardiol.* 46 (4), 100767. doi:10.1016/j.cpcardiol.2020.100767

Rochfort, S. (2005). Metabolomics reviewed: A new "omics" platform technology for systems biology and implications for natural products research. *J. Nat. Prod.* 68 (12), 1813–1820. doi:10.1021/np050255w

Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., et al. (2020). Global, regional, and national sepsis incidence and mortality, 1990-2017: Analysis for the global burden of disease study. *Lancet* 395 (10219), 200–211. doi:10.1016/S0140-6736(19)32989-7

She, H., Tan, L., Zhou, Y., Zhu, Y., Ma, C., Wu, Y., et al. (2022). The landscape of featured metabolism-related genes and imbalanced immune cell subsets in sepsis. *Front. Genet.* 13, 821275. doi:10.3389/fgene.2022.821275

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 315 (8), 801–810. doi:10.1001/jama.2016.0287

Song, P., Shen, D. F., Meng, Y. Y., Kong, C. Y., Zhang, X., Yuan, Y. P., et al. (2020). Geniposide protects against sepsis-induced myocardial dysfunction through AMPKα-dependent pathway. *Free Radic. Biol. Med.* 152, 186–196. doi:10.1016/j.freeradbiomed.2020.02.011

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* 270 (5235), 484–487. doi:10.1126/science.270.5235.484

Virlon, B., Cheval, L., Buhler, J. M., Billon, E., Doucet, A., and Elalouf, J. M. (1999). Serial microanalysis of renal transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 96 (26), 15286–15291. doi:10.1073/pnas.96.26.15286

Wang, Z., Liu, M., Ye, D., Ye, J., Wang, M., Liu, J., et al. (2021). Il12a deletion aggravates sepsis-induced cardiac dysfunction by regulating macrophage polarization. *Front. Pharmacol.* 12, 632912. doi:10.3389/fphar.2021.632912

Wasyluk, W., Nowicka-Stazka, P., and Zwolak, A. (2021). Heart metabolism in sepsis-induced cardiomyopathy-unusual metabolic dysfunction of the heart. *Int. J. Environ. Res. Public Health* 18 (14), 7598. doi:10.3390/ijerph18147598

Wood, F. C., Jr., and Cahill, G. F., Jr. (1963). Mannose utilization in man. *J. Clin. Invest.* 42, 1300–1312. doi:10.1172/JCI104814

Yang, H., and Zhang, Z. (2021). Sepsis-induced myocardial dysfunction: The role of mitochondrial dysfunction. *Inflamm. Res.* 70 (4), 379–387. doi:10.1007/s00011-021-01447-0

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 29 August 2022
DOI 10.3389/fmolb.2022.916128

OPEN ACCESS

EDITED BY
Sergio Oller Moreno,
University Medical Center Hamburg-Eppendorf, Germany

REVIEWED BY
Andre Kahles,
ETH Zürich, Switzerland
Timuçin Avşar,
Bahçeşehir University, Turkey

*CORRESPONDENCE
Maria C. Jenmalm,
maria.jenmalm@liu.se
Mika Gustafsson,
mika.gustafsson@liu.se

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 08 April 2022
ACCEPTED 25 July 2022
PUBLISHED 29 August 2022

CITATION
Magnusson R, Rundquist O, Kim MJ, Hellberg S, Na CH, Benson M, Gomez-Cabrero D, Kockum I, Tegnér JN, Piehl F, Jagodic M, Mellergård J, Altafini C, Ernerudh J, Jenmalm MC, Nestor CE, Kim M-S and Gustafsson M (2022), RNA-sequencing and mass-spectrometry proteomic time-series analysis of T-cell differentiation identified multiple splice variants models that predicted validated protein biomarkers in inflammatory diseases.
Front. Mol. Biosci. 9:916128.
doi: 10.3389/fmolb.2022.916128

RNA-sequencing and mass-spectrometry proteomic time-series analysis of T-cell differentiation identified multiple splice variants models that predicted validated protein biomarkers in inflammatory diseases

Rasmus Magnusson[1†], Olof Rundquist[1†], Min Jung Kim[2], Sandra Hellberg[3], Chan Hyun Na[4], Mikael Benson[5], David Gomez-Cabrero[6], Ingrid Kockum[7], Jesper N. Tegnér[8,9,10], Fredrik Piehl[7], Maja Jagodic[7], Johan Mellergård[11,12], Claudio Altafini[13], Jan Ernerudh[12,14], Maria C. Jenmalm[3]*, Colm E. Nestor[3], Min-Sik Kim[15] and Mika Gustafsson[1]*

[1]Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden, [2]Department of Applied Chemistry, College of Applied Sciences, Kyung Hee University, Yongin, South Korea, [3]Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden, [4]Department of Neurology, Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, United States, [5]Centre for Personalised Medicine, Linköping University, Linköping, Sweden, [6]Navarrabiomed, Complejo Hospitalario de Navarra, Universidad Pública de Navarra, IdiSNA, Pamplona, Spain, [7]Department of Clinical Neuroscience, Center for Molecular Medicine, Karolinska Institute, Stockholm, Sweden, [8]Biological and Environmental Sciences and Engineering Division, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, [9]Unit of Computational Medicine, Department of Medicine, Solna, Center for Molecular Medicine, Karolinska Institutet, Solna, Sweden, [10]Science for Life Laboratory, Solna, Sweden, [11]Department of Neurology, Linköping University, Linköping, Sweden, [12]Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden, [13]Department of Automatic Control, Linköping University, Linköping, Sweden, [14]Department of Clinical Immunology and Transfusion Medicine, Linköping University, Linköping, Sweden, [15]Department of New Biology, Daegu Gyeongbuk Institute of Science and Technology, Daegu, South Korea

Profiling of mRNA expression is an important method to identify biomarkers but complicated by limited correlations between mRNA expression and protein abundance. We hypothesised that these correlations could be improved by mathematical models based on measuring splice variants and time delay in protein translation. We characterised time-series of primary human naïve CD4$^+$ T cells during early T helper type 1 differentiation with RNA-sequencing and mass-spectrometry proteomics. We performed computational time-series analysis in this system and in two other key human and murine immune cell types. Linear mathematical mixed time delayed splice variant models were used to predict protein abundances, and the models were validated using out-of-

sample predictions. Lastly, we re-analysed RNA-seq datasets to evaluate biomarker discovery in five T-cell associated diseases, further validating the findings for multiple sclerosis (MS) and asthma. The new models significantly out-performing models not including the usage of multiple splice variants and time delays, as shown in cross-validation tests. Our mathematical models provided more differentially expressed proteins between patients and controls in all five diseases. Moreover, analysis of these proteins in asthma and MS supported their relevance. One marker, sCD27, was validated in MS using two independent cohorts for evaluating response to treatment and disease prognosis. In summary, our splice variant and time delay models substantially improved the prediction of protein abundance from mRNA expression in three different immune cell types. The models provided valuable biomarker candidates, which were further validated in MS and asthma.

KEYWORDS

proteomics, RNA-seq, T-cell differentiation, biomarkers, multiple sclerosis

1 Introduction

Identifying biomarkers that can be used in clinical routine to diagnose patients, monitor disease and response to treatment is required for more precision-based medicine (Mayeux, 2004; Chase Huizar et al., 2020). The complex etiology behind many diseases, potentially involving multiple genes and proteins across multiple cell types, renders biomarker discovery for most complex diseases challenging (Rifai et al., 2006).

Proteins are regarded as optimal biomarkers as they are often directly connected to patho-physiological processes as well as serving as targets for many therapeutic interventions (Ek et al., 2021). Whereas measuring global protein levels in a clinical setting remains challenging, gene expression profiling can be readily performed on the limited amount of material obtained from most clinical sampling procedures. Combinations of mRNAs can have high diagnostic efficacy in multiple diseases (Gustafsson et al., 2014; Mao et al., 2018; Gawel et al., 2019; Cha et al., 2020). Ideally, mRNA profiling of clinical samples could be used to identify protein biomarkers for diagnoses, subtyping of diseases and evaluating treatment response.

mRNA expression has often been used to determine corresponding protein levels, even though the accuracy of such estimations can be very imprecise (Gygi et al., 1999; Fortelny et al., 2017). Indeed, the correlation between mRNA and protein expression is often poor (Gygi et al., 1999; de Sousa Abreu et al., 2009; Maier et al., 2009; Vogel and Marcotte, 2012; Fortelny et al., 2017), which becomes highly problematic when using mRNA expression as proxy for protein levels. Several strategies have been proposed to circumvent this issue using more dynamic approaches, as compared to steady-state approximations, accounting for example for spatial and temporal variations in both mRNA and protein expression (Liu et al., 2016; Kuchta et al., 2018). The discrepancy between mRNA and protein abundance is also due to several other factors, including but not limited to differences in the rates of translation and degradation between proteins and cell types (Wethmar et al., 2010). The large number of potential transcript isoforms that can be generated from the same gene due to alternative splicing as well as cell type-specific differences in splice variant use represent additional layers of complexity that complicate the correlation between mRNA to protein (Barbosa-Morais et al., 2012; Floor and Doudna, 2016). To our knowledge, leveraging the contribution and dynamics of different splice variants to infer protein abundance remains largely unexplored.

Here, we developed a novel method incorporating time delay and splice variants to improve protein level inference from mRNA expression. To test our approach, we performed RNA-seq and mass spectrometry proteomics analysis during early human T_H1 differentiation and used a machine learning modelling approach to infer the relationship between mRNA and protein abundance. T_H differentiation is an optimal model system to dissect the relationship between mRNA and protein as 1) primary human naïve T_H (NT_H) cells can be isolated with high purity and in large quantity from human blood (ii), all NT_H cells are synchronised in the G_1 phase of the cell cycle, further reducing inter-cell heterogeneity (Sprent and Tough, 1994) and 3) easy access to large quantities of material enabling relative quantification of mRNA and associated protein abundance to be assayed over time (Schmidt et al., 2018). Moreover, T_H cells are important regulators of immunity and thereby associated with many complex diseases, and T_H1 differentiation itself is pathogenetically relevant in several diseases (Raphael et al., 2015). The utilised models were based on a time delayed linear model between mRNA splice variants of the same gene and protein levels. We generalised the model by applying it

10.3389/fmolb.2022.916128

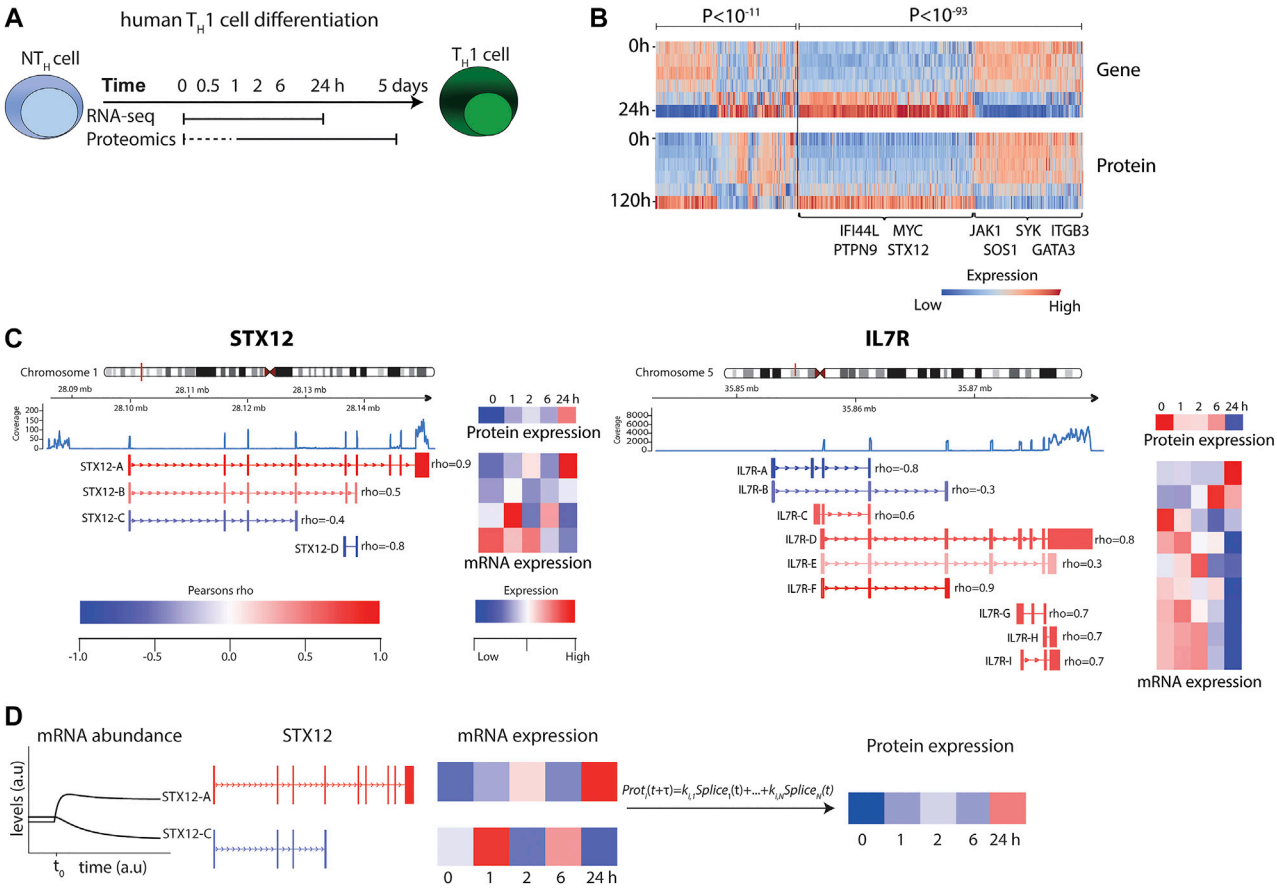


FIGURE 1
RNA-seq and mass-spectrometry analysis of T_H1 differentiation revealed highly variable correlations. **(A)** Experimental design. **(B)** Heat map of transcript and protein abundance dynamics in genes that show significant negative (left) and positive (right) correlations. Genes of particular relevance for T cells and T cell differentiation are highlighted in the figure. **(C)** Examples of transcript splice variants showing that both *STX12* (left) and *IL7R* (right) were significantly negatively and positively correlated with protein levels. **(D)** Illustration of the modelling procedure for resolving the poor correlation, using *STX12* as an example.

onto recent data from human regulatory T (T_{reg}) cell and murine B cell differentiation. By combining the strength of time-series analysis and RNA-sequencing, we noted a much better agreement between our mRNA-based measures and proteomics. To test our models, we showed the potential clinical usefulness by predicting potential biomarkers in five complex diseases using our derived models. Analysis of these predicted proteins in asthma and multiple sclerosis (MS) supported their biological relevance. Finally, we validated one of the predicted biomarkers, sCD27, using two independent cohorts of MS patients, which showed a remarkably better stratification between patients and controls than any of our previously reported protein biomarkers. The application of our approach to multiple different cell types, species and diseases shows its general applicability to increase the power of mRNA-based studies for biomarker discovery.

2 Materials and methods

2.1 Isolation of naïve CD4$^+$ T helper cells and T_H1 polarization

Peripheral blood mononuclear cells (PBMC) were isolated from blood donor derived buffy coats (n = 12), purchased at the blood bank facility at Linköping University Hospital, through gradient centrifugation using Lymphoprep™ (Axis Shields Diagnostics, Dundee, Scotland). Naïve CD45RA$^+$ CD4$^+$ T cells were isolated with negative immunomagnetic selection using the "Naive CD4$^+$ T Cell Isolation Kit II, human" (Miltenyi Biotec, Bergisch Gladbach, Germany) according to the instructions provided by the manufacturer. Cells were suspended in RPMI 1640 media containing L-glutamine, 10%FBS and 1% Penicillin/Streptomycin mixture (all from Gibco, Thermo Fisher Scientific, Waltham, MA, United States) and subsequently activated and

polarized towards T_H1 using Dynabeads™ Human T-Activator CD3/CD28 (1 bead/cell) (Dynal AS, Lillestøm, Norway), 5 ng/µl recombinant human IL-12p70, 10 ng/µl recombinant human IL-2 and 5 µg/µl anti-IL-4 antibodies (clone MAB204; all three from Bio-Techne, Minneapolis, MN, United States). The cells were cultured and differentiated at 37°C, with 5% CO_2 for 0 min, 0.5, 1, 2, 6 and 24 h for RNA-seq and 0 min, 1, 2, 6, 24 h and 5 days for proteomics (Figure 1A). The earliest time point for the RNA-seq time series was determined based on the change in expression of *IL2, IFNG* and *TBX21* at 3, 5, 10, 15, 30 and 60 min of T_H1 differentiation, measured by qPCR, where the expression of *IL2* and *IFNG* was significantly increased after 30 and 60 min ($p < 0.05$, Student's t-test) (See Supplementary Methods and Supplementary Figure S1). After cell culture, the cells were processed for RNA and protein extraction. An overview of the study is shown in Figure 1A and Supplementary Figure S2.

2.2 RNA-sequencing

2.2.1 Extraction of RNA

RNA was isolated using the ZR-Duet DNA/RNA kit (Zymo Research, Irvine, CA, United States) following the protocol provided by the manufacturer. The RNA was stored at −80°C until library preparation.

2.2.2 Library preparation and sequencing

The RNA library preparation and subsequent RNA-sequencing (RNA-seq) were carried out by the Beijing Genomics Institute (https://www.bgi.com/global/). Library preparation was performed using the TruSeq RNA Library Prep Kit v2 (Illumina, San Diego, CA, United States). Each sample was sequenced to the depth of 40 million reads per samples with pair end sequencing and a read length of 100 bp on an Illumina 2500 instrument (Illumina).

2.2.3 RNA-seq analysis

All RNA-seq data, both in-house and public, were processed similarly using the following pipeline: Sample qualities were assessed with fastQC (Version 0.11.8) and the mRNA reads were subsequently aligned using STAR (version 2.6.0c) (Dobin et al., 2013), with the parameter "--outSAMstrandField intronMotif" and "--out Filter Intron Motifs Remove Noncanonical," to the "Homo_sapiens.GRCh37.75.dna.primary_assembly.fa" from Ensemble. The resulting read alignment bam files were assembled into transcripts with StringTie (version 1.3.4d) (Pertea et al., 2015), with default parameters, using the GRCh37.75 gtf annotation from Ensemble. To evaluate mRNA to protein relationship, the mRNA reads were mapped to the mass spectrometry signal of protein abundance using the Homo.sapiens and Mus.musculus package in R (BC., T., 2015a; BC., T., 2015b). Correlations were calculated using Pearson correlations across gene expressions, i.e., one coefficient per gene.

2.3 Mass spectrometry

2.3.1 Protein extraction

The cells were thawed and resuspended in 100 µl of 8 M Urea in 40 mM Tris-HCl (pH 7.6) (Sigma-Aldrich, Saint Louis, MO, United States). Ten million cells per time point and biological replicate were pooled from 3–5 samples from different individuals to reach the necessary amount of material required for subsequent analysis steps. In total, cells were isolated from 12 different individuals to achieve the necessary amount of material. The suspension was sonicated using focus sonicator (Sonic Dismembrator 500, Thermo Fisher Scientific, Waltham, MA, United States) for 3 cycles of 10 s pulse with 10 s intervals at 10% of power. After sonication, a magnetic rack was used to remove the T-Activator beads used for the polarization. Protein concentration was measured using the Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific). 40 ug of each sample were used for digestion.

2.3.2 In solution digestion

Reduction and alkylation of disulfide bonds on proteins were carried out using 1 M dithiothreitol (Roche, Switzerland; final sample concentration 10 mM) for 45 min and 1 M Iodoacetamide (Sigma-Aldrich; final sample concentration 30 mM) for 30 min in a dark, respectively. Following alkylation and reduction, the samples were diluted with ammonium bicarbonate buffer (pH 8.0) until the urea concentration was 1 M (Sigma-Aldrich). The proteins were digested with trypsin (MS grade; Promega, Madison, WI, United States) overnight at 37°C at an enzyme to protein ratio of 1:20. Finally, the peptides were acidified with 100% Trifluoroacetic acid (TFA; Sigma-Aldrich) to a final concentration of 1% TFA and then desalted using macro spin columns (Harvard apparatus, Holliston, MA, United States).

2.3.3 TMT labeling

Peptides were labeled with 6-plex TMT reagent using manufacturer's protocol with some modification (Thermo Fisher Scientific). The six peptide samples from each time series were resuspended in 100 µl of 100 mM TEAB buffer (pH 8.0; Sigma-Aldrich) and a unit of each TMT reagent was resuspended in 40 µl of acetonitrile. Subsequently, the prepared TMT reagent was transferred to the peptide sample and then vortexed. The samples were incubated for 2 h at room temperature (RT). The labelled peptide samples from each time series were pooled and concentrated by vacuum centrifugation. The labelled sample was resuspended 100 µl with 10 mM ammonium formate (Sigma-Aldrich) in water (pH 10).

2.3.4 High pH fractionation

The TMT labelled samples were separated using an analytical column (Xbridge, Waters, MA, United States; C18, 5 µm, 4.6 mm × 250 mm) on the Agilent 1200 series HPLC system (Agilent Technologies, Santa Clara, CA, United States). Peptides were eluted using following gradient over 115 min: 0–10 min 0% B, 10–20 min 5% B, 20–80 min 35% B, 80–95 min 70% B, 95–105 min 70% B, 105–115 min 0% B; 10 mM ammonium formate (pH 10; Sigma-Aldrich) was mobile phase A, and 10 mM ACN (pH 10) was mobile phase B. The 96 fractions were added up into 24 fractions, vacuum dried and stored at −80°C after desalting.

2.3.5 LC-MS analysis

The fractionated peptides were analysed on an Orbitrap Fusion Lumos Tribrid Mass Spectrometer (Thermo Fisher Scientific) coupled with the Easy-nLC 1200 nano-flow liquid chromatography system (Thermo Fisher Scientific). The peptides from each fraction were reconstituted in 0.1% formic acid and loaded on an Acclaim PepMap100 Nano-Trap Column (100 µm × 2 cm; Thermo Fisher Scientific) packed with 5 µm C18 particles at a flow rate of 5 µl per minute. Peptides were resolved at 250-nl/min flow rate using a linear gradient of 10%–35% solvent B (0.1% formic acid in 95% acetonitrile) over 95 min on an EASY-Spray column (50 cm × 75 µm ID), PepMap RSLC C18 and 2 µm C18 particles (Thermo Fisher Scientific), which was fitted with an EASY-Spray ion source that was operated at a voltage of 2.3 kV. Mass spectrometry analysis was carried out in a data-dependent manner with a full scan in the mass-to-charge ratio (*m/z*) range of 350 to 1,800 in the "Top Speed" setting, 3 seconds per cycle. MS1 and MS2 were acquired for the precursor ions and the peptide fragmentation ions, respectively. MS1 scans were measured at a resolution of 120,000 at an *m/z* of 200. MS2 scan was acquired by fragmenting precursor ions using the higher-energy collisional dissociation method and detected at a mass resolution of 30,000, at an *m/z* of 200. Automatic gain control for MS1 was set to one million ions and for MS2 was set to 0.1 million ions. A maximum ion injection time was set to 50 ms for MS1 and 100 ms for MS2. Higher-energy collisional dissociation was set to 35 for MS2. Precursor isolation window was set to 0.7 *m/z*. Dynamic exclusion was set to 35 s, and singly charged ions were rejected. Internal calibration was carried out using the lock mass.

2.3.6 Peptide and protein identification

The obtained data were analysed using MaxQuant (version 1.6.0.1). MS raw data were searched using Andromeda algorithm with matching to the Uniprot human reference (released in November 2017). A specificity of trypsin was determined at up to 2 missed cleavages. In modification, carbamidomethylation, TMT 6-plex modification at lysine and N-termination were set as the fixed modifications, and oxidation of methionine was set as a variable modification. The false discovery rate (FDR) for peptide level was evaluated to 0.01 for removing false positive data. For highly confident quantifications of protein, protein ratios were calculated from two or more unique quantitative peptides in each replicate. Data was normalized and removed contaminant and razor peptide. To enrich differentially expressed proteins (DEPs), we analysed the quantitative ratios (as the Log2 value). The fold-change ratio cut off was more than 2 or less than 0.5 based on intensity of 0 min. Searched data went through statistical process with Perseus (version 1.5.1.6).

2.4 Mathematical modelling

2.4.1 Splice variant model construction

We hypothesized that protein abundance could be predicted using a linear combination of the corresponding splice variants. To predict protein abundance, we used the Sklearn (Pedregosa et al., 2011) implementation of the LASSO (Tibshirani, 1996), an L1-penalized linear regression model.

$$\underset{\beta,\in Re}{min}\left\{\frac{1}{N}\left\|Y-\beta X\right\|_2+\boldsymbol{\lambda}\left\|\beta\right\|_1\right\}$$

Here, the time series of one protein is denoted the vector Y, and the corresponding time series of the splice variants are denoted by the matrix X. The rate constant for each splice variant is contained in the vector β. Furthermore, the λ parameter regulates the influence of the L1 term and was determined individually for each protein. The λ term was chosen to minimize the prediction error of a leave-one-out cross validation. In the T_H1 dataset, the time points differed such that the mRNA abundance also had a measurement at t = 30 min, while the protein data instead had a measurement of t = 120 h. For comparison, the protein data for 30 min was interpolated, while the 120 h time point was omitted. The same procedure was performed using the T_{reg} data from (Schmidt et al., 2018) where T_{reg} were induced by either TGF-β, TGF-β and ATRA, or TGF-β and butyrate. Lastly, the same procedure was performed for mice B cells where B cell differentiation was induced by the Ikaros transcription factor (Gomez-Cabrero et al., 2019) (GSE75417).

2.4.2 Time delay analysis

The effect of time delays between mRNA and protein was analysed since this might affect the prediction of protein abundance. First, we considered the T_H1 data and linearly interpolated between 0 and 24 h for both the mRNA expression and protein abundance data with a quadratically increasing distribution between the time delays. In total, 200 time series were interpolated, such that the difference between the first time points was 43 s, and the difference

between the last samples was 15 min. In the updated model, we added a protein specific time delay τ to regulate which time point of splice variant expression should be used. As an example, a τ = 0.5 h would result in splice variant abundance of t = [0, 1, 2, 6, 24 h] predict protein abundance interpolated at t = [0.5, 1.5, 2.5, 6.5, 24.5 h]. Full details on the models can be found in Supplementary Table S1.

$$min\left\{\frac{1}{N}\left\|Y\left(t+\tau\right)-\beta X\left(t\right)\right\|_2+\boldsymbol{\lambda}\left\|\beta\right\|_1\right\}$$

2.4.3 Cross validation

To select the values of λ and τ, a double cross-validation was performed (Supplementary Figure S3). First, one of the time points of the protein measurements was removed from the set, leaving only 5 data points. Secondly, a leave-one out cross-validation was performed on the remaining 5 time points, giving an estimate of the accuracy of the model approach given a time delay and a lambda value for the penalty term in the Lasso operator. We used the 200-time delays ranging between 0 and 24 h, and a varying set of lambda parameters (increased until all parameters equaled zero). Thirdly, the time delay and penalization that generated the smallest average squared residuals between the second cross-validation and the data were chosen and used to predict the sixth data point from splice variants. Fourth, this double cross-validation procedure was repeated for all 6 data points.

2.5 Differential expression analysis

The raw counts of each transcript were z normalized, and, in the case of predicted protein, combined using the transcript-specific coefficient from the linear model. Next, differential expression was analysed using a non-parametric Kruskal-Wallis test as implemented in the SciPy Python package. We used the Benjamini Hochberg false discovery rate (FDR) when accounting for multiple testing.

2.6 Disease prediction

Disease relevance of the splice variant models was tested by re-analysis of RNA-seq case and control material of samples containing conventional $CD4^+$ T-cells, i.e., $CD4^+$ T-cells with all its sub-types. We found T-cell prolymphocytic leukaemia (T-PLL, GSE100882), asthma in obese children (GSE86430), and allergic rhinitis/asthma (GSE75011) studies through a Gene Expression Omnibus (GEO) repository search and MS through collaboration (James et al., 2018). For each of the studies, we used the T_H1 and T_{reg} derived models on how to combine mRNA splice variants to predict protein abundance. The resulting sets of predicted protein levels were tested for differential expression between patients and controls using a non-parametric Kruskal-Wallis test. We also applied Kruskal-Wallis tests to the individual splice variants that were used by the models. We assessed model effects by measuring the increase in nominally differential expression from model predictions compared to ingoing splice variants into the model. In the study of MS, we performed a specific gene selection and performed FDR correction using the Benjamini Hochberg selection procedure (FDR < 0.05). Using protein data from two of the largest biomarker studies in MS (Huang et al., 2020; Mahler et al., 2020), we compared the protein measurements with our predicted proteins. One study reported 36 out of 92 proteins as significant (Huang et al., 2020) and another study (Mahler et al., 2020) reported the expression of four proteins whereof two were significant. We found that the expression of all our predicted differentially expressed protein agreed with the two studies (9/9 negatively reported from first study and 1/1 negatively and 1/1 positively reported from second study) and the corresponding P-value was calculated as $((92–36)/92)^9$ x $(2/4)^2 = 2.9 \times 10^{-3}$.

2.7 Protein validation

2.7.1 Patients and controls

Cerebrospinal fluid (CSF) was collected from a cohort of 41 patients with newly diagnosed clinically isolated syndrome (CIS) or relapsing remitting MS (RRMS) (Supplementary Table S2) that has been described in more detail elsewhere (Håkansson et al., 2018). All patients fulfilled the revised McDonald criteria from 2010 (Polman et al., 2011). The patients were followed, and new samples obtained after one, two and 4 years. Disease activity was assessed using “no evidence of disease activity” (NEDA), defined by no clinical relapses, no sustained EDSS progression and no new T2 or Gadolinium enhancing lesions. 12 patients at the two year- and 7 patients at the 4-year follow-up were classified as NEDA, whereas patients with relapses, brain MRI activity and sustained disease progression were classified as “evidence of disease activity” (EDA; n = 27 and n = 32 at two and 4 years, respectively). Two patients did not complete the study (Håkansson et al., 2018). Twenty-three healthy age-and sex-matched blood donors were included as controls. A second cohort of CSF samples from 16 Natalizumab-treated patients with RRMS or secondary progressive MS (SPMS) was also included. CSF samples were obtained (out of a total of ≈70 included patients with RRMS or SPMS) before and after 1 year of treatment with Natalizumab (Supplementary Table S2). This study cohort has been described previously (Mellergård et al., 2010; Mellergård et al., 2013; Gustafsson et al., 2014). All

patients were recruited at the Department of Neurology, Linköping, University Hospital Sweden and both patients and controls gave written consent prior to inclusion. The study was approved by The Regional Ethics Committee in Linköping.

2.7.2 Protein measurements

Quantification of sCD27 was performed using the Human Instant ELISA™ kit from eBioscience (Thermo Fischer Scientific) according to the instructions provided by the manufacturer. The optical densities (O.D.) were read at 450 nm with a wavelength correction at 620 nm in a Sunrise™ microplate reader (Tecan, Männedorf, Switzerland). Data acquisition was performed using Magellan™ version 7.1 computer software (Tecan). The lowest detection limit was 0.63 U/ml and values below the detection limit were given half the value of the detection limit. Statistical differences were determined using Mann-Whitney U-test or Wilcoxon matched-pairs signed rank test (Graphpad Prism v7.04, San Diego, CA, United States). Annexin A1, measured by the human Annexin A1 ELISA kit (Abcam, Cambridge, United Kingdom), was undetectable in all analysed samples (n = 32, of whom n = 16 samples were included before and n = 16 after 1 year of treatment with Natalizumab). Multiplex Bead Technology (MILLIPLEX® MAP Kit, Cat. #: HCYTOMAG-60K-01, Merck Millipore, Burlington, MA, United States) was used to measure soluble CD40L according to the manufacturer's description. The samples were analysed on a Luminex®200™ instrument (Invitrogen, Carlsbad, CA, United States) and data was collected using xPONENT 3.1™ (Luminex Corporation, Austin, TX, United States) analysed using the MasterPlex® Reader Fit (MiraiBio Group, Hitachi Solutions America Ltd., San Bruno, CA, United States). The lowest detection limit was 1.6 pg/ml and values below the detection limit were given half the value of the detection limit. sCD40L concentration was below the lowest detection limit in 71 out of 96 samples (74% undetectable) and was therefore considered as undetectable.

3 Results

3.1 A significant portion of T-cell genes showed diverse correlations between RNA splice variants and proteins

To generate accurate mRNA and protein models, considering the major factors of time delay and splice variant usage, we first developed a model by analysing early $T_H1$differentiation. This was done by performing time series transcriptomic (RNA-seq) and proteomics (mass spectrometry) analysis at six different time points, from 30 min to 5 days, during T_H1 differentiation, whereof five time points were paired between the omics and could be further used to infer correlations between mRNA and protein (Figure 1A and Supplementary Figures S3, S4). We found a total of 15,699 genes and 6,909 proteins to be expressed during early T_H1 differentiation. Out of the 6,909 expressed proteins, 5,749 could be mapped to genes and out of those, 4,920 were also found to be expressed at the transcriptomic level. As expected, a significant proportion of the 4,920 genes showed a significant positive correlation between mRNA and protein levels (n = 407, expected 123 out of 4,920, binomial test $p < 10^{-93}$) during T_H1 cell differentiation. Interestingly, a significant fraction of negatively correlated genes was also observed (n = 205, expected 123, $p < 10^{-11}$) (Figure 1B and Supplementary Table S1). Notably, the overall median Pearson correlation (rho) between mRNA and protein was only 0.21. Analysis of the distribution of the correlation coefficients revealed significant enrichments of both positive and negative correlations between splice variants and their corresponding proteins (binomial test for enrichment of significant negative correlation $p < 1.3 \times 10^{-3}$, odds ratio = 1.48) (Figure 1C and Supplementary Figure S5). For example, the known T-cell associated genes, *IL7R* and *STX12* (Kanduri et al., 2015), contained multiple splice variants, of which several were positively or negatively correlated to their corresponding protein levels (Figure 1C). Given the large variation in correlation between different splice variants of a given gene and its corresponding protein, we proceeded to construct predictive splice variant models of protein abundance.

3.2 A linear model combining the expressions of multiple splice variant transcripts showed substantially stronger correlations with protein abundance than individual transcripts

In order to construct generally applicable and predictive mRNA-to-protein models, we applied a simple linear relation between the protein abundance of a gene and its associated mRNA splice variants. Furthermore, we allowed for different translation times for each gene. Firstly, we used a cross-validated L1 penalised linear regression model to favour simple models using single splices without any time delays (Figure 1D). The rationale for the L1 penalty was to effectively remove splice variants that carry little or no predictive power over protein abundance. In practice this resulted in maximum of three splice variants per protein for the T_H1 model, which is a method limitation due to the few data points and our regularisation. This simple model resulted in a median gene-protein correlation of rho_{TH1} = 0.86 using cross-validated predictions (Figure 2A). Likewise, to test the generality of the approach we also trained similar models for two existing mRNA-protein time-series datasets with similar results, that is from human T_{reg} cells (Schmidt et al., 2018) (rho_{Treg} = 0.79) and mice B cells (Gomez-Cabrero et al., 2019) (GSE75417) (rho_{Bcell} = 0.94) (Figure 2A). Next, to test whether the increase in correlation was due to the incorporation of negatively correlating splice variants, multiple transcripts, or time delay, we also constructed

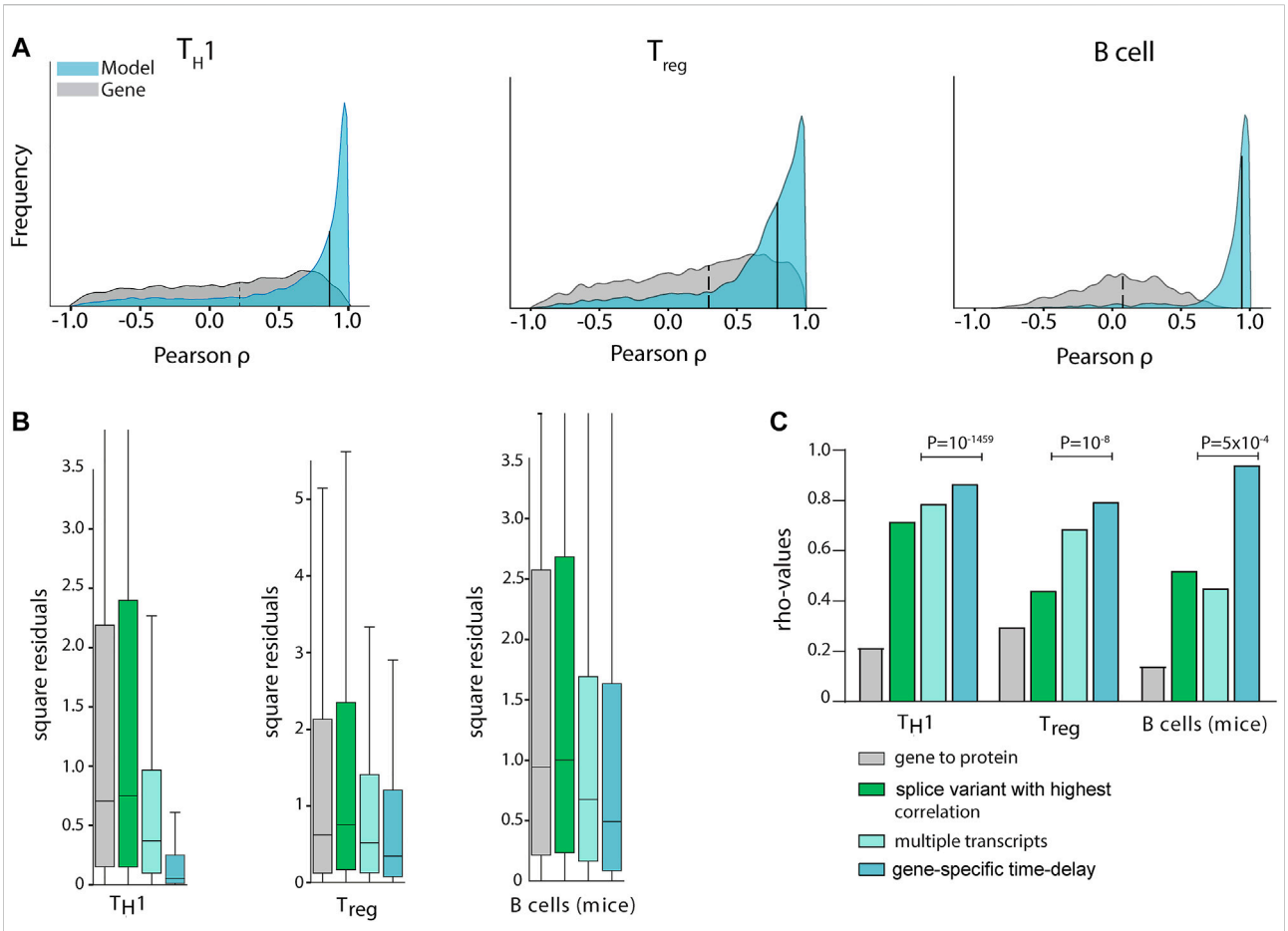


FIGURE 2
Multiple transcripts and time delays increased mRNA and protein correlations significantly in multiple cell types. **(A)** Gene/protein Pearson correlations in T_H1 (left), T_{reg} (middle), and murine B-cell (right) differentiation. In the histogram, the grey curve shows the correlation distribution when the sum of all splice variant expressions of a transcript (Fortelny et al., 2017) is used to quantify mRNA abundance (median: dashed line), while in the blue histogram our time delayed multiple splice variant based models are used (medians: solid lines at 0.86, 0.79, and 0.94 for T_H1, T_{reg} and murine B-cells, respectively). Only cross-validated protein predictions are shown for the proteins for which the null-model could be rejected. **(B)** Out-of-sample cross validation prediction of the three models. Aiming to quantify the predictive power of each added input to the model, we observed that a linear model with gene-specific time delays was the model that generated predictions with the smallest sum of squared residuals. **(C)** Median correlation coefficients (rho) for different mathematical protein prediction models derived from mRNA with increasing protein abundance correlations. P-values were derived from predictions using leave-one-out cross-validation.

such models without each of these parameters. Importantly, our model outperformed the models using only the most highly correlated splice variant for each gene (rho_{TH1} = 0.71, rho_{Treg} = 0.44, rho_{Bcell} = 0.52), and the models using multiple transcripts but without a time delay (rho_{TH1} = 0.74, rho_{Treg} = 0.69, rho_{Bcell} = 0.45) (Figures 2B,C), thus demonstrating that both multiple dynamical splice variants and time delay increase the fit of data and are needed for optimal performance.

To define the optimal time delays between splice variants and proteins, we analysed the time delay distributions and found it to have a mean of 8 h 17 min, 6 h 18 min and 8 h 49 min for T_H1, T_{reg} and mice B cells, respectively. The detailed parameters of our models are fully displayed in Supplementary Table S1. Next, by using double cross-validation we confirmed that our models could do out-of-sample prediction significantly better than conventional gene expression-based models of protein abundance (binomial test; $p_{TH1} = 10^{-297}$ (expected 14.4 of 28.9, observed 18.0), $p_{Treg} = 10^{-247}$ (expected 21.2 of 43.5, observed 25.2), $p_{mice\ B} = 10^{-59}$ (expected 2.3 of 5.5, observed 3.3)), and better than static splice variant models which did not include time delays ($p_{TH1} = 10^{-1459}$ (expected 14.8 of 29.6, observed 21.8), $p_{Treg} = 10^{-8}$ (expected 22199 of 44397, observed 22811), $p_{mice\ B} = 5 \times 10^{-4}$ (expected 2.6 of 5.5, observed 2.9), Figure 2C). Moreover, we used time-point scrambling and dynamical correlation analysis to show that our analysis was not seriously affected by time-dependences within the time-series (data not shown). In summary, we have identified simple linear models of mRNA splice variants and time

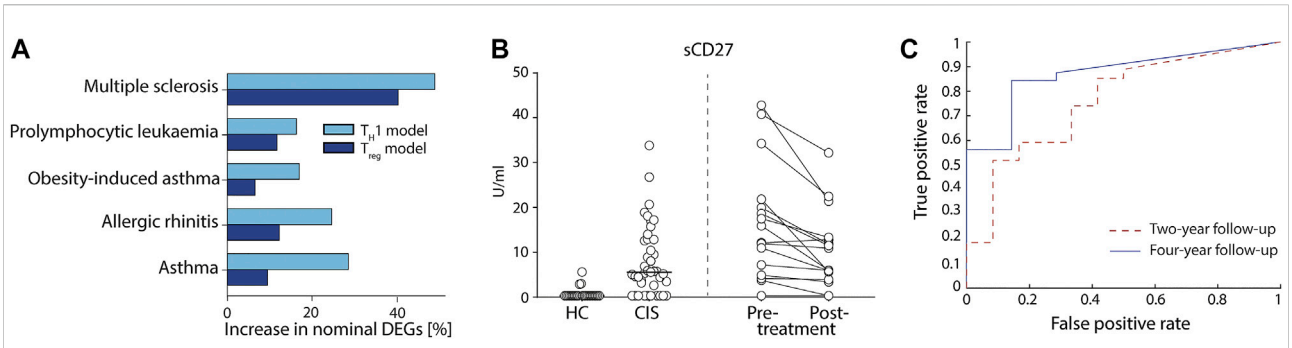


FIGURE 3
Proteins models led to the discovery of new potential biomarkers of complex diseases that were validated in multiple sclerosis (MS). **(A)** Differential predicted protein (PP) analysis of five diseases using the T_H1 (light blue) and T_{reg} (dark blue) models showed higher fraction of nominally significant genes than that of normal differential gene expression tests. **(B)** Measurement of actual protein levels of the predicted proteins in a cohort of patients with early MS [clinically isolated syndrome (CIS)] vs. healthy controls (HC) (left side of the figure) and from a cohort of MS patients pre vs. post 1-year treatment with Natalizumab (right side of the figure). sCD27 was measured in cerebrospinal fluid (CSF) using ELISA. **(C)** Receiver operating curve using sCD27 concentration as a single prognostic marker of NEDA at four (solid line) and 2 years (dashed line) after CIS.

delay which could be used to model the time courses in T- and B-cell differentiation (see the full models in Supplementary Table S1). We would like to emphasize that this is a minimal requirement for mRNA-protein models to be meaningful, so we proceeded to analyse if the models were useful to translational research by identifying biomarkers in complex diseases.

3.3 The models showed increased biomarker sensitivity which were further verified in multiple sclerosis and asthma

Lastly, we aimed to test the potential usefulness of our derived models for the identification of protein biomarkers by applying them on available RNA-seq datasets from human total $CD4^+$ T cells. We found datasets for five different diseases (Seumois et al., 2016; James et al., 2018; Johansson et al., 2018; Rastogi et al., 2018); asthma, allergic rhinitis, obesity-induced asthma, pro-lymphocytic leukaemia, and MS, as well as corresponding controls. Because our models correlated well to protein abundances, we hypothesised that differential expression tests using the predicted proteins between patients and controls would be more sensitive than testing directly on the mRNA expression for all splice variants individually. Indeed, we observed that the fraction of nominally differentially expressed genes was higher than using an individual differential expression analysis in all comparisons (binomial $p < 9.8 \times 10^{-4}$). Moreover, we consistently observed a higher enrichment for the T_H1 model compared to the T_{reg} model ($p < 0.03$) (Figure 3A), with the highest enrichments in MS and asthma. We therefore proceeded to use our T_H1 model on MS and asthma.

First, we compared our MS findings with previously reported proteins using two large biomarker studies (Huang et al., 2020; Mahler et al., 2020) of MS and found a significant agreement comparing our nominal predictions (binomial $p < 2.9 \times 10^{-3}$; see Methods). Then, we found 20 genes with FDR<0.05, of which none were detected at 20% FDR level by testing for differential expression on the mRNA expression data directly (Supplementary Table S3). Interestingly, eight of the 20 genes had previously been associated with MS (Figure 4 and Supplementary Table S3). To further justify the relevance of the added genes we analysed if CSF levels of these proteins were related to clinical outcome and immunomodulatory treatment in two independent cohorts, newly diagnosed MS patients (clinically isolated syndrome (CIS) and relapsing/remitting MS, $n = 41$) *vs.* healthy controls (HC, $n = 23$), and response to Natalizumab treatment in relapsing remitting MS patients ($n = 16$). In both cohorts, only sCD27 was present in CSF at a detectable level (Supplementary Table S4), while Annexin A1 and sCD40L were not. Analysis of all patients ($n = 57$) *vs.* HC ($n = 23$) showed high separation (AUC = 0.88, non-parametric $p = 3.0 \times 10^{-8}$, Figure 3B), and treatment with Natalizumab reduced the sCD27 levels by 34% ($p = 4.9 \times 10^{-4}$). Notably, sCD27 levels at baseline of newly diagnosed MS and CIS patients were able to predict disease activity after 4 years follow up (AUC = 0.87, $p = 1.2 \times 10^{-3}$, Figure 3C), which was a stronger prediction than that of all our previously reported 14 biomarkers (Håkansson et al., 2018). Taken together, using the splice variants-to-protein model we were able to uniquely identify and validate biomarkers of MS in an independent patient cohort, while these genes could not be discovered using previous state-of-the-art test for differential gene expression.

For asthma we found six of the top 20 genes that were differentially expressed (determined by conventional mRNA expression) to be previously associated with the disease (Supplementary Table S5). Next, we analysed asthma-associated genes uniquely identified by our model and found seven additional genes to be associated with asthma

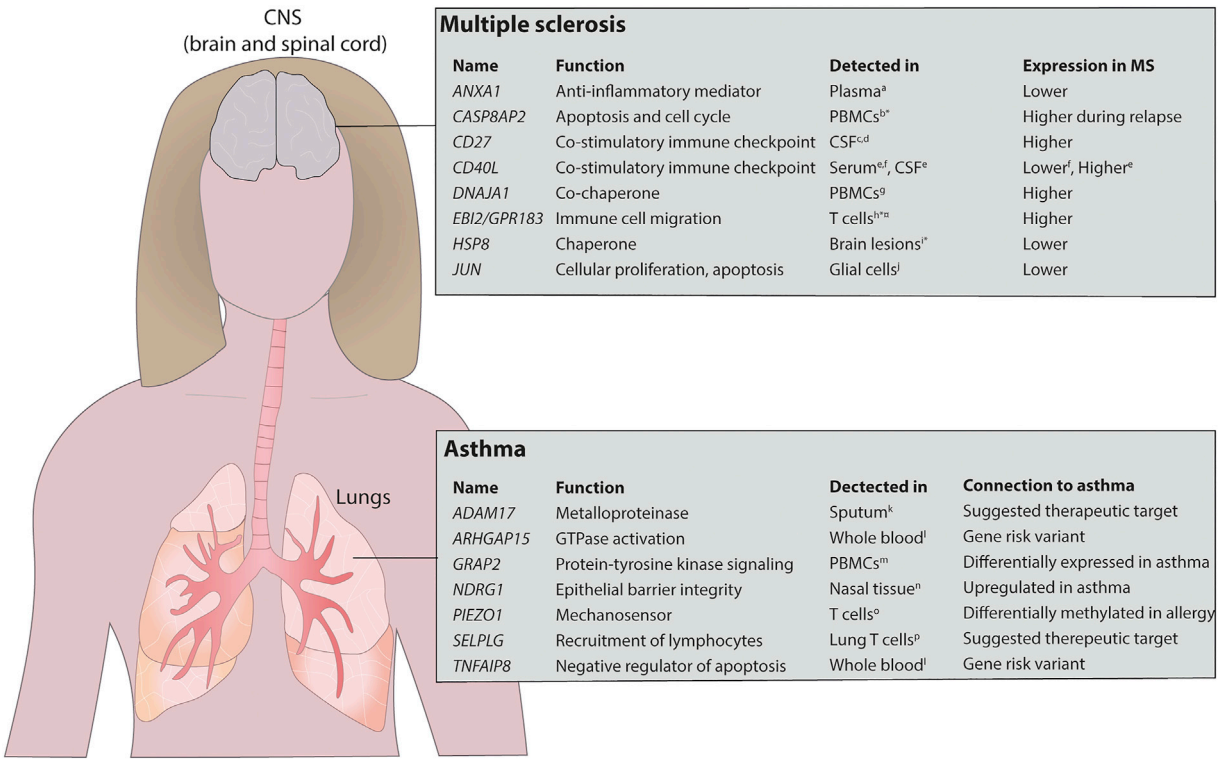


FIGURE 4
Overview of detected potential biomarkers in asthma and MS. The model identified several proteins that have previously been identified in MS and asthma. The upper panel shows the potential biomarkers identified in MS and the lower panel shows the same in asthma. *mRNA expression, ¤ identified in mice. PBMCs, peripheral blood mononuclear cells. References stated in the figure [a]Colamatteo A et al., J Immunol, 2019; [b]Achiron A et al., Ann N Y Acad Sci, 2007; [c]van der Vuurst de Vries RM et al., JAMA Neurol, 2017; [d]Wong YYM et al., Mult Scler, 2018; [e]Masuda H et al., J Neuroimmunol, 2017; [f]de JG-GJ et al., Immunobiology, 2018; [g]Bomprezzi R et al., Hum Mol Genet, 2003; [h]Wanke F et al., Cell Rep, 2017; [i]Aquino DA et al., J Neuropathol Exp Neurol, 1997; [j]Bonetti B et al., Am J Pathol, 1999; [k]Enomoto Y et al., J Allergy Clin Immunol, 2009; [l]Ferreira MA et al., Nat Genet. 2017; [m]Persson H et al., J Allergy Clin Immunol, 2015; [n]Murray JT et al., Biochem J, 2004; [o]Nestor CE et al., PLoS Genet, 2014; [p]Purwar R et al., PLoS One, 2011.

(Supplementary Table S6). Interestingly, these genes had previously also been reported to be relevant for the disease (Enomoto et al., 2009; Nestor et al., 2014; Poole et al., 2014; Dreymueller et al., 2015; Persson et al., 2015; Ferreira et al., 2017), and are currently being evaluated as potential therapeutic targets (Figure 4). Examples of those genes include *NDRG1*, which regulates T_H2 differentiation, a key driver in asthmatic disease, downstream of the mTORC2 complex (Murray et al., 2004; Heikamp et al., 2014), *ADAM17*, a metalloproteinase involved in lung inflammation (Dreymueller et al., 2015), *PIEZ O 1*, a mechanosensor regulating T cell activation (Liu et al., 2018) and pulmonary inflammatory responses (Solis et al., 2019), and the P-selectin ligand encoding gene *SELPLG*, important for recruitment of lymphocytes to the airways (Leath et al., 2005; Purwar et al., 2011). Furthermore, the immunomodulatory genes *TNFAIP8* and *ARHGAP15* were identified in GWAS studies as shared risk variants for several IgE-mediated diseases including asthma, allergic rhinitis and atopic eczema (Ferreira et al., 2017). Thus, we have validated that our model can identify relevant biomarker candidates and therapeutical targets also in the context of another immune-mediated disease, i.e., asthma.

4 Discussion

In the present study we have shown that simple mRNA-protein models, in which the protein expression is defined as a linear combination of the splice variants of a gene with a time delay accounting for the dynamical effect induced by post-transcriptional processes and protein synthesis, can improve our ability to predict protein abundance from mRNA expression. Furthermore, we demonstrated the impact that this finding can have within genome medicine by predicting and validating biomarkers for MS and asthma. Throughout the paper we aimed to increase the sensitivity in RNA-seq differential expression analysis. Sensitivity was measured using the fraction of nominally ($p < 0.05$) differentially expressed genes. This

application revealed significantly more predicted biomarkers than by using off-the-shelf methods for RNA-seq data analysis only, which suggests increased sensitivity.

Despite being part of the central dogma and of uttermost importance in biology and medicine, the prediction of protein levels from mRNA levels has long been associated with low precision, which has been a matter of debate (Fortelny et al., 2017). Due to the complex process of mRNA-to-protein translation, there are several aspects that need to be considered (Liu et al., 2016). In this paper we thoroughly addressed two presumed main aspects; 1) how to incorporate splice variants into the prediction protein expression, and 2) how to deal with the time delay of the translation between mRNA and protein expression. Interestingly, both aspects were found to impact prediction of protein abundance, as shown in our combined model, although the incorporation of splice variants influenced the protein abundance prediction the most. Herein, we report splice variants to have a wider correlation profile, both positive and negative, than what would be expected, and our novel approach takes advantage of this anti-correlation between splice variants and proteins. In previous work, the impact of incorporating splice variants into protein predictions has been analysed. These studies have focused on mechanistic cell type independent factors such as splice variant-specific degradation rates (Eraslan et al., 2019). Instead, we found that the correlations were cell type-specific, and we constructed data-driven predictive models. To construct those models, we performed activation of NT_H cells followed by time-series analysis, which enabled us to infer the system based on its dynamics. A necessary requirement for such as model was dynamical data covering a decent number of time-points that allowed for the possibility of including modelling of intermediate time-points and the inference of time delays. However, the resulting Pearson correlations from our model need to be taken cautiously as we could not do a complete test as parts of the longitudinal data was visible to the model. From our models we proposed a biomarker discovery strategy which was validated in three steps. First, we found that usage of these models in complex disease enabled identification of more differentially expressed genes, which we therefore predicted as potential biomarkers. Second, we noted that many of the predicted proteins had previously been associated with MS and asthma, confirming that our strategy predicts relevant disease genes. Third, we validated one such protein as a biomarker in MS, namely sCD27. While sCD27 has already been associated with MS (van der Vuurst de Vries et al., 2017; Wong et al., 2018; Mahler et al., 2020), our clinical analysis of two independent cohorts yielded novel findings of remarkably good prognostic capabilities for treatment response and 4 years disease activity, which is important areas for early MS treatment selection.

Although incorporating splice variant information into the model was the main influential factor on the correlation, time delay also had an impact. The kinetics in translation of mRNA to protein is of general interest given its crucial importance in the design of experiments, for example in verifying relevance of mRNA expression to protein expression. Such models should ideally be functionally validated based on mechanistic principles, described by ordinary differential equations, such as the ones presented by for example Jovanovic et al. (2015). However, given that time-series experiments are time- and labor intensive, as well as expensive and predictive large-scale models are highly needed for biomarker discoveries, a database that provides the relevant time delay between mRNA expression and the expression of its corresponding protein would be immensely valuable. Here, we present such an atlas, comprising almost 5000 gene expression-to-protein translation kinetics (Supplementary Table S1).

A limitation with the paper is that we investigated few key cell types, namely T_H1 cells, T_{REG} cells and B cells whereof wet lab experiments was only performed in one of these cell types. However, we were able to transfer the approach to two other cell type re-using data of other studies, demonstrating the robustness of the model assumptions. Furthermore, the chosen cell types are central in regulation of immune responses, and the T_H cells indeed are involved in many complex and common illnesses, like infectious, allergic, autoimmune and cardiovascular diseases and cancer (Farber, 2020).

In conclusion, we have constructed data-driven linear models incorporating splice variant information and time delay to predict protein expression from mRNA. We showed the general applicability of our approach by developing robust models for datasets from several cell types, and therefore the general principle of the model should be applicable to other cell types. For example, we expect this modelling strategy to be generally applicable to other cellular differentiation systems, such as embryonic stem cell differentiation, and to be increasingly useful for understanding basic biology and identification of new biomarkers as more RNA-seq and proteomic data sets become publicly available. Finally, we have shown that our proposed approach is of clinical relevance for prediction of validated biomarkers.

Data availability statement

The raw and processed RNA-seq data were submitted to the EMBL-EBI sequencing archive ArrayExpress and is available under the accession number E-MTAB-7775. The proteomics data were submitted to the EMBL-EBI proteomics repository PRIDE under the accession PXD013361. Pipeline and code for the mathematical modelling and bioinformatics analysis available from https://gitlab.com/Gustafsson-lab/splice_protein_predictions.

Ethics statement

The study was approved by the Regional Ethics Committee in Linköping, Sweden (Dnr M180-07 and M2-09). All patients were

recruited at the Department of Neurology, Linköping, University Hospital Sweden and both patients and controls gave written consent prior to inclusion.

Author contributions

MG initiated and supervised the study. RM and OR performed bioinformatics analyses. RM performed the modelling. These analyses were led by MG, CA, JT, and DG-C. OR performed experimental work on T-cell differentiation, which were supervised by CEN, MCJ, JE, and MB. MJK and CHN performed the proteomics analysis, which was supervised by M-SK. FP and JM recruited patients and collected clinical material, and SH performed and analysed the biomarker validation assays, which were led by IK, MCJ, and JE. All authors contributed to and approved the final draft for publication.

Funding

This work was supported by the Swedish foundation for strategic research (SB16-0011), Swedish Cancer Society grants (CAN 2017/625), East Gothia Regional Funding, Åke Wiberg foundation, Neuro Sweden, the Swedish Research Council grants 2015-02575, 2015-03495, 2015-03807, 2016-07108, and 2018-02776, and National Research Foundation of Korea (NRF-2016K1A3A1A47921601, 2017M3C7A1027472).

Acknowledgments

We would like to thank Jun Hyung Lee for his contribution to the proteomics sample preparation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.916128/full#supplementary-material

References

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338 (6114), 1587–1593. doi:10.1126/science.1230612

BC., T. (2015a). *Homo.sapiens: Annotation package for the Homo.sapiens object*. R package version 1.3.1. Available at: https://bioconductor.org/packages/release/data/annotation/html/Homo.sapiens.html.

BC., T. (2015b). *Mus.musculus: Annotation package for the Mus.musculus object*. R package version 1.3.1. Available at: https://bioconductor.org/packages/release/data/annotation/html/Mus.musculus.html.

Cha, B. S., Park, K. S., and Park, J. S. (2020). Signature mRNA markers in extracellular vesicles for the accurate diagnosis of colorectal cancer. *J. Biol. Eng.* 14, 4. doi:10.1186/s13036-020-0225-9

Chase Huizar, C., Raphael, I., and Forsthuber, T. G. (2020). Genomic, proteomic, and systems biology approaches in biomarker discovery for multiple sclerosis. *Cell. Immunol.* 358, 104219. doi:10.1016/j.cellimm.2020.104219

de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* 5 (12), 1512–1526. doi:10.1039/b908315d

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29 (1), 15–21. doi:10.1093/bioinformatics/bts635

Dreymueller, D., Uhlig, S., and Ludwig, A. (2015). ADAM-Family metalloproteinases in lung inflammation: Potential therapeutic targets. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 308 (4), L325–L343. doi:10.1152/ajplung.00294.2014

Ek, W. E., Karlsson, T., Hoglund, J., Rask-Andersen, M., and Johansson, A. (2021). Causal effects of inflammatory protein biomarkers on inflammatory diseases. *Sci. Adv.* 7 (50), eabl4359. doi:10.1126/sciadv.abl4359

Enomoto, Y., Orihara, K., Takamasu, T., Matsuda, A., Gon, Y., Saito, H., et al. (2009). Tissue remodeling induced by hypersecreted epidermal growth factor and amphiregulin in the airway after an acute asthma attack. *J. Allergy Clin. Immunol.* 124 (5), 913–917. doi:10.1016/j.jaci.2009.08.044

Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Hallstrom, B. M., Uhlen, M., et al. (2019). Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol. Syst. Biol.* 15 (2), e8513. doi:10.15252/msb.20188513

Farber, D. L. (2020). Form and function for T cells in health and disease. *Nat. Rev. Immunol.* 20 (2), 83–84. doi:10.1038/s41577-019-0267-8

Ferreira, M. A., Vonk, J. M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J. D., et al. (2017). Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* 49 (12), 1752–1757. doi:10.1038/ng.3985

Floor, S. N., and Doudna, J. A. (2016). Tunable protein synthesis by transcript isoforms in human cells. *Elife* 5, e10921. doi:10.7554/eLife.10921

Fortelny, N., Overall, C. M., Pavlidis, P., and Freue, G. V. C. (2017). Can we predict protein from mRNA levels? *Nature* 547 (7664), E19–E20. doi:10.1038/nature22293

Gawel, D. R., Serra-Musach, J., Lilja, S., Aagesen, J., Arenas, A., Asking, B., et al. (2019). A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. *Genome Med.* 11 (1), 47. doi:10.1186/s13073-019-0657-3

Gomez-Cabrero, D., Tarazona, S., Ferreiros-Vidal, I., Ramirez, R. N., Company, C., Schmidt, A., et al. (2019). STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci. Data* 6 (1), 256. doi:10.1038/s41597-019-0202-7

Gustafsson, M., Edström, M., Gawel, D., Nestor, C. E., Wang, H., Zhang, H., et al. (2014). Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment. *Genome Med.* 6 (2), 17. doi:10.1186/gm534

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19 (3), 1720–1730. doi:10.1128/MCB.19.3.1720

Håkansson, I., Tisell, A., Cassel, P., Blennow, K., Zetterberg, H., Lundberg, P., et al. (2018). Neurofilament levels, disease activity and brain volume during follow-up in multiple sclerosis. *J. Neuroinflammation* 15 (1), 209. doi:10.1186/s12974-018-1249-7

Heikamp, E. B., Patel, C. H., Collins, S., Waickman, A., Oh, M. H., Sun, I. H., et al. (2014). The AGC kinase SGK1 regulates TH1 and TH2 differentiation downstream of the mTORC2 complex. *Nat. Immunol.* 15 (5), 457–464. doi:10.1038/ni.2867

Huang, J., Khademi, M., Fugger, L., Lindhe, O., Novakova, L., Axelsson, M., et al. (2020). Inflammation-related plasma and CSF biomarkers for multiple sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* 117 (23), 12952–12960. doi:10.1073/pnas.1912839117

James, T., Linden, M., Morikawa, H., Fernandes, S. J., Ruhrmann, S., Huss, M., et al. (2018). Impact of genetic risk loci for multiple sclerosis on expression of proximal genes in patients. *Hum. Mol. Genet.* 27 (5), 912–928. doi:10.1093/hmg/ddy001

Johansson, P., Klein-Hitpass, L., Choidas, A., Habenberger, P., Mahboubi, B., Kim, B., et al. (2018). SAMHD1 is recurrently mutated in T-cell prolymphocytic leukemia. *Blood Cancer J.* 8 (1), 11. doi:10.1038/s41408-017-0036-5

Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., et al. (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 347 (6226), 1259038. doi:10.1126/science.1259038

Kanduri, K., Tripathi, S., Larjo, A., Mannerstrom, H., Ullah, U., Lund, R., et al. (2015). Identification of global regulators of T-helper cell lineage specification. *Genome Med.* 7, 122. doi:10.1186/s13073-015-0237-0

Kuchta, K., Towpik, J., Biernacka, A., Kutner, J., Kudlicki, A., Ginalski, K., et al. (2018). Predicting proteome dynamics using gene expression data. *Sci. Rep.* 8 (1), 13866. doi:10.1038/s41598-018-31752-4

Leath, T. M., Singla, M., and Peters, S. P. (2005). Novel and emerging therapies for asthma. *Drug Discov. Today* 10 (23-24), 1647–1655. doi:10.1016/S1359-6446(05)03646-9

Liu, C. S. C., Raychaudhuri, D., Paul, B., Chakrabarty, Y., Ghosh, A. R., Rahaman, O., et al. (2018). Cutting edge: Piezo1 mechanosensors optimize human T cell activation. *J. Immunol.* 200 (4), 1255–1260. doi:10.4049/jimmunol.1701118

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell* 165 (3), 535–550. doi:10.1016/j.cell.2016.03.014

Mahler, M. R., Sondergaard, H. B., Buhelt, S., von Essen, M. R., Romme Christensen, J., Enevold, C., et al. (2020). Multiplex assessment of cerebrospinal fluid biomarkers in multiple sclerosis. *Mult. Scler. Relat. Disord.* 45, 102391. doi:10.1016/j.msard.2020.102391

Maier, T., Guell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583 (24), 3966–3973. doi:10.1016/j.febslet.2009.10.036

Mao, Z., Ji, A., Yang, K., He, W., Hu, Y., Zhang, Q., et al. (2018). Diagnostic performance of PCA3 and hK2 in combination with serum PSA for prostate cancer. *Med. Baltim.* 97 (42), e12806. doi:10.1097/MD.0000000000012806

Mayeux, R. (2004). Biomarkers: Potential uses and limitations. *NeuroRx.* 1 (2), 182–188. doi:10.1602/neurorx.1.2.182

Mellergård, J., Edström, M., Jenmalm, M. C., Dahle, C., Vrethem, M., and Ernerudh, J. (2013). Increased B cell and cytotoxic NK cell proportions and increased T cell responsiveness in blood of natalizumab-treated multiple sclerosis patients. *PLoS One* 8 (12), e81685. doi:10.1371/journal.pone.0081685

Mellergård, J., Edström, M., Vrethem, M., Ernerudh, J., and Dahle, C. (2010). Natalizumab treatment in multiple sclerosis: Marked decline of chemokines and cytokines in cerebrospinal fluid. *Mult. Scler.* 16 (2), 208–217. doi:10.1177/1352458509355068

Murray, J. T., Campbell, D. G., Morrice, N., Auld, G. C., Shpiro, N., Marquez, R., et al. (2004). Exploitation of KESTREL to identify NDRG family members as physiological substrates for SGK1 and GSK3. *Biochem. J.* 384 (3), 477–488. doi:10.1042/BJ20041057

Nestor, C. E., Barrenäs, F., Wang, H., Lentini, A., Zhang, H., Bruhn, S., et al. (2014). DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4+ T-cell population structure. *PLoS Genet.* 10 (1), e1004059. doi:10.1371/journal.pgen.1004059

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi:10.5555/1953048.2078195

Persson, H., Kwon, A. T., Ramilowski, J. A., Silberberg, G., Soderhall, C., Orsmark-Pietras, C., et al. (2015). Transcriptome analysis of controlled and therapy-resistant childhood asthma reveals distinct gene expression profiles. *J. Allergy Clin. Immunol.* 136 (3), 638–648. doi:10.1016/j.jaci.2015.02.026

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33 (3), 290–295. doi:10.1038/nbt.3122

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., et al. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann. Neurol.* 69 (2), 292–302. doi:10.1002/ana.22366

Poole, A., Urbanek, C., Eng, C., Schageman, J., Jacobson, S., O'Connor, B. P., et al. (2014). Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. *J. Allergy Clin. Immunol.* 133 (3), 670–678. e12 e612. doi:10.1016/j.jaci.2013.11.025

Purwar, R., Campbell, J., Murphy, G., Richards, W. G., Clark, R. A., and Kupper, T. S. (2011). Resident memory T cells (T(RM)) are abundant in human lung: Diversity, function, and antigen specificity. *PLoS One* 6 (1), e16245. doi:10.1371/journal.pone.0016245

Raphael, I., Nalawade, S., Eagar, T. N., and Forsthuber, T. G. (2015). T cell subsets and their signature cytokines in autoimmune and inflammatory diseases. *Cytokine* 74 (1), 5–17. doi:10.1016/j.cyto.2014.09.011

Rastogi, D., Nico, J., Johnston, A. D., Tobias, T. A. M., Jorge, Y., Macian, F., et al. (2018). CDC42-related genes are upregulated in helper T cells from obese asthmatic children. *J. Allergy Clin. Immunol.* 141 (2), 539–548. e7 e537. doi:10.1016/j.jaci.2017.04.016

Rifai, N., Gillette, M. A., and Carr, S. A. (2006). Protein biomarker discovery and validation: The long and uncertain path to clinical utility. *Nat. Biotechnol.* 24 (8), 971–983. doi:10.1038/nbt1235

Schmidt, A., Marabita, F., Kiani, N. A., Gross, C. C., Johansson, H. J., Elias, S., et al. (2018). Time-resolved transcriptome and proteome landscape of human regulatory T cell (Treg) differentiation reveals novel regulators of FOXP3. *BMC Biol.* 16 (1), 47. doi:10.1186/s12915-018-0518-3

Seumois, G., Zapardiel-Gonzalo, J., White, B., Singh, D., Schulten, V., Dillon, M., et al. (2016). Transcriptional profiling of Th2 cells identifies pathogenic features associated with asthma. *J. Immunol.* 197 (2), 655–664. doi:10.4049/jimmunol.1600397

Solis, A. G., Bielecki, P., Steach, H. R., Sharma, L., Harman, C. C. D., Yun, S., et al. (2019). Mechanosensation of cyclical force by PIEZO1 is essential for innate immunity. *Nature* 573 (7772), 69–74. doi:10.1038/s41586-019-1485-8

Sprent, J., and Tough, D. F. (1994). Lymphocyte life-span and memory. *Science* 265 (5177), 1395–1400. doi:10.1126/science.8073282

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

van der Vuurst de Vries, R. M., Mescheriakova, J. Y., Runia, T. F., Jafari, N., Siepman, T. A., and Hintzen, R. Q. (2017). Soluble CD27 levels in cerebrospinal fluid as a prognostic biomarker in clinically isolated syndrome. *JAMA Neurol.* 74 (3), 286–292. doi:10.1001/jamaneurol.2016.4997

Vogel, C., and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13 (4), 227–232. doi:10.1038/nrg3185

Wethmar, K., Smink, J. J., and Leutz, A. (2010). Upstream open reading frames: Molecular switches in (patho)physiology. *Bioessays.* 32 (10), 885–893. doi:10.1002/bies.201000037

Wong, Y. Y. M., van der Vuurst de Vries, R. M., van Pelt, E. D., Ketelslegers, I. A., Melief, M. J., Wierenga, A. F., et al. (2018). T-cell activation marker sCD27 is associated with clinically definite multiple sclerosis in childhood-acquired demyelinating syndromes. *Mult. Scler.* 24 (13), 1715–1724. doi:10.1177/1352458518786655

frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 08 September 2022
DOI 10.3389/fmolb.2022.962799

OPEN ACCESS

EDITED BY
Ornella Cominetti,
Nestlé Research Center, Switzerland

REVIEWED BY
Bharat Mishra,
University of Alabama at Birmingham, United States
Han Wang,
Northeast Normal University, China

*CORRESPONDENCE
Arnaud Droit,
arnaud.droit@crchuq.ulaval.ca

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 06 June 2022
ACCEPTED 16 August 2022
PUBLISHED 08 September 2022

CITATION
Robin V, Bodein A, Scott-Boyer M-P, Leclercq M, Périn O and Droit A (2022), Overview of methods for characterization and visualization of a protein–protein interaction network in a multi-omics integration context. *Front. Mol. Biosci.* 9:962799. doi: 10.3389/fmolb.2022.962799

Overview of methods for characterization and visualization of a protein–protein interaction network in a multi-omics integration context

Vivian Robin[1], Antoine Bodein[1], Marie-Pier Scott-Boyer[1], Mickaël Leclercq[1], Olivier Périn[2] and Arnaud Droit[1]*

[1]Molecular Medicine Department, CHU de Québec Research Center, Université Laval, Québec, QC, Canada, [2]Digital Sciences Department, L'Oréal Advanced Research, Aulnay-sous-bois, France

At the heart of the cellular machinery through the regulation of cellular functions, protein–protein interactions (PPIs) have a significant role. PPIs can be analyzed with network approaches. Construction of a PPI network requires prediction of the interactions. All PPIs form a network. Different biases such as lack of data, recurrence of information, and false interactions make the network unstable. Integrated strategies allow solving these different challenges. These approaches have shown encouraging results for the understanding of molecular mechanisms, drug action mechanisms, and identification of target genes. In order to give more importance to an interaction, it is evaluated by different confidence scores. These scores allow the filtration of the network and thus facilitate the representation of the network, essential steps to the identification and understanding of molecular mechanisms. In this review, we will discuss the main computational methods for predicting PPI, including ones confirming an interaction as well as the integration of PPIs into a network, and we will discuss visualization of these complex data.

KEYWORDS

interactome, biological network, computational prediction, integrated strategies, graphic view, protein-protein interaction

Introduction

Proteins are essential to life, controlling molecular and cellular mechanisms. Their main role is to carry out cellular biological functions through interactions with molecules or macromolecules (Pellegrini et al., 1999; Vinayagam et al., 2014; Fionda, 2019). These interactions are organized in networks (Bersanelli et al., 2016) of various molecular elements (e.g., protein–DNA and protein–drug) involved in physical and biochemical processes in structured environments. Biological networks have been highlighted by the work of Barabási and Oltvai (2004), who showed that cellular networks are governed by universal laws. This new concept revolutionized the vision of system biology, initiating

creation and analysis of the first protein–protein interaction (PPI) network of yeast *Saccharomyces cerevisiae* (Dezso, Oltvai and Barabási, 2003).

In the PPI network, proteins are represented by nodes, and interactions between proteins by edges (Gursoy, Keskin and Nussinov, 2008; Zou et al., 2018). The size of the network and the amount of information (e.g., discovered node) varies between species (Kotlyar, Rossos and Jurisica, 2017; Wang and Jin, 2017). The number of PPIs is constantly changing due to complexity of the genome and many interactions remain undiscovered (Safari-Alighiarloo et al., 2014; Thanasomboon et al., 2020). PPIs can be determined by high-throughput experiments such as co-immunoprecipitation, two-hybrid screening, pull-down assays (MacDonald, 1998; Lin and Lai, 2017; Louche, Salcedo and Bigot, 2017), or by computational methods. Experimental methods are time-consuming, relatively expensive, and difficult to reproduce (von Mering et al., 2002; Piehler, 2005; Browne et al., 2010; Ngounou Wetie et al., 2013). In response to these challenges, computational methods have emerged, showing promising results in terms of performance to integrate functional (i.e., same biochemical reaction) and physical interactions. A physical interaction describes a physical contact between proteins, as a result of biochemical events steered by interactions including electrostatic forces, hydrogen bonding, and the hydrophobic effect (Berne, Weeks and Zhou, 2009; Nitzan, Casadiego and Timme, 2017). These computational methods allow a more specific identification of interactions than experimental prediction methods (Droit, Poirier and Hunter, 2005; Shoemaker and Panchenko, 2007; Zhou, Li and Wang, 2016).

Although PPIs from computational methods provide a better prediction of physical interactions, PPI databases contain a few false positive interactions (Peng et al., 2017; Luck et al., 2020). One way to remove these false interactions is through integration methods (as can be seen in session integration of a PPI network). Following the integration of the data, it becomes possible to filter PPI. To observe the resulting network and the proteins having a role in mechanisms, visualization is a key step.

Visual representation allows to understanding PPIs and to analyze networks (Iranzo, Krupovic and Koonin, 2016; Armanious et al., 2020; Schneider et al., 2021; Sejdiu and Tieleman, 2021). However, due to complexity of proteomes of different organisms, visualization is a challenge (Crowther, Wipat and Goñi-Moreno, 2021). Moreover, the density of the graph representing the proportion of interactions in the network compared to the total number of possible interactions makes representation more difficult (Ren et al., 2013; Franzese et al., 2019; Wu et al., 2019). To facilitate representation, the network is divided into sub-networks (He and Chan, 2018; Farahani, Karwowski and Lighthall, 2019). These sub-networks are obtained by filtration or by decomposing the network according to proteins of interest, with the concept of ego network (Liu et al., 2019; Tian, Ju and Yang, 2019). Ego networks are subgraphs centered on a seed node and comprise all nodes connected at a defined distance from the ego (seed node) (Zhou, Miao and Yuan, 2018; Malek, Zorzan and Ghoniem, 2020). Sub-networks facilitate representation and allow identification and understanding of cellular mechanisms, core proteins, or biomarkers (Gehlenborg et al., 2010; Laniau, 2017; Hao et al., 2019).

In this review, we will discuss computational methodologies for construction of PPI networks as well as integration and validation of these networks. Next, we will discuss the visualization aspect of a network by discussing its roles and advantages and disadvantages of different visualization tools.

Computational methods for PPI construction

Computational methods for predicting PPIs can be classified into three prediction methods: based on the genomic context, machine learning algorithm, and text mining (Table 1).

The methods can be combined to refine the prediction of PPIs. Alachram et al. (2021) exploited text mining algorithms mixed with machine learning algorithms to capture biologically significant relationships between entities, including PPIs.

Methods based on genomic context

The genomic context refers to the structure of genomic data (e.g., genes), as well as the statistical or mathematical methods to test for gene, protein set association (Dimitrieva and Bucher, 2012; Mooney et al., 2014). Genomic context methods are usually based on gene sequences, structure, and organization of genes on the chromosome (Skrabanek et al., 2008; De Las Rivas and Fontanillo, 2010; Reimand et al., 2012; Rao et al., 2014).

Domain fusion interaction prediction method

Gene fusion leads to fusion proteins, which are an assembly of several proteins encoded by different genes created by joining (fusion) of one or more genes (Morilla et al., 2010; Latysheva et al., 2016). This fusion results in a single or multiple polypeptides that takes on the functional properties of each in original proteins. The existence of a functional interaction between protein A and protein B is based on the hypothesis that if protein domains A and B of one species have fused homologs in a single AB polypeptide in another species, then domains A and B are functionally linked (Truong and Ikura, 2003; Chia and Kolatkar, 2004). The gene fusion method marked a major turning point in methods for predicting PPIs. This computational method, developed by Eisenberg et al. (2000),

TABLE 1 Summary table of computational methods for the prediction of a protein–protein interaction. Computational methods for predicting PPIs are grouped into three distinct categories: genomic context–based methods, machine learning, and text mining. Within each of these approaches, several sub-methods exist. A database can be composed of interactions obtained by several prediction methods.

	Main method	Main advantage	Main disadvantage	Database
Genomic context	Domain fusion, conserved gene neighborhood, phylogenetic profiles, and co-evolution (De Las Rivas and Fontanillo, 2010; Raman, 2010; Rao et al., 2014)	Interspecies comparison requires few IT resources, fast calculation	Low coverage rate, prediction, using only genomic features	String (Szklarczyk et al., 2019), BioGRID (Oughtred et al., 2021), Hippie (Alanis-Lobato, Andrade-Navarro and Schaefer, 2017), IntAct (Hermjakob et al., 2004a), HPRD (Keshava Keshava Prasad et al., 2009)
Machine learning algorithm	Supervised learning: support vector machine, artificial neural networks, naïve Bayes learning, decision trees (Sarkar and Saha, 2019; Chakraborty et al., 2021) Unsupervised learning: K-means, hierarchical clustering (Bello-Orgaz, Menéndez and Camacho, 2012; Lu et al., 2021)	Handling multi-dimensional and multi-variety data, high efficiency	Data acquisition (massive datasets), High error susceptibility, requires significant IT resources	String, BioGRID, IID (Kotlyar et al., 2019), Hitpredict (Patil, Nakai and Nakamura, 2011)
Text mining	Extracting information from scientific studies and references databases as PubMed Using natural language processing (NLP) technology (Raja, Subramani and Natarajan, 2013; Vyas et al., 2016; Badal, Kundrotas and Vakser, 2018)	Many publications are available, rapidity of execution, inexpensive, easily accessible data	Requests that the interactions be cited in the articles	String, BioGRID, MINT (Chatr-aryamontri et al., 2007), IntAct, HPRD (Keshava Prasad et al., 2009)

was the first computational method to find PPIs from the genome of distinct species based on polypeptides (Marcotte et al., 1999).

The comparison of inter-species sequences can show AB sequences, which are also called Rosetta stones because they allow the interaction between A and B to be deciphered (Date, 2007). This method assumes that if the affinity of A and B increases as B increases when A is fused to B, then pairs of proteins may have evolved from proteins with A and B interaction domains on the same polypeptide (Chia and Kolatkar, 2004; Kamisetty et al., 2011). To improve this method, Veitia, (2002) integrated eukaryotic gene sequences. This incorporation increases robustness of AB polypeptide prediction due to the larger volume of sequences in eukaryotes. A question of equilibrium explains this increase in robustness: the required concentrations of proteins A and B cannot be higher than the equilibrium concentration of AB polypeptides, proteins A and B cannot be separated. Despite the addition of these sequences, few PPIs are found explaining a limited interactome or many PPIs are missing (Latysheva et al., 2016). This method is usually combined with other methods such as machine learning methods (De Braekeleer, Douet-Guilbert and De Braekeleer, 2014; Birtles and Lee, 2021). The accuracy values, therefore, take several methods and are not specific to the domain fusion method. Tagore et al. (2019) have developed the ProtFus tool which combines machine learning, protein fusion, and text mining methods to obtain accuracy values between 75% and 83% to predict PPIs.

Conserved gene neighborhood

This method relies on neighbor gene conservation at the genomic scale. This method compares the position of genes from different genomes to predict potential interactions (Dandekar et al., 1998). For example, a gene is always next to the B gene. Two direct neighboring genes in different genomes suggest interactions. This method is widely used in the prediction of PPIs in eukaryotes (Rogozin et al., 2002). Nomenclature discrepancies in ortholog genes, as well as the search of orthologs that are adjacent on chromosome, explain the low predictive coverage of PPIs (Raman, 2010; Lv et al., 2021). Recently, this method in multi-omics integration has confirmed that bacterial genomes are not randomly organized and can form clusters depending on the local genomic context (Esch and Merkl, 2020). They obtained an accuracy value of 55%. As they mention, this type of method is not intended for the discovery of direct interactions. Recently, a new tool: GENPPI (Anjos et al., 2021), allowing the generation of PPI networks by

taking into account evolutionary relationships that can only be annotated from genomes, namely, conserved gene neighborhoods (CN), phylogenetic profiles (PPs), and gene fusions, has been introduced, showing that these three methods mainly allow the annotation of missing data and thus the understanding of a limited number of interactions. At present, the tool is being tested in their laboratory.

Phylogenetic profiles

This method is based on the comparison of phylogenetic data between gene families of different organisms (Pellegrini et al., 1999; Škunca and Dessimoz, 2015). The phenotypic profile is represented by a binary vector composed of values 0 and 1, corresponding to the absence and presence of proteins in an organism, respectively. Proteins with close or similar phylogenetic profiles tend to be strongly functionally related (Pellegrini, 2019). Ding and Kihara (2018) recently implemented this approach to predict new interactions from known *Arabidopsis thaliana* interactions. The phylogenetic profile approach is combined with machine learning approaches. This method allowed the detection of PPIs with high precision and accuracy. In their work, the performance values range from 75% to 93.2% accuracy.

Coevolution

Coevolution is a fundamental principle of evolutionary theory. Coevolution is defined as the chain of transformation events during the evolution of two species in a mutually dependent manner (de Juan, Pazos and Valencia, 2013). Coevolution results from selective pressure between two or more species (Anderson and de Jager, 2020; Takagi et al., 2020). The interactions of coevolved proteins can be kept either by direct binding or by functional associations (Tillier and Charlebois, 2009). If there is an interaction between two proteins, when one protein mutates, the other protein might have a compensatory mutation, otherwise; two proteins cannot support stability or functions of the interaction during evolution. The evolutionary pressure resulted in the elaboration of co-evolutionary protein pairs in cells that keep the interaction and therefore the function of the protein (Pazos et al., 1997; Goh and Cohen, 2002; Xia et al., 2008).

The global advantage of methods based on the genomic context is the interspecies comparison that requires high computing resources (Sun et al., 2008; Pattin and Moore, 2009). The limitations of these methods are a limited number of predicted PPIs, using only genomic features (Chiang et al., 2007; Raman, 2010; Rao et al., 2014). Recent work by Green et al. (2021) using coevolution had accuracy values of the order of 80% showing promising results for the prediction of protein interaction structures and interfaces. The work of Croce et al. (2019) offered similar results in terms of accuracy for the prediction of protein domain interactions.

The methods based on the genomic context are relevant for evolutionary history analysis, small proteome size, or for experimental verification, agronomic analysis on mutations, or other variants (Koh et al., 2012; Zahiri, Bozorgmehr and Masoudi-Nejad, 2013; Malik, Sharma and Khatri, 2017). On the other hand, these prediction methods are less appropriate for medical data analysis, especially for the search of driving proteins in mechanisms due to the high complexity of the human proteome (Kuzmanov and Emili, 2013; Zhong et al., 2019; Swamy, Schuyler and Leu, 2021).

Methods based on the machine learning algorithm

Machine learning (ML) belongs to the field of artificial intelligence (AI) and computer science. ML algorithms learn from already obtained data to predict outcomes in a specific context (El Naqa and Murphy, 2015; Murdoch et al., 2019). This field has undergone a considerable revolution in the last 10 years with the emergence of promising new methods for PPI prediction (Ding and Kihara, 2018; Kotlyar et al., 2019; Das et al., 2020). ML can be classified into two subclasses: supervised and unsupervised learning. Supervised learning can be defined as a machine learning task that learns to predict from labeled data, conversely; unsupervised learning will learn to predict an outcome on unlabeled data (Zhao, Wang and Wu, 2017; Sarkar and Saha, 2019; Razaghi-Moghadam and Nikoloski, 2020).

Supervised learning method for PPI prediction

Support vector machines

Support vector machines, developed by Vapnik, (1963); (Cortes and Vapnik, 1995), build the best hyperplane to separate training sample classes by a maximal margin, with all positive samples lying on one side and all negative samples lying on the other side. Hyperplane, in the framework of a PPI network, will classify the protein pairs as a binary problem. Protein pairs serve as input, and it classifies if an interaction is possible or not. Protein pairs that are close to the hyperplane are called support vectors and predict an interaction between that pair of proteins (Sarkar and Saha, 2019; Chakraborty et al., 2021).

Ma et al. (2020) developed a method called ACT-SVM for predicting PPIs. This model maps protein sequences to numerical features. Extraction of numerical features is performed twice on the protein sequence to obtain two vectors: a vector and descriptor CT (composition and

transformation) are combined to form a single vector. Feature vectors of a protein pair will be the input of the SVM. The closer these feature vectors of a pair of proteins are to to hyperplane, the higher the probability of an interaction between these proteins.

Dunham and Ganapathiraju, (2021) benchmarked different PPI prediction algorithms, and show how well they perform on realistically proportioned datasets. Based on verified interactions and a known false interaction rate, 16 datasets using the SVM method are generated. Accuracy values ranged from 51 to 96%, which highlights false interactions predicted or not predicted by the SVM methods.

Artificial neural networks

Artificial neural networks (ANNs) are inspired by neural networks in the brain (Wang, 2003; Zhang, 2018). An artificial neural network is composed of different layers with a variable number of neurons, and each layer is connected between them (Yann Lecun, 1986). To simplify, an ANN network works like an artificial neuron that can receive and send information as a signal to the neurons connected to it. This signal is represented by a real number calculated by a non-linear function of the sum of the inputs to a neuron. Neurons and edges can be weighted, and the weighting is adjusted during the learning process. Weight varies according to the intensity of the signal. Signals travel from the first to the last layer, and this results in the output of active neurons (those with a high intensity) (Baxt, 1995; Krogh, 2008; Dongare, Kharde, and Kachare, 2012).

In the context of PPI prediction, artificial neurons represent pairs of proteins. The signal propagates between different artificial neurons. Neurons and edges with high intensity suggest a connection between proteins. A suggested input for these algorithms is the protein sequences of two proteins, other inputs can be put such as 3D structures of proteins (Xie, Deng and Shu, 2020; Pan et al., 2021). The prediction of PPIs based on their amino acid sequences as well as their physiochemical properties is of great interest to understand the probabilistic constraints of the prediction (Ahmed, Witbooi and Christoffels, 2018; Tang et al., 2021). Sharma and Shrivastava (2015) applied an ANN approach that takes the animated acidic sequences of protein pairs as inputs and returns as output whether the pair interacts or not.

The ANN method had quite similar results to the SVM methods. The accuracy values are variable, Hu et al. (2021) showed an accuracy of 71.5% for the prediction of hot spots in a PPI while Pan et al. (2022) observed an accuracy of about 90% in predicting protein interactions in *Arabidopsis thaliana* as a result of this work.

ANNs are exploited as a reference method in several classification tasks (Rohani and Eslahchi, 2019; Baek et al., 2021), but they suffer from some limitations. Artificial neurons that are interaction pairs are checked to limit the introduction of bias during the prediction step (H. Li et al., 2018a; Wu et al., 2021).

Naïve Bayes classifier

A naïve Bayes classifier (NBC) relies on the simple probability of the Bayes' theorem (Bayes et al., 1763). NBC classifies an item by taking each feature of the item independently (e.g., color and shape). To predict a PPI interaction, protein sequences are split into several sub-sequences of n residues. Bayes classifier establishes a probability matrix allowing to classify the different residues; residues that will interact with each other and the non-interface residues. This method is based on conditional probabilities, the probability that is an interaction knowing that an interaction has already occurred. This method will predict interaction sites from protein sequence information alone (Murakami and Mizuguchi, 2010; Geng, Chen and Wang, 2021). Accuracy values are generally lower than those of the SVM and ANN methods, due to the difference in the amount of information available on the proteins.

In PPI prediction, each observation is represented by a vector Z (X_1; X_2;X_3;. . .; X_m,Y), where X{X1,X_2,X_3,. . .,X_m} is the m-dimensional input variable and Y is the output variable taking {0,1}. As input, this method can take either protein interaction datasets or genomic interaction datasets (Jansen et al., 2003; Alashwal, Deris and Othman, 2009; Lin et al., 2021). In the end, the classifier gives a binary response, a zero indicating the interaction is not verified, and a one when there is a potential interaction. Geng et al. (2015) adopted naive Bayes classification to predict site interactions between two proteins. Each pair of proteins is split into several residues, with two residues of two proteins in the same cluster interacting. In terms of performance, they achieved an accuracy value of 60%, which is generally lower than those of the SVM and ANN methods, due to the difference in the amount of information available on the proteins (Ahmed, 2020; Jonathan et al., 2021; Lin et al., 2021).

Identification of interface residues by this method is less expensive and gives results comparable to experimental methods for the prediction of interactions (Murakami and Mizuguchi, 2010; Amirkhah et al., 2015).

Decision trees

A decision tree is a statistical tool that will represent a set of choices as a hierarchical tree. According to different choices made, the algorithm ranks the input elements according to distinctive features: domain presence, spatial folding, site fixation, etc. The decision tree will classify the pair of proteins either as interacting (the proteins in the pair interact with each other) or as non-interacting. Each pair of proteins is characterized by several information and subdomains forming a vector. An interaction is predicted as true if the probability of interactions between two different protein domains is high (Chen and Liu, 2005).

Lee and Oh, (2014) exploited the decision tree method to find discriminating biological features that allow the identification and identify true positive interaction. They have acquired

accuracy averages of 97%. This classification helps to understand the biological context of an interaction. The performance of these methods is dependent on the amount of information available for a biological entity and the projection of low-dimensional features (Xuan et al., 2019; Blassel et al., 2021; Zhou et al., 2021). Li et al. (2021) presented challenges of these methods in terms of performance.

Within supervised methods, a sub-class of methods has emerged in recent years: self-supervised learning methods (Chen et al., 2022; Murphy, Jegelka and Fraenkel, 2022), able to train themselves to learn and predict the output of one part of the input data from another part of the data (Wang et al., 2021; Guo et al., 2022). A graph neural network is a self-supervised method for predicting interactions and in particular PPIs (Mahdipour and Ghasemzadeh, 2021; Jha, Saha and Singh, 2022; Y. Wu et al., 2022b). They are based on machine learning algorithms that extract important information from graphs and use this information to make predictions (Li et al., 2020b; Shen et al., 2021). Jha, Saha, and Singh (2022) developed a method for predicting PPI interactions based on structural information contained in the PDB (Burley et al., 2021) and the sequence characteristics of proteins. The molecular graph of a protein has nodes representing the amino acids (also called residues) of which proteins are made up of. A PPI is formed when pairs of atoms contained in two different residues, have a Euclidean distance less than the threshold distance set, here 6 angstroms. They obtained accuracy values after training of 99.5%. The results of this work show better prediction effectiveness than traditional machine learning methods such as SVM and ANN. Although this method is recent, the resulting accuracy values for interaction prediction are promising such as the prediction of drug–target interactions with an average accuracy value of 89.76% (Zhao et al., 2021), and the prediction of ncRNA–protein interactions with an accuracy value of 93.3% (Shen et al., 2021).

Unsupervised learning method for PPI prediction

The unsupervised analysis includes several methods. The most widely used method is clustering, which aimed to group data into clusters. We will focus on two main clustering methods in the context of creating PPI networks (Malouche, 2013; Creusier and Biétry, 2014).

Clustering methods

K-means clustering and hierarchical clustering methods are unsupervised learning techniques, the most used in the prediction of PPIs (Johansson-Åkhe, Mirabello and Wallner, 2019; Nath and Leier, 2020; Wang et al., 2020; Shirmohammady, Izadkhah and Isazadeh, 2021). Proteins will be clustered according to common characteristics (Ou-Yang, Yan and Zhang, 2017). Clustering steps are repeated to refine the clusters and improve prediction of PPIs (Bello-Orgaz, Menéndez and Camacho, 2012; Lu et al., 2021). Proteins in the same cluster have a high probability of interaction (Geng, Chen and Wang, 2021).

The input data can be of various nature for the prediction of PPIs (Krause, Stoye and Vingron, 2005; Zhao, Wang and Wu, 2017; Wang et al., 2020). Sun et al. (2008) relied on the phylogenetic profile of a protein as input. The phylogenetic profile is a comparative genomic method that predicts the large-scale biological molecule function through evolution information (Mikkelsen, Galagan and Mesirov, 2005). Liu et al. (2018) resorted to hot spot residues databases and in particular the Alanine Thermodynamic Scanning Database. Hot spot residues are functional sites in protein interaction interfaces, and these sites allow the understanding of the type of interactions and are highly conserved in proteins to ensure the functions. Itraq (K. Wang et al., 2018a) used protein sequences as input and hierarchical clustering to identify age-related biomarkers of dental caries. Protein interactions were then successfully validated by multiple reaction control mass spectrometry.

Each of these two clustering methods has sub-methods. For example, hierarchical clustering methods can be divided into two sub-families: “bottom-up” and “top-down” methods (Maimon and Rokach, 2006; Wang et al., 2010; S Bhowmick and Seah, 2015).

Clustering methods are known to be sensitive to noisy data due to experimental bias during acquisition of protein sequences (Arnau, Mars and Marín, 2005; Brohée and van Helden, 2006; Wang et al., 2008). As a result, false-positive interactions appear in the clusters (Sloutsky et al., 2013; Pizzuti and Rombo, 2014; Aghakhani, Qabaja and Alhajj, 2018; Stacey, Skinnider and Foster, 2021).

The global advantage of methods based on machine learning is the processing of multidimensional and multivariate data from several omics or horizontal omics (Das et al., 2020; Jamasb et al., 2021). Prediction of interactions is highly efficient (Terayama et al., 2019; Balogh et al., 2022), but machine learning requires large computational resources and large datasets of good quality (Hashemifar et al., 2018; Y. Wang et al., 2018b).

Machine learning–based approaches are approaches that will be scalable in different domains, these approaches offer very promising results (Casadio, Martelli and Savojardo, 2022; Huang et al., 2022; Pan et al., 2022). However, as we have seen in the articles, many sequences or interactions are necessary to train the model (Li M. et al, 2022; Hu et al., 2022; Jha, Saha and Singh, 2022). So, these approaches will be preferred for large-scale omics approaches, prediction of new interactions, or identification of clusters or hubs (protein with many interactions) (Pei et al., 2021; Song et al., 2022; Stringer et al., 2022). Different studies on PPI by

You et al. (2013), Shirmohammady, Izadkhah, and Isazadeh (2021), and Kusuma et al. (2019), respectively, showed an accuracy of 88%, 63.8%, and 84.6% for clustering methods. This difference in accuracy is explained by the fact that clustering methods depend on the annotations and missing data contained in them (Wang et al., 2010; Zhou et al., 2022).

Methods based on text mining

Text mining is a technique for exploring and transforming unstructured text into structured data (e.g., tables). In PPI prediction, text mining allowed to extracting information about proteins and their interactions from scientific studies and reference databases. Text mining techniques try to automate the extraction of sentence-related proteins from abstracts or paragraphs of text corpora (Papanikolaou et al., 2015). Several text mining methods exist, some are based on statistical matches between gene names, protein names in public repositories, and online resources. Links and types of interactions between proteins are defined by action verbs, for example, interact, interfering, and reacting. He, Wang and Li (2009) benefited from this technique through the PPI finder tool that was developed to extract human PPIs from PubMed abstracts based on their co-occurrences and interaction words, the retrieved interactions are then validated by the occurrence of Gene Ontology (GO) terms. More complex text mining methodologies use advanced dictionaries and generate natural language processes (NLPs) to build networks. The networks generated by these methods have as nodes the names of the genes or proteins, and as edges the verbs found. By these methods, a semantic notion is added (Raja, Subramani and Natarajan, 2013; Badal, Kundrotas and Vakser, 2018; Roth, Subramanian and Ganapathiraju, 2018). Newer methods utilized kernel methods, a class of algorithms for pattern analysis, to predict PPIs from the text. Vyas et al. (2016) applied this method and data mining for disease-related protein identification, functional annotation, and other proteomic studies. The overall advantage of text mining–based methods is the amount of information available and the extremely low cost to acquire PPIs (Alanis-Lobato, 2015; Zhu and Schmotzer, 2017). The main limitation is that the interactors must be close together or in the same sentence (Badal, Kundrotas and Vakser, 2015; Bajpai et al., 2020). Text mining methods have generally high accuracies because PPIs come from the text published as a result of experiments, thus reducing false interactions. For example, the InfersentPPI (Li X. et al, 2022) tool gave an accuracy value of 0.89 for humans, and the ModEx (Farahmand, Riley and Zarringhalam, 2020) tool gave an assurance value of 0.88.

Interaction prediction methods based on text mining are highlighted in the literature because of the large amount of data available in all domains (Jia et al., 2018; Khashan, Tropsha and Zheng, 2022). These methods are recommended for the study of molecular mechanisms and for a large and fast statistical analysis. But in the context of new experiments where little information is available, these methods do not seem to be very suitable (Elangovan, Davis and Verspoor, 2020; Piereck et al., 2020; Shi et al., 2021).

Integration of a PPI network

A set of interactions between different biological entities that allows the study of biological systems is called an interactome (Cusick et al., 2005; Tieri et al., 2014; Guney et al., 2016; Pinu et al., 2019; Halder et al., 2020; Castillo-Arnemann et al., 2021; Wörheide et al., 2021). Understanding molecular interactions and how they give rise to higher-level functions or diseases is important, especially for repositioning drugs, finding new biomarkers, and potentially developing new therapies or elucidating biological and functional processes (Tieri et al., 2014; Guney et al., 2016; Zhou, Miao and Yuan, 2018; Halder et al., 2020; Castillo-Arnemann et al., 2021; Dimitrakopoulos et al., 2021; L. Wu et al., 2022a). These PPI networks can be integrated horizontally and/or vertically (Lercher and Pál, 2008; Ma and Zhang, 2019). Horizontal integration aimed to create a PPI network from different PPI databases for many interactions (Hibbs et al., 2007; Subramanian et al., 2020), whereas vertical integration will assemble information from different omics (genomics, proteomics, metabolomics, etc.) databases for a given interaction (Wang and Jin, 2017; Ulfenborg, 2019; Das et al., 2020; Welch et al., 2021). All interactions can be modeled into a multi-layered graph structure (Kinsley et al., 2020) where each layer represents a network associated with omic-specific information (Hammoud and Kramer, 2020). PPI networks are a central layer in the multi-omics integration process (Mosca and Milanesi, 2013; Hammoud and Kramer, 2020; Dugourd, Christoph Kuppe and Marco Sciacovelli, 2021) (Figure 1).

Horizontal and vertical integration took advantage of topological properties of the network to facilitate construction of different interactomes, to improve classification and evaluation of a PPI (Peng et al., 2017; Kim, Jeong and Sohn, 2019; Halder et al., 2020; Novkovic et al., 2020). Network topology helps in understanding inter/intracellular interactions and functionality, identifying sub-networks (Banerjee et al., 2020; Pournoor et al., 2020; Mishra, Kumar and Mukhtar, 2021). Thus, the topological properties of a PPI network give insight into dynamics of the network and sub-networks and allow the detection of proteins whose roles can be key in complex central biological mechanisms (Yu et al., 2004; Chen et al., 2019; Wahab Khattak et al., 2021). Filtering the network on topological properties allows the acquisition of highly connected nodes and thus facilitates analysis against the topological data. For example, it is possible to filter network by keeping only proteins of a certain degree (Wu et al., 2009; Navlakha et al., 2014; Azevedo and Moreira-Filho, 2015), or by other topological properties from the graph theory such as, degree distribution (Han et al., 2004; Pablo Porras et al., 2020), shortest path (Du

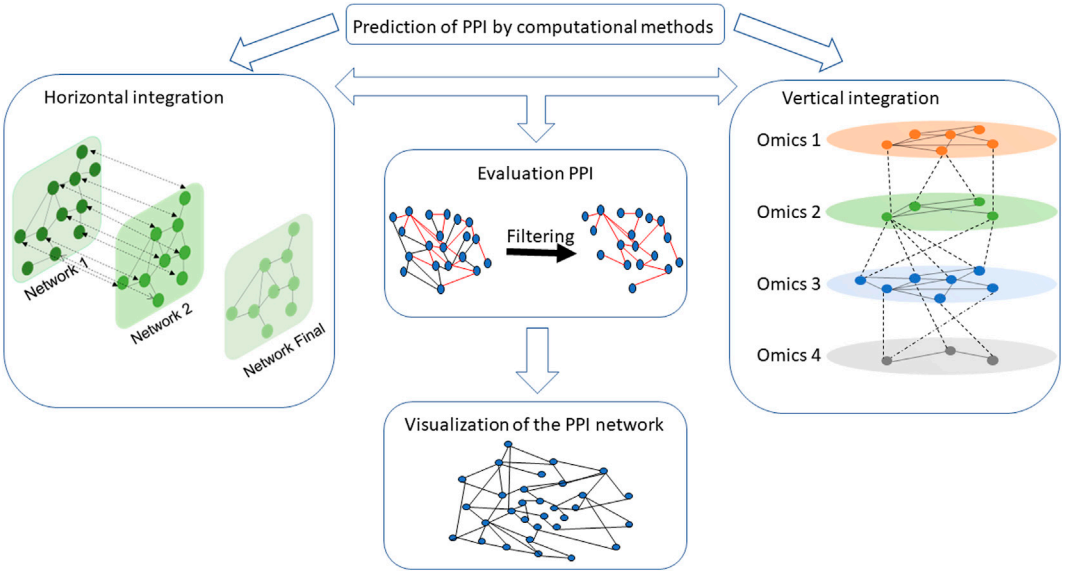


FIGURE 1
Workflow of key steps to design a PPI network assembly. PPI networks can be integrated horizontally and/or vertically. Horizontal integration creates a PPI network by concatenating interaction information from different PPI databases (here networks 1 and 2 represent two PPI networks from two different databases), while vertical integration gathered information from different omics databases for a given interaction. In the vertical integration box, each omics network represents different interactomes such as protein–protein, drug–protein, and RNA–protein. Once the networks are generated, it is necessary to evaluate its interactions confidence to filter the network. Interactions in red are interactions with a high confidence score. After narrowing the network, specialized tools can be used to visualize the network and information about the connected entities (e.g., identify proteins with a central role in the mechanisms).

et al., 2014), and transitivity (Hakes et al., 2008; Lynn and Bassett, 2021).

Integration of a PPI network in a multi-omics context is nowadays an essential issue in the understanding of biological mechanisms (Hawe, Theis and Heinig, 2019; Bodein et al., 2021; Dimitrakopoulos et al., 2021). To integrate an interaction into a network, it must first be estimated by a so-called confidence score (Stelzl and Wanker, 2006; Li et al., 2016; Xu et al., 2021), representing probability that the interaction is accurately identified by algorithms and is expressed as a percentage (Kamburov et al., 2012; Peng et al., 2017). This score is usually a ratio of the measured value to the total number of the measured value for each interaction. For example, the Mi-score measures the number of publications observed for an interaction out of the overall number of publications available to the network (Villaveces et al., 2015a). Sub-networks represent a part of the network retaining only interactions with a high confidence score (Flórez et al., 2010; Pietrosemoli and Dobay, 2018; Hao et al., 2019), which can also be extracted to facilitate visualizations. Proteins forming groups called clusters in the sub-networks are recovered. By modifying the threshold of the confidence score, we can better define new clusters and the impact size of the sub-network.

Horizontal integration of a PPI network

Horizontal integration is a solution to eliminate these false interactions and allows to find missing data, thus adjusting the resulting confidence score (Everson et al., 2019; Gebreyesus et al., 2022). Horizontal integration methods have contributed to development of various types of databases based on organism-specific diseases, biological processes, and detection methods, such as the Integrated Interactions Database (IID) (Kotlyar et al., 2019), IntAct (Hermjakob et al., 2004a), and StringDB (Szklarczyk et al., 2019). PPI is usually redundant in different databases. A PPI found in one database may also be found in others such as BioGRID (Oughtred et al., 2021) or Reactome (Gillespie et al., 2022). This communication between the different databases corresponds to horizontal data integration (Zitnik and Leskovec, 2017; Cowman et al., 2020).

Assembly and merging are the main algorithms for horizontal integrations (De Las Rivas, Alonso-López and Arroyo, 2018; Amanatidou and Dedoussis, 2021). Two PPI networks are assembled by alignment algorithms. Alignment of PPI networks aimed at finding topological and functional similarities between different PPI networks (Kazemi et al., 2016; Ma and Liao, 2020). In a first step, the alignment algorithm looks for overlapping regions in two networks. These regions form clusters that will be assembled to make a

local alignment. Then, using local interactions between clusters, a second alignment is performed: global alignment (Malod-Dognin, Ban and Pržulj, 2017; Alcalá et al., 2020; Chow et al., 2021). Other horizontal integration algorithms applied propagation algorithms as the random walk with restart (RWR) process (detailed in vertical integration of a PPI network). Xu et al. (2018) drawled on these propagation methods to reconstruct a multi-level PPI network and identify protein complexes.

Through these different network alignment algorithms, many PPI databases have been updated or created. The most exploited are BioGRID (Oughtred et al., 2021), IntAct (Hermjakob et al., 2004b), String (Szklarczyk et al., 2019), and UniprotKB (The UniProt Consortium, 2019). A large set of databases is referenced in startbioinfo.org (Kshitish et al., 2013) and pathguide.org (Bader, Cary and Sander, 2006). Following the revolution in NGS technology and the increase in PPI datasets, the integration of a single cell with PPI networks is showing promising results. Indeed, the single-cell method coupled PPI network will allow the understanding of gene regulation, cellular heterogeneity (Cha and Lee, 2020), tissue-specific networks, identification of ligand–receptor interactions, functional interactions, and cell–cell communication (Armingol et al., 2021; Johnson et al., 2021; F. Ma et al., 2021a). Cell–cell interactions mediated by ligand–receptor complexes are essential for the coordination of various biological processes, such as development, differentiation, and inflammation. These interactions subsequently ensure that physiological processes are carried out (Vento-Tormo et al., 2018; Efremova et al., 2020). Using single-cell data and PPI networks, it will be possible to understand this crucial interaction and thus to create new therapies targeting these ligand–receptor interactions in future (Ji et al., 2020; Lee et al., 2021). The applications of single cell PPI are numerous and in many fields such as health (Qi et al., 2022) and agronomy (Zhang et al., 2019). These methods will help in the understanding of cellular mechanisms, regulation according to the environment, and in the development of new therapy (Ryu et al., 2019; Mahdessian et al., 2021). Single-cell data can also be used to filter and weight the PPI network following a differential analysis or by filtering according to fluorescence (Dünkler et al., 2015; Wu et al., 2017). Recently, Klimm et al. (2020) have developed SCPPIN, a method of integrating single-cell RNA-seq data with protein–protein interaction networks. By filtering the network by differentially expressed genes and maximum subgraph weight, they detected active modules in cells of different transcriptional states.

However, horizontal integration faces problems such as uniformity of protein interaction identifiers and redundancy of information, data structure, and organization (Dohrmann, Puchin and Singh, 2015; L. Liu et al., 2020a).

Vertical integration of a PPI network

Vertical integration of networks is generally represented by multi-layer networks (Lv et al., 2021; Watson, Schwartz and Francavilla, 2021). Each layer represents an interactome (protein, gene, and drug). Biological relationships between biological entities and types of interactions form the relationships between different omics layers (Lee and Nam, 2018). Network propagation (or diffusion) algorithms are commonly promoted in omics vertical integration (Di Nanni et al., 2020; Pak et al., 2021). By integrating the information from the different omics and by diffusion algorithms, it is possible to understand the most probable interactions where the diffusion signal has strongly transited (Zhao et al., 2018). Propagation algorithms are a class of algorithms that integrate input data information across connected nodes of a given network. Propagation is usually performed by random walk with restart (RWR) algorithms, inspired by the work of Page et al. (1999) to classify web pages in an objective and mechanical way. RWR is the state-of-the-art approach to infer the relationship: as the name suggests, a random walker, starting from a set of nodes of interest (starting nodes), jumps to neighboring nodes, or nodes in another layer according to a certain probability assigned to the edges of the nodes (Lee and Yoon, 2018). In addition, the walker has a certain probability, known as the damping factor, such that for each step taken in any direction, there is a probability associated with returning to one of the original sets of nodes (Valdeolivas et al., 2019; Nguyen et al., 2021; Qu et al., 2021; Wen et al., 2021). The probability is calculated from a transition matrix from one node to the other, allowing to obtain a weight for each interaction. This node-dependent weight will reflect an interaction between two omics layers (Bhatia, 2019; Dupré, 2022). Lei et al. (2019a) adjusted this method to detect essential proteins. In this method, PPIs are weighted according to network topology, gene expression, and GO annotation data. Then, an initial score is assigned to each protein in a PPI network by exploiting information on subcellular localization and protein complexes. Then the RWR algorithm is applied to the weighted PPI networks to iteratively score the proteins, allowing the filtration of interactions with high weight.

The main other algorithms based on topological properties use integration strategies from two classes: empirical methods and machine learning method (Jin et al., 2014; Haas et al., 2017; Eicher et al., 2020). Empirical methods simply assembled different layers of the network, whereas machine learning methods tried to find missing information about how information flows between the omics layers (Picard et al., 2021; Santiago-Rodriguez and Hollister, 2021). MoGCN (Li X. et al., 2022) is a tool for multi-omics integration based on a convolutional graph network. This tool allows the classification and analysis of cancer subtypes. MoGCN can extract the most significant topological features and properties of each omic layer for downstream biological knowledge discovery.

Integration of PPI networks into multi-layer networks has a central role (Liang et al., 2019; Huang and Zitnik, 2021). Indeed, projection of PPI and layer connectivity allows improvement of the mechanistic and functional knowledge of a cell, identifying key proteins and repositioning drugs (F. Li et al., 2020c). Silverbush and Sharan (2019) created an approach to direct the human PPI network using the drug response and cancer genomic data. A directed graph is a graph in which the edges have a direction. The direction of the relationships or edges is found by diffusion methods. The oriented network allows the detection of key genes in cancers.

In vertical or horizontal integration, the PPI layer must be reliable. The topological properties of the network can allow the establishment of a confidence score for a given interaction. It is essential to understand these properties to build the most robust network possible (Zhang, Xu and Xiao, 2013; Sardiu et al., 2019).

Validation of PPI

An important question persists in network analysis: can we trust on the network of interactions to be a true biological interaction? PPIs from these methods have supplied insights into functions of individual proteins, regulatory pathways, molecular mechanisms, and entire biological systems. Noise inherent in the interactome information hinders evaluation of PPI data (Correia et al., 2019). Several PPIs are, in fact, false positives in these methods and even in methods using strict criteria to define a positive (Yu et al., 2004; Scott and Barton, 2007). It should be noted that the coverage of the interactome is also incomplete and uneven, so we cannot always filter out the less reliable evidence (Han et al., 2005; Stelzl and Wanker, 2006). Many different methods exist for finding reliability and giving a measure of confidence. These techniques can be classified into three main categories.

Contextual biological information

This strategy for assessing the veracity of an interaction looked for different information, for example, overlapping patterns of co-expression, conservation of structure, and sequences (Aytuna, Gursoy and Keskin, 2005; Tirosh and Barkai, 2005). As an example, Schaefer et al. (2013) seek biological information based on influenza virus knowledge to validate PPIs.

Scores based on the literature

Acts as an orthogonal validation and analyzed how often a PPI is cited in publications. The main problem with implementing this method is the application of thresholds, so that only interactions with a sufficiently high score are retained (Bozhilova et al., 2019). Well-studied proteins will have a greater number of interactions and associated publications than proteins that are new or have little information. Hence, thresholds need to be standardized. In order to normalize thresholds among different databases, the MI-score method was created (Villaveces et al., 2015a). This method allows to merge data from different databases that are in the PSI–MI(Proteomics Standards Initiative–Molecular Interaction) format (Hermjakob et al., 2004a; Bader, et al., 2006; Kerrien et al., 2007), and link an interaction to a notation system. This method generates three different scores: publication score (number of different publications on an interaction), method score (considers the different methods of detecting an interaction), and the type of score which refers to the type of interaction. The type of interaction follows the nomenclature of the PSI-MI controlled vocabulary, for example, genetic interaction, physical association, and co-location.

Aggregated methods

Use different score calculation strategies and combine these strategies into a single score. Several scoring methods exist, including the toolkit developed by Braun et al. (2009) that includes four statistical tests to verify a PPI from a high-throughput experiment. The results of the four tests are then combined to calculate the probability that a new pair of interactions is a true biophysical interaction. Intscore is a reference aggregation tool, which calculates confidence scores for user-specified sets of interactions. Its scoring system is based on network topology and annotations. The aggregated score can be computed by machine learning approaches (Kamburov et al., 2012). Recently, Paul and Anand (2022) developed several similarity measures using GO to create a confidence score for PPIs.

Apart from these three distinct categories, to measure the confidence of PPIs, robust measures resulting from data provenance and network topology are needed, such as the average redundancy difference between various sources, natural connectivity of the PPI network as well as the number of edges in a protein-centered sub-arrays (ego networks) (Bozhilova et al., 2019; Wang et al., 2019). The main problem with all these methods is that a score is mainly specific to one database, so threshold values are highly database dependent (Kamburov et al., 2012; Dahiya et al., 2019; Xu et al., 2019). To address this issue, consensus networks appeared such as HugGan (Huang et al., 2022) which is a tool that gathers 31 data sources using deep learning approaches to keep only interactions with a high confidence score resulting in a network with high coverage and quality.

Visualization of protein–protein networks

Networks are a powerful way to visualize complex systems (Charitou, Bryan and Lynn, 2016; Mlecnik, Galon and Bindea, 2018). Visualization of PPI networks is crucial for the understanding of pathways, sub-graphs, sub-network, and

TABLE 2 Summary table of tool for visualizing of protein–protein interaction network. Visualization methods to analyze network are grouped into three distinct categories: visualization through downloadable tools, visualization by libraries integrated with languages, and visualization through graph-oriented databases. The user has to choose his tools according to his study context. For analysis of high dimensional data containing a large amount of information, it is advisable to manipulate tools based on graph databases. Conversely, if the user wants to have a quick representation, we recommend the user to turn more to visualization libraries or downloadable software.

	Tool	Advantage	Disadvantage
Visualization through downloadable tools	Cytoscape (Otasek et al., 2019), Gephi (Bastian, Heymann and Jacomy, 2009), Tulip (Auber et al., 2017), Graphviz (Ellson et al., 2001), Pajek (Mrvar and Batagelj, 2016)	Many add-on features, flexibility for network analysis, easy to handle, open source and free	Difficult to set up automation interface, working with big networks requires big memory and computing power
Visualization by libraries integrated with languages	Igraph (Csárdi and Nepusz, 2006), NetworkX (Hagberg et al., 2008), graph-tool (Peixoto Tiago, 2014), NetView (Neuditschko, Khatkar and Raadsma, 2012)	Open source and free, well documented, accessible, import and export graphs easily, easy to implement	Graphic possibilities are limited, restricted number of nodes
Visualization through graph-oriented databases	Neo4j (Gong et al., 2018), ArangoDB (*ArangoDB NoSQL Multi-Model Database: Graph, Document, Key/Value*, 2022), JanusGraph (Sharp. 2017), OrientDB (Tesoriero, 2013), Elasticsearch (Shay Banon, 2014), Siren (Giovanni Tummarello and Renaud, 2015)	Speed of calculation, adapted big networks, integrated search engine, Flexible and agile structures	Request for calculation servers. Not very scalable as it is designed for a single server architecture

central proteins (Sharan and Ideker, 2006; Fionda et al., 2009; Snider et al., 2015; De Las Rivas, Alonso-López and Arroyo, 2018; Vella et al., 2018; Marai et al., 2019). The simplistic and rapid visualization of networks makes it a tool of choice (Gillis, Ballouz and Pavlidis, 2014; Chung et al., 2015; Xu et al., 2021). This has led to the development of methods and tools that allow visualization. The integration of PPI networks and their visualizations in a multi-omics context has helped in the modeling of complex systems such as Parkinson's disease (Tomkins and Manzoni, 2021), identifying central proteins in diseases (Narayanan et al., 2011; Deng, Xu and Wang, 2019), understanding protein clusters linked to cellular function (Zhao, Wang and Wu, 2017; Amanatidou and Dedoussis, 2021), understanding mechanisms of action (Jia et al., 2021; Yuan et al., 2021), and drug repositioning (Lee and Yoon, 2018; Soleimani Zakeri, Pashazadeh and MotieGhader, 2021).

Larger and complex networks are more difficult to visualize. This is the case of the most popular source offering a representation of PPI networks such as StringDB (Szklarczyk et al., 2019). This online database is intended for the inspection of small networks or sub-networks (less than 500 interactions). Therefore, because of their size and topology, the PPI network requires specialized tools (Bosque et al., 2014; Freilich et al., 2018; Aihaiti et al., 2021).

The methods for visualizing a network can be divided into three categories (Table 2).

The methods can be combined to take advantage of each of the benefits of these categories. This is the case with cyNeo4j (Summer et al., 2015) which combines Cytoscape (Otasek et al., 2019) and Neo4j (Gong et al., 2018) for fast visualization of large networks based on a graph-oriented database. Cytoscape is the most widely used tool for the visualization of large networks (Shannon et al., 2003). Other visualization systems do not fit into these categories and are based on web-based visualization interfaces and on a relational database (Salazar-Ciudad and Jernvall, 2013; Salazar et al., 2014; Hayashi et al., 2018). This is the case of the PINA 3.0 (Du et al., 2021) tool, which is a consensus database containing five interactomes and offering a web visualization service allowing the identification of interacting protein pairs in different cancer types. The weaknesses of these methods are the size of the networks, the execution time of a query, and their limited applicability (Jeanquartier, Jean-Quartier and Holzinger, 2015; Zhou and Xia, 2018; Perlasca et al., 2020).

Visualization tools are evaluated by four criteria: compatibility (available on which OS (operating systems): Windows, Mac Os, and Linux, analytic functions (presence of functions measuring the topological properties of the network, weak interactions of external data, etc.), visualizations (graph layout, dynamics, and parallel implementation), and the extensibility of the tool (addition of plugins, type of input, and output file) forming distinct classes (Sanz-Pamplona et al., 2012; Agapito, Guzzi and Cannataro, 2013; Dallago et al., 2020). In the context of biological network analysis and in particular protein networks, one of the essential criteria is dynamic visualization tools (Xia, Benner and Hancock, 2014; Zhou and Xia, 2018). PPI networks have a dynamic organization of biological sub-networks (Yang, Wagner and Beli, 2015). In other words, the molecular interactions in a cell vary in time, as do the signals from the environment surrounding an interaction (Przytycka, Singh and Slonim, 2010; M. Li et al., 2018b).

In order to overcome the limitation of network size and consider the dynamics of the networks, several tools have been developed over the last decades (Sanz-Pamplona et al., 2012; Winkler et al., 2021). The success of Cytoscape is due to the large number of plugins/features that can be added directly from the tool (Saito et al., 2012; Lotia et al., 2013). The calculation of overrepresented GO terms in a network can be performed by Bingo (Maere, Heymans and Kuiper, 2005), a widely downloaded Cytoscape plugin. Through Cytoscape, we also find plugins allowing the understanding of the dynamic organization of biological networks such as TVNViewer (Curtis et al., 2011), KDDN (Tian et al., 2015), and Dynetviewer. Another downloadable software offering a visual representation of PPI networks is the Gephi (Bastian, Heymann and Jacomy, 2009). Downloadable network visualization tools have difficulties with the implementation of data (Villaveces et al., 2015b; Li et al., 2016). Visualization libraries such as igraph (Csárdi and Nepusz, 2006) and NetworkX (Hagberg et al., 2008) will make it easier to import and export networks but are limited in terms of adding new functionality and graphic possibilities (Pandey, 2018; L. Wu et al., 2022a).

Network visualization tools are specific to the detection method (Ashtiani et al., 2018). HPIminer (Subramani et al., 2015) extracts information from human PPIs and PPI pairs in biomedical literature and provides a visualization of interactions, networks, and associated pathways using two databases, namely, HPRD (Goel et al., 2012) and KEGG (Kanehisa et al., 2016). Another area of improvement for online or general-purpose visualization tools and libraries is the addition of a visualization engine or search engine (Chisanga et al., 2017). Tools integrating visualization engines such as NAViGaTOR (Brown et al., 2009) and MIST (Hu et al., 2018) have been developed. These tools allow the acceleration of the visualization of large PPI networks (Yu and Zhang, 2008; Gerasch et al., 2014; Zaki and Tennakoon, 2017). It is also possible to improve the speed of visualizations by connecting directly to graph databases such as Neo4j (Gong et al., 2018, p. 4) and ArangoDB (Touré et al., 2016; Timón-Reina, Rincón and Martínez-Tomás, 2021; *ArangoDB NoSQL Multi-Model Database: Graph, Document, Key/Value*, 2022). Since graph databases store data directly in a graph form, they are becoming a preferred resource for storing complex relationships of heterogeneous biological data (Yoon, Kim and Kim, 2017; Jupe et al., 2018; Castillo-Arnemann et al., 2021). Flexibility of multi-omics integration offered by graph databases facilitates data mining to support different hypotheses (Lysenko et al., 2016; Brandizi et al., 2018; Wandy and Daly, 2021).

All these tools for the visualization of PPI networks are based on different visualization algorithms (Koutrouli et al., 2020; Sandoval and Orlando, 2021). Visualization algorithms can be based on simplistic approaches such as adjacent matrices (Fekete, 2009), circular layouts (Suderman and Hallett, 2007), or complex approaches such as force-directed algorithms (Liu et al., 2021). The main differences between simple and complex algorithms for visualization depend on the size of the network, the topology of the network, and the dimensionality of the information (Heberle et al., 2017; Becker et al., 2020; Raja et al., 2020). The selection of the appropriate visualization algorithm will depend on the nature of the network. In the context of single networks, in particular PPI networks, visualization algorithms focus on the identification of protein sub-clusters or hub proteins (Li et al., 2020b; H. Ma et al., 2021b). Cytoscape's Cytohubba (Chin et al., 2014) plugin is commonly dedicated for sub-network identification and central protein identification. The most powerful method of Cytohubba for better sub-network visualization is the maximum clique centrality (MCC) method. This algorithm allows the visualization of groups of proteins called clusters, based on the assumption that essential proteins tend to be grouped together (Lu et al., 2010; Lei et al., 2019b; Kim, Jeong and Sohn, 2019). Recently, Zu et al. (2017) used this plugin's method to visualize six target genes for quercetin (an organic compound of the flavonoid family), suggesting a therapeutic potential in type 2 diabetes mellitus (T2DM) and Alzheimer's disease.

However, in a multi-omics integrations context one seeks above all to connect information from different omics fields (transcriptomics, proteomics, metabolomics, lipidomics, and metabolomics (Haas et al., 2017; Fan, Zhou and Ressom, 2020; Cansu Demirel, Kaan Arici and Tuncbag, 2022). In this context, multi-layer algorithms for visualization are preferable to force-directed algorithms (Bodein et al., 2021; Dursun, Kwitek and Bozdag, 2021; Marín-Llaó et al., 2021). There are several algorithms for implementing multi-layer networks, in the context of multi-omics integration, the most highlighted implementation is the one named by Hammoud and Kramer, (2020): "Interactive/Interconnected/Interdependent Networks and Networks of Networks Implementation." This implementation has as input a set of monoplex networks (single layer networks, e.g., PPI network). Each network interacts with the other networks. The different monoplex networks will form distinct layers which will be connected by the inter-side nodes (Rappoport and Shamir, 2018; Yan et al., 2018; Zoppi et al., 2021; Cuenca et al., 2022). Recently Arena3dweb (Karatzas et al., 2021), a web application incorporating these algorithms and offering a visualization of multi-layer graphs in a 3D space, has enabled GPCR signaling pathways implicated in melanoma.

Summary and outlook

In this review, different computational strategies for predicting PPI, from integration to visualization to methods for validating interactions have been studied. Many computational prediction approaches rely on experimental methods to predict a PPI interaction (Rao et al., 2014; Peng

et al., 2017; Ding and Kihara, 2018; Tanwar and George Priya Doss, 2018). Although this increases the coverage of the network, it can disrupt the horizontal integration process (Browne et al., 2010; Ngounou Wetie et al., 2013). Sets of PPI interactions from different datasets are constructed and transformed independently, which can lead to information gaps, redundant information, and poor identifier compatibility when aligning two PPI networks. Ideally, at any point in the overall integration process (including vertical and horizontal), each omics data set should be evaluated in the context of the other datasets, so that complementary information can be fully exploited, and added information can be identified (Bozhilova et al., 2019; Bajpai et al., 2020). Implementation of validation scores based on topological properties allows to limit the redundancy of edges and will allow to filter the PPI network (Pietrosemoli and Dobay, 2018; Sardiu et al., 2019).

Information redundancy is the repetition of information without adding additional information in different databases. The increase in omics data and PPI integration methods has contributed to the growth of many PPI databases. However, this increase in the number of databases increases the redundancy of information, making it difficult for the user to choose a PPI database (Rabbani et al., 2018; Hawe, Theis and Heinig, 2019; Zahiri et al., 2020). In addition, information redundancy slows down the calculation time for the construction and visualization of networks (Chen et al., 2019, 2019). To limit and remove redundancy, different information scores have been set up (Silverbush and Sharan, 2019; Mahdipour and Ghasemzadeh, 2021). The Mi-score (Villaveces et al., 2015b) consisting of three scores, is increasingly used to validate a PPI.

The study of PPI networks is a growing field of systems biology. Due to their significant role, PPI networks are used to understand cellular functions or biological mechanisms (Stelzl and Wanker, 2006; Jordán, Nguyen and Liu, 2012; Safari-Alighiarloo et al., 2014). The integration of these networks, both vertically and horizontally, can highlight clusters of proteins with central roles, aiding the understanding of drug action mechanisms (Martin, Roe and Faulon, 2005; Dimitrakopoulos et al., 2021; Marín-Llaó et al., 2021; Tomkins and Manzoni, 2021). PPI networks offer prospects in many fields, such as medicine, health and also in agri-food (Hao et al., 2019; Hasan et al., 2020; Thanasomboon et al., 2020; Charmpi et al., 2021). Vertical and horizontal integration algorithms are mainly based on propagation and alignment algorithms but are often combined with machine learning methods to predict the probability of reliability of an interaction (Li and Ilie, 2017; Lee and Nam, 2018; Zhang et al., 2018; Das and Chakrabarti, 2021). These propagation algorithms will allow to focus on sub-networks, keeping only the interactions where the propagation signal is high (Gehlenborg et al., 2010; Laniau, 2017).

By focusing on sub-networks as opposed to complete networks, visualization is facilitated allowing the identification of sub-groups of interactions (Tian, Ju and Yang, 2019; T.-H. Liu et al., 2020b). The visualization of networks is a problematic issue for networks and especially for PPI networks (Du et al., 2021). Visualization tools depend mainly on the size of our networks (Summer et al., 2015; Zou et al., 2017). Currently, multilayer network visualization is limited to small networks and requires a consequent pre-formatting of the data (Smith-Aguilar et al., 2019; Hammoud and Kramer, 2020; Sebestyén, Domokos and Abonyi, 2020).The study of multilayer networks based on the PPI network is constantly evolving and will become more powerful with advancement of more powerful mathematical models offering better predictions (Kapadia et al., 2019; Karatzas et al., 2021; Cuenca et al., 2022). Different perspectives on the integration of PPI networks can be imagined. The visualization of multilayer multi-omics networks and creation of consensus networks for each omics dimension to understanding new mechanisms of multi-omics integration. A consensus network is the result of the horizontal integration of different databases (Berto et al., 2016; Mosca et al., 2021). Through this network, it will be possible to homogenize the different thresholds of the different databases and to eliminate the recurrence of information (Leblanc et al., 2013; Affeldt et al., 2016; Zohra Smaili et al., 2021). Recently, Woo and Yoon (2021) created a Monaco aligner that can find multiple alignments with high accuracy to identify functional modules. In the era of big data and NGS (next generating sequencing) technologies, it is difficult to know which information is needed to build a PPI network. Machine learning and deep learning methods offer novel perspectives in the prediction and standardization of information in PPI networks (Gligorijević and Pržulj, 2015; Borhani et al., 2022; Cervantes-Gracia, Chahwan and Husi, 2022). Standardizing and evaluating the relevance of interactions will facilitate integration of PPI networks (Fiorentino et al., 2021; Nadeau, Byvsheva and Lavallée-Adam, 2021).

On the visualization side, several perspectives can be imagined, a tool to visualize each layer independently and globally in a multilayer network (Kanai, Maeda and Okada, 2018; McGee et al., 2019). As the size and complexity of PPI networks increases, more efficient visualization algorithms are needed (Chong, Wishart and Xia, 2019; Koutrouli et al., 2020). Augmented reality technologies and virtual reality (VR) remove the constraints of 2D/3D space constraints (Pirch et al., 2021; Hütter et al., 2022). Moreover, the notable advances in the prediction of the structure of proteins from their sequence in amino acids with alphafold (Jumper et al., 2021), which could lead to a revolution in the PPI prediction algorithm. In view of the generous size of PPI networks, visualization tools focus on specific networks, including

Mechnetor (González-Sánchez et al., 2021), a tool for visualization of biological mechanisms. At the moment, there are no tools available to visualize the interactome protein specific to a tissue, but there are different databases on this subject (Islam et al., 2013; Basha et al., 2018).

Author contributions

VR wrote the manuscript, VR designed the figures and tables, and VR, AB, MPSB, ML, OP and AD revised the manuscript. AD supervised the research.

Funding

This work was supported by Research and Innovation chair L'Oréal in Digital Biology.

Conflict of interest

Author OP is employed by company L'Oréal.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Affeldt, S., Sokolovska, N., Prifti, E., and Zucker, J. D. (2016). Spectral consensus strategy for accurate reconstruction of large biological networks. *BMC Bioinforma.* 17 (16), 493. doi:10.1186/s12859-016-1308-y

Agapito, G., Guzzi, P. H., and Cannataro, M. (2013). Visualization of protein interaction networks: Problems and solutions. *BMC Bioinforma.* 14 (1), S1. doi:10.1186/1471-2105-14-S1-S1

Aghakhani, S., Qabaja, A., and Alhajj, R. (2018). Integration of k-means clustering algorithm with network analysis for drug-target interactions network prediction. *Int. J. Data Min. Bioinform.* 20, 185. doi:10.1504/IJDMB.2018.10016075

Ahmed, I., Witbooi, P., and Christoffels, A. (2018). Prediction of human-Bacillus anthracis protein-protein interactions using multi-layer neural network. *Bioinforma. Oxf. Engl.* 34 (24), 4159–4164. doi:10.1093/bioinformatics/bty504

Ahmed, M. (2020). Modified naive Bayes classifier for classification of protein- protein interaction sites. *J. Biosci. Agric. Res.* 26, 2177–2184. doi:10.18801/jbar.260220.266

Aihaiti, Y., Song Cai, Y., Tuerhong, X., Ni Yang, Y., Ma, Y., Shi Zheng, H., et al. (2021). Therapeutic effects of naringin in rheumatoid arthritis: Network pharmacology and experimental validation. *Front. Pharmacol.* 12, 672054. doi:10.3389/fphar.2021.672054

Alachram, H., Chereda, H., BeiBbarth, T., Wingender, E., and Stegmaier, P. (2021). Text mining-based word representations for biomedical data analysis and protein-protein interaction networks in machine learning tasks. *PloS One* 16 (10), e0258623. doi:10.1371/journal.pone.0258623

Alanis-Lobato, G., Andrade-Navarro, M. A., and Schaefer, M. H. (2017). HIPPIE v2.0: Enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res.* 45, D408–D414. doi:10.1093/nar/gkw985

Alanis-Lobato, G. (2015). Mining protein interactomes to improve their reliability and support the advancement of network medicine. *Front. Genet.* 6. Availableat: https://www.frontiersin.org/article/10.3389/fgene.2015.00296 (Accessed: February 25, 2022).

Alashwal, H., Deris, S., and Othman, R. M. (2009). *A bayesian kernel for the prediction of protein- protein interactions*, 6.

Alcalá, A., Alberich, R., Llabres, M., Rossello, F., and Valiente, G. (2020). AligNet: Alignment of protein-protein interaction networks. *BMC Bioinforma.* 21 (6), 265. doi:10.1186/s12859-020-3502-1

Amanatidou, A. I., and Dedoussis, G. V. (2021). Construction and analysis of protein-protein interaction network of non-alcoholic fatty liver disease. *Comput. Biol. Med.* 131, 104243. doi:10.1016/j.compbiomed.2021.104243

Amirkhah, R., Farazmand, A., Gupta, S. K., Ahmadi, H., Wolkenhauer, O., and Schmitz, U. (2015). Naïve Bayes classifier predicts functional microRNA target interactions in colorectal cancer. *Mol. Biosyst.* 11 (8), 2126–2134. doi:10.1039/c5mb00245a

Anjos, W. F., Lanes, G. C., Azevedo, V. A., and Santos, A. R. (2021). Genppi: Standalone software for creating protein interaction networks from genomes. *BMC Bioinforma.* 22 (1), 596. doi:10.1186/s12859-021-04501-0

ArangoDB NoSQL Multi-Model Database: Graph, Document, Key/Value (2022). ArangoDB. Availableat: https://www.arangodb.com/(Accessed March 2, 2022).

Armanious, D., Schuster, J., Tollefson, G. A., Agudelo, A., DeWan, A. T., Istrail, S., et al. (2020). Proteinarium: Multi-sample protein-protein interaction analysis and visualization tool. *Genomics* 112 (6), 4288–4296. doi:10.1016/j.ygeno.2020.07.028

Armingol, E., Officer, A., Harismendy, O., and Lewis, N. E. (2021). Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* 22 (2), 71–88. doi:10.1038/s41576-020-00292-x

Arnau, V., Li, S., and Marín, I. (2005). MarsIterative cluster Analysis of protein interaction data. *Bioinformatics* 21 (3), 364–378. doi:10.1093/bioinformatics/bti021

Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H., Wolkenhauer, O., Mirzaie, M., et al. (2018). A systematic survey of centrality measures for protein-protein interaction networks. *BMC Syst. Biol.* 12, 80. doi:10.1186/s12918-018-0598-2

Auber, D., Archambault, D., and Bourqui, R. (2017). "Tulip 5," in *Encyclopedia of social network analysis and mining*. Editors R. Alhajj and J. Rokne (Springer), 1–28. doi:10.1007/978-1-4614-7163-9_315-1

Aytuna, A., Gursoy, A., and Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21. Oxford, England, 2850–2855. doi:10.1093/bioinformatics/bti443

Azevedo, H., and Moreira-Filho, C. A. (2015). Topological robustness analysis of protein interaction networks reveals key targets for overcoming chemotherapy resistance in glioma. *Sci. Rep.* 5 (1), 16830. doi:10.1038/srep16830

Badal, V. D., Kundrotas, P. J., and Vakser, I. A. (2018). Natural language processing in text mining for structural modeling of protein complexes. *BMC Bioinforma.* 19 (1), 84. doi:10.1186/s12859-018-2079-4

Badal, V. D., Kundrotas, P. J., and Vakser, I. A. (2015). Text mining for protein docking. *PLoS Comput. Biol.* 11 (12), e1004630. doi:10.1371/journal.pcbi.1004630

Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: A pathway resource list. *Nucleic Acids Res.* 34, D504–D506. doi:10.1093/nar/gkj126

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Sci. (New York, N.Y.)* 373 (6557), 871–876. doi:10.1126/science.abj8754

Bajpai, A. K., Davuluri, S., Tiwary, K., Narayanan, S., Oguru, S., Basavaraju, K., et al. (2020). Systematic comparison of the protein-protein interaction databases from a user's perspective. *J. Biomed. Inf.* 103, 103380. doi:10.1016/j.jbi.2020.103380

Balogh, O. M., Benczik, B., Horvath, A., Petervari, M., Csermely, P., Ferdinandy, P., et al. (2022). Efficient link prediction in the protein–protein interaction network

using topological information in a generative adversarial network machine learning model. *BMC Bioinforma.* 23 (1), 78. doi:10.1186/s12859-022-04598-x

Banerjee, K., Jana, T., Ghosh, Z., and Saha, S. (2020). PSCRIdb: A database of regulatory interactions and networks of pluripotent stem cell lines. *J. Biosci.* 45, 53. doi:10.1007/s12038-020-00027-4

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* 5 (2), 101–113. doi:10.1038/nrg1272

Basha, O., Shpringer, R., Argov, C. M., and Yeger-Lotem, E. (2018). The DifferentialNet database of differential protein–protein interactions in human tissues. *Nucleic Acids Res.* 46 (1), D522–D526. doi:10.1093/nar/gkx981

Bastian, M., Heymann, S., and Jacomy, M. (2009). *Gephi : An open source software for exploring and manipulating networks*, 2.

Baxt, W. G. (1995). Application of artificial neural networks to clinical medicine. *Lancet* 346 (8983), 1135–1138. doi:10.1016/S0140-6736(95)91804-3

Bayes, M., Moivre, A., and Price, M. (1763). An essay towards solving a problem in the doctrine of chances. By the late rev. Mr. Bayes, F. R. S. Communicated by mr. Price, in a letter to john canton, A. M. F. R. S.'. *Philos. Trans.* 53, 370–418.

Becker, M., Lippel, J., Stuhlsatz, A., and Zielke, T. (2020). Robust dimensionality reduction for data visualization with deep neural networks. *Graph. Models* 108, 101060. doi:10.1016/j.gmod.2020

Bello-Orgaz, G., Menéndez, H. D., and Camacho, D. (2012). Adaptive k-means algorithm for overlapped graph clustering. *Int. J. Neural Syst.* 22 (5), 1250018. doi:10.1142/S0129065712500189

Berne, B. J., Weeks, J. D., and Zhou, R. (2009). Dewetting and hydrophobic interaction in physical and biological systems. *Annu. Rev. Phys. Chem.* 60, 85–103. doi:10.1146/annurev.physchem.58.032806.104445

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinforma.* 17 (2), S15. doi:10.1186/s12859-015-0857-9

Berto, S., Perdomo-Sabogal, A., Gerighausen, D., Qin, J., and Nowick, K. (2016). A consensus network of gene regulatory factors in the human frontal lobe. *Front. Genet.* 7, 31. doi:10.3389/fgene.2016.00031

Bhatia, C. (2019). 'Random walk with restart and its applications'. *Medium* 8. Availableat: https://medium.com/@chaitanya_bhatia/random-walk-with-restart-and-its-applications-f53d7c98cb9 (Accessed: April 4, 2022).

Bhowmick, S., and Seah, B.-S. (2015). Clustering and summarizing protein-protein interaction networks: A survey. *IEEE Trans. Knowl. Data Eng.* 28, 638–658. doi:10.1109/TKDE.2015.2492559

Birtles, D., and Lee, J. (2021). Identifying distinct structural features of the SARS-CoV-2 spike protein fusion domain essential for membrane interaction. *Biochemistry* 60 (40), 2978–2986. doi:10.1021/acs.biochem.1c00543

Blassel, L., Tostevin, A., Villabona-Arenas, C. J., Peeters, M., Hue, S., Gascuel, O., et al. (2021). Using machine learning and big data to explore the drug resistance landscape in HIV. *PLoS Comput. Biol.* 17 (8), e1008873. doi:10.1371/journal.pcbi.1008873

Bodein, A., Scott-Boyer, M. P., Perin, O., Le Cao, K. A., and Droit, A. (2021). Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic Acids Res.* 50, e27. doi:10.1093/nar/gkab1200

Borhani, N., Ghaisari, J., Abedi, M., Kamali, M., and Gheisari, Y. (2022). A deep learning approach to predict inter-omics interactions in multi-layer networks. *BMC Bioinforma.* 23 (1), 53. doi:10.1186/s12859-022-04569-2

Bosque, G., Folch-Fortuny, A., Pico, J., Ferrer, A., and Elena, S. F. (2014). Topology analysis and visualization of Potyvirus protein-protein interaction network. *BMC Syst. Biol.* 8, 129. doi:10.1186/s12918-014-0129-8

Bozhilova, L. V., Whitmore, A. V., Wray, J., Reinert, G., and Deane, C. M. (2019). Measuring rank robustness in scored protein interaction networks. *BMC Bioinforma.* 20 (1), 446. doi:10.1186/s12859-019-3036-6

Brandizi, M., Singh, A., Rawlings, C., and Hassani-Pak, K. (2018). Towards FAIRer biological knowledge networks using a hybrid linked data and graph database approach. *J. Integr. Bioinform.* 15 (3). doi:10.1515/jib-2018-0023

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., et al. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* 6 (1), 91–97. doi:10.1038/nmeth.1281

Brohée, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinforma.* 7 (1), 488. doi:10.1186/1471-2105-7-488

Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B., et al. (2009). NAViGaTOR: Network analysis, visualization and graphing toronto. *Bioinformatics* 25 (24), 3327–3329. doi:10.1093/bioinformatics/btp595

Browne, F., Wang, H., and Zheng, H. (2010). From experimental approaches to computational techniques: A review on the prediction of protein-protein interactions. *Adv. Artif. Intell.* 2010, e924529. doi:10.1155/2010/924529

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., et al. (2021). RCSB protein data bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* 49 (1), D437–D451. doi:10.1093/nar/gkaa1038

Cansu Demirel, H., Kaan Arici, M., and Tuncbag, N. (2022). Computational approaches leveraging integrated connections of multi-omic data toward clinical applications. *Mol. Omics* 18 (1), 7–18. doi:10.1039/D1MO00158B

Casadio, R., Martelli, P. L., and Savojardo, C. (2022). Machine learning solutions for predicting protein–protein interactions. *WIREs Comput. Mol. Sci.*, e1618. doi:10.1002/wcms.1618

Castillo-Arnemann, J. J., Solodova, O., Dhillon, B. K., and Hancock, R. E. W. (2021). PaIntDB: Network-based omics integration and visualization using protein–protein interactions in *Pseudomonas aeruginosa*. *Bioinformatics* 37 (22), btab363–4281. doi:10.1093/bioinformatics/btab363

Cervantes-Gracia, K., Chahwan, R., and Husi, H. (2022). Integrative OMICS data-driven procedure using a derivatized meta-analysis approach. *Front. Genet.* 13, 828786. doi:10.3389/fgene.2022.828786

Cha, J., and Lee, I. (2020). Single-cell network biology for resolving cellular heterogeneity in human diseases. *Exp. Mol. Med.* 52 (11), 1798–1808. doi:10.1038/s12276-020-00528-0

Chakraborty, A., Mitra, S., De, D., Pal, A. J., Ghaemi, F., Ahmadian, A., et al. (2021). Determining protein–protein interaction using support vector machine: A review. *IEEE Access* 9, 12473–12490. doi:10.1109/ACCESS.2021.3051006

Charitou, T., Bryan, K., and Lynn, D. J. (2016). Using biological networks to integrate, visualize and analyze genomics data. *Genet. Sel. Evol.* 48 (1), 27. doi:10.1186/s12711-016-0205-1

Charmpi, K., Chokkalingam, M., Johnen, R., and Beyer, A. (2021). Optimizing network propagation for multi-omics data integration. *PLoS Comput. Biol.* 17 (11), e1009161. doi:10.1371/journal.pcbi.1009161

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., et al. (2007). Mint: The molecular INTeraction database. *Nucleic Acids Res.* 35, D572–D574. doi:10.1093/nar/gkl950

Chen, J., Zhang, L., Cheng, K., Jin, B., Lu, X., and Che, C. (2022). Predicting drug-target interaction via self-supervised learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1. doi:10.1109/TCBB.2022.3153963

Chen, S.-J., Liao, D. L., Chen, C. H., Wang, T. Y., and Chen, K. C. (2019). Construction and analysis of protein-protein interaction network of heroin use disorder. *Sci. Rep.* 9 (1), 4980. doi:10.1038/s41598-019-41552-z

Chen, X.-W., and Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 21 (24), 4394–4400. doi:10.1093/bioinformatics/bti721

Chia, J.-M., and Kolatkar, P. R. (2004). Implications for domain fusion protein-protein interactions based on structural information. *BMC Bioinforma.* 5, 161. doi:10.1186/1471-2105-5-161

Chiang, T., Scholtens, D., Sarkar, D., Gentleman, R., and Huber, W. (2007). Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol.* 8 (9), R186. doi:10.1186/gb-2007-8-9-r186

Chin, C.-H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8 (4), S11. doi:10.1186/1752-0509-8-S4-S11

Chisanga, D., Keerthikumar, S., and Mathivanan, S. (2017). Network tools for the analysis of proteomic data. *Methods Mol. Biol.* 1549, 177–197. doi:10.1007/978-1-4939-6740-7_14

Chong, J., Wishart, D. S., and Xia, J. (2019). Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Curr. Protoc. Bioinforma.* 68 (1), e86. doi:10.1002/cpbi.86

Chow, K., Sarkar, A., Elhesha, R., Cinaglia, P., Ay, A., and Kahveci, T. (2021). Anca: Alignment-based network construction algorithm. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18 (2), 512–524. doi:10.1109/TCBB.2019.2923620

Chung, S. S., Pandini, A., Annibale, A., Coolen, A. C. C., Thomas, N. S. B., and Fraternali, F. (2015). Bridging topological and functional information in protein interaction networks by short loops profiling. *Sci. Rep.* 5 (1), 8540. doi:10.1038/srep08540

Correia, F. B., Coelho, E. D., and Oliveira, J. L. (2019). Handling noise in protein interaction networks. *BioMed Res. Int.* 2019, 8984248. doi:10.1155/2019/8984248

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20 (3), 273–297. doi:10.1007/BF00994018

Cowman, T., Coskun, M., Grama, A., and Koyuturk, M. (2020). Integrated querying and version control of context-specific biological networks. *Database.*, 2020, baaa018. doi:10.1093/database/baaa018

Creusier, J., and Biétry, F. (2014). Analyse comparative des méthodes de classifications. *RIMHE Revue Interdiscip. Manag. Homme & Entreprise* 103 (1), 105–123. doi:10.3917/rimhe.010.0105

Croce, G., Gueudre, T., Ruiz Cuevas, M. V., Keidel, V., Figliuzzi, M., Szurmant, H., et al. (2019). A multi-scale coevolutionary approach to predict interactions between protein domains. *PLoS Comput. Biol.* 15 (10), e1006891. doi:10.1371/journal.pcbi.1006891

Crowther, M., Wipat, A., and Goñi-Moreno, Á. (2021). *Network visualisation of synthetic biology designs*. bioRxiv, 2021. doi:10.1101/2021.09.14.460206

Csárdi, G., and Nepusz, T. (2006). The igraph software package for complex network research. Availableat: https://www.semanticscholar.org/paper/The-igraph-software-package-for-complex-network-Cs%C3%A1rdi-Nepusz/1d2744b83519657f5f2610698a8ddd177ced4f5c (Accessed January 29, 2022).

Cuenca, E., Sallaberry, A., Ienco, D., and Poncelet, P. (2022). VERTIGo: A visual platform for querying and exploring large multilayer networks. *IEEE Trans. Vis. Comput. Graph.* 28 (3), 1634–1647. doi:10.1109/TVCG.2021.3067820

Curtis, R. E., Yuen, A., Song, L., Goyal, A., and Xing, E. P. (2011). TVNViewer: An interactive visualization tool for exploring networks that change over time or space. *Bioinforma. Oxf. Engl.* 27 (13), 1880–1881. doi:10.1093/bioinformatics/btr273

Cusick, M. E., Klitgord, N., Vidal, M., and Hill, D. E. (2005). Interactome: Gateway into systems biology. *Hum. Mol. Genet.* 14, R171–R181. doi:10.1093/hmg/ddi335

Dahiya, S., Saini, V., Kumar, P., and Kumar, A. (2019). Protein-Protein interaction network analyses of human WNT proteins involved in neural development. *Bioinformation* 15 (5), 307–314. doi:10.6026/97320630015307

Dallago, C., Goldberg, T., Andrade-Navarro, M. A., Alanis-Lobato, G., and Rost, B. (2020). Visualizing human protein-protein interactions and subcellular localizations on cell images through CellMap. *Curr. Protoc. Bioinforma.* 69 (1), e97. doi:10.1002/cpbi.97

Dandekar, T., Snel, B., HuynenM.and Bork, P. (1998). Conservation of gene order: A fingerprint of proteins that physically interact. *Trends biochem. Sci.* 23 (9), 324–328. doi:10.1016/s0968-0004(98)01274-2

Das, S., and Chakrabarti, S. (2021). Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Sci. Rep.* 11 (1), 1761. doi:10.1038/s41598-020-80900-2

Das, T., Andrieux, G., and Ahmed, M. (2020). Integration of online omics-data resources for cancer research. *Front. Genet.* 11, 578345. doi:10.3389/fgene.2020.578345

Date, S. V. (2007). "Estimating protein function using protein-protein relationships," in *Gene function analysis*. Editor M. F. Ochs (Totowa, NJ: Humana Press), 109–127. doi:10.1007/978-1-59745-547-3_7

De Braekeleer, E., Douet-Guilbert, N., and De Braekeleer, M. (2014). RARA fusion genes in acute promyelocytic leukemia: A review. *Expert Rev. Hematol.* 7 (3), 347–357. doi:10.1586/17474086.2014.903794

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14 (4), 249–261. doi:10.1038/nrg3414

De Las Rivas, J., Alonso-López, D., and Arroyo, M. M. (2018). "Chapter nine - human interactomics: Comparative analysis of different protein interaction resources and construction of a cancer protein–drug bipartite network," in *Advances in protein chemistry and structural biology*. Editor R. Donev (Academic Press (Protein-Protein Interactions in Human Disease, Part B), 263–282. doi:10.1016/bs.apcsb.2017.09.002

De Las Rivas, J., and Fontanillo, C. (2010). Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* 6 (6), e1000807. doi:10.1371/journal.pcbi.1000807

Deng, J.-L., Xu, Y.-H., and Wang, G. (2019). Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. *Front. Genet.* 10, 695. doi:10.3389/fgene.2019.00695

Dezso, Z., Oltvai, Z. N., and Barabási, A.-L. (2003). Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. *Genome Res.* 13 (11), 2450–2454. doi:10.1101/gr.1073603

Di Nanni, N., Bersanelli, M., and Milanesi, L. (2020). Network diffusion promotes the integrative analysis of multiple omics. *Front. Genet.* 11, 106. doi:10.3389/fgene.2020.00106

Dimitrakopoulos, C., Hindupur, S. K., Colombi, M., Liko, D., Ng, C. K. Y., Piscuoglio, S., et al. (2021). Multi-omics data integration reveals novel drug targets in hepatocellular carcinoma. *BMC genomics* 22 (1), 592. doi:10.1186/s12864-021-07876-9

Dimitrieva, S., and Bucher, P. (2012). Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. *Bioinformatics* 28 (18), i395–i401. doi:10.1093/bioinformatics/bts400

Ding, Z., and Kihara, D. (2018). Computational methods for predicting protein-protein interactions using various protein features. *Curr. Protoc. Protein Sci.* 93 (1), e62. doi:10.1002/cpps.62

Dohrmann, J., Puchin, J., and Singh, R. (2015). Global multiple protein-protein interaction network alignment by combining pairwise network alignments. *BMC Bioinforma.* 16 (13), S11. doi:10.1186/1471-2105-16-S13-S11

Dongare, A. D., Kharde, R. R., and Kachare, A. D. (2012). *Introd. Artif. Neural Netw.* 2 (1), 6.

Droit, A., Poirier, G. G., and Hunter, J. M. (2005). Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function. *J. Mol. Endocrinol.* 34 (2), 263–280. doi:10.1677/jme.1.01693

Du, Y., Cai, M., Xing, X., Ji, J., Yang, E., and Wu, J. (2021). Pina 3.0: Mining cancer interactome. *Nucleic Acids Res.* 49 (D1), D1351–D1357. doi:10.1093/nar/gkaa1075

Du, Z.-P., Wu, B. L., Wang, S. H., Shen, J. H., Lin, X. H., Zheng, C. P., et al. (2014). Shortest path analyses in the protein-protein interaction network of NGAL (neutrophil gelatinase-associated lipocalin) overexpression in esophageal squamous cell carcinoma. *Asian pac. J. Cancer Prev.* 15 (16), 6899–6904. doi:10.7314/apjcp.2014.15.16.6899

Dugourd, C., Sciacovelli, M., Gjerga, E., Gabor, A., and Emdal, K. B. (2021). Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol. Syst. Biol.* 17 (1), e9730. doi:10.15252/msb.20209730

Dunham, B., and Ganapathiraju, M. K. (2021). Benchmark evaluation of protein–protein interaction prediction algorithms. *Molecules* 27 (1), 41. doi:10.3390/molecules27010041

Dünkler, A., Rösler, R., and Kestler, H. A. (2015). "Spliff: A single-cell method to map protein-protein interactions in time and space," in *Single cell protein analysis: Methods and protocols*. Editors A. K. Singh and A. Chandrasekaran (New York, NY: Springer), 151–168. doi:10.1007/978-1-4939-2987-0_11

Dupré, X. (2022). Random walk with restart (système de recommandations) — Papierstat. Availableat: http://www.xavierdupre.fr/app/papierstat/helpsphinx/notebooks/tinygraph_rwr.html (Accessed: April 4, 2022).

Dursun, C., Kwitek, A., and Bozdag, S. (2021). PhenoGeneRanker: Gene and phenotype prioritization using multiplex heterogeneous networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1. doi:10.1109/TCBB.2021.3098278

Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020). CellPhoneDB: Inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* 15 (4), 1484–1506. doi:10.1038/s41596-020-0292-x

Eicher, T., Kinnebrew, G., Patt, A., Spencer, K., Ying, K., Ma, Q., et al. (2020). Metabolomics and multi-omics integration: A survey of computational methods and resources. *Metabolites* 10 (5), E202. doi:10.3390/metabo10050202

Eisenberg, D., Marcotte, E. M., XenarIos, I., and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* 405 (6788), 823–826. doi:10.1038/35015694

El Naqa, I., and Murphy, M. J. (2015). "What is machine learning?," in *Machine learning in radiation oncology: Theory and applications*. Editors I. El Naqa, R. Li, and M. J. Murphy (Cham: Springer International Publishing), 3–11. doi:10.1007/978-3-319-18305-3_1

Elangovan, A., Davis, M., and Verspoor, K. (2020). *Assigning function to protein-protein interactions: A weakly supervised BioBERT based approach using PubMed abstracts*, 6.

Ellson, J., Gansner, E., and Koutsofios, L. (2001). "Graphviz — Open source graph drawing tools," in *Lecture notes in computer science* (Springer-Verlag), 483–484. doi:10.1007/3-540-45848-4_57

Esch, R., and Merkl, R. (2020). Conserved genomic neighborhood is a strong but no perfect indicator for a direct interaction of microbial gene products. *BMC Bioinforma.* 21, 5. doi:10.1186/s12859-019-3200-z

Everson, J., Richards, M. R., and Buntin, M. B. (2019). Horizontal and vertical integration's role in meaningful use attestation over time. *Health Serv. Res.* 54 (5), 1075–1083. doi:10.1111/1475-6773.13193

Fan, Z., Zhou, Y., and Ressom, H. W. (2020). Mota: Network-based multi-omic data integration for biomarker discovery. *Metabolites* 10 (4), 144. doi:10.3390/metabo10040144

Farahani, F. V., Karwowski, W., and Lighthall, N. R. (2019). Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Front. Neurosci.* 13, 585. doi:10.3389/fnins.2019.00585

Farahmand, S., Riley, T., and Zarringhalam, K. (2020). ModEx: A text mining system for extracting mode of regulation of transcription factor-gene regulatory interaction. *J. Biomed. Inf.* 102, 103353. doi:10.1016/j.jbi.2019.103353

Fekete, J.-D. (2009). "Visualizing networks using adjacency matrices: Progresses and challenges," in 2009 11th IEEE International Conference on Computer-Aided Design and Computer Graphics, 32. doi:10.1109/CADCG.2009.5246808

Fionda, V. (2019). "Networks in biology," in *Encyclopedia of bioinformatics and computational biology*. Editor S. Ranganathan (Oxford: Academic Press), 915–921. doi:10.1016/B978-0-12-809633-8.20420-2

Fionda, V., Palopoli, L., and Panni, S. (2009). "Extracting similar subgraphs across PPI networks," in 2009 24th International Symposium on Computer and Information Sciences, 183–188. doi:10.1109/ISCIS.2009.5291845

Fiorentino, G., Visintainer, R., Domenici, E., Lauria, M., and Marchetti, L. (2021). Mousse: Multi-omics using subject-specific SignaturEs. *Cancers* 13 (14), 3423. doi:10.3390/cancers13143423

Flórez, A. F., Park, D., Bhak, J., Kim, B. C., Kuchinsky, A., Morris, J. H., et al. (2010). Protein network prediction and topological analysis in Leishmania major as a tool for drug target selection. *BMC Bioinforma.* 11, 484. doi:10.1186/1471-2105-11-484

Franzese, N., Groce, A., Murali, T. M., and Ritz, A. (2019). Hypergraph-based connectivity measures for signaling pathway topologies. *PLoS Comput. Biol.* 15 (10), e1007384. doi:10.1371/journal.pcbi.1007384

Freilich, R., Arhar, T., Abrams, J. L., and Gestwicki, J. E. (2018). Protein-protein interactions in the molecular chaperone network. *Acc. Chem. Res.* 51 (4), 940–949. doi:10.1021/acs.accounts.8b00036

Gebreyesus, S. T., Siyal, A. A., Kitata, R. B., Chen, E. S. W., Enkhbayar, B., Angata, T., et al. (2022). Streamlined single-cell proteomics by an integrated microfluidic chip and data-independent acquisition mass spectrometry. *Nat. Commun.* 13 (1), 37. doi:10.1038/s41467-021-27778-4

Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., et al. (2010). Visualization of omics data for systems biology. *Nat. Methods* 7, S56–S68. doi:10.1038/nmeth.1436

Geng, H., Chen, X., and Wang, C. (2021). Systematic elucidation of the pharmacological mechanisms of Rhynchophylline for treating epilepsy via network pharmacology. *BMC Complement. Med. Ther.* 21, 9. doi:10.1186/s12906-020-03178-x

Geng, H., Lu, T., and Lin, X. (2015). Prediction of protein-protein interaction sites based on naive bayes classifier. *Biochem. Res. Int.* 2015, 978193. doi:10.1155/2015/978193

Gerasch, A., Faber, D., Kuntzer, J., Niermann, P., Kohlbacher, O., Lenhof, H. P., et al. (2014). BiNA: A visual analytics tool for biological network data. *PLOS ONE* 9 (2), e87397. doi:10.1371/journal.pone.0087397

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., et al. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 50 (1), D687–D692. doi:10.1093/nar/gkab1028

Gillis, J., Ballouz, S., and Pavlidis, P. (2014). Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *J. Proteomics* 100, 44–54. doi:10.1016/j.jprot.2014.01.020

Giovanni, T., and Renaud, D. (2015). Siren investigate. Availableat: https://siren.io/.

Gligorijević, V., and Pržulj, N. (2015). Methods for biological data integration: Perspectives and challenges. *J. R. Soc. Interface* 12 (112), 20150571. doi:10.1098/rsif.2015.0571

Goel, R., Harsha, H. C., Pandey, A., and Prasad, T. S. K. (2012). Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Mol. Biosyst.* 8 (2), 453–463. doi:10.1039/c1mb05340j

Goh, C.-S., and Cohen, F. E. (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. *J. Mol. Biol.* 324 (1), 177–192. doi:10.1016/s0022-2836(02)01038-0

Gong, F., Ma, Y., Gong, W., Li, X., Li, C., and Yuan, X. (2018). Neo4j graph database realizes efficient storage performance of oilfield ontology. *PloS One* 13 (11), e0207595. doi:10.1371/journal.pone.0207595

González-Sánchez, J. C., Ibrahim, M. F. R., Leist, I. C., Weise, K. R., and Russell, R. B. (2021). Mechnetor: A web server for exploring protein mechanism and the functional context of genetic variants. *Nucleic Acids Res.* 49 (W1), W366–W374. doi:10.1093/nar/gkab399

Green, A. G., Elhabashy, H., Brock, K. P., Maddamsetti, R., Kohlbacher, O., and Marks, D. S. (2021). Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* 12 (1), 1396. doi:10.1038/s41467-021-21636-z

Guney, E., Menche, J., Vidal, M., and Barabasi, A. L. (2016). Network-based *in silico* drug efficacy screening. *Nat. Commun.* 7 (1), 10331. doi:10.1038/ncomms10331

Guo, Y., Wu, J., and Ma, H. (2022). "Self-supervised pre-training for protein embeddings using tertiary structures,", Proceedings of the AAAI Conference on Artificial Intelligence, 9. doi:10.1609/aaai.v36i6.20636

Gursoy, A., Keskin, O., and Nussinov, R. (2008). Topological properties of protein interaction networks from a structural perspective. *Biochem. Soc. Trans.* 36 (6), 1398–1403. doi:10.1042/BST0361398

Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S., and Ralser, M. (2017). Designing and interpreting "multi-omic" experiments that may change our understanding of biology. *Curr. Opin. Syst. Biol.* 6, 37–45. doi:10.1016/j.coisb.2017.08.009

Hagberg, A., Swart, P., and S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx. LA-UR-08-05495; LA-UR-08-5495*. Los Alamos National Lab. LANL, Los Alamos, NM United States. Availableat: https://www.osti.gov/biblio/960616-exploring-network-structure-dynamics-function-using-networkx (Accessed March 14, 2022).

Hakes, L., Pinney, J. W., Robertson, D. L., and Lovell, S. C. (2008). Protein-protein interaction networks and biology—what's the connection? *Nat. Biotechnol.* 26 (1), 69–72. doi:10.1038/nbt0108-69

Halder, A. K., Denkiewicz, M., Sengupta, K., Basu, S., and Plewczynski, D. (2020). Aggregated network centrality shows non-random structure of genomic and proteomic networks. *Methods (San Diego, Calif.)* 181–182, 5–14. doi:10.1016/j.ymeth.2019.11.006

Hammoud, Z., and Kramer, F. (2020). Multilayer networks: Aspects, implementations, and application in biomedicine. *Big Data Anal.* 5 (1), 2. doi:10.1186/s41044-020-00046-0

Han, J.-D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93. doi:10.1038/nature02555

Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.* 23 (7), 839–844. doi:10.1038/nbt1116

Hao, T., Zhao, L., Wu, D., Wang, B., Feng, X., Wang, E., et al. (2019). The protein-protein interaction network of Litopenaeus vannamei haemocytes. *Front. Physiol.* 10, 156. doi:10.3389/fphys.2019.00156

Hasan, Md.R., Paul, B. K., Ahmed, K., and Bhuyian, T. (2020). Design protein-protein interaction network and protein-drug interaction network for common cancer diseases: A bioinformatics approach. *Inf. Med. Unlocked* 18, 100311. doi:10.1016/j.imu.2020.100311

Hashemifar, S., Neyshabur, B., Khan, A. A., and Xu, J. (2018). Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* 34 (17), i802i802–i810. doi:10.1093/bioinformatics/bty573

Hawe, J. S., Theis, F. J., and Heinig, M. (2019). Inferring interaction networks from multi-omics data. *Front. Genet.* 10, 535. doi:10.3389/fgene.2019.00535

Hayashi, T., Matsuzaki, Y., Yanagisawa, K., Ohue, M., and Akiyama, Y. (2018). MEGADOCK-web: An integrated database of high-throughput structure-based protein-protein interaction predictions. *BMC Bioinforma.* 19 (4), 62. doi:10.1186/s12859-018-2073-x

He, M., Wang, Y., and Li, W. (2009). PPI finder: A mining tool for human protein-protein interactions. *PLOS ONE* 4 (2), e4554. doi:10.1371/journal.pone.0004554

He, T., and Chan, K. C. C. (2018). Evolutionary graph clustering for protein complex identification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15 (3), 892–904. doi:10.1109/TCBB.2016.2642107

Heberle, H., Carazzolle, M. F., Telles, G. P., Meirelles, G. V., and Minghim, R. (2017). CellNetVis: A web tool for visualization of biological networks using force-directed layout constrained by cellular components. *BMC Bioinforma.* 18 (10), 395. doi:10.1186/s12859-017-1787-5

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., et al. (2004b). The HUPO PSI's molecular interaction format—A community standard for the representation of protein interaction data. *Nat. Biotechnol.* 22 (2), 177–183. doi:10.1038/nbt926

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004a) 'IntAct: An open source molecular interaction database', *Nucleic Acids Res.*, 32pp. D452, D455. doi:10.1093/nar/gkh052

Hibbs, M., Wallace, G., and Li, K., (2007) 'Viewing the larger context of genomic data through horizontal integration', in 2007 11th International Conference Information Visualization (IV '07), 04-06 July 2007, Zurich, Switzerland, IEEEpp. 326–334. doi:10.1109/IV.2007.120

Hu, J., Zhou, L., Li, B., Zhang, X., and Chen, N. (2021). Improve hot region prediction by analyzing different machine learning algorithms. *BMC Bioinforma.* 22 (3), 522. doi:10.1186/s12859-021-04420-0

Hu, X., Feng, C., Zhou, Y., Harrison, A., and Chen, M. (2022). DeepTrio: A ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks. *Bioinformatics* 38 (3), 694–702. doi:10.1093/bioinformatics/btab737

Hu, Y., Vinayagam, A., Nand, A., Comjean, A., Chung, V., Hao, T., et al. (2018). Molecular interaction search tool (MIST): An integrated resource for mining gene and protein interaction data. *Nucleic Acids Res.* 46 (1), D567–D574. doi:10.1093/nar/gkx1116

Huang, K., and Zitnik, M. (2021). *Graph meta learning via local subgraphs*. arXiv: 2006.07889. Availableat: http://arxiv.org/abs/2006.07889 (Accessed March 29, 2021).

Huang, X.-T., Jia, S., Gao, L., and Wu, J. (2022). Reconstruction of human protein-coding gene functional association network based on machine learning. *Brief. Bioinform.* 23, bbab552. doi:10.1093/bib/bbab552

Hütter, C. V. R., Sin, C., Muller, F., and Menche, J. (2022). Network cartographs for interpretable visualizations. *Nat. Comput. Sci.* 2 (2), 84–89. doi:10.1038/s43588-022-00199-z

Iranzo, J., Krupovic, M., and Koonin, E. V. (2016). The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *Mbio* 7 (4), e0097816. doi:10.1128/mBio.00978-16

Islam, M. F., Hoque, M. M., Banik, R. S., Roy, S., Sumi, S. S., Hassan, F. M. N., et al. (2013). Comparative analysis of differential network modularity in tissue specific normal and cancer protein interaction networks. *J. Clin. Bioinforma.* 3 (1), 19. doi:10.1186/2043-9113-3-19

Jamasb, A. R., Day, B., Cangea, C., Lio, P., and Blundell, T. L. (2021). Deep learning for protein-protein interaction site prediction. *Methods Mol. Biol.* 2361, 263–288. doi:10.1007/978-1-0716-1641-3_16

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., et al. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302 (5644), 449–453. doi:10.1126/science.1087361

Jeanquartier, F., Jean-Quartier, C., and Holzinger, A. (2015). Integrated web visualizations for protein-protein interaction databases. *BMC Bioinforma.* 16 (1), 195. doi:10.1186/s12859-015-0615-z

Jha, K., Saha, S., and Singh, H. (2022). Prediction of protein–protein interaction using graph neural networks. *Sci. Rep.* 12 (1), 8360. doi:10.1038/s41598-022-12201-9

Ji, A. L., Rubin, A. J., Thrane, K., Jiang, S., Reynolds, D. L., Meyers, R. M., et al. (2020). Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* 182 (2), 497–514. doi:10.1016/j.cell.2020.05.039

Jia, D., Li, S., Li, D., Xue, H., Yang, D., and Liu, Y. (2018). Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. *Aging* 10 (4), 592–605. doi:10.18632/aging.101415

Jia, Y., Zou, J., Wang, Y., Zhang, X., Shi, Y., Liang, Y., et al. (2021). Action mechanism of Roman chamomile in the treatment of anxiety disorder based on network pharmacology. *J. Food Biochem.* 45 (1), e13547. doi:10.1111/jfbc.13547

Jin, N., Wu, D., and Gong, Y. (2014). Integration strategy is a key step in network-based analysis and dramatically affects network topological properties and inferring outcomes. *BioMed Res. Int.* 2014, e296349. doi:10.1155/2014/296349

Johansson-Åkhe, I., Mirabello, C., and Wallner, B. (2019). Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Sci. Rep.* 9 (1), 4267. doi:10.1038/s41598-019-38498-7

Johnson, K. L., Qi, Z., Yan, Z., Wen, X., Nguyen, T. C., Zaleta-Rivera, K., et al. (2021). Revealing protein-protein interactions at the transcriptome scale by sequencing. *Mol. Cell* 81 (19), 4091–4103. doi:10.1016/j.molcel.2021.07.006

Jonathan, J., Sanga, C., Mwita, M., and Mgode, G. (2021). Visual analytics of tuberculosis detection rat performance. *Online J. Public Health Inf.* 13 (2), e12. doi:10.5210/ojphi.v13i2.11465

Jordán, F., Nguyen, T.-P., and Liu, W. (2012). Studying protein–protein interaction networks: A systems view on diseases. *Brief. Funct. Genomics* 11 (6), 497–504. doi:10.1093/bfgp/els035

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Jupe, S., Ray, K., Roca, C. D., Varusai, T., Shamovsky, V., Stein, L., et al. (2018). Interleukins and their signaling pathways in the Reactome biological pathway database. *J. Allergy Clin. Immunol.* 141 (4), 1411–1416. doi:10.1016/j.jaci.2017.12.992

Kamburov, A., Grossmann, A., Herwig, R., and Stelzl, U. (2012). Cluster-based assessment of protein-protein interaction confidence. *BMC Bioinforma.* 13 (1), 262. doi:10.1186/1471-2105-13-262

Kamisetty, H., Ramanathan, A., Bailey-Kellogg, C., and Langmead, C. J. (2011). Accounting for conformational entropy in predicting binding free energies of protein-protein interactions: Entropy and Protein-Protein Interactions. *Proteins* 79 (2), 444–462. doi:10.1002/prot.22894

Kanai, M., Maeda, Y., and Okada, Y. (2018). Grimon: Graphical interface to visualize multi-omics networks. *Bioinformatics* 34 (22), 3934–3936. doi:10.1093/bioinformatics/bty488

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44 (1), D457–D462. doi:10.1093/nar/gkv1070

Kapadia, P., Khare, S., and Priyadarshini, P. (2019). "Predicting protein-protein interaction in multi-layer blood cell PPI networks," in *Advanced informatics for computing research*. Editor A. K. Luhach (Singapore: Springer), 240–251. doi:10.1007/978-981-15-0111-1_22

Karatzas, E., Baltoumas, F. A., Panayiotou, N. A., Schneider, R., and Pavlopoulos, G. A. (2021). Arena3Dweb: Interactive 3D visualization of multilayered networks. *Nucleic Acids Res.* 49 (W1), W36–W45. doi:10.1093/nar/gkab278

Kazemi, E., Hassani, H., Grossglauser, M., and Pezeshgi Modarres, H. (2016). Proper: Global protein interaction network alignment through percolation matching. *BMC Bioinforma.* 17 (1), 527. doi:10.1186/s12859-016-1395-9

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., et al. (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* 5, 44. doi:10.1186/1741-7007-5-44

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database--2009 update. *Nucleic Acids Res.* 37, D767–D772. doi:10.1093/nar/gkn892

Khashan, R., Tropsha, A., and Zheng, W. (2022). Data mining meets machine learning: A novel ANN-based multi-body interaction docking scoring function (MBI-score) based on utilizing frequent geometric and chemical patterns of interfacial atoms in native protein-ligand complexes. *Mol. Inf.*, e2100248. doi:10.1002/minf.202100248

Kim, T. R., Jeong, H.-H., and Sohn, K.-A. (2019). Topological integration of RPPA proteomic data with multi-omics data for survival prediction in breast cancer via pathway activity inference. *BMC Med. Genomics* 12 (5), 94. doi:10.1186/s12920-019-0511-x

Klimm, F., Toledo, E. M., Monfeuga, T., Zhang, F., Deane, C. M., and Reinert, G. (2020). Functional module detection through integration of single-cell RNA sequencing data with protein–protein interaction networks. *BMC Genomics* 21 (1), 756. doi:10.1186/s12864-020-07144-2

Koh, G. C. K. W., Porras, P., Aranda, B., Hermjakob, H., and Orchard, S. E. (2012). Analyzing protein–protein interaction networks. *J. Proteome Res.* 11 (4), 2014–2031. doi:10.1021/pr201211w

Kotlyar, M., Pastrello, C., Malik, Z., and Jurisica, I. (2019). IID 2018 update: Context-specific physical protein-protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* 47 (1), D581–D589. doi:10.1093/nar/gky1037

Kotlyar, M., Rossos, A. E. M., and Jurisica, I. (2017). Prediction of protein-protein interactions. *Curr. Protoc. Bioinforma.* 60, 821–8. doi:10.1002/cpbi.38

Koutrouli, M., Karatzas, E., and Espino, D. (2020). A guide to conquer the biological network era using graph theory. *Front. Bioeng. Biotechnol.* 8, 31. doi:10.3389/fbioe.2020.0003

Krause, A., Stoye, J., and Vingron, M. (2005). Large scale hierarchical clustering of protein sequences. *BMC Bioinforma.* 6 (1), 15. doi:10.1186/1471-2105-6-15

Krogh, A. (2008). What are artificial neural networks? *Nat. Biotechnol.* 26 (2), 195–197. doi:10.1038/nbt1386

Kshitish, A., Salaingambi, S., Raksha, H. N., Deepika, T. S., and Preeti, G. (2013). Startbioinfo contributors. Availableat: https://startbioinfo.org/contributors.html (Accessed February 23, 2022).

Kusuma, W., F Ahmad, H., and Suryono, M. (2019). Clustering of protein-protein interactions (PPI) and gene ontology molecular function using Markov clustering and fuzzy K partite algorithm. *IOP Conf. Ser. Earth Environ. Sci.* 299 (1), 012034. doi:10.1088/1755-1315/299/1/012034

Kuzmanov, U., and Emili, A. (2013). Protein-protein interaction networks: Probing disease mechanisms using model systems. *Genome Med.* 5 (4), 37. doi:10.1186/gm441

Laniau, J. (2017). *Structure de réseaux biologiques : Rôle des nœuds internes vis-à-vis de la production de composés*. Theses. Inria Rennes - Bretagne Atlantique. Availableat: https://hal.archives-ouvertes.fr/tel-01656474 (Accessed: January 26, 2022).

Latysheva, N. S., Oates, M. E., Maddox, L., Flock, T., Gough, J., Buljan, M., et al. (2016). Molecular principles of gene fusion mediated rewiring of protein interaction networks in cancer. *Mol. Cell* 63 (4), 579–592. doi:10.1016/j.molcel.2016.07.008

Leblanc, H. J., Zhang, H., Koutsoukos, X., and Sundaram, S. (2013). Resilient asymptotic consensus in robust networks. *IEEE J. Sel. Areas Commun.* 31 (4), 766–781. doi:10.1109/jsac.2013.130413

Lee, I., and Nam, H. (2018). Identification of drug-target interaction by a random walk with restart method on an interactome network. *BMC Bioinforma.* 19 (8), 208. doi:10.1186/s12859-018-2199-x

Lee, J. J., Bernard, V., Semaan, A., Monberg, M. E., Huang, J., Stephens, B. M., et al. (2021). Elucidation of tumor-stromal heterogeneity and the ligand-receptor interactome by single-cell transcriptomics in real-world pancreatic cancer biopsies. *Clin. Cancer Res.* 27 (21), 5912–5921. doi:10.1158/1078-0432.CCR-20-3925

Lee, M. S., and Oh, S. (2014). Alternating decision tree algorithm for assessing protein interaction reliability. *Vietnam J. Comput. Sci.* 1 (3), 169–178. doi:10.1007/s40595-014-0018-5

Lee, T., and Yoon, Y. (2018). Drug repositioning using drug-disease vectors based on an integrated network. *BMC Bioinforma.* 19 (1), 446. doi:10.1186/s12859-018-2490-x

Lei, X., Wang, S., and Wu, F. (2019a). Identification of essential proteins based on improved HITS algorithm. *Genes* 10 (2), 177. doi:10.3390/genes10020177

Lei, X., Yang, X., and Fujita, H. (2019b). Random walk based method to identify essential proteins by integrating network topology and biological characteristics. *Knowledge-Based Syst.* 167, 53–67. doi:10.1016/j.knosys.2019.01.012

Lercher, M. J., and Pál, C. (2008). Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* 25 (3), 559–567. doi:10.1093/molbev/msm283

Li, F., Zhu, F., Ling, X., and Liu, Q. (2020c). Protein interaction network reconstruction through ensemble deep learning with attention mechanism. *Front. Bioeng. Biotechnol.* 8, 390. doi:10.3389/fbioe.2020.00390

Li, H., Gong, X. J., Yu, H., and Zhou, C. (2018a). Deep neural network based predictions of protein interactions using primary sequences. *Molecules* 23 (8), 1923. doi:10.3390/molecules23081923

Li, M., Jiang, Y., and Ryu, K. H. (2022b). InfersentPPI: Prediction of protein-protein interaction using protein sentence embedding with gene ontology information. *Front. Genet.* 13. doi:10.3389/fgene.2022.82754

Li, M., Yang, J., Wu, F. X., Pan, Y., and Wang, J. (2018b). DyNetViewer: A cytoscape app for dynamic network construction, analysis and visualization. *Bioinformatics* 34 (9), 1597–1599. doi:10.1093/bioinformatics/btx821

Li, X., Ma, J., Han, M., and He, F. (2022a). MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype Analysis. *Front. Genet.* 13. doi:10.3389/fgene.2022.806842

Li, X., Tran, K. M., Aziz, K. E., Sorokin, A. V., Chen, J., and Wang, W. (2016). Defining the protein-protein interaction network of the human protein tyrosine phosphatase family. *Mol. Cell. Proteomics* 15 (9), 3030–3044. doi:10.1074/mcp.M116.060277

Li, Y., Liang, Y., Ma, T., and Yang, Q. (2020b). Identification of DGUOK-AS1 as a prognostic factor in breast cancer by bioinformatics analysis. *Front. Oncol.* 10, 1092. doi:10.3389/fonc.2020.01092

Li, Y., Qian, B., Zhang, X., and Liu, H. (2020a). Graph neural network-based diagnosis prediction. *Big Data* 8 (5), 379–390. doi:10.1089/big.2020.0070

Li, Y., and Ilie, L. (2017). Sprint: Ultrafast protein-protein interaction prediction of the entire human interactome. *BMC Bioinforma.* 18 (1), 485. doi:10.1186/s12859-017-1871-x

Li, Y., Wang, Z., Li, L. P., You, Z. H., Huang, W. Z., Zhan, X. K., et al. (2021). Robust and accurate prediction of protein-protein interactions by exploiting evolutionary information. *Sci. Rep.* 11 (1), 16910. doi:10.1038/s41598-021-96265-z

Liang, L., Chen, V., Zhu, K., Fan, X., Lu, X., and Lu, S. (2019). Integrating data and knowledge to identify functional modules of genes: A multilayer approach. *BMC Bioinforma.* 20, 225. doi:10.1186/s12859-019-2800-y

Lin, H.-H., Zhang, Q. R., Kong, X., Zhang, L., Zhang, Y., Tang, Y., et al. (2021). Machine learning prediction of antiviral-HPV protein interactions for anti-HPV pharmacotherapy. *Sci. Rep.* 11 (1), 24367. doi:10.1038/s41598-021-03000-9

Lin, J.-S., and Lai, E.-M. (2017). Protein-protein interactions: Co-immunoprecipitation. *Methods Mol. Biol.* 1615, 211–219. doi:10.1007/978-1-4939-7033-9_17

Liu, L., Zhu, X., Ma, Y., Piao, H., Yang, Y., Hao, X., et al. (2020a). Combining sequence and network information to enhance protein–protein interaction prediction. *BMC Bioinforma.* 21 (16), 537. doi:10.1186/s12859-020-03896-6

Liu, Q., Chen, P., Wang, B., Zhang, J., and Li, J. (2018). Hot spot prediction in protein-protein interactions by an ensemble system. *BMC Syst. Biol.* 12 (9), 132. doi:10.1186/s12918-018-0665-8

Liu, T.-H., Chen, W. H., Chen, X. D., Liang, Q. E., Tao, W. C., Jin, Z., et al. (2020b). Network pharmacology identifies the mechanisms of action of TaohongSiwu decoction against essential hypertension. *Med. Sci. Monit.* 26, e920682. doi:10.12659/MSM.920682

Liu, X., Chang, C., Han, M., Yin, R., Zhan, Y., Li, C., et al. (2019). PPIExp: A web-based platform for integration and visualization of protein-protein interaction data and spatiotemporal proteomics data. *J. Proteome Res.* 18 (2), 633–641. doi:10.1021/acs.jproteome.8b00713

Liu, Y., Zhu, Y., and He, C. (2021). BENviewer: A gene interaction network visualization server based on graph embedding model, Database. 2021, baab033. doi:10.1093/database/baab033

Lotia, S., Montojo, J., Dong, Y., Bader, G. D., and Pico, A. R. (2013). Cytoscape app store. *Bioinforma. Oxf. Engl.* 29 (10), 1350–1351. doi:10.1093/bioinformatics/btt138

Louche, A., Salcedo, S. P., and Bigot, S. (2017). Protein-protein interactions: Pull-down assays. *Methods Mol. Biol.* 1615, 247–255. doi:10.1007/978-1-4939-7033-9_20

Lu, C., Hu, X., Wang, G., Leach, L. J., Yang, S., Kearsey, M. J., et al. (2010). Why do essential proteins tend to be clustered in the yeast interactome network? *Mol. Biosyst.* 6 (5), 871–877. doi:10.1039/b921069e

Lu, X., Liu, F., Miao, Q., Liu, P., Gao, Y., and He, K. (2021). A novel method to identify gene interaction patterns. *BMC Genomics* 22 (1), 436. doi:10.1186/s12864-021-07628-9

Luck, K., Kim, D. K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., et al. (2020). A reference map of the human binary protein interactome. *Nature* 580 (7803), 402–408. doi:10.1038/s41586-020-2188-x

Lv, Y., Huang, S., and Zhang, T. (2021). Application of multilayer network models in bioinformatics. *Front. Genet.* 12. doi:10.3389/fgene.2021.664860

Lynn, C. W., and Bassett, D. S. (2021). Quantifying the compressibility of complex networks. *Proc. Natl. Acad. Sci. U. S. A.* 118 (32), e2023473118. doi:10.1073/pnas.2023473118

Lysenko, A., Roznovat, I. A., Saqi, M., Mazein, A., Rawlings, C. J., and Auffray, C. (2016). Representing and querying disease networks using graph databases. *BioData Min.* 9 (1), 23. doi:10.1186/s13040-016-0102-8

Ma, C.-Y., and Liao, C.-S. (2020). A review of protein–protein interaction network alignment: From pathway comparison to global alignment. *Comput. Struct. Biotechnol. J.* 18, 2647–2656. doi:10.1016/j.csbj.2020.09.011

Ma, F., Zhang, S., Song, L., Wang, B., Wei, L., and Zhang, F. (2021a). Applications and analytical tools of cell communication based on ligand-receptor interactions at single cell level. *Cell Biosci.* 11 (1), 121. doi:10.1186/s13578-021-00635-z

Ma, H., He, Z., Chen, J., Zhang, X., and Song, P. (2021b). Identifying of biomarkers associated with gastric cancer based on 11 topological analysis methods of CytoHubba. *Sci. Rep.* 11 (1), 1331. doi:10.1038/s41598-020-79235-9

Ma, T., and Zhang, A. (2019). Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC Genomics* 20 (11), 944. doi:10.1186/s12864-019-6285-x

Ma, W., Cao, Y., and Bao, W. (2020). ACT-SVM: Prediction of protein-protein interactions based on support vector basis model. *Sci. Program.* 2020, e8866557. doi:10.1155/2020/8866557

MacDonald, P. N. (1998). "A two-hybrid protein interaction system to identify factors that interact with retinoid and vitamin D receptors," in *Retinoid protocols*. Editor C. P. F. Redfern (Totowa, NJ: Humana Press), 359–375. doi:10.1385/0-89603-438-0:359

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21 (16), 3448–3449. doi:10.1093/bioinformatics/bti551

Mahdessian, D., Cesnik, A. J., Gnann, C., Danielsson, F., Stenstrom, L., Arif, M., et al. (2021). Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. *Nature* 590 (7847), 649–654. doi:10.1038/s41586-021-03232-9

Mahdipour, E., and Ghasemzadeh, M. (2021). The protein-protein interaction network alignment using recurrent neural network. *Med. Biol. Eng. Comput.* 59 (11), 2263–2286. doi:10.1007/s11517-021-02428-5

Maimon, O., and Rokach, L. (2006). *Data mining and knowledge discovery handbook*. Springer Science & Business Media.

Malek, M., Zorzan, S., and Ghoniem, M. (2020). A methodology for multilayer networks analysis in the context of open and private data: Biological application. *Appl. Netw. Sci.* 5 (1), 41–28. doi:10.1007/s41109-020-00277-z

Malik, S., Sharma, D., and Khatri, S. K. (2017). Reconstructing phylogenetic tree using a protein-protein interaction technique. *IET Nanobiotechnol.* 11 (8), 1005–1016. doi:10.1049/iet-nbt.2016.0177

Malod-Dognin, N., Ban, K., and Pržulj, N. (2017). Unified alignment of protein-protein interaction networks. *Sci. Rep.* 7 (1), 953. doi:10.1038/s41598-017-01085-9

Malouche, D. (2013). *Méthodes de classifications*, 32.

Marai, G. E., Pinaud, B., Buhler, K., Lex, A., and Morris, J. H. (2019). Ten simple rules to create biological network figures for communication. *PLoS Comput. Biol.* 15 (9), e1007244. doi:10.1371/journal.pcbi.1007244

Marcotte, E. M., Pellegrini, M., and Ng, H. L. (1999). 'Detecting protein function and protein-protein interactions from genome sequences'. *Science* 285. doi:10.1126/science.285.5428.751

Marín-Llaó, J., Mubeen, S., Perera-Lluna, A., Hofmann-Apitius, M., Picart-Armada, S., and Domingo-Fernandez, D. (2021). MultiPaths: A Python framework for analyzing multi-layer biological networks using diffusion algorithms. *Bioinformatics* 37 (1), 137–139. doi:10.1093/bioinformatics/btaa1069

Martin, S., Roe, D., and Faulon, J.-L. (2005). Predicting protein–protein interactions using signature products. *Bioinformatics* 21 (2), 218–226. doi:10.1093/bioinformatics/bth483

McGee, F., Ghoniem, M., Melancon, G., Otjacques, B., and Pinaud, B. (2019). The state of the art in multilayer network visualization. *Comput. Graph. Forum* 38 (6), 125–149. doi:10.1111/cgf.13610

Mikkelsen, T. S., Galagan, J. E., and Mesirov, J. P. (2005). Improving genome annotations using phylogenetic profile anomaly detection. *Bioinformatics* 21 (4), 464–470. doi:10.1093/bioinformatics/bti027

Mishra, B., Kumar, N., and Mukhtar, M. S. (2021). Network biology to uncover functional and structural properties of the plant immune system. *Curr. Opin. Plant Biol.* 62, 102057. doi:10.1016/j.pbi.2021.102057

Mlecnik, B., Galon, J., and Bindea, G. (2018). Comprehensive functional analysis of large lists of genes and proteins. *J. Proteomics* 171, 2–10. doi:10.1016/j.jprot.2017.03.016

Mooney, M. A., Nigg, J. T., McWeeney, S. K., and Wilmot, B. (2014). Functional and genomic context in pathway analysis of GWAS data. *Trends Genet.* 30 (9), 390–400. doi:10.1016/j.tig.2014.07.004

Morilla, I., Lees, J. G., Reid, A. J., Orengo, C., and Ranea, J. A. G. (2010). Assessment of protein domain fusions in human protein interaction networks prediction: Application to the human kinetochore model. *N. Biotechnol.* 27 (6), 755–765. doi:10.1016/j.nbt.2010.09.005

Mosca, E., Bersanelli, M., Matteuzzi, T., Di Nanni, N., Castellani, G., Milanesi, L., et al. (2021). Characterization and comparison of gene-centered human interactomes. *Brief. Bioinform.* 22 (6), bbab153. doi:10.1093/bib/bbab153

Mosca, E., and Milanesi, L. (2013). Network-based analysis of omics with multi-objective optimization. *Mol. Biosyst.* 9 (12), 2971–2980. doi:10.1039/c3mb70327d

Mrvar, A., and Batagelj, V. (2016). Analysis and visualization of large networks with program package Pajek. *Complex adapt. Syst. Model.* 4 (1), 6. doi:10.1186/s40294-016-0017-8

Murakami, Y., and Mizuguchi, K. (2010). Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics* 26 (15), 1841–1848. doi:10.1093/bioinformatics/btq302

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U. S. A.* 116 (44), 22071–22080. doi:10.1073/pnas.1900654116

Murphy, M., Jegelka, S., and Fraenkel, E. (2022). Self-supervised learning of cell type specificity from immunohistochemical images. *Bioinformatics* 38 (1), i395–i403. doi:10.1093/bioinformatics/btac263

Nadeau, R., Byvsheva, A., and Lavallée-Adam, M. (2021). Pignon: A protein-protein interaction-guided functional enrichment analysis for quantitative proteomics. *BMC Bioinforma.* 22 (1), 302. doi:10.1186/s12859-021-04042-6

Narayanan, T., Gersten, M., Subramaniam, S., and Grama, A. (2011). Modularity detection in protein-protein interaction networks. *BMC Res. Notes* 4 (1), 569. doi:10.1186/1756-0500-4-569

Nath, A., and Leier, A. (2020). Improved cytokine-receptor interaction prediction by exploiting the negative sample space. *BMC Bioinforma.* 21 (1), 493. doi:10.1186/s12859-020-03835-5

Navlakha, S., He, X., Faloutsos, C., and Bar-Joseph, Z. (2014). Topological properties of robust biological and computational networks. *J. R. Soc. Interface* 11 (96), 20140283. doi:10.1098/rsif.2014.0283

Neuditschko, M., Khatkar, M. S., and Raadsma, H. W. (2012). NetView: A high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. *PLOS ONE* 7 (10), e48375. doi:10.1371/journal.pone.0048375

Ngounou Wetie, A. G., Sokolowska, I., Woods, A. G., Roy, U., Loo, J. A., and Darie, C. C. (2013). Investigation of stable and transient protein–protein interactions: Past, present, and future. *PROTEOMICS* 13 (3–4), 538–557. doi:10.1002/pmic.201200328

Nguyen, V. T., Le, T. T. K., Than, K., and Tran, D. H. (2021). Predicting miRNA-disease associations using improved random walk with restart and integrating multiple similarities. *Sci. Rep.* 11 (1), 21071. doi:10.1038/s41598-021-00677-w

Nitzan, M., Casadiego, J., and Timme, M. (2017). Revealing physical interaction networks from statistics of collective dynamics. *Sci. Adv.* 3 (2), e1600396. doi:10.1126/sciadv.1600396

Novkovic, M., Onder, L., Bocharov, G., and Ludewig, B. (2020). Topological structure and robustness of the lymph node conduit system. *Cell Rep.* 30 (3), 893–904. doi:10.1016/j.celrep.2019.12.070

Otasek, D., Morris, J. H., Boucas, J., Pico, A. R., and Demchak, B. (2019). Cytoscape automation: Empowering workflow-based network analysis. *Genome Biol.* 20 (1), 185. doi:10.1186/s13059-019-1758-4

Ou-Yang, L., Yan, H., and Zhang, X.-F. (2017). A multi-network clustering method for detecting protein complexes from multiple heterogeneous networks. *BMC Bioinforma.* 18 (13), 463. doi:10.1186/s12859-017-1877-4

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* 30 (1), 187–200. doi:10.1002/pro.3978

Pablo Porras, M., Ochoa, D., and Rogon, M. (2020). *Network analysis of protein interaction data: An introduction*. Hinxton, Cambridgeshire, UK: EBI : european Bioinformatic Institute. doi:10.6019/TOL.Networks_t.2016.00001.1

Page, L., Brin, S., and Motwani, R. (1999). The PageRank citation ranking: Bringing order to the web. *Stanf. InfoLab*. Availableat: http://ilpubs.stanford.edu:8090/422/(Accessed: March 21, 2022).

Pak, M., Jeong, D., and Moon, J. (2021). Network propagation for the analysis of multi-omics data. *Recent Adv. Biol. Netw. Analysis*, 185–217. doi:10.1007/978-3-030-57173-3_9

Pan, J., Li, P., and Hong, Z. (2021). Prediction of protein–protein interactions in Arabidopsis, maize, and rice by combining deep neural network with discrete hilbert transform. *Front. Genet.* 12. doi:10.3389/fgene.2021.745228

Pan, J., You, Z. H., and Li, L. P. (2022). Dwppi: A deep learning approach for predicting protein–protein interactions in plants based on multi-source information with a large-scale biological network. *Front. Bioeng. Biotechnol.* 10. doi:10.3389/fbioe.2022.807522

Pandey, B. (2018). *Analysis of protein-protein interaction networks using high performance scalable tools*, 33.

Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T., and Iliopoulos, I. (2015). Protein–protein interaction predictions using text mining methods. *Methods* 74, 47–53. doi:10.1016/j.ymeth.2014.10.026

Patil, A., Nakai, K., and Nakamura, H. (2011). HitPredict: A database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res.* 39, D744–D749. doi:10.1093/nar/gkq897

Pattin, K. A., and Moore, J. H. (2009). Role for protein–protein interaction databases in human genetics. *Expert Rev. Proteomics* 6 (6), 647–659. doi:10.1586/epr.09.86

Paul, M., and Anand, A. (2022). A new family of similarity measures for scoring confidence of protein interactions using gene ontology. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19 (1), 19–30. doi:10.1109/TCBB.2021.3083150

Pazos, F., HelMer-Citterich, M., Ausiello, G., and VAlenciA, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* 271 (4), 511–523. doi:10.1006/jmbi.1997.1198

Pei, F., Shi, Q., Zhang, H., and Bahar, I. (2021). Predicting protein–protein interactions using symmetric logistic matrix factorization. *J. Chem. Inf. Model.* 61 (4), 1670–1682. doi:10.1021/acs.jcim.1c00173

PeixotoTiago, P. (2014). The graph-tool python library. Availableat: http://figshare.com/articles/graph_tool/1164194.

Pellegrini, M. (2019). "Community detection in biological networks," in *Encyclopedia of bioinformatics and computational biology*. Editor S. Ranganathan (Oxford: Academic Press), 978–987. doi:10.1016/B978-0-12-809633-8.20428-7

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96 (8), 4285–4288. doi:10.1073/pnas.96.8.4285

Peng, X., Wang, J., Peng, W., Wu, F. X., and Pan, Y. (2017). Protein–protein interactions: Detection, reliability assessment and applications. *Brief. Bioinform.* 18 (5), 798–819. doi:10.1093/bib/bbw066

Perlasca, P., Frasca, M., Ba, C. T., Gliozzo, J., Notaro, M., Pennacchioni, M., et al. (2020). Multi-resolution visualization and analysis of biomolecular networks

through hierarchical community detection and web-based graphical tools. *PloS One* 15 (12), e0244241. doi:10.1371/journal.pone.0244241

Picard, M., Scott-Boyer, M. P., Bodein, A., Perin, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi:10.1016/j.csbj.2021.06.030

Piehler, J. (2005). New methodologies for measuring protein interactions *in vivo* and *in vitro*. *Curr. Opin. Struct. Biol.* 15 (1), 4–14. doi:10.1016/j.sbi.2005.01.008

Piereck, B., Oliveira-Lima, M., Benko-Iseppon, A. M., Diehl, S., Schneider, R., Brasileiro-Vidal, A. C., et al. (2020). LAITOR4HPC: A text mining pipeline based on HPC for building interaction networks. *BMC Bioinforma.* 21 (1), 365. doi:10.1186/s12859-020-03620-4

Pietrosemoli, N., and Dobay, M. P. (2018). "Optimized protein–protein interaction network usage with context filtering," in *Computational cell biology: Methods and protocols*. Editors L. von Stechow and A. Santos Delgado (New York, NY: Springer), 33–50. doi:10.1007/978-1-4939-8618-7_2

Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., et al. (2019). Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. *Metabolites* 9 (4), E76. doi:10.3390/metabo9040076

Pirch, S., Muller, F., Iofinova, E., Pazmandi, J., Hutter, C. V. R., Chiettini, M., et al. (2021). The VRNetzer platform enables interactive network analysis in Virtual Reality. *Nat. Commun.* 12 (1), 2432. doi:10.1038/s41467-021-22570-w

Pizzuti, C., and Rombo, S. E. (2014). Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics* 30 (10), 1343–1352. doi:10.1093/bioinformatics/btu034

Pournoor, E., Mousavian, Z., Dalini, A. N., and Masoudi-Nejad, A. (2020). Identification of key components in colon adenocarcinoma using transcriptome to interactome multilayer framework. *Sci. Rep.* 10 (1), 4991. doi:10.1038/s41598-020-59605-z

Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Brief. Bioinform.* 11 (1), 15–29. doi:10.1093/bib/bbp057

Qi, J., Sun, H., Zhang, Y., Wang, Z., Xun, Z., Li, Z., et al. (2022). Single-cell and spatial analysis reveal interaction of FAP+ fibroblasts and SPP1+ macrophages in colorectal cancer. *Nat. Commun.* 13, 1742. doi:10.1038/s41467-022-29366-6

Qu, J., Wang, C. C., Cai, S. B., Zhao, W. D., Cheng, X. L., and Ming, Z. (2021). Biased random walk with restart on multilayer heterogeneous networks for MiRNA-disease association prediction. *Front. Genet.* 12, 720327. doi:10.3389/fgene.2021.720327

Rabbani, G., Baig, M. H., Ahmad, K., and Choi, I. (2018). Protein-protein interactions and their role in various diseases and their prediction techniques. *Curr. Protein Pept. Sci.* 19 (10), 948–957. doi:10.2174/1389203718666170828122927

Raja, K., Natarajan, J., and Kuusisto, F. (2020). Automated extraction and visualization of protein-protein interaction networks and beyond: A text-mining protocol. *Methods Mol. Biol.* 2074, 13–34. doi:10.1007/978-1-4939-9873-9_2

Raja, K., Subramani, S., and Natarajan, J. (2013). PPInterFinder—A mining tool for extracting causal relations on human proteins from literature. *Database*, 2013, bas052. doi:10.1093/database/bas052

Raman, K. (2010). Construction and analysis of protein–protein interaction networks. *Autom. Exp.* 2, 2. doi:10.1186/1759-4499-2-2

Rao, V. S., Srinivas, K., Sujini, G. N., and Kumar, G. N. S. (2014). Protein-protein interaction detection: Methods and analysis. *Int. J. Proteomics*, 2014, 147648. doi:10.1155/2014/147648

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res.* 46 (20), 10546–10562. doi:10.1093/nar/gky889

Razaghi-Moghadam, Z., and Nikoloski, Z. (2020). Supervised learning of gene regulatory networks. *Curr. Protoc. Plant Biol.* 5 (2), e20106. doi:10.1002/cppb.20106

Reimand, J., Hui, S., Jain, S., Law, B., and Bader, G. D. (2012). Domain-mediated protein interaction prediction: From genome to network. *FEBS Lett.* 586 (17), 2751–2763. doi:10.1016/j.febslet.2012.04.027

Ren, J., Wang, J., Li, M., and Wang, L. (2013). Identifying protein complexes based on density and modularity in protein-protein interaction network. *BMC Syst. Biol.* 7 (4), S12. doi:10.1186/1752-0509-7-S4-S12

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., et al. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.* 30 (10), 2212–2223. doi:10.1093/nar/30.10.2212

Rohani, N., and Eslahchi, C. (2019). Drug-drug interaction predicting by neural network using integrated similarity. *Sci. Rep.* 9 (1), 13645. doi:10.1038/s41598-019-50121-3

Roth, A., Subramanian, S., and Ganapathiraju, M. K. (2018). Towards extracting supporting information about predicted protein-protein interactions. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15 (4), 1239–1246. doi:10.1109/TCBB.2015.2505278

Ryu, J. Y., Kim, J., Shon, M. J., Sun, J., Jiang, X., Lee, W., et al. (2019). Profiling protein–protein interactions of single cancer cells with *in situ* lysis and co-immunoprecipitation. *Lab. Chip* 19 (11), 1922. doi:10.1039/C9LC00139E

Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B., and Peyvandi, A. A. (2014). Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol. Hepatol. Bed Bench* 7 (1), 17–31.

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., Lotia, S., et al. (2012). A travel guide to Cytoscape plugins. *Nat. Methods* 9 (11), 1069–1076. doi:10.1038/nmeth.2212

Salazar, G. A., Meintjes, A., Mazandu, G. K., Rapanoel, H. A., Akinola, R. O., and Mulder, N. J. (2014). A web-based protein interaction network visualizer. *BMC Bioinforma.* 15 (1), 129. doi:10.1186/1471-2105-15-129

Salazar-Ciudad, I., and Jernvall, J. (2013). The causality horizon and the developmental bases of morphological evolution. *Biol. Theory* 8 (3), 286–292. doi:10.1007/s13752-013-0121-3

Sandoval, O., and Orlando, O. (2021). Analysis and visualization of signal execution in network-driven biological processes. Thesis. Availableat: https://ir.vanderbilt.edu/handle/1803/16761 (Accessed: March 7, 2022).

Santiago-Rodriguez, T. M., and Hollister, E. B. (2021). Multi 'omic data integration: A review of concepts, considerations, and approaches. *Semin. Perinatol.* 45 (6), 151456. doi:10.1016/j.semperi.2021.151456

Sanz-Pamplona, R., Berenguer, A., Sole, X., Cordero, D., Crous-Bou, M., Serra-Musach, J., et al. (2012). Tools for protein-protein interaction network analysis in cancer research. *Clin. Transl. Oncol.* 14 (1), 3–14. doi:10.1007/s12094-012-0755-9

Sardiu, M. E., Gilmore, J. M., Groppe, B. D., Dutta, A., Florens, L., and Washburn, M. P. (2019). Topological scoring of protein interaction networks. *Nat. Commun.* 10, 1118. doi:10.1038/s41467-019-09123-y

Sarkar, D., and Saha, S. (2019). Machine-learning techniques for the prediction of protein–protein interactions. *J. Biosci.* 44 (4), 104. doi:10.1007/s12038-019-9909-z

Schaefer, M. H., Lopes, T. J. S., Mah, N., Shoemaker, J. E., Matsuoka, Y., Fontaine, J. F., et al. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.* 9 (1), e1002860. doi:10.1371/journal.pcbi.1002860

Schneider, L., Guo, Y. K., Birch, D., and Sarkies, P. (2021). Network-based visualisation reveals new insights into transposable element diversity. *Mol. Syst. Biol.* 17 (6), e9600. doi:10.15252/msb.20209600

Scott, M. S., and Barton, G. J. (2007). Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinforma.* 8, 239. doi:10.1186/1471-2105-8-239

Sebestyén, V., Domokos, E., and Abonyi, J. (2020). Multilayer network based comparative document analysis (MUNCoDA). *MethodsX* 7, 100902. doi:10.1016/j.mex.2020

Sejdiu, B. I., and Tieleman, D. P. (2021). ProLint: A web-based framework for the automated data analysis and visualization of lipid-protein interactions. *Nucleic Acids Res.* 49 (W1), W544–W550. doi:10.1093/nar/gkab409

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Sharan, R., and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* 24 (4), 427–433. doi:10.1038/nbt1196

Sharma, P., and Shrivastava, N. (2015). "Artificial neural network to prediction of protein-protein interactions in yeast," in 2015 International Conference on Computer, Communication and Control, 1–5. doi:10.1109/IC4.2015.73756384

Sharp, A. (2017). Janusgraph. Availableat: https://janusgraph.org/.

Shay Banon, B. (2014). Elastic. Availableat: https://www.elastic.co/.

Shen, Z.-A., Luo, T., Zhou, Y. K., Yu, H., and Du, P. F. (2021). NPI-GNN: Predicting ncRNA–protein interactions with deep graph neural networks. *Brief. Bioinform.* 22 (5), bbab051. doi:10.1093/bib/bbab051

Shi, Q., Chen, W., Huang, S., Wang, Y., and Xue, Z. (2021). Deep learning for mining protein data. *Brief. Bioinform.* 22 (1), 194–218. doi:10.1093/bib/bbz156

Shirmohammady, N., Izadkhah, H., and Isazadeh, A. (2021). PPI-GA: A novel clustering algorithm to identify protein complexes within protein-protein interaction networks using genetic algorithm. *Complexity* 2021, e2132516. doi:10.1155/2021/2132516

Shoemaker, B. A., and Panchenko, A. R. (2007). Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.* 3 (3), e42. doi:10.1371/journal.pcbi.0030042

Silverbush, D., and Sharan, R. (2019). A systematic approach to orient the human protein–protein interaction network. *Nat. Commun.* 10 (1), 3015. doi:10.1038/s41467-019-10887-6

Skrabanek, L., Saini, H. K., Bader, G. D., and Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Mol. Biotechnol.* 38 (1), 1–17. doi:10.1007/s12033-007-0069-2

Škunca, N., and Dessimoz, C. (2015). Phylogenetic profiling: How much input data is enough? *PLOS ONE* 10 (2), e0114701. doi:10.1371/journal.pone.0114701

Sloutsky, R., Jimenez, N., Swamidass, S. J., and Naegle, K. M. (2013). Accounting for noise when clustering biological data. *Brief. Bioinform.* 14 (4), 423–436. doi:10.1093/bib/bbs057

Smith-Aguilar, S. E., Aureli, F., Busia, L., Schaffner, C., and Ramos-Fernandez, G. (2019). Using multiplex networks to capture the multidimensional nature of social structure. *Primates.* 60 (3), 277–295. doi:10.1007/s10329-018-0686-3

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* 11 (12), 848. doi:10.15252/msb.20156351

Soleimani Zakeri, N. S., Pashazadeh, S., and MotieGhader, H. (2021). 'Drug repurposing for alzheimer's disease based on protein-protein interaction network'. *BioMed Res. Int.* 2021, 1280237. doi:10.1155/2021/1280237

Song, B., Luo, X., Luo, X., Liu, Y., Niu, Z., and Zeng, X. (2022). Learning spatial structures of proteins improves protein-protein interaction prediction. *Brief. Bioinform.* 23 (2), bbab558. doi:10.1093/bib/bbab558

Stacey, R. G., Skinnider, M. A., and Foster, L. J. (2021). On the robustness of graph-based clustering to random network alterations. *Mol. Cell. Proteomics.* 20, 100002. doi:10.1074/mcp.RA120.002275

Stelzl, U., and Wanker, E. E. (2006). The value of high quality protein–protein interaction networks for systems biology. *Curr. Opin. Chem. Biol.* 10 (6), 551–558. doi:10.1016/j.cbpa.2006.10.005

Stringer, B., de Ferrante, H., Abeln, S., Heringa, J., Feenstra, K. A., and Haydarlou, R. (2022). Pipenn: Protein interface prediction from sequence with an ensemble of neural nets. *Bioinformatics* 38 (8), 2111–2118. doi:10.1093/bioinformatics/btac071

Subramani, S., Kalpana, R., Monickaraj, P. M., and Natarajan, J. (2015). HPIminer: A text mining system for building and visualizing human protein interaction networks and pathways. *J. Biomed. Inf.* 54, 121–131. doi:10.1016/j.jbi.2015.01.006

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* 14, 1177932219899051. doi:10.1177/1177932219899051

Suderman, M., and Hallett, M. (2007). Tools for visually exploring biological networks. *Bioinformatics* 23 (20), 2651–2659. doi:10.1093/bioinformatics/btm401

Summer, G., Kelder, T., Ono, K., Radonjic, M., Heymans, S., and Demchak, B. (2015). cyNeo4j: connecting Neo4j and Cytoscape. *Bioinformatics* 31 (23), 3868–3869. doi:10.1093/bioinformatics/btv460

Sun, P., Ma, Y., Lu, L., and Ma, Z. (2008). "Application of improved K-mean clustering in predicting protein-protein interactions," in 2008 International Conference on BioMedical Engineering and Informatics, Sanya, China, 83–86. doi:10.1109/BMEI.2008.82

Swamy, K. B. S., Schuyler, S. C., and Leu, J.-Y. (2021). Protein complexes form a basis for complex hybrid incompatibility. *Front. Genet.* 12. doi:10.3389/fgene.2021.609766

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi:10.1093/nar/gky1131

Tagore, S., Gorohovski, A., Jensen, L. J., and Frenkel-Morgenstern, M. (2019). ProtFus: A comprehensive method characterizing protein-protein interactions of fusion proteins. *PLoS Comput. Biol.* 15 (8), e1007239. doi:10.1371/journal.pcbi.1007239

Tang, M., Wu, L., and Yu, X. (2021). Prediction of protein–protein interaction sites based on stratified attentional mechanisms. *Front. Genet.* 12. doi:10.3389/fgene.2021.784863

Tanwar, H., and George Priya Doss, C. (2018). Computational resources for predicting protein-protein interactions. *Adv. Protein Chem. Struct. Biol.* 110, 251–275. doi:10.1016/bs.apcsb.2017.07.006

Terayama, K., Shinobu, A., Tsuda, K., Takemura, K., and Kitao, A. (2019). evERdock Bai: Machine-learning-guided selection of protein-protein complex structure. *J. Chem. Phys.* 151 (21), 215104. doi:10.1063/1.5129551

Tesoriero, C. (2013). *Getting started with OrientDB*. Availableat: http://orientdb.com/docs/last/index.html.

Thanasomboon, R., Kalapanulak, S., Netrphan, S., and Saithong, T. (2020). Exploring dynamic protein-protein interactions in cassava through the integrative interactome network. *Sci. Rep.* 10 (1), 6510. doi:10.1038/s41598-020-63536-0

The UniProt Consortium (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* 47 (1), D506–D515. doi:10.1093/nar/gky1049

Tian, X., Ju, H., and Yang, W. (2019). An ego network analysis approach identified important biomarkers with an association to progression and metastasis of gastric cancer. *J. Cell. Biochem.* 120 (9), 15963–15970. doi:10.1002/jcb.28873

Tian, Y., Zhang, B., Hoffman, E. P., Clarke, R., Zhang, Z., Shih, I. M., et al. (2015). Kddn: An open-source cytoscape app for constructing differential dependency networks with significant rewiring. *Bioinforma. Oxf. Engl.* 31 (2), 287–289. doi:10.1093/bioinformatics/btu632

Tieri, P., Zhou, X., Zhu, L., and Nardini, C. (2014). Multi-omic landscape of rheumatoid arthritis: Re-evaluation of drug adverse effects. *Front. Cell Dev. Biol.* 2, 59. doi:10.3389/fcell.2014.00059

Tillier, E. R. M., and Charlebois, R. L. (2009). The human protein coevolution network. *Genome Res.* 19 (10), 1861–1871. doi:10.1101/gr.092452.109

Timón-Reina, S., Rincón, M., and Martínez-Tomás, R. (2021). An overview of graph databases and their applications in the biomedical domain. *Database.* 2021, baab026. doi:10.1093/database/baab026

Tirosh, I., and Barkai, N. (2005). Computational verification of protein-protein interactions by orthologous co-expression. *BMC Bioinforma.* 6 (1), 40. doi:10.1186/1471-2105-6-40

Tomkins, J. E., and Manzoni, C. (2021). Advances in protein-protein interaction network analysis for Parkinson's disease. *Neurobiol. Dis.* 155, 105395. doi:10.1016/j.nbd.2021.105395

Touré, V., Mazein, A., Waltemath, D., Balaur, I., Saqi, M., Henkel, R., et al. (2016). Ston: Exploring biological pathways using the SBGN standard and graph databases. *BMC Bioinforma.* 17 (1), 494. doi:10.1186/s12859-016-1394-x

Truong, K., and Ikura, M. (2003). Domain fusion analysis by applying relational algebra to protein sequence and domain databases. *BMC Bioinforma.* 4 (1), 16. doi:10.1186/1471-2105-4-16

Ulfenborg, B. (2019). Vertical and horizontal integration of multi-omics data with miodin. *BMC Bioinforma.* 20 (1), 649. doi:10.1186/s12859-019-3224-4

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., et al. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 35 (3), 497–505. doi:10.1093/bioinformatics/bty637

Vapnik, V. (1963). Pattern recognition using generalized portrait method. Availableat: https://www.semanticscholar.org/paper/Pattern-recognition-using-generalized-portrait-Vapnik/7cabbdf6a7288d15e26fa6ea504009bab3d1edf4 (Accessed January 16, 2022).

Veitia, R. A. (2002). Rosetta stone proteins: "chance and necessity". *Genome Biol.* 3 (2). doi:10.1186/gb-2002-3-2-interactions1001

Vella, D., Marini, S., Vitali, F., Di Silvestre, D., Mauri, G., and Bellazzi, R. (2018). Mtgo: PPI network analysis via topological and functional module identification. *Sci. Rep.* 8 (1), 5499. doi:10.1038/s41598-018-23672-0

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563 (7731), 347–353. doi:10.1038/s41586-018-0698-6

Villaveces, J. M., Jimenez, R. C., Porras, P., Del-ToroN.DuesburyM.DuMousseauM., et al. (2015a). Merging and scoring molecular interactions utilising existing community standards: Tools, use-cases and a case study. *Database.* 2015, bau131. doi:10.1093/database/bau131

Villaveces, J. M., Koti, P., and Habermann, B. H. (2015b). Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Adv. Appl. Bioinform. Chem.* 8, 11–22. doi:10.2147/AABC.S63534

Vinayagam, A., Zirin, J., Roesel, C., Hu, Y., Yilmazel, B., Samsonova, A. A., et al. (2014). Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nat. Methods* 11 (1), 94–99. doi:10.1038/nmeth.2733

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., et al. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417 (6887), 399–403. doi:10.1038/nature750

Vyas, R., Bapat, S., Jain, E., Karthikeyan, M., Tambe, S., and Kulkarni, B. D. (2016). Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis. *Comput. Biol. Chem.* 65, 37–44. doi:10.1016/j.compbiolchem.2016.09.011

Wahab Khattak, F., Salamah Alhwaiti, Y., Ali, A., Faisal, M., and Siddiqi, M. H. (2021). Protein-protein interaction analysis through network topology (oral cancer). *J. Healthc. Eng.* 2021, 6623904. doi:10.1155/2021/6623904

Wandy, J., and Daly, R. (2021). GraphOmics: An interactive platform to explore and integrate multi-omics data. *BMC Bioinforma.* 22 (1), 603. doi:10.1186/s12859-021-04500-1

Wang, J., Li, M., Deng, Y., and Pan, Y. (2010). Recent advances in clustering methods for protein interaction networks. *BMC Genomics* 11 (3), S10. doi:10.1186/1471-2164-11-S3-S10

Wang, K., Wang, X., Zheng, S., Niu, Y., Zheng, W., Qin, X., et al. (2018a). iTRAQ-based quantitative analysis of age-specific variations in salivary proteome of caries-susceptible individuals. *J. Transl. Med.* 16 (1), 293. doi:10.1186/s12967-018-1669-2

Wang, R.-S., Zhang, S., Wang, Y., Zhang, X. S., and Chen, L. (2008). Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. *Neurocomputing* 72 (1), 134–141. doi:10.1016/j.neucom.2007.12.043

Wang, S.-C. (2003). "Artificial neural network," in *Interdisciplinary computing in java programming*. Editor S.-C. Wang (Boston, MA: Springer US (The Springer International Series in Engineering and Computer Science), 81–100. doi:10.1007/978-1-4615-0377-4_5

Wang, X., and Jin, Y. (2017). Predicted networks of protein-protein interactions in Stegodyphus mimosarum by cross-species comparisons. *BMC Genomics* 18, 716. doi:10.1186/s12864-017-4085-8

Wang, X., Yang, Y., Li, K., Li, W., Li, F., and Peng, S. (2021). BioERP: Biomedical heterogeneous network-based self-supervised representation learning approach for entity relationship predictions. *Bioinforma. Oxf. Engl.* 37, 4793–4800. doi:10.1093/bioinformatics/btab565

Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2019). Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 35 (14), 2395–2402. doi:10.1093/bioinformatics/bty995

Wang, Y., Li, L., and Li, C. (2018b). Predicting protein interactions using a deep learning method-stacked sparse autoencoder combined with a probabilistic classification vector machine. *Complexity* 2018, e4216813. doi:10.1155/2018/4216813

Wang, Z., Li, Y., and Pan, J. (2020). Prediction of protein-protein interactions from protein sequences by combining MatPCA feature extraction algorithms and weighted sparse representation models. *Math. Problems Eng.* 2020, e5764060. doi:10.1155/2020/5764060

Watson, J., Schwartz, J.-M., and Francavilla, C. (2021). Using multilayer heterogeneous networks to infer functions of phosphorylated sites. *J. Proteome Res.* 20 (7), 3532–3548. doi:10.1021/acs.jproteome.1c00150

Welch, N., Singh, S. S., Kumar, A., Dhruba, S. R., Mishra, S., Sekar, J., et al. (2021). Integrated multiomics analysis identifies molecular landscape perturbations during hyperammonemia in skeletal muscle and myotubes. *J. Biol. Chem.* 297 (3), 101023. doi:10.1016/j.jbc.2021.101023

Wen, Y., Song, X., Yan, B., Yang, X., Wu, L., Leng, D., et al. (2021). Multi-dimensional data integration algorithm based on random walk with restart. *BMC Bioinforma.* 22 (1), 97. doi:10.1186/s12859-021-04029-3

Winkler, J., Mylle, E., De Meyer, A., Pavie, B., Merchie, J., Grones, P., et al. (2021). Visualizing protein–protein interactions in plants by rapamycin-dependent delocalization. *Plant Cell* 33 (4), 1101–1117. doi:10.1093/plcell/koab004

Woo, H.-M., and Yoon, B.-J. (2021). Monaco: Accurate biological network alignment through optimal neighborhood matching between focal nodes. *Bioinforma. Oxf. Engl.* 37 (10), 1401–1410. doi:10.1093/bioinformatics/btaa962

Wörheide, M. A., Krumsiek, J., Kastenmuller, G., and Arnold, M. (2021). Multi-omics integration in biomedical research – a metabolomics-centric review. *Anal. Chim. Acta* 1141, 144–162. doi:10.1016/j.aca.2020.10.038

Wu, J., Khodaverdian, A., Weitz, B., and Yosef, N. (2019). Connectivity problems on heterogeneous graphs. *Algorithms Mol. Biol.* 14, 5. doi:10.1186/s13015-019-0141-z

Wu, L., Wang, X., Zhang, J., Luan, T., Bouveret, E., and Yan, X. (2017). Flow cytometric single-cell analysis for quantitative *in vivo* detection of protein–protein interactions via relative reporter protein expression measurement. *Anal. Chem.* 89 (5), 2782–2789. doi:10.1021/acs.analchem.6b03603

Wu, L., Xie, X., Liang, T., Ma, J., Yang, L., Yang, J., et al. (2022a). Integrated multi-omics for novel aging biomarkers and antiaging targets. *Biomolecules* 12 (1), 39. doi:10.3390/biom12010039

Wu, X., Zeng, W., Lin, F., and Zhou, X. (2021). NeuRank: Learning to rank with neural networks for drug–target interaction prediction. *BMC Bioinforma.* 22 (1), 567. doi:10.1186/s12859-021-04476-y

Wu, Y., Gao, M., Zeng, M., Zhang, J., and Li, M. (2022b). BridgeDPI: A novel graph neural network for predicting drug-protein interactions. *Bioinforma. Oxf. Engl.* 38, 2571–2578. doi:10.1093/bioinformatics/btac155

Wu, Y., Zhang, X., Yu, J., and Ouyang, Q. (2009). Identification of a topological characteristic responsible for the biological robustness of regulatory networks. *PLoS Comput. Biol.* 5 (7), e1000442. doi:10.1371/journal.pcbi.1000442

Xia, J., Benner, M. J., and Hancock, R. E. W. (2014). NetworkAnalyst - integrative approaches for protein–protein interaction network analysis and visual exploration. *Nucleic Acids Res.* 42 (W1), W167–W174. doi:10.1093/nar/gku443

Xia, K., Fu, Z., Hou, L., and Han, J. D. J. (2008). Impacts of protein–protein interaction domains on organism and network complexity. *Genome Res.* 18 (9), 1500–1508. doi:10.1101/gr.068130.107

Xie, Z., Deng, X., and Shu, K. (2020). Prediction of protein–protein interaction sites using convolutional neural network and improved data sets. *Int. J. Mol. Sci.* 21 (2), 467. doi:10.3390/ijms21020467

Xu, B., Guan, J., Wang, Y., and Wang, Z. (2019). Essential protein detection by random walk on weighted protein-protein interaction networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16 (2), 377–387. doi:10.1109/TCBB.2017.2701824

Xu, B., Liu, Y., and Dong, J. (2018). Reconstruction of the protein-protein interaction network for protein complexes identification by walking on the protein pair fingerprints similarity network. *Front. Genet.* 9. doi:10.3389/fgene.2018.00272

Xu, W., Gao, Y., Wang, Y., and Guan, J. (2021). Protein-protein interaction prediction based on ordinal regression and recurrent convolutional neural networks. *BMC Bioinforma.* 22 (6), 485. doi:10.1186/s12859-021-04369-0

Xuan, P., Sun, C., Zhang, T., Ye, Y., Shen, T., and Dong, Y. (2019). Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Front. Genet.* 10, 459. doi:10.3389/fgene.2019.00459

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Brief. Bioinform.* 19 (6), 1370–1381. doi:10.1093/bib/bbx066

Yang, J., Wagner, S. A., and Beli, P. (2015). Illuminating spatial and temporal organization of protein interaction networks by mass spectrometry-based proteomics. *Front. Genet.* 6. Availableat: https://www.frontiersin.org/article/10.3389/fgene.2015.00344 (Accessed: January 28, 2022).

Yann Lecun, L. (1986). A Learning Scheme for assymetric threshold network. Availableat: http://yann.lecun.com/exdb/publis/pdf/lecun-85.pdf (Accessed: February 18, 2022).

Yoon, B.-H., Kim, S.-K., and Kim, S.-Y. (2017). Use of graph database for the integration of heterogeneous biological data. *Genomics Inf.* 15 (1), 19–27. doi:10.5808/GI.2017.15.1.19

You, Z.-H., Lei, Y. K., Zhu, L., Xia, J., and Wang, B. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinforma.* 14 (8), S10. doi:10.1186/1471-2105-14-S8-S10

Yu, H., and Zhang, Z. (2008). *BioReact: Visualization of systems biology network*, 7.

Yu, H., Zhu, X., Greenbaum, D., Karro, J., and Gerstein, M. (2004). TopNet: A tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res.* 32 (1), 328–337. doi:10.1093/nar/gkh164

Yuan, C., Wang, M. H., and Wang, F. (2021). Network pharmacology and molecular docking reveal the mechanism of Scopoletin against non-small cell lung cancer. *Life Sci.* 270, 119105. doi:10.1016/j.lfs.2021.119105

Zahiri, J., Bozorgmehr, J. H., and Masoudi-Nejad, A. (2013). Computational prediction of protein–protein interaction networks: Algo-rithms and resources. *Curr. Genomics* 14 (6), 397–414. doi:10.2174/1389202911314060004

Zahiri, J., Emamjomeh, A., Bagheri, S., Ivazeh, A., Mahdevar, G., Sepasi Tehrani, H., et al. (2020). Protein complex prediction: A survey. *Genomics* 112 (1), 174–183. doi:10.1016/j.ygeno.2019.01.011

Zaki, N., and Tennakoon, C. (2017). BioCarian: Search engine for exploratory searches in heterogeneous biological databases. *BMC Bioinforma.* 18 (1), 435. doi:10.1186/s12859-017-1840-4

Zhang, J., Suo, Y., Liu, M., and Xu, X. (2018). Identification of genes related to proliferative diabetic retinopathy through RWR algorithm based on protein-protein interaction network. *Biochim. Biophys. Acta. Mol. Basis Dis.* 1864 (6), 2369–2375. doi:10.1016/j.bbadis.2017.11.017

Zhang, X., Xu, J., and Xiao, W. (2013). A new method for the discovery of essential proteins. *PloS One* 8 (3), e58763. doi:10.1371/journal.pone.0058763

Zhang, Y., Natale, R., Domingues, A. P., Toleco, M. R., Siemiatkowska, B., Fabregas, N., et al. (2019). Rapid identification of protein-protein interactions in plants. *Curr. Protoc. Plant Biol.* 4 (4), e20099. doi:10.1002/cppb.20099

Zhang, Z. (2018). "Artificial neural network," in *Multivariate time series analysis in climate and environmental research*. Editor Z. Zhang (Cham: Springer International Publishing), 1–35. doi:10.1007/978-3-319-67340-0_1

Zhao, B., Wang, J., and Wu, F.-X. (2017). Computational methods to predict protein functions from protein-protein interaction networks. *Curr. Protein Pept. Sci.* 18 (11), 1120–1131. doi:10.2174/1389203718666170505121219

Zhao, Q., Zhang, Y., and Hu, H. (2018). Irwnrlpi: Integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction. *Front. Genet.* 9. doi:10.3389/fgene.2018.00239

Zhao, T., Hu, Y., Valsdottir, L. R., Zang, T., and Peng, J. (2021). Identifying drug-target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22 (2), 2141–2150. doi:10.1093/bib/bbaa044

Zhong, M., Lee, G. M., Sijbesma, E., Ottmann, C., and Arkin, M. R. (2019). Modulating protein-protein interaction networks in protein homeostasis. *Curr. Opin. Chem. Biol.* 50, 55–65. doi:10.1016/j.cbpa.2019.02.012

Zhou, G., and Xia, J. (2018). OmicsNet: A web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Res.* 46 (1), W514–W522. doi:10.1093/nar/gky510

Zhou, J., Guo, Y., Fu, J., and Chen, Q. (2022). Construction and validation of a glioma prognostic model based on immune microenvironment, Neuroimmunomodulation, 30. 1–12. doi:10.1159/000522529

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). LPI-deepGBDT: A multiple-layer deep framework based on gradient boosting decision trees for lncRNA–protein interaction identification. *BMC Bioinforma.* 22, 479. doi:10.1186/s12859-021-04399-8

Zhou, M., Li, Q., and Wang, R. (2016). Current experimental methods for characterizing protein–protein interactions. *Chemmedchem* 11 (8), 738–756. doi:10.1002/cmdc.201500495

Zhou, W.-Z., Miao, L.-G., and Yuan, H. (2018). Identification of significant ego networks and pathways in rheumatoid arthritis. *J. Cancer Res. Ther.* 14 (1), S1024–S1028. doi:10.4103/0973-1482.189250

Zhu, M.-L., and Schmotzer, C. (2017). Writing the genome: Are we ready? *Clin. Chem.* 63 (4), 929–930. doi:10.1373/clinchem.2016.270066

Zu, G., Sun, K., Li, L., Zu, X., Han, T., Huang, H, et al. (2017). Direct visualization of interaction between calmodulin and connexin45. *Biochem. J.* 474, 22959. doi:10.1038/s41598-021-02248-5

frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 11 October 2022
DOI 10.3389/fmolb.2022.962743

OPEN ACCESS

EDITED BY
Ornella Cominetti,
Nestlé Research Center, Switzerland

REVIEWED BY
Leonidas Salichos,
New York Institute of Technology, United States
Qiuyu Lian,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Raoul Santiago,
raoul.santiago@crchudequebec.ulaval.ca
Arnaud Droit,
arnaud.droit@crchudequebec.ulaval.ca

†These authors share senior authorship

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 06 June 2022
ACCEPTED 21 September 2022
PUBLISHED 11 October 2022

CITATION
Raufaste-Cazavieille V, Santiago R and Droit A (2022), Multi-omics analysis: Paving the path toward achieving precision medicine in cancer treatment and immuno-oncology. *Front. Mol. Biosci.* 9:962743. doi: 10.3389/fmolb.2022.962743

Multi-omics analysis: Paving the path toward achieving precision medicine in cancer treatment and immuno-oncology

Virgile Raufaste-Cazavieille[1], Raoul Santiago[1,2*†] and Arnaud Droit[1*†]

[1]CHU de Québec Research Center, Université Laval, Québec, QC, Canada, [2]Division of Pediatric Hematology-Oncology, Centre Hospitalier Universitaire de L'Université Laval, Charles Bruneau Cancer Center, Québec, QC, Canada

The acceleration of large-scale sequencing and the progress in high-throughput computational analyses, defined as omics, was a hallmark for the comprehension of the biological processes in human health and diseases. In cancerology, the omics approach, initiated by genomics and transcriptomics studies, has revealed an incredible complexity with unsuspected molecular diversity within a same tumor type as well as spatial and temporal heterogeneity of tumors. The integration of multiple biological layers of omics studies brought oncology to a new paradigm, from tumor site classification to pan-cancer molecular classification, offering new therapeutic opportunities for precision medicine. In this review, we will provide a comprehensive overview of the latest innovations for multi-omics integration in oncology and summarize the largest multi-omics dataset available for adult and pediatric cancers. We will present multi-omics techniques for characterizing cancer biology and show how multi-omics data can be combined with clinical data for the identification of prognostic and treatment-specific biomarkers, opening the way to personalized therapy. To conclude, we will detail the newest strategies for dissecting the tumor immune environment and host–tumor interaction. We will explore the advances in immunomics and microbiomics for biomarker identification to guide therapeutic decision in immuno-oncology.

KEYWORDS

multi-omics, machine learning, immunology, immunomics, microbiome, cancer, precision medicine

Introduction

Cancer is an important cause of death worldwide, even if its mortality has declined during the last decade (Santucci et al., 2020). The International Agency for Research on Cancer (IARC) GLOBOCAN cancer statistics predicted an increase of 50% for the cancer incidence and an increase of 62.5% of the mortality from now to 2040 worldwide (Ferlay et al., 2020). There is an urgent need to better cancer survival rate; however, to cure this disease, we first need to understand its underlying mechanisms.

Until recently, cancer was widely considered as organ dependent and characterized by the site of apparition of the tumor. This hypothesis was slowly abandoned due to the heterogeneity present between two patients with a similar tumor (Dagogo-Jack and Shaw, 2018) that was identified through the emergence and democratization of next-generation sequencing (NGS). NGS allows the generation of a large amount of data in a short period of time, marking the beginning of a new era for cancer research: the genomics era (Knox, 2010; Lakshmanan et al., 2020). This led to a new characterization of cancer, not based on the tumor site but founded on molecular classification (pan-classification) (Hoadley et al., 2014; Campbell et al., 2018). Following the expansion of genomics, the exploration of the other layers of cancer biology started. The apparition of different omics such as transcriptomics, epigenomics, and proteomics allowed the possibility to understand the underlying complexity of tumors (Alyass et al., 2015). The dawn of omics studies led to the discovery of further complexity with cancer presenting intra-tumoral heterogeneity at a cellular level (Dagogo-Jack and Shaw, 2018). The study of these new omics highlighted an unsuspected complexity inside the tumor architecture but also furthered our understanding of interactions between cancer and its environment (gene-environment, microenvironmental interaction, and immune system interaction) (McAllister et al., 2017; Zhou et al., 2017; Gonzalez et al., 2018; Barriga et al., 2019).

The use of single omics, such as genomics and transcriptomics, uncovered many driver genes to better comprehend the genomic landscape of cancer. Interestingly, some of these studies revealed the wide complementarity of genomics and transcriptomics, many of the driver genes being identified by either one or the other modality (Wong et al., 2020; Berlanga et al., 2022). This fosters the need of combining the interconnected biological elements of cancer to move from single-omics to multi-omics analysis. The multi-omics integration is defined by the modelization of more than one biological element in order to characterize biological systems in its globality at the phenomenological level (de Anda-Jáuregui and Hernández-Lemus, 2020). The purpose of doing this is to look at how the different biological layers of the cells interact with each other, leading to the creation of an interconnected network highlighting the underlying complexity of cancer. Data integration in cancer have three main goals: understanding the molecular mechanism of cancer, clustering disease samples, and predicting an outcome (survival or therapy efficacy) (Hiley et al., 2014; Jamal-Hanjani et al., 2014; Tebani et al., 2016; Sharifi-Noghabi et al., 2019). Computational methods were needed to integrate the large diversity of data prompting the development of new algorithms to overcome the intricacies of multi-omics integration.

To this day, even with the new information that multi-omics approaches can bring, understanding how cancer develops and maintains is puzzling. Indeed, cancer cells interact with many different components including the host immune system (Witkowski et al., 2020). These interactions define the tumor immune microenvironment (TiME) that can be beneficial or detrimental to cancer cells, leaning toward more tolerogenicity or immunogenicity (Iwai et al., 2002; Garrido and Aptsiauri, 2019; Gou et al., 2020; Tang et al., 2020). To understand the TiME provides insights on how the cancer cells hijack the immune system to survive, but also might predict if a tumor is likely to respond to immunotherapy by immune checkpoint blockade (ICB) (Riaz et al., 2017). Indeed, biomarkers derived from the study of TiME appeared to be helpful to anticipate cancer sensitivity to immunotherapy (Goswami et al., 2020). Immunomics, the field of omics-based analysis that aims to describe the reaction of the immune system to another biological component (pathogen or cancer), can be used to analyze the interactions existing between the host immune system and the cancer cells and how it leads to immune recognition or immune ignorance (Arnaout et al., 2021). However, to fully depict the interconnection that exists between the tumor and the immune system, other players need to be taken into the equation. Indeed, it has been shown that the microbiome can influence the sensitivity of cancers to ICB, suggesting an interplay between the host microbiota and the immune constitution of tumor microenvironment (Baruch et al., 2021). Immunomics and microbiomics are two novel components that should be included into multi-omics models to integrate the tumor/host interaction in cancer complexity.

In the first part, due to the growing interest in multi-omics, this review will address the challenges raised by omics and how to overcome them. In the second part, we will provide an overview of different databases that are useful in cancer research. In the third part, this review will attempt to illustrate how to integrate mutli-omics to clarify cancer complexity, and how the use of machine learning can help to predict survival and treatment response. Finally, we will give an overview of new methods to decipher the interaction between the host and the cancer cells and how it can bring opportunity for personalized therapy (Figure 1).

Challenges

To integrate the massive data flow generated through the different biological elements explored by NGS, innovative computational tools are needed for diminishing the data dimension, then making them manipulable by human hands. These tools allow researchers to fulfill the gap of missing knowledges and contribute to the discovery of novel biomarkers, deciphering the complexity of cancers (Subramanian et al., 2020; Poirion et al., 2021; Wörheide et al., 2021; Zeng et al., 2021). Despite the progresses in these new instruments, multiple challenges remain.

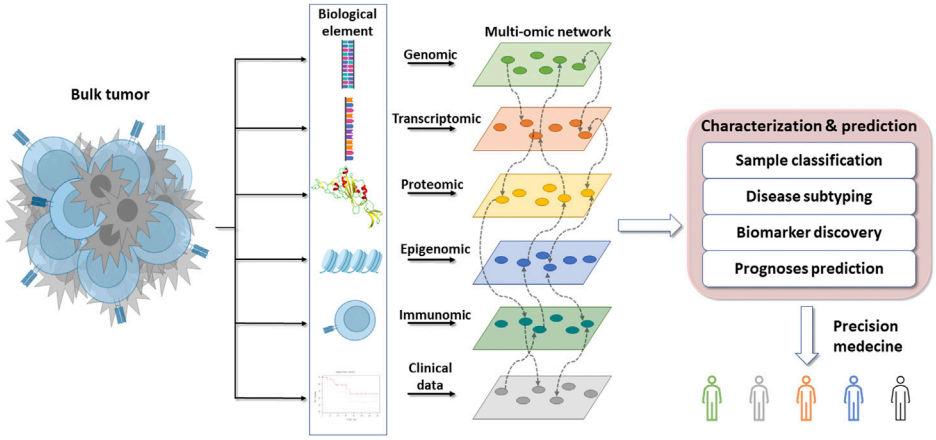


FIGURE 1
Overflow of the omics integration for precision medicine. The bulk tumor is composed of many different biological elements that can be classified and used to find novel interactions between interconnected elements. These interactions can be used to molecularly characterize and subtype cancers with the aim of anticipating their clinical evolution and treatment sensitivity to promote precision medicine.

TABLE 1 Description of different tools for multi-omics integration with their application and their major strength and limits.

Method	Principle	Aim	Omics element	Pros	Cons
JIVE	Matrix factorization	Disease subtyping, systemic knowledge, module detection	Genomics and epigenomics	Integrate large amount of data	Sensitive to outliers and missing values
NMF		Disease subtyping, module detection, biomarker discovery	Genomics and epigenomics	Filtering weak signal. Integrate large amount of data. Detection of cluster of small size	Time and memory consuming. Underperforming on missing values
nNMF					
jNMF					
intNMF					
SLIDE		Disease subtyping, module detection, biomarker discovery	Genomics, epigenomics and proteomics	Integrate large amount of data	Underperforming with missing values. Optimum solution is not guaranteed
MALA	Logic data mining	Sample classification	Genomics and transcriptomics	Works well on experimental data. Integrate large amount of data	Phenotype number must be delivered with data. Sensitive to missing values
iCluster	Gaussian latent variable model		Genomics, epigenomics and transcriptomics		Needs to test a large amount of solution to find the most relevant
iCluster+	Generalized linear regression	Disease subtyping	Genomics, transcriptomics, proteomics and epigenomics	Handle missing values	No evaluation of statistical significance for selected features
iClusterBayes	Bayesian integrative clustering	Biomarker discovery	Genomics, transcriptomics, and epigenomics	Good performance in the presence of explicative data	Underperform with outliers
MOFA	Bayesian factor analysis	Biomarker discovery, systemic knowledge	Proteomics, metabolomics and lipidomics	Handle well missing values	Linear model can miss linear relation
MOFA+			Genomics and epigenomics	The use of continuous learning enabling MOFA to recover different trajectory	Need of multi-modal measurement for the same set of cells

JIVE: joint and individual variation explained.
(n, j, int) NMF: (network, joint, integrative) non-negative matrix factorization.
SLIDE: structural learning and integrative decomposition.
MALA: micro array logic analyzer.
MOFA: multi-omics factor analysis.

One of these challenges is the heterogeneity that exists across the biological layers. It may differ in type, with numerical or categorical features, discrete or continuous variables with different ranges. The number of features can also vary between the different data sources, creating novel struggles to consider. Another burden, operating during sample collection, are missing values, setting up uninterpretable elements difficult to handle by some algorithms. Also, processing outliers and highly correlated variables might be burdensome to integrate as they might be irrelevant, create noise, or overfeed a system (Mirza et al., 2019; Song et al., 2020a).

To face these challenges, different tools have been created that we will summarize (Canzler et al., 2020; Nicora et al., 2020; Subramanian et al., 2020). Two main types of algorithms can be distinguished: 1) the "exploratory matrix" and 2) the "probabilistic matrix". The first one is based on the matrix and its result. It performs best with a well-defined matrix, can handle outliers, but is less accurate when data are missing (Devarajan, 2008; Chauvel et al., 2020; Hamamoto et al., 2022). The probabilistic matrix utilizes probabilistic formulas (such as Gaussian and Bayesian). These algorithms shine when the dataset is incomplete, with missing values, but lose accuracy with outliers (Needham et al., 2007; Yuan et al., 2021; Chu et al., 2022). This approach aims to reduce the size of the matrix to identify patterns within the dataset, allowing for powerful classification models (Table 1).

Exploratory matrix

Starting with exploratory matrices, herein is a brief overview of two algorithms using this approach: matrix factorization and logic data mining. These methods would be preferred for well-defined exhaustive datasets containing extreme values that need to be considered.

Matrix factorization: joint and individual variation explained (JIVE), non-negative factorization (NMF), and structural learning and integrative decomposition (SLIDE) are some of the tools based on this approach. This algorithm uses a system of multiplication between columns and rows on a dataset decomposed in multiple matrices of smaller dimension. Each matrices have n columns, referring to n common objects, that will be multiplied by m rows, referring to other common objects. The matrix can contain different biological elements, that is, gene expression and miRNA measurement. The strength of matrix factorization is to cluster different matrices together but requires a dataset with scarce missing values (Brunet et al., 2004; Devarajan, 2008; Lock et al., 2013; Pierre-Jean et al., 2022).

Logic data mining is a multistep procedure: 1) feature binarization, that assigns a threshold to convert each feature in binary values (0 or 1); 2) feature selection, that uses a machine learning approach to select a subset of data with relevant features introduced in a model; 3) extraction of logic formulas to build the final classification model. As an example, microarray logic analyzer (MALA) is based on this approach and has been used to identify overexpressed interrelated genes and proteins involved in a common pathway (Bertolazzi et al., 2008; Wang and éditeur, 2009; Weitschek et al., 2012).

Probabilistic matrix

We will now detail four types of algorithms to better conceptualize the probabilistic matrix approach. These models follow the principles of Gaussian probability, linear regression, or Bayesian statistics and are very useful to troubleshoot datasets with missing values.

Gaussian latent variable model uses a probabilistic matrix starting with the dimension reduction of the dataset. The matrix composed of N columns and D rows will be decreased to a matrix with a lower dimensionality on D. D will be reduced in Q most relevant data. Relevant data will then be extracted to identify a pattern explaining the results. Because the algorithm is based on Gaussian probability, dataset with a Gaussian distribution perform better with this instrument than expression profiles near 0. Lower outliers close to the null value may cause misinterpretation during the analysis. However, missing values are well managed by the algorithm. A tool using this algorithm, iCluster, will be further explained in the next chapter (Li and Chen, 2016).

Generalized linear regression is one of the subcategories found in the generalized linear model. This algorithm is composed of three components: 1) random component; 2) systematic component; and 3) link function. The first step is the matrix generation, then a linear regression will be applied to the matrix finding the best possible linear relation to fit the expected values. The systematic components are distributed on the line while random components are outside of the line. The link function will identify a relationship between the linear predictor and the distribution of the random components. It aims to explain why some values are following a linear repartition while others are not. This model performs well when using data that are already explained on unexplained data. The accuracy might, however, be hindered by outliers that may introduce misinterpretation. This model is used in the icluster + tools (Mo et al., 2013; Song et al., 2022).

Bayesian integrative clustering: the objective of this approach is to capture the major variations of multiple omics datasets, using a reduction of the high-dimensional space to a low-dimensional subspace. This could be vulgarized into a compaction of the data. The algorithm will extract the principal variations of the datasets to integrate matrices of different dimensions to a single matrix called Z with n x k dimensions, following the rule of multivariate normal distribution. A joint integrative clustering will capture relevant features to individualize distinct clusters across omics datasets.

TABLE 2 Publicly available cancer databases with their main characteristics.

Database	Multi-omics data available	Disease	References
TCGA (The Cancer Genome Atlas)	20,000 individual tumor samples RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA, clinical data (treatment received and response to treatment), histological data	Cancers	(Weinstein et al., 2013; Tomczak et al., 2015)
TCIA (The Cancer Immunome Database)	8,000 tumor samples Genomic immune-related gene set, immune infiltrate, neoantigens, cancer antigens, HLA types, and tumor heterogeneity	Cancers	Charoentong et al. (2017)
ICGC (International Cancer Genomics Consortium)	20,383 samples listed Somatic and germline whole genome sequencing, genomic variation data	Cancers	(Hudson et al., 2010; Australian Pancreatic Cancer Genome InitiativeBailey et al., 2016; Thompson et al., 2018)
METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)	2,503 breast tumor samples Clinical traits, gene expression, SNP, and CNV.	Breast cancers	Curtis et al. (2012)
TARGET (Therapeutically Applicable Research to Generate Effective Treatments)	24 different types of pediatric cancers Gene expression, miRNA expression, CNV, and DNA-seq data	Pediatric cancers	(Ma et al., 2018; Rajbhandari et al., 2018)
CRI (Cancer Research Institute) **iAtlas**	10,000 tumors samples Clinical data (immunotherapy responses and clinical phenotypes), genomics, immunomodulatory genes, neoantigens load	Cancers	(Thorsson et al., 2018; Eddy et al., 2020)

DNA-seq: deoxyribo nucleic acid-sequencing.
RNA-seq: ribonucleic acid-sequencing.
miRNA-seq: micro ribonucleic acid-sequencing.
SNV: single nucleotide variant.
CNV: copy number variation.
SNP: single nucleotide polymorphisms.
DNA, methylation: deoxyribo nucleic acid methylation.
RPPA: reverse-phase proteomic arrays.
miRNA, expression: micro ribonucleic expression.
HLA: human leukocyte antigen.

The use of Baye's theorem enables the analysis of factors with various distributions and correlation among datasets. The probabilistic model is also permissive of missing values (Needham et al., 2007; Fang et al., 2018; Mo et al., 2018).

Bayesian factor analysis: this is a key component of the multi-omics factor analysis (MOFA) family tools used for multi-omics data integration. This unsupervised method infers principal component-based factors to decompose each matrix of M different omics components. The matrices will be transformed into a Z matrix of factors for each sample and M weight (W) matrices with features in rows and factors in columns, for each omics element. Downstream analysis will identify inference across Z and W matrices. Also, because the algorithm is based on a Bayesian framework, the distributions of the data are placed on the unobserved variables of the models and the algorithm will keep running until all the data have been characterized (Needham et al., 2007; Argelaguet et al., 2018; Min et al., 2018).

To finish, despite the numerous tools that are now available, the perfect tool does not exist. The continual development of computational methods necessitates systematic evaluation (benchmarking) of the omics data analyses tools and methods (Mangul et al., 2019). The major issue on this benchmarking is the lack of "gold standard" datasets, providing unbiased ground truth. This lack of gold standard hinders the possibility to establish generalizable benchmarks to test novel complex software (Mangul et al., 2019; Weber, 2019; Marx, 2020). Different notable benchmarks are available comparing: multi-omics and multi-view clustering algorithms (Rappoport and Shamir, 2018), multi-omics dimensionality reduction (Cantini et al., 2021), multi-omics for cancer subtyping (Duan et al., 2021),

and multi-omics survival prediction methods (Herrmann et al., 2021). Regardless of the use of the same dataset for comparison, the results obtained are not necessarily reproducible on another database and cannot be apply to the integration of omics that were not included in the initial dataset. The lack of a gold standard dataset impairs the success of the benchmarking (Krassowski et al., 2020).

Multi-omics database

The expansion and acceleration of NGS have generated a tremendous amount of data. In a collaborative effort, extensive omics databases have been created to stock and make those data publicly accessible to researchers, allowing for large meta-analysis for a tumor category or across cancer types. These datasets can host the sequencing output of different biological elements from bulk tumor or single-cell sequencing as well as clinical and treatment information. We will provide a summary of the main databases that are available for cancer research and detail the characteristics of each library with tumor types included and the biological and clinical contents provided (Table 2).

The Cancer Genome Atlas (TCGA) is the largest pan-cancer multi-omics database with clinical annotation allowing for large meta-analysis. TCGA is widely used by the research community, promoting new discoveries on tumor biology, evolution, and treatment specific biomarkers validated on meta-analysis across cancer types (Weinstein et al., 2013; Tomczak et al., 2015).

The Cancer Immunome Database (TCIA) is a new database of immunogenic analysis of NGS data from 20 different types of solid tumors derived from TCGA database. This database has served for the elaboration of a pan-cancer immunogenomic classification for checkpoint blockade sensitivity (Charoentong et al., 2017).

The International Cancer Genomics Consortium (ICGC) is the most ambitious biomedical efforts research since the human genome project. This database has permitted novel observation of cancer biology through the whole genome annotated alterations from 2,800 samples (Hudson et al., 2010; Australian Pancreatic Cancer Genome InitiativeBailey et al., 2016; Thompson et al., 2018).

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) contains genomic data from breast cancers. This database helped improving the classification and subtyping of breast tumors and already led to the discovery of 10 subgroups (Curtis et al., 2012).

Therapeutically Applicable Research to Generate Effective Treatments (TARGET) is a clinically annotate database hosting genomics and transcriptomics data from pediatric tumors. This dataset empowered the description of the driving process of childhood cancers, and was used, for example, for characterizing the immune environment of pediatric cancers and its prognostic impact (Ma et al., 2018; Rajbhandari et al., 2018; Sherif et al., 2021).

Cancer Research Institute (CRI) iAtlas is a database platform allowing the study of interaction between tumors and immune microenvironment, granting the possibility to explore immune response across genomics and clinical phenotypes. iAtlas (first version) allows researchers to explore these readouts and the relation between tumor types and immune response (Thorsson et al., 2018). CRI iAtlas now shares information about immunotherapy response useful for biomarker identification (Eddy et al., 2020).

The dramatic acceleration in data production and the policy for data sharing have increased the comprehension of cancer complexity and fostered the integration of interrelated biological elements of omics analysis in more and more complex computational models that we will show in detail in the following.

Multi-omics for greater accuracy

Multi-omics for a better classification of cancer types

Tumor heterogeneity remains a hurdle to understand tumor biology. For example, for a same tumor type, there is a wide variation in clinical evolution across population of patients, highlighting differences in tumor cells to progress or mutate (Lagoa et al., 2020; Wu et al., 2020). These differences are barriers to develop efficient therapies (Marusyk et al., 2012; Cyll et al., 2017; Marusyk et al., 2020). Cancer is relatively easy to classify as the classification is based on the histotype and site of origin of the tumor. However, the classification system has increased in complexity in the last decades with the introduction of genomics and molecular features to better account for the clinical evolution and foster the identification of histologic groups (Carbone, 2020). Therefore, omics data can be used to individualize tumor types. In this aim, a variety of bioinformatic tools have been developed to refine and accelerate sample classification.

One of these tools that allows the sample classification is a microarray logic analyzer (MALA). MALA is based on logic data mining algorithm. Its follows a three-step process: (Santucci et al., 2020) discrete cluster analysis, (Ferlay et al., 2020) selection of the most relevant (cluster of) genes, and (Dagogo-Jack and Shaw, 2018) logical formulas characterizing the samples (Weitschek et al., 2012). In the context of cancer research, a comparative study came out with the purpose to evaluate which tools could be the most accurate in treating multilevel omics data. MALA with sparse canonical correlation analysis (SCCA) and non-negative matrix factorization (NMF) were compared. When using the experimental data, MALA performed the best sample classification compared to the others. However, on a larger data set of simulated data, the efficiency of the model decreased (Pucher et al., 2019).

Multi-omics for cancer subtyping

To better understand cancer, researchers need to characterize the different cancer subtypes. Subtyping cancer is more complicated and remains one of the major challenges in cancer research. The identification of subtypes can provide an understanding of the underlying molecular mechanisms and thereby help design precise treatment strategies for efficient cancer management. Contrary to classification that is more histologic, the subtypes are influenced by oncogenic alterations and/or modifications in the gene/protein expression (Anderson et al., 2021). Molecular classification has partially elucidated tumor heterogeneity, however, different subtypes can be identified depending on different layers of biological elements: genomics, alterations, gene and protein expression profile as well as cellular composition (Skoulidis and Heymach, 2019). The main challenge is to be able to accurately analyze the large amount of data and to determine and isolate predictive patterns. Multi-omics provides a powerful approach to process, treat, and characterize the large quantity of data required for cancer subtyping (Menyhárt and Győrffy, 2021).

Cancer subtyping can be efficiently addressed by matrix factorization, such as joint and individual variation explained (JIVE) algorithm. This tool aims to individualize two types of structures: joint and individual structures. The former is the biological patterns of the samples that are shared between the different component types (*i.e.*, gene expression and miRNA), while the latter is intrinsic of one component and unrelated to others (Lock et al., 2013). One structure can interfere in finding a signal in the second structure and *vice versa*. JIVE was developed to distinguish these possible interfering effects by decomposing the datasets into a sum of three terms: (Santucci et al., 2020) low-rank approximation capturing joint variation across the biological components, (Ferlay et al., 2020) low-rank approximation capturing structured variation specific to a given component, and (Dagogo-Jack and Shaw, 2018) residual noise. It explores the information provided by a specific biological data type or by the interaction between several data for subtyping. A logical weakness of this method is its sensitivity to outliers. JIVE algorithm was tested on 234 glioblastoma samples, with an input of miRNA and gene expression matrices. It showed that the gene expression introduced more structured variation than miRNA and that joint structure variation between the gene expression and miRNA was more accurate to classify the biological subgroups. Overall, the multi-component integration improved the subclassification of glioblastoma samples.

In the optic of diseases subtyping, a growing diversity of tools based on matrix factorization were developed. As an example, non-negative factorization (NMF) consists in multiplication between the columns and the rows of a matrix. The input data sets are formed by a matrix called A, composed of N genes and M samples. Genes that regulate the expression of downstream genes will be identified and labeled k. In the second step, the matrix A will be split in two matrices: W composed of the N genes and k, and H composed of k and M samples. Finally, W and H matrices will be combined in a new matrix. The advantage of the NMF technics is to easily cluster gene and samples, but this is a time and memory consuming tool that does not handle negative input and is not designed to integrate multilayer components (Lee and Seung, 1999; Brunet et al., 2004; Zhang et al., 2012; Yang and Michailidis, 2015).

Briefly, NMF extensions were implemented to allow multi-omics profiling. Multi-omics integration with jNMF was compared to single omics to class clinical data from TCGA into different subgroups and outperformed the single element model. Moreover, the number of groups will depend on the clinical data used (Zhang et al., 2012). A second major update was intNMF, while jNMF only considers homogenous effects in *WHI* and intNMF considers heterogenous effects (Yang and Michailidis, 2015). As an example, breast cancer subtyping with multi-omics data integration of five biological components by intNMF refined the subclassification from four classical subsets to six unique clusters (Chalise and Fridley, 2017).

The last update of NMF is called network-based integrative (nNMF). nNMF involves a two-step process with network generation and integration of the network. To generate the network, a consensus matrix is built with a binary value attributed to each sample to reflect the connectivity between samples. Successive cycles are performed attributing a new binary value for every novel entry until the matrix is stabilized. The mean of the consensus matrix is made for each iterative cycle until the last cycle. This generates a consensus matrix for each data type integrated. The larger elements of the consensus matrix reflect the higher similarity between samples. After the networks are generated, samples can be considered as vertices and the consensus values as the edges. The network integration is based on the message passing theory (updating and combining network) that is processed on two ways: strong signals present in any data are conserved and consistent signals in multiple data are added up during an iterative process. Weak signals disappear while filtering out the noise. An advantage of nNMF is its ability to detect true cluster of small size with high reliability. Clinical application of nNMF demonstrated the capacity to establish novel clusters on head and neck squamous cell carcinoma, glioblastoma, and low-grade glioma data sets, suggesting a new comprehensive subtyping that eliminates previously unclassified samples (Chalise et al., 2020).

NMF and its extensions are based on matrix factorization algorithm that are efficient tools to treat a massive amount of data but are underperforming with missing values. To complement this part, iCluster family tools are good alternatives. iCluster is based on probabilistic matrices algorithm and iCluster is based on a Gaussian latent variable model. The basic concept of

iCluster is to jointly estimate the link between the data with a dimension reduction principle: data and features are clustered together to maximize the correlation between data types (Shen et al., 2009). iCluster was used to classify novel subtypes of esophageal cancers based on genomics, epigenomics, and transcriptomics data. The classification varied depending on the type of omics. Compared to previous classification, the samples were consistently classified into three groups with different biological traits and prognostic significance (Ma et al., 2021). It was also used in breast cancer to integrate DNA and RNA data leading to novel subtyping with noticeable clinical outcomes beyond classic expression profiling (Curtis et al., 2012).

The first extension of iCluster was iCluster+ (or iClusterplus), a model based on a generalized linear regression combined to the basic algorithm of iCluster. Compared to different models, iCluster + produces the best classification when integrating unknown datasets. It has been useful to construct two molecular subtypes and identified two core genes (*CNTN4* and *RFTN1*) for lung adenocarcinoma (Zhao et al., 2021). Two limitations to this tools can be pinpointed: it needs to test hundreds of values to tune the optimal solution parameters and there is no evaluation of statistical significance for a selected feature (Sathyanarayanan et al., 2020).

The most recent upgrade is iClusterBayes, which uses a Bayesian integrative clustering algorithm. The main advances of iClusterBayes are to overcome the limitations of iCluster + that were priorly exposed. This tool was tested on kidney cancer and glioblastoma and grants the possibility for tumor subtyping. However, iClusterBayes was not compared to another method, so that its capacity to discover novel subtypes could not be assessed (Mo et al., 2018).

SLIDE is a tool based on the concept of the multi-view of data, using a matrix factorization algorithm. SLIDE was developed in the continuity of JIVE, trying to investigate on shared and individual structure. Compared to JIVE, SLIDE allows the creation of partially shared scores in addition to the individual ones. SLIDE was able to upgrade breast cancer subtyping using gene, methylation, miRNA, and protein profiling (Gaynanova and Li, 2019).

The integration of multiple layers of interconnected biological elements through a wide palette of multi-omics tools available is a great opportunity to better classify biological-relevant cancer subtypes and to reduce unclassified samples. The choice of the algorithm should judiciously fit the characteristic of a given dataset to optimize the model performance.

Biomarker discovery

The tumor biology is intimately related to disease evolution and treatment response (Marusyk et al., 2020). The assimilation of clinical data as additional features to feed multilayer integrated models enables association between molecular subtypes and clinical outcome. Discovering biomarkers associated with prognosis and treatment sensitivity/resistance is a keystone for risk-group classification and therapeutic decision. The identification of treatment specific biomarkers is also granting the opportunities to provide therapies tailored to the biological trait of a specific tumor, opening the path for precision medicine. Machine learning and deep learning models, trained on Kaplan–Meier derived survival data, are powerful classifiers to predict the clinical evolution. In this part, we aim to illustrate how multi-omics integration alongside machine and deep learning approaches might facilitate biomarker discovery and guide treatment decision.

Application of multi-omics to biomarker discovery

Multi-omics tools have been developed to discover novel biomarkers in oncology. For example, jNMF allows biomarker discovery for prediction of drug response through pathway signature analysis. The identification of novel connections between tumor biology and drug response highlighted an association between BRAF inhibitors efficacy and *BRAF*/*MITF* overexpression in breast cancer (Fujita et al., 2018). Also, iClusterBayes demonstrated its ability to discover biomarkers, revealing the role of *MTAP*/*CDKN2A*/*2B* expression for PD-L1 blockade sensitivity with a proportional relation of Kaplan–Meier survival to the gene expression level (Mo et al., 2020).

Lemon-tree is another tool for multi-omics processing that runs a series of tasks that are self-contained step in the learning and clustering process. The workflow of the tools is as follow: ask biological question → preprocess data → clusterization → builds modules of cluster → compute score → results. Using this tool to discover biomarkers was operated on glioblastoma genes, looking at the amplification levels and copy-loss level of genes. The results show that genes that have copy number alteration (CNA) of glioblastoma oncogene *EGFR* and tumor suppressors *CDKN2A* and *PTEN*, but also novel candidates such as *KRIT1* and *PAOX* were assimilated with a worse prognosis (Bonnet et al., 2015). iProFun, is a method analyzing the "cascade effects" of the genes. It takes as input statistics associated with the data, aiming to detect the joint variation between each data. This study highlighted potential therapeutic candidates (*AKT1*, *KRT8*, and *MAP2*) in ovarian cancers. But it also demonstrated the role of *BIN2* in ovarian cancer, and how it can be a favorable survival outcome (Song et al., 2019). Finally, AMARETTO is used to identify driver genes by integrating genomics and epigenomics data. This is a three-step process, identifying candidate cancer driver genes, modeling effect on gene expression, and association of drivers with their

targets. In other words, it creates clusters depending on driver genes and expression of the genes. AMARETTO was able to highlight different driver genes such as *GPX2* for smoking induced cancer (lung squamous cell carcinoma), but also identified *OAS2* and *TRIM22* as modulator of the immune response (Champion et al., 2018).

Machine learning model for biomarker discovery

Machine learning are algorithms that can predict models using statistical methods based on training data. Algorithms can be trained in two ways: *via* supervised or unsupervised learning. Supervised learning relies on a labelled input dataset used to train and develop a function to predict the outcome on an experimental dataset. The algorithm attempts to find patterns relating the features to the given label. The second step is to integrate data test and to classify the validation data in the right label by following the identified patterns. The correlation between the given and predicted labels can be compared to assess the accuracy of the model. Unsupervised learning clusters the samples by identifying and regrouping different features following a similar pattern (Huang et al., 2015). The main difference of supervised learning is that the input data are not labelled. Machine learning approaches algorithms containing only one hidden layer.

A powerful tool developed for multi-omics integration is the artificial neural network (ANN). It can be used for machine learning and for deep learning. ANN is a simple approach to create an artificial model, composed of neurons organized in layers. This is composed of an input layer, hidden layer, and an output layer. The input layer corresponds to the dataset given to the algorithm and the output layer corresponds to the possible outputs of the algorithm, in this case, a classification of the data. The hidden layer is what defines the complexity of the algorithm, it represents possible pathways linking the input layer to the output layer. For the integration of multi-omics, different blocks, specific to each type of omics data, are created. Depending on the biological question, different machine learning algorithms can be preferred.

A vast diversity of algorithms are available for machine learning; random forest classifier and k-nearest neighbor are commonly supervised classifiers that are easily suitable for biomarker discovery. A random forest algorithm builds multiple decision trees that are randomly generated. Each decision tree provides a classification. The classification that is chosen in a majority of cases is used to classify each datapoint of the dataset (Tin Kam, 1995). K-nearest neighbor follows three main steps: 1) a reference dataset is clustered into the different classes that need to be distinguished, 2) the experimental data are integrated into the dataset, and 3) the experimental data are classified by calculating the distance to each defined class and classifying each datapoint in the nearest cluster (Fixt and Hodges, 2022).

These different approaches have been successfully applied to biological data to improve prognoses prediction. For example, ANN has already been used to analyze the survival in two breast cancer datasets. The model was able to predict the prognoses (favorable or unfavorable) and relapse probability (Chi et al., 2022). In the second example, machine learning approaches were compared to histopathological grading to classify and predict patients' outcome in glioblastoma. In this test, a k-nearest neighbor-based ANN approach out scaled the histopathological grading for tumor classification and survival prediction. (Petalidis et al., 2008). In the third example, a random forest algorithm allowed the discovery of a novel prognostic biomarker in the Ewing Sarcoma. As an example, a lower Ki67 expression was associated to a better prognosis for the subset of samples with low-CD99 expression (Bühnemann et al., 2014). In the final example, a gene expression-based algorithm of k-nearest neighbor and random forest identified novel prognostic genes had increased the accuracy for the classification of the poorly defined group of soft tissue sarcomas. They were also able to annotate the samples in molecular subsets with potential therapeutic susceptibility (van IJzendoorn et al., 2019).

Deep learning for treatment guiding

Deep learning approaches are a subset of machine learning that contains multiple hidden layers organized in a network allowing for progressive learning (LeCun et al., 2015; Bi et al., 2019). Deep learning can be deep neural network (DNN) or convolutional neural network (CNN). DNN is similar to ANN with an increased number of hidden layers, and CNN is a class of ANN used for imaging analysis (Simonyan and Zisserman, 2015; Szegedy et al., 2015; He et al., 2016; Chollet, 2017).

Survival analysis learning with multi-omics neural network (SALMON) attempts to aggregate and simplify the gene expression data to enable prognosis prediction. Deep learning approaches typically introduce an extensive number of parameters from a limited sample set that can introduce data overfitting rendering the models ineffective. In comparison, SALMON favorizes the use of eigengene matrices of gene co-expression modules to feed the model. This model was compared to other survival prognostic models on breast cancer datasets, including DeepSurv and random survival forest. SALMON predictor reached a higher survival concordance index than other methods and showed that the integration of multi-omics data combined with clinical and demographical features can increase the prediction performance. This algorithm was used to develop an age-specific prognostic score for breast cancer (Huang et al., 2019).

CNN is very useful for deconvolutional observations of histological images used to classify and predict cancer prognoses. CNN was successful to detect and to predict

patient outcomes on haematoxylin-eosin-stained tumor tissue microarray with a higher accuracy than the visual prognostic prediction. For that, the image was cut into spots that were characterized with a pre-trained convolutional neural network (Bychkov et al., 2018). MesoNet is another tool developed to analyze larger images. To do so, the whole-slide images are subdivided in tiles that have a score assigned. Based on the score, the algorithm can select the most relevant tiles for the prediction. The cumulated score of the selected tiles informs the prediction of a patient's overall survival (Courtiol et al., 2019). These tools perform well to analyze cancer heterogeneity. To finish, survival convolutional neural network (SCNN) is a CNN-based algorithm combined to an integration tool. This model identifies visual patterns from regions of interest isolated in biopsies and relates them to patient's outcome to create a model. This comprehensive tool can also integrate genomics biomarkers. This technology was able to develop a better prognostic prediction tool compared to WHO genomics classification (Mobadersany et al., 2018).

DrugCell is a visible neural network for drug response prediction. The model is built with two branches. The first one modulates the hierarchical organization of molecular subsystems in human cells. Each subsystem, involving small protein complexes, is connected to larger pathways and to cellular functions assigned from a bank of artificial neurons. The input layer is the mutation status from genomics data and the output corresponds to the state of the whole cell based on the genotype. The second branch of the system is a CNN assessing the fingerprint of a drug. The output from the two branches of the model is combined in a single layer of neurons integrating the response of a genotype to a certain treatment. To test on the first side of the model, they used two large drug screening resources: the Cancer Therapeutics Response Portal (CTRP) v2 and the Genomics of Drug Sensitivity in Cancer (GDSC) database. Using these two resources, it covers 684 drugs and 1,235 cell lines. Each cell-line genotype was represented with a binary vector (mutated or non-mutated). The chemical structure of the drug was represented by an average of 81 activated bits in the Morgan fingerprint vector. Drug cell was trained to associate each genotype-drug paired with its corresponding drug–response curve. As a last point, drug cell was clinically tested stratify cancer patients depending on their response to treatment. They tested drug cell on clinical trial data from 221 estrogen receptor positive metastatic breast cancer patients and showed that the model was able to predict the response to mTOR and CDK4/6 inhibitors. Drug cell also includes a feature for predicting the sensitivity to drug combination (Kuenzi et al., 2020).

In a number of studies, DNN was used to predict prognoses. As an example, DeepSurv is a tool defined as a prognostic model. With the integration of time-dependent and treatment-sensitive survival data to the model, it was trained to provide treatment recommendations. DeepSurv was tested on four real-life datasets: Worcester Heart Attack Study (WHAS), the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT), the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), and Rotterdam & the German Breast Cancer Study Group (GBSG). DeepSurv takes input as the patient's baseline data. The output is composed of one node estimating the log-risk function in the Cox model (estimation of the effect parameters without any consideration of the hazard function). As the second part, it uses a treatment recommender system. This system takes into account a group of patients assigned to a specific treatment group to predict the log-risk of a given treatment. Using the same base line hazard functions, it is possible to calculate the personal risk-ratio of prescribing one treatment and then, determine a treatment recommendation algorithm. To assess the possibility of a treatment recommendation system, DeepSurv was trained on the Rotterdam tumor bank to build the recommendation system that was then tested on GBSG. The treatment recommendation system demonstrated that, on a real clinical dataset, the patient's survival would be increased by following the treatment recommendation of the algorithm. In comparison, a random survival forest-based recommendation system was not able to improve the patient's outcome (Katzman et al., 2018). To conclude, DeepSurv provides strong modeling capabilities with the ability of guiding the therapeutic decision and bringing the clinician closer to computer-assisted technology for precision medicine in oncology.

Other examples of machine and deep learning models applicable for biomarker discovery are detailed in a recent review articles devoted to this subject (Nicora et al., 2020) (Zhu et al., 2020). So far, these powerful algorithms have been used to successfully develop models for prognosis and treatment response prediction. However, these cutting-edge technologies are not yet ready for personalized treatment recommendation in the perspective of prospective clinical implementation.

New hope in omics: deciphering host–tumor interplay

Novel fields of interest have emerged to unravel host–tumor interactions (Hui, 1989) through the immunologic cancer response (Nisar et al., 2020) (Binnewies et al., 2018) and to understand how tumor cells can trick the host to hijack its defensive system and favor cancer immune evasion. The comprehension of the immune escape mechanisms opened the path to immune-mediated therapeutic opportunities. ICB are molecules that restore the immune system's ability to recognize and eliminate tumor cells such as PD1/PDL1 (program cell-death protein one and its ligand) or CTLA4 (cytotoxic T-lymphocyte associated protein 4) inhibitors have revolutionized the treatment of some adult cancers (Nishino et al., 2017). However, not all the patients with a same disease will respond to these treatments and anticipating the patients who

will be responders remains challenging. Indeed, ICB efficacy is highly dependent of the constitution of the tumor cell immunogenicity, the host immune recognition and, ultimately, the TiME (Bonaventura et al., 2019). The advances in NGS and the growing number of high-throughput tumor sequencing datasets alongside the constant improvement of computational analyzes opened the way to new disciplines like immunomics and microbiomics.

ICB biomarker identification

The host-tumor interaction comprises a succession of interrelated steps facilitating or refraining an efficient immune recognition of tumor cells. These steps are dependent of the tumor biology, the host ability to recognize the tumor, and the mechanisms that will promote immune inflammation or immune evasion. Briefly, the somatic mutations of tumors result in the expression of neoantigens that can be exposed to and recognized by the host antigen presenting cells (APCs). This will activate an immune cascade responsible for the activation and recruitment of effector T cells in the tumor (Chen and Mellman, 2017). These primary steps determine the infiltration of the tumor by effector immune cells: the tumor infiltrated lymphocytes (TILs). The anti-tumor activity of TILs is influenced by the presence of stimulatory or inhibitory signaling. PDL1, expressed by tumor and immune cells, binds to the PD1 receptor on T cell and generates an inhibitory signal leading to T-cell exhaustion and loss of activity (Chen and Mellman, 2017). To be fully efficient, ICBs needs the tumor to be infiltrated by inflamed T cell rending possible to develop biomarkers of efficacy (Bonaventura et al., 2019).

The identification of biomarkers was first made possible by immunohistochemistry (IHC) assay that has the advantage of spatial annotation but is limited by the sparse number of features that can be assessed. A classification in three IHC immune states for tumor infiltrated lymphocytes (TILs) density is usually admitted with immune inflamed (or "hot"), immune excluded, and immune deserted (or "cold") tumors (PMID: 30,179,157). Few biomarkers are so far broadly recognized. Some from the IHC like tumor infiltrated lymphocytes (TILs) density, mostly for CD8, and the protein expression of PD-L1 in the tumor. Other biomarkers derive from NGS test: tumor mutation burden (TMB) and the level of expression of interferon-gamma (*IFNγ*) or *INFγ* gene expression profile signature (Ayers et al., 2017; Nishino et al., 2017). These biomarkers are known to be independent predictors of response to ICB, however the predictive value of cumulative biomarkers remains partially solved, so that the integration of multilayers omics analysis for dissecting the immune environment through immunomics was brought to the fore.

Immunomics for TiME profiling

Immunomics is the part of omics integration that aims to comprehend the host-tumor interaction and to profile the TiME. Integrative immunomics allows for quantitative, functional, spatial, and clonal annotation of the immune environment to comprehend the host/cancer interaction and discover biomarkers for ICB efficacy.

To integrate and study immune cell-host interactions, novel bioinformatic tools were developed. The quantification and enumeration of immune cell infiltrate can be deducted from bulk-tumor RNAseq data by deconvolution-based tools, such as MCP-counter (Becht et al., 2016) and CIBERSORT (Newman et al., 2019), or through gene expression scores (Danaher et al., 2017). It is also possible to use computational analyzes to infer the somatic mutations and the neoantigens presented by the tumor cells as well as the immunogenicity of these neoantigens, for example, the TMB assesses the mutational load of a tumor (Endris et al., 2019). The enumeration of intra-tumor T-cell and B-cell clonotype expansion, through T-cell receptor (TCR) and B-cell receptor (BCR) rearrangement, are particularly interesting as a surrogate of the immune reactivity potentially induced by the tumor. Different tools exist to extract the T-cell and B-cell clonotype repertoire from genomics or transcriptomics data, such as MiXCR (Bolotin et al., 2015) and immuneDB (Rosenfeld et al., 2018). These methods align VDJ sequences, quantify them, and sort them in distinct clonotypes. With the development of these tools, high-throughput NGS dataset can be analyzed through the spectre of immunomics and for discovering new biomarkers of ICB efficacy. We will, herein, present some of the latest studies of immunomics integration.

A pan-cancer analysis, from the publicly available TARGET dataset, aimed to characterize and classify immune subsets of pediatric solid cancers based on transcriptomics and survival data (Sherif et al., 2021). The gene expression profile and gene set enrichment analysis (ssGSEA) individualized six distinct immune groups with diagnostic and prognostic association. The study used the data from 408 samples from five pediatric cancers: neuroblastoma, osteosarcoma, and three kinds of renal cancers.

The immune infiltrate was first assessed by the immunologic constant of rejection (ICR) method (Thorsson et al., 2018) in three classes from low- to high-immune score. Kidney rhabdoid tumors had the higher immune score and Wilms tumor the lowest. Correlation clustering of ssGSEA-based immune signatures identified five main modules: interferon-gamma (IFN-G) and tumor growth factor beta (TGF-B) signaling, macrophages, lymphocytes, and wound healing. These five modules were used for the identification of six immune subtypes: T-cell helper 2 (Th2) dominant, inflammatory, immunologically quiet, wound healing dominant, macrophages dominant, and lymphocyte suppressed subtypes. Each immune subtype was composed by a specific immune

infiltrate by CIBERSORT enumeration, that mirrored the immune signature.

Tumor types were unevenly distributed between the immune groups, and thus highlighting the association of the immune phenotype and the diagnosis. Furthermore, the immune subtyping was correlated with the prognosis, inflammatory subtype having the best clinical outcome while wound healing dominant subtype had the worse survival. The clinical impact was observed across cancers and within a same cancer type. This work demonstrates that access to NGS, well clinically annotated large tumor database, and immunogenomics integration help understand the interplay of host immune system and tumor cells. The immune landscape characterization can also improve tumor classification and prognostication as shown in this pediatric database.

A systematic meta-analysis of tumor and microenvironmental biomarkers for ICB sensitivity has investigated the relative impact of independent and combined biomarkers. An immunomics analysis from bulked-tumor transcriptomics and genomics was performed on a dataset from seven cancer types from 1,008 adult patients (CPI1000+) treated with immune checkpoint blockade (ICB) (Litchfield et al., 2021). The goal of this work was to select a large panel of biomarkers, aggregated by a systematic literature search, to assign a Z score for each of them and test their prediction impact for ICB efficacy in a large pan-cancer population. The different biomarkers explored the T-cell response, the mechanisms for immune evasion and infiltration, and the host factors.

In univariate analysis, clonal and total TMB and *CXCL9* expression, a CD8 attractive chemokine, were the strongest predictors for ICB response, followed by *CD8A* expression, T-cell inflamed *INFG* gene expression, and *CD274* expression. Due to the high prevalence of TMB in the literature, researchers decided to subdivide the somatic mutation by mutations fitting in an immunogenic signature. Some signatures like dinucleotide variants, that are source of amino-acid changes and generate immunogenic epitopes, ultraviolet (UV), or tobacco mutation signatures also came out as significant determinant of ICB response. In somatic copy number analysis, two copy number anomalies were identified as positive or negative predictor of response. Surprisingly, host factors like the loss of HLA heterogeneity or HLA subtypes did not show any impact on ICB sensitivity. They also performed single-cell RNA sequencing of a reactive CD8 TILs from patients to identify T-cell intrinsic markers of ICB sensitivity and highlighted *CXCL13* and *CCR5* gene involvement. The immune cell enumeration or BCR/TCR clonality assay were not assessed in this study.

A machine learning model for multivariate analysis including the significant determinant of response was a better predictor of ICB response than TMB alone in the initial cohort and three external independent validation datasets. Also, a two-parameter biomarker model combining clonal TMB and *CXCL9* expression had a better prediction accuracy than clonal TMB alone, yet inferior to the multivariate model. The integration of multiparameter biomarkers could explain approximatively 60% of the response to ICB in the different cancer types.

This study showed that multilayers integration of immunogenomics data and multiparameter models can improve the prediction of biomarker to anticipate the response to immunotherapy. It is open the opportunity to mechanistically solve host-tumor interaction and to understand how ICB remodel the tumor environment.

Others studies of multilayer immunomics analysis depicted the role of B cell in promoting ICB efficacy (Anagnostou et al., 2020; Helmink et al., 2020). In Helmink's study, conducted on multiple cohorts of melanoma and renal cell carcinoma treated with ICB, for comparison between responders and non-responders, or immunotherapy naïve to test the prognostic impact of B-cell infiltrate. The input data comprised RNAseq, deconvolution immune cell enumeration, BCR clonality, single-cell RNA sequencing, and mass cytometry (CyTOF) for functional analysis and histological evaluation for spatial annotation. The gene expression profile showed an increase in the activation marker genes of B cells (*IFNG*, *MZB1*, *JCHAIN*, and *IGLL5*) in ICB responders. Using TRUST-algorithm for BCR clonotype enumeration, they observed an increase of clonal counts for heavy and light chains and in BCR diversity in responders. MCP-counter deconvolution algorithm also confirmed the enrichment of B cells in responders compared to non-responders. To confirm the role of B cell in ICB activity, they used IHC to assess B-cell density and to interrogate the spatial repartition of B cell and they performed functional study. The IHC confirmed an increase in B-cell density in responders and revealed a spatial organization in tertiary lymphoid structure (TLS) where $CD20^+$ B cells colocalize with $CD4^+$, $CD8^+$, and FOXP3+ T cells. A previous study similarly showed a correlation between B-cell signature and increased the expression of *CD8A* and $CD8^+$ T-cell infiltration (Griss et al., 2019). Functional characterization of tumor B cell by single-cell RNA-seq and CyTOF, first confirmed the B-cell enrichment, but also decipher a unique immune activity of B cell (increased activity in CXCR4 signaling, cytokine receptor interaction, and chemokine signaling pathways) and a switch in immune activated CXCR3+ memory B cells in responders. The use of single-cell RNAseq and spatial omics in cancer analyze allowed, in this case, to discover unsuspected novel interaction. This comprehensive approach has revealed the structural role of B cell and tertiary lymphoid structures in responses to ICB treatment.

Multi-omics integration for immune characterization has dramatically furthered the identification of biomarkers raising the possibility of personalizing ICB treatment based on the tumor immune constitution.

Microbiomics as a modulator of the immune environment

The microbiome is the community of microbial species that inhabit a human body (Bhatt et al., 2017). The study of interactions between cancer and the microbiome has gained popularity in the past few years, showing the interplay between cancer features and species colonizing the intestinal flora. The intestinal flora can, indeed, play a role on cancer (Sepich-Poore et al., 2021) by influencing the risk of cancer apparition (Song et al., 2020b) and cancer immune response, it also may be useful for cancer detection (Chen et al., 2021). Theses interactions can lead to mucosal inflammation or systemic metabolic/immune dysregulation and modulate immune responses by altering anti-cancer immunity and response to therapy (Bhatt et al., 2017; Gopalakrishnan et al., 2018; Riquelme et al., 2019; Dohlman et al., 2021; Jackson et al., 2022). In this part, we will show that immunomics and microbiomics data should be integrated into classic omics analysis to master cancer complexity.

There are different sequencing techniques to elucidate the microbiome composition. The first one is the sequencing of the bacterial 16 S ribosomal RNA (rRNA) that is abundant and specific to each species. The 16 S rRNA genes are isolated and amplified from microbiome containing samples and then sequenced for species characterization. The downside of this technique is the need of high accuracy in the primer for amplification (Janda and Abbott, 2007; Pei et al., 2010). A second technique, majorly used in microbiomics, is shotgun sequencing that uses the taxonomic, functional, and genomics profile of the bacteria to deduct the microbiome composition (Quince et al., 2017). The study of microbiome can also be done using classic omics techniques like transcriptomics (metatranscriptomics), proteomics (metaproteomics), and metabolomics used to detect dysregulation of genes, proteins, and metabolites (Franzosa et al., 2014; Bikel et al., 2015; Singhal et al., 2015; Daliri et al., 2017).

Lately, in the aim of cancer treatment, the study of microbiome allowed significant discoveries. It was first stated that antibiotics can disrupt the activity of immunotherapy inducing loss of response to immune checkpoint blockade (ICB). In murine model of melanoma, it was shown that the use of different antibiotics decreases the response of PD-1 treatment compromising its tumor effect (Routy et al., 2018). The same effect was observed in patients with non-small cell lung cancer (NSCLC), renal cell carcinoma (RCC), and urothelial carcinoma. The use of those antibiotics negatively impacted their overall survival and progression-free survival (Derosa et al., 2022) (Routy et al., 2018). This observation suggested that a balanced microbiota is necessary for inducing and maintaining the tumor immune response and prompted further evaluation of microbiome in patients treated with ICB. Microbiome exploration by metagenomics in NSCLC patients treated with ICB confirmed the impact of intestinal flora, showing an enrichment of some species in responders, notably Akkermansia muciniphila (AkM) and Firmicutes, and in non-responder (Prevotella, *Clostridium* species) (Derosa et al., 2018; Routy et al., 2018). A recent work confirmed the role of AkM in response to ICB and introduced the notion relative abundance of bacteria in cancer (Derosa et al., 2022). They demonstrated that a high abundance of AkM can improve the patient outcomes in NSCLC, but if present in high quantity, AkM will be deleterious for ICB response. This raised the hypothesis that restoring microbiota equipoise could restore the response. This was assessed by fecal microbiota transplantation of responder flora in non-responders that was able to shift from unsensitive to sensitive phenotype and restore ICB efficacy. After operating the transplantation, 16 S RNA sequencing revealed that a higher abundance of Veillonellaceae family and poor in *Bifidobacterium bifidum* increased the sensitivity to the treatment (Baruch et al., 2021).

The role of microbiomics for modeling the host-tumor interaction and the immune response to cancer is now established, however, microbiomics has not yet been routinely implemented in multi-omics analysis but is probably a key element of cancer biology. Considering the high plasticity of the microbiome constitution, longitudinal analysis of the flora should be preferred (Turnbaugh et al., 2009; Khoruts et al., 2010; Spencer et al., 2011; Kong et al., 2012). Most importantly, the gut microbiota is highly intricated with the host immune reaction to the tumor warranting its integration in the models of tumor immune environment analysis.

Improving multi-omics integration in the context of immunomics

As we have shown, combining multiple biological elements of the tumor and of the host environment is an extremely powerful way to anatomize the TiME and to discover new biomarkers. However, there is no standardized method for immunomics and most of the computational integration tools are home made. To improve the generalization of the observation and to accelerate the possibility of patient selection for personalized therapy, standardization of the method is urging.

In that sense, an available R package for computational tool for immuno-oncology biological research (IOBR) has been developed to comprehensively combine and interpret multi-omics data in the context of immuno-oncology (Zeng et al., 2021). IOBR has been built to easily integrate whole exome and RNA sequencing from bulk tumor as well as single-cell RNA sequencing and long non-coding RNA data. It consists of a four-module pipeline with a signature/deconvolution, phenotype, mutation, and model construction. The signature module can identify immune or tumor specific signatures through expression gene profile and deconvolution estimate of the immune infiltrate. The deconvolution part directly integrates CIBERSORT (Newman et al., 2019), TIMER (Li et al., 2020), MCP-Counter (Becht et al.,

2016), xCELL (Aran et al., 2017), EPIC (Racle and Gfeller, 2020), and quanTIseq (Finotello et al., 2019). The phenotype module tests a large set of immune and non-immune-related published phenotypes, and the mutation part is able to determine association between different gene alterations and signatures. Finally, this tool has a model construction module that provides robust biomarker identification and model construction from the prior modules. The utilization of standardized ready-to-use package for onco-immunomics analysis is crucial to hasten biomarker discovery and test their inter-cohort reproducibility.

Future directions to fully characterize the immune environment and its dynamic would be to implement microbiomics data to such model and to construct longitudinal models with the introduction of time dependent models (Bodein et al., 2022). The incorporation of spatial annotation and single-cell omics to explore the cancer architecture at the cell level in all its heterogeneity and to complement with functional analysis offers a new venue to elucidate the TiME (Roh et al., 2018; Helmink et al., 2020; Zheng et al., 2021).

Conclusion

Cancer is a complex and highly heterogenous disease that is yet partially solved. Despite the molecular characterization made possible by the advances of NGS technologies, many of the underlying oncogenic mechanisms remain puzzling. The integration of the multilayers of the omics elements is an avenue to further elucidate cancer biology. In this review, we highlighted the challenges of computational modelization of large multi-omics datasets and we provided some clues for how to overcome these barriers. We presented some innovative bioinformatic tools developed to enlighten the implication of all the interconnected biological elements of cancer and showed how determinant they are to refine the disease subtyping and classification, and to discover biomarkers. Multi-omics integration has also heightened the field of immunomics and microbiomics, and thus has dramatically accelerated the identification of robust biomarkers for ICB efficacy toward the development of tailored immunotherapy.

The constant modernization of the models endows analyses of increasingly larger datasets with a growing number of components. Collaborative effort for high-quality and clinically well-annotated databases, combining all the elements of high-throughput sequencing, is crucial to feed the models. Finally, the standardization of the methods would aid in replicating and confirming the results to increase the global knowledge and, ultimately, improve cancer treatments.

Author contributions

VR-C, RS, and AD wrote the manuscript. VR-C designed the figures. VR-C, RS, and AD revised the manuscript. AD and RS supervised this research.

Funding

This work was supported by the Charles Bruneau Foundation.

Acknowledgments

The authors wish to acknowledge Lucie Leclair and Elloïse Coyle for the reviewing of grammar and syntaxes.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alyass, A., Turcotte, M., and Meyre, D. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Med. Genomics* 8 (1), 33. doi:10.1186/s12920-015-0108-y

Anagnostou, V., Bruhm, D. C., Niknafs, N., White, J. R., Shao, X. M., Sidhom, J. W., et al. (2020). Integrative tumor and immune cell multi-omic analyses predict response to immune checkpoint blockade in melanoma. *Cell Rep. Med.* 1 (8), 100139. doi:10.1016/j.xcrm.2020.100139

Anderson, P., Gadgil, R., Johnson, W. A., Schwab, E., and Davidson, J. M. (2021). Reducing variability of breast cancer subtype predictors by grounding deep learning models in prior knowledge. *Comput. Biol. Med.* 138, 104850. doi:10.1016/j.compbiomed.2021.104850

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 18 (1), 220. doi:10.1186/s13059-017-1349-1

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-omics factor Analysis—A framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol. [Internet]* 14 (6), 1. doi:10.15252/msb.20178124

Arnaout, R. A., Prak, E. T. L., Schwab, N., and Rubelt, F. (2021). The future of blood testing is the immunome. *Front. Immunol.* 12, 626793. doi:10.3389/fimmu.2021.626793

Australian Pancreatic Cancer Genome InitiativeBailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A. M., et al. (2016). Genomic analyses identify

molecular subtypes of pancreatic cancer. *Nature* 531 (7592), 47–52. doi:10.1038/nature16965

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., et al. (2017). IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* 127 (8), 2930–2940. doi:10.1172/JCI91190

Barriga, V., Kuol, N., Nurgali, K., and Apostolopoulos, V. (2019). The complex interaction between the tumor micro-environment and immune checkpoints in breast cancer. *Cancers* 11 (8), 1205. doi:10.3390/cancers11081205

Baruch, E. N., Youngster, I., Ben-Betzalel, G., Ortenberg, R., Lahat, A., Katz, L., et al. (2021). Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science* 371 (6529), 602–609. doi:10.1126/science.abb5920

Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., et al. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* 17 (1), 218. doi:10.1186/s13059-016-1070-5

Berlanga, P., Pierron, G., Lacroix, L., Chicard, M., Adam de Beaumais, T., Marchais, A., et al. (2022). The European MAPPYACTS trial: Precision medicine program in pediatric and adolescent patients with recurrent malignancies. *Cancer Discov.* 12 (5), 1266–1281. doi:10.1158/2159-8290.CD-21-1136

Bertolazzi, P., Felici, G., Festa, P., and Lancia, G. (2008). Logic classification and feature selection for biomedical data. *Comput. Math. Appl.* 55 (5), 889–899. doi:10.1016/j.camwa.2006.12.093

Bhatt, A. P., Redinbo, M. R., and Bultman, S. J. (2017). The role of the microbiome in cancer development and therapy. *Ca. Cancer J. Clin.* 67 (4), 326–344. doi:10.3322/caac.21398

Bi, Q., Goodman, K. E., Kaminsky, J., and Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *Am. J. Epidemiol.* 188, 2222–2239. doi:10.1093/aje/kwz189

Bikel, S., Valdez-Lara, A., Cornejo-Granados, F., Rico, K., Canizales-Quinteros, S., Soberón, X., et al. (2015). Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiome. *Comput. Struct. Biotechnol. J.* 13, 390–401. doi:10.1016/j.csbj.2015.06.001

Binnewies, M., Roberts, E. W., Kersten, K., Chan, V., Fearon, D. F., Merad, M., et al. (2018). Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* 24 (5), 541–550. doi:10.1038/s41591-018-0014-x

Bodein, A., Scott-Boyer, M. P., Perin, O., Lê Cao, K. A., and Droit, A. (2022). timeOmics: an R package for longitudinal multi-omics data integration. *Bioinformatics* 38 (2), 577–579. doi:10.1093/bioinformatics/btab664

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., et al. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* 12 (5), 380–381. doi:10.1038/nmeth.3364

Bonaventura, P., Shekarian, T., Alcazer, V., Valladeau-Guilemond, J., Valsesia-Wittmann, S., Amigorena, S., et al. (2019). Cold tumors: A therapeutic challenge for immunotherapy. *Front. Immunol.* 10, 168. doi:10.3389/fimmu.2019.00168

Bonnet, E., Calzone, L., and Michoel, T. (2015). Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput. Biol.* 11 (2), e1003983. doi:10.1371/journal.pcbi.1003983

Brunet, J. P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.* 101 (12), 4164–4169. doi:10.1073/pnas.0308531101

Bühnemann, C., Li, S., Yu, H., Branford White, H., Schäfer, K. L., Llombart-Bosch, A., et al. (2014). Quantification of the heterogeneity of prognostic cellular biomarkers in ewing sarcoma using automated image and random survival forest analysis. *PLoS ONE* 9 (9), e107105. doi:10.1371/journal.pone.0107105

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., et al. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* 8 (1), 3395. doi:10.1038/s41598-018-21758-3

Campbell, J. D., Yau, C., Bowlby, R., Liu, Y., Brennan, K., Fan, H., et al. (2018). Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell Rep.* 23 (1), 194–212. doi:10.1016/j.celrep.2018.03.063

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., et al. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* 12 (1), 124. doi:10.1038/s41467-020-20430-7

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U. E., Seitz, H., et al. (2020). Prospects and challenges of multi-omics data integration in toxicology. *Arch. Toxicol.* 94 (2), 371–388. doi:10.1007/s00204-020-02656-y

Carbone, A. (2020). Cancer classification at the crossroads. *Cancers* 12 (4), 980. doi:10.3390/cancers12040980

Chalise, P., and Fridley, B. L. (2017). Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One* 12 (5), e0176278. doi:10.1371/journal.pone.0176278

Chalise, P., Ni, Y., and Fridley, B. L. (2020). Network-based integrative clustering of multiple types of genomic data using non-negative matrix factorization. *Comput. Biol. Med.* 118, 103625. doi:10.1016/j.compbiomed.2020.103625

Champion, M., Brennan, K., Croonenborghs, T., Gentles, A. J., Pochet, N., and Gevaert, O. (2018). Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine* 27, 156–166. doi:10.1016/j.ebiom.2017.11.028

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* 18 (1), 248–262. doi:10.1016/j.celrep.2016.12.019

Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F., and Becker, J. (2020). Evaluation of integrative clustering methods for the analysis of multi-omics data. *Brief. Bioinform.* 21 (2), 541–552. doi:10.1093/bib/bbz015

Chen, D. S., and Mellman, I. (2017). Elements of cancer immunity and the cancer-immune set point. *Nature* 541 (7637), 321–330. doi:10.1038/nature21349

Chen, H., Ma, Y., Liu, Z., Li, J., Li, X., Yang, F., et al. (2021). Circulating microbiome DNA: An emerging paradigm for cancer liquid biopsy. *Cancer Lett.* 521, 82–87. doi:10.1016/j.canlet.2021.08.036

Chi, C. L., Street, W. N., and Wolberg, W. H. (2022). *Application of artificial neural network-based survival analysis on two breast cancer datasets*, 5.

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet], Honolulu, HI, cité 10 janv 2022 (IEEE), 1800–1807. Disponible sur: http://ieeexplore.ieee.org/document/8099678/.

Chu, J., Sun, N., Hu, W., Chen, X., Yi, N., and Shen, Y. (2022). The application of bayesian methods in cancer prognosis and prediction. *Cancer Genomics Proteomics* 19 (1), 1–11. doi:10.21873/cgp.20298

Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* 25 (10), 1519–1525. doi:10.1038/s41591-019-0583-3

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2, 000 breast tumours reveals novel subgroups. *Nature* 486 (7403), 346–352. doi:10.1038/nature10983

Cyll, K., Ersvær, E., Vlatkovic, L., Pradhan, M., Kildal, W., Avranden Kjær, M., et al. (2017). Tumour heterogeneity poses a significant challenge to cancer biomarker research. *Br. J. Cancer* 117 (3), 367–375. doi:10.1038/bjc.2017.171

Dagogo-Jack, I., and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* 15 (2), 81–94. doi:10.1038/nrclinonc.2017.166

Daliri, E. B. M., Wei, S., Oh, D. H., and Lee, B. H. (2017). The human microbiome and metabolomics: Current concepts and applications. *Crit. Rev. Food Sci. Nutr.* 57 (16), 3565–3576. doi:10.1080/10408398.2016.1220913

Danaher, P., Warren, S., Dennis, L., D'Amico, L., White, A., Disis, M. L., et al. (2017). Gene expression markers of tumor infiltrating leukocytes. *J. Immunother. Cancer* 5 (1), 18. doi:10.1186/s40425-017-0215-8

de Anda-Jáuregui, G., and Hernández-Lemus, E. (2020). Computational oncology in the multi-omics era: State of the art. *Front. Oncol.* 10, 423. doi:10.3389/fonc.2020.00423

Derosa, L., Hellmann, M. D., Spaziano, M., Halpenny, D., Fidelle, M., Rizvi, H., et al. (2018). Negative association of antibiotics on clinical activity of immune checkpoint inhibitors in patients with advanced renal cell and non-small-cell lung cancer. *Ann. Oncol.* 29 (6), 1437–1444. doi:10.1093/annonc/mdy103

Derosa, L., Routy, B., Thomas, A. M., Iebba, V., Zalcman, G., Friard, S., et al. (2022). Intestinal Akkermansia muciniphila predicts clinical response to PD-1 blockade in patients with advanced non-small-cell lung cancer. *Nat. Med.* 28 (2), 315–324. doi:10.1038/s41591-021-01655-5

Devarajan, K. (2008). Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Comput. Biol.* 4 (7), e1000029. doi:10.1371/journal.pcbi.1000029

Dohlman, A. B., Arguijo Mendoza, D., Ding, S., Gao, M., Dressman, H., Iliev, I. D., et al. (2021). The cancer microbiome atlas: A pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* 29 (2), 281–298. e5. doi:10.1016/j.chom.2020.12.001

Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., et al. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS Comput. Biol.* 17 (8), e1009224. doi:10.1371/journal.pcbi.1009224

Eddy, J. A., Thorsson, V., Lamb, A. E., Gibbs, D. L., Heimann, C., Yu, J. X., et al. (2020). CRI iAtlas: An interactive portal for immuno-oncology research. *F1000Res* 9, 1028. doi:10.12688/f1000research.25141.1

Endris, V., Buchhalter, I., Allgäuer, M., Rempel, E., Lier, A., Volckmar, A., et al. (2019). Measurement of tumor mutational burden (TMB) in routine molecular diagnostics: *In silico* and real-life analysis of three larger gene panels. *Int. J. Cancer* 1, 2303–2312. doi:10.1002/ijc.32002

Fang, Z., Ma, T., Tang, G., Zhu, L., Yan, Q., Wang, T., et al. (2018). Bayesian integrative model for multi-omics data with missingness. *Bioinformatics* 34 (22), 3801–3808. doi:10.1093/bioinformatics/bty775

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., et al. (2020). *Cancer statistics for the year 2020: An overview*, 12.

Finotello, F., Mayer, C., Plattner, C., Laschober, G., Rieder, D., Hackl, H., et al. (2019). Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* 11 (1), 34. doi:10.1186/s13073-019-0638-6

Fixt, E., and Hodges, J. L. (2022). Discriminatory analysis. *Nonparametric Discrim. Consistency Prop.* 11, 1.

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* 111 (22), E2329–E2338. doi:10.1073/pnas.1319284111

Fujita, N., Mizuarai, S., Murakami, K., and Nakai, K. (2018). Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Sci. Rep.* 8 (1), 9743. doi:10.1038/s41598-018-28066-w

Garrido, F., and Aptsiauri, N. (2019). Cancer immune escape: MHC expression in primary tumours versus metastases. *Immunology* 158 (4), 255–266. doi:10.1111/imm.13114

Gaynanova, I., and Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics* 75 (4), 1121–1132. doi:10.1111/biom.13108

Gonzalez, H., Hagerling, C., and Werb, Z. (2018). Roles of the immune system in cancer: From tumor initiation to metastatic progression. *Genes Dev.* 32 (19-20), 1267–1284. doi:10.1101/gad.314617.118

Gopalakrishnan, V., Helmink, B. A., Spencer, C. N., Reuben, A., and Wargo, J. A. (2018). The influence of the gut microbiome on cancer, immunity, and cancer immunotherapy. *Cancer Cell* 33 (4), 570–580. doi:10.1016/j.ccell.2018.03.015

Goswami, S., Chen, Y., Anandhan, S., Szabo, P. M., Basu, S., Blando, J. M., et al. (2020). *ARID1A* mutation plus CXCL13 expression act as combinatorial biomarkers to predict responses to immune checkpoint therapy in mUCC. *Sci. Transl. Med.* 12 (548), eabc4220. doi:10.1126/scitranslmed.abc4220

Gou, Q., Dong, C., Xu, H., Khan, B., Jin, J., Liu, Q., et al. (2020). PD-L1 degradation pathway and immunotherapy for cancer. *Cell Death Dis.* 11 (11), 955. doi:10.1038/s41419-020-03140-2

Griss, J., Bauer, W., Wagner, C., Simon, M., Chen, M., Grabmeier-Pfistershammer, K., et al. (2019). B cells sustain inflammation and predict response to immune checkpoint blockade in human melanoma. *Nat. Commun.* 10 (1), 4186. doi:10.1038/s41467-019-12160-2

Hamamoto, R., Takasawa, K., Machino, H., Kobayashi, K., Takahashi, S., Bolatkan, A., et al. (2022). Application of non-negative matrix factorization in oncology: One approach for establishing precision medicine. *Brief. Bioinform.* 23 (4), bbac246. doi:10.1093/bib/bbac246

He, K., Zhang, X., Ren, S., and Sun, J. (2016). *Deep residual learning for image recognition* in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet], Las Vegas, NV, USA, cité 10 janv 2022 (IEEE), 770–778. Disponible sur: http://ieeexplore.ieee.org/document/7780459/.

Helmink, B. A., Reddy, S. M., Gao, J., Zhang, S., Basar, R., Thakur, R., et al. (2020). B cells and tertiary lymphoid structures promote immunotherapy response. *Naturejanv* 23577 (7791), 549–555. doi:10.1038/s41586-019-1922-8

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A. L. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Brief. Bioinform.* 22 (3), bbaa167. doi:10.1093/bib/bbaa167

Hiley, C., de Bruin, E. C., McGranahan, N., and Swanton, C. (2014). Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine. *Genome Biol.* 15 (8), 453. doi:10.1186/s13059-014-0453-8

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158 (4), 929–944. doi:10.1016/j.cell.2014.06.049

Huang, G., Huang, G. B., Song, S., and You, K. (2015). Trends in extreme learning machines: A review. *Neural Netw.* 61, 32–48. doi:10.1016/j.neunet.2014.10.001

Huang, Z., Zhan, X., Xiang, S., Johnson, T. S., Helm, B., Yu, C. Y., et al. (2019). Salmon: Survival analysis learning with multi-omics neural networks on breast cancer. *Front. Genet.* 10, 166. doi:10.3389/fgene.2019.00166

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., et al. (2010). International network of cancer genome projects. *Nature* 464 (7291), 993–998. doi:10.1038/nature08987

Hui, K. M. (1989). Re-expression of major histocompatibility complex (MHC) class I molecules on malignant tumor cells and its effect on host-tumor interaction. *Bioessays* 11 (1), 22–26. doi:10.1002/bies.950110107

Iwai, Y., Ishida, M., Tanaka, Y., Okazaki, T., Honjo, T., and Minato, N. (2002). Involvement of PD-L1 on tumor cells in the escape from host immune system and tumor immunotherapy by PD-L1 blockade. *Proc. Natl. Acad. Sci. U. S. A.* 99 (19), 12293–12297. doi:10.1073/pnas.192461099

Jackson, L. A., Wang, S. P., Nazar-Stewart, V., Grayston, J. T., and Vaughan, T. L. (2022). *Association of Chlamydia pneumoniae immunoglobulin A seropositivity and risk of lung cancer*, 5.

Jamal-Hanjani, M., Hackshaw, A., Ngai, Y., Shaw, J., Dive, C., Quezada, S., et al. (2014). Tracking genomic cancer evolution for precision medicine: The lung TRACERx study. *PLoS Biol.* 12 (7), e1001906. doi:10.1371/journal.pbio.1001906

Janda, J. M., and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J. Clin. Microbiol.* 45 (9), 2761–2764. doi:10.1128/JCM.01228-07

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* 18 (1), 24. doi:10.1186/s12874-018-0482-1

Khoruts, A., Dicksved, J., Jansson, J. K., and Sadowsky, M. J. (2010). Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. *J. Clin. Gastroenterol.* 44 (5), 354–360. doi:10.1097/MCG.0b013e3181c87e02

Knox, S. S. (2010). From « omics » to complex disease: A systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int.* 10 (1), 11. doi:10.1186/1475-2867-10-11

Kong, H. H., Oh, J., Deming, C., Conlan, S., Grice, E. A., Beatson, M. A., et al. (2012). Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res.* 22 (5), 850–859. doi:10.1101/gr.131029.111

Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the field in multi-omics research: From computational needs to data mining and sharing. *Front. Genet.* 11, 610798. doi:10.3389/fgene.2020.610798

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., et al. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 38 (5), 672–684. e6. doi:10.1016/j.ccell.2020.09.014

Lagoa, R., Marques-da-Silva, D., Diniz, M., Daglia, M., and Bishayee, A. (2020). Molecular mechanisms linking environmental toxicants to cancer development: Significance for protective interventions with polyphenols. *Seminars Cancer Biol.* 1, 1. doi:10.1016/j.semcancer.2020.02.002

Lakshmanan, V. K., Ojha, S., and Jung, Y. D. (2020). A modern era of personalized medicine in the diagnosis, prognosis, and treatment of prostate cancer. *Comput. Biol. Med.* 126, 104020. doi:10.1016/j.compbiomed.2020.104020

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (6755), 788–791. doi:10.1038/44565

Li, P., and Chen, S. (2016). A review on Gaussian process latent variable models. *CAAI Trans. Intell. Technol.* 1 (4), 366–376. doi:10.1016/j.trit.2016.11.004

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48 (1), W509–W514. doi:10.1093/nar/gkaa407

Litchfield, K., Reading, J. L., Puttick, C., Thakkar, K., Abbosh, C., Bentham, R., et al. (2021). Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* 184 (3), 596–614. e14. doi:10.1016/j.cell.2021.01.002

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* 7, 523–542. doi:10.1214/12-AOAS597

Ma, M., Chen, Y., Chong, X., Jiang, F., Gao, J., Shen, L., et al. (2021). Integrative analysis of genomic, epigenomic and transcriptomic data identified molecular subtypes of esophageal carcinoma. *Aging* 13 (5), 6999–7019. doi:10.18632/aging.202556

Ma, X., Liu, Y., Liu, Y., Alexandrov, L. B., Edmonson, M. N., Gawad, C., et al. (2018). Pan-cancer genome and transcriptome analyses of 1, 699 paediatric leukaemias and solid tumours. *Nature* 555 (7696), 371–376. doi:10.1038/nature25795

Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K. M., Distler, M. G., Zelikovsky, A., et al. (2019). Systematic benchmarking of omics computational tools. *Nat. Commun.* 10 (1), 1393. doi:10.1038/s41467-019-09406-4

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* 12 (5), 323–334. doi:10.1038/nrc3261

Marusyk, A., Janiszewska, M., and Polyak, K. (2020). Intratumor heterogeneity: The rosetta stone of therapy resistance. *Cancer Cell* 37 (4), 471–484. doi:10.1016/j.ccell.2020.03.007

Marx, V. (2020). Bench pressing with genomics benchmarkers. *Nat. Methods* 17 (3), 255–258. doi:10.1038/s41592-020-0768-1

McAllister, K., Mechanic, L. E., Amos, C., Aschard, H., Blair, I. A., Chatterjee, N., et al. (2017). Current challenges and new opportunities for gene-environment interaction studies of complex diseases. *Am. J. Epidemiol.* 186 (7), 753–761. doi:10.1093/aje/kwx227

Menyhárt, O., and Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.* 19, 949–960. doi:10.1016/j.csbj.2021.01.009

Min, E. J., Chang, C., and Long, Q. (2018). "Generalized bayesian factor Analysis for integrative clustering with applications to multi-omics data," in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) [Internet], Turin, Italy, cité 16 août 2022 (IEEE), 109–119. Disponible sur: https://ieeexplore.ieee.org/document/8631499/.

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes* 10 (2), 87. doi:10.3390/genes10020087

Mo, Q., Li, R., Adeegbe, D. O., Peng, G., and Chan, K. S. (2020). Integrative multi-omics analysis of muscle-invasive bladder cancer identifies prognostic biomarkers for frontline chemotherapy and immunotherapy. *Commun. Biol.* 3 (1), 784. doi:10.1038/s42003-020-01491-2

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K. S., and Hilsenbeck, S. G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* 19 (1), 71–86. doi:10.1093/biostatistics/kxx017

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U. S. A.* 110 (11), 4245–4250. doi:10.1073/pnas.1208949110

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., et al. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U. S. A.* 115 (13), E2970–E2979. doi:10.1073/pnas.1717139115

Needham, C. J., Bradford, J. R., Bulpitt, A. J., and Westhead, D. R. (2007). A primer on learning in Bayesian networks for computational biology. *PLoS Comput. Biol.* 3 (8), e129. doi:10.1371/journal.pcbi.0030129

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 37 (7), 773–782. doi:10.1038/s41587-019-0114-2

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., and Bellazzi, R. (2020). Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. *Front. Oncol.* 10, 1030. doi:10.3389/fonc.2020.01030

Nisar, S., Bhat, A. A., Hashem, S., Yadav, S. K., Rizwan, A., Singh, M., et al. (2020). Non-invasive biomarkers for monitoring the immunotherapeutic response to cancer. *J. Transl. Med.* 18 (1), 471. doi:10.1186/s12967-020-02656-7

Nishino, M., Ramaiya, N. H., Hatabu, H., and Hodi, F. S. (2017). Monitoring immune-checkpoint blockade: Response evaluation and biomarker development. *Nat. Rev. Clin. Oncol.* 14 (11), 655–668. doi:10.1038/nrclinonc.2017.88

Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., et al. (2010). Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl. Environ. Microbiol.* 76 (12), 3886–3897. doi:10.1128/AEM.02953-09

Petalidis, L. P., Oulas, A., Backlund, M., Wayland, M. T., Liu, L., Plant, K., et al. (2008). Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Mol. Cancer Ther.* 7 (5), 1013–1024. doi:10.1158/1535-7163.MCT-07-0177

Pierre-Jean, M., Mauger, F., Deleuze, J. F., and Le Floch, E. P. IntM. F. (2022). PIntMF: Penalized integrative matrix factorization method for multi-omics data. *Bioinformatics* 38 (4), 900–907. doi:10.1093/bioinformatics/btab786

Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., and Garmire, L. X. (2021). DeepProg: An ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* 13 (1), 112. doi:10.1186/s13073-021-00930-x

Pucher, B. M., Zeleznik, O. A., and Thallinger, G. G. (2019). Comparison and evaluation of integrative methods for the analysis of multilevel omics data: A study based on simulated and experimental cancer data. *Brief. Bioinform.* 20 (2), 671–681. doi:10.1093/bib/bby027

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35 (9), 833–844. doi:10.1038/nbt.3935

Racle, J., and Gfeller, D. (2020). Epic: A tool to estimate the proportions of different cell types from bulk gene expression data. *Methods Mol. Biol.* 2120, 233–248. doi:10.1007/978-1-0716-0327-7_17

Rajbhandari, P., Lopez, G., Capdevila, C., Salvatori, B., Yu, J., Rodriguez-Barrueco, R., et al. (2018). Cross-cohort analysis identifies a TEAD4–MYCN positive feedback loop as the core regulatory element of high-risk neuroblastoma. *Cancer Discov.* 8 (5), 582–599. doi:10.1158/2159-8290.CD-16-0861

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res.* 46 (20), 10546–10562. doi:10.1093/nar/gky889

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., et al. (2017). Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell* 171 (4), 934–949. e16. doi:10.1016/j.cell.2017.09.028

Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., et al. (2019). Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell* 178 (4), 795–806. e12. doi:10.1016/j.cell.2019.07.008

Roh, V., Abramowski, P., Hiou-Feige, A., Cornils, K., Rivals, J. P., Zougman, A., et al. (2018). Cellular barcoding identifies clonal substitution as a hallmark of local recurrence in a surgical model of head and neck squamous cell carcinoma. *Cell Rep.* 25 (8), 2208–2222. e7. doi:10.1016/j.celrep.2018.10.090

Rosenfeld, A. M., Meng, W., et al. (2018). ImmuneDB, a novel tool for the analysis, storage, and dissemination of immune repertoire sequencing data. *Front. Immunol.* 9, 2107. doi:10.3389/fimmu.2018.02107

Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P. M., Alou, M. T., Daillère, R., et al. (2018). Gut microbiome influences efficacy of PD-1–based immunotherapy against epithelial tumors. *Science* 359 (6371), 91–97. doi:10.1126/science.aan3706

Santucci, C., Carioli, G., Bertuccio, P., Malvezzi, M., Pastorino, U., Boffetta, P., et al. (2020). Progress in cancer mortality, incidence, and survival: A global overview. *Eur. J. Cancer Prev.* 29 (5), 367–381. doi:10.1097/CEJ.0000000000000594

Sathyanarayanan, A., Gupta, R., Thompson, E. W., Nyholt, D. R., Bauer, D. C., and Nagaraj, S. H. (2020). A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Brief. Bioinform.* 21 (6), 1920–1936. doi:10.1093/bib/bbz121

Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., and Knight, R. (2021). The microbiome and human cancer. *Science* 371 (6536), eabc4552. doi:10.1126/science.abc4552

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. (2019). Moli: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 35 (14), i501–i509. doi:10.1093/bioinformatics/btz318

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25 (22), 2906–2912. doi:10.1093/bioinformatics/btp543

Sherif, S., Roelands, J., Mifsud, W., Ahmed, E., Mifsud, B., Bedognetti, D., et al. (2021). The Immune landscape of pediatric solid tumors [Internet]. *Cancer Biol.* 1, 1. doi:10.1101/2021.05.04.442503

Simonyan, K., and Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. arXiv:14091556 [cs] [Internet]. 10 avr 2015 [cité 10 janv 2022]; Disponible sur: http://arxiv.org/abs/1409.1556.

Singhal, N., Kumar, M., Kanaujia, P. K., and Virdi, J. S. (2015). MALDI-TOF mass spectrometry: An emerging technology for microbial identification and diagnosis. *Front. Microbiol. [Internet]* 5, 6. doi:10.3389/fmicb.2015.00791/abstract

Skoulidis, F., and Heymach, J. V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nat. Rev. Cancer* 19 (9), 495–509. doi:10.1038/s41568-019-0179-8

Song, M., Chan, A. T., and Sun, J. (2020). Influence of the gut microbiome, diet, and environment on risk of colorectal cancer. *Gastroenterology* 158 (2), 322–340. doi:10.1053/j.gastro.2019.06.048

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Shen, H., et al. (2020). A review of integrative imputation for multi-omics datasets. *Front. Genet.* 11, 570255. doi:10.3389/fgene.2020.570255

Song, Q., Zhu, X., Jin, L., Chen, M., Zhang, W., and Su, J. (2022). Smgr: A joint statistical method for integrative analysis of single-cell multi-omics data. *Nar. Genom. Bioinform.* 4 (3), lqac056. doi:10.1093/nargab/lqac056

Song, X., Ji, J., Gleason, K. J., Yang, F., Martignetti, J. A., Chen, L. S., et al. (2019). Insights into impact of DNA copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. *Mol. Cell. Proteomics* 18 (8), S52–S65. doi:10.1074/mcp.RA118.001220

Spencer, M. D., Hamp, T. J., Reid, R. W., Fischer, L. M., Zeisel, S. H., and Fodor, A. A. (2011). Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology* 140 (3), 976–986. doi:10.1053/j.gastro.2010.11.049

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* 14, 1. doi:10.1177/1177932219899051

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet], Boston, MA, USA, cité 10 janv 2022 (IEEE), 1–9. Disponible sur: http://ieeexplore.ieee.org/document/7298594/.

Tang, S., Ning, Q., Yang, L., Mo, Z., and Tang, S. (2020). Mechanisms of immune escape in the cancer immune cycle. *Int. Immunopharmacol.* 86, 106700. doi:10.1016/j.intimp.2020.106700

Tebani, A., Afonso, C., Marret, S., and Bekri, S. (2016). Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *Int. J. Mol. Sci.* 17 (9), 1555. doi:10.3390/ijms17091555

Thompson, J., Christensen, B., and Marsit, C. (2018). Pan-cancer analysis reveals differential susceptibility of bidirectional gene promoters to DNA methylation, somatic mutations, and copy number alterations. *Int. J. Mol. Sci.* 19 (8), 2296. doi:10.3390/ijms19082296

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., et al. (2018). The immune landscape of cancer. *Immunity* 48 (4), 812–830. e14. doi:10.1016/j.immuni.2018.03.023

Tin Kam, H. (1995). "Random decision forests," in Proceedings of 3rd International Conference on Document Analysis and Recognition [Internet], Montreal, Que., Canada, cité 10 janv 2022 (IEEE Comput. Soc. Press), 278–282. Disponible sur: http://ieeexplore.ieee.org/document/598994/.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* 1A, 68–77. doi:10.5114/wo.2014.47136

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009). The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med. [Internet]* 1 (6), 1. doi:10.1126/scitranslmed.3000322

van Ijzendoorn, D. G. P., Szuhai, K., Briaire-de Bruijn, I. H., Kostine, M., Kuijjer, M. L., and Bovée, J. V. M. G. (2019). Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput. Biol.* 15 (2), e1006826. doi:10.1371/journal.pcbi.1006826

Wang, J., and éditeur (2009). *Encyclopedia of data warehousing and mining*. Second Edition. Hershey, Pennsylvania: IGI Global. [Internet]cité 16 août 2022]. Disponible sur: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-010-3.

Weber, L. M. (2019). *Essential guidelines for computational method benchmarking*, 12.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Weitschek, E., Felici, G., and Bertolazzi, P. (2012). "Mala: A microarray clustering and classification software," in 2012 23rd International Workshop on Database and Expert Systems Applications [Internet], Vienna, Austria, cité 5 janv 2022 (IEEE), 201–205. Disponible sur: http://ieeexplore.ieee.org/document/6327426/.

Witkowski, M. T., Dolgalev, I., Evensen, N. A., Ma, C., Chambers, T., Roberts, K. G., et al. (2020). Extensive remodeling of the immune microenvironment in B cell acute lymphoblastic leukemia. *Cancer Cell* 37 (6), 867–882. e12. doi:10.1016/j.ccell.2020.04.015

Wong, M., Mayoh, C., Lau, L. M. S., Khuong-Quang, D. A., Pinese, M., Kumar, A., et al. (2020). Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high-risk pediatric cancer. *Nat. Med.* 26 (11), 1742–1753. doi:10.1038/s41591-020-1072-4

Wörheide, M. A., Krumsiek, J., Kastenmüller, G., and Arnold, M. (2021). Multi-omics integration in biomedical research – a metabolomics-centric review. *Anal. Chim. Acta* 1141, 144–162. doi:10.1016/j.aca.2020.10.038

Wu, B., Lu, X., Shen, H., Yuan, X., Wang, X., Yin, N., et al. (2020). Intratumoral heterogeneity and genetic characteristics of prostate cancer. *Int. J. Cancer* 146 (12), 3369–3378. doi:10.1002/ijc.32961

Yang, Z., and Michailidis, G. (2015). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 32, 1–8. doi:10.1093/bioinformatics/btv544

Yuan, L., Zhao, J., Sun, T., and Shen, Z. (2021). A machine learning framework that integrates multi-omics data predicts cancer-related LncRNAs. *BMC Bioinforma.* 22 (1), 332. doi:10.1186/s12859-021-04256-8

Zeng, D., Ye, Z., Shen, R., Yu, G., Wu, J., Xiong, Y., et al. (2021). Iobr: Multi-Omics immuno-oncology biological research to decode tumor microenvironment and signatures. *Front. Immunol.* 12, 687975. doi:10.3389/fimmu.2021.687975

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 40 (19), 9379–9391. doi:10.1093/nar/gks725

Zhao, Y., Gao, Y., Xu, X., Zhou, J., and Wang, H. (2021). Multi-omics analysis of genomics, epigenomics and transcriptomics for molecular subtypes and core genes for lung adenocarcinoma. *BMC Cancer* 21 (1), 257. doi:10.1186/s12885-021-07888-4

Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., et al. (2021). Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* 374 (6574), abe6474. doi:10.1126/science.abe6474

Zhou, J. X., Taramelli, R., Pedrini, E., Knijnenburg, T., and Huang, S. (2017). Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. *Sci. Rep.* 7 (1), 8815. doi:10.1038/s41598-017-09307-w

Zhu, W., Xie, L., Han, J., and Guo, X. (2020). The application of deep learning in cancer prognosis prediction. *Cancers* 12 (3), 603. doi:10.3390/cancers12030603

frontiers | Frontiers in Molecular Biosciences

TYPE Perspective
PUBLISHED 13 October 2022
DOI 10.3389/fmolb.2022.974799

OPEN ACCESS

EDITED BY
Ornella Cominetti,
Nestlé Research Center, Switzerland

REVIEWED BY
Marie-Pier Scott-Boyer,
L'Institut de Radiobiologie Cellulaire et Moléculaire, France

*CORRESPONDENCE
Gary Saunders,
garysaunders@eatris.eu
Emanuela Oldoni,
emanuelaoldoni@eatris.eu

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Frontiers in Molecular Biosciences

RECEIVED 21 June 2022
ACCEPTED 08 September 2022
PUBLISHED 13 October 2022

CITATION
Oldoni E, Saunders G, Bietrix F, Garcia Bermejo ML, Niehues A, 't Hoen PAC, Nordlund J, Hajduch M, Scherer A, Kivinen K, Pitkänen E, Mäkela TP, Gut I, Scollen S, Kozera Ł, Esteller M, Shi L, Ussi A, Andreu AL and van Gool AJ (2022), Tackling the translational challenges of multi-omics research in the realm of European personalised medicine: A workshop report.
Front. Mol. Biosci. 9:974799.
doi: 10.3389/fmolb.2022.974799

Tackling the translational challenges of multi-omics research in the realm of European personalised medicine: A workshop report

Emanuela Oldoni [1]*, Gary Saunders [1]*, Florence Bietrix [1], Maria Laura Garcia Bermejo [2], Anna Niehues [3,4], Peter A. C. 't Hoen [4], Jessica Nordlund [5], Marian Hajduch [6], Andreas Scherer [7], Katja Kivinen [7,8,9], Esa Pitkänen [7,8,9], Tomi Pekka Mäkela [8,9], Ivo Gut [10], Serena Scollen [11], Łukasz Kozera [12], Manel Esteller [13,14,15,16], Leming Shi [17], Anton Ussi [1], Antonio L. Andreu [1] and Alain J. van Gool [3]

[1]European Infrastructure for Translational Medicine (EATRIS), Amsterdam, Netherlands, [2]Biomarkers and Therapeutic Targets Group, Ramon and Cajal Health Research Institute (IRYCIS), Madrid, Spain, [3]Translational Metabolomic Laboratory, Department of Laboratory Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands, [4]Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands, [5]Department of Medical Sciences, Molecular Precision Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden, [6]Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University and University Hospital in Olomouc, Olomouc, Czechia, [7]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland, [8]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki, Helsinki, Finland, [9]HiLIFE-Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland, [10]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, [11]ELIXIR Hub, Hinxton, United Kingdom, [12]Biobanking and BioMolecular Resources Research Infrastructure-European Research Infrastructure Consortium (BBMRI-ERIC), Graz, Austria, [13]Josep Carreras Leukemia Research Institute (IJC), Badalona, Spain, [14]Centro de Investigacion Biomedica en Red Cancer (CIBERONC), Madrid, Spain, [15]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain, [16]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Barcelona, Spain, [17]State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, Shanghai, China

Personalised medicine (PM) presents a great opportunity to improve the future of individualised healthcare. Recent advances in -omics technologies have led to unprecedented efforts characterising the biology and molecular mechanisms that underlie the development and progression of a wide array of complex human diseases, supporting further development of PM. This article reflects the outcome of the 2021 EATRIS-Plus Multi-omics Stakeholder Group workshop organised to 1) outline a global overview of common promises and challenges that key European stakeholders are facing in the field of multi-omics research, 2) assess the potential of new technologies, such as artificial intelligence (AI), and 3) establish an initial dialogue between key initiatives in this space. Our focus is on the alignment of agendas of European initiatives in

multi-omics research and the centrality of patients in designing solutions that have the potential to advance PM in long-term healthcare strategies.

KEYWORDS

personalised medicine, translational medicine, multi-omics, EU initiatives, research infrastructures, bottlenecks in health data

Introduction

Definition of personalised medicine and -omics technologies

The 'Personalised Medicine' (PM) field has evolved rapidly over recent years, and now plays an increasingly important role in disease prevention, diagnosis, prognosis, and the unearthing of novel therapeutics. The European Commission recently defined PM as: "a *medical model using characterization of individuals' phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention*" (European Commission, 2021). This definition has also been adopted by the European PERMIT project (Banzi et al., 2020).

The application of PM has the clear potential to aid routine clinical decision making processes based on (an) individual patient profile(s) and condition(s) in order to minimise harmful side effects, ensure a more successful outcome, assure more efficient patient management, and at the same time provide an economic advantage (Goetz and Schork, 2018; Kisor and Ehret, 2020). A strong contribution to this comes from new molecular biomarker analysis technologies, often summarised under the term -"omics" (e.g., genomics, transcriptomics, proteomics, metabolomics, radiomics, and lipidomics) (Karczewski and Snyder, 2018; Krassowski et al., 2020). Following landmark observations of genetic variants for use as stratification biomarkers to select best responding patients, including Her-2 amplification in breast cancer Pegram et al. (1998) and BRAF mutation in melanoma Davies et al. (2002), multiple examples of omics impact in PM were described, such as in oncology van 't Veer et al. (2002), diabetes Chen et al. (2012), inflammatory bowl disease Lloyd-Price et al. (2019) and rare neurometabolic disease Tarailo-Graovac et al. (2016). The resulting data enable scientists and healthcare professionals to obtain mechanistic insights in a patients' disease state and determine the correct course of action (Chen and Snyder, 2013).

However, the integration of multiple different types of -omics data to identify composite biomarker signatures is a current bottleneck in PM. This integration is further complicated when we consider the integration with other data such as imaging data, phenotypic data, medical data (Electronic Health Records (EHRs) and patient-related outcomes (Abul-Husn and Kenny, 2019). This level of integrated multi-modal omics and phenotypic (and/or health) data application facilitates a more precise understanding of disease biology. When disease mechanisms are better understood, drug/therapeutic target identification and the selection of course of treatment for a specific subgroup of patients is possible (Karczewski and Snyder, 2018; Krassowski et al., 2020). The integration of multi-omics and multi-modal data marks a significant step closer to PM (Wenk, 2005; Aslam et al., 2017; Li et al., 2017; Kalisky et al., 2018; Olivier et al., 2019), although many challenges remain before the further development and implementation of these integrated data in routine clinical care.

European workshop to discuss PM potentials, bottlenecks and challenges, and propose solutions

The EATRIS-Plus project, funded by the EU's research and innovation funding programme Horizon 2020, aims to build capabilities and deliver innovative scientific tools to support the long-term sustainability of EATRIS as one of Europe's key PM research infrastructures.

As part of this project, the EATRIS-Plus Multi-omics Stakeholder Group was established in 2020 with the intention of facilitating best practices for omics research and bringing better alignment in goals and objectives in large-scale multi-omics initiatives across Europe and beyond. Moreover, the group seeks to become a key opinion leader group in PM.

The EATRIS-Plus Multi-omics Stakeholder Group held its first virtual meeting on 4 March 2021. The group brings together close to 20 experts from world-leading European institutions and associated with EATRIS-Plus. Further expert members include key stakeholders in the multi-omics landscape, such as Ivo Gut representing the EASI Genomics initiative, Manel Esteller from the International Human Epigenome Consortium, Katja Kivinen from the 1 + MG initiative, Esa Pitkänen and Tomi Mäkelä from the iCAN Digital Precision Cancer Medicine project, and Leming Shi (Fudan University, China) representing the International Human Phenome Consortium. In addition to EATRIS, a further two European Research Infrastructures were represented by Lukasz Kozera and Michaela Mayrhofer from BBMRI and Jennifer Harrow and Serena Scollen from ELIXIR.

The focus of the workshop was to explore common bottlenecks and start a dialogue around potential areas of collaboration across participating organisations. As mentioned above, the overall aim of this multidisciplinary and

cross-institutional working group is to become a European reference group for fully implementing PM across Europe.

This first publication from the group reports the main conclusions on the identification of current bottlenecks, pitfalls and potential solutions in multi-omics research in support of PM.

Bottlenecks in the application of multi-omics to PM

Moving beyond genomics to integrated multi-omics and multi-modal complex biomarker generation

Diseases are caused by a complex combination of genetic and environmental factors. As a consequence, the uncovering of precise molecular processes by which these factors result in the disease phenotype(s) is vital, yet difficult.

Historically, genomics (e.g., using an individual patient's genotypic information) laid the foundation for PM (Sadee, 2011). PM has been successfully applied in the areas where strong genetic drivers provide an excellent platform for developing personalised approaches, for example in oncology (Berger and Mardis, 2018) and rare diseases (Alves et al., 2018). Indeed, there are nowadays many reported successes in the application of genomics to clinical care, and this portfolio of success continues to grow (Shendure et al., 2019).

However, the application of PM for other areas of disease where solely genetic factors are less of a driver, such as neurological or metabolic disorders, can be considered as in its infancy. Analysis and application of PM that goes beyond genomics alone and passes through the development and validation of integrated multi-omics biomarker signatures including all the biological layers (effectors and regulators) rather than small sets of putative biomarkers has been demonstrated (Prasser et al., 2018; Glaab et al., 2021). A significant bottleneck, a significant bottleneck in the application to PM remains the required multi-modal data integration that can exploit and integrate multiple molecular and clinical data types in order to improve our understanding of disease mechanisms, stratify patients and inform clinicians about optimised strategies for therapeutic intervention.

New technologies, new challenges, and digital health

A key element discussed during the first EATRIS-Plus Multi-omics Stakeholder Group was the role of artificial intelligence (AI) in boosting the use of multi-omics in the PM domain. AI can be harnessed to deconvolute high-dimensional data from multi-omics data profiling and resolve molecular profiles that are indicative of treatment response and/or potential drug toxicity. However, examples of the application of AI in omics analyses are scarce which optimally need sufficient multi-scale, multi-modal and longitudinal omics data to reasonably capture relationships that may exist between input and output features. Also, significant challenges are faced in the application of AI to the PM domain in the areas of interoperability, data quality and result reproducibility.

Similar challenges are faced by digital health, being a promising, multidisciplinary approach under development encompassing the use of medical technologies (wearable devices, digital healthcare programs, etc), that permit disease monitoring, management, health risk assessment and/or prevention (Jandoo, 2020). Digital health offers great potential in PM. It improves medical outcomes and enhances efficiency, empowering patients to make better-informed decisions about their own health and providing new options for the prevention, early diagnosis, and management of chronic conditions outside of traditional health care settings (Jandoo, 2020). Additionally, digital health is a potential source of data that can be useful for designing public health policies and epidemiological programs with impact in PM.

Data standardisation to enable multi-modal integration and AI supported drug modelling

Data modelling can predict many patient characteristics, but its prediction accuracy depends on the quantity and quality of available data, as well as the interoperability of the tools (algorithms, code) used. Using AI to build effective evidence-based decisions requires the collection of significant volume of complex standardised data (multi-omics, imaging, EHRs, etc.) that need to be reliable. The volume of biomedical data being collected has increased exponentially over the past years, but these data are not always readily available for AI-based approaches. This is partly due to the sensitive nature of clinical phenotype data and to the difficulties in obtaining standardised and structured experimental datasets and information from EHRs and research databases (Panahiazar et al., 2015; Abul-Husn and Kenny, 2019; Conesa and Beck, 2019).

Mathematical models should be both flexible and dynamic; as data is continuously provided the model should improve and be more accurate. However, data that is not of high quality will produce results that are not actionable or insightful, and that can even be misleading and useless in clinical practice. Therefore, high-quality multi-omics, clinical, and epidemiological data are fundamental for generating, establishing, and sustaining algorithms that are sufficient for application.

In order to increase quality, standardisation, and reusability of scientific data, and specifically for these data to be machine

actionable, the FAIR principles were published in 2016 (Wilkinson et al., 2016). For data to be truly FAIR (Findable, Accessible, Interoperable, Reusable), the principles need to be applied to data at source, including information relating to samples, experimental methods, and data analyses. This is imperative to ensure the required results' reproducibility needed for the application of AI in the field of PM and to promote the reuse of multi-omics data in patient management (Wilkinson et al., 2016; Abul-Husn and Kenny, 2019; FAIR, 2021).

The FAIR data principles have provided a valuable route forward to the standardisation of data enabling the application of AI in PM. However, due to varying data qualities and multiple different standards used across the landscape, heterogeneity and reduced interoperability, e.g., a timely and secure access, is still common. Integration and use of EHR data so that it can be used to optimise health outcomes for individuals and populations is difficult, and the true application of multi-modal data to PM remains a challenge. Wider adoption of health data standards and models such as the Fast Healthcare Interoperability Resources (FHIR), the Clinical Data Interchange Standards Consortium (CDISC) and the Observational Medical Outcomes Partnership (OMOP), and continuing efforts to map these data models to each other, is needed (Fischer et al., 2020).

Variability in omics data at source

A key factor for the quality and reusability of multi-omics and clinical data is the availability and feasibility of relevant quality assurance (QA) and quality control (QC) schemes for laboratories in the field. Limited participation in QA/QC schemes due to budget or other constraints leads to decreased harmonisation of sample and data processing methods across laboratories, potentially resulting in lower reliability of analytical results (Freedman et al., 2015). Although standardisation of genomics data is improving (Endrullat et al., 2016; Lubin et al., 2017; Corpas et al., 2018) for other -omics data and EHRs there is a distinct lack of common standards and/or reference benchmark values for assessing complex tests and for determining clinical validity and utility of, for example, biomarkers (Quackenbush, 2004; Simons, 2018; Veenstra, 2021).

Many multi-omics focussed initiatives, such as the EATRIS-Plus project are aimed at tackling the challenges of data interoperability and increasing data FAIRness with the delivery of standard operating procedures (SOPs) that are ready to be implemented by the scientific and clinical communities, enabling standardisation among methods and technical controls in order to increase results reproducibility and improve reliability of the techniques.

An additional critical consideration for the application of multi-omics and clinical data to PM is population diversity. As the biological determinants of health are strongly influenced by environmental and sociocultural factors, and European populations are characterised by genetic and biological diversity, population-tailored reference values for multi-omic and clinical data are required. Currently there is a need of large cohorts representing human population genotypes diversity that can be analysed in order to deliver accurate reference values. The EATRIS-Plus project is aiding this situation by providing a proof of concept with a focus on the Caucasian population (EATRIS, 2022a).

Data privacy and regulatory aspects, and economic implications

Many data stewardship aspects provide significant challenges to data privacy, for example in the areas of data harmonisation and data curation, use of Common Data Models, standardised nomenclature and data transfer specifications. Such challenges and data management decisions have critical implications for data access both within and outside of jurisdictional regions and, for example, on design, creation and implementation of extract transform load workflows enabling data source maintenance and governance, quality assessment and testing.

Additionally, general concerns around data privacy and regulatory compliance-related restrictions as well as ethical and legal aspects must not be overlooked. When working with multi-omics and/or clinical data there are multiple data security, ethical, and personal information barriers that can present potential roadblocks (Knowles et al., 2017; Adamo et al., 2018; Adamo et al., 2020). Moreover, each European country has its own national implementations of General Data Protection Regulation (GDPR) for processing personal data (Vlahou et al., 2021).

The regulatory framework of PM is still emerging, and the lack of clear regulation continues to discourage investment in the field, especially since developing and implementing personalised approaches is costly - as described in the 2020 report on the current state of PM from the Personalised Medicine Coalition (PMC). In fact, coverage and payment policies both in the public and private sectors play an important role in ensuring patient access and encouraging continued innovation. To tackle the rising health care costs, often policy makers and payers do not promote PM (Kisor and Ehret, 2020). However, the costs of PM could be justified as an investment: PM approaches supporting disease prevention and identifying best therapeutic treatment will likely improve patient life quality and reduce cost in the healthcare management (Gavan et al., 2018). In fact, PM could be considered as social responsibility since most of us will become patients in our lifetimes.

TABLE 1 EATRIS project portfolio in the Data field.

Project	Scope	Goals
Beyond 1 million Genomes (B1MG)	Access to the genome information of at least one million European citizens for joint European research by 2022	• To make the genome information of at least one million European citizens accessible for joint European research as if it were one large cohort, while the data will be made accessible using a federated infrastructure
EATRIS-Plus	To build further capabilities and deliver innovative scientific tools to support the long-term sustainability strategy of EATRIS as one of Europe's key research infrastructures for PM	• To develop a multi-omic toolbox to support cross omic analysis and data integration in clinical samples
HealthyCloud	To support the creation of a European Health Data Space	• To deliver a Strategic Agenda including a Ready-to-implement Roadmap for the European Health Data Space ecosystem • The project has been organized around four fundamental objectives that cover: • interactions with stakeholders to ensure their voices are included as part of the Strategic Agenda • the inclusion of Ethical, Legal and Societal aspects in the design of the future Health Research and Innovation Cloud (HRIC) ecosystem • the sustainable access, use and re-use of health-related data considering a progressive adoption of the FAIR principles • the technological solutions in terms of computational facilities and mechanisms to enable distributed health data analysis across Europe
EOSC-Life	To create an open collaborative digital space for life science	• To publish 'FAIR' data and a catalogue of services provided by participating RIs for the management, storage and reuse of data in the European Open Science Cloud (EOSC) • To implement workflows across disciplines and address the needs of interdisciplinary science • To address the data policies needed for human research data under GDPR
EOSC-Future	To demonstrate an operational EOSC Platform ('System of Systems') with an integrated execution environment consisting of data, professionally provided services, and open research products and infrastructure that will be accessed and used by the European researchers	• To realise a EOSC-Core and EOSC-Exchange with interoperable data and resources • To allow the integration of data and resources from the Science Cluster communities into the EOSC Platform • To involve users in the co-design and implementation of the EOSC Platform
BY-COVID	To connect well-established data resources and deliver access to heterogeneous yet interlinked and organised data across domains and jurisdictions via the components of the COVID-19 Data Platform (https://www.covid19dataportal.org/)	• To create a flexible and interlinked core of FAIR data capable of addressing the constantly evolving questions during a pandemic

Tackling the challenges for implementation of PM in routine clinical care

As for all novel technologies, for assuring an effective implementation in clinical care a series of factors need to be taken in account: 1) the benefits, 2) the risks, 3) associated ethical and social aspects and 4) room for innovation. The integration of these four components requires the strong and effective communication between all stakeholders involved in the PM pipeline in order to successfully tackle all challenges. In particular, patients should be placed at the centre and empowered working in concert with researchers, clinicians, industry, and regulators. Patients should be directly involved in research, actively participating in projects, and they should take control as much as possible of their treatment.

It is true to say that healthcare professionals cannot manage a patient properly without taking into account his or her value or the patient's lifestyle, an important aspect of PM. This is not always straightforward and requires a cultural change that has already started. EATRIS is contributing to this transformation via different initiatives involving the European Patients' Academy on Therapeutic Innovation (EUPATI) and the European Patient Forum (EPF) (EATRIS, 2021; EATRIS, 2022b; EATRIS, 2022c).

Furthermore, the EATRIS Data Pillar, a key structure of the EATRIS strategy, is devoted not only to the harmonisation of data management and application but also to the unveiling and/or designing of new pathways for uncovering novel therapeutic strategies, with patients at the centre of considerations, continuing to ensure patient privacy. This is reflected in the project portfolio of EATRIS data, empowering the community to drive towards data standardisation, validation, and

TABLE 2 European initiatives focused on omics research for PM.

Initiative	Scope	Goals
EATRIS-Plus	To build further capabilities and deliver innovative scientific tools to support the long-term sustainability strategy of EATRIS as one of Europe's key research infrastructures for PM	• To develop a multi-omic toolbox to support cross omic analysis and data integration in clinical samples • To drive patient empowerment through active involvement in the infrastructure's operations • To expand strategic partnerships with research infrastructures and other relevant stakeholders
1 + Million Genomes (1 + MG)	Access to the genome information of at least one million European citizens for joint European research by 2022	• To make the genome information of at least one million European citizens accessible for joint European research as if it were one large cohort, while the data will be made accessible using a federated infrastructure
Beyond 1 million Genomes (B1MG)	To make it easier to share human health data around Europe. The project provides coordination and support to 1 + MG.	• To create the infrastructure, the legal guidance and the best practices to enable cross border genetic and phenotypic data access
International Human Epigenome Consortium (IHEC)	To provide free access to high-resolution References human epigenome maps for normal and disease cell types to the research community	• To coordinate the production of References maps of human epigenomes for key cellular states relevant to health and diseases • To coordinate rapid distribution of the data to the entire research community with minimal restrictions, to accelerate translation of this new knowledge into health and diseases • To coordinate the development of common bioinformatics standards, data models and analytical tools to organize, integrate and display whole epigenomic data generated from this important international effort
iCAN Digital Precision Cancer Medicine project	To improve outcomes and quality of life of cancer patients	• To integrate tumor molecular profiling and patients' health data • To improve cancer diagnostics and treatments • To accelerate world-class scientific innovation with the patient in focus
X-omics	To establish an integrated multi-omics research infrastructure across Netherlands with expertise in molecular biology research (genomics, proteomics, metabolomics, data integration and analysis and their combination)	• To advance X-omics technologies far beyond state-of-art • To realize an integrated X-omics infrastructure in Netherlands
PERMIT	To develop recommendations for robust and reproducible personalised medicine research	To develop recommendation for • the application and types (supervised or unsupervised) of different stratification algorithms, and the robustness and validation of the stratification methods • translational research establishing a link between data-driven stratification and the choice of treatment options • randomised clinical trials needed to test treatment strategies for each of the identified patient clusters, and to test the added value of the personalised approach vs non-personalised standard of care
Deutsche COVID-19 OMICS Initiative (DeCOI)	To use NGS-based omics data in COVID-19 research	• To establish an infrastructure addressing short-term, but also mid- and long-term challenges of the current pandemics • To prepare the NGS sector in Germany for future threats
NeurOmics	To revolutionise diagnostics and develop new treatments for ten major neuromuscular and neurodegenerative diseases	To use the most sophisticated -omics technologies in order to • increase the number of patients with a genetic diagnosis • develop biomarkers for clinical application • improve understanding of pathophysiology and identify drug targets • identify disease modifiers • develop targeted therapies • translate findings to other, related disease groups
Personal Genome Project United Kingdom	To provide open genome, trait, and health data	• Open access data to enable the timely development of tools for personalised medicine and provide a resource for advancing research

(Continued on following page)

TABLE 2 (*Continued*) European initiatives focused on omics research for PM.

Initiative	Scope	Goals
ICPerMed	To provide a platform to initiate and support communication and exchange on personalised medicine research, funding and implementation	• To contribute to the reasonable and fair implementation of personalised medicine approaches into the health systems for the benefit of patients, citizens and society as a whole • To provide a flexible framework for cooperation between member organisations
The Personal Health Train Network	To learn from each other's experiences and solutions. Netherlands PHT network is actively engaged in promoting the FAIR movement, such as the GO FAIR implementation network and committed to the principles for development of the PHT as outlined in the PHT Manifesto	• Promoting data FAIRification
EASI Genomics	To provide easy and seamless access to cutting-edge DNA sequencing technologies within a framework compliant with ethical and legal requirements, as well as FAIR and secure data management	• To build an infrastructure for enabling omics analyses (genomics, transcriptomics, epigenomics, metagenomics, immunogenomics, etc.)
IMPACT	Strategic Action designed to offer services to the Spanish R&D&I landscape, oriented to Precision Medicine, through 3 programs	• To promote generation and transfer of high-quality knowledge to the National Health System
	• Predictive Medicine	
	• Data Science	• To ensure excellence in science and technology
	• Genomic Medicine	• To assure equity and efficiency in the use of available resources

reproducibility to deliver transformative revolution in the translational medicine domain (Table 1).

Moreover, the adaption of the regulatory, ethical and legal landscape for facilitating the exploitation of patient data in the context of PM is simultaneously occurring across many large European organisations. For example, the EATRIS regulatory service and support centre is available to guide key stakeholders through this complex world, especially for complex and hybrid products for which clear regulatory guidance may not be available.

European cooperation for tackling the challenges of multi-omics in the realm of PM

In order to overcome the aforementioned challenges of PM implementation several national and international initiatives are working towards providing solutions in the further development and implementation of multi-omics research (Table 2). However, a strong synergy between such initiatives is needed to further ensure successful outcomes and to tackle any potential defragmentation (Van Gool et al., 2017). Despite this commonly understood need for alignment and cooperation among ongoing initiatives, misalignment of agendas, priorities and deliverables are potential obstacles that must be understood in order to overcome.

In Europe, Research Infrastructures (RIs) can and do facilitate this process. In particular, the Alliance of Medical Research Infrastructures (AMRI, https://eu-amri.org/), consisting of EATRIS (the RI for translational medicine, https://eatris.eu/), ECRIN (the RI for clinical research, https://www.ecrin.org/) and BBMRI (the RI for biobanking https://www.bbmri-eric.eu/) works to support the development of PM by expanding strategic partnerships with relevant stakeholders and expediting interactions. Because of the high-level of participation from countries all over Europe, AMRI allows an alignment of research activities at pan-European scale, and even beyond.

One focus of AMRI is to ensure and implement a common quality framework across the multi-omics domain. The AMRI RIs are already committed to quality in science (Freedman et al., 2015) and lead various actions addressing reproducibility, best practice guidelines, benchmarking, standards and reference materials for generation, sharing and management of multi-omics data and metadata. To date, a lot of effort has been put into enabling genomics as part of the PM pipeline (Sadee, 2011; Berger and Mardis, 2018). In particular, the European Commission driven 1 + Million Genomes initiative (1 + MG) aims to enable access to the genomic information of at least one million European citizens for joint European research as if it were one large cohort, while the data remains safely stored locally (Saunders et al., 2019). To support this effort, the Beyond 1 Million Genomes (B1MG) project is creating the federated infrastructure, including shared legal guidance and best practices for cross-borders data access. Additionally, EASI Genomics, provides easy and seamless access to cutting-edge DNA sequencing technologies to researchers from academia and industry.

However, to go beyond genomics and truly integrate multi-omics in the PM pipeline requires the facilitation of the process of data sharing, federated data analysis and

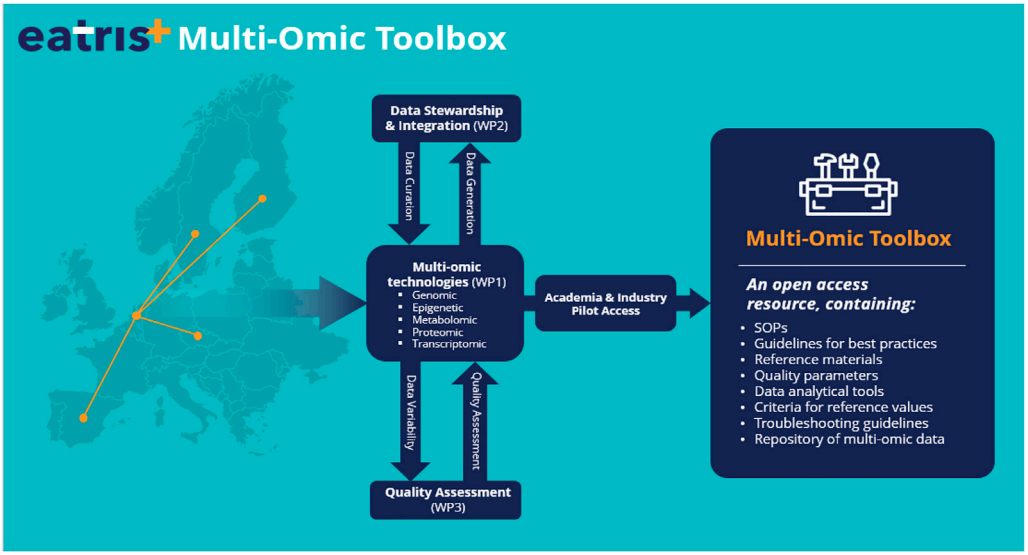


FIGURE 1
The EATRIS+ multi-omics toolbox. The multi-omics toolbox will be developed and tested with a real-setting demonstrator, an already established cohort of 1,000 healthy individuals in Czechia upon whom genomic sequencing has been already performed. Information available on this healthy individual cohort will be augmented during the project with transcriptomic, proteomic and metabolomic data. By providing such toolbox to the research community, EATRIS-Plus will be the engine to enable high-quality research in the context of patient stratification and accelerate the implementation of Personalised Medicine solutions. EATRIS is the European Infrastructure for Translational Medicine providing services for accelerating biomedical innovation.

TABLE 3 Summary of challenges and recommendations from the EATRIS + multi-omics workshop (March 2021).

Challenge	Recommendation
Moving beyond genomics	• Communicate and educate on the pros and cons of other omics technologies such as proteomics, metabolomics and lipidomics
	• Develop multi-modal data integration models that showcase the added value of multi-omics approaches in Personlized Medicine
New technologies, new challenges	• Share lessons-learned, failures and successes when evaluating new technologies in Personalized Medicine
	• Evaluate the added value of Artificial Intelligence and Digital health in Personalized Medicine, particularly in combination with multi-omics data
Data standardisation	• Adopt international standards of health data and models including the FAIR principles of data stewardship (e.g., OMOP, FHIR, CDISC)
	• Define criteria for quantity, quality and FAIR levels of data prior to multi-modal data analyses for a specific objective in Personalized Medicine
	• Work with flexible and dynamic mathematical models to adapt to changing data collections in Personalized Medicine
Variability in omics data at source	• Use internationally recognised laboratory standards and standard operating procedures for omics analyses
	• Adopt and apply quality assurance and control schemes for laboratories, such as the EATRIS Certificate of Commitment to Quality
	• Include confounding factors such as population diversity in biological systems in the multi-modal data analysis
Data privacy and regulatory aspects	• Consider ethical, legal, societal aspects when designing multi-omics Personalized Medicine studies
	• Comply with international standards on data security, including the General Data Protection Regulation in personal data
	• Report of the successes and failures of implementations from the European landscape
Implementation of Personalized Medicine in routine clinical care	• Consider well prior to multi-omics Personalized Medicine implementation: 1) the benefits, 2) the risks, 3) associated ethical and social aspects, 4) room for innovation

integration for other omics technologies. In this regard, EATRIS-Plus (EATRIS, 2022a) is developing a multi-omic toolbox (Figure 1) to support data integration and joint analysis in clinical samples. By providing such a toolbox to the research community, EATRIS-Plus will act as an engine to enable high-quality research in the context of patient stratification and accelerate the implementation of PM solutions.

There are several similarly focussed initiatives looking to blossom the integration of multi-omics in the PM pipeline at the national level. For example, in Netherlands, the X-omics initiative has established a national research infrastructure consisting of several facilities (genomics, proteomics, metabolomics, and data analysis, integration and stewardship), with the aim of generating data that is FAIR at source and ready for multi-omics integration in a customizable cloud-based digital research environment.

In Finland, the iCAN project is developing a platform for enabling the integration of cutting-edge molecular profiling information from tumours with rich longitudinal health data.

In Spain, the IMPACT initiative in PM are favouring the building of national platforms focused on the implementation of omics techniques and data exploitation in daily clinical practice, through the Health Research Institutes which are members of EATRIS.

Although the outcomes of such initiatives are undoubtably useful and have an impact in PM development and implementation, rapid development also requires fast and flexible ethical, legal and regulatory policy making as well as tackling some technical challenges (Adamo et al., 2018; Misra et al., 2019). To support healthcare providers and patients with new tools, it is crucial to facilitate data access, pilot studies for PM, and incorporate learned lessons into future policymaking.

Common strategies for the implementation of omics technologies in the PM field should be developed in the early stages of projects, even considered during the design of the call topics and proposals, with all relevant stakeholders (researchers, clinicians, patients, regulators, funders), efficiently communicating in order to align agendas and priorities. Stepping into already running processes limits the potential of common understanding and therefore only with a close collaboration from the start, consortia will be able to truly efficiently and effectively support PM development and deliver impactful and transformational solutions. Common objectives, milestones and achievements need to be defined and should always be considered and oriented from the citizens and patient perspectives. Finally, it is fundamental that policymakers engage to bridge the gap between science, medicine and the policy agenda, since we all are eventual patients of our combined European healthcare systems.

Conclusion

The recent advances in -omics technologies and their integration holds great promise for further development and implementation in the PM pipeline in order to revolutionise European healthcare. This journey is still in its infancy and many complex challenges and issues must be understood and addressed before the true benefits of PM can be seen in full implementation at the clinical setting. The sharing of knowledge on multi-omics capabilities, challenges and potential solutions is imperative for this field to mature and evolve. Here we describe how the EATRIS-Plus Multi-omics Stakeholder Group workshop has brought together relevant stakeholders to work towards PM implementation and commitment to achieve this goal. Key observations are summarized in Table 3.

An improved environment for innovation and for the integration of -omics requires a cultural and educational shift to be embraced by the entire scientific community. One of the biggest challenges will be to convince citizens, patients, healthcare communities and national regulators to allow the sharing of personal, clinical and multi-omics data to enable and accelerate PM.

Data quality and result reproducibility, a good cooperation and communication between multi-omics consortia, and an alignment with the policy agenda, are all essential aspects for facilitating the translation of multi-omics-related discoveries from bench to clinic and only following this approach we will be able to make PM an accessible reality, where the European citizen and patient is at the centre.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

The first draft of the manuscript was written by EO and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement N. 871096.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abul-Husn, N. S., and Kenny, E. E. (2019). Personalized medicine and the power of electronic health Records. *Cell* 177, 58–69. doi:10.1016/J.CELL.2019.02.039

Adamo, J. E., Bienvenu Ii, R. V., Dolz, F., Liebman, M., Nilsen, W., and Steele, S. J. (2020). Translation of digital health technologies to advance precision medicine: Informing regulatory science. *Digit. Biomark.* 4, 1–12. doi:10.1159/000505289

Adamo, J. E., Bienvenu, R. V., Fields, F. O., Ghosh, S., Jones, C. M., Liebman, M., et al. (2018). The integration of emerging omics approaches to advance precision medicine: How can regulatory science help? *J. Clin. Transl. Sci.* 2, 295–300. doi:10.1017/CTS.2018.330

Alves, I. T. S., Condinho, M., Custódio, S., Pereira, B. F., Fernandes, R., Goncalves, V., et al. (2018). Genetics of personalized medicine: Cancer and rare diseases. *Cell. Oncol.* 41, 335–341. doi:10.1007/S13402-018-0379-3

Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., and Rasool, M. H. (2017). Proteomics: Technologies and their applications. *J. Chromatogr. Sci.* 55, 182–196. doi:10.1093/CHROMSCI/BMW167

Banzi, R., Gerardi, C., and Fratelli, M. (2020). Methodological approaches for personalised medicine: Protocol for a series of scoping reviews. *Protocol* 2. doi:10.5281/ZENODO.3770937

Berger, M. F., and Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nat. Rev. Clin. Oncol.* 15, 353–365. doi:10.1038/S41571-018-0002-6

Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y. K., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148 (6), 1293–1307. doi:10.1016/j.cell.2012.02.009

Chen, R., and Snyder, M. (2013). Promise of personalized omics to precision medicine. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 5, 73–82. doi:10.1002/WSBM.1198

Conesa, A., and Beck, S. (2019). Making multi-omics data accessible to researchers. *Sci. Data* 6, 251–254. doi:10.1038/s41597-019-0258-4

Corpas, M., Kovalevskaya, N. V., McMurray, A., and Nielsen, F. G. G. (2018). A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS Comput. Biol.* 14, e1005873. doi:10.1371/JOURNAL.PCBI.1005873

Davies, H., Bignell, G. R., Cox, C., Watt, S., Andrew, P., Hall, S., et al. (2002). Mutations of the BRAF gene in human cancer. *Nature* 417 (6892), 949–954. doi:10.1038/nature00766

EATRIS (2022). EATRIS and EUPATI sign collaboration agreement, ensuring stronger patient training in translational research - EATRIS. Available at: https://eatris.eu/news/eatris-and-eupati-sign-collaboration-agreement-ensuring-stronger-patient-training-in-translational-research/(accessed Apr 25, 2022).

EATRIS (2022). EATRIS and the European Patients' Forum sign collaboration agreement, ensuring stronger patient involvement throughout the research process. Available at: https://eatris.eu/wp-content/uploads/2019/07/EPF_EATRIS_Collaboration_Release.pdf (accessed Apr 25, 2022).

EATRIS (2022). EATRIS-plus - flagship in personalised medicine. Available at: https://eatris.eu/projects/eatris-plus/(accessed Apr 25, 2022).

EATRIS (2021). Patient engagement open Forum 2021. Available at: https://eatris.eu/events/patient-engagement-open-forum-2021/(accessed Apr 25, 2022).

Endrullat, C., Glökler, J., Franke, P., and Frohme, M. (2016). Standardization and quality management in next-generation sequencing. *Appl. Transl. Genom.* 10, 2–9. doi:10.1016/J.ATG.2016.06.001

European Commission. 2021 Personalised medicine | public health. Available at: https://ec.europa.eu/health/human-use/personalised-medicine_en (accessed Dec 20, 2021).

FAIR (2021). FAIR principles - go. Available at: https://www.go-fair.org/fair-principles/(accessed Dec 20, 2021).

Fischer, P., Stöhr, M. R., Gall, H., Michel-Backofen, A., and Majeed, R. W. (2020). Data integration into OMOP CDM for heterogeneous clinical data collections via HL7 FHIR bundles and XSLT. *Stud. Health Technol. Inf.* 270, 138–142. doi:10.3233/SHTI200138

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biol.* 13, e1002165–e1002169. doi:10.1371/JOURNAL.PBIO.1002165

Gavan, S. P., Thompson, A. J., and Payne, K. (2018). The economic case for precision medicine. *Expert Rev. Precis. Med. Drug Dev.* 3, 1–9. doi:10.1080/23808993.2018.1421858

Glaab, E., Rauschenberger, A., Banzi, R., Gerardi, C., Garcia, P., and Demotes, J. (2021). Biomarker discovery studies for patient stratification using machine learning analysis of omics data: A scoping review. *BMJ Open* 11, e053674. doi:10.1136/bmjopen-2021-053674

Goetz, L. H., and Schork, N. J. (2018). Personalized medicine: Motivation, challenges and progress. *Fertil. Steril.* 109, 952–963. doi:10.1016/J.FERTNSTERT.2018.05.006

Jandoo, T. (2020). WHO guidance for digital health: What it means for researchers. *Digit. Health* 6, 2055207619898984. doi:10.1177/2055207619898984

Kalisky, T., Oriel, S., Bar-Lev, T. H., Ben-Haim, N., Trink, A., Wineberg, Y., et al. (2018). A brief review of single-cell transcriptomic technologies. *Brief. Funct. Genomics* 17, 64–76. doi:10.1093/BFGP/ELX019

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19, 299–310. doi:10.1038/nrg.2018.4

Kisor, D., and Ehret, M. (2020). The personalized medicine report. *Oppor. Challenges Future*. https://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PMC_The_Personalized_Medicine_Report_Opportunity_Challenges_and_the_Future.pdf.

Knowles, L., Luth, W., and Bubela, T. (2017). Paving the road to personalized medicine: Recommendations on regulatory, intellectual property and reimbursement challenges. *J. Law Biosci.* 4, 453–506. doi:10.1093/JLB/LSX030

Krassowski, M., Das, V., Sahu, S. K., and Misra, B. (2020). State of the field in multi-omics research: From computational needs to data mining and sharing. *Front. Genet.* 11, 1598. doi:10.3389/fgene.2020.610798

Li, B., He, X., Jia, W., and Li, H. (2017). Novel applications of metabolomics in personalized medicine: A mini-review. *Molecules* 22, E1173. doi:10.3390/MOLECULES22071173

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Poon, W., Holly, C., Kevin, S., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569 (7758), 655–662. doi:10.1038/s41586-019-1237-9

Lubin, I. M., Aziz, N., Babb, L. J., Ballinger, D., Bisht, H., Church, D. M., et al. (2017). Principles and recommendations for standardizing the use of the next-generation sequencing variant file in clinical settings. *J. Mol. Diagn.* 19, 417–426. doi:10.1016/J.JMOLDX.2016.12.001

Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: Tools, advances and future approaches. *J. Mol. Endocrinol.* 62, R21–R45. doi:10.1530/JME-18-0055

Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D., and Cox, L. A. (2019). The need for multi-omics biomarker signatures in precision medicine. *Int. J. Mol. Sci.* 20, 4781. doi:10.3390/IJMS20194781

Panahiazar, M., Taslimitehrani, V., Jadhav, A., and Pathak, J. (2015). "Empowering personalized medicine with big data and semantic web technology: Promises, challenges, and use cases," in Proc. IEEE Int. Conf. Big Data, Washington, DC, USA, 27-30 October 2014, 790–795. doi:10.1109/BIGDATA.2014.7004307

Pegram, M. D., Pauletti, G., and Slamon, D. J. (1998). HER-2/neu as a predictive marker of response to breast cancer therapy. *Breast Cancer Res. Treat.* 52 (1-3), 65–77. doi:10.1023/a:1006111117877

Prasser, F., Kohlbacher, O., Mansmann, U., Bauer, B., and Kuhn, K. A. (2018). Data integration for future medicine (DIFUTURE). *Methods Inf. Med.* 57, e57–e65. doi:10.3414/ME17-02-0022

Quackenbush, J. (2004). Data standards for 'omic' science. *Nat. Biotechnol.* 22, 613–614. doi:10.1038/nbt0504-613

Sadee, W. (2011). Genomics and personalized medicine. *Int. J. Pharm.* 415, 2–4. doi:10.1016/J.IJPHARM.2011.04.048

Saunders, G., Baudis, M., Becker, R., Beltran, S., Beroud, C., Birney, E., et al. (2019). Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* 20, 693–701. doi:10.1038/S41576-019-0156-9

Shendure, J., Findlay, G. M., and Snyder, M. W. (2019). Genomic medicine–progress, pitfalls, and promise. *Cell* 177, 45–57. doi:10.1016/J.CELL.2019.02.003

Simons, K. (2018). How can omic science be improved? *Proteomics* 18, e1800039. doi:10.1002/PMIC.201800039

Tarailo-Graovac, M., Shyr, C., Ross, C. J., Horvath, G. A., Salvarinova, R., Ye, X. C., et al. (2016). Exome sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* 374 (23), 2246–2255. doi:10.1056/NEJMoa1515792

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, D., Hart, M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415 (6871), 530–536. doi:10.1038/415530a

Van Gool, A. J., Bietrix, F., Caldenhoven, E., Zatloukal, K., Scherer, A., Litton, J. E., et al. (2017). Bridging the translational innovation gap through good biomarker practice. *Nat. Rev. Drug Discov.* 16, 587–588. doi:10.1038/NRD.2017.72

Veenstra, T. D. (2021). Omics in systems biology: Current progress and future outlook. *Proteomics* 21, e2000235. doi:10.1002/pmic.202000235

Vlahou, A., Hallinan, D., Apweiler, R., Argiles, A., Beige, J., Benigni, A., et al. (2021). Data sharing under the general data protection regulation: Time to harmonize law and research ethics? *Hypertens. 1979* 77, 1029–1035. doi:10.1161/HYPERTENSIONAHA.120.16340

Wenk, M. R. (2005). The emerging field of lipidomics. *Nat. Rev. Drug Discov.* 4, 594–610. doi:10.1038/NRD1776

Wilkinson, M. D., Dumontier, M., Aalbersberg, IjJ., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018–160019. doi:10.1038/sdata.2016.18

frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 26 October 2022
DOI 10.3389/fmolb.2022.962644

OPEN ACCESS

EDITED BY
Sergio Oller Moreno,
University Medical Center Hamburg-Eppendorf, Germany

REVIEWED BY
Marie-Pier Scott-Boyer,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Martin Treppner,
martin.treppner@uniklinik-freiburg.de

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 06 June 2022
ACCEPTED 12 October 2022
PUBLISHED 26 October 2022

CITATION
Brombacher E, Hackenberg M, Kreutz C, Binder H and Treppner M (2022), The performance of deep generative models for learning joint embeddings of single-cell multi-omics data.
Front. Mol. Biosci. 9:962644.
doi: 10.3389/fmolb.2022.962644

The performance of deep generative models for learning joint embeddings of single-cell multi-omics data

Eva Brombacher[1,2,3,4,5†], Maren Hackenberg[1,2†], Clemens Kreutz[1,2,4], Harald Binder[1,2] and Martin Treppner[1,2*]

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany, [2]Freiburg Center for Data Analysis and Modeling University of Freiburg, Freiburg, Germany, [3]Spemann Graduate School of Biology and Medicine (SGBM) University of Freiburg, Freiburg, Germany, [4]Centre for Integrative Biological Signaling Studies (CIBSS) University of Freiburg, Freiburg, Germany, [5]Faculty of Biology University of Freiburg, Freiburg, Germany

Recent extensions of single-cell studies to multiple data modalities raise new questions regarding experimental design. For example, the challenge of sparsity in single-omics data might be partly resolved by compensating for missing information across modalities. In particular, deep learning approaches, such as deep generative models (DGMs), can potentially uncover complex patterns *via* a joint embedding. Yet, this also raises the question of sample size requirements for identifying such patterns from single-cell multi-omics data. Here, we empirically examine the quality of DGM-based integrations for varying sample sizes. We first review the existing literature and give a short overview of deep learning methods for multi-omics integration. Next, we consider eight popular tools in more detail and examine their robustness to different cell numbers, covering two of the most common multi-omics types currently favored. Specifically, we use data featuring simultaneous gene expression measurements at the RNA level and protein abundance measurements for cell surface proteins (CITE-seq), as well as data where chromatin accessibility and RNA expression are measured in thousands of cells (10x Multiome). We examine the ability of the methods to learn joint embeddings based on biological and technical metrics. Finally, we provide recommendations for the design of multi-omics experiments and discuss potential future developments.

KEYWORDS

multi-omics, deep learning, experimental design, sample size, transcriptomics, proteomics, epigenomics

1 Introduction

Many diseases, such as cancer, affect complex molecular pathways across different biological layers. Consequently, there iscurrently an ongoing surge in multi-omics techniques that study the interaction of biomolecules across various omics layers (Veenstra, 2021b; Picard et al., 2021). Multi-omics techniques have been used, e.g., to infer mechanistic insights about molecular regulation, the discovery of new cell types, and the delineation of cellular differentiation trajectories (Colomé-Tatché and Theis, 2018; Adossa et al., 2021; Veenstra, 2021a; Tarazona et al., 2021). However, because performing multi-omics experiments in the same cell is still costly and experimentally complex, many experiments have been carried out with comparatively small numbers of cells so far. Additionally, single-cell multi-omics data suffer from the sparseness and noisiness of the measured modalities, differences in sequencing depth, and batch effects. Data analysis is further complicated by differing feature spaces as well as shared and modality- or batch-specific variation (Lance et al., 2022).

Deep learning approaches, known for their ability to learn complex non-linear patterns from data, have become a popular building block for integrating different data types (Grapov et al., 2018; Erfanian et al., 2021). For example, in 2021s Conference on Neural Information Processing Systems (NeurIPS) competition (https://openproblems.bio/neurips_2021), which addressed the topic of multimodal single-cell data integration, neural networks proved to be the most popular model choice, with shallow deep learning models being among the best-performing methods (Lance et al., 2022). Specifically, deep generative models (DGMs), such as variational autoencoders (VAEs), are

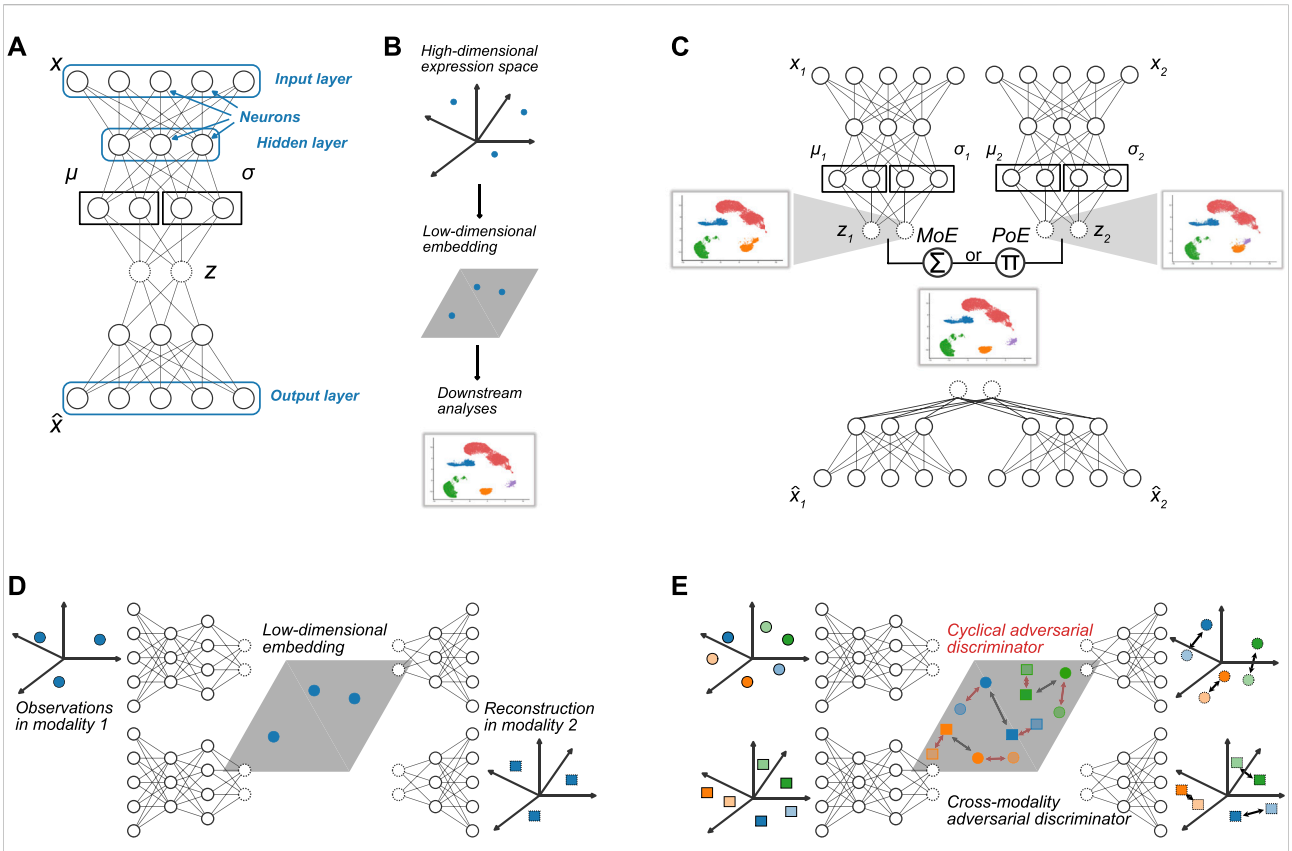


FIGURE 1
Neural network architectures. **(A)** Exemplary network architecture of vanilla VAE, where x represents the input data and $\hat{x}$ is the reconstructed data. Random variables z in the bottleneck layer are indicated by dashed circles. μ and σ represent the mean and standard deviation of the distributions, typically Gaussian distributions with diagonal covariance matrices, learned in the bottleneck layer. **(B)** Typical workflow: High-dimensional omics data are mapped to a low-dimensional embedding, which can then be utilized for visualization and downstream analyses such as clustering or trajectory inference. **(C)** General architecture of multimodal VAEs. **(D)** Cross-modality translation: High-dimensional measurements from one modality are mapped to a low-dimensional embedding with the modality-specific encoder. The latent representation is then used as input for the decoder of the respective other modality. **(E)** Adversarial training principles: Adversarial discriminators can be employed (1) to align low-dimensional embeddings of different modalities (squares vs. circles) of the same cell (same color) in the latent space (black arrows), (2) to align reconstructed profiles with the cross-modal reconstructions (lighter colors) obtained by decoding low-dimensional embeddings of one modality with the decoder of the other modality (black arrows in the reconstruction space, or (3) to align re-embedded decoder outputs from intra-modal and/or cross-modal reconstruction (lighter colors) with the original embeddings (red arrows in the latent space).

increasingly employed to infer joint embeddings, i.e., low-dimensional representations, from multi-omics datasets. This allows for performing all further downstream analyses simultaneously within this joint latent space (Figure 1B).

This review provides a systematic overview of current DGM-based approaches for learning joint embeddings from multi-omics data and illustrates how small sample sizes impact the amount of information that can be recovered from multi-omics datasets. Specifically, we examine how the performance of popular DGM-based approaches to infer joint low-dimensional representations from such data is influenced by varying numbers of cells. The required number of cells is particularly relevant at the stage of designing an experiment (Treppner et al., 2021). To tackle the challenging task of evaluating the quality of a latent representation with respect to the conservation of biological signal and batch correction capabilities, we draw on the guidelines provided by Luecken et al. (2021a).

The training of DGMs on multi-omics data is challenging due to the inherent high dimensionality and low sample size of multi-omics data and the large number of model parameters that need to be estimated while avoiding overfitting and bias (Kang et al., 2021). Thus, we investigate the impact of cell numbers on the performance of selected single-cell multi-omics integration algorithms. We consider eight popular VAE-based tools that incorporate different integration paradigms and training strategies for this illustration. Specifically, we included product-of-experts- and mixture-of-experts-based approaches and techniques that employ additional, commonly used integration techniques, such as cross-modality translation and adversarial training. Also, we chose models with different degrees of architectural complexity, including one model (Li et al., 2022) with (self-) attention modules and additional regularization by clustering consistency. We thus created an exemplary selection that represents the range of architectural choices, additional training and regularization strategies, and levels of complexity currently used for the task at hand. Thus, viewing the selected models as representatives of the current landscape of DGMs for multi-omics integration, our case study enables us to draw conclusions on the performance of the investigated tools in small sample size scenarios, and to give recommendations regarding architectural choices, integration strategies, and regularization paradigms.

2 Deep learning background

As the number of experimental methods in molecular biology is exploding, immense amounts of data are produced. Machine learning techniques can help in extracting information from such data to make it human-interpretable.

In recent years, deep learning has emerged as a potent tool for analyzing such high-throughput biological data. At the core of these approaches are artificial neural networks (ANNs) that provide powerful yet versatile building blocks to learn complex non-linear transformations and thus uncover underlying structures from high-dimensional data.

In particular, a networks' architecture comprises interconnected layers of neurons. Each neuron is connected to all of the neurons in the preceding layer. The depth of the network is determined by the number of hidden layers, i.e., the layers between the input and output layers. In contrast, the number of neurons in one layer determines a network's width (Figure 1A). With deep architectures, ANNs are especially effective at learning increasingly complicated patterns from large volumes of data based on non-linear transformations. Specifically, each individual neuron computes a weighted sum of its inputs, where the weighted total is then subjected to an activation function, typically producing a nonlinear transformation of the neuron's output. The weights of an ANN, which link the neurons between layers and make up the model's parameters, are a crucial part of the model. Training an ANN amounts to finding model weights that optimize a loss function, which represents how well the model fits the data. However, one of the major difficulties in training ANNs is optimizing the loss function as it is typically complex and non-convex and the parameter space is high-dimensional (Angermueller et al., 2016).

While supervised deep learning relies on labeled data to solve, e.g., classification problems, unsupervised deep learning can be employed in exploratory analyses to uncover central structure in data. For example, researchers frequently aim to understand cell-type compositions, for which they usually rely on unlabelled data. Hence, unsupervised deep learning methods have become increasingly popular in omics data analysis. Specifically, DGMs have been used for imputation (Lopez et al., 2018; Xu et al., 2020), visualization of the underlying structure of single-cell RNA-sequencing (scRNA-seq) data (Ding et al., 2018), and synthetic data generation (Marouf et al., 2020; Treppner et al., 2021).

Many computational approaches for processing scRNA-seq data use dimensionality reduction to produce a compressed representation of the high-dimensional transcription space. Grouping cells based on some measure of distance is a typical step in scRNA-seq research since these analyses usually attempt to understand the cell type composition of tissues or samples. However, conventional distance metrics, such as Euclidean distance, are unsuited to accurately represent similarity relations between cells due to the high dimensionality of the gene expression space, which is commonly referred to as the curse of dimensionality. As a result, the solution usually adopted is to reduce the number of dimensions based on the assumption that such a low-dimensional space captures the underlying biological phenomena. As an illustration, a transcription factor may be responsible for the activation of many genes. Therefore, one variable characterizing the activation of genes

through the transcription factor would be adequate to describe the patterns of gene expression rather than modeling the high-dimensional space spanned by all genes and their combinations (Kharchenko, 2021). Principal component analysis (PCA) is one method for reducing the dimensionality of scRNA-seq data. However, applying PCA to scRNA-seq data has a number of drawbacks since it assumes a symmetric distribution, which is typically not satisfied in scRNA-seq data, and only learns linear relationships. As a result, researchers have developed DGMs that accurately represent the distributional assumptions of scRNA-seq data while accurately portraying the data's inherent complexity (Lopez et al., 2018; Grønbech et al., 2020).

An autoencoder is the basis for many DGMs and is composed of three modules: an encoder, a bottleneck layer, and a decoder. The encoder reduces the input to a lower dimension (through the bottleneck layer), and the decoder reconstructs the original input from the bottleneck. This design also forms the foundation for the variational autoencoder and effectively compresses the essential information needed for data reconstruction (Lopez et al., 2020), which is mainly used to eliminate noise from data by compressing and re-compressing and reducing data to lower dimensions for visualization. In contrast, a variational autoencoder aims to infer the parameters of the probability distribution assumed to underly the source data, which can subsequently be used to generate realistic *in silico* data.

Specifically, DGMs are trained to capture the joint probability distribution over all features in the input data, thus allowing to also generate new synthetic data with the same patterns as the training data by sampling from the learned distribution. This is typically done by introducing latent random variables z in addition to the observed data x. In single-cell transcriptomics applications, these latent variables might encode complex gene programs based on non-linear relationships between genes. Typically, the joint distribution $p_\theta(x, z)$ of observed and latent variables is described through a parametric model, where θ represents the model parameters. The joint probability can be factorized into a prior probability $p_\theta(z)$ and a posterior $p_\theta(x|z)$ and can thus be written as $p_\theta(x, z) = p_\theta(z)$ $p_\theta(x|z)$. Inferring the data likelihood $p_\theta(x) = \int p_\theta(x, z)dz$ from the joint distribution requires marginalizing over all possible values of z, which is typically computationally intractable (Kingma and Welling, 2019). Hence, approximate inference techniques are employed to efficiently optimize the model parameters (Blei et al., 2017).

Two methods are frequently used in the machine learning literature to aggregate distributions, such as data from various single-cell modalities like gene expression and surface proteins. One strategy involves multiplying the density functions of the two modalities to create a product of experts (PoE) approach. On the other side, a mixture of experts (MoE) approach can blend the modalities using a weighted sum. In Section 2.2, we go over these strategies' benefits and drawbacks.

In single-cell applications, the most frequently used DGMs to date are Variational autoencoders (VAEs) (Kingma and Welling, 2013) and generative adversarial networks (GANs) (Goodfellow et al., 2014), which we present in more detail below.

2.1 Variational autoencoders

VAEs employ two independently parameterized but jointly optimized neural network models to learn an explicit parametrization of the underlying probability distributions. This is achieved by non-linearly encoding the data into a lower-dimensional latent space and reconstructing back to the data space. Specifically, the encoder (or recognition model) maps the input data x to a lower-dimensional representation given by a sample of the latent variable z, while the decoder network performs a reverse transformation and aims to reconstruct the input data based on the lower-dimensional latent representation (Figure 1A).

To approximate the underlying data distribution $p_\theta(x)$, the encoder and decoder parameterize the conditional distributions $p_\theta(z|x)$ and $p_\theta(x|z)$, respectively. Since $p_\theta(x)$ and $p_\theta(z|x)$ are intractable, a variational approximation $q_\phi(z|x)$ is employed, typically given by a Gaussian distribution with diagonal covariance matrix.

Intuitively, the model is trained by reconstructing its inputs based on the lower-dimensional data representation, such that the latent space recovers the central factors of variation that allow for approximating the data distribution as closely as possible. Formally, a training objective for the model can be derived based on variational inference (Blei et al., 2017). The parameters ϕ and θ of the encoder and decoder distributions can be optimized by maximizing the evidence lower bound (ELBO), a lower bound for the true data likelihood $p_\theta(x)$, with respect to ϕ and θ. Denoting with KL the Kullback-Leibler divergence $\mathrm{KL}[q\|p] \coloneqq \mathbb{E}_q[\log \frac{q}{p}]$ for probability distributions q and p, the ELBO is given by

$$\begin{aligned} ELBO(x;\phi,\theta) &= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathrm{KL}\left[q_\phi(z|x)\|p(x)\right] \quad (1) \end{aligned}$$

Here, the likelihood of a single observation (i.e., cell) x indicates how well it is supported by the model. The first term on the right side of Eq. 1 describes the reconstruction error indicating how well the generated samples from the model resemble the input. The KL-divergence on the right-hand side quantifies the difference between the approximate posterior to the true posterior, and, therefore, defines the tightness of the bound—meaning the difference between the ELBO and the marginal likelihood.

The decoder network is typically built to learn the parameters of specific distributions, which best describe the underlying biological data. For scRNA-seq and surface protein data

(CITE-seq) a negative binomial distribution is frequently assumed, while single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) data usually requires an additional modeling term that accounts for the increased sparsity of the data, e.g., in the form of a zero-inflated negative binomial (ZINB) distribution (Minoura et al., 2021). Other approaches use a binarized version of the scATAC-seq data (Ashuach et al., 2021; Wu et al., 2021; Zuo et al., 2021; Zhang R. et al., 2022).

The typical workflow for analyzing high-dimensional (single- or multi-) omics data with a VAE is illustrated in Figure 1B. The data is embedded with the encoder to obtain a low-dimension representation, which can subsequently be used for downstream analysis, such as clustering or trajectory inference.

2.2 Multimodal variational autoencoders

Several approaches already exist in which multimodal VAEs (Shi et al., 2019) are used to map different omics measurements into a common latent representation (Gong et al., 2021; Minoura et al., 2021; Lotfollahi et al., 2022). Each of these methods uses different approaches to combine the latent variables of the respective modalities. We can usually distinguish between MoE and PoE models (Figure 1C). Hence, we describe a MoE and a PoE model in more detail below and examine their performance in our analyses.

We denote a single-cell multimodal dataset as $x_{1:M}$, where two modalities ($M = 2$) is the most common case. The joint generative model can therefore be written as $p_\theta(x_{1:M}, z) = p(z)\prod_{m=1}^{M} p_{\theta_m}(x_m|z)$, where $p_{\theta_m}(x_m|z)$ represents the likelihood of the decoder network for modality m, and $\theta = \{\theta_1, \ldots, \theta_M\}$.

For the MoE model, the resulting joint variational posterior can be factorized into $q_\phi(z|x_{1:M}) = \sum_{m=1}^{M} \alpha_m q_{\phi_m}(z|x_m)$, with $\alpha_m = 1/M$ and $\phi = \{\phi_1, \ldots, \phi_M\}$. This results in the following ELBO:

$$ELBO = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{z_m \sim q_{\phi_m}(z|x_m)}\left[\log \frac{p_\theta(x_{1:M}, z)}{q_\phi(z|x_{1:M})}\right]$$
$$= \frac{1}{M}\sum_{m=1}^{M}\left\{\mathbb{E}_{z_m \sim q_{\phi_m}(z|x_m)}\left[\log p_\theta(x_{1:M}|z_m)\right] - KL\left[q_{\phi_m}(z|x_m)\|p(x)\right]\right\} \quad (2)$$

which is similar to Eq. 1, but the ELBOs of the individual modalities are combined by a weighted average. In contrast, PoE approaches (Gong et al., 2021; Lotfollahi et al., 2022) combine the variational posteriors of the individual modalities as products $q_\phi(z|x_{1:M}) = \prod_{m=1}^{M} q_{\phi_m}(z|x_m)$.

Shi et al. (2019) argue that PoE approaches suffer from potentially overconfident experts, i.e., experts with lower standard deviations will tend to have a more considerable influence on the combined posterior, as experts with lower precision come with lower marginal posteriors. In contrast, in the MoE approach we consider here, both modalities receive equal weighting, reflecting the assumption that both modalities are of similar importance. Intuitively, employing a PoE approach corresponds to taking the 'intersection' of the individual posteriors, as a single posterior assigning a near-zero likelihood to a specific observation is enough to cause the product to be near-zero. In contrast, an MoE approach corresponds to taking the 'union' of all posteriors. Additionally, the weights α_m assigned to each modality can be adjusted to reflect prior assumptions on their relative importance or be learned from the data during training.

2.3 Cross-modality translation

In addition to architectural choices regarding the integration of the modality-specific sub-networks *via* a PoE or MoE approach, many VAE-based methods introduce training objectives that facilitate specific functionality such as cross-modality translation or encourage particular properties of the embedding, such as clustering consistency between the modality-specific latent representations. On a higher level, these components can be seen as regularizers that push the embeddings found by the model towards certain desired properties.

A prominent example of such an additional feature to direct a joint embedding is cross-modality translation. Here, a cell's measurements of one modality, say, gene expression, are mapped to the joint latent space with the respective modality-specific encoder. Then, the decoder of another modality, say, chromatin accessibility, is employed to map the latent representation of the gene expression profile to a corresponding chromatin accessibility profile (Figure 1D). This is only possible due to the integration of both modalities into a shared latent space, in which 1 cell's encoded representations of different modalities align.

When paired measurements of both modalities in the same cell are available, the translated reconstructions in the respective other modality can be compared to the cell's observed profile during training. The model learns a latent embedding that facilitates consistent cross-modality predictions. Thus, the model is explicitly pushed towards an embedding from which both modality-specific profiles can be reconstructed equally well, and that can, therefore, help in better capturing general underlying biological cell states as defined by the interplay of both modalities.

After training, cross-modality translation can be used to impute measurements of cells for which a specific modality is missing or to answer counterfactual questions such as 'based on this specific gene expression profile, what would the corresponding chromatin accessibility profile have looked like?'. This could be further combined with *in silico*

perturbations, i.e., generating synthetic profiles of one modality and using the model to infer corresponding profiles in other modalities. Additionally, this technique can be used to query the trained model for, e.g., subpopulations of cells where the cross-modality predictions are particularly well or particularly poorly aligned with the true measurements and to further characterize them, thus, also facilitating interpretability.

Examples of approaches that employ this technique are given by, e.g., Minoura et al. (2021), Wu et al. (2021), and Zhao et al. (2022), and will be presented in more detail below in Section 3 and in the experimental Section 6.

2.4 Adversarial training strategies

Another commonly used regularization technique is given by adversarial training, which is closely related to cross-modality translation and is often employed concurrently. Such adversarial components are often integrated into a variational or standard autoencoder framework and are inspired by generative adversarial networks (GANs) (Goodfellow et al., 2014), another form of DGMs that differs from VAEs in how the joint probability distribution over all input features is specified. While VAEs learn an explicit parameterization of (an approximation of) this distribution (see 2.1), in GANs, this distribution is available only implicitly *via* sampling. A GAN consists of a generator and a discriminator neural network that can be thought of as playing a zero-sum minimax game: The generator simulates synthetic observations that are presented to the discriminator together with real data observations. The discriminator then has to decide whether a given sample is a real observation or a synthetic one from the generator.

In multi-omics data integration, such adversarial approaches are typically integrated into (V)AE models as additional components to regularize the latent representation and/or the decoder reconstructions (Liu et al., 2021; Xu et al., 2021a; Hu et al., 2022; Zhao et al., 2022), while, e.g., Amodio and Krishnaswamy (2018); Amodio et al. (2022) present purely GAN-based approaches. More specifically, a discriminator is typically employed to distinguish between two omics modalities, either based on samples from their latent representations or based on reconstructed samples from cross-modal decoders (Figure 1E, black arrows). The objective of the discriminator then is to maximize the probability of correctly identifying the original modality a sample comes from, while the encoder and decoder of the (V)AE model are trained to fool the discriminator by producing samples that are indistinguishable. By training all components jointly, the (V)AE model is encouraged to find a latent embedding in which the different modalities are better aligned and integrated, and/or learn decoders that allow for accurate cross-modal predictions well aligned with the intra-modal predictions. In practice, this is achieved by incorporating adversarial penalty terms into the loss function.

Such adversarial components can also be used to train the model in a cyclical fashion for additional intra-modal and cross-modal consistency. For intra-modal consistency, the low-dimensional embeddings of samples of one modality are decoded with the modality-specific decoder. Subsequently, the reconstructions are re-encoded with the modality-specific encoder and compared to the original embedding of the sample. An adversarial discriminator can be employed to align the embedding of the original sample with the embedding of the re-encoded reconstruction of that sample (Figure 1E, red arrows). For cross-modal consistency, the low-dimensional embeddings from samples of one modality are decoded and subsequently re-encoded with the decoder and encoder of the other modality. By aligning these cross-modal embeddings with the original embeddings using an adversarial discriminator, the model can learn to produce cross-modal translations that are consistent with the original sample when re-embedded in the latent space.

3 Literature review

Although recently, several available deep learning-based applications for the integration of single-cell multi-omics data have been reviewed in (Erfanian et al., 2021) and (Stanojevic et al., 2022), there is still a lack of a more comprehensive review focusing specifically on DGMs. In the following, we are going to survey approaches for paired (both modalities measured in the same cell in one experiments) and unpaired (modalities measured in different cells in separate experiments) single-cell data. An overview is given in Table 1, where we list recent deep learning-based approaches for multi-omics data integration. We remark whether the methods are designed for paired or unpaired datasets and compare the basic network architectures and demonstrated modalities on which the respective methods have been demonstrated. Additionally, we comment on the integration tasks tackled by each model and provide a reference to the implementation.

We exclusively included methods that learn a joint embedding based on DGMs and have been demonstrated on multi-omics data of different modalities (not just, e.g., single-cell RNA-seq from different protocols).

3.1 Approaches for paired data

The Cobolt model (Gong et al., 2021) learns shared representations between modalities and is based on a multimodal VAE, where an independent encoder network is used for each modality and the learned parameters of the posterior distributions are combined using a PoE approach.

TABLE 1 Overview of recently published deep learning-based methods to integrate single-cell multi-omics data. [1]Only for mapping single-omics to multi-omics; [2]Only when converting peaks to activity scores.

Name	References	Un-paired	Network architecture	Demonstrated modalities	Integration type	Code
MAGAN	Amodio and Krishnaswamy (2018)	yes	Two GANs, both unsupervised and semi-supervised training	Flow cytometry + scRNA-seq; Multiple CyTOF Panels; Multiple CyTOF Replicates	Integration of single-omics data	https://github.com/KrishnaswamyLab/MAGAN
SCIM	Stark et al. (2020)	yes	multimodal autoencoders with an adversarial objective	scRNA + CyTOF, more modalities possible	Integration of multi-omics data	https://github.com/ratschlab/scim
BABEL	Wu et al. (2021)	no	VAE with separate encoders and decoders, trained by cross-prediction	SNARE-seq, SHAREseq, CITE-seq, scRNA-seq, scATAC-seq	Cross-modality translation	https://github.com/wukevin/babel
Cobolt	Gong et al. (2021)	yes	MVAE (direct fusion/concatenation)	SNARE-seq, 10x multiome (treated as different modalities)	Integration of multi-omics data and multi- with single-omics data	https://github.com/epurdom/cobolt
DAVAE	Hu et al. (2022)	yes	VAE, shared encoder + adversarial classifier	scRNA-seq from different samples/protocols (SmartSeq2, 10X), scRNA + scATAC-seq, 10X/Visium. Requires common input features	Integration of multiple scRNA-seq into an atlas References, transfer learning	https://github.com/jhu99/scbean
DCCA	Zuo et al. (2021)	no	VAE with separate mutually supervised encoders and decoder	scRNA-seq + scATAC-seq (10x, SNARE-seq, SHARE-seq, scNMT-seq)	Transfer learning, impute missing modalities	https://github.com/cmzuo11/DCCA
MultiVI	Ashuach et al. (2021)	no[1]	VAE (distributional average and penalization to mix the latent representations)	scRNA-seq + scATAC-seq (PBMC 10x)	Integration of multi-omics data and multi-omics with single-omics data, imputation of missing modalities	https://github.com/YosefLab/scvi-tools
p/mp SMILE	Xu et al. (2021b)	no	Modality-specific encoders trained by noise-contrastive estimation	scRNA-seq + scATAC-seq, scMethyl + scHi-C, SNARE-seq, sci-CAR, SHARE-seq, (integration of >2 modalities possible)	Integration of single-omics and multi-omics data	https://github.com/rpmccordlab/SMILE
SCALEX	Xiong et al. (2021)	(yes)[2]	VAE with batch-free encoder and a batch-specific decoder	CITE-seq, spatial transcriptome MERFISH data, scRNA-seq + scATAC-seq	Integration of single-omics data, integration of multi-omics data	https://github.com/jsxlei/SCALEX
scMM	Minoura et al. (2021)	no	VAE (mixture of experts)	CITE-seq + SHARE-seq	Integration of multi-omics data, cross-modal prediction	https://github.com/kodaim1115/scMM
scMVAE	Zuo and Chen (2021)	no	MVAE (3 strategies: product of experts, neural network, direct concatenation)	SNARE-seq	Integration of multi-omics data	https://github.com/cmzuo11/scMVAE
TotalVI	Gayoso et al. (2021b)	no	VAE	CITE-seq	Integration of multi-omics data, missing protein imputation	https://github.com/YosefLab/scvi-tools
Con-AAE	Wang et al. (2022)	no	Two autoencoders, using adversarial loss and latent cycle-consistency loss	sci-CAR, SNAREseq	Integration of single-omics data, integration of multi-omics data	https://github.com/kakarotcq/RNA-Seq-and-ATAC-Seq-mapping
MIRA	Lynch et al. (2022)	no	VAE	SHARE-seq and 10X	Integration of multi-omics data	https://github.com/cistrome/MIRA
Polarbear	Zhang et al. (2022a)	yes	VAE with semi-supervised cross-domain translation	SNARE-seq (+snATAC-seq, scATAC-seq, scRNA-seq)	Cross-modality translation, align single-modality data, predict missing modalities	https://github.com/Noble-Lab/Polarbear
Multigrate	Lotfollahi et al. (2022)	yes	VAE (product of experts)	CITE-seq and scRNA-seq + scATAC-seq (adaptable to other modalities)	Mapping of novel multi-omic query datasets to a References atlas, imputation of missing modalities, integration of multi-omics	https://github.com/theislab/multigrate
Portal	Zhao et al. (2022)	yes	AE + GAN: adversarial discriminators on latent spaces	Various single-cell RNA-seq (Drop-seq, 10X, SmartSeq2), scRNA (10X, DropSeq) + snRNA-seq (Split-Seq), scRNA + scATAC-seq	Integration of multi-omics and single-omics data, cross-modality translation	https://github.com/YangLabHKUST/Portal

(Continued on following page)

TABLE 1 (*Continued*) Overview of recently published deep learning-based methods to integrate single-cell multi-omics data. [1]Only for mapping single-omics to multi-omics; [2]Only when converting peaks to activity scores.

Name	References	Un-paired	Network architecture	Demonstrated modalities	Integration type	Code
scMVP	Li et al. (2022)	no	Multimodal VAE with Gausian mixture prior and attention modules	SNARE-seq, sci-CAR, Paired-seq, SHARE-seq, 10X (could be extended to parallel profiling of other epigenomic data)	Integration of multi-omics data	https://github.com/bm2-lab/scMVP

Additionally, Cobolt can jointly integrate single-modality datasets with multi-omics datasets, allowing one to draw on the many publicly available scRNA-seq or scATAC-seq datasets.

Multigrate (Lotfollahi et al., 2022) is another model that employs a PoE to combine the posteriors of different modalities. Additional datasets can be integrated into the model by minimizing the maximum mean discrepancy (MMD) loss between joint representations of different datasets.

Similar to Cobolt and Multigrate, scMM (Minoura et al., 2021) is a VAE-based method that trains an encoder network for each modality independently. However, instead of combining the parameters of the posterior distributions using a PoE, a MoE is used. By equally mixing information from both modalities through the MoE, the model avoids putting too much emphasis on one individual modality only (Minoura et al., 2021). In addition, scMM provides a method for model interpretability that uses latent traversals, where synthetic cells are generated by the learned decoder and one latent variable is modified continually, while the others remain fixed. The Spearman correlations calculated between each latent variable and the features of each modality then allow relevant features to be identified. Additionally, by using a Laplace prior, scMM learns disentangled representations, with correlations between latent variables being penalized, which allows for better interpretation of individual features (Treppner et al., 2022).

Similarly, the MultiVI model presented by Ashuach et al. (2021) is also based on a MoE with $\alpha_m = 1/M$ where M denotes the number of modalities, as the authors use individual encoders for each data modality and then average the resulting variational posteriors. However, a regularization term is added to the ELBO, which penalizes the distance between the learned latent representations such that a joint representation can be inferred (Ashuach et al., 2021).

While the single-cell multi-view profiler (scMVP) (Li et al., 2022) is also based on a multimodal VAE architecture with modality-specific encoders and decoders and a joint latent space, it more explicitly accounts for the much higher sparsity of single-cell measurements from joint profiling protocols, with a throughput of only one-tenth to one-fifth of that of single-modality assays (Li et al., 2022). Specifically, the authors employ attention-based building blocks for both the encoder and decoder. Attention mechanisms have first been proposed in computer science in the context of machine translation (Bahdanau et al., 2014; Kim et al., 2017) and are based on the idea of using flexible weighting of an input observation, to have the model specifically 'attend to' the most important parts of the observation. In the context of omics data, attention scores are assigned to the observed features (e.g., genes, chromatin loci) of each cell, to enhance the effect and interplay of specific features. In contrast to fixed weights, the attention scores are learned during model training and can thus adapt to highlight the most informative features for learning, e.g., latent representations. Attention-based mechanisms have specifically been popularized by transformer models (Vaswani et al., 2017) due to their high performance on sparse datasets in the area of natural language processing or protein structure prediction. In scMVP, the authors build on that by using multi-head self-attention transformer modules to capture local, long-distance correlation in the encoder and decoder of the term frequency-inverse document frequency-transformed (Stuart et al., 2021) scATAC-seq data while using simple attention blocks in the RNA encoder and decoder. Given the latent embedding, the modality-specific decoders are weighted according to the posterior probabilities of cell-type or cluster identity. To encourage consistency of the shared latent space, the decoder-reconstructed values of each modality are again embedded into the latent space, and the KL-divergence between the joint latent embedding and the modality-specific re-embedding from the reconstructed data is minimized as an additional loss term. This corresponds to the idea of cyclical adversarial training as described in Section 2.4 and Figure 1E. More generally, this concept is based on a cycle GAN (Zhu et al., 2017) and is also present in, e.g., Xu et al. (2021a); Zhao et al. (2022); Khan et al. (2022); Wang et al. (2022) and Zuo et al. (2021).

SCALEX (Xiong et al., 2021) builds on SCALE (Single-Cell ATAC-seq Analysis *via* Latent feature Extraction) (Xiong et al., 2019), a tool for analyzing scATAC-seq data. The developers of SCALE found that its encoder could be beneficial in disentangling cell-type- and batch-related features, which would allow for online integration of different batches. Specifically, using a VAE, SCALEX integrates different batches into a batch-invariant embedding through simultaneous learning of a batch-free encoder and a batch-specific decoder. The latter contains a domain-specific batch normalization layer. This

allows the encoder to concentrate only on batch-invariant biological data components while being oblivious to batch-specific variations. The resulting generalizability of the encoder further allows for the integration of new single-cell data in an online manner, i.e., without the need to retrain the model. The authors demonstrate this property of SCALEX by generating multiple expandable single-cell atlases.

Another subgroup of models addresses the task of translating between different modalities. These cross-modality translation approaches, however, often do not learn a common latent representation of the data. For example, Polarbear (Zhang R. et al., 2022) trains VAEs on each of two modalities (here: scRNA-seq and scATAC-seq data) and then links the respective encoders to the decoders of the other modality. The authors intend that the training in the first stage, i.e., the training of the individual VAEs, takes place on publicly available single-assay data, whereby the translation task is carried out on SNARE-seq data in a supervised manner.

Another such model called BABEL (Wu et al., 2021) similarly employs distinct modality-specific encoders and decoders for scRNA- and scATAC-seq data but utilizes a shared latent space. In contrast to PoE/MoE approaches, this joint representation is not constructed from separate spaces from each modality, but the encoders directly project onto the common latent space. Mutual cross-modal translation together with single-modality reconstruction are then used to train the model, i.e., from each modality-specific encoder, a sample of the joint latent representation is obtained and subsequently passed through both decoders to reconstruct both the scRNA and the scATAC profiles of the respective cell. Thus, both the reconstruction of the modality itself and the respective other modality based on the joint latent embedding are evaluated for each modality.

A similar approach is taken by Portal (Zhao et al., 2022), where a domain translation framework is combined with an adversarial training mechanism to integrate scRNA- and scATAC-seq data. Specifically, as in (Wu et al., 2021), modality-specific encoders directly embed the data in a shared latent space and cross-modal generators are introduced to decode the latent representation to the respective other modality. The resulting domain translation networks for each modality are then trained to compete against adversarial discriminators on the domain of each modality that aims to distinguish between original cells from the respective modality and cells translated from the other modality. The discriminators are specifically designed to adaptively distinguish between domain-shared and domain-unique cells by thresholding the discriminator scores. Since, according to the authors, domain-unique cell populations are prone to be assigned with extreme discriminator scores, discriminators are, thus, made effectively inactive on cells with a high probability of being modality-specific, which avoids the risk of over-correction by enforced alignment of domain-unique cells. Further, additional regularizers are employed: an autoencoder loss based on the within-modality reconstructions, a latent alignment loss to encourage the consistency of a specific cell's embedding and the embedding of its cross-modal reconstruction, and a cosine similarity loss between cells and their cross-modal reconstructions. Notably, Portal uses the first 30 principal components of a joint PCA as inputs for the model and employs a 20-dimensional latent space, such that the dimension reduction component is less pronounced than for the other models, and the data are not modeled as counts.

The authors of Zuo and Chen (2021) have extended scMVAE and proposed Deep Cross-Omics Cycle Attention (DCCA) (Zuo et al., 2021), which improves some of the weaknesses of scMVAE. DCCA combines VAEs with attention transfer. While scMVAE combines two modalities into a shared embedding, which potentially attenuates modality-specific patterns, in the case of DCCA, each data modality is processed by a separate VAE. These VAEs can then learn from each other through mutual supervision based on semantic similarity between the embeddings of each omics modality.

In the sciCAN model presented by Xu et al. (2021a), modality-specific autoencoders map the input data to a latent space for each modality, and a discriminator is employed to distinguish between the two modalities based on their latent representations. Additionally, a cross-modal generator is employed that generates synthetic scATAC-seq data based on the scRNA-seq latent representation, and a second discriminator is employed to distinguish between generated and real scATAC-seq samples. Additionally, the generated scATAC-seq data can be fed to the encoder again, and the latent representation is compared with the original latent representation from the scRNA-seq data used for generating the scATAC-seq data, thus introducing a cycle consistency loss (see Figure 1E, Section 2.4). Notably, the model does not necessarily expect paired measurements from the same cell but employs a shared encoder for both modalities, and, thus, requires a common feature set.

The authors of Hu et al. (2022) propose the DAVAE model based on domain-adversarial and variational approximation to integrate multiple single-cell datasets and paired scRNA-seq and scATAC-seq data. The model employs an adversarial training strategy to remove batch effects and enable transfer learning between modalities, by incorporating a domain classifier that tries to determine the batch or modality label based on the latent representation of VAE and training the VAE encoder to 'fool' the classifier *via* an adversarial loss component. Similarly to Portal and sciCAN, the DAVAE model also employs a shared encoder and thus requires a common set of input features.

Similarly, the scDEC model proposed by Liu et al. (2021) is based a pair of generative adversarial models to learn a latent representation. While focusing on scATAC-seq data analysis, this approach also allows for integrative analysis of multi-modal scATAC and scRNA-seq datasets for trajectory inference during

differentiation processes and cell type identification based on the joint latent representation.

Finally, MIRA (Lynch et al., 2022) combines probabilistic cell-level topic modeling (Blei, 2012) with gene-level regulatory potential (RP) modeling (Wang et al., 2013; Qin et al., 2020) to determine key regulators responsible for fate decisions at lineage branch points. The topic model uses a VAE with a Dirichlet prior to learn both the topic of the gene transcription and the topic of gene accessibility for each cell to derive the cell's identity. Complementing MIRA's topic model, its RP model integrates the transcription and accessibility information for each gene locus to infer how the expression of the respective gene is influenced by surrounding regulators. To this end, the topic model learns the rate with which the regulatory influence of enhancers decays with increasing genomic distance. In addition, the identity of key regulators is identified by analyzing transcription factor motif enrichment or occupancy.

3.2 Approaches for unpaired data

Since the generation of multi-omics measurements in the same cell is still costly and experimentally complex, many methods for integrating datasets measured in different cells are being developed.

Because of the difficulty of linking latent representations learned from variational autoencoders in the absence of measurement pairing information, Lin et al. (2022) proposed a transfer learning approach. Although not a DGM, it is worth mentioning in this article because of its usefulness and the possibility of adapting it to unsupervised settings. Notably, it represents a method for a horizontal alignment task, i.e., it relies on a common set of features as anchors and thus requires the translation of scATAC peaks to gene activity scores.

In a similar spirit, the scDART model proposed by Zhang Z. et al. (2022) learns a neural network-based joint embedding or unpaired scRNA-seq and scATAC-seq data by composing the embedding network with a gene-activity module network that maps scATAC peaks to genes. In addition, scDART can leverage partial cell matching information by using it as a prior to inform the training of the gene activity function.

Similar to the sciCAN model presented by Xu et al. (2021a), scAEGAN (Khan et al., 2022) also embraces the concept of cycle consistency, integrating the adversarial training mechanism of a cycle GAN (Zhu et al., 2017) into an autoencoder framework. Specifically, for each modality, a discriminator and a generator are defined. In addition to the standard GAN loss for each modality, a cycle loss is calculated by mapping a cell from one modality to the second modality with the second modality's generator and mapping it back to the first modality with the first modality's generator and comparing that to the original observation. Unlike for Xu et al. (2021a), the model does not rely on a common feature set but first trains an autoencoder model independently for each modality before training a cycle GAN on the two latent spaces to enforce their consistency.

A similar approach is employed in the Contrastive Cycle Autoencoder (Con-AAE) proposed by Wang et al. (2022). Again, the consistency between latent spaces of modality-specific autoencoders is enforced by a cycle consistency loss. However, here, it is more tightly integrated within the AE architecture, as the modality-specific encoder and decoders are used as generators, i.e., samples from one modality are embedded with the modality-specific encoder but decoded with the decoder of the other modality, and subsequently encoded with the other modality encoder back to the latent space, where they are compared with the original latent representation from the original encoder of the modality.

A purely GAN-based approach to integrating unpaired data by aligning the respective manifolds is presented in Amodio and Krishnaswamy (2018).

Another line of research for the integration of unpaired multi-omics data focuses on the concept of optimal transport (Peyré and Cuturi, 2019). A separate embedding or distance matrix is constructed from each modality, and the alignment task is formulated to find an optimal coupling between the two embeddings or distance matrices. An optimal coupling corresponds to finding a map along which one modality can be "transported" with minimal cost to the other, which can be formalized as an optimal transport problem (Peyré and Cuturi, 2019). Examples for such optimal transport-based methods are UnionCom (Cao et al., 2020), SCOT (Demetci et al., 2022) and Pamona (Cao et al., 2021). While these approaches typically rely on computing a coupling between modality-specific distance matrices and are not deep learning-based, a recent approach called uniPort employs a VAE architecture and solves an optimal transport problem in the latent space. More specifically, a shared encoder that requires a common input feature set across modalities is used to project the data into a common latent space, is combined with modality-specific decoders for reconstruction, and an optimal transport loss is minimized between the latent cell embeddings from different modalities.

Finally, the recently published Graph-Linked Unified Embedding (GLUE) framework (Cao and Gao, 2022) is based on the construction of a guidance graph based on prior knowledge of the relations between features of the different modalities to explicitly model regulatory interactions across different modalities with distinct feature spaces. This is achieved by learning joint feature embeddings from the knowledge graph with a graph VAE and linking them to modality-specific autoencoders. Specifically, the decoder of these modality-specific AEs is given by the inner product of the feature embeddings and the cell embeddings from the latent space of the respective modality. Additionally, the cell embeddings of different modalities are aligned using an adversarial discriminator.

4 Benchmark dataset

To acquire an objective performance estimate of the ability of different multi-omics integration approaches to describe the biological state of a cell through learning a joint embedding from multiple modalities, we used the benchmark dataset which was provided in the course of the NeurIPS 2021 competition and for which the ground-truth cell identity labels are known (Luecken et al., 2021a). This dataset was the first available multi-omics benchmarking dataset for single-cell biology. It mimics realistic challenges researchers are faced with when integrating single-cell multi-omics data, e.g., by incorporating nested donor and site batch effects (Lance et al., 2022).

Specifically, the NeurIPS benchmark dataset is a multi-donor (10 donors), multi-site (4 sites), multi-omics bone marrow dataset comprising two data types (Lance et al., 2022):

- CITE-seq data with 81,241 cells, where for each cell RNA gene expression (GEX) and cell surface protein markers using antibody-derived tags (ADT) are jointly captured.
- 10X Multiome assay data with 62,501 cells, where nucleus GEX and chromatin accessibility measured by assay for transposase-accessible chromatin (ATAC) are jointly captured.

In total, this dataset contained information on the accessibility of 119,254 genomic regions, the expression of 15,189 genes, and the abundance of 134 surface proteins, and has been preprocessed as described in Luecken et al. (2021a). We acquired the benchmark dataset from the NeurIPS 2021 website (https://openproblems.bio/neurips_2021), it can, however, also be accessed *via* https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122.

As recommended in Luecken and Theis (2019), we filtered this dataset for highly variable genes, as they are considered to be most informative of the variability in the data. In addition, analogous to the FindTopFeatures function of Signac (Stuart et al., 2021), we filtered the ATAC data such that we retained only peaks with the 25% highest overall counts. Finally, to determine the effect of the number of cells, we randomly subsampled the original NeurIPS dataset to subsamples containing information on 500, 1,000, 2,500, 5,000, and 10,000 cells, where for each number of cells we sampled 10 subsamples of that size.

5 Performance metrics

Generating a highly resolved, interpretable, low-dimensional embedding capturing the underlying biological cell states is pivotal for the analysis of multi-omics data (Lähnemann et al., 2020; Lance et al., 2022). We assess the performance of the compared integration approaches based on six metrics capturing the conservation of biological variation (normalized mutual information (NMI), cell type average silhouette width (ASW), trajectory conservation) and the degree of batch removal (batch ASW, site ASW, graph connectivity) (Lance et al., 2022). These metrics are described in detail in Luecken et al. (2021b) and are briefly introduced below:

- NMI compares the overlap of two clusterings. It is used to compare the Louvain clustering of the joint embedding to the cell type labels. It ranges from 0 (uncorrelated clustering) to 1 (perfect match).
- Cell type ASW is used to evaluate the compactness of cell types in the joint embedding. It is based on the silhouette width, which measures the compactness of observations with the same labels. Here, the ASW was computed on cell identity labels and scaled to a value between 0 (strong misclassification) and 1 (dense and well-separated clusters).
- The trajectory conservation assesses the conservation of a continuous biological signal in the joint embedding. Trajectories computed using diffusion pseudotime after integration for relevant cell types are compared. Based on a diffusion map space embedding of the data, an ordering of cells in this space can be derived. Using Spearman's rank correlation coefficient between the pseudotime values before and after integration, the conservation of the trajectory can be quantified, with the scaled score ranging from 0 (reverse order of cells on the trajectory before and after integration) to 1 (same order).
- Batch ASW describes the ASW of batch labels per cell. The scaled score ranges from 0 to 1, where 1 indicates well-mixed batches and any deviation from 1 indicates a batch effect.
- Site ASW describes the ASW of site labels per cell and can be interpreted analogously to batch ASW.
- The graph connectivity score evaluates whether cells of the same type from different batches are close to each other in the embedding by assessing if they are all connected in this embedding's k-nearest neighbor (kNN) graph. It ranges from 0 (no cell is connected) and 1 (all cells with the same cell identity are connected).

6 Results

We use various metrics to quantify the preservation of biological variation and metrics for the removal of technical effects based on the 10-dimensional embeddings obtained when applying Cobolt, scMM, TotalVI, and SCALEX to subsamples of the NeurIPS CITE-seq dataset, and Cobolt, scMM, MultiVI, scMVP, DAVAE, and Portal to subsamples of the NeurIPS Multiome dataset. We randomly sampled 500, 1,000, 2,500, 5,000, and 10,000 cells ten times each and applied the models to the respective datasets. We refrain from extensive parameter

optimisation as we put ourselves in the position of a user new to the field of deep learning, who will, most likely, leave the default parameters unchanged and use the same parameters as the original authors in their application of their proposed method. Thus, we used the default hyperparameters of the respective models as reported by the authors who originally proposed them where possible (Supplementary Material: Hyperparameters).

When applying scMM to the CITE-seq data, we frequently observed non-converging training runs, in particular for larger sample sizes. Here, we refer to the convergence of the iterative optimization procedure by stochastic gradient descent on the loss function of the respective model (see also Section 2). Convergence is achieved if towards the end of the training, the changes in the loss function in each iteration become smaller and eventually level out, whereas in non-converging runs we observe exploding gradients of the loss function. This is often due to suboptimal hyperparameter choices. For scMM, lowering the learning rate for sample sizes above 2,500 by one order of magnitude and increasing the batch size from 128 (default used by scMM) to 200 achieved convergence of the model training on all subsamples.

In general, similar performances were achieved irrespective of which of the two data types we used for deriving a joint embedding (Figures 2, 3). For the Multiome dataset, two of the considered tools, DAVAE and Portal, employ a shared encoder based on a common set of features across both modalities (top 30 principal components of a joint PCA on both datasets for Portal and common highly variable genes when converting scATAC peaks to gene activity scores for DAVAE) and thus embed each cell's profiles in the two modalities separately. To keep the evaluation as comparable as possible to the other tools, we thus created a joint embedding by calculating the mean of each cell's embedded profiles in the two modalities in a mixture-of-experts approach.

We compare our results with the metric values achieved by the models of the NeurIPS 2021 competition for the integration of the Multiome dataset (data points were extracted *via* WebPlotDigitizer-4.5 (Rohatgi, 2021) from Supplementary Figure S6 of (Lance et al., 2022)). However, as we merely used a subset of at most 10,000 cells of the original benchmark dataset, we expect our investigated algorithms to score higher for most metrics if they were to be subjected to the complete benchmark dataset.

By visual inspection of the Uniform Manifold Approximation and Projection (UMAP) (Becht et al. (2019); Konopka and Konopka (2018) version 0.2.9.0 with default parameters) plots of one exemplary subsample (Figure 4 and Figure 5), we see that MultiVI shows no obvious clustering for 500 cells (2, top panel). In contrast, defined cluster structures are beginning to build at this low cell number, and become more refined for 10,000 cells, for all other investigated tools. This behavior of MultiVI for smaller numbers of cells is also reflected in lower values for most of the investigated performance metrics (Figures 4, 5). Interestingly, the TotalVI tool, which is built on a similar architecture and was used for the CITE-seq dataset does not show such behavior (4, top panel).

UMAP plots including further meta information on the embedded cells are given in Supplementary Figures S3–22 for the exemplary subsample.

To ensure that the number of parameters in the respective models is not the determining factor for decreasing performance on small sample sizes, we calculated the Spearman correlation coefficient between the ranks of the models from Figures 2, 3 and the evaluation metrics. The predominantly negative correlations, i.e., lower rank (better performance) with an increasing number of trainable parameters, indicate that more complex models also deliver better performance regardless of the number of observations.

6.1 Preserving biological information

We assess the preservation of biological variation based on the NMI, cell type ASW, and the trajectory conservation scores (Figure 2). In addition, we show boxplots of the metrics for all models and sample sizes for both Multiome and CITE-seq data in the Supplementary Figures S1, 2, to show the variability of each metric across the 10 replicates of each dataset size.

NMI, as a measure of cluster overlap, reaches values of approx. 0.7 for all Multiome and CITE-seq integrating models. The NMI is slightly lower than what was achieved during the NeurIPS 2021 competition, where the best competition entries reached an NMI of close to 0.8 for the complete Multiome dataset (Lance et al., 2022) (see Supplementary Figure S1). This is to be expected as we evaluate the models in a low sample size scenario. MultiVI profits greatly from a larger cell number, while an increasing cell number only slightly increases the performance of the other models. Across most sample sizes, Cobolt performed best for the CITE-seq datasets, while Portal performs best on the Multiome datasets for all sample sizes but does not profit much from increasing sample size. For larger sample sizes, scMVP shows only slightly worse performance than Portal on the Multiome dataset.

Cell type ASW is a measure of cluster compactness and overlap. We see values of around 0.5 for the Multiome and CITE-seq datasets, which implies overlapping of clusters and only a moderate separation. This is slightly lower than the 0.6 that models have achieved in the NeurIPS 2021 competition (Lance et al., 2022). For Multiome data, the impact of cell numbers was minor in Cobolt, scMM, Portal, DAVAE and scMVP representations and higher for MultiVI. For CITE-seq data, only scMM and TotalVI show a dependence between cell type ASW and cell number. As expected, increasing the number of cells leads to a decrease in variance.

The trajectory conversation score measures the preservation of a biological signal, e.g. in the form of developmental processes. For the CITE-seq dataset, all models reach comparable scores of

FIGURE 2
Biological preservation metrics. NMI, cell type ASW, and trajectory conservation score indicate the biological preservation quality achieved by joint embeddings from various models. On the 10x Multiome data, the performance of Portal, scMVP, Cobolt, scMM, DAVAE, MultiVI is shown (top), whereas the performance on CITE-seq data is shown for Cobolt, TotalVI, SCALEX, and scMM (bottom). Median scores across all iterations are shown inside the tiles.

around 0.9 irrespective of the cell numbers, with a substantial decrease in variance for larger cell numbers. In contrast, for the Multiome dataset, an increase in cell numbers affects the trajectory conservation score for all models except DAVAE. In particular MultiVI shows a large improvement in performance with increasing cell numbers, while for Portal, scMVP, Cobolt and scMM, the scores increase from around 0.87 to around 0.96. Cobolt performs best for higher cell numbers, while the performance of Portal and scMVP is on par and slightly better than Cobolt for lower cell numbers. The maximum score that models reach in our analysis slightly exceeds the median of the trajectory conservation scores of around 0.9 achieved by models of the NeurIPS 2021 competition (Lance et al., 2022).

Taken together, Cobolt is the strongest performing model based on almost all biology preservation metrics on the CITE-seq data and regarding cell type ASW on the Multiome data, performing well even in scenarios with small sample sizes. Portal is the strongest performing model on the Multiome data based on NMI and trajectory conservation and performs well on cell type ASW, also showing consistently high performance across sample sizes.

6.2 Removing technical effects

We assess the removal of technical artifacts based on the batch ASW and graph connectivity score (Figure 3). As a measure of between-site technical variation and to account for the shortcomings of batch ASW (which does not sufficiently account for the nested batch effects of donors and sites) and graph connectivity (which is not sufficiently challenging) (Lance et al., 2022), we also assess batch ASW with the site as a covariate ('Site ASW'), as has been suggested by Lance et al. (2022).

The Batch ASW score of around 0.8 that we observe in our results indicates only a minor batch effect, although the score is slightly lower than the 0.9 that models achieved in the course of the NeurIPS 2021 competition (Lance et al., 2022) (see Supplementary Figure S2). There is a slight increase in performance for increasing cell numbers across both datasets. For the Multiome dataset, Portal consistently performed best, closely followed by scMVP in particular for larger cell numbers, while MultiVI scored lowest for most cell number settings. For the CITE-seq dataset, SCALEX shows the highest Batch ASW score across all cell number settings, implying superior handling of batch effects even with small sample sizes. This is in line with SCALEX being specifically designed to separate batch-related from batch-invariant components (Xiong et al., 2021).

The graph connectivity score indicates how well cells of the same cell type and cells coming from different batches are connected in the joint embedding. For the Multiome dataset, MultiVI's graph connectivity score is considerably lower for small sample sizes, while all models improve performance with an increasing number of cells. Portal and scMVP are the best performing models, reaching a score of almost 1 for higher cell numbers in the case of the Multiome dataset in line with the scores achieved by the models of the NeurIPS competition (Lance et al., 2022). For the CITE-seq dataset, the performance of

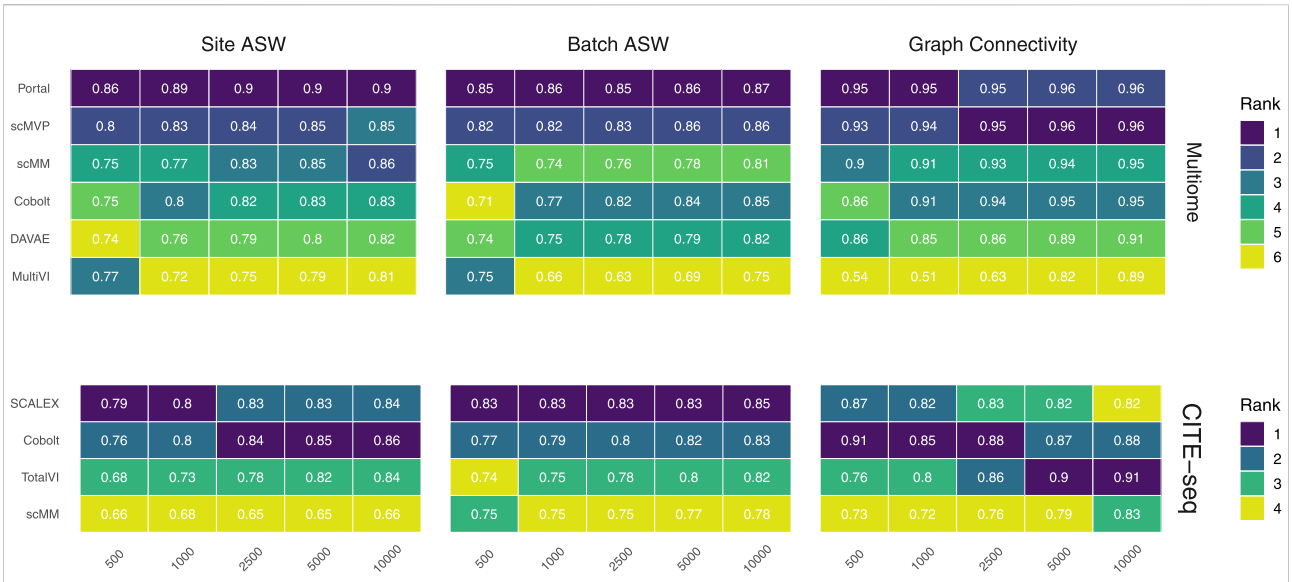


FIGURE 3
Technical effect removal metrics. Site ASW, batch ASW, and graph connectivity score indicate the quality of technical effects removal achieved by joint embeddings from various models. On the 10x Multiome data, the performance of Portal, scMVP, scMM, Cobolt, DAVAE, and MultiVI is shown (top), whereas the performance on CITE-seq data is shown for the SCALEX, Cobolt, TotalVI, and scMM (bottom). Median scores across all iterations are shown inside the tiles.

TotalVI increased with increasing cell numbers, achieving the highest graph connectivity score for 5,000 and 10,000 cells. In contrast, the number of cells had only a minor effect on the other models. scMM consistently had the lowest graph connectivity score for the CITE-seq dataset.

Site ASW captures site-specific batch effects. Compared to Batch ASW, the performance differences between the models that we applied to the CITE-seq dataset are enhanced. For the CITE-seq dataset, Cobolt and SCALEX perform best, with Cobolt surpassing SCALEX for increasing cell numbers. scMM consistently has the lowest scores on the CITE-seq data. For the Multiome data, the spread of the investigated models is comparable to the one of Batch ASW. Portal achieved the highest Site ASW scores followed by scMVP, which is in agreement with their high Batch ASW score.

Portal and scMVP are the best performing models for metrics considering the removal of technical effects on the Multiome data, whereas MultiVI's performance suffers. On the CITE-seq data, SCALEX and Cobolt are among the best performing models, while scMM shows consistently low scores across metrics and cell numbers.

6.3 Usability

The scMM model by (Minoura et al., 2021) was easily usable. The authors provide both a command line interface and a script that is straightforward to adapt and run. However, HDF5-based data (such as the popular "AnnData" objects) has to be manually restructured to separate files to be used as input for the model. For CITE-seq-data, model training did not always converge, in particular for larger sample sizes, which could be addressed by lowering the learning rate and changing the batch size. While this behavior did not occur with very small learning rates (2 orders of magnitude smaller than the default used by Minoura et al. (2021)), this also tended to substantially lower the performance.

To run the scMVP model by (Li et al., 2022), package dependency issues had to be resolved manually. Here, too, data had to be restructured manually to fit the custom input data structs defined by the authors. Adapting and running the model and extracting the learned embedding was straightforward.

All in all, all investigated tools were relatively easy to use and adapt, though in most cases not without at least intermediate programming skills (e.g., to transform own data into rather specific and often largely undocumented data structs defined by the authors).

Finally, looking at the time the tools need for their calculations, we found that the central processing unit (CPU) time (without preprocessing) of Cobolt considerably exceeds the CPU time of the other tools especially for the Multiome dataset (Supplementary Figures S1, 2). Of note, the tools were run on different machines, which hinders a direct comparison. However, it should give the reader a rough idea about the processing time each tools requires, and it is useful to see how well the different investigated tools scale timewise for increasing cell numbers.

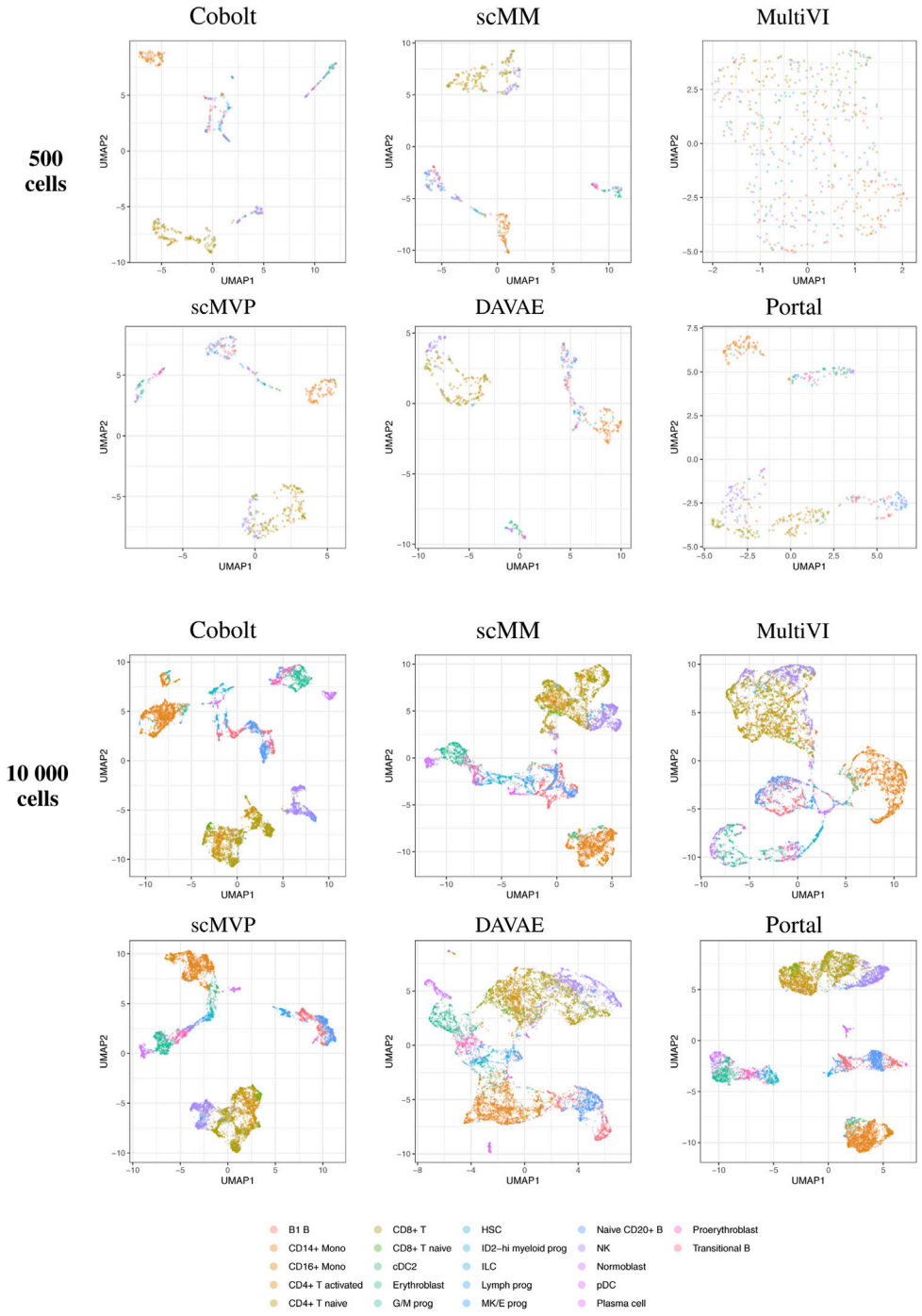


FIGURE 4
UMAP of the 10-dimensional latent space of Cobolt, scMM, MultiVI, scMVP, DAVAE, and Portal based on 500 (top) and 10,000 (bottom) cells of one exemplary subsample from the Multiome dataset each. The color coding corresponds to manually annotated cell types as provided by Luecken et al. (2021a).

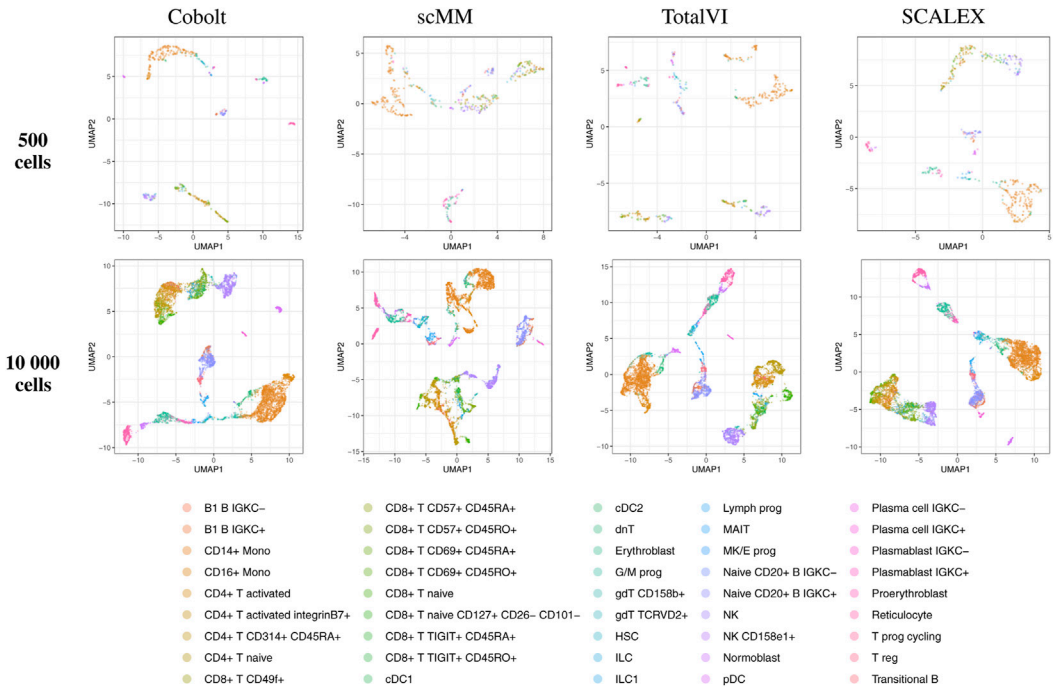


FIGURE 5
UMAP of the 10-dimensional latent space of Cobolt, scMM, TotalVI, and SCALEX based on 500 (top) and 10,000 (bottom) cells of one exemplary subsample from the CITE-seq dataset each. The color coding corresponds to manually annotated cell types as provided by Luecken et al. (2021a). The following cell types are not present in the 500 cell sample: $CD4^+$ T CD314+ CD45RA+, $CD8^+$ T naive CD127+ $CD26^-$ CD101-, cDC1, dnT, Plasma cell IGKC-, Plasma cell IGKC+, Plasmablast IGKC-, T prog cycling.

7 Outlook and discussion

The rapid emergence of experimental protocols for profiling several omics layers from the same cell or in independent experiments is closely followed by the development of corresponding computational models for analyzing and integrating such data. These methods promise to answer biological questions previously out of reach. Still, they have so far been hampered by often rather small and sparse datasets and the lack of a systematic overview and comparison. In particular, considering the sparsity and high dimensionality inherent to single-cell (multi-)omics data, researchers seek to identify a low-dimensional embedding that integrates the information from multiple modalities and can be used for further downstream analyses. Consequently, many computational tools to infer such a joint latent representation have recently been proposed, often based on deep learning approaches due to their success in identifying complex structures from data in unsupervised settings. Specifically, deep generative models such as VAEs that infer a low-dimensional, compressed representation of the input data in an unsupervised way are among the most popular solutions, often including additional components or custom architectures to accommodate the properties of single-cell multi-omics data and facilitate specific characteristics of the learned embedding.

Due to the rapidly growing number of complex methodological proposals for solving the challenging task of computationally integrating multi-omics data, an overview and categorization of such models are essential for understanding the advantages and disadvantages of the different methods. We have compiled a comprehensive review of the literature on DGMs for learning joint embeddings of multi-omics data and categorized the different models according to their architectural choices.

In addition to this overview, we have also illustrated the robustness of selected models to small sample sizes, where sample size refers to the number of cells in the dataset. For evaluating model performance, we have relied on the guidelines of a comprehensive benchmarking project (Luecken et al., 2021a). We have evaluated the models based on established metrics concerning their ability to adjust for technical effects while maintaining biological signals. Our analyses have shown that Cobolt, an approach that uses a multimodal VAE with products of experts to combine individual embeddings, and Portal, an approach that uses the principal components of a

joint PCA on both modalities as input to an autoencoder with an adversarial training strategy, deliver the best performance for most biological preservation metrics, particularly for small numbers of cells. On the other hand, Portal and scMVP, an approach that employs attention-based components and a dedicated architecture to deal with the sparsity of scATAC-seq data, score highest for metrics related to removing technical artifacts on the 10x Multiome data, while SCALEX performs best on the CITE-seq data.

To consider the usability of the approaches from the perspective of a user who is not an expert in tuning deep learning models, we employed the default hyperparameters of the models as proposed by their original authors. While this could potentially introduce bias and dedicated tuning of hyperparameters might improve the results, our focus was on comparing the different approaches relative to each other and relative to the sample size of the respective dataset rather than absolute values of a metric which might be improved by hyperparameter tuning.

Especially for users with little programming experience, some of the models investigated will be difficult to apply, as they require, e.g., the use of command line tools. Here, libraries such as scvi-tools (Gayoso et al., 2021a) offer a significant benefit by providing extensive documentation and exemplary applications.

Interpretability is an aspect that is of great importance for the application of DGMs (Treppner et al., 2022). Some of the models we have reviewed already offer the possibility of making the corresponding outputs interpretable for users. For example, post-hoc methods such as applying archetypal analysis (Cutler and Breiman, 1994) to the joint embedding as conducted by TotalVI (Gayoso et al., 2021b), can make the models explainable after they have been trained. On the other hand, model-based interpretability can be directly incorporated into the model architecture to allow for immediate interpretation, such as the latent traversals and specification of a dedicated prior to facilitate disentanglement in (Minoura et al., 2021). However, no dominant approach has yet emerged in this area, providing scope for new developments.

We would like to stress that our review should not be understood as a comprehensive benchmark but rather as an illustrative case study, as we merely looked at the investigated DGM tools in the scope of representative examples of the landscape of state-of-the-art approaches, with a focus on potential differences in the number of cells they require to perform well.

In this work, we merely discussed some of all available omics modalities, and the performance of the models may be affected for the better or the worse if applied to other data types due to differing data characteristics, e.g., in the degree of sparsity.

The performances we obtained by running the investigated tools on a benchmark dataset may well deviate if applying those tools to other datasets of differing biological backgrounds, e.g., in terms of cell type composition, tissue types, *etc.* Although a focus on specific cell types is beyond the scope of our review, we invite others to use our findings as a stepping stone to explore the performance of DGMs for specific biological scenarios.

In the future, linking information from measurements of transcriptomes, epigenomes, proteomes, chromatin organization, *etc.*, could lead to a deeper understanding of cellular processes. Scientists could then further enhance their understanding of these processes by information on the spatial context.

Author contributions

EB, MH, and MT conceived the idea for the manuscript, conducted the analyses, and wrote the manuscript. CK and HB contributed to writing and proofread the manuscript. All authors read and approved the final manuscript.

Funding

The work of MH is funded by the DFG (German Research Foundation)—322977937/GRK2344. MT and HB are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 431984000 - SFB 1453. CK is funded by the German Ministry of Education and Research by grant EA: Sys [FKZ031L0080]. CK and EB are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (CIBSS-EXC-2189-2100249960-390939984). We acknowledge support by the Open Access Publication Fund of the University of Freiburg.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.962644/full#supplementary-material

References

Adossa, N., Khan, S., Rytkönen, K. T., and Elo, L. L. (2021). Computational strategies for single-cell multi-omics integration. *Comput. Struct. Biotechnol. J.* 19, 2588–2596. doi:10.1016/j.csbj.2021.04.060

Amodio, M., and Krishnaswamy, S. (2018). "Magan: Aligning biological manifolds," in *International conference on machine learning* (PMLR), 215

Amodio, M., Youlten, S. E., Venkat, A., San Juan, B. P., Chaffer, C., and Krishnaswamy, S. (2022). *Single-cell multi-modal gan (scmmgan) reveals spatial patterns in single-cell data from triple negative breast cancer*. bioRxiv.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878. doi:10.15252/msb.20156651

Ashuach, T., Gabitto, M. I., Jordan, M. I., and Yosef, N. (2021). *Multivi: Deep generative model for the integration of multi-modal data*. bioRxiv.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473*.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nat. Biotechnol.* 37, 38–44. doi:10.1038/nbt.4314

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi:10.1080/01621459.2017.1285773

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM* 55, 77–84. doi:10.1145/2133806.2133826

Cao, K., Bai, X., Hong, Y., and Wan, L. (2020). Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* 36, i48–i56. doi:10.1093/bioinformatics/btaa443

Cao, K., Hong, Y., and Wan, L. (2021). Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics* 38, 211–219. doi:10.1093/bioinformatics/btab594

Cao, Z.-J., and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* 40, 1458–1466. doi:10.1038/s41587-022-01284-4

Colomé-Tatché, M., and Theis, F. J. (2018). Statistical single cell multi-omics integration. *Curr. Opin. Syst. Biol.* 7, 54–59. doi:10.1016/j.coisb.2018.01.003

Cutler, A., and Breiman, L. (1994). Archetypal analysis. *Technometrics* 36, 338–347. doi:10.1080/00401706.1994.10485840

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. (2022). Scot: Single-cell multi-omics alignment with optimal transport. *J. Comput. Biol.* 29, 3–18. doi:10.1089/cmb.2021.0446

Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9, 2002–2013. doi:10.1038/s41467-018-04368-5

Erfanian, N., Heydari, A. A., Iañez, P., Derakhshani, A., Ghasemigol, M., Farahpour, M., et al. (2021). *Deep learning applications in single-cell omics data analysis*. bioRxiv.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Wu, K., Jayasuriya, M., et al. (2021a). *Scvi-tools: A library for deep probabilistic analysis of single-cell omics data* Cold Spring Harbor Laboratory.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., et al. (2021b). Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nat. Methods* 18, 272–282. doi:10.1038/s41592-020-01050-x

Gong, B., Zhou, Y., and Purdom, E. (2021). Cobolt: Integrative analysis of multimodal single-cell sequencing data. *Genome Biol.* 22, 351–421. doi:10.1186/s13059-021-02556-z

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. neural Inf. Process. Syst.* 27.

Grapov, D., Fahrmann, J., Wanichthanarak, K., and Khoomrung, S. (2018). Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. *Omics a J. Integr. Biol.* 22, 630–636. doi:10.1089/omi.2018.0097

Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., and Winther, O. (2020). scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics* 36, 4415–4422. doi:10.1093/bioinformatics/btaa293

Hu, J., Zhong, Y., and Shang, X. (2022). A versatile and scalable single-cell data integration algorithm based on domain-adversarial and variational approximation. *Brief. Bioinform.* 23, bbab400. doi:10.1093/bib/bbab400

Kang, M., Ko, E., and Mersha, T. B. (2021). A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform.* 23, bbab454. doi:10.1093/bib/bbab454

Khan, S. A., Lehmann, R., Martinez-de Morentin, X., Ruiz, A. M., Lagani, V., Kiani, N. A., et al. (2022). *scaegan: Unification of single-cell genomics data by adversarial learning of latent space correspondences*. bioRxiv. doi:10.1101/2022.04.19.488745

Kharchenko, P. V. (2021). The triumphs and limitations of computational methods for scrna-seq. *Nat. Methods* 18, 723–732. doi:10.1038/s41592-021-01171-x

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. (2017). "Structured attention networks," in 5th International Conference on Learning Representations (ICLR) 2017, Toulon, France, April 24–26, 2017

Kingma, D. P., and Welling, M. (2019). *An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691*.

Kingma, D. P., and Welling, M. (2013). *Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114*.

Konopka, T., and Konopka, M. T. (2018). *R-Package: Umap*. Uniform Manifold Approximation and Projection.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21, 31–35. doi:10.1186/s13059-020-1926-6

Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A. C., et al. (2022). *Multimodal single cell data integration challenge: Results and lessons learned*. bioRxiv.

Li, G., Fu, S., Wang, S., Zhu, C., Duan, B., Tang, C., et al. (2022). A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome Biol.* 23, 20–23. doi:10.1186/s13059-021-02595-6

Lin, Y., Wu, T.-Y., Wan, S., Yang, J. Y., Wong, W. H., and Wang, Y. (2022). Scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nat. Biotechnol.* 40, 703–710. doi:10.1038/s41587-021-01161-6

Liu, Q., Chen, S., Jiang, R., and Wong, W. H. (2021). Simultaneous deep generative modeling and clustering of single cell genomic data. *Nat. Mach. Intell.* 3, 536–544. doi:10.1038/s42256-021-00333-y

Lopez, R., Gayoso, A., and Yosef, N. (2020). Enhancing scientific discoveries in molecular biology with deep generative models. *Mol. Syst. Biol.* 16, e9198. doi:10.15252/msb.20199198

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058. doi:10.1038/s41592-018-0229-2

Lotfollahi, M., Litinetskaya, A., and Theis, F. J. (2022). *Multigrate: Single-cell multi-omic data integration*. bioRxiv.

Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Lance, C., Agrawal, A., and Aliee, H. (2021a). "A sandbox for prediction and integration of dna, rna, and proteins in single cells," in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., et al. (2021b). Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* 19, 41–50. doi:10.1038/s41592-021-01336-8

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell rna-seq analysis: A tutorial. *Mol. Syst. Biol.* 15, e8746. doi:10.15252/msb.20188746

Lynch, A. W., Theodoris, C. V., Long, H. W., Brown, M., Liu, X. S., and Meyer, C. A. (2022). Mira: Joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat. Methods* 19, 1097–1108. doi:10.1038/s41592-022-01595-z

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., et al. (2020). Realistic *in silico* generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nat. Commun.* 11, 166–212. doi:10.1038/s41467-019-14018-z

Minoura, K., Abe, K., Nam, H., Nishikawa, H., and Shimamura, T. (2021). A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell. Rep. Methods* 1, 100071. doi:10.1016/j.crmeth.2021.100071

Peyré, G., and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *FNT. Mach. Learn.* 11, 355–607. doi:10.1561/2200000073

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi:10.1016/j.csbj.2021.06.030

Qin, Q., Fan, J., Zheng, R., Wan, C., Mei, S., Wu, Q., et al. (2020). Lisa: Inferring transcriptional regulators through integrative modeling of public chromatin

accessibility and chip-seq data. *Genome Biol.* 21, 32–14. doi:10.1186/s13059-020-1934-6

Rohatgi, A. (2021). *Webplotdigitizer*. Available at: https://automeris.io/WebPlotDigitizer

Shi, Y., Paige, B., Torr, P., et al. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Adv. Neural Inf. Process. Syst.* 32.

Stanojevic, S., Li, Y., and Garmire, L. X. (2022). *Computational methods for single-cell multi-omics integration and alignment. arXiv preprint arXiv:2201.06725*.

Stark, S. G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., et al. (2020). Scim: Universal single-cell matching with unpaired feature sets. *Bioinformatics* 36, i919–i927. doi:10.1093/bioinformatics/btaa843

Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., and Satija, R. (2021). Single-cell chromatin state analysis with signac. *Nat. Methods* 18, 1333–1341. doi:10.1038/s41592-021-01282-5

Tarazona, S., Arzalluz-Luque, A., and Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat. Comput. Sci.* 1–8, 395–402. doi:10.1038/s43588-021-00086-z

Treppner, M., Binder, H., and Hess, M. (2022). Interpretable generative deep learning: An illustration with single cell gene expression data. *Hum. Genet.* 141, 1481–1498. doi:10.1007/s00439-021-02417-6

Treppner, M., Salas-Bastos, A., Hess, M., Lenz, S., Vogel, T., and Binder, H. (2021). Synthetic single cell rna sequencing data from small pilot studies using deep generative models. *Sci. Rep.* 11, 9403–9411. doi:10.1038/s41598-021-88875-4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in neural information processing Systems*. Editors I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Curran Associates, Inc.), 30.

Veenstra, T. D. (2021a). Omics in systems biology: Current progress and future outlook. *Proteomics* 21, 2000235. doi:10.1002/pmic.202000235

Veenstra, T. D. (2021b). Systems biology and multi-omics. *Proteomics* 21, 2000306. doi:10.1002/pmic.202000306

Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., et al. (2013). Target analysis by integration of transcriptome and chip-seq data with beta. *Nat. Protoc.* 8, 2502–2515. doi:10.1038/nprot.2013.150

Wang, X., Hu, Z., Yu, T., Wang, Y., Wang, R., Wei, Y., et al. (2022). *Contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration*. bioRxiv. doi:10.1101/2021.12.12.472268

Wu, K. E., Yost, K. E., Chang, H. Y., and Zou, J. (2021). Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2023070118. doi:10.1073/pnas.2023070118

Xiong, L., Tian, K., Li, Y., and Zhang, Q. C. (2021). *Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space*. bioRxiv.

Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., et al. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nat. Commun.* 10, 4576–4610. doi:10.1038/s41467-019-12630-7

Xu, Y., Begoli, E., and McCord, R. P. (2021a). *scican: Single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network*. bioRxiv. doi:10.1101/2021.11.30.470677

Xu, Y., Das, P., and McCord, R. P. (2021b). Smile: Mutual information learning for integration of single-cell omics data. *Bioinformatics* 38, 476–486. doi:10.1093/bioinformatics/btab706

Xu, Y., Zhang, Z., You, L., Liu, J., Fan, Z., and Zhou, X. (2020). scigans: single-cell rna-seq imputation using generative adversarial networks. *Nucleic Acids Res.* 48, e85. doi:10.1093/nar/gkaa506

Zhang, R., Meng-Papaxanthos, L., Vert, J.-P., and Noble, W. S. (2022a). "Semi-supervised single-cell cross-modality translation using polarbear," in *Research in computational molecular biology*. Editor I. Peer (Cham: Springer International Publishing), 20–35.

Zhang, Z., Yang, C., and Zhang, X. (2022b). *Integrating unmatched scrna-seq and scatac-seq data and learning cross-modality relationship simultaneously*. bioRxiv. doi:10.1101/2021.04.16.440230

Zhao, J., Wang, G., Ming, J., Lin, Z., Wang, Y., Consortium, T. T. M., et al. (2022). Adversarial domain translation networks for fast and accurate integration of large-scale atlas-level single-cell datasets. *Nat. Comput. Sci.* 2, 317–330.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2223–2232.

Zuo, C., and Chen, L. (2021). Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief. Bioinform.* 22, bbaa287. doi:10.1093/bib/bbaa287

Zuo, C., Dai, H., and Chen, L. (2021). Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. *Bioinformatics* 37, 4091–4099. doi:10.1093/bioinformatics/btab403

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 26 October 2022
DOI 10.3389/fmolb.2022.962431

OPEN ACCESS

EDITED BY
Ornella Cominetti,
Nestlé Research Center, Switzerland

REVIEWED BY
José Camacho,
University of Granada, Spain
Federico Marini,
Sapienza University of Rome, Italy

*CORRESPONDENCE
Anders Hagen Jarmund,
anders.h.jarmund@ntnu.no

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 06 June 2022
ACCEPTED 20 September 2022
PUBLISHED 26 October 2022

CITATION
Jarmund AH, Madssen TS and Giskeødegård GF (2022), ALASCA: An R package for longitudinal and cross-sectional analysis of multivariate data by ASCA-based methods.
Front. Mol. Biosci. 9:962431.
doi: 10.3389/fmolb.2022.962431

ALASCA: An R package for longitudinal and cross-sectional analysis of multivariate data by ASCA-based methods

Anders Hagen Jarmund[1,2]*, Torfinn Støve Madssen[3] and Guro F. Giskeødegård[4]

[1]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, [2]Centre of Molecular Inflammation Research (CEMIR), NTNU, Trondheim, Norway, [3]Department of Circulation and Medical Imaging, NTNU, Trondheim, Norway, [4]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Trondheim, Norway

The increasing availability of multivariate data within biomedical research calls for appropriate statistical methods that can describe and model complex relationships between variables. The extended ANOVA simultaneous component analysis (ASCA$^+$) framework combines general linear models and principal component analysis (PCA) to decompose and visualize the separate effects of experimental factors. It has recently been demonstrated how linear mixed models can be included in the framework to analyze data from longitudinal experimental designs with repeated measurements (RM-ASCA$^+$). The ALASCA package for R makes the ASCA$^+$ framework accessible for general use and includes multiple methods for validation and visualization. The package is especially useful for longitudinal data and the ability to easily adjust for covariates is an important strength. This paper demonstrates how the ALASCA package can be applied to gain insights into multivariate data from interventional as well as observational designs. Publicly available data sets from four studies are used to demonstrate the methods available (proteomics, metabolomics, and transcriptomics).

KEYWORDS

R, omics analysis, statistical method, ASCA, longitudinal data analysis, multivariate analysis

1 Introduction

The increasing availability of high-dimensional data through omics-technologies can yield new insights into how intricate biological systems evolve and how they respond to various experimental conditions. However, there is a need for parallel development of novel statistical methods that can deal with the increased complexity of such data. The methods must be valid for multidimensional data sets, flexible for different experimental settings, as well as interpretable. Commonly used methods for multivariate data analysis, such as principal component analysis (PCA) and partial least squares (PLS) regression, are

not able to fully account for more complex experimental designs. Multilevel PLS-DA, for instance, can only handle two time points, and adjusting for confounders can only be handled by subgroup analysis. One powerful approach for analysis of multivariate data is the ANOVA simultaneous component analysis (ASCA) framework that combines ANOVA with PCA (Smilde et al., 2005; Smilde et al., 2012). More recently, extended ASCA methods such as ASCA$^+$ (Thiel et al., 2017), LiMM-PCA, and repeated measures ASCA$^+$ (RM-ASCA$^+$, Martin and Govaerts, 2020; Madssen et al., 2021) have emerged that combine general linear (mixed) models with PCA. In this way the flexibility of regression models are merged with the visualization of multivariate analysis, providing excellent interpretability by allowing to separate and display the complex multivariate patterns originating from different experimental factors. Despite these benefits, the availability of software implementations of ASCA$^+$, and thus the use of the framework, has been limited.

In short, (RM-)ASCA$^+$ comprises three steps: first, linear regression with or without random effects produces regression coefficients (β) which are summarized into a fixed effect parameter matrix (**B**, also including fixed intercepts). In the case of K measurements of J variables in I individuals, the linear mixed model based regression with R random effect coefficients (γ, including intercepts) and p fixed effect coefficients (β, including the intercept) can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E}, \tag{1}$$

where **Y** is an $IK \times J$ response matrix, **X** is an $IK \times p$ design matrix, **B** is a $p \times J$ parameter matrix, **Z** is an $IK \times R$ design matrix for random effects, **U** is an $R \times J$ random parameters matrix, and **E** is an $IK \times J$ residual matrix. Equation 1 can also be written as

$$\begin{matrix}\text{Response matrix} \\ \text{Variables} \rightarrow \\ \text{Observations} \downarrow\end{matrix}\begin{bmatrix} y_{(1,1),1} & y_{(1,1),2} & \cdots & y_{(1,1),J} \\ y_{(1,2),1} & y_{(1,2),2} & \cdots & y_{(1,2),J} \\ \vdots & \vdots & \vdots & \vdots \\ y_{(I,K),1} & y_{(I,K),2} & \cdots & y_{(I,K),J} \end{bmatrix} = \tag{2}$$

$$\begin{matrix}\text{Design matrix} \\ \text{Int} \quad x_1 \quad \ldots \quad x_{p-1}\end{matrix}\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}\begin{matrix}\text{Fixed effects} \\ \end{matrix}\begin{bmatrix} \beta_{0,1} & \beta_{0,2} & \cdots & \beta_{0,J} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{p-1,1} & \beta_{p-1,2} & \cdots & \beta_{p-1,J} \end{bmatrix} +$$

$$\begin{matrix}\text{Design matrix} \\ z_1 \quad z_2 \quad \ldots \quad z_R\end{matrix}\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \ldots \end{bmatrix}\begin{matrix}\text{Random effects} \\ \end{matrix}\begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,J} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,J} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{R,1} & \gamma_{R,2} & \cdots & \gamma_{R,J} \end{bmatrix} + \begin{matrix}\text{Residuals} \\ \end{matrix}\begin{bmatrix} \epsilon_{(1,1),1} & \epsilon_{(1,1),2} & \cdots & \epsilon_{(1,1),J} \\ \epsilon_{(1,2),1} & \epsilon_{(1,2),2} & \cdots & \epsilon_{(1,2),J} \\ \vdots & \vdots & \vdots & \vdots \\ \epsilon_{(I,K),1} & \epsilon_{(I,K),2} & \cdots & \epsilon_{(I,K),J} \end{bmatrix}$$

where the design matrices are filled with custom values for demonstration, $y_{(i,k),j}$ is the kth measurement of variable j in individual i, and $\epsilon_{(i,k),j}$ the corresponding residuals. It will in many cases be sufficient to include a random intercept for participant. **ZU** is then simplified to an $IK \times J$ matrix with one intercept per individual per variable ($\gamma_{r,j} \rightarrow \gamma_{i,j}$), repeated for K rows. The subject-specific random intercepts ($\gamma_{i,j}$) and the residuals ($\epsilon_{(i,k),j}$) are assumed to be normally distributed with mean zero and variations σ_u^2 and σ_e^2, respectively. Ordinary ASCA$^+$ represents the special case when no random effects are included. The second step in RM-ASCA$^+$ is to decompose the **XB** matrix into effect matrices $\mathbf{M}_h$ representing specific parts of the regression model,

$$\mathbf{XB} = \mathbf{M}_0 + \sum_h \mathbf{M}_h. \tag{3}$$

Here, $\mathbf{M}_0$ represents the intercept and is typically of little interest. In ordinary ASCA, $\mathbf{M}_0$ usually represents the grand mean matrix, whereas in RM-ASCA$^+$ it typically either represents the baseline mean of all, or one of the groups, depending on how the effects are coded in the model. The effects h reflect the statistical and experimental design (for examples, *see* Madssen et al., 2021). In the context of a longitudinal study, an effect matrix $\mathbf{M}_T$ would represent the effect of time, i.e., the change from baseline. If the study comprises multiple groups, additional effect matrices describing group differences ($\mathbf{M}_G$) and time-group interaction ($\mathbf{M}_{T:G}$) would be appropriate. Other covariates included in the regression model, such as gender or body mass index (BMI), would also require a separate effect matrix. The final step in RM-ASCA$^+$ is to apply PCA to individual or combined effect matrices, depending on the research question, and extract scores and loadings. The resulting scores and loadings can then be plotted to visualize how variables are affected by the selected effects.

Providing an estimate of uncertainty and robustness is an important feature for all statistical techniques. There is a risk of overfitting when using (RM-)ASCA$^+$, as (RM-) ASCA$^+$ is a supervised method applied to labeled data (Bertinetto et al., 2020). To mitigate the risk of overfitting, the confidence of the estimated scores and loadings from (RM-)ASCA$^+$, reflecting the effects of factors and possibly their interaction, should be tested. Most common are resampling methods such as bootstrap, jack-knife and permutation (Vis et al., 2007; Bertinetto et al., 2020). The latter involves random shuffling of the data labels before applying (RM-)ASCA$^+$, often 1,000–10,000 times. As no systematic relationships should exist in the data when measurements are shuffled across experimental conditions, it establishes null-distributions for scores, loadings, or other metrics. A p-value can then be calculated by comparing the metric from the unaltered model to the null-distributions. While exact permutation tests exist for main effects, only approximate tests are available for interaction effects (Anderson and Braak, 2003; Bertinetto et al., 2020). In contrast to the permutation test, the bootstrap and jack-knife methods conserve the data labels. Here, the robustness of the metrics are tested by applying (RM-) ASCA$^+$ to either a subset of the original data set, where a proportion of the participants are excluded (jack-knife), or a resampled data set, where individual participants are selected at random with replacement (bootstrap). When this is repeated in the order of 1,000–10,000 times, confidence intervals can be estimated for the scores and loadings by extracting upper and

lower percentiles from the results of the resampled data sets. Multiple strategies exist for permutation testing (Anderson and Braak, 2003), and their suitability for RM-ASCA$^+$ models with various designs is currently under investigation.

The Assorted Linear functions for ASCA (ALASCA) package for R has been developed to make the ASCA$^+$ and RM-ASCA$^+$ frameworks accessible for the general researcher. The package does not require advanced programming skills and is publicly available from the Github code repository (https://github.com/andjar/ALASCA). Although the ALASCA package supports both ASCA$^+$ and RM-ASCA$^+$ analysis, the main focus of this paper will be analysis of repeated measures of multivariate data with RM-ASCA$^+$ due to the increasing need for flexible methods to deal with longitudinal experimental designs. The package utilizes well-known R syntax for defining regression models, offers options for predefined or custom scaling, includes multiple validation methods (jack-knifing and bootstrapping), and produces publication-ready figures. While the package is designed to be easy to use, it provides a wide range of customizable options available for advanced users. Further, the package includes several options for exporting the resulting models for archival, post-processing, external visualization, or sharing. Earlier versions of the ALASCA package has been used to reveal how serum cytokine levels change throughout pregnancy in healthy women (Jarmund et al., 2021) and in women with polycystic ovary syndrome (Stokkeland et al., 2022), and to show how the cytokine development is sensitive to maternal and fetal factors. The flexibility of the RM-ASCA$^+$ framework was crucial for the combination of multiple cohorts and for making complex relationships available for interpretation. Since then, the package has been further developed for general use and includes new functions for validation and visualization.

In this paper, we demonstrate how the ALASCA package can be used to analyze various multivariate omics-data using RM-ASCA$^+$. Three publicly available data sets are analyzed to illustrate each modeling step, including appropriate choice of scaling, model setup, and validation technique, and to demonstrate how the results can be easily visualized and interpreted. The data sets are diverse in terms of biological level (proteomics, metabolomics, transcriptomics) and experimental design (repeated measures within observational and randomized-controlled intervention studies). This practical and integrated approach will demonstrate the flexibility of the ALASCA package for data exploration and analysis.

1.1 Related works

Previous implementations of ASCA and ASCA-related methods exist for several common statistical software such as R and Matlab (Bertinetto et al., 2020). The first implementation of ASCA was published as Matlab scripts by Smilde et al. (2005). For R, the earliest implementations include ASCA-genes (Nueda et al., 2007, the scripts are no longer available) and the lmdme package (Fresno et al., 2014). Later options include MetStaT (removed from CRAN but available as archive https://cran.r-project.org/src/contrib/Archive/MetStaT/) for R and the PLS_toolbox and MetaboAnalyst (Xia et al., 2015) for Matlab (Bertinetto et al., 2020).

The multiblock package for R offers a comprehensive set of methods for multiblock analysis, including various ASCA-based methods such as LiMM-PCA, generalized ASCA, RM-ASCA$^+$, and covariates ASCA (Liland, 2022; Smilde et al., 2022). A Matlab implementation of RM-ASCA$^+$ has been published by Madssen et al. (2021), (scripts available at https://github.com/ntnu-mr-cancer/RM_ASCA). An extension of RM-ASCA$^+$ has been proposed in the case of zero-inflated count data, namely the zero-inflated counts (ZIC)RM-ASCA$^+$ by applying zero-inflated negative binomial mixed models, with code available for R (https://github.com/AukeHaver/ZICRM-ASCA_plus).

The ALASCA package offers several distinct features compared to existing implementations such as integrated scaling and validation, option to force equal baseline (important for randomized designs), supports both sum and contrast coding, precise yet simple specification of effect matrices, and diverse options for visualization.

2 Materials and methods

2.1 Package overview

The main functions of the ALASCA package are described in Table 1 and a typical work flow is illustrated in Figure 1. The `ALASCA()` function is used to define the regression model, scaling, and validation strategy. The resulting ALASCA object can then be visualized in several ways.

The `ALASCA()` function accepts a range of arguments related to the regression model and validation (Table 2). Recommended arguments for various study designs and research questions are demonstrated in the examples below. ALASCA will fit linear mixed models if the regression formula contains terms with | such as `(1|ID)` (i.e., random effects) and ordinary linear regression models otherwise. Regression coefficients are estimated with one of three algorithms, depending on the specific model to be fitted, namely, the Rfast package (Papadakis et al., 2021), the lme4 package (Bates et al., 2015), or base lm (R Core Team, 2020). Coefficients are estimated by Rfast as default due to performance, but Rfast has some limitations on which regression models can be fitted. Therefore, lme4 and lm can be used as alternatives when more complex regression models are used. The two latter can be applied by specifying

TABLE 1 Important functions in the ALASCA package.

Function	Description
`ALASCA()`	Initialize and create the ALASCA model
`flip()`	Invert the signs of scores and loadings
`plot(..., type = "effect")`	Plot scores and loadings from a model
`plot(..., type = "prediction")`	Plot marginal means from the underlying regression models
`plot(..., type = "validation")`	Plot score and loading for all validation runs
`plot(..., type = "histogram")`	Plot score and loading for all validation runs as histograms
`plot(..., type = "residuals")`	Plot regression residuals
`plot(..., type = "covars")`	Plot regression coefficients of covariates
`plot(..., type = "2D")`	Plot the main results of the model
`plot(..., type = "participants")`	Plot measurements from individual participants
`summary()`	Returns key information about the model
`get_scores()`	Returns the scores of the model
`get_loadings()`	Returns the loadings of the model
`get_covars()`	Returns additional regression coefficients
`get_predictions()`	Returns marginal means from the model

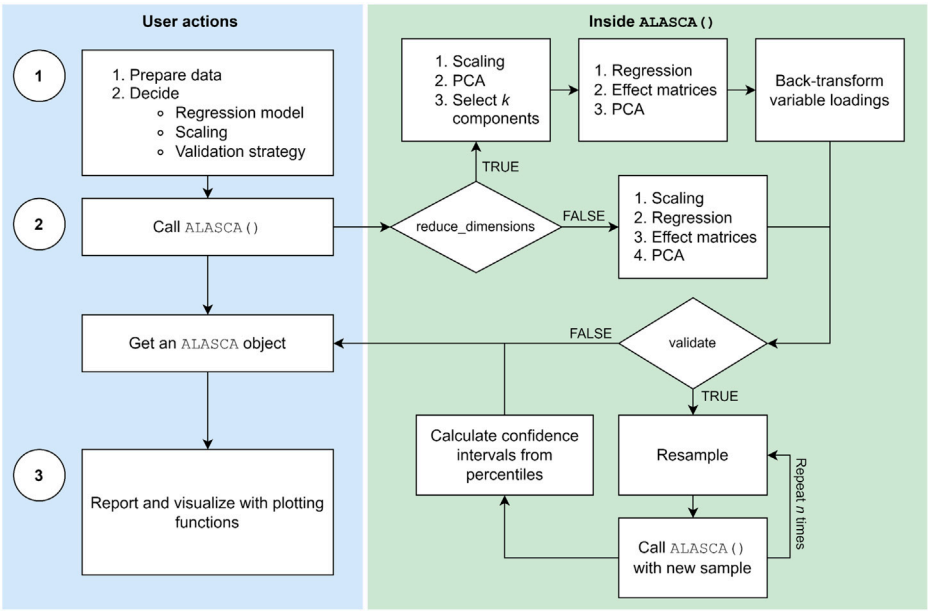


FIGURE 1
The typical workflow in the ALASCA package involves three stages: (1) Preparation, (2) execution, and (3) visualization. When the user has prepared the data and decided regression model, scaling, and validation strategy, the `ALASCA()` function is called. The `ALASCA()` function will then scale the data, perform regression analyses, apply principal component analysis (PCA) to the effect matrices, and extract loadings and scores. The option `reduce_dimensions = TRUE` will use PCA to reduce the number of variables to k, and loadings are automatically transformed back to the original variable space. Validation is performed if `validate = TRUE` is specified. The validation consists of performing (RM-)ASCA$^+$ on n resampled data sets, and using percentiles for loadings and scores for confidence intervals. When the model is constructed, the user can visualize and report results in various ways.

`use_Rfast = FALSE` and will also produce p-values and additional information such as R^2 for each regression model. When lme4 is used, p-values are calculated with Satterthwaite's degrees of freedom method with the lmerTest package (Kuznetsova et al., 2017). The data.table package is extensively used to improve performance by doing data

TABLE 2 Important arguments for the `ALASCA()` function. A full list of arguments can be shown in R using `?ALASCA::ALASCA()`.

Function	Default	Description
`df`	—	Data frame containing the data set to be analyzed
`formula`	—	Regression formula
`scale_function`	`"sdall"`	Function to scale data. See description of possible defaults in the text
`separate_effects`	`FALSE`	When `TRUE`, separate effect terms
`equal_baseline`	`FALSE`	When `TRUE`, remove interaction at baseline
`validate`	`FALSE`	When `TRUE`, validate the model
`reduce_dimensions`	`FALSE`	When `TRUE`, use principal component analysis to reduce the number of variables
`wide`	`FALSE`	Set to `TRUE` if data are provided in wide format
`stratification_column`	`NULL`	Name of the column to be used for stratification during validation. By default, use `group` or first the effect term
`validation_method`	`"bootstrap"`	Set to `"jack-knife"` to use jack-knife resampling for validation
`n_validation_runs`	`1000`	Number of validation runs
`save`	`FALSE`	When `TRUE`, automatically save the model and subsequent plots
`limitsCI`	`c(0.025, 0.975)`	Lower and upper percentiles for confidence intervals

manipulation by reference and other optimizations (Dowle and Srinivasan, 2021). ALASCA objects are also manipulated by reference with help of the R6 package (Chang, 2021). Traditionally in R, functions will not modify variables in place but requires that variables are reassigned. ALASCA objects, however, can be modified without re-assignment. For instance, both `flip(model)` and `model <- flip(model)` will modify the `model` object.

Currently, model validation can be performed with cluster bootstrap or jack-knife, both with stratification. During validation, the `ALASCA()` function will call itself using a modified data set `n_validation_runs` times (Figure 1). The default is 1,000 runs. If cluster bootstrap is selected (default), each participant is replaced by a randomly selected participant from the same stratification group, with replacement, and all measurements from the sampled participant are added to the modified data set. If jack-knife is chosen, the stratification groups are iterated and one out of q (defaults to $q = 7$) participants are excluded at random from the iteration. By default, any column named `group` in the data set `df` will be used for stratification, i.e., the relative group sizes are kept during validation. Alternatively, another column in `df` can be specified for stratification as `stratification_column`. If there is no `group` column and `stratification_column` is not specified, the first effect term will be used for stratification. Loadings from the validation runs are rotated towards loadings from the initial run using procrustes rotation, and the rotation matrix is applied to the scores from the validation run as well. As the sign of loadings and scores in PCA is arbitrarily defined, `ALASCA()` will test whether changing the signs of each principal component (PC) improves the fit of the scores from validation runs and the initial run, and choose the signs minimizing the summed distance of the scores. Only PCs explaining more than 5% variance are used for rotation. Finally, 95% confidence intervals (CIs) are calculated for scores and loadings by selecting the 2.5% and 97.5% percentiles from the validation runs.

Visualizations are made within the popular ggplot2 framework (Wickham, 2016; Kassambara, 2020; Slowikowski, 2021). The default color palette for figures is the viridis palette which is designed to be readable and perceptually uniform despite gray scale printing and the most common forms of color blindness (Wickham and Seidel, 2020; Garnier et al., 2021). Custom ggplot2 themes can be used by specifying `plot.my_theme`. If `save = TRUE` was used during initialization of the model, the `plot()` function will automatically save all plots that are produced.

For megavariate data sets, the large number of measured variables makes individual regression too time consuming for validation with sufficient numbers of iterations. If `reduce_dimensions = TRUE`, `ALASCA()` will perform an initial PCA on the measurements, prior to regression, so that the original variables are replaced by PCs (Figure 1), similar as for Limm-PCA (Martin and Govaerts, 2020). The number of PCs kept from the initial PCA is selected so that 95% of the variance in the measurements is explained. The limit can be changed by specifying `reduce_dimensions.limit`. Additionally, one can prevent ALASCA from running out of memory by saving results from the validation runs directly to a duckdb or sqlite3 database instead of keeping all the results in memory with `save_to_disk = TRUE` (R Special Interest Group on Databases et al., 2021; Müller et al., 2021; Mühleisen and Raasveldt, 2022).

Logging of important events, such as estimated time for validation or error messages, is performed with the log4r package and written to file by default (White and Jacobs, 2021).

2.2 Installation and data preparation

The ALASCA package is freely available at the Github code repository and can be installed in R with the following commands:

```
install.packages("devtools")
devtools:install_github("andjar/ALASCA", ref = "main")
```

Version 1.0.0 of ALASCA was used for this paper. The code to reproduce all results in this paper, including data preparation and figures, can be found in the supplementary materials, and simplified function calls are given below. The full code in the supplementary materials utilizes additional packages such as here and reshape2.

The `ALASCA()` function requires at minimum a data frame or data table `df` and a regression formula. Generally, data can be organized in two formats (Supplementary Figure S1): long (all measured variables have separate rows) or wide (all observations have separate rows, with the different variables as separate columns). If data are provided to `ALASCA()` in long format with one row for each measured variable (Supplementary Figure S1A and examples 1 and 3 below), the variable names (i.e., the measured variables) must be in a column named `variable`. If wide format is used (one row per measured sample, with variables as separate columns, Supplementary Figure S1B and example 2 below), `wide = TRUE` must be provided to `ALASCA()` and all columns not mentioned in the formula or being specified otherwise (Table 2) will be treated as columns containing measurements of interest. At least two other columns are required, regardless of format: One column must contain an identifier for the experimental unit, typically the study numbers of the participants. By default, this column is either derived from the random intercept in the formula or, in case there are no or multiple random intercepts in the formula, it is assumed to be named `ID`. If another column is to be used, it must be specified as `participant_column`. Secondly, one column must contain the first effect of interest and will be used to label the *x*-axis in subsequent score plots. By default, this is assumed to be the first term in the formula. If another column is to be used, it must be specified as `x_column`. General data preparation is demonstrated in the supplementary files. For example, the function call

```
ALASCA(
df,
formula = value ~ v1 + v2 + (1|ID),
validate = TRUE)
```

will assume that the provided data (`df`) is organized in long format (Supplementary Figure S1A) and includes the columns `variable`, `value`, `v1`, `v2`, and `ID` (random intercept). The regression formula `value~v1 + v2 + (1|ID)` corresponds to a model with `value` as outcome, `ID` as random intercept, and `v1` and `v2` as main effect terms. Bootstrap validation will also be applied as `validate = TRUE` with 1,000 iterations (default). If `df` contains a column called `group`, the observations will be stratified by `group` during bootstrapping, otherwise they are statified by `v1`. Since scaling has not been specified (*see* below), the outcome data will be scaled by the default method (i.e., division by the standard deviation, by variable).

The effects of interest can be specified (e.g., `effects = c("v1", "v1:v2")` where `v1`, `v2`, ... are terms in the regression formula) or inferred by ALASCA. In the latter case, the first formula term is assumed to be of interest. Next, ALASCA will look for an interaction term, and, if it exists, include the interaction and second main effect. For example, if the formula is `value~v1*v2 + v3 + (1|ID)`, ALASCA will assume that `v1`, `v2`, and `v1:v2` (interaction) are all of interest. How they are combined depends on `separate_effects`. If `separate_effects = FALSE` (default), only one combined effect is extracted (i.e., `v1*v2` or `v1+v2+v1:v2`). If `separate_effects = TRUE`, two separate effect matrices will be produced: `v1` and `v2+v1:v2`. ALASCA will explicitly state which effects that are assessed when ran.

Columns representing effects of interest, typically the `time` and `group` columns, are expected to contain factors, i.e., categorical data with ordered levels. For example, `df$group <- factor(df$group)` will convert the group column to factors with the factor levels ordered alphabetically. The first levels of time and group are used as baseline or reference group. Level order can be specified explicitly, `factor(..., levels = c("Male," "Female"))`, or by specifying just the reference, `relevel(..., ref = "Male")`.

The data should not be normalized or scaled as part of the preparation. Instead, a scaling function must be specified and provided to the `ALASCA()` function. This prevents data leak during validation where a subset of the data set is used to determine scaling factors that are independently applied to the remaining data for validation. Four predefined options are currently available (Timmerman et al., 2015):

- `scale_function = "sdall"` will divide the value column by the standard deviation of all samples, by variable:

$$\hat{y}_{(\cdot,\cdot),j} = y_{(\cdot,\cdot),j} \Big/ \mathrm{SD}\left(y_{(\cdot,\cdot),j}\right)$$

- `scale_function = "sdt1"` will divide the value column by the standard deviation of all baseline samples, by variable:

$$\hat{y}_{(\cdot,\cdot),j} = y_{(\cdot,\cdot),j} \Big/ \mathrm{SD}\left(y_{(\cdot,k),j}\right), \quad k = 1$$

- `scale_function = "sdref"` will divide the value column by the standard deviation of all samples in the reference group, by variable:

$$\hat{y}_{(\cdot,\cdot),j} = y_{(\cdot,\cdot),j} \Big/ \mathrm{SD}\left(y_{(i,\cdot),j}\right), \; i \in \text{Reference group}$$

- `scale_function = "sdreft1"` will divide the value column by the standard deviation of all baseline samples in the reference group, by variable:

$$\hat{y}_{(\cdot,\cdot),j} = y_{(\cdot,\cdot),j} \Big/ \mathrm{SD}\left(y_{(i,k),j}\right),\ i \in \text{Reference group},\quad k = 1$$

where SD refers to the standard deviation, $\hat{y}_{(i,k),j}$ is the scaled and $y_{(i,k),j}$ the raw value of variable j for individual i at time point k (*see* Eq. 2). Mean centering is by default performed before scaling. In addition, a custom scaling function can be provided. The scaling function should have the data frame as argument and return a data frame with scaled values:

```
scale_function <- function(df){
... # Scale the value column
return(df) }
```

2.3 Example 1: Observational design with repeated measurements

To illustrate the analysis of longitudinal, observational data, we use two publicly available proteomics data sets (Erez et al., 2017; Tarca et al., 2019) to approach the following research questions:

1. How does the plasma proteome develop throughout normal pregnancy?
2. How does smoking affect the plasma proteome development throughout normal pregnancy, when accounting for body mass index (BMI)?
3. Does the plasma proteome of pregnancies that are later complicated by early- or late-onset preeclampsia follow distinct developmental trajectories?

2.3.1 Materials

The two data sets contain repeated measurements of 1,125 plasma proteins from pregnant women, and share the same control group (n = 90 women). The first study, by Tarca et al. (2019), focused on early-onset preeclampsia (n = 33 women), whereas the second study, by Erez et al. (2017), investigated late-onset preeclampsia (n = 76 women). BMI, smoking status, age, and parity were available for controls and early-onset preeclampsia cases only.

For the two first analyses, we selected control cases to visualize the normal plasma proteome development throughout pregnancy. To utilize as many serum samples as possible, the control samples were divided into five time intervals: first trimester ($\leq 13^{+6}$ weeks, n = 76), early second trimester (14^{+0}–21^{+6} weeks, n = 87), late second trimester (22^{+0}–27^{+6} weeks, n = 43), early third trimester (28^{+0}–33^{+6} weeks, n = 40), and late third trimester ($\geq 34^{+0}$ weeks, n = 32). Only the first sample from each participant at each time interval was included.

For the second analysis, the data from the previous example are reused as BMI and smoking status were available for the all healthy women. Smoking was coded as a factor in the `group` column with non-smokers acting as reference. Pre-pregnancy BMI was included as a continuous covariate as BMI is a potential confounder in the analysis.

For the third analysis the data sets from Erez et al. (2017) and Tarca et al. (2019) were merged to assess whether the plasma proteome of EO- and LO-preeclamptic pregnancies developed along distinct trajectories. The two data sets shared the same control group. Since women who developed EO-PE did not deliver plasma samples in late pregnancy, we restricted the analysis to samples collected before week 32^{+0}. The remaining plasma samples were divided by gestational age into four time intervals: before week 14^{+0} ($\leq 13^{+6}$ weeks), week 14–21 (14^{+0}–20^{+6}), week 21–28 (21^{+0}–27^{+6}), and week 28–32 (28^{+0}–31^{+6}).

2.4 Example 2: Randomized intervention with repeated measurements

To demonstrate how data from randomized intervention studies with repeated measurements can be analyzed with RM-ASCA$^+$, we investigated a publicly available metabolomics data set from Euceda et al. (2017). In this data set, we aimed to assess the following research questions:

1. How is the metabolomic response in breast cancer affected by adding the drug bevacizumab to standard neoadjuvant chemotherapy?
2. How does the metabolomic response in breast cancer differ between responders and non-responders receiving neoadjuvant cheomtehrapy with or without bevacizumab?

Whereas Example 1 focused on the interpretation of models, this example will review scaling and validation strategies.

2.4.1 Materials

The publicly available metabolomics data set from Euceda et al. (2017) contains measurements of 16 metabolites from 270 tumor biopsies from 122 patients randomized to either bevacizumab + chemotherapy (n = 60) or chemotherapy alone (n = 62). Biopsies were taken before treatment (T_1), at 12 weeks into treatement (T_2), and at tumor removal at 24 weeks (T_3) and profiled with high resolution magic angle spinning MR spectroscopy (HR MAS MR). In total, 46 participants provided three biopsies, 21 in the chemotherapy group and 25 in the bevacizumab group. By time point, 105 (50% later received bevacizumab), 78 (47% receiving bevacizumab), and 87 (55% receiving bevacizumab) biopsies were available at T_1, T_2, and T_3, respectively. Madssen et al. (2021) used this data set in the

original description of RM-ASCA$^+$ and their results are reproduced and further explored here using the ALASCA package.

For the second analysis, participants were classified as responders (n = 44) or non-responders (n = 78) on basis of tumor size at surgery (T_3). In the chemotherapy group, there were 20 responders and 42 non-responders, and the corresponding numbers for the bevacizumab group were 24 and 36, respectively.

2.5 Example 3: Megavariate data

This example introduces dimension reduction which makes analysis of megavariate data sets manageable. A publicly available transcriptomics data set by Skaug et al. (2021) was analyzed to answer the following research questions:

1. Does skin gene expression differ between patients with systemic sclerosis (SSc) and healthy controls?
2. Does longitudinal skin gene expression differ between patients with limited and diffuse SSc?

2.5.1 Materials

Skaug et al. (2021) collected forearm skin biopsies from 113 unique patients with limited (n = 43) or diffuse (n = 70) SSc and 44 matched healthy controls. Two additional biopsies were subsequently collected from a subset of the patients. A fourth biopsy was excluded due to the low sample size (n = 1). Gene expression was measured by RNA sequencing and microarrays. Variables with more than 10% missing values were excluded (1,065 genes), and the remaining missing values were replaced by half of the lowest measured value for the corresponding variable. To avoid duplicated gene names, "(d)" was added to the gene name when multiple probes assessed the same genes. In sum, 26,910 genes were kept for analysis.

3 Results and discussion

3.1 Example 1: Observational design with repeated measurements

3.1.1 How does the plasma proteome develop throughout normal pregnancy?

Longitudinal plasma samples were collected from 90 healthy pregnancies and analyzed for 1,125 proteins. A possible model to assess normal proteome development throughout pregnancy involves a main effect for time (k) and a random intercept for each participant i. In R, this model can be specified as `value~time + (1|ID)`, where `value` is outcome, `time` the predictor, and `ID` the random intercepts. Since the first time point acts as baseline, protein levels were scaled by the standard deviation of the baseline samples by setting `scale_function = "sdt1"`. The RM-ASCA$^+$ model was then initialized as

```
mod <- ALASCA(
df = df,
formula = value ~ time + (1|ID),
scale_function = "sdt1",
validate = TRUE
)
```

The corresponding design matrix is shown in Supplementary Table S1.

RM-ASCA$^+$ extracted two general patterns of change as represented by the first (PC1) and second (PC2) principal component, explaining 87% and 9%, respectively, of the variability in the data set (Supplementary Figure S2). Each component is associated with positive and negative loadings describing how each plasma protein is related to the corresponding PC. Proteins with positive loadings have higher concentration in time points with higher score values, and *vice versa* for proteins with negative loading values.

The first component represents a monotone increase (for positive loadings) or decrease (for negative loadings) in plasma level throughout pregnancy (Figure 2). The largest change takes place in the first and second trimester before stabilizing in the third trimester, as can be validated by assessing the underlying regression models (Figure 3). Bone morphogenetic protein 1 (BMP-1), epithelial discoidin domain-containing receptor 1 (EDDR1), and placenta growth factor (PlGF) showed the strongest positive loading on the first component, and therefore increase the most during the first trimesters. The increase of BMP1, EDDR1, and PlGF levels in plasma is clearly visible from the raw data itself (Supplementary Figure S3). In the opposite end, dual specificity mitogen-activated protein kinase kinase 4 (MAP2K4), histidine-rich glycoprotein (HRG), and endothelin-converting enzyme 1 (ECE1) showed the strongest negative loadings on the first component (Figure 2). This pattern is also evident from inspection of raw data (Supplementary Figure S3).

The second component represents a non-linear development with either peak (for positive loadings) or dip (for negative loadings) in the second trimester (Supplementary Figure S4). The first pattern is seen for proteins such as vascular endothelial growth factor A (VEGF-A), C1q and PAPPA-A. C1q did, however, show significant variability and had a CI for the loading that included zero. In contrast, the concentration of sialic acid-binding Ig-like lectin (siglec-) 6, Activin A, and IL-1 R4 showed a u-shaped dipping in the second trimester. These patterns are visible in the raw data as well (Supplementary Figure S5). Some variables had high loadings on both PC1 and PC2. Their trajectory is a combination of the two, as can be seen as flattening of the curve PlGF in the third trimester (Figure 3 and Supplementary Figure S3).

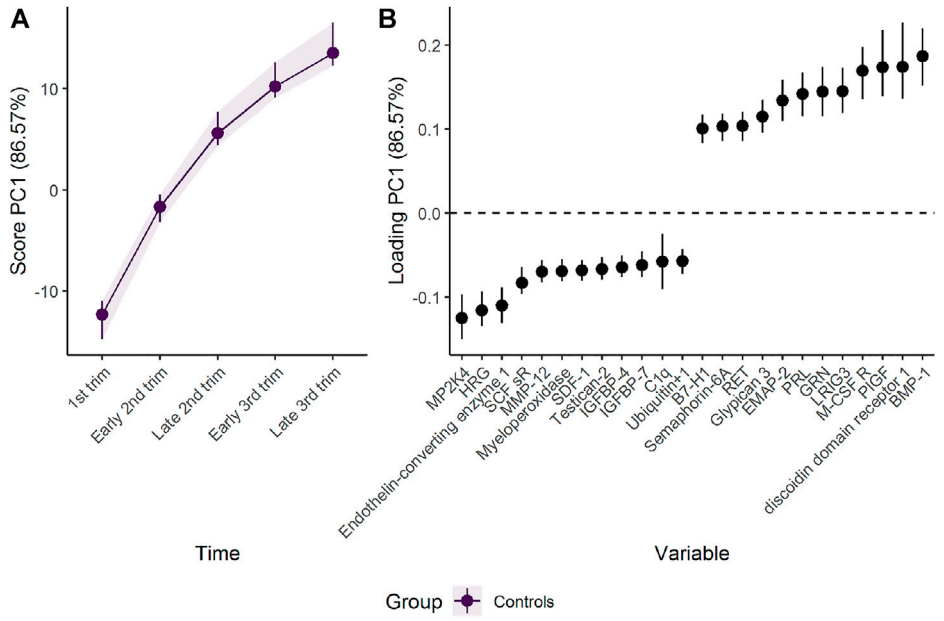


FIGURE 2
Time development of the plasma proteome throughout pregnancy as **(A)** scores and **(B)** loadings. The plasma level of proteins with high loading is increasing when the scores increase and *vice versa*. Only the 12 proteins with highest and lowest loadings, separated by the vertical dotted line, are shown due to the large number of assessed proteins.

3.1.2 How does smoking affect the plasma proteome development throughout normal pregnancy, when accounting for BMI?

The impact of smoking and pre-pregnancy BMI on plasma proteome development was examined in the same group of women as the analysis above (Section 3.1.1). Of the 90 pregnant women, 18 (20%) were smoking. Samples were collected from 76 (17% smoking), 87 (20% smoking), 43 (16% smoking), 40 (18% smoking), and 32 (19% smoking) women in the first trimester, early and late second trimester, and early and late third trimester, respectively. The BMI was 29 ± 7.8 and 28.1 ± 6.8 kg m^{-2} in the smoking and non-smoking group, respectively, and 28.3 ± 7.0 kg m^{-2} overall. The influence of BMI on the protein profile was assumed to be constant during pregnancy and thus there was no interaction with time in the regression model. In contrast, the effect of smoking was allowed to vary with time.

The regression formula was expanded to include a group term and time-group interaction: `time*group` is shorthand for `time + group + time:group`, where the two first terms represent the main effects of time and group, respectively, and the latter their interaction. Similarly, BMI was added as a covariate and the corresponding column kept as numerical values. The time and group effect matrices from Eq. 3 can be analyzed either separately or combined, so the model was ran twice, with `separate_effects = TRUE`, i.e., PCA is applied separately to $\mathbf{M}_T$ and $\mathbf{M}_{G+T:G}$, specified in the second run. The RM-ASCA$^+$ models were initialized as

```
mod <- ALASCA(
df = df,
formula = value ~ time*group + BMI + (1|ID),
scale_function = "sdt1",
validate = TRUE
)
```

and

```
mod <- ALASCA(
df = df,
formula = value ~ time*group + BMI + (1|ID),
separate_effects = TRUE,
scale_function = "sdt1",
validate = TRUE
)
```

The corresponding design matrix is shown in Supplementary Table S2.

RM-ASCA$^+$ offers two approaches to compare the time development of distinct groups of individuals. When the time and group effects are analyzed as a combined unit, i.e., the effect matrices for time, group, and time-group interaction in Eq. 3 are subjected to the same PCA, the resulting components will describe the common development of the groups. When the time and group effects are analyzed as separate units, i.e., the effect matrix for time is separated from the effect matrices for group and

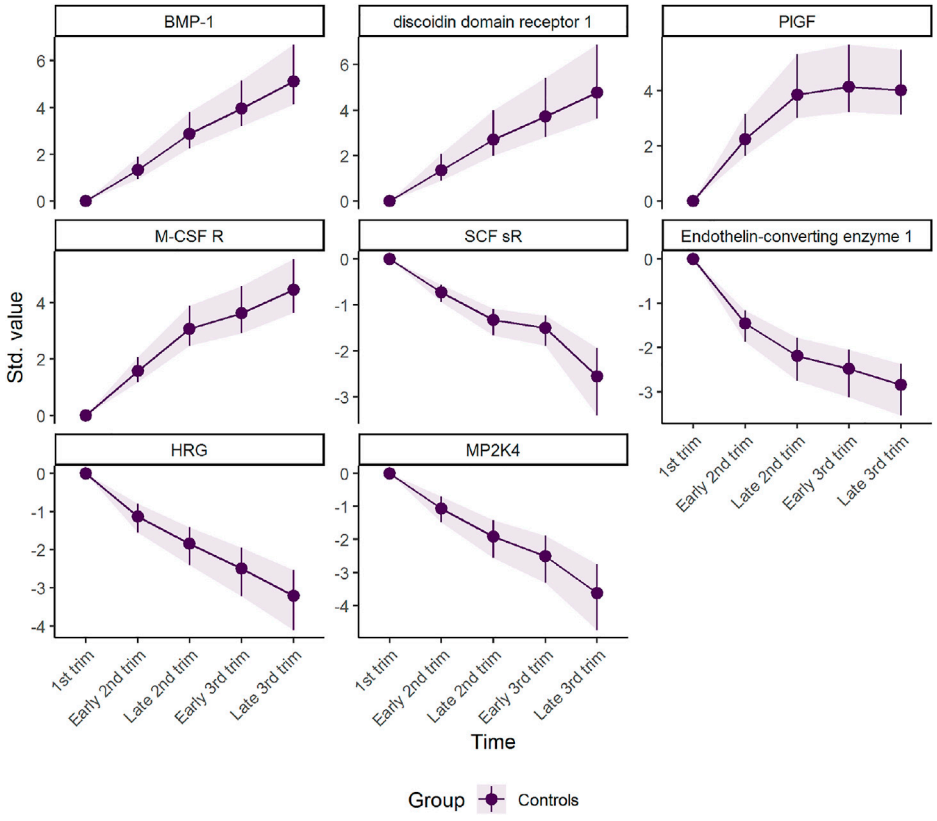


FIGURE 3
Marginal means for scaled protein concentration from linear mixed models. The intercept has been removed to highlight the robustness of development over time. The plot was made with the `plot(..., type = "prediction")` function.

time-group interaction in Eq. 3 and analyzed separately by PCA, two sets of scores and loadings are extracted. The first set of scores and loadings describes the development of the reference group, whereas the second set describes how the other groups diverge from the reference group. The underlying regression models, as well as the resulting regression coefficients, are, however, the same for the two approaches as the matrices **X** and **B** in Eq. 3 remain unchanged.

Analysis of the combined effect of time and group shows that smoking and non-smoking women demonstrate similar development in plasma proteome in pregnancy, with a tendency to lower scores for the smoking group (Figure 4). The parallel lines in Figure 4 suggest that the differences between the groups are stable over time, with somewhat lower levels of proteins such as BMP-1 and higher levels of proteins such as MP2K4 in smoking women. However, the confidence intervals are overlapping, suggesting that the effect of time is stronger than the effect of smoking, and no group specific development is evident.

Separating the effect of time and group changes the focus from common trajectories to divergent trajectories. The isolated time development of the non-smoking group, acting as reference, is similar to the time development of the combined group shown in Figure 2. The isolated group and time-group effect demonstrates how the plasma proteome of smoking women diverge from non-smoking women during pregnancy (Figure 5). The first component shows a stable and reliable difference between the two groups, with higher scores for the smoking women. Higher scores corresponds to higher plasma levels of proteins with positive loadings and *vice versa*. Thus, smoking women showed higher levels of proteins such as casein kinase II 2-alpha':2-beta heterotetamer (CK2-A2:B) and roundabout homolog 3 (ROBO3), and lower levels of proteins such as apolipoprotein A-I (Apo A-I) and siglec-9. Apolipoprotein A-I is an important constituent of high-density cholesterol, and is known to be decreased by smoking (Richard et al., 1997; Meenakshisundaram et al., 2010; Slagter et al., 2013).

The ability to adjust for covariates is one of the main advantages of (RM-)ASCA$^+$ when compared to other

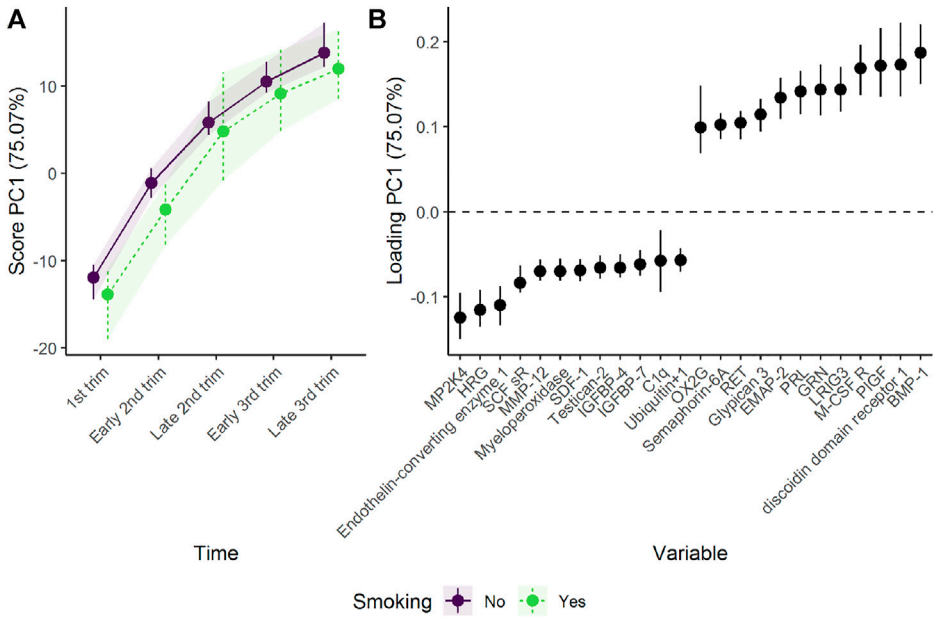


FIGURE 4
Time development of the plasma proteome throughout pregnancy in smoking and non-smoking women as **(A)** scores and **(B)** loadings. The plasma level of proteins with high loading is increasing when the scores increase and *vice versa*. Only the 12 proteins with highest and lowest loadings, separated by the vertical dotted line, are shown due to the large number of assessed proteins.

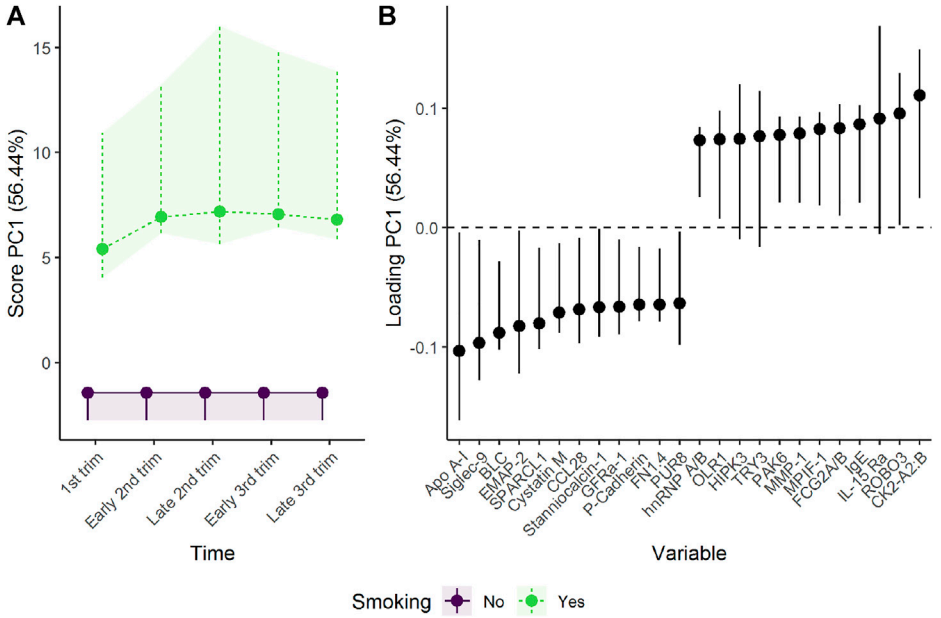


FIGURE 5
Time development of the plasma proteome throughout pregnancy in smoking and non-smoking women as **(A)** scores and **(B)** loadings. The time development of the non-smoking women has been removed to highlight the effect of smoking. The plasma level of proteins with high loading is increasing when the scores increase and *vice versa*. Only the 12 proteins with highest and lowest loadings, separated by the vertical dotted line, are shown due to the large number of assessed proteins.

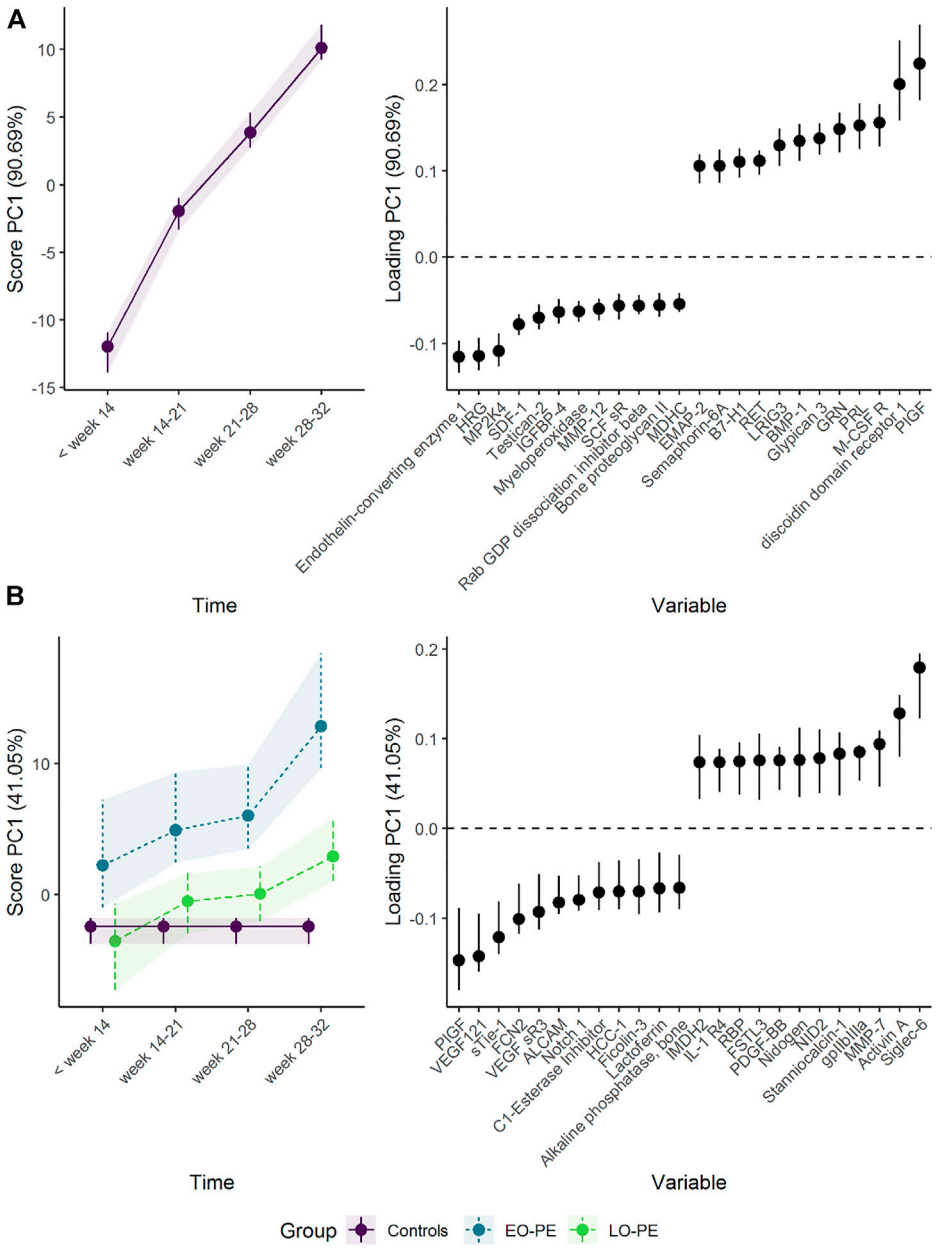


FIGURE 6
Time development of the plasma proteome throughout pregnancy in **(A)** healthy women and **(B)** women developing early-onset (EO-) or late-onset (LO-) preeclampsia (PE). The time development of healthy women is isolated in the upper panels, whereas the lower panels visualize how the plasma proteome differs between the groups. The plasma level of proteins with high loading is increasing when the scores increase and *vice versa*. Only the 12 proteins with highest and lowest loadings, separated by the vertical dotted line, are shown due to the large number of assessed proteins.

multivariate methods such as PLS. Continuous covariate adjustment was first introduced with ASCA$^+$ and with RM-ASCA$^+$ this ability has been extended to longitudinal data. For longitudinal trials, adjusting for covariates can offer both more precise and less biased effect estimates, and increase statistical power. Although covariate adjustment can be achieved for methods such as PLS by including it as part of data preprocessing, the ASCA$^+$ framework leverages the users' existing intuitions and knowledge of both linear regression and PCA together in a cohesive approach. With RM-ASCA$^+$ the effect of BMI can be isolated by including BMI as a covariate in the regression model, but not in the effect matrices subjected

to PCA. The effect of BMI is thus presented as ordinary β coefficients (Supplementary Figure S6). The β coefficients are the same regardless of whether the time and group effects are assessed separately or not, and represent the adjustment for BMI. High BMI was associated with higher plasma levels of leptin, and the complement components C1s and C5a. In contrast, lower levels of kallistatin, soluble receptor for advanced glycation end products (sRAGE) and neural cell adhesion Molecule (Nr-CAM) were observed with increasing BMI. Obesity is related to low-grade inflammation with lower levels of both the anti-inflammatory kallistatin (Zhu et al., 2013; Frühbeck et al., 2018) and the cardioprotective sRAGE (Norata et al., 2009), and leptin is strongly linked to obesity and correlate with body fat percentage (Obradovic et al., 2021). In addition, the strong effect of BMI on leptin, IGFBP2, and SHBG is in line with previous research on plasma proteomics (Goudswaard et al., 2021).

3.1.3 Does the plasma proteome of pregnancies that are later complicated by early- (EO-) or late-onset (LO-) preeclampsia (PE) follow distinct developmental trajectories?

To assess the developmental trajectories of preeclamptic women, the full data sets of Erez et al. (2017) and Tarca et al. (2019) were used. In total, 572 plasma samples were included for analysis. Of 199 participants, 33 (17%) developed early-onset preeclampsia (EO-PE) and 76 (38%) developed late-onset preeclampsia (LO-PE). For the different time points, 151 (12% EO-PE and 27% LO-PE), 157 (16% EO-PE and 39% LO-PE), 135 (20% EO-PE and 54% LO-PE), and 129 (13% EO-PE and 56% LO-PE) samples were analyzed. The disease groups were coded in the `group` column with the controls acting as reference and the previous regression formula was similar to the previous example (Section 3.1.1) except that the BMI term was removed. To isolate the potentially distinct trajectories of the preeclamptic pregnancies, the time and group effect matrices were separated by setting `separate_effects = TRUE`. The RM-ASCA$^+$ model was thus initialized as

```
mod <- ALASCA(
df = df,
formula = value ~ time*group + (1|ID),
separate_effects = TRUE,
scale_function = "sdt1",
validate = TRUE
)
```

The corresponding design matrix is shown in Supplementary Table S3.

Women developing EO-PE showed lower plasma levels of proteins such as PlGF, VEGF-121, and soluble tyrosine-protein kinase receptor Tie-1 (sTie-1), and higher plasma levels of proteins such as Siglec-6, activin A, and matrilysin/MMP-7 (Figure 6 and Supplementary Figure S7). These findings support the original results by Tarca et al. (2019) (Supplementary Figure S8). The differences from the control group were present from early pregnancy for some proteins, and increased steadily as the pregnancy progressed. The development of the reference group is similar as in sections 3.1.1, 3.1.2 except minor changes of scores and loadings due to redefined time points.

Interestingly, women developing LO-PE showed a similar but delayed shift in plasma proteome (Figure 6). It is, however, necessary to also investigate PC2, as PC1 explained only 41% of the group variation. PC2 demonstrates a clear difference between women developing LO-PE, and the remaining women (Supplementary Figure S9). Women developing LO-PE seem to have higher levels of proteins such as MMP-7, RAN and PPID from early pregnancy, and lower levels of proteins such as HSP70, BMP10, and integrin aVb5 (Supplementary Figure S10). These findings are consistent with the original results by Erez et al. (2017). It is useful to visualize the marginal means from the underlying regression models when a protein has strong loading on multiple PCs and there are robust differences in score in the corresponding PCs. From Supplementary Figures S7, S10, it can be seen that women developing PE had clearly higher MMP-7 throughout pregnancy.

3.2 Example 2: Randomized intervention with repeated measurements

3.2.1 How is the metabolomic response in breast cancer affected by adding bevacizumab to standard neoadjuvant chemotherapy?

In contrast to the previous example with observational data, studies with randomized intervention assume that the groups are equal prior to intervention. Thus, the regression model should not include a main effect for treatment (Twisk et al., 2018). A regression model with a time effect, a time-group interaction, and a random intercept can in R be defined as `value~time + time:group + (1|ID)`. By default, however, the interaction term between time and group (`time:group`) will include the interaction between the first time point (i.e., baseline) and group, which has to be removed. This can be achieved by providing `equal_baseline = TRUE` to the `ALASCA()` function. Thus, the function call

```
mod <- ALASCA(
df = df,
formula = value ~ time + time:group + (1|ID),
equal_baseline = TRUE,
scale_function = "sdt1",
validate = TRUE
)
```

reproduce the findings of Madssen et al. (2021). The corresponding design matrix is shown in Supplementary Table S4.

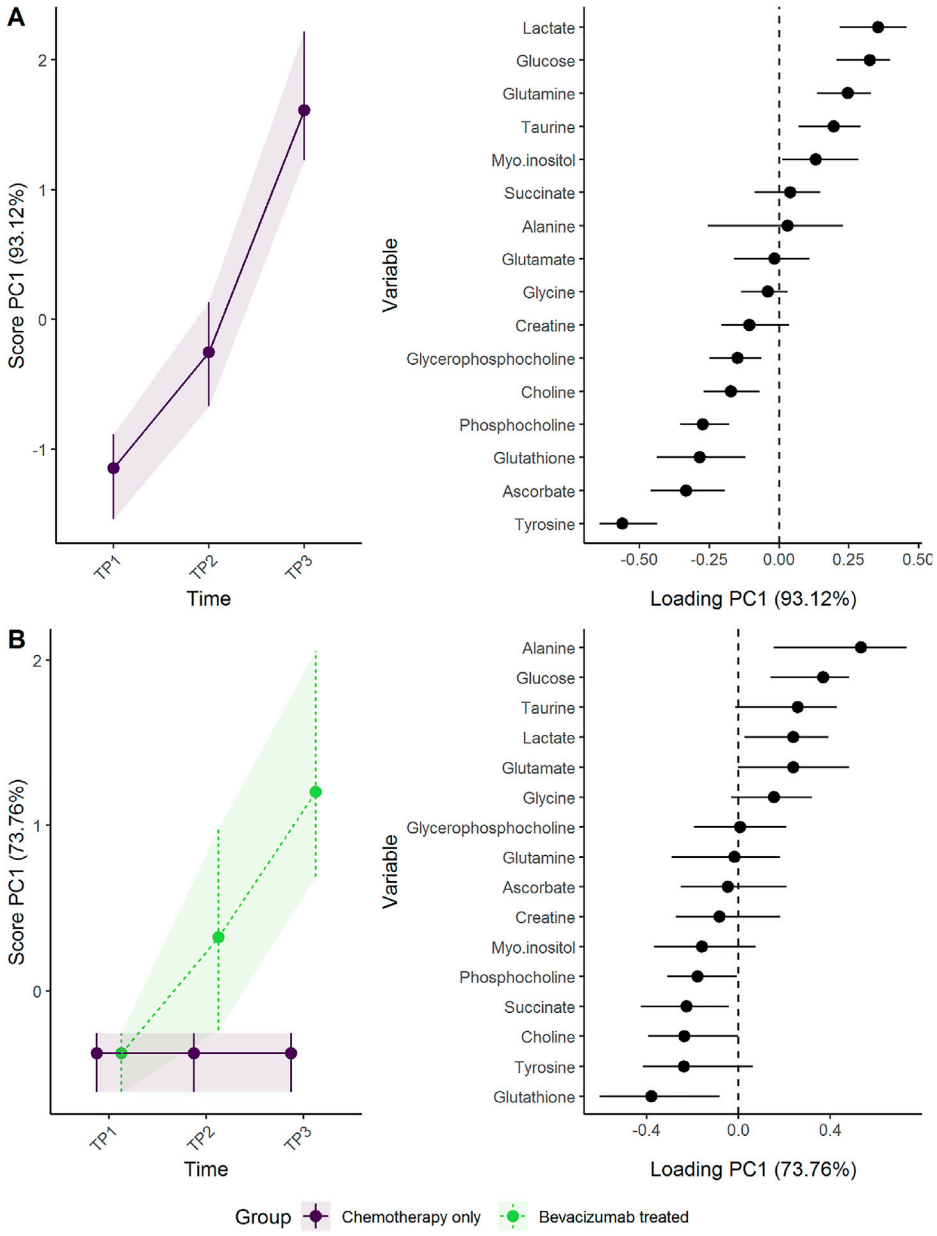


FIGURE 7
Time development of tumor biopsy metabolome before and during cancer treatment. **(A)** The time development of the participants receiving chemotherapy only is isolated in the upper panels, whereas **(B)** the lower panels visualize how the metabolome differs between the groups. The levels of metabolites with high loading is increasing when the scores increase and *vice versa*.

To illustrate how scaling and validation strategy impact the analysis, the model was generated for all 16 combinations of scaling (sdall, sdt1, sdref, and sdfref1), resampling (bootstrap and jack-knife), and extraction of effect matrices (combined and separate). The bootstrap and jack-knife samples were reused for each model to make the results comparable.

To assess the effect of adding the drug bevacizumab to standard neoadjuvant chemotherapy to treat breast cancer, the effect matrix for time and the effect matrix for time-group interaction were analyzed separately by PCA (Figure 7). The addition of bevacizumab led to higher concentrations of alanine, glucose, and lactate, and lower concentrations of gluthatione, succinate, and phosphocoline. The increased alanine and glucose levels, and decreased gluthatione levels, were statistically significant at T_3 following bevacizumab treatment in univariate models (Supplementary Figure S11) and the

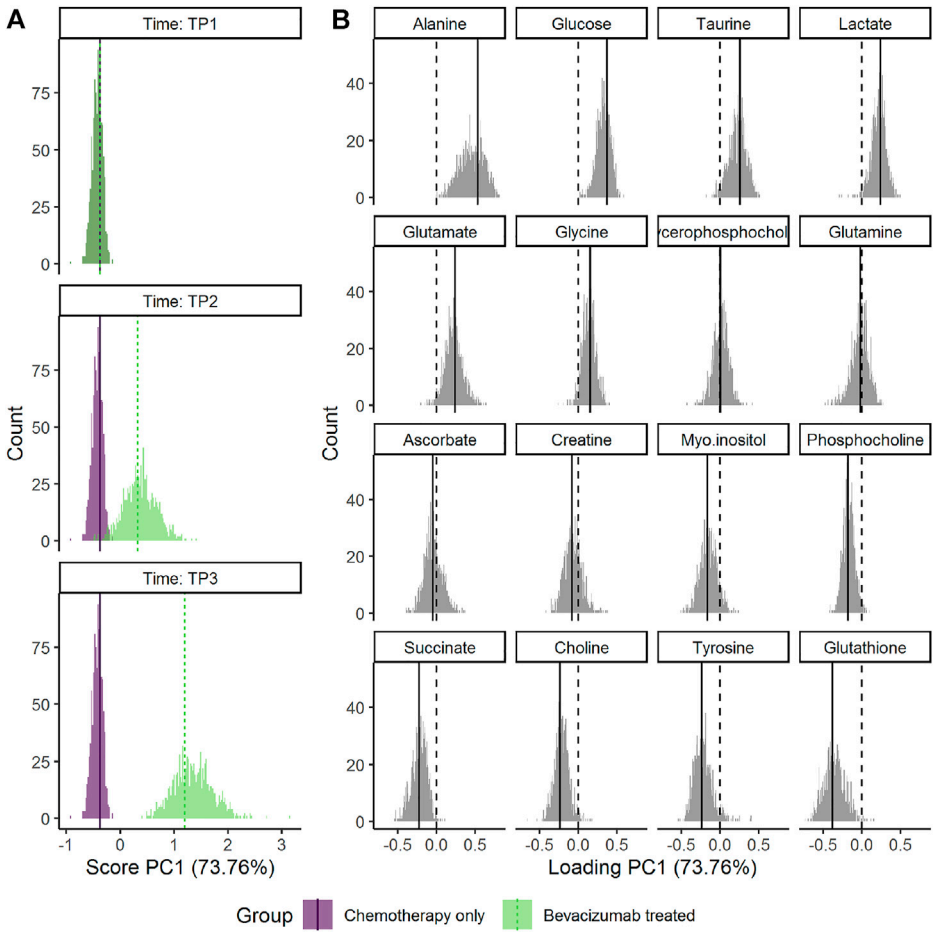


FIGURE 8
Distribution of the bootstrapped parameters **(A)** scores and **(B)** loadings for the RM-ASCA⁺ model shown in Figure 7. Main model estimates are shown as vertical line. The dotted lines mark zero. The plot was made with the `plot(..., type = "histogram")` function.

residuals showed acceptable normal distribution (Supplementary Figure S12). These results are discussed in detail by Madssen et al. (2021). ALASCA also allows the results to be displayed as a more classical ASCA analysis, by plotting the first and second PC against each other, as in Supplementary Figure S13.

The choice of scaling and validation strategy has strong impact on uncertainty estimates (Supplementary Figures S14–S16). Jack-knife resulted in markedly smaller CIs for both scores and loadings than bootstrap. The choice of scaling does not alter how the results are interpreted but using baseline samples for scaling (sdt1 or sdreft1) enhanced the separation of the groups at the third time point. ALASCA provides two additional visualizations of the validation results: either the scores and loadings for each individual iteration (Supplementary Figure S17) or the distribution of scores and loadings as histograms (Figure 8).

In general, Timmerman et al. (2015) advice that "scaling factors should be free from the effect of interest." The argument is that if the effect of interest actually increases between-group variation, then we have to avoid that this effect is damped by scaling. I.e., the between-group variation introduced by experimental manipulation should not be part of the scaling factor. In this specific example with a randomized trial, the baseline measurements constitute a subset of data where no such between-group variation has yet been introduced. In other cases, however, it may be less clear which groups that are affected by the experimental condition of interest. In addition, the scaling factor must be based on a sufficiently large group. In this paper, we are primarily using the baseline measurement for scaling to balance the need for a sample free from the effect of interest (typically the effect of time and time-group interaction) and sample size. In example 3, however, where a healthy and a diseased population are compared at a single time point and where the disease is manifest, the scaling factor is based on the healthy controls only.

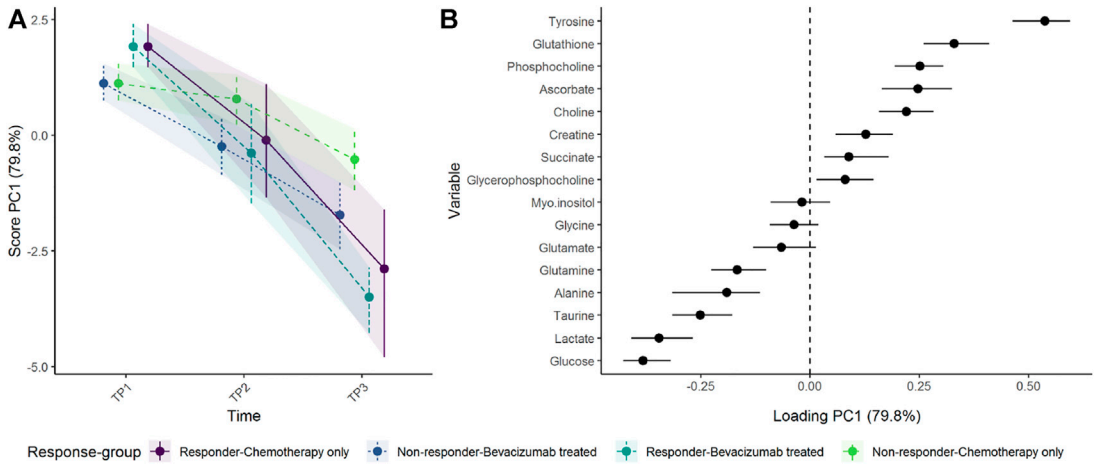


FIGURE 9
Time development of tumor biopsy metabolome before and during cancer treatment, in responders and non-responders as **(A)** scores and **(B)** loadings. Since participants were randomized to treatment but not response, the baseline levels are equal for all participants within each response groups. Non-responders receiving chemotherapy only show the smallest change of metabolome. The level of metabolites with high loading is increasing when the scores increase and *vice versa*.

Bootstrapping seems the preferable resampling strategy despite jack-knifing resulting in smaller CIs and clearer separation between groups. Targeted studies are needed to assess the performance and coverage of specific validation strategies for (RM-)ASCA$^+$, and the most conservative approach seems reasonable until such studies emerge. A possible explanation for the smaller CIs from jack-knife may be that bootstrapping "'shakes' the original data more violently than jackknifing" (Efron and Hastie, 2016, p. 161); on average, bootstrapping leaves out approximately 37% of the participants compared to 14% for jack-knife when 1/7 participants are excluded. Many refined strategies exist for resampling and CI calculation for multilevel models and may be implemented in later versions of ALASCA when the strengths and weaknesses have been thoroughly mapped (van der Leeden et al., 2008). Similarly, permutation tests exist in exact or approximate form for general ASCA models and provide means to calculate p values for model terms and interactions (Anderson and Braak, 2003; Bertinetto et al., 2020), and may be implemented in ALASCA when their performance under various model design have been thoroughly explored.

3.2.2 How does the metabolomic response in breast cancer differ between responders and non-responders receiving neoadjuvant chemotherapy with or without bevacizumab?

To investigate whether the metabolomic changes in tumors from patients having a good response to either chemotherapy alone or chemotherapy+bevacizumab differed from non-responders, a main effect for response and a three-way interaction between time, group, and response was added. In R, the model can be specified as value ~ time + response + time:response + time:group + time:group:response + (1|ID). Since `equal_baseline = TRUE`, the *treatment* groups are similar at baseline, whereas the *response* groups can differ. In this case, the effect matrix is specified manually. If not, the response effect would be separated as for BMI in example 1. The `ALASCA()` call was:

```
mod <- ALASCA(
df = df,
formula = value ~ time + response +
time:response + time:group +
time:group:response + (1|ID),
equal_baseline = TRUE,
effects = "time + response + time:response +
time:group + time:group:response",
scale_function = "sdt1",
validate = TRUE
)
```

The corresponding design matrix is shown in Supplementary Table S5.

The regression model including a three-way-interaction between time, response, and treatment showed that responders had somewhat higher concentrations of tyrosine and glutathione, and lower concentrations of glucose and lactate at baseline and showed a larger shift in metabolomic profile than non-responders (Figure 9). After 12 weeks of treatment (T_2), the metabolomic shift seems similar in the reponder group as well as non-responders receiving bevacizumab. At 24 weeks, however, the responders had the largest change in metabolic profile, followed by non-responders receiving bevacizumab, whereas non-responders receiving chemotherapy only had the smallest change.

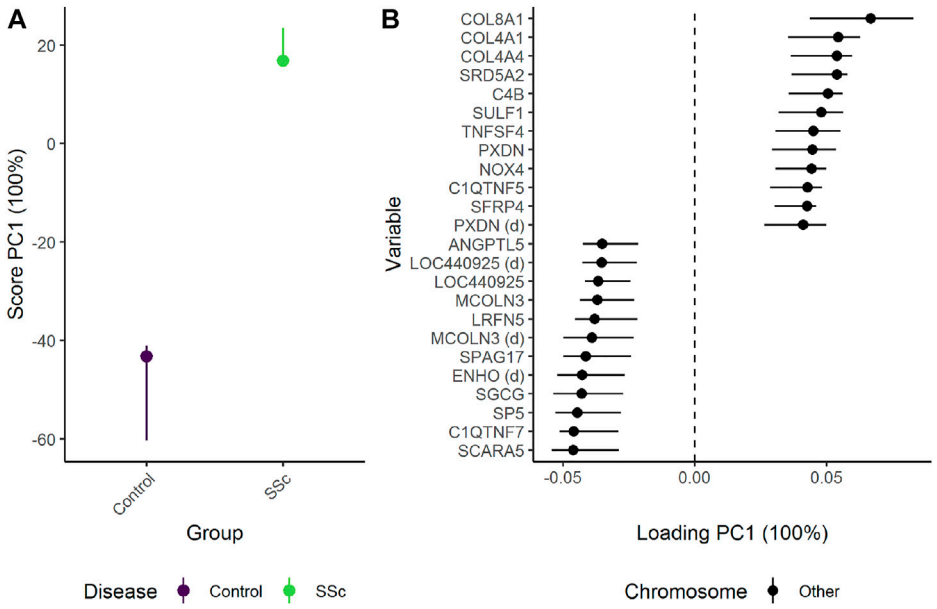


FIGURE 10
The difference in skin biopsy transciptome between healthy controls and patients with systemic sclerosis (SSc) as **(A)** scores and **(B)** loadings. The effects of age and gender have been adjusted for (Figure 11). The expression of genes with high loading is increasing when the scores increase and *vice versa*. Only the 12 genes with highest and lowest loadings, separated by the vertical dotted line, are shown due to the large number of assessed genes.

One should note that the baseline levels shown in Figure 9 reflect a more complex statistical model than the previous example, where the treatment groups shared the same baseline. Since the tumors from responders and non-responders may have had some distinct properties from the beginning, the baseline levels of responders and non-responders are allowed to vary, whereas the baseline levels of the treatment groups are kept equal. Thus, the three-way interaction between time, treatment, and response could not have been reproduced by simply creating four groups (treatment×response) and using the same regression model as above (`value~time + time:group + (1|ID)`).

3.3 Example 3: Megavariate data

3.3.1 Does skin gene expression differ between patients with systemic sclerosis (SSc) and healthy controls?

Since control samples were only available for a single time point, skin gene expression in healthy controls were compared to patients with limited or diffuse SSc at baseline. Reduction of dimensions by PCA was applied due to the size of the data set.

Although the ALASCA package is primarily designed for longitudinal data sets, it also supports ordinary linear models without random effects. When there is no time term in the regression formula, the first term will be used as abscissa. Gender and age were included as covariates to demonstrate adjustment of categorical and continuous variables. In R, the regression model can be defined as `value~disease + gender + age`:

```
mod <- ALASCA(
df = df,
formula = value ~ disease + gender + age,
scale_function = "sdref",
reduce_dimensions = TRUE,
validate = TRUE
)
```

The corresponding design matrix is shown in Supplementary Table S6.

ALASCA can be used to compare multivariate data from experimental designs with single measurements and adjust for confounders such as gender. When only two groups are compared, the difference between the groups is fully explained by PC1 (Figure 10). Patients with SSc showed stronger expression of several genes related to collagen alpha proteins such as COL8A1, COL4A1, and COL4A4. In contrast, the healthy controls showed stronger expression of genes such as SCARA5 (Scavenger Receptor Class A Member 5), C1QTNF7 (Complement C1q Tumor Necrosis Factor-Related Protein 7), SP5 (Transcription Factor Sp5), SGCG (sarcoglycan gamma), and ENHO (Energy Homeostasis-Associated Protein). The genes with highest and lowest loading showed some overlap with the

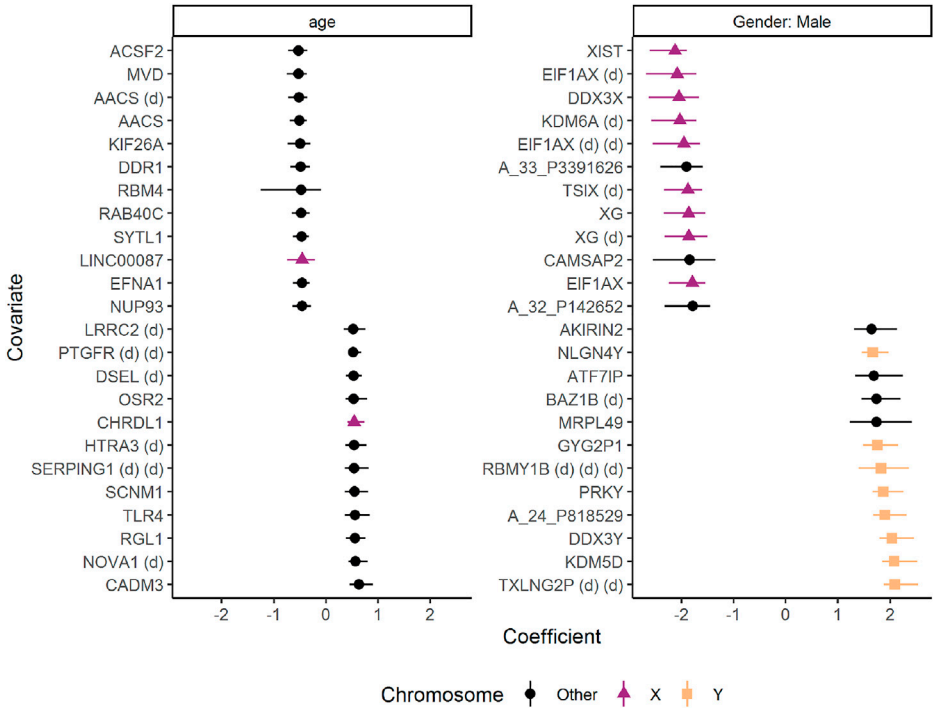


FIGURE 11
The effects of age and gender on gene expression in skin biopsies from healthy controls and patients with systemic sclerosis. The coefficients are regression coefficients from linear regression models, colored by chromosome location. Some genes were associated with mulitple probes, and are marked with "(d)" to avoid duplicated names. The error bars reflect 95% confidence intervals from bootstrapping. Only the 12 genes with highest and lowest coefficients are shown. The figure was made with the `plot(..., type = "covars")` function.

genes with the highest/lowest fold-change as reported in the original study, but ALASCA also identified several new genes of possible interest (Supplementary Figure S18). In addition, the original study did not adjust for gender and age.

Many of the genes differently expressed in males and females were located on the sex chromosomes (Figure 11). Male participants had stronger expression of genes such as TXLNG2P (Taxilin Gamma Pseudogene, Y-Linked), Lysine Demethylase 5D (KDM5D), and DDX3Y (DEAD-Box Helicase 3, Y-Linked). Females, on the other hand, showed stronger expression of genes such as XIST (X Inactive Specific Transcript), EIF1AX (Eukaryotic Translation Initiation Factor 1A, X-Linked), and DDX3X (DEAD-Box Helicase 3, X-Linked). Increasing age was associated with stronger expression of genes such as CADM3 (Cell Adhesion Molecule 3) and NOVA1 (NOVA Alternative Splicing Regulator 1), whereas genes such as ACSF2 (Acyl-CoA Synthetase Family Member 2) and MVD (Mevalonate Diphosphate Decarboxylase) showed the opposite pattern.

The default settings in the ALASCA package are suggestions and should not be treated as authoritative recommendations. The user's choice of parameters and settings should be informed by the research question and the data. For example, by reducing the number of variables through PCA as in this example, one improves efficiency at the cost of accuracy. Currently, there are many opinions on how to select the number of necessary components (Abdi and Williams, 2010), and the performance of various methods depends on the nature of the data being studied (Peres-Neto et al., 2005). The number of components selected by the ALASCA package during dimension reduction depends on how much variance wish to retain (by default, `reduce_dimensions.limit = 0.95` so that 95% of the variance will be kept). A good strategy would be to compare the results from multiple models with various limits to see how sensitive the results are to that specific parameter. A similar strategy can be employed to gain confidence in other parameters as well.

3.3.2 Does longitudinal skin gene expression differ between patients with limited and diffuse SSc?

The longitudinal skin gene expression from patients with limited or diffuse SSc was assessed with the limited variant as reference group. To reduce the number of variables subjected to regression by applying an initial PCA prior to regression, `reduce_dimensions` was set to `TRUE`. As the default

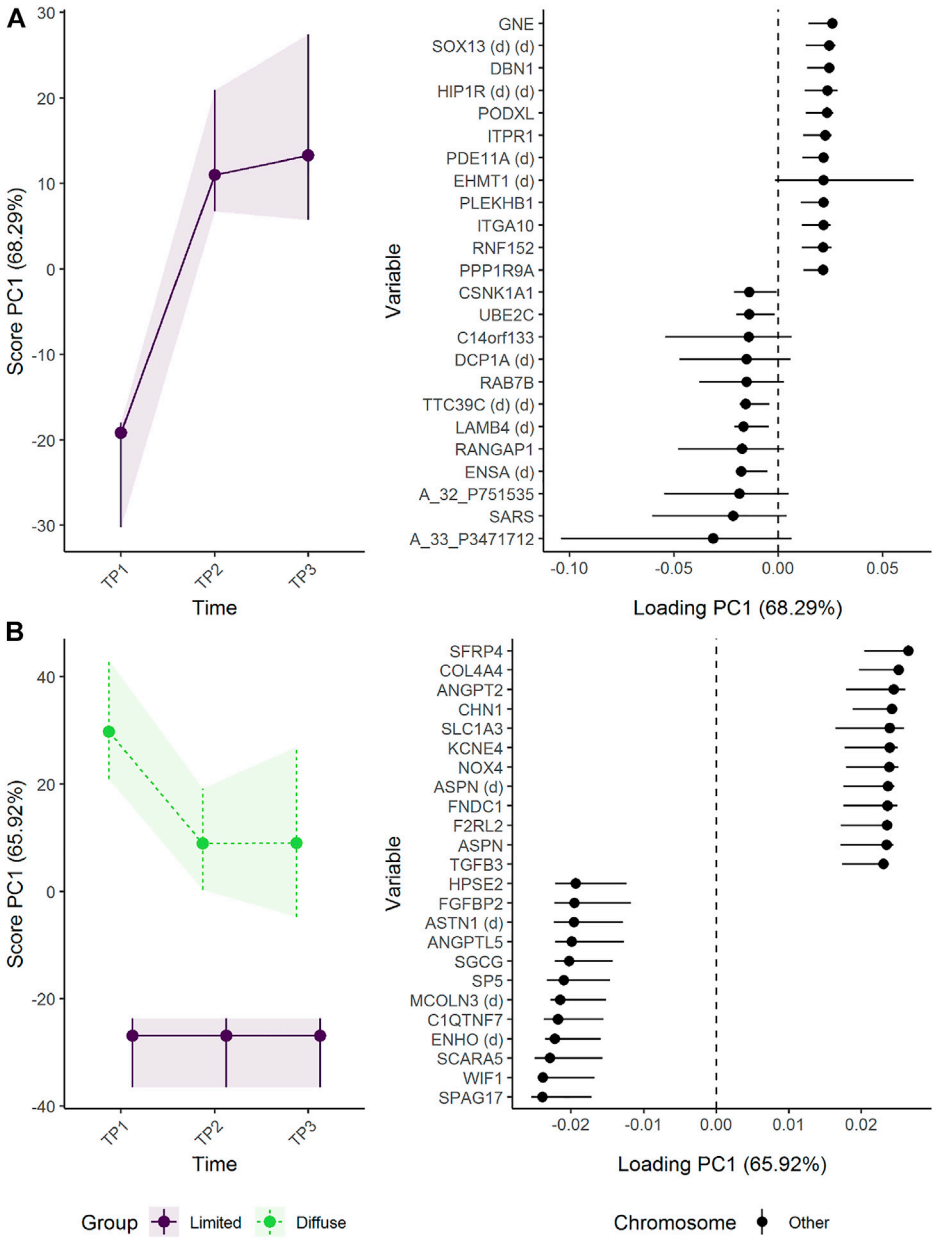


FIGURE 12
Time development of skin biopsy genome in patients with limited or diffuse systemic sclerosis. **(A)** The time development of patients with limited systemic sclerosis is isolated in the upper panels, whereas **(B)** the lower panels visualize how the skin biopsy genome develop distinctly between the groups. The expression of genes with high loadings is increasing when the scores increase and *vice versa*. Only the 12 genes with highest and lowest loadings, separated by the vertical dotted line, are shown due to the large number of assessed genes.

PCA algorithm in R sometimes stops due to internal errors, an alternative PCA function can be provided by specifying `pca_function` (Baglama et al., 2021). The regression model is similar to the final model in Example 1 with separated effects for time and group:

```
mod <- ALASCA(
df = df,
value ~ time * group + (1|ID),
scale_function = "sdt1",
pca_function = "irlba",
reduce_dimensions = TRUE,
separate_effects = TRUE,
validate = TRUE
)
```

The corresponding design matrix is shown in Supplementary Table S7.

The initial skin biopsy from patients with limited SSc differed from the two subsequent biopsies with a tendency to increased expression of genes such as GNE (Bifunctional UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase), SOX13 (SRY-Box Transcription Factor 13), and DBN1 (drebin 1) with time (Figure 12A). The difference in gene expression between the patient groups was stable over time (Figure 12B). Patients with diffuse SSc showed stronger expression of genes such as SFRP4 (Secreted Frizzled Related Protein 4), ANGPT2 (Angiopoietin 2), and COL4A4 (Collagen Type IV Alpha 4 Chain) than patients with limited SSc. In contrast, genes such as SPAG17 (Sperm Associated Antigen 17), SCARA5, and WIF1 (WNT Inhibitory Factor 1) were more strongly expressed in skin from patients with limited SSc than patients with diffuse SSc. Although SFRP4 was reported to have the highest fold-change between diffuse and limited SSc in the original publication (Skaug et al., 2021), ALASCA identifies several the genes of possible interest (Supplementary Figure S19).

4 Conclusion

The (RM-)ASCA$^+$ framework offers a flexible and robust method to quickly discover patterns in multivariate data. Advantages with (RM-)ASCA$^+$ compared to other methods such as PLS-DA include the possibility to model longitudinal changes from multiple timepoints, to incorporate advanced experimental designs, and to include confounders in the analysis. The ALASCA package for R makes the (RM-)ASCA$^+$ available for general use by offering a simple interface to model complex relationships, to scale the data, to perform model validation, and to produce a variety of publication-ready visualizations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material and at https://doi.org/10.6084/m9.figshare.21362979.v1. Further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

AJ, TM, and GG contributed to conception and design of the study. AJ wrote the software. TM validated the statistical results. AJ wrote the first draft of the manuscript. AJ, TM, and GG wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgments

We would like to thank Live M. T. Stokkeland and Mariell Ryssdal for testing the ALASCA package and providing constructive feedback during the development of the package.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.962431/full#supplementary-material

References

Abdi, H., and Williams, L. J. (2010). Principal component analysis: Principal component analysis. *WIREs. Comp. Stat.* 2, 433–459. doi:10.1002/wics.101

Anderson, M., and Braak, C. T. (2003). Permutation tests for multi-factorial analysis of variance. *J. Stat. Comput. Simul.* 73, 85–113. doi:10.1080/00949650215733

Baglama, J., Reichel, L., and Lewis, B. W. (2021). *Irlba: Fast truncated singular value decomposition and principal components analysis for large dense and sparse matrices*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi:10.18637/jss.v067.i01

Bertinetto, C., Engel, J., and Jansen, J. (2020). ANOVA simultaneous component analysis: A tutorial review. *Anal. Chim. Acta. X* 6, 100061. doi:10.1016/j.acax.2020.100061

Chang, W. (2021). *R6: Encapsulated classes with reference semantics*. R package version 2.5.1.

Dowle, M., and Srinivasan, A. (2021). *Data.Table: Extension of 'data.Frame'*.

Efron, B., and Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. New York, NY: Institute of Mathematical Statistics Monographs, Cambridge University Press.

Erez, O., Romero, R., Maymon, E., Chaemsaithong, P., Done, B., Pacora, P., et al. (2017). The prediction of late-onset preeclampsia: Results from a longitudinal proteomics study. *PLOS ONE* 12, e0181468. doi:10.1371/journal.pone.0181468

Euceda, L. R., Haukaas, T. H., Giskeødegård, G. F., Vettukattil, R., Engel, J., Silwal-Pandit, L., et al. (2017). Evaluation of metabolomic changes during neoadjuvant chemotherapy combined with bevacizumab in breast cancer using MR spectroscopy. *Metabolomics* 13, 37. doi:10.1007/s11306-017-1168-0

Fresno, C., Balzarini, M. G., and Fernández, E. A. (2014). Lmdme : Linear models on designed multivariate experiments in *R. J. Stat. Softw.* 56. doi:10.18637/jss.v056.i07

Frühbeck, G., Gómez-Ambrosi, J., Rodríguez, A., Ramírez, B., Valentí, V., Moncada, R., et al. (2018). Novel protective role of kallistatin in obesity by limiting adipose tissue low grade inflammation and oxidative stress. *Metabolism.* 87, 123–135. doi:10.1016/j.metabol.2018.04.004

Garnier, S., Ross, N., Rudis, B., Filipovic-Pierucci, A., Galili, T., Timelyportfolio, Greenwell, B., et al. (2021). *Viridis - colorblind-friendly color maps for r*. doi:10.5281/zenodo.4679424

Goudswaard, L. J., Bell, J. A., Hughes, D. A., Corbin, L. J., Walter, K., Davey Smith, G., et al. (2021). Effects of adiposity on the human plasma proteome: Observational and Mendelian randomisation estimates. *Int. J. Obes.* 45, 2221–2229. doi:10.1038/s41366-021-00896-1

Jarmund, A. H., Giskeødegård, G. F., Ryssdal, M., Steinkjer, B., Stokkeland, L. M. T., Madssen, T. S., et al. (2021). Cytokine patterns in maternal serum from first trimester to term and beyond. *Front. Immunol.* 12, 752660. doi:10.3389/fimmu.2021.752660

Kassambara, A. (2020). *Ggpubr: 'ggplot2' based publication ready plots*.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi:10.18637/jss.v082.i13

Liland, K. H. (2022). *multiblock: Multiblock data fusion in statistics and machine learning*. https://github.com/khliland/multiblock/.

Madssen, T. S., Giskeødegård, G. F., Smilde, A. K., and Westerhuis, J. A. (2021). Repeated measures ASCA+ for analysis of longitudinal intervention studies with multivariate outcome data. *PLoS Comput. Biol.* 17, e1009585. doi:10.1371/journal.pcbi.1009585

Martin, M., and Govaerts, B. (2020). LiMM-PCA: Combining ASCA+ and linear mixed models to analyse high-dimensional designed data. *J. Chemom.* 34, e3232. doi:10.1002/cem.3232

Meenakshisundaram, R., Rajendiran, C., and Thirumalaikolundusubramanian, P. (2010). Lipid and lipoprotein profiles among middle aged male smokers: A study from southern India. *Tob. Induc. Dis.* 8, 11. doi:10.1186/1617-9625-8-11

Mühleisen, H., and Raasveldt, M. (2022). *Duckdb: DBI package for the DuckDB database management system*. R package version 0.3.2-2.

Müller, K., Wickham, H., James, D. A., and Falcon, S. (2021). *RSQLite: SQLite interface for r*.

Norata, G. D., Garlaschelli, K., Grigore, L., Tibolla, G., Raselli, S., Redaelli, L., et al. (2009). Circulating soluble receptor for advanced glycation end products is inversely associated with body mass index and waist/hip ratio in the general population. *Nutr. Metab. Cardiovasc. Dis.* 19, 129–134. doi:10.1016/j.numecd.2008.03.004

Nueda, M. J., Conesa, A., Westerhuis, J. A., Hoefsloot, H. C. J., Smilde, A. K., Talón, M., et al. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA–SCA. *Bioinformatics* 23, 1792–1800. doi:10.1093/bioinformatics/btm251

Obradovic, M., Sudar-Milovanovic, E., Soskic, S., Essack, M., Arya, S., Stewart, A. J., et al. (2021). Leptin and obesity: Role and clinical implication. *Front. Endocrinol.* 12, 585887. doi:10.3389/fendo.2021.585887

Papadakis, M., Tsagris, M., Dimitriadis, M., Fafalios, S., Tsamardinos, I., Fasiolo, M., et al. (2021). *Rfast: A collection of efficient and extremely fast r functions*.

Peres-Neto, P. R., Jackson, D. A., and Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Statistics Data Analysis* 49, 974–997. doi:10.1016/j.csda.2004.06.015

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richard, F., Marécaux, N., Dallongeville, J., Devienne, M., Tiem, N., Fruchart, J. C., et al. (1997). Effect of smoking cessation on lipoprotein A-I and lipoprotein A-I:A-II levels. *Metabolism.* 46, 711–715. doi:10.1016/s0026-0495(97)90018-4

Skaug, B., Lyons, M. A., Swindell, W. R., Salazar, G. A., Wu, M., Tran, T. M., et al. (2021). Large-scale analysis of longitudinal skin gene expression in systemic sclerosis reveals relationships of immune cell and fibroblast activity with skin thickness and a trend towards normalisation over time. *Ann. Rheum. Dis.* 81, 516–523. doi:10.1136/annrheumdis-2021-221352

Slagter, S. N., van Vliet-Ostaptchouk, J. V., Vonk, J. M., Boezen, H. M., Dullaart, R. P., Kobold, A. C. M., et al. (2013). Associations between smoking, components of metabolic syndrome and lipoprotein particle size. *BMC Med.* 11, 195. doi:10.1186/1741-7015-11-195

Slowikowski, K. (2021). *Ggrepel: Automatically position non-overlapping text labels with 'Ggplot2'*.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R.-J. A. N., van der Greef, J., and Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* 21, 3043–3048. doi:10.1093/bioinformatics/bti476

Smilde, A. K., Næs, T., and Liland, K. H. (2022). *Multiblock data fusion in statistics and machine learning*. Chichester, West Sussex, UK: John Wiley & Sons.

Smilde, A. K., Timmerman, M. E., Hendriks, M. M., Jansen, J. J., and Hoefsloot, H. C. (2012). Generic framework for high-dimensional fixed-effects ANOVA. *Brief. Bioinform.* 13, 524–535. doi:10.1093/bib/bbr071

Stokkeland, L. M. T., Giskeødegård, G. F., Ryssdal, M., Jarmund, A. H., Steinkjer, B., Madssen, T. S., et al. (2022). Changes in serum cytokines throughout pregnancy in women with polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* 107, 39–52. doi:10.1210/clinem/dgab684

Tarca, A. L., Romero, R., Benshalom-Tirosh, N., Than, N. G., Gudicha, D. W., Done, B., et al. (2019). The prediction of early preeclampsia: Results from a longitudinal proteomics study. *PLOS ONE* 14, e0217273. doi:10.1371/journal.pone.0217273

Thiel, M., Féraud, B., and Govaerts, B. (2017). ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *J. Chemom.* 31, e2895. doi:10.1002/cem.2895

Timmerman, M. E., Hoefsloot, H. C. J., Smilde, A. K., and Ceulemans, E. (2015). Scaling in ANOVA-simultaneous component analysis. *Metabolomics.* 11, 1265–1276. doi:10.1007/s11306-015-0785-8

Twisk, J., Bosman, L., Hoekstra, T., Rijnhart, J., Welten, M., and Heymans, M. (2018). Different ways to estimate treatment effects in randomised controlled trials. *Contemp. Clin. Trials Commun.* 10, 80–85. doi:10.1016/j.conctc.2018.03.008

van der Leeden, R., Meijer, E., and Busing, F. M. (2008). "Resampling multilevel models," in *Handbook of multilevel analysis*. Editors J. de Leeuw and E. Meijer (New York, NY: Springer), 401–433. doi:10.1007/978-0-387-73186-5_11

Vis, D. J., Westerhuis, J. A., Smilde, A. K., and van der Greef, J. (2007). Statistical validation of megavariate effects in ASCA. *BMC Bioinforma.* 8, 322. doi:10.1186/1471-2105-8-322

White, J. M., and Jacobs, A. (2021). *log4r: A fast and lightweight logging system for R, based on 'log4j'*. R package version 0.4.2.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

R Special Interest Group on Databases (R-SIG-DB)Wickham, H., and Müller, K. (2021). *Dbi: R database interface*.

Wickham, H., and Seidel, D. (2020). *Scales: Scale functions for visualization*.

Xia, J., Sinelnikov, I. V., Han, B., and Wishart, D. S. (2015). MetaboAnalyst 3.0—Making metabolomics more meaningful. *Nucleic Acids Res.* 43, W251–W257. doi:10.1093/nar/gkv380

Zhu, H., Chao, J., Kotak, I., Guo, D., Parikh, S. J., Bhagatwala, J., et al. (2013). Plasma kallistatin is associated with adiposity and cardiometabolic risk in apparently healthy African American adolescents. *Metabolism.* 62, 642–646. doi:10.1016/j.metabol.2012.10.012

frontiers | Frontiers in Molecular Biosciences

TYPE Review
PUBLISHED 14 November 2022
DOI 10.3389/fmolb.2022.967205

OPEN ACCESS

EDITED BY
Sumeet Agarwal,
Indian Institute of Technology Delhi, India

REVIEWED BY
Eve Syrkin Wurtele,
Iowa State University, United States
Patrick May,
University of Luxembourg, Luxembourg

*CORRESPONDENCE
Thomas H. A. Ederveen,
Tom.Ederveen@radboudumc.nl
Emile R. Chimusa,
emile.chimusa@northumbria.ac.uk
Peter A. C. 't Hoen,
Peter-Bram.tHoen@radboudumc.nl

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 12 June 2022
ACCEPTED 20 October 2022
PUBLISHED 14 November 2022

CITATION
Agamah FE, Bayjanov JR, Niehues A, Njoku KF, Skelton M, Mazandu GK, Ederveen THA, Mulder N, Chimusa ER and 't Hoen PAC (2022), Computational approaches for network-based integrative multi-omics analysis. *Front. Mol. Biosci.* 9:967205. doi: 10.3389/fmolb.2022.967205

Computational approaches for network-based integrative multi-omics analysis

Francis E. Agamah[1,2], Jumamurat R. Bayjanov[3], Anna Niehues[3], Kelechi F. Njoku[1], Michelle Skelton[2], Gaston K. Mazandu[1,2,4], Thomas H. A. Ederveen[3]*, Nicola Mulder[2], Emile R. Chimusa[5]* and Peter A. C. 't Hoen[3]*

[1]Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [2]Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, CIDRI-Africa Wellcome Trust Centre, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [3]Center for Molecular and Biomolecular Informatics (CMBI), Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands, [4]African Institute for Mathematical Sciences, Cape Town, South Africa, [5]Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle, United Kingdom

Advances in omics technologies allow for holistic studies into biological systems. These studies rely on integrative data analysis techniques to obtain a comprehensive view of the dynamics of cellular processes, and molecular mechanisms. Network-based integrative approaches have revolutionized multi-omics analysis by providing the framework to represent interactions between multiple different omics-layers in a graph, which may faithfully reflect the molecular wiring in a cell. Here we review network-based multi-omics/multi-modal integrative analytical approaches. We classify these approaches according to the type of omics data supported, the methods and/or algorithms implemented, their node and/or edge weighting components, and their ability to identify key nodes and subnetworks. We show how these approaches can be used to identify biomarkers, disease subtypes, crosstalk, causality, and molecular drivers of physiological and pathological mechanisms. We provide insight into the most appropriate methods and tools for research questions as showcased around the aetiology and treatment of COVID-19 that can be informed by multi-omics data integration. We conclude with an overview of challenges associated with multi-omics network-based analysis, such as reproducibility, heterogeneity, (biological) interpretability of the results, and we highlight some future directions for network-based integration.

KEYWORDS

multi-omics, data integration, multi-modal network, machine learning, network diffusion/propagation, network causal inference

Introduction

Studies that implement large-scale molecular profiling techniques (-omics technologies) have increased our understanding of disease mechanisms and led to the discovery of new biological pathways, genetic loci underpinning disease progression, biomarkers, and targets for therapeutic development (Horgan and Kenny, 2011; Sun and Hu, 2016; Karczewski and Snyder, 2018). Until recently, these studies have mostly relied on single omics investigations. Dependencies between biological features and the relationships between different molecular layers (for example transcriptome, proteome, metabolome, microbiome, and lipidome) remain mostly elusive. The holistic understanding of the molecular and cellular bases of disease phenotypes and normal physiological processes requires integrated investigations of the contributions and associations between multiple (different but parallel) molecular layers driving the observed outcome. Most importantly, genetic information flows from the genome to traits and involves several molecular layers (Sun and Hu, 2016; Hasin et al., 2017). Thus, understanding the genetic architecture of complex phenotypes would involve integrating and investigating the interactions between different molecular layers (Buescher and Driggers, 2016; Hasin et al., 2017; Chakravorty et al., 2018; Zapalska-Sozoniuk et al., 2019).

Multi-omics datasets require appropriate computational methods for data integration and analysis. These methods/models implement statistical, network-based, and/or machine learning (ML) techniques on different omics layers to elucidate key omics features associated with diseases at various molecular levels and predict phenotypic traits and outcomes with increased accuracy (Ritchie et al., 2015; Bersanelli et al., 2016; Zeng and Lumley, 2018).

Based on the hypothesis that molecular features within a system establish functional connections or are part of modules to carry out processes, network-based methods offer a framework to conceptualize the complex interactions in a system as a collection of connected nodes (molecular features). They further suggest possible connections (e.g., genotype to phenotype relationships) and/or subnetworks (e.g., biological pathways) that are informative of an observed phenotype (Chakravorty et al., 2018). Therefore, network-based methods are particularly useful to assess complex interactions within multi-omics datasets and illustrate dependencies among multiple features. In addition, some network-based methods can incorporate prior information to guide the integrative analysis. For this reason, network-based methods have attracted considerable attention in multi-omics data integration around understanding disease mechanisms and drug discovery (Wu et al., 2018; Agamah et al., 2021). Previous reviews have mostly focused on the network-based analysis of single-omics data (Camacho et al., 2018; Yan et al., 2018; Zitnik et al., 2019) or different approaches toward multi-omics data integration (Cavill et al., 2016; Duruflé et al., 2021). Here, we review different integrative network-based approaches and some tools for multi-omics data analysis.

The outline of the review is as follows; we begin with a discussion on integrative multi-omics approaches, where we highlight the approaches for network-based analyses. We then discuss the different classes of methods for multi-modal network analysis. Next, we describe several network-based integrative multi-omics tools. This is followed by a discussion on the application of network-based tools to pertinent biological questions. This section provides guidance on the choice of the most appropriate network-based tools to answer a given biological question. As further examples, we show how some tools have been applied to COVID-19 research, which is currently one of the research areas benefiting from multi-omics integration approaches. Finally, we conclude with a discussion on some challenges associated with multi-omics analysis and the possible directions to mitigate such challenges.

Integrative multi-omics approaches

After initial data selection, processing, and quality assurance, an appropriate data analysis approach needs to be selected. We categorize integrative multi-omics analysis approaches into two main categories, multi-stage and multi-dimensional (multi-modal) analytical approaches (Figure 1) (Holzinger and Ritchie, 2012; Wen et al., 2021). The multi-stage integration involves integrating data from different technologies using a stepwise approach. In this approach, omics layers are analysed separately before investigating statistical correlations between different biological features from the datasets under consideration. This analytical approach puts an initial emphasis on the relationships of features within an omics layer and how they relate to the phenotype of interest (Ritchie et al., 2015). The multi-modal analytical approach involves integrating multiple omics profiles in a simultaneous analysis (Holzinger and Ritchie, 2012; Ritchie et al., 2015; Karczewski and Snyder, 2018; Ulfenborg, 2019).

Methods for multi-modal network analysis

In this review, we focus on (i) machine learning-driven network-based methods, (ii) network-based diffusion/propagation methods, and (iii) causality- and network-based inference methods. The selection criteria were based on the fact that these multi-omics/multi-modal network-based methods implement network architectures together with statistical and mathematical models for integrative multi-omics data analysis. Most of these methods can be

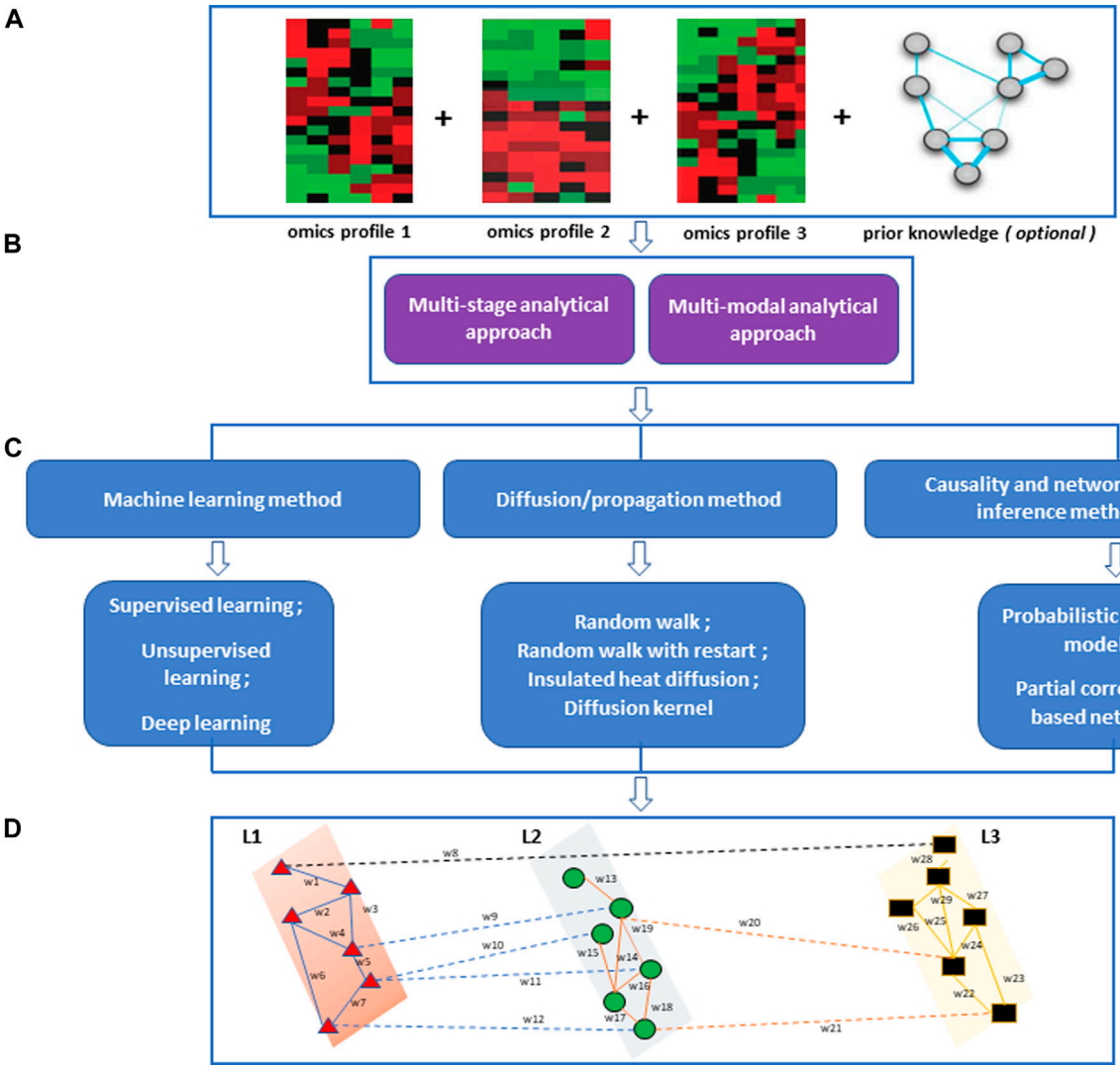


FIGURE 1
An overview of the multi-omics integration approach and the methods for network-based integration. **(A)** Processed omics data and prior knowledge for integrative analysis. **(B)** An integrative multi-omics approach that could be implemented. **(C)** Integrative network-based methods **(D)** Multi-layered network showing intra-layer interaction (solid lines) and crosstalk (dashed lines) across different layers (L1, L2, L3). The nodes are shaped and coloured to represent different omics features within the omics layers they are involved in. The edges are coloured to show different interactions within and between omics layers.

implemented in both multi-stage and multi-dimensional multi-omics analysis (Figure 1).

Machine learning-driven network-based methods

ML is a collection of data-driven techniques for fitting an analytical model to a given dataset. ML methods do not only provide the framework to automatically learn models from large multi-omics data and make accurate predictions but also implement network architectures to exploit interaction across the different omics layers e.g., for exploring omics-phenotype associations (Reel et al., 2021). ML comprises mainly supervised and unsupervised learning methods. Supervised learning uses labelled datasets to train models to yield the desired output and emphasizes predictions by inferring discriminating rules from the data. Supervised learning model training requires comprehensive data and can be time-consuming, while unsupervised learning uses unlabelled data, to find latent structures or patterns in the data.

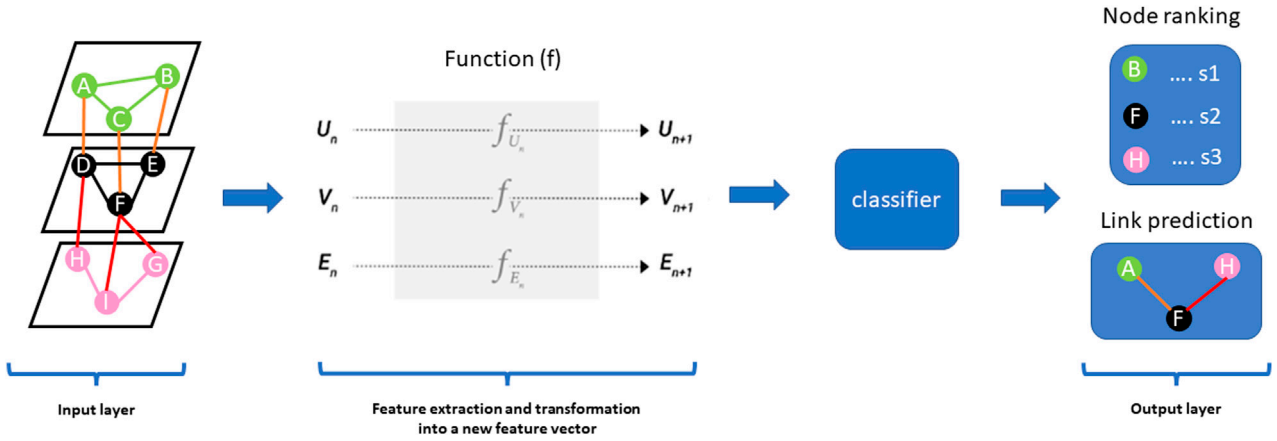


FIGURE 2
Graph Neural Networks (GNNs) are a class of deep learning methods designed to perform inference and predictions on graph data by learning embeddings for graph attributes (nodes, edges, global-context). The concept behind the architecture of these methods is such that it accepts graph data as input and produces the same input graph with updated embeddings before making predictions. GNN uses a function (f) on each graph component vector [nodes vector (Vn), edge vector (En), global-context vector (Un)] in the input graph to learn abstract feature representations of the graph to compute a new feature vector for nodes (Vn+1), edges (En+1) and global-context (Un+1)). The output layer could predict nodes ranked according to a particular score (s1, s2, s3) and also predict edges (links) in the input network.

Classical graph-based ML methods (e.g., label propagation, a method for assigning labels to unlabelled points) can be used for a variety of tasks including generating graph edges, estimating node weights (quantitative measure of node importance) as well as estimating and optimizing edge weights (quantitative measure of the importance of the pairwise interaction between nodes) in a network to exploit the structure of graphs and learn models from the data (Karasuyama and Mamitsuka, 2017). Subsequent network optimization techniques introduce perturbations into the network and identify highly perturbed subnetworks to prioritize the most relevant features that correlate with the biological processes under study.

Multiview/multi-modal ML is an emerging method for multi-omics data integration used to exploit information captured in each omics dataset and infer from the associations between the different data types (Nguyen and Wang, 2020). Multi-view learning implements the alignment-based framework and the factorization-based framework (Nguyen and Wang, 2020). The alignment-based framework is a method based on the supervised setting for seeking pairwise alignment among different omics data whereas the factorization-based framework is based on an unsupervised setting for seeking a common representation of features across different omics layers. Deep learning methods, an example of multiview/multi-modal learning, have become one of the more promising integration methods not only because of their ability to exploit the structure of graph neural networks/graph or convolutional networks in both supervised and unsupervised settings with high sensitivity, specificity, and efficiency compared to classical ML methods but also, the predictive performance and capability to capture nonlinear and hierarchical representative features (Martorell-Marugán et al., 2019; Kang et al., 2022). The hierarchical feature processing can capture complex nonlinear associations in a multi-layered manner. The architecture of deep learning models consists of the input layer, hidden layer(s), and output layer. From the perspective of multiomics data integration, most deep learning methods follow the steps of (i) feature selection, (ii) transforming high dimensional multiomics data into low-ranked latent variables, (iii) concatenating multi-omics features into a larger dataset and (iv) analysing the data for the desired task such as node ranking, link prediction, node classification and clustering (Figure 2) (Kang et al., 2022). It is worth noting that the deeper the hidden layer, the more it can learn complex patterns in the data. A major challenge for deep learning methods is the problem of overfitting due to large features and the small sample size of multi-omics data. In addition, a large amount of cleaned data is required to train and validate the model, thus influencing how the model is interpreted (Kang et al., 2022). We refer the reader to a current review on deep learning in multi-omics data integration by Kang et al. (Kang et al., 2022).

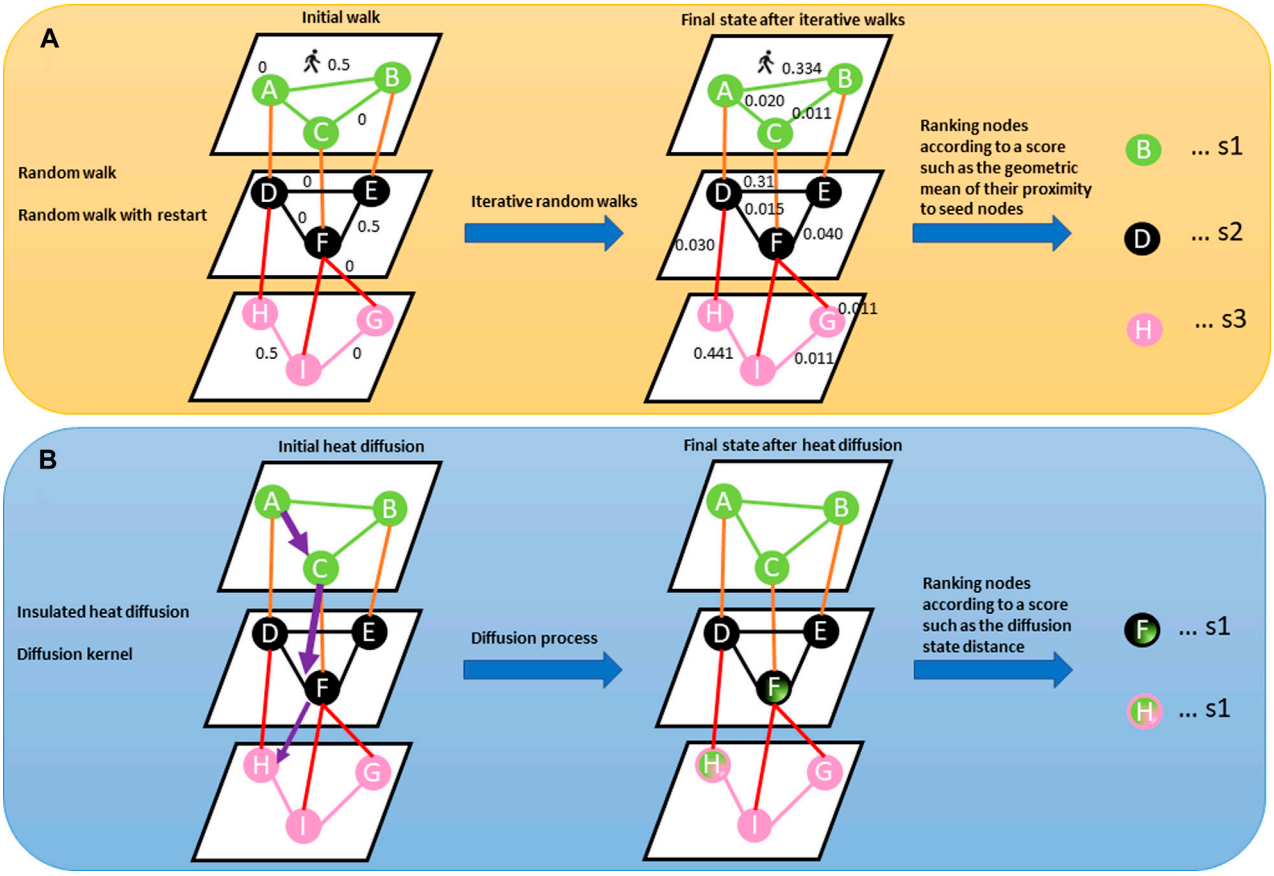


FIGURE 3
(A) Describes a random walk from the seed node (e.g., node A). The concept behind random walk is a guilt-by-association approach where an imaginary particle explores the network structure from seed nodes. The direction of movement of the particle is completely independent of the previous directions moved. At each step, the particle transition from any node in the graph with a certain probability (shown on the edges). The probability flow of random walks on a network is used as a proxy for information flows in the network to study the function of features, subnetworks, and prioritize features in the network. After several iterations, we are interested in the distribution of our position (Stationary distribution) in the graph (final state after iterative walks). The stationary probability distribution can be seen as a measure of the proximity between the seed(s) and all the other nodes in the graph. Nodes within the network can be prioritized using a specific metric (s1, s2, s3) such as the geometric mean of their proximity to seed nodes. **(B)** Describes heat diffusion from a reference query (e.g., node A). The concept behind heat diffusion in biological networks is perturbing nodes and simulating how the disturbance flows across edges within the network. Node disturbance means adding a scalar value (e.g., log fold changes from gene expression experiment, copy number variations) to node(s). Within a biological network, heat diffusion allows for the assessment of connectivity and topology of features which can allow the identification of relevant/dysregulated pathways and/or mutational effects across edges to neighbouring nodes. The purple arrow means diffusion jumps across different layers. The thickness of the purple arrow signifies the effect of query node (A) on nodes (F) and (H) as shown in nodes (F) and (H) in the final state graph after diffusion. Nodes within the network can be prioritized using a specific metric such as diffusion state distance.

Network-based diffusion/propagation methods

Network-based diffusion/propagation is a technique for detecting the spread of biological information throughout the network along network edges, thanks to its ability to amplify feature associations based on the hypothesis that node proximity within a network is a measure of their relatedness and contribution to biological processes (Cowen et al., 2017; Di Nanni et al., 2020). The method has been exploited in many network-based analysis pipelines and is suitable for analysing patient-level molecular profiles with different aims including disease subtyping because of its label propagation (Di Nanni et al., 2020). Propagation methods, including random walk, random walk with restart, insulated heat diffusion, and diffusion kernel networks, provide a quantitative estimation of proximity between features associated with different data types by considering all possible paths beyond the shortest paths (Figure 3) (Cowen et al., 2017; Di Nanni et al., 2020).

From a data analysis perspective, the network diffusion (ND) methods require omics data and network data. The network data could be obtained from *a priori* knowledge, inferred from omics data, or generated using a mixed

approach of *a priori* and novel knowledge (Di Nanni et al., 2020). Omics data information, e.g., genetic aberration events underlying differential expression and/or a biological phenotype, are superimposed on the nodes (source nodes) within the network before the information is propagated *via* the edges until convergence and consensus features are found (Cowen et al., 2017; Di Nanni et al., 2020).

ND methods transform input vectors of scores obtained from the omics data into dense vectors to eliminate missing values and ties. This transformation process can be applied before, after, or during the integration step to refine the results based on molecular network data (Di Nanni et al., 2020). In the ND-before integration approach, the diffusion method is applied to a collection of scores (scores obtained from the omics data) that represent the multi-omics data. The ND-after integration approach is implemented when the various multi-omics data have been initially integrated into a unique structure. The ND-during integration approach is implemented in an instance where each layer exchanges information during the diffusion process. Box 1 provides a summary of the equations related to the diffusion methods.

BOX 1 Summary equations of the network propagation/ diffusion methods

Random Walk

$$x_T = \left[AD^{-1}\right]^k . x_o$$

Random Walk with Restart (RWR)

$$x_T = \alpha\left[I - (1-\alpha)\,AD^{-1}\right]^{-1} . x_0$$

Insulated Heat Diffusion

$$x_T = \alpha\left[I - (1-\alpha)\,AD^{-\frac{1}{2}}\right]^{-1} . x_0$$

Diffusion Kernel

$$x_{T=}e^{\alpha(D-A)} . x_o$$

Where,

x_T is the final state of the network after the propagation of information throughout the network

x_0 is the initial biological information (initial state vector of aberration scores e.g., gene expression scores). A is the adjacency matrix of the network. D is the diagonal matrix of the out-degrees of nodes. AD^{-1} is the normalized adjacency matrix. k is the number of time steps, α is the restart probability and I is an identity matrix

Causality- and network-based inference methods

The mechanism of action within a biological system is fundamental to understanding such a system. For this reason, biological network inference and causal learning can be used to investigate the direct and indirect multi-layer associations and possible causal relations between omics data features in the system (Griffin et al., 2018).

Causal networks are generally graphical representations that demonstrate likely causal relations between nodes by capturing directional interactions and modelling dependencies between biological variables. The method enables researchers to put directionality between features in a network as well as decipher modules (subnetworks) and/or features associated with patient survival, disease processes, or pinpoint sources of perturbations within multi-omics biological network data (Hawe et al., 2019).

Partial correlation-based networks enable the inference of features regulating co-expression or the activities of other features within the network by estimating conditional dependencies (partial correlations) (Hawe et al., 2019). Partial correlation corrects for spurious associations among features that are mediated by other variables measured in the dataset, thereby reducing the density of the network and enhancing its interpretability (Hawe et al., 2019). These methods have been implemented to infer mechanistic regulatory interactions or predict markers in biological networks (Hawe et al., 2019).

Alternatively, network-based computational frameworks that implement probabilistic graphical models offer attractive solutions for causal reasoning and inference over multi-omics data (Friedman, 2004; Koller and Friedman, 2009; Griffin et al., 2018). A probabilistic graphical model (PGM) is a graph technique for modelling joint probability distributions and (in)dependencies over a set of random variables (Koller and Friedman, 2009). From a data analysis perspective, PGM uses graph-based representation (nodes as features and edges as direct probabilistic interactions between node pairs) as the basis to encode the complex distribution of the data for probabilistic reasoning and inference (Koller and Friedman, 2009). The framework of probabilistic graphical models includes a variety of directed and undirected models (Koller and Friedman, 2009). Directed models (e.g., Bayesian networks) require pre-defined directionality or capture conditional (in)dependencies to assert an influence on features. Undirected models (e.g., Markov networks) are undirected graphical models that offer a simpler perspective on directed models, especially in instances where the directionality of the interactions between features cannot be determined. Compared to directed models which can be used for causal reasoning and inference, undirected models are limited to inference tasks because they fail to capture the influence of nodes on neighbouring nodes.

In addition to partial correlation and probabilistic graphical models, advanced ML models and frameworks that are more computationally efficient have been explored for inferring causal relationships between multi-modal data (Peters et al., 2017; Badsha and Fu, 2019; Luo et al., 2020; Wein et al., 2021). Also, new methods

TABLE 1 Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
Machine learning-driven network-based tools									
mixOmics	An R toolkit dedicated to the exploration and integration of biological data sets with a specific focus on variable selection. The package contains suite of algorithms and functions. The function network is used for graph visualization	1) Receives as input multiple matrices each representing a different omics 2) Perform network analysis using the network function	Infer interactions between nodes by using a pairwise association score	Leverages on measurements of variables	Relevance networks	Supervised and unsupervised ML	most omics types (genes, mRNA, metabolites, miRNomics data, proteomics)	2012	González et al. (2012)
Similarity network fusion	A network-based framework that uses networks of samples as a basis for integration. It fuses individual networks from each omics layer to represent the full spectrum of underlying data	1) SNF first creates a sample-similarity network for each omics level and then fuses these into one network using a nonlinear combination method	Uses a scaled exponential similarity kernel to determine the edge weight. The weighted edges represent pairwise sample similarities	Nodes represent samples and the node size represents a phenotype like survival	Identifies disease subtypes, performs survival prediction	Unsupervised ML	most omics types (mRNA, DNA methylation, and microRNA (miRNA) expression data)	2014	Wang et al. (2014b)
Lemon-Tree	A multi-omics module network inference software suite that finds co-expressed gene clusters and reconstructs regulatory programs involving other upstream omics data	1) Infer co-expressed gene clusters 2) Build consensus modules using the spectral edge clustering algorithm 3) Build module network 4) Module learning	Computes edge weight which represents the frequency with which pairs of genes belong to the same cluster	Compute the regulator score and considers the number of trees a regulator is assigned to, with what score (posterior probability), and at which level of the tree	Predicts driver genes/biomarker	Unsupervised ML	expression data, copy number, microRNA, epigenetic profiles	2015	Bonnet et al. (2015)
Multiscale Embedded Gene Co-expression Network Analysis (MEGENA)	An R package co-expression network analysis framework that effectively and efficiently constructs and analyses co-expression networks	1) Constructs fast planar filtered network 2) Identify multi-scale clustering structures 3) Perform multiscale hub analysis 4) Perform cluster-trait association analysis	Computes a similarity score between node pair	Compute node degree as node weight/size	Predicts subnetworks, driver hubs	Unsupervised ML	Genes, mRNA, Fast planar filtered network	2015	Song and Zhang, (2015)
Omics Integrator	The approach applies advanced network optimization algorithms to a network to find	1) Garnet identifies a set of transcriptional factors associated with mRNA expression changes by	Uses least-squares regression to relate the transcription factor	Transcription factors with motifs exhibiting statistically significant regression coefficients	Predicts subnetworks that connect changes	Supervised ML	most omics types (mRNA, epigenetic changes, proteins, metabolites)	2016	Tuncbag et al. (2016)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
	high-confidence, interpretable subnetworks that best explain the data	incorporating epigenetic changes nearby expressed genes	affinity scores to mRNA expression changes	are given a weight of–log (*p*-value)	observed in omics data				
	The software is comprised of the Garnet and Forest tools	2) Garnet scans regions proximal to transcribed genes for transcription factor binding sites and then regresses transcription factor affinity scores against gene expression changes Forest provides perturbation strategies for perturbation analyses to determine the robustness of a network	Forest converts uniform edge weights to costs using a scoring function	The prize function assigns negative weights to nodes based on the number of connections they have in the interactome					
		3) Forest identifies a condition-specific functional sub-network from user data and a confidence-weighted interactome							
		4) The confidence-weighted interactome is integrated with the 'omic' hits using the prize-collecting Steiner forest algorithm, where the data is either connected directly or *via* intermediate nodes, called 'Steiner nodes'							
Weighted Similarity Network Fusion	A method that implements a modified similarity network approach to identify disease subtypes. It	1) Build a regulatory network from the input data 2) Calculating the weight for each feature	Considers the similarity of two patients by considering the overall difference between the expression levels of all	Computes feature weights by first ranking features using a modified PageRank algorithm followed by	Identifies disease subtypes, performs survival prediction	Unsupervised ML	miRNA, mRNA, transcription factors	2016	Xu et al. (2016)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
	accounts for feature weights when clustering patients	and ranking the features based on network information and the expression variation of the features 3) Obtain weighted sample similarity networks from genes (mRNAs, TFs) and miRNAs separately using the weights and expression data of the features 4) Perform network fusion and clustering to find patient groups that imply disease subtypes	their features and the weight of each feature	Integrating feature ranking and feature variation					
iOmicsPASS	A method for integrating multi-omics profile over genome-scale biological networks and identifying predictive subnetworks that provides the mechanistic interpretation of a specific phenotype The tool considers molecular interactions within and between omics data types as a data feature	1) Integrates quantitative multi-omics data by computing interaction scores for a network 2) Discover molecular interactions whose joint expression patterns predict phenotypic subnetworks/groups 3) Report biological pathways enriched in the subnetworks using a modified nearest shrunken centroid algorithm	Computes scores for each molecular interaction. The scores are derived in the context of the type of interactions data (TF regulatory network and protein-protein interaction network with or without DNA copy number)	Utilizes measurement of each molecule in their respective omics data sets as node score	Predicts phenotypic group-specific subnetworks, feature selection	Supervised ML	Biological network, mRNA, proteomics data, DNA copy number, sample metainformation	2019	Koh et al. (2019)
Sparse CRossmodal Superlayered Neural Network (SCR-SNN)	A subtype classification model that represents a sparse version of a cross-modal super-layered neural network	1) Biomarker filtering 2) Biomarker selection, using a cross-modal, super-layered neural network	Estimates connection between nodes	Compute weight for nodes	Predicts disease subtype	Neural network	DNA methylation, mRNA	2020	Joshi et al. (2020)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
		3) Integration of selected biomarkers from omics data 4) Prediction model building							
Integrative Network Fusion	A framework for high-throughput omics data integration that leverages machine learning models to extract multi-omics predictive biomarkers	1) A set of top-ranked features is extracted by juxtaposition by Random Forest (RF) and linear Support Vector Machine (LSVM) classifiers 2) A feature ranking scheme is computed on similarity network fusion-integrated features 3) A random forest model is trained on the intersection of two sets of top-ranked features from the juxtaposition and feature ranking scheme (rSNF) and provides compact predictive biomarkers	Uses a scaled exponential Euclidean distance kernel to compute edges weight	Implements a feature ranking scheme on similarity network fusion integrated features	Identifies disease subtypes and predictive biomarkers	Supervised ML	mRNA, microRNA expression, protein levels, copy number variants, DNA Methylation	2020	Tuncbag et al., (2016); Chierici et al., (2020)
Discovery of active Modules In Networks using Omics (DOMINO)	A network-based active module identification algorithm used for identifying subnetworks that show significant over-representation of accrued activity signal ("active modules")	1) Receives as input a set of genes flagged as the active genes in a dataset and a network of gene interactions 2) Partition the network into disjoint, highly connected subnetworks 3) Detect relevant subnetworks containing active over-represented genes 4) Further, refine subnetworks into compartments	Uses the confidence scores of the tissue-specific functional interactions as weights of edges	Uses gene activity scores	Predicts subnetworks	Unsupervised ML	gene network and transcriptomics data	2021	Levi et al. (2021)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
		5) Repartition's subnetwork compartments in putative modules 6) Reports final modules that are over-represented by active genes							
multi-source information super network	A network-based framework for constructing a single network from multi-source data	1) Constructs a super network based on the weighted sum of the pairwise weighted edge vectors (for each pair of genes)	Computes edge weights	Computes gene-specific scores based on characteristics and topology of the super network	Predicts subnetworks	Unsupervised ML	Genes, pathway information, CNVs, Drug data, mRNA, miRNA, PPI	2018	Zachariou et al. (2018)
i-Modern	A deep learning network framework for integrating multi-omics data	1) Feature extraction using optimized autoencoder 2) Low-dimensional feature extraction *via* Cox-PH models 3) Patient subgroup classification	Estimate connection between nodes	Implements a randomization approach to explore node weight	predict omics signatures, patient subgroup classification	Neural network	miRNA, somatic mutations, copy number variation (CNV), DNA methylation, proteins	2022	Pan et al. (2022)
OmicsNet 2.0	A network-based multi-omics analysis platform and an R package (OmicsNetR) to easily build, visualize, and analyze multi-omics networks	1) Accepts different data types as input 2) Search different molecular interaction database 3) Creates multi-omics networks 4) Performs network visual analytics	The methodology does not take edge directionality orweights into account	Uses feature activity scores	Predicts sub-networks, crosstalk	Unsupervised ML	Genes, proteins, transcription factors, miRNAs, metabolites, SNPs, Taxa, lc-ms Peaks	2022	Zhou et al. (2022)
multi-omics data integration for clustering to identify cancer subtypes (MDICC)	A method for multi-omics data integration that implements affinity matrix and network fusion methods	1) Construct an affinity matrix for different omics data based on a Gaussian kernel function 2) Fuse affinity matrices into a new relational matrix with low rank 3) Cluster fused network	Computes edge weight as a measure of the Euclidean distance between samples	Utilizes measurement of each molecule in their respective omics data	Predicts disease subtypes	Unsupervised ML	mRNA, miRNA, proteomics data, DNA methylation	2022	Yang et al. (2022)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
Network-based diffusion/propagation tools									
Tied Diffusion of Interacting Events (TieDIE)	TieDIE method extends the heat diffusion strategies by leveraging different types of genomic inputs to find relevant genes on a background network with high specificity	1) Computes scores for each node in the graph 2) Utilizes multiple diffusion processes to predict disease-related genes, subnetworks, and pathways	The diffusion approach is used to describe the edge score between node pairs (1 and -1). Aij = 1 if node i activates node j, Aij = −1 if node i represses or inactivates node j, and 0 otherwise, where A is an adjacency matrix	Scores between -1 and +1 are assigned to the nodes reflecting a positive or negative association with the disease state A node score of 0 reflects genes not known to be associated with the disease process Nodes scores could represent experimental measurements	Predicts biomarkers and disease-specific subnetworks	Diffusion-based	genes, proteins, biological pathway features, mRNA, DNA methylation	2013	Paull et al. (2013)
Network-based Integration of Multi-omics Data (NetICS)	A gene prioritization method that is a framework for per-sample network-based integration of diverse data types on a directed functional interaction network NetICS provides insight into how aberration events that are different between samples of the same disease type cause similar expression changes in other genes	1) Constructs a directed functional interaction network from input functional interactions 2) Diffuse aberration scores from the aberrant genes following the directionality of the network interactions 3) Diffuse differential expression scores from differentially expressed genes 4) Predicts how aberration events cause expression changes through gene interaction	Compute connectivity scores between node pairs	Compute a ranking score for all genes	Predicts biomarkers	Random walk	miRNA-gene interaction, mRNA, DNA methylation, genetic aberrations, protein levels	2018	Dimitrakopoulos et al. (2018)
Hierarchical HotNet	An algorithm that simultaneously combines network interactions and vertex scores to construct, identify, and rank statistically significant high-weight altered subnetworks across different omics datasets.	1) Combines network topology and vertex scores 2) Defines a similarity matrix from the network using a random walk-based approach 3) Implements hierarchical clustering to	Defines a similarity measure between node pairs using both network topology and vertex scores	Uses vertex scores in the input network	Predicts a hierarchy of mutated subnetworks	Random walk	Interaction network with vertex scores	2018	Paull et al., (2013); Reyna et al., (2018)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
	It addresses the limitations of HotNet (Vandin et al., 2012), HotNet2 (Leiserson et al., 2015) by combating ascertainment bias in data and integrating both network topology and vertex score	construct a hierarchy of clusters consisting of highly connected components 4) Assesses the statistical significance of clusters							
regNet	regNet R package utilizes gene expression and copy number data to learn regulatory networks to estimate the potential impacts of individual gene expression alterations on clinically relevant signature genes	1) RegNet learns a regulatory network from a large collection of paired gene expression and copy number profiles 2) Uses network propagation to quantify the impacts of altered genes sample-specific gene expression changes on other clinically relevant target genes	Compute a connectivity table that represents learned links between genes	Compute impact score for regulator genes, describing the contribution to expression changes in another gene	Predicts driver genes or disease biomarkers	Diffusion-based	transcription factors, mRNA, copy number data	2018	Seifert and Beyer, (2018); Marín-Llaó et al., (2020)
Integrative multi-cohort and multi-omics meta-analysis framework	A multi-omics meta-analysis framework that can identify robust molecular subnetworks and biomarkers for a given disease condition	1) Module (A) takes multiple independent mRNA datasets and performs a leave-one-out meta-analysis to identify reliable differentially expressed genes 2) Module (B) takes multiple independent DNA methylation datasets and identifies differentially methylated genes 3) Module (C) identifies methylation-driven genes 4) Methylation-driven genes are used as inputs in a network propagation algorithm	The confidence score for each protein-protein interaction is obtained from the STRING database	Utilizes experimental values from differential expression and methylation for omics features	Predicts biomarkers and subnetworks describing patients' clinical outcome	Diffusion-based	mRNA, DNA methylation, protein-protein interactions	2019	Shafi et al. (2019)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
		to identify the proposed subnetworks							
Random walk with restart on multiplex and heterogeneous biological networks	A random walk algorithm able to exploit multiple biological interaction sources to integrate multiplex-heterogeneous networks	1) Define adjacency matrix for input networks 2) Compute transition probabilities of the random walk with restart 3) Performs propagation from seed nodes	Generates weighted or unweighted adjacency matrix	Scores nodes according to their proximity to the seed nodes	Predicts candidate features and subnetworks	Random walk	Multi-modal data	2019	Valdeolivas et al. (2019)
MultiPaths	A Python framework to build customized harmonized multi-omics networks from multiple biological databases. MultiPaths framework contains two independent Python packages: DiffuPy and DiffuPath useful for interpreting and contextualizing results from multi-omics experiments	1) DiffuPy implements four existing network propagation algorithms and five graph kernels and enables propagating user-defined labels, either as lists of entities or lists of entities with their corresponding quantitative values 2) DiffuPath, wraps the generic diffusion algorithms from DiffuPy and applies them to construct biological networks	The methodology does not take edge directionality or weights into account for propagation	Compute node scores using a function of graph kernel and input scores	Predicts subnetworks	Diffusion-based	genes, mRNA, metabolites, miRNomics data, biological pathway/ processes data	2020	Reyna et al., (2018); Marín-Llaó et al., (2020)
Analytic and integration framework for multi-omics longitudinal datasets	An integrative framework for building multi-omics networks from longitudinal datasets. It consists of multi-omics kinetic clustering and multi-layer network-based analysis. The method is based on the modeling and clustering of expression profiles with similar behaviours using the timeOmics (Bodein et al., 2019) approach	1) Performs network reconstruction 2) Perform over-representation analysis	Infers correlations between molecules based on multi-omics data	Uses experimental measurements as node scores	Identify crosstalk, key biological functions, or mechanisms	Random walk	Metabolites, genes, protein abundance, mRNA	2020	Bodein et al. (2020)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
Random Walk with Restart for multi-dimensional data Fusion (RWRF)	The method uses a similarity network of samples as the basis for integration	1) Construct a similarity network for each data type 2) Fuse similarity networks 3) Performs random walk with restart on the multiplex network 4) Performs network clustering	Edge weight is estimated by calculating the similarity measure	Estimate stationary probability distribution which indicates similarity between the seed node and other nodes	Identify disease subtypes	Random walk with restart	mRNA, DNA methylation, microRNA	2021	Wen et al. (2021)
Causality- and network-based inference tools									
Differential network analysis in genomics (DINGO)	DINGO is a pathway-based model for estimating patient group-specific networks and making inferences on differential network activation between patient-specific groups. DINGO jointly estimates the group-specific conditional dependencies by decomposing them into global and group-specific components	1) Estimates global component, which represents the relations common to both patient-specific groups 2) Estimates local group-specific component which represents the differential unique relations in each patient-specific group 3) Determines significant differential edges	Constructs differential scores for group-specific edges	The vertices are ordered by their degree centrality	Predicts driver genes	Differential network approach	mRNA, DNA copy number, DNA methylation, microRNA	2015	Ha et al. (2015)
Permutation-based Causal Inference Algorithms with Interventions	The non-parametric algorithm is used to learn directed acyclic graphs comprising both observational and interventional data. An example is the greedy sparsest permutation algorithm	1) Generate an interventional distribution 2) Search for a permutation 3) Learn from interventions	Estimates edge weight	Utilizes experimental measurements of features	Allows for inference of causal graphs	Unsupervised ML	Multi-modal data (omics, clinical data)	2017	Wang et al. (2017)
iDINGO	iDINGO R package is an expansion of DINGO. The package estimates group-specific dependencies between different omics data and make inferences on the	1) Integrate ordered data platforms using the chain graph model 2) Constructs differential scores for group-specific edges to	Constructs differential scores for group-specific edges	The vertices are ordered by degrees (number of connections)	Predicts hub omics features characterized by the number of differential edges	Differential network approach	mRNA, DNA copy number, DNA methylation, microRNA	2018	Class et al. (2018)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
	integrative differential networks, considering the biological hierarchy among the omics platforms. It integrates omics data using the chain graph model	determine the significant differential edges							
prior incorporation Mixed Graphical Model (piMGM)	Can learn with accuracy the structure of probabilistic graphs over mixed data by appropriately incorporating priors from multiple sources Identifies gene pathways associated with disease subtype	1) Incorporates prior information from multiple sources 2) Score the reliability of prior information by using a weighted scheme 3) Merge prior information into a single prior distribution for each edge 4) Learning the structure of probabilistic graphs 5) Uses separate regularization parameters for edges with and without priors 6) Determine active pathways	Leverage conditional dependencies to estimate the strength of edges	Utilizes experimental measurements of features	Identify disease subtypes, active pathways in healthy and disease samples	Probabilistic graphical model	Multi-modal data (omics, clinical data)	2018	Ha et al., (2015); Manatakis et al. (2018)
CausalMGM	A method for learning a causal graph over variables of mixed type linked to disease diagnosis and progression	1) Learn the undirected graph over mixed data types 2) Perform local directionality determinations with conditional independence tests	Leverage conditional dependencies to estimate the strength of edges	Leverages on measurements of variables	Identify causal pathways, biomarkers, and patient stratification	Probabilistic graphical model	Multi-modal data (omics, clinical data)	2019	Sedgewick et al. (2019)
Multi-Omic inTegrative Analysis (MOTA)	A network-based method that uses data acquired at multiple layers from the same set of samples to rank	1) Builds a differential network 2) Computes partial correlation between	The weight of edges represents the partial correlation (above threshold) between node pairs	Computes an activity score (MOTA Score) for each node based on its *p*-value and its connected nodes	Predicts driver genes or disease biomarkers	Differential network approach	mRNA, metabolite, glycomics data, proteins	2020	Class et al., (2018); Fan et al., (2020)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
	candidate disease biomarkers	node pairs using graphical LASSO 3) Calculates the differential partial correlation to determine intra-omics connections for the network							
Integrative multi-omics network-based approach (IMNA)	An integrative multi-omics framework for regulatory network analysis	1) SNP-gene mapping pairs collection 2) Construct SNP-gene bipartite network 3) Construct a functional interaction network 4) Computes signature score for nodes in the network 5) Computes composite score to provide quantitative evidence of node to evaluate the importance of the regulatory function 6) Perform key driver analysis on tissue-specific gene interaction networks	Uses the confidence scores of the tissue-specific functional interactions as edge weight	Computes signature scores for each node from different networks. Signature scores for a gene from different networks are combined and normalized to get a composite score for each gene	Identifies tissue-specific gene interaction networks and key nodes	Bayesian network approach	GWAS signals, eQTLs, epigenomic regulatory annotations, mRNA, protein interactome, and chromatin long-range interactions	2020	Chen et al. (2020)
MRPC	An R package that learns causal graphs and allows for inference	1) Learning the graph skeleton 2) Orienting edges in the skeleton 3) Simulating continuous and discrete data 4) Assessment of inferred graphs	Incorporates the principle of Mendelian randomization as constraints on edge direction	Utilizes experimental measurements of features	Allows for inference of causal graphs	Unsupervised ML	Genomic data, mRNA	2021	Badsha et al. (2021)

(Continued on following page)

10.3389/fmolb.2022.967205

TABLE 1 (*Continued*) Network-based multi-omics integrative tools for predicting biomarkers, crosstalk, disease subtypes, and subnetworks/enriched modules.

Tool	Description	Major steps of the tool	Edge weighting component	Node weighting component	Outcome	Method/ Approach	Input data type	year	References
MIMOSA2	An R package and web application metabolic network-based tool for inferring relationships in microbiome-metabolome data	1) Construct a community metabolic model by linking microbiome data features to reference databases 2) Compute community metabolic potential (CMP) scores for each taxon, sample, and metabolite 3) Aggregate CMP scores at the community level 4) Evaluates the relationship between total CMP scores and metabolites by fitting a linear regression model	The method does not take edge directionality or weights into account	Utilizes CMP score for each feature	Allows inference of microbe-metabolite relationships and predicts disease-associated features	Unsupervised ML	Metabolite, microbiome data	2022	Noecker et al. (2022)

that extend Bayesian networks have been developed for causal inference. For instance, Zheng et al. (Zheng et al., 2018) developed a new method to estimate the structure and inference from a Bayesian network by transforming the structure learning problem into a continuous optimization formulation that does not impose any structural assumptions on the graph. In another instance, Lachapelle et al. (Lachapelle et al., 2019) proposed a novel score-based approach to learning from Bayesian networks *via* the edge weights of neural networks. The approach developed by the authors adapts the optimization method presented by Zheng et al. (Zheng et al., 2018) to allow for non-linear relationships between variables using neural networks. Box 2 provides a summary of the equations related to the Bayesian and Markov methods. Given that the underlying principles behind network-based approaches for analysis vary, combining such approaches is feasible and may increase prediction accuracy as shown by Zheng et al. (Zheng et al., 2018) and Lacapelle et al. (Lachapelle et al., 2019).

BOX 2 Summary equations of the Bayesian and Markov network.

Bayesian Network

Each node in a Bayesian network is represented as a probability distribution of ***cause*** given the observed ***evidence*** which is built from the **Bayes theorem** shown below (Kotiang and Eslami, 2020).

$$P[Cause\,|Evidence] = P[Evidence\,|Cause]\,.\,\frac{P[Cause]}{P[Evidence]}$$

Thus, the full probability model for a Bayesian network is obtained by specifying the joint probability distribution (i.e., a series of the conditional probability distribution of the nodes in the network) (Kotiang and Eslami, 2020).

Markov Network

$$P(X = x) = \frac{1}{z}\exp\Big(\sum_i \omega_i f_i(x_{\{i\}})\Big)$$

Where x is the feature vector, Z is the normalization constant calculated as

$$Z = \sum_{x\epsilon X} \exp\,\Big(\sum_i \omega_i f_i(x_{\{i\}})\Big)$$

f_i is the feature function defined as

$$f_i(x_{\{i\}}) = \begin{cases} 1 & F_i(x\{i\}) = true \\ 0 & otherwise \end{cases}$$

ω_i is the non-negative real-valued weight which reflects constraints on nodes

F_i is the logistic formula.

Review of network-based integrative multi-omics tools

We systematically reviewed literature primarily published between 2010 and 2022 that report on ML-driven network-based tools, network-based diffusion/propagation tools, and causality-

and network-based inference tools. We further highlight the tool's uniqueness in terms of (i) input data types, (ii) method/algorithm implemented, (iii) most important analytical steps, (iv) potential node and/or edge weighting, and (v) predicted outcome (crosstalk, disease subtypes, biomarkers, subnetworks, and patient survival). The tools presented in this review (Table 1) (i) have broad biomedical data applications and are not restricted to specific (disease) research topics only, (ii) are implemented as standalone software like R, MATLAB, Python libraries, or as part of a pipeline and, (iii) account for the weight of nodes and/or edges within the network.

Research questions explored using integrative multi-omics network approaches

Understanding how crosstalk between omics layers impacts a biological process or disease phenotype

A perturbed biological system is characterized by deviations in the behaviour of the molecules (omics data features) causing changes in crosstalk (Figure 1). These changes could become apparent in multiple (connected and dependent) omics levels and may represent a wide range of molecular events responsible for disease phenotype or impaired biological processes.

Network-based diffusion/propagation tools (described in Table 1) offer a framework to identify aberrant omics features (e.g., gene expression, somatic mutations, copy number variations, molecular subnetworks informative of disease subtype) and how their presence and activities within the network induce possible (downstream) changes that might underpin disease phenotype.

In a study to understand the molecular function of SARS-CoV-2 and SARS-CoV proteins and their interaction with the human host, Stukalov et al. (Stukalov et al., 2021) profiled the interactomes of both virus groups and investigated the effect of viral infection on the transcriptome, proteome, ubiquitinome, and phosphoproteome of a lung-derived human cell line. Functional analysis of the various biomolecules within a molecular network revealed crosstalk between the cellular processes during perturbations taking place upon infection at different omics layers and pathway levels. The authors (Stukalov et al., 2021) implemented the Hierarchical HotNet ND method to explore host-SARS-CoV-2 protein interactions during viral infection and its impact on omics levels and cell lines to understand how that could influence molecular pathways. Importantly, the group observed that the transforming growth factor beta (TGF-β) signalling pathway, known for its involvement in tissue fibrosis as one of the hallmarks of COVID-19 (Mo et al., 2020), was specifically dysregulated by SARS-CoV-2 ORF8. Further results revealed that autophagy, one of the mechanisms for controlling SARS-CoV-2 replication and monitoring the progression of viral infection (Sargazi et al., 2021), was specifically dysregulated by SARS-CoV-2 ORF3. These findings highlight the biological relevance of crosstalk and the insights it provides to understanding disease mechanisms.

Identifying modules/subnetworks for disease or disease progression prediction/prognosis

Modular organizations within a network, characterized by clusters of neighbouring nodes highlight features that are functionally related or involved in similar activities within the system. In contrast to identifying (crosstalk of) features informative of disease mechanism, the focus here is on identifying different omics data features that cluster together to inform molecular transitions that describe disease severity level and/or disease subtypes.

Network-based tools that predict disease subtypes or subnetworks informative of a phenotype or a phenotypic group (described in Table 1) are useful for answering such questions and can help in e.g., estimating survival rates across different patient groups. Tools that implement ML and ND-based methods are useful to identify clusters in a network (see Table 1). It is noteworthy that the approach or steps, algorithms, and input data types implemented by such tools to predict subnetworks vary (as described in Table 1). In a recent application of a network-based method to COVID-19 research, Sun et al. (Sun et al., 2021a), employed MEGENA (Song and Zhang, 2015), an unsupervised ML method, to perform protein-metabolite-lipid multi-omics network analysis based on the differential co-expression (correlation between pair of omics features) of these omics data features. The network analysis indicated that tryptophan metabolism and melatonin, a metabolite related to tryptophan metabolism may contribute to molecular transitions in critical COVID-19 patients. Studies have shown that tryptophan and melatonin can improve the immune system and reduce inflammation in COVID-19, suggesting that function disorder may cause impairment to tryptophan metabolism and immune response (Essa et al., 2020; Shneider et al., 2020). Interestingly, activation of tryptophan metabolism has been clinically shown to be selectively enhanced in severe patients (Takeshita and Yamamoto, 2022). The authors further identified pathologically-relevant lipid modules which are being altered among mild COVID-19 patients.

Interestingly, connections between clusters/modules in the omics data may explain the crosstalk of biological features which are specific to the disease state and may serve as biomarkers for monitoring disease progression, treatment, and management (Yan et al., 2016; Overmyer et al., 2020; Su et al., 2020).

TABLE 2 Useful network-based integrative multi-omics tools for drug discovery.

Tool/Method	Description	Major steps of tool	Outcome	Method/ Approach	Input data type	Year	References
DTINet	A computational pipeline focuses on learning a low-dimensional vector representation of features, which accurately explains the topological properties of individual nodes in the heterogeneous network, and then makes prediction based on these representations *via* a vector space projection scheme	1) Integrates a variety of drug-related information sources to construct a heterogeneous network 2) Applies a compact feature learning algorithm to obtain a low-dimensional vector representation of the features 3) Finds the best projection from drug space onto protein space 4) Infers new drug-target interactions	Drug–target interactions	Unsupervised ML	drug-related information protein-protein interactome	2017	Luo et al. (2017)
DrugComboExplorer	A tool for identifying driver signalling pathways and inferring the polypharmacy efficacies and synergy mechanisms through drug functional module-induced regulation of target expression analysis	1) Identify the seed (driver) genes 2) Explore networks from the seed genes by integrating the RNA-seq profiles and pathway knowledge 3) Explore networks from the seed genes by integrating the methylation profiles and pathway data 4) Combine the networks generated from the RNA-seq data and the methylation data	Prioritize synergistic drug combinations, Uncover potential mechanisms of drug synergy	Unsupervised ML	DNA sequencing, gene copy number, DNA methylation, RNA-seq data	2019	Huang et al. (2019)
Reciprocal nearest neighbour and contextual information encoding (RNCE)	A network integration approach accounting for network structure by a reciprocal nearest neighbour and contextual information encoding (RNCE) approach	1) Applies the similarity network fusion (SNF) approach to fuse drug networks 2) Generate contextual information network 3) Compensate for the contextual information network with the initial SNF network	Predicts drug targets, drug mechanism of action	Unsupervised ML	Pharmacogenomic data such as gene expression data under drug perturbation or drug sensitivity data at the cell-line level	2021	Chen and Wong, (2021)

Identifying candidate drivers of disease mechanisms

The contributory effect of features (nodes) within a system varies and depends on factors including but not limited to the level of feature expression or abundance, the level of interaction with other features, and the (background) state of the system. While some of these omics data features are passive (i.e., have little or no effect on system stability), others may have a significant effect on the observed phenotype.

In many biological disease-related problems, exploring relationships between multi-omics data extends beyond measuring marginal associations between features. Thus, identifying biologically relevant nodes that influence changes

within the system could serve as candidate disease-related nodes responsible for an underlying phenotype (Dimitrakopoulos et al., 2018). Causal and network inference methods described in Table 1 can be implemented to explore likely causal features, potential causal relationships, and infer networks that differentiate severe disease from mild in a multi-modal network. Although causal methods provide insights into likely causal agents, investigating and confirming true causality extends beyond computational analysis to experimental validation in relevant models. Also, ML and diffusion-based methods can be used to explore candidate drivers. We describe in Table 1 some network-based tools that predict candidate disease-related nodes. In a recent COVID-19-related study, Tomazou et al. (Tomazou et al., 2021) implemented a network-based multi-omics data integration approach based on a multi-source information super-network scheme (described in Table 1) to prioritize COVID-19-related genes that could be useful as drug targets. The super network was constructed based on the weighted sum of the pairwise weighted edge vectors (for each pair of features) obtained from different sources. The method then prioritizes genes in the network by calculating a characteristic score known as the Multi-source Information Gain (MIG). Some of the genes identified by the authors include Serum Amyloid A (SAA1, SAA2, SAA3) which has been clinically verified as a sensitive biomarker in evaluating the severity and prognosis of COVID-19 (Li et al., 2020), C-reactive protein (CRP) clinically shown to be a marker of systemic inflammation associated with adverse outcomes in COVID-19 patients (Smilowitz et al., 2021), Serine proteinase inhibitor A3 (SERPINA3) shown to be a biomarker for COVID-19-related organ damage (coronary artery disease) and erythropoiesis impairment (Demichev et al., 2021), and vascular cell adhesion molecule (VCAM1) shown to be a vascular and inflammatory implicated in the inflammatory response to sever COVID-19 (Birnhuber et al., 2021).

Drug discovery

Network-based methods that employ systematic integration of disease-specific omics profiles coupled with drug-related data (e.g., FDA-approved, experimental drugs, drug-target interactions) into a heterogeneous network have been shown to provide answers to biological questions related to drug development (Wang et al., 2014a; Vitali et al., 2016; Luo et al., 2017). In this type of network analysis, nodes could represent both omics data features and non-omics data features such as drugs, diseases, and drug targets. The edges represent the functional association between the data types such as pharmacological or phenotypic information.

The network-based view of drug discovery and development may involve multiple methods or tools at different steps. ND and ML methods have been widely implemented in this research area to make predictions (Luo et al., 2017; Tomazou et al., 2021). Predictions from such methods present an effective way to complement experimental methods with the aim of, (i) identifying drug targets, (ii) understanding the disease-drug relationship, (iii) investigating drug-target interactions, (iv) identifying potential drug candidates, (v) drug response prediction, (vi) drug-drug relations, and (vii) predict effective drug combinations. Of note, driver nodes or subnetworks as predicted by tools described in Table 1 might also inform on drug targets. An interesting application of network-based methods for drug discovery is the COVID-19 study by Tomazou et al. (Tomazou et al., 2021), whereby some of the predicted candidate compounds including dexamethasone, atorvastatin, beta-estradiol, cyclosporin-A, imatinib, and remdesivir have been found to generate promising results in clinical trials (https://clinicaltrials.gov/). We describe in Table 2, some useful integrative multi-modal network-based tools that are specifically for drug discovery.

Current challenges and recommendations

Design of experiment

The choice of a network-based integration method does not only depend on the biological question but also the experimental design. Certain network-based methods can only deal with paired data, whereas others can also deal with sparse datasets where there is no or only partial overlap between the samples profiled with the different omics layers. Importantly, the scope of the research will inform the type of data that should be generated. For instance, the paired data, herein referring to different omics data measurements from the same biological sample, is preferred when establishing a holistic picture of systems biology underpinning molecular mechanisms linked to disorders, whereas non-paired data (data generated from different biological samples) is more appropriate for comparative (meta)analysis of samples or omics data measurements. It is therefore recommended to consider the scope of research and the network-based methods that fit.

Reproducibility

Researchers routinely expect that results generated by applying network models are reproducible. For network-based methods, the key issues related to reproducibility are non-harmonized data, biased model evaluation, and lack of transferable code or software. First, multi-omics network-based integration involves the use of heterogeneous data, and some sort of data harmonization is required. A promising approach to harmonize multi-omics research is to ensure that the data comply with FAIR data

principles (findability, accessibility, interoperability, and reusability). The data FAIRification process ensures that a (meta)data schema/method which captures relations between (omics) measurements, data structure, and concepts are clearly defined and easily interpretable by both humans and computers. The metadata schema provides information about the omics data structure and facilitates easy mapping of measured features onto persistent identifiers and established biological networks to investigate the connection between network elements (Krassowski et al., 2020). Second, confidence in multi-omics network-based methods requires systematic evaluation and validation of both datasets and models as a prerequisite for benchmarking toward reproducibility (Krassowski et al., 2020). This approach requires harmonized datasets of quality and quantity that provide unbiased ground truth to ensure that the model at least predicts biologically verified features or edges. Given that there is no gold standard metric for validation, it is critical to validate on a variety of data sources and use metrics that are robust to the level of missing data. Third, to replicate results from previous studies, a detailed report of the analysis together with executable analysis code is important to achieve this purpose. The report and code could be hosted in repositories (e.g., GitHub, Bitbucket, GitLab), reproducible scientific workflow management systems (e.g., Nextflow, Galaxy), environment sharing avenues (e.g., Conda, Docker), or packaged as libraries for programming languages (Canzler et al., 2020). In addition to the key issues, adapting general best practices in the computational analysis will aid reproducibility.

Heterogeneity

Heterogeneity (a measure of variation) of multi-omics datasets, characterized by diverse data sources, data types, and data structure results in computational complexity, analysis bias, and hampers a robust and reproducible integrative network analysis (Lee et al., 2021). There is an increasing awareness of controlling heterogeneity across multi-omics integrative analysis, but most of them are focused on paired data rather than non-paired data.

In the context of network-based integrative analysis developing models and algorithms that could account for non-uniformity by identifying the most robust signals encompassing data, heterogeneity is important. This could be in the form of variable selection models to identify important covariates with the strength of multiple datasets, and yet maintain the flexibility of variable selection between the datasets to account for the data heterogeneity (Lima et al., 2020).

(Biological) Interpretation of results

Interpreting results from an integrative multi-omics analysis is a process of disentangling multiple functional relationships. Primarily, the systematic interpretation of results depends on the kind of biological question and the type of omics measurements used for the analysis. Different omics technologies may have different levels of completeness and sensitivity in terms of detecting biological features. This might result in some omics data types containing more information than others as well as impact the results significantly (Jung et al., 2020). It is important to consider the inherent relationship between the omics profiles used during the interpretation of the results. More often functional annotation of features is based on generalized information which allows a less comprehensive understanding of the molecular mechanisms underlying a phenotype. For this reason, incorporating relevant contextualized pathway information (e.g., tissue-specific or cell-specific) in the analysis has been useful to assess the functional relevance of nodes and subnetworks on the disease/phenotypic landscape, thereby facilitating interpretation.

The capacity to interpret predicted features and interactions of known biological relevance may take the form of deductive reasoning or semantic similarities to support a hypothesis (Guo et al., 2022). In the context of algorithms, robust node weighting and edge weighting metrics measured based on known evidence (e.g., text mining, contextualized pathway information) is important to make an inference that is potentially biologically grounded and experimentally confirmable, knowing that the association between omics layers extends from one-to-one and one-to-many to many-to-many.

Sparsity

There is sparsity at the sample level (not all samples have been profiled with the same assays) and at the feature level. The latter is far more prominent in metabolomics and proteomics than in DNA and RNA sequencing. This is mainly due to the selection of peaks (intensities observed in MS1 survey scans) for fragmentation by data-dependent acquisition (DDA) or data-independent acquisition (DIA) tandem mass spectrometry (LC-MS/MS) approaches (Guo and Huan, 2020; Davies et al., 2021). Typically, an ideal acquisition mode ought to produce spectra of high quality for as many of the ions present in the sample as possible, however, that is not the case, resulting in sparsity at the feature level. This issue is partly but not completely resolved in the newer DIA and integrated DDA-DIA modes which operate in a less-selective manner and have higher coverage as compared to the older DDA mode (Sun et al., 2021b; Davies et al., 2021).

Another contributing factor to sparsity in omics data in tandem with omics technologies is the absence of accumulation of a molecule to a detectable level by omics platforms (evidenced even across platforms of the same omics technology (*e.g.,* next-generation RNA and DNA sequencing). This is partly associated with experimental design, poor biological sample quality, and sample processing.

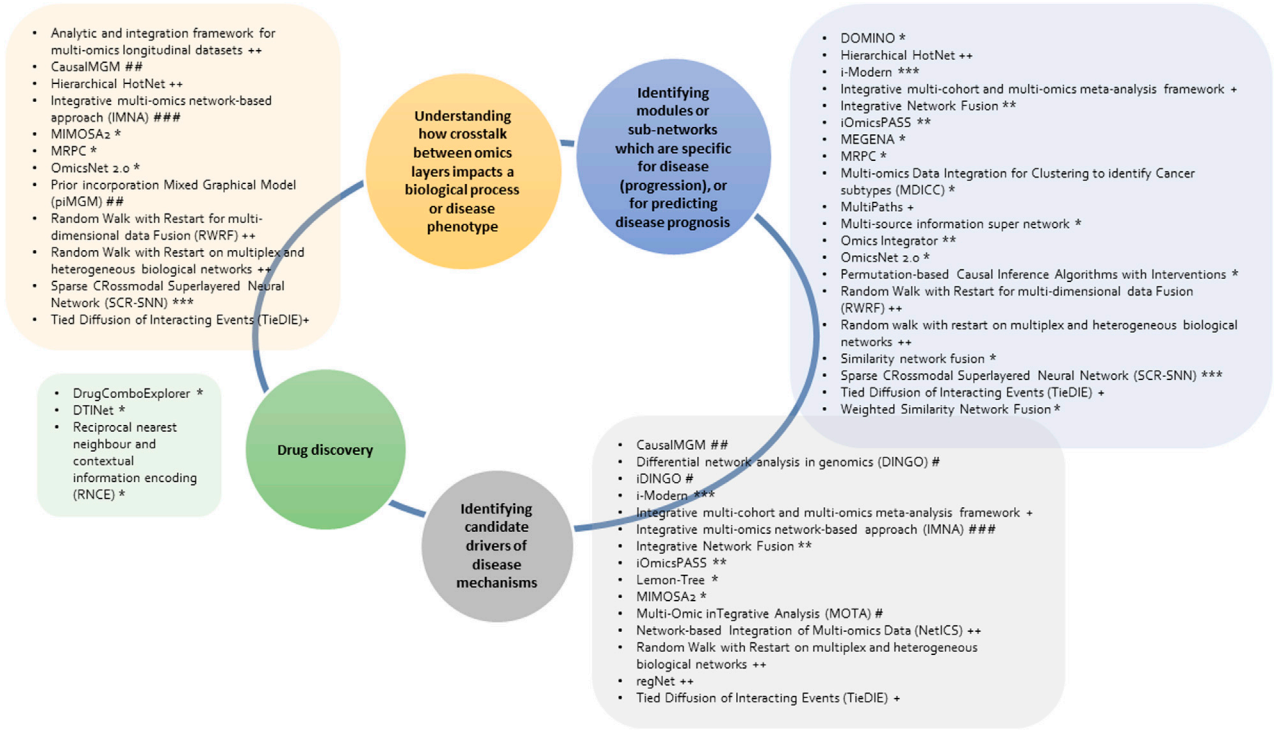


FIGURE 4
Overview of the discussed network-based multi-omics integrative tools and research questions (in the circle) that they can be applied to. The tools implement different methods including unsupervised machine learning (*), supervised machine learning (**), neural networks (***), diffusion-based (+), random walk (++), differential network (#), probabilistic graphical model (##) and Bayesian methods (###).

For computational analysis purposes, mputation can be used to solve missing value problems; however, imputation does not apply to all omics data types (Folch-Fortuny et al., 2015). In addition to imputation, sample similarity measurement methods such as matrix calibration (Li, 2015) and the Mahalanobis distance approach (Sitaram et al., 2015) could be useful to extrapolate for missing values, however, these methods are also limited to specific omics data types. Thus, a feature may have values only in a small percentage of samples leading to sparse matrices, where features may have a wide variety of distributions. Some multi-omics data integration methods can handle sparse data and also feature reduction methods; however, skewed estimates might result in a biased interpretation of results (Greenland et al., 2016). To address the issue of sparsity in the context of networks, network integration aggregates independent data sources to form a more comprehensive attributed interactome, where the edges are qualified by specific semantic relations or similarity correlation, and the level of confidence in the node pair relationship based on evidence from similarity scores, literature and graph databases (Guo et al., 2022). Also, incorporating autoencoders, a deep learning approach, and its denoising and variational variants autoencoders (e.g., sparse autoencoders) have been used to address this issue in graph neural networks (Ng, 2011). Autoencoders learn a representation of the data from the input layer, enforce sparsity constraints and try to reproduce it at the output layer. During this process, the model can learn from incomplete data and generate new plausible values for imputation (Pereira et al., 2020).

Future directions

An area of prospect for integrative multi-omics network-based research, which remains an important opportunity, is making efforts to limit the challenges linked with network-based multi-omics integration in the context of heterogeneity, reproducibility, sparsity, and interpretation of results as discussed above. Another area of importance is building hybrid integrative models which are capable of handling paired and non-paired omics data, as well as other biomedical data. Furthermore, efforts to develop a framework tool or metadata schema that standardizes or harmonizes various multi-omics approaches for data integration could be useful. For example, such a framework may leverage an optimized approach to weigh and prioritize genes, pathways, biological processes, drug targets, and relationships between various other biological

features from the multi-omics datasets. However, such framework tools will also require the expertise of domain experts, as well as the detailed and uniform characterization of statistical and technical attributes of the data (Krassowski et al., 2020).

Discussion

Network-based integrative multi-omics analysis offers the opportunity to elucidate interactions that can occur among all classes of molecules in a biological system as well as information flow between and within multiple omics levels. In addition, it potentially provides substantial improvement of biological understanding by helping in the interpretation of results, as compared to single omics analysis, although collecting multi-omics data from different sources does not guarantee that it will be possible to learn about (all of) the relationships present.

Various graph-based multi-omics methods have been developed for network analysis; however, their application is dependent on the scope of the research question of interest and the (omics) data types available. Consequently, this will inform the choice of an integrative analytical approach and tools. The network-based methods discussed use different scoring metrics, algorithms, and data types which together translate into a comprehensive data source/graph to be employed for interpretation into biological knowledge. The overviewand description of the tools for network-based integrative analysis (Table 1) show that different approaches can be implemented in different ways to achieve similar results. Additionally, the classification of tools (Figure 4) highlights that some tools can be applied to more than one research question. However, due to the difference in approaches of these methods, we recommend the use of multiple analytical and methodological approaches during integrative data analysis, to compare and validate the study results in different ways before interpretation for further downstream tests or follow-up studies.

Author contributions

FA conceived the study and prepared the first draft. JB, AN, KN, MS, GM, NM, EC, TE, and PH contributed to the revision of the article. GM, EC, TE, and PH supervised the work. All authors contributed to the article and approved the submitted version.

Funding

This work was partially funded by an LSH HealthHolland grant to the TWOC consortium, a large-scale infrastructure grant from the Dutch Organization of Scientific Research (NWO) to the Netherlands X-omics initiative (184.034.019), and a Horizon2020 research grant from the European Union to the EATRIS-Plus infrastructure project (grant agreement: No 871096).

Acknowledgments

We acknowledge members of the Trusted World of Corona (TWOC) Consortium. We also acknowledge the staff and colleagues from the Division of Human Genetics and Division of Computational Biology, University of Cape Town, and colleagues from the Center for Molecular and Biomolecular Informatics (CMBI), Radboud University Medical Center, Nijmegen. In memorial of GM who passed away due to COVID-19 complications before the submission of this manuscript. His contribution as supervisor of FA, supervising this work, writing, and revising multi sections of this manuscript will never be forgotten.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Agamah, F. E., Damena, D., Skelton, M., Ghansah, A., Mazandu, G. K., and Chimusa, E. R. (2021). Network-driven analysis of human–plasmodium falciparum interactome: Processes for malaria drug discovery and extracting *in silico* targets. *Malar. J.* 20 (1), 421. doi:10.1186/s12936-021-03955-0

Badsha, M., and Fu, A. Q. (2019). Learning causal biological networks with the principle of Mendelian randomization. *Front. Genet.* 10, 460. doi:10.3389/fgene.2019.00460

Badsha, M. B., Martin, E. A., and Fu, A. Q. (2021). Mrpc: An R package for inference of causal graphs. *Front. Genet.* 12, 460. doi:10.3389/fgene.2019.00460

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinforma.* 17 (2), 15–77. doi:10.1186/s12859-015-0857-9

Birnhuber, A., Fliesser, E., Gorkiewicz, G., Zacharias, M., Seeliger, B., David, S., et al. (2021). Between inflammation and thrombosis: Endothelial cells in COVID-19. *Eur. Respir. J.* 58 (3), 2100377. doi:10.1183/13993003.00377-2021

Bodein, A., Chapleur, O., Droit, A., and Lê Cao, K-A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Front. Genet.* 10, 963. doi:10.3389/fgene.2019.00963

Bodein, A., Scott-Boyer, M. P., Perin, O., Le Cao, K-A., and Droit, A. (2020). Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic acids Res.* 50, e27. doi:10.1093/nar/gkab1200

Bonnet, E., Calzone, L., and Michoel, T. (2015). Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput. Biol.* 11 (2), e1003983. doi:10.1371/journal.pcbi.1003983

Buescher, J. M., and Driggers, E. M. (2016). Integration of omics: More than the sum of its parts. *Cancer Metab.* 4 (1), 4–8. doi:10.1186/s40170-016-0143-y

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell.* 173 (7), 1581–1592. doi:10.1016/j.cell.2018.05.015

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U. E., Seitz, H., et al. (2020). Prospects and challenges of multi-omics data integration in toxicology. *Arch. Toxicol.* 94 (2), 371–388. doi:10.1007/s00204-020-02656-y

Cavill, R., Jennen, D., Kleinjans, J., and Briedé, J. J. (2016). Transcriptomic and metabolomic data integration. *Brief. Bioinform.* 17 (5), 891–901. doi:10.1093/bib/bbv090

Chakravorty, D., Banerjee, K., and Saha, S. (2018). *Integrative omics for interactomes. Synthetic Biology*. Berlin, Germany: Springer, 39–49.

Chen, J., and Wong, K-C. R. N. C. E. (2021). RNCE: Network integration with reciprocal neighbors contextual encoding for multi-modal drug community study on cancer targets. *Brief. Bioinform.* 22 (3), bbaa118. doi:10.1093/bib/bbaa118

Chen, Y. X., Chen, H., Rong, Y., Jiang, F., Chen, J. B., Duan, Y. Y., et al. (2020). An integrative multi-omics network-based approach identifies key regulators for breast cancer. *Comput. Struct. Biotechnol. J.* 18, 2826–2835. doi:10.1016/j.csbj.2020.10.001

Chierici, M., Bussola, N., Marcolini, A., Francescatto, M., Zandonà, A., Trastulla, L., et al. (2020). Integrative network fusion: A multi-omics approach in molecular profiling. *Front. Oncol.* 10, 1065. doi:10.3389/fonc.2020.01065

Class, C. A., Ha, M. J., Baladandayuthapani, V., and Do, K-A. (2018). iDINGO—integrative differential network analysis in genomics with Shiny application. *Bioinformatics* 34 (7), 1243–1245. doi:10.1093/bioinformatics/btx750

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: A universal amplifier of genetic associations. *Nat. Rev. Genet.* 18 (9), 551–562. doi:10.1038/nrg.2017.38

Davies, V., Wandy, J., Weidt, S., Van Der Hooft, J. J., Miller, A., Daly, R., et al. (2021). Rapid development of improved data-dependent acquisition strategies. *Anal. Chem.* 93 (14), 5676–5683. doi:10.1021/acs.analchem.0c03895

Demichev, V., Tober-Lau, P., Lemke, O., Nazarenko, T., Thibeault, C., Whitwell, H., et al. (2021). A time-resolved proteomic and prognostic map of COVID-19. *Cell. Syst.* 12 (8), 780–794.e7. doi:10.1016/j.cels.2021.05.005

Di Nanni, N., Bersanelli, M., Milanesi, L., and Mosca, E. (2020). Network diffusion promotes the integrative analysis of multiple omics. *Front. Genet.* 11, 106. doi:10.3389/fgene.2020.00106

Dimitrakopoulos, C., Hindupur, S. K., Häfliger, L., Behr, J., Montazeri, H., Hall, M. N., et al. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 34 (14), 2441–2448. doi:10.1093/bioinformatics/bty148

Duruflé, H., Selmani, M., Ranocha, P., Jamet, E., Dunand, C., and Déjean, S. (2021). A powerful framework for an integrative study with heterogeneous omics data: From univariate statistics to multi-block analysis. *Brief. Bioinform.* 22 (3), bbaa166. doi:10.1093/bib/bbaa166

Essa, M. M., Hamdan, H., Chidambaram, S. B., Al-Balushi, B., Guillemin, G. J., Ojcius, D. M., et al. (2020). *Possible role of tryptophan and melatonin in COVID-19*. London, England: SAGE Publications Sage UK, 1178646920951832.

Fan, Z., Zhou, Y., and Ressom, H. W. (2020). MOTA: Network-based multi-omic data integration for biomarker discovery. *Metabolites* 10 (4), 144. doi:10.3390/metabo10040144

Folch-Fortuny, A., Villaverde, A. F., Ferrer, A., and Banga, J. R. (2015). Enabling network inference methods to handle missing data and outliers. *BMC Bioinforma.* 16 (1), 283. doi:10.1186/s12859-015-0717-7

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* 303 (5659), 799–805. doi:10.1126/science.1094068

González, I., Lê Cao, K-A., Davis, M. J., and Déjean, S. (2012). Visualising associations between paired 'omics' data sets. *BioData Min.* 5 (1), 19–23. doi:10.1186/1756-0381-5-19

Greenland, S., Mansournia, M. A., and Altman, D. G. (2016). Sparse data bias: A problem hiding in plain sight. *BMJ* 352, i1981. doi:10.1136/bmj.i1981

Griffin, P. J., Zhang, Y., Johnson, W. E., and Kolaczyk, E. D. (2018). Detection of multiple perturbations in multi-omics biological networks. *Biometrics* 74 (4), 1351–1361. doi:10.1111/biom.12893

Guo, J., and Huan, T. (2020). Comparison of full-scan, data-dependent, and data-independent acquisition modes in liquid chromatography–mass spectrometry based untargeted metabolomics. *Anal. Chem.* 92 (12), 8072–8080. doi:10.1021/acs.analchem.9b05135

Guo, M. G., Sosa, D. N., and Altman, R. B. (2022). Challenges and opportunities in network-based solutions for biological questions. *Brief. Bioinform.* 23 (1), bbab437. doi:10.1093/bib/bbab437

Ha, M. J., Baladandayuthapani, V., and Do, K. A. (2015). DINGO: Differential network analysis in genomics. *Bioinformatics* 31 (21), 3413–3420. doi:10.1093/bioinformatics/btv406

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18 (1), 83–15. doi:10.1186/s13059-017-1215-1

Hawe, J. S., Theis, F. J., and Heinig, M. (2019). Inferring interaction networks from multi-omics data. *Front. Genet.* 10, 535. doi:10.3389/fgene.2019.00535

Holzinger, E. R., and Ritchie, M. D. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics* 13 (2), 213–222. doi:10.2217/pgs.11.145

Horgan, R. P., and Kenny, L. C. (2011). 'Omic' technologies: Genomics, transcriptomics, proteomics and metabolomics. *Obstetrician Gynaecol.* 13 (3), 189–195. doi:10.1576/toag.13.3.189.27672

Huang, L., Brunell, D., Stephan, C., Mancuso, J., Yu, X., He, B., et al. (2019). Driver network as a biomarker: Systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. *Bioinformatics* 35 (19), 3709–3717. doi:10.1093/bioinformatics/btz109

Joshi, P., Jeong, S., and Park, T. (2020). Sparse superlayered neural network-based multi-omics cancer subtype classification. *Int. J. Data Min. Bioinform.* 24 (1), 58–73. doi:10.1504/ijdmb.2020.109500

Jung, G. T., Kim, K-P., and Kim, K. (2020). How to interpret and integrate multi-omics data at systems level. *Anim. Cells Syst.* 24 (1), 1–7. doi:10.1080/19768354.2020.1721321

Kang, M., Ko, E., and Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform.* 23 (1), bbab454. doi:10.1093/bib/bbab454

Karasuyama, M., and Mamitsuka, H. (2017). Adaptive edge weighting for graph-based learning algorithms. *Mach. Learn.* 106 (2), 307–335. doi:10.1007/s10994-016-5607-3

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19 (5), 299–310. doi:10.1038/nrg.2018.4

Koh, H. W., Fermin, D., Vogel, C., Choi, K. P., Ewing, R. M., and Choi, H. (2019). iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. *NPJ Syst. Biol. Appl.* 5 (1), 22–10. doi:10.1038/s41540-019-0099-y

Koller, D., and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA, USA: MIT press.

Kotiang, S., and Eslami, A. (2020). A probabilistic graphical model for system-wide analysis of gene regulatory networks. *Bioinformatics* 36 (10), 3192–3199. doi:10.1093/bioinformatics/btaa122

Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the field in multi-omics research: From computational needs to data mining and sharing. *Front. Genet.* 11, 610798. doi:10.3389/fgene.2020.610798

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2019). *Gradient-based neural dag learning*. arXiv preprint arXiv:02226.

Lee, D., Park, Y., and Kim, S. (2021). Towards multi-omics characterization of tumor heterogeneity: A comprehensive review of statistical and machine learning approaches. *Brief. Bioinform.* 22 (3), bbaa188. doi:10.1093/bib/bbaa188

Leiserson, M. D., Vandin, F., Wu, H-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47 (2), 106–114. doi:10.1038/ng.3168

Levi, H., Elkon, R., and Shamir, R. (2021). DOMINO: A network-based active module identification algorithm with reduced rate of false calls. *Mol. Syst. Biol.* 17 (1), e9593. doi:10.15252/msb.20209593

Li, H., Xiang, X., Ren, H., Xu, L., Zhao, L., Chen, X., et al. (2020). Serum Amyloid A is a biomarker of severe Coronavirus Disease and poor prognosis. *J. Infect.* 80 (6), 646–655. doi:10.1016/j.jinf.2020.03.035

Li, W., (2015). "Estimating jaccard index with missing observations: A matrix calibration approach," in Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2.

Lima, E., Davies, P., Kaler, J., Lovatt, F., and Green, M. (2020). Variable selection for inferential models with relatively high-dimensional data: Between method heterogeneity and covariate stability as adjuncts to robust selection. *Sci. Rep.* 10 (1), 8002–8011. doi:10.1038/s41598-020-64829-0

Luo, Y., Peng, J., and Ma, J. (2020). When causal inference meets deep learning. *Nat. Mach. Intell.* 2 (8), 426–427. doi:10.1038/s42256-020-0218-x

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8 (1), 573. doi:10.1038/s41467-017-00680-8

Manatakis, D. V., Raghu, V. K., and Benos, P. V. (2018). piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks. *Bioinformatics* 34 (17), i848–i856. doi:10.1093/bioinformatics/bty591

Marín-Llaó, J., Mubeen, S., Perera-Lluna, A., Hofmann-Apitius, M., Picart-Armada, S., and Domingo-Fernández, D. (2020). MultiPaths: A Python framework for analyzing multi-layer biological networks using diffusion algorithms. *Bioinformatics* 37, 137. doi:10.1093/bioinformatics/btaa1069

Martorell-Marugán, J., Tabik, S., Benhammou, Y., del Val, C., Zwir, I., Herrera, F., et al. (2019). *Deep learning in omics data analysis and precision medicine*. Brisbane City, Australia: Exon Publications, 37–53.

Mo, X., Jian, W., Su, Z., Chen, M., Peng, H., Peng, P., et al. (2020). Abnormal pulmonary function in COVID-19 patients at time of hospital discharge. *Eur. Respir. J.* 55 (6), 2001217. doi:10.1183/13993003.01217-2020

Ng, A. (2011). Sparse autoencoder. *CS294A Lect. notes* 72 (2011), 1–19.

Nguyen, N. D., and Wang, D. (2020). Multiview learning for understanding functional multiomics. *PLoS Comput. Biol.* 16 (4), e1007677. doi:10.1371/journal.pcbi.1007677

Noecker, C., Eng, A., Muller, E., and Borenstein, E. (2022). MIMOSA2: A metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data. *Bioinformatics* 38 (6), 1615–1623. doi:10.1093/bioinformatics/btac003

Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., et al. (2020). Large-scale multi-omic analysis of COVID-19 severity. *Cell. Syst.* 12, 23–40.e7. doi:10.1016/j.cels.2020.10.003

Pan, X., Burgman, B., Wu, E., Huang, J. H., Sahni, N., and Yi, S. S. (2022). i-Modern: Integrated multi-omics network model identifies potential therapeutic targets in glioma by deep learning with interpretability. *Comput. Struct. Biotechnol. J.* 20, 3511–3521. doi:10.1016/j.csbj.2022.06.058

Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion through Interacting Events (TieDIE). *Bioinformatics* 29 (21), 2757–2764. doi:10.1093/bioinformatics/btt471

Pereira, R. C., Santos, M. S., Rodrigues, P. P., and Abreu, P. H. (2020). Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. *J. Artif. Intell. Res.* 69, 1255–1285. doi:10.1613/jair.1.12312

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. Cambridge, MA, USA: The MIT Press.

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/j.biotechadv.2021.107739

Reyna, M. A., Leiserson, M. D., and Raphael, B. J. (2018). Hierarchical HotNet: Identifying hierarchies of altered subnetworks. *Bioinformatics* 34 (17), i972–i980. doi:10.1093/bioinformatics/bty613

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* 16 (2), 85–97. doi:10.1038/nrg3868

Sargazi, S., Sheervalilou, R., Rokni, M., Shirvaliloo, M., Shahraki, O., and Rezaei, N. (2021). The role of autophagy in controlling SARS-CoV-2 infection: An overview on virophagy-mediated molecular drug targets. *Cell. Biol. Int.* 45 (8), 1599–1612. doi:10.1002/cbin.11609

Sedgewick, A. J., Buschur, K., Shi, I., Ramsey, J. D., Raghu, V. K., Manatakis, D. V., et al. (2019). Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics* 35 (7), 1204–1212. doi:10.1093/bioinformatics/bty769

Seifert, M., and Beyer, A. (2018). regNet: An R package for network-based propagation of gene expression alterations. *Bioinformatics* 34 (2), 308–311. doi:10.1093/bioinformatics/btx544

Shafi, A., Nguyen, T., Peyvandipour, A., Nguyen, H., and Draghici, S. (2019). A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures. *Front. Genet.* 10, 159. doi:10.3389/fgene.2019.00159

Shneider, A., Kudriavtsev, A., and Vakhrusheva, A. (2020). Can melatonin reduce the severity of COVID-19 pandemic? *Int. Rev. Immunol.* 39 (4), 153–162. doi:10.1080/08830185.2020.1756284

Sitaram, D., Dalwani, A., Narang, A., Das, M., and Auradkar, P. (2015). "A measure of similarity of time series containing missing data using the mahalanobis distance," in 2015 second international conference on advances in computing and communication engineering, 01-02 May 2015 (Dehradun, India: IEEE). doi:10.1109/ICACCE.2015.14

Smilowitz, N. R., Kunichoff, D., Garshick, M., Shah, B., Pillinger, M., Hochman, J. S., et al. (2021). C-reactive protein and clinical outcomes in patients with COVID-19. *Eur. Heart J.* 42 (23), 2270–2279. doi:10.1093/eurheartj/ehaa1103

Song, W. M., and Zhang, B. (2015). Multiscale embedded gene co-expression network analysis. *PLoS Comput. Biol.* 11 (11), e1004574. doi:10.1371/journal.pcbi.1004574

Stukalov, A., Girault, V., Grass, V., Karayel, O., Bergant, V., Urban, C., et al. (2021). Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* 594 (7862), 246–252. doi:10.1038/s41586-021-03493-4

Su, Y., Chen, D., Yuan, D., Lausted, C., Choi, J., Dai, C. L., et al. (2020). Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell.* 183, 1479–1495. doi:10.1016/j.cell.2020.10.037

Sun, C., Sun, Y., Wu, P., Ding, W., Wang, S., Li, J., et al. (2021). Longitudinal multi-omics transition associated with fatality in critically ill COVID-19 patients. *Intensive Care Med. Exp.* 9 (1), 13–14. doi:10.1186/s40635-021-00373-z

Sun, F., Tan, H., Li, Y., De Boevre, M., Zhang, H., Zhou, J., et al. (2021). An integrated data-dependent and data-independent acquisition method for hazardous compounds screening in foods using a single UHPLC-Q-Orbitrap run. *J. Hazard. Mat.* 401, 123266. doi:10.1016/j.jhazmat.2020.123266

Sun, Y. V., and Hu, Y-J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* 93, 147–190. doi:10.1016/bs.adgen.2015.11.004

Takeshita, H., and Yamamoto, K. (2022). Tryptophan metabolism and COVID-19-induced skeletal muscle damage: Is ACE2 a key regulator? *Front. Nutr.* 9, 868845. doi:10.3389/fnut.2022.868845

Tomazou, M., Bourdakou, M. M., Minadakis, G., Zachariou, M., Oulas, A., Karatzas, E., et al. (2021). Multi-omics data integration and network-based analysis drives a multiplex drug repurposing approach to a shortlist of candidate drugs against COVID-19. *Briefings Bioinforma.* 22, bbab114. doi:10.1093/bib/bbab114

Tuncbag, N., Gosline, S. J., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E. (2016). Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLoS Comput. Biol.* 12 (4), e1004879. doi:10.1371/journal.pcbi.1004879

Ulfenborg, B. (2019). Vertical and horizontal integration of multi-omics data with miodin. *BMC Bioinforma.* 20 (1), 649. doi:10.1186/s12859-019-3224-4

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., et al. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 35 (3), 497–505. doi:10.1093/bioinformatics/bty637

Vandin, F., Clay, P., Upfal, E., and Raphael, B. J. (2012). Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp. Biocomput* 2012, 55–66.

Vitali, F., Cohen, L. D., Demartini, A., Amato, A., Eterno, V., Zambelli, A., et al. (2016). A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer. *PloS one* 11 (9), e0162407. doi:10.1371/journal.pone.0162407

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11 (3), 333–337. doi:10.1038/nmeth.2810

Wang, W., Yang, S., Zhang, X., and Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 30 (20), 2923–2930. doi:10.1093/bioinformatics/btu403

Wang, Y., Solus, L., Yang, K. D., and Uhler, C. (2017). Permutation-based causal inference algorithms with interventions. *Adv. Neural Inf. Process. Syst.*

Wein, S., Malloni, W. M., Tomé, A. M., Frank, S. M., Henze, G-I., Wüst, S., et al. (2021). A graph neural network framework for causal inference in brain networks. *Sci. Rep.* 11 (1), 8061. doi:10.1038/s41598-021-87411-8

Wen, Y., Song, X., Yan, B., Yang, X., Wu, L., Leng, D., et al. (2021). Multi-dimensional data integration algorithm based on random walk with restart. *BMC Bioinforma.* 22 (1), 97–22. doi:10.1186/s12859-021-04029-3

Wu, Z., Li, W., Liu, G., and Tang, Y. (2018). Network-based methods for prediction of drug-target interactions. *Front. Pharmacol.* 9, 1134. doi:10.3389/fphar.2018.01134

Xu, T., Le, T. D., Liu, L., Wang, R., Sun, B., and Li, J. (2016). Identifying cancer subtypes from mirna-tf-mrna regulatory networks and expression data. *PloS one* 11 (4), e0152792. doi:10.1371/journal.pone.0152792

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Brief. Bioinform.* 19 (6), 1370–1381. doi:10.1093/bib/bbx066

Yan, W., Xue, W., Chen, J., and Hu, G. (2016). Biological networks for cancer candidate biomarkers discovery. *Cancer Inf.* 15, S39458. doi:10.4137/CIN.S39458

Yang, Y., Tian, S., Qiu, Y., Zhao, P., and Zou, Q. (2022). MDICC: Novel method for multi-omics data integration and cancer subtype identification. *Brief. Bioinform.* 23 (3), bbac132. doi:10.1093/bib/bbac132

Zachariou, M., Minadakis, G., Oulas, A., Afxenti, S., and Spyrou, G. M. (2018). Integrating multi-source information on a single network to detect disease-related clusters of molecular mechanisms. *J. Proteomics* 188, 15–29. doi:10.1016/j.jprot.2018.03.009

Zapalska-Sozoniuk, M., Chrobak, L., Kowalczyk, K., and Kankofer, M. (2019). Is it useful to use several "omics" for obtaining valuable results? *Mol. Biol. Rep.* 46 (3), 3597–3606. doi:10.1007/s11033-019-04793-9

Zeng, I. S. L., and Lumley, T. (2018). Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinform. Biol. Insights* 12, 1177932218759292. doi:10.1177/1177932218759292

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *Adv. Neural Inf. Process. Syst.*

Zhou, G., Pang, Z., Lu, Y., Ewald, J., and Xia, J. (2022). OmicsNet 2.0: A web-based platform for multi-omics integration and network visual analytics. *Nucleic Acids Res.* 50, W527–W533. doi:10.1093/nar/gkac376

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf. Fusion* 50, 71–91. doi:10.1016/j.inffus.2018.09.012

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 20 February 2023
DOI 10.3389/fmolb.2023.1000248

OPEN ACCESS

EDITED BY
Ornella Cominetti,
Nestlé Research Center, Switzerland

REVIEWED BY
Elzbieta Radzikowska,
National Institute of Tuberculosis and Lung Diseases (Poland), Poland
Laura Schmidt,
National Cancer Institute at Frederick (NIH), United States

*CORRESPONDENCE
Wei Sun,
sunwei1018@sina.com
Lulu Jia,
jluyu@126.com
Yushi Zhang,
beijingzhangyushi@126.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Metabolomics, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 21 July 2022
ACCEPTED 10 February 2023
PUBLISHED 20 February 2023

CITATION
Wang Z, Liu X, Wang W, Xu J, Sun H, Wei J, Yu Y, Zhao Y, Wang X, Liao Z, Sun W, Jia L and Zhang Y (2023), UPLC-MS based integrated plasma proteomic and metabolomic profiling of TSC-RAML and its relationship with everolimus treatment. *Front. Mol. Biosci.* 10:1000248. doi: 10.3389/fmolb.2023.1000248

UPLC-MS based integrated plasma proteomic and metabolomic profiling of TSC-RAML and its relationship with everolimus treatment

Zhan Wang[1†], Xiaoyan Liu[2†], Wenda Wang[1†], Jiyu Xu[2], Haidan Sun[2], Jing Wei[3], Yuncui Yu[3], Yang Zhao[1], Xu Wang[1], Zhangcheng Liao[1], Wei Sun[2]*, Lulu Jia[3]* and Yushi Zhang[1]*

[1]Department of Urology, Peking Union Medical College Hospital, Chinese Academy of Medical Science and Peking Union Medical College, Beijing, China, [2]School of Basic Medical College, Core facility of instrument, Institution of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing, China, [3]Clinical Research Center, National Center for Children's Health, Beijing Children's Hospital, Capital Medical University, Beijing, China

Aim: To profile the plasma proteomics and metabolomics of patients with renal cysts, sporadic angiomyolipoma (S-AML) and tuberous sclerosis complex related angiomyolipoma (TSC-RAML) before and after everolimus treatment, and to find potential diagnostic and prognostic biomarkers as well as reveal the underlying mechanism of TSC tumorigenesis.

Materials and Methods: We retrospectively measured the plasma proteins and metabolites from November 2016 to November 2017 in a cohort of pre-treatment and post-treatment TSC-RAML patients and compared them with renal cyst and S-AML patients by ultra-performance liquid chromatography-mass spectrometer (UPLC-MS). The tumor reduction rates of TSC-RAML were assessed and correlated with the plasma protein and metabolite levels. In addition, functional analysis based on differentially expressed molecules was performed to reveal the underlying mechanisms.

Results: Eighty-five patients with one hundred and ten plasma samples were enrolled in our study. Multiple proteins and metabolites, such as pre-melanosome protein (PMEL) and S-adenosylmethionine (SAM), demonstrated both diagnostic and prognostic effects. Functional analysis revealed many dysregulated pathways, including angiogenesis synthesis, smooth muscle proliferation and migration, amino acid metabolism and glycerophospholipid metabolism.

Conclusion: The plasma proteomics and metabolomics pattern of TSC-RAML was clearly different from that of other renal tumors, and the differentially expressed

Abbreviations: AML, angiomyolipoma; CY, renal cyst; DE, differentially expressed; GO, gene oncology; GSEA, gene set enrichment analysis; LAM, lymphangioleiomyomatosis; mTOR, mammalian target of rapamycin; PMEL, pre-melanosome protein; QC, Quality control; RA, rhabdomyosarcoma; SAM, S-Adenosylmethionine; S-AML, sporadic angiomyolipoma; SEGA, subependymal giant cell astrocytoma; TSC, Tuberous sclerosis complex; TSC-RAML, TSC related AML; t-SNE, t-Distributed Stochastic Neighbor Embedding; UPLC-MS, ultra-performance liquid chromatography-mass spectrometer; WGCNA, weighted gene correlation network analysis; WM-CC, whole metabolome co-expression clusters; WP-CC, whole proteome co-expression clusters.

plasma molecules could be used as prognostic and diagnostic biomarkers. The dysregulated pathways, such as angiogenesis and amino acid metabolism, may shed new light on the treatment of TSC-RAML.

KEYWORDS

UPLC-MS, proteomics, metabolomics, tuberous sclerosis complex, everolimus

Background

Tuberous sclerosis complex (TSC) is a rare disease caused by germline mutations of tumor suppressor genes in either the *TSC1* gene on chromosome 9 or the *TSC2* gene on chromosome 16 (Henske et al., 2016). Its incidence is approximately 1 in 6,000–10,000, and there are around 2 million patients worldwide, although the rate may be greatly underestimated due to large numbers of undiagnosed patients (Lam et al., 2018). TSC threatens multiple organs throughout the body and causes corresponding distinctive manifestations, including subependymal giant cell astrocytoma (SEGA) in the brain, rhabdomyosarcoma (RA) in the heart, lymphangioleiomyomatosis (LAM) in the lung, angiomyolipoma (AML) in the kidney and so on.

For the underlying mechanism, the most widely acknowledged theory is that silencing of the TSC complex caused by mutations could lead to overactivation of the mammalian target of rapamycin (mTOR) signaling pathway, which has been proven to be critical in various physiological processes, such as regulating cell growth, metabolism and autophagy (Yang et al., 2013; Ranek et al., 2019). Aberrant constitutive mTOR pathway activation could result in unregulated cell proliferation, migration, and invasion and finally cause hamartoma in different organs (Bottolo et al., 2020). Based on the above mechanism, mTOR inhibitors, including rapamycin and everolimus, have been developed to control the various manifestations, including renal AML (Bissler et al., 2013; Cai et al., 2018), brain SEGA (Krueger et al., 2010; Franz et al., 2013) and pulmonary LAM (McCormack et al., 2011).

As the most common cause of early death among patients with TSC (Shepherd et al., 1991; Amin et al., 2017), the renal lesions have three main forms, namely, AML (the most common, making up more than 80%), renal cysts and renal cell carcinoma. The abrupt rupture of TSC related AML (TSC-RAML) is a common cause of mortality and is sometimes referred to as a "ticking bomb" within the body. Exist-2 is thus far the largest multi-center randomized controlled trial assessing the effect of everolimus on TSC-RAML (Bissler et al., 2013). This trial validated its efficacy with a 42% response rate and an acceptable safety profile, making everolimus the only drug approved by the Food and Drug Administration of America to treat TSC-RAML. Our center has also conducted a 2-year, nonrandomized, open-label, phase 2 clinical trial, and the result showed that 50% volume reduction rate reached 52.94% at 3 months and 58.82% at 6 months, further confirming the favorable effect of mTOR inhibitors on TSC-RAML (Cai et al., 2018).

In the EXIST-2 trial, plasma VEGF-D and collagen IV levels were found to be potential prognostic as well as diagnostic biomarkers, and these results have been validated by subsequent studies not only in TSC-RAML (Dabora et al., 2011; Malinowska et al., 2013) but also in the TSC-LAM (Young et al., 2010; Xu et al., 2013; Amaral et al., 2019). So far, very few efficient biomarkers have been discovered to guide clinical treatment or follow-up of patients with TSC.

mTOR is an atypical serine/threonine protein kinase that forms two distinct signaling complexes, mTORC1 and mTORC2, which are distinguished primarily by their association with Raptor or Rictor, respectively (Martin et al., 2014). Through direct phosphorylation and activation of S6 kinase 1 (S6K) and inactivation of 4E-BP1, mTORC1 regulates many cellular metabolisms, such as amino acid, glucose, nucleotide, fatty acid and lipid metabolism (Morita et al., 2015; Mossmann et al., 2018). During this process, massive proteomic and metabolomic hallmarks will be produced if mTOR continuously activated. As one of the most commonly used high-throughput approaches to detect proteome and metabolome in biofluids, ultra-performance liquid chromatography-mass spectrometer (UPLC-MS) has been widely applied in searching for candidate diagnostic and prognostic biomarkers and potential drug targets (Wang et al., 2019; Blomme et al., 2020; Sovio et al., 2020; Behsaz et al., 2021; de la Calle Arregui et al., 2021; Wang C. Y. et al., 2021; Wang Z. et al., 2021).

Therefore, the aim of our study was to retrospectively analyze the plasma proteomic and metabolomic profiles with UPLC-MS and to search for diagnostic and prognostic markers of TSC-RAML to guide clinical management.

Materials and methods

Human samples and clinical data

This study was conducted at Peking Union Medical College Hospital from November 2016 to November 2017 and was approved by the Institutional Review Board of Peking Union Medical College Hospital and the Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences (Approval number: KS2020127). This research was carried out according to the Code of Ethics of the World Medical Association (Declaration of Helsinki), and formally written consent documents were provided by every participant before been enrolled in this study.

The inclusion criteria were as follows: 1) The TSC patients met the clinical or genetic diagnosis of TSC according to the International Tuberous Sclerosis Complex Consensus Conference in 2012 and took oral everolimus at a dose of 10 mg/qd for at least 6 months. 2) The S-AML patients received partial nephrectomy and the diagnosis was confirmed as AML. 3) All plasma samples from the TSC patients were collected pre-treatment and 3 or 6 months after initiating everolimus treatment. 4) Patients with renal cysts were considered healthy controls, and blood samples were collected preoperatively during the same period.

The exclusion criteria were: 1) Those who had no pre-operative or pre-treatment plasma in our sample bank. 2) Patients who had malignant tumors or metabolomic diseases such as diabetes and hyperlipidemia.

All the enrolled TSC patients were assessed independently by two radiologists to determine the tumor volume at baseline, every 3 months within the first year, every 6 months within the second year and yearly thereafter. The maximum AML volume was used to calculate the tumor response, and >50% volume reduction was regarded as effective. All TSC patients received next-generation gene sequencing (NGS) to assess their *TSC* gene mutations. The exact process of NGS were described in our previously published article (Wang et al., 2022a).

Sample collection

Whole blood samples were collected in the morning at 07: 00 am–09:00 am with at least 10 h of fasting to eliminate the impact of diet. The 4 mL EDTA tubes with whole blood were transferred and separated by density gradient centrifugation within 1 h after collection. The plasma was stored at −80°C until conducting the formal experiment.

Sample preparation for proteomics

To remove the highly abundant proteins (including albumin, IgA and IgD) from the plasma, High Select™ Top14 Abundant Protein Depletion Mini Spin Columns (Thermo Fisher Scientific, MA. United States) were applied according to the manufacturer's instructions (as attached in Supplementary Material S1). After this procedure, we obtained 300 μL samples with the highly abundant proteins removed. Ten microliters of each sample were removed to measure the protein concentration by the BCA assay (Pierce).

Every 100 mg of protein was reduced with 20 mM dithiothreitol (DTT) for 5 min at 95°C and subsequently alkylated with 50 mM iodoacetamide for 45 min at room temperature in the dark. Protein digestion was carried out using the filter-aided sample preparation technique (FASP). Proteins were loaded onto 30 kDa filter devices (Pall, Port Washington, NY, United States). Trypsin (Trypsin Gold, mass spec grade, Promega, WI, United States) was added (enzyme to protein ratio of 1:50), and the samples were incubated at 37°C overnight. The samples were centrifuged at ×14,000 g, and approximately 30 μL of the liquid was used for analysis.

For the quality control (QC) samples, 3 μL was taken from 23 randomly selected representative samples and mixed with the testing samples together, and the mixture was loaded with the testing samples. The QC samples were injected every 10 samples. All the samples were loaded on the autosamplers with a mixture of iRT.

ESI-LC-MS/MS for proteome library generation

The pooled peptide samples of each group were separated by high-pH RPLC columns (4.6 mm × 250 mm, C18, 3 μm; Waters, United States). Each pooled sample was loaded onto the column in buffer A1 (H2O, pH 10). The elution gradient was 5%–30% buffer B1 (90% ACN, pH 10; flow rate, 1 mL/min) for 30 min. The eluted peptides were collected at one fraction per minute. After lyophilization, the 30 fractions were resuspended in 0.1% formic acid and then concatenated into 10 fractions by combining fractions 1, 11, 21, and so on. To generate the spectral library, the fractions from RPLC were analyzed in DDA mode. The parameters were set as follows: the MS was recorded at 350–1,500 m/z at a resolution of 60,000 m/z; the maximum injection time was 50 ms, the auto gain control (AGC) was 1e6, and the cycle time was 3 s. MS/MS scans were performed at a resolution of 15,000 with an isolation window of 1.6 Da and a collision energy at 32% (HCD); the AGC target was 50,000, and the maximum injection time was 30 ms.

ESI-LC-MS/MS for proteome data-independent acquisition analysis

The digested peptides were dissolved in 0.1% formic acid and separated on an RP C18 self-packing capillary LC column (75 μm × 150 mm, 3 μm). The elution gradient was 5%–30% buffer B2 (0.1% formic acid, 99.9% ACN; flow rate, 0.3 μL/min) for 60 min. For MS acquisition, the variable isolation window DIA method with 38 windows was developed. The specific window lists were constructed based on the DDA experiment of the pooled sample. The full scan was set at a resolution of 120,000 over the m/z range of 400 to 900, followed by DIA scans with a resolution of 30,000; the HCD collision energy was 32%, the AGC target was 1E6, and the maximal injection time was 50 ms.

Spectral library generation

To generate a comprehensive spectral library, the pooled sample from each group was processed. The DDA data were processed using Proteome Discoverer (Thermo Scientific, Germany) software and searched against the human SwissProt database appended with the iRT fusion protein sequence (Biognosys). A maximum of two missed cleavages for trypsin was used, cysteine carbamidomethylation was set as a fixed modification, and methionine oxidation deamination and +43 on Kn (carbamyl) were used as variable modifications. The parent and fragment ion mass tolerances were set to 10 ppm and 0.02 Da, respectively. The applied false discovery rate (FDR) cutoff was 0.01 at the protein level. The results were then imported into Spectronaut Pulsar (Biognosys, Switzerland) software to generate the library. Additionally, DIA data were imported into Spectronaut Pulsar software and searched against the human SwissProt database to generate the DIA library. The final library was generated by combining the DDA and DIA libraries of all the enrolled samples.

Data analysis

The DIA-MS data were analyzed using the Spectronaut Pulsar (Biognosys, Switzerland) with the default settings. All of the results were filtered with a Q-value cutoff of 0.01 (corresponding to an FDR of 1%). Proteins identified in more than 50% of the samples in at

least one subgroup were retained for further analysis. Missing values were imputed based on the k-nearest neighbor method or by the minimum value (details provided in Supplementary Figure S1A).

Raw proteomics data were log10 transformed and then centralized. Student's t-test was used, and the software was R (version 4.1.1). Any differential proteins that fulfilled all of the limitations were considered significant: 1) p-value <0.05; and 2) Fold change ≥2.

Sample preparation for metabolomics

First, each mixture of plasma sample (50 μL) and H2O (150 μL) was vortexed for 30 s. Then, 400 μL acetonitrile was added to the mixture, vortexed for another 30 s and centrifuged at ×14,000 g for 10 min. The samples were dried under vacuum, and the supernatant was then blended with 200 μL of 2% acetonitrile. Before being transferred to the autosamplers, 10 kDa molecular weight cutoff ultracentrifugation filters (Millipore Amicon Ultra, MA) were applied to separate the blood metabolites from the larger molecules. QC samples were prepared by mixing aliquots of one hundred and ten representative samples and they were injected every ten samples throughout the analytical run to assess the method stability and repeatability.

UPLC-MS analysis for metabolomics

The Waters ACQUITY H-class LC system coupled with an AB Sciex TripleTOF 5600 (AB Sciex, United States) was launched to perform the ultra-performance LC-MS analyses of the plasma samples. We separated the plasma metabolites with a 17 min gradient on a Waters HSS C18 column (3.0 × 100 mm, 1.7 μm), and the flow speed was 0.5 mL/min. Mobile phases A and B were 0.1% formic acid in H2O and acetonitrile, respectively. The gradient was as follows: 0–1 min, 2% solvent B; 1–3 min, 2%–15% solvent B; 3–6 min, 15%–50% solvent B; 6–9 min, 50%–95% solvent B; 9–9.1 min, 95%–100% solvent B; 9.1–12 min, 100% solvent B; 12–12.1 min, 100%–2% solvent B; and 12–17 min, 2% solvent B. The column temperature was 45°C. Data dependent acquisition mode was used to acquire the MS and MS/MS spectra. The 10 most abundant ions were submitted for MS/MS fragmentation with a collision energy of 35+-15 eV.

Data processing for metabolomics

Progenesis QI (Waters, Milford, MA, United States) software was applied to analyze the raw data. The data handling and metabolite identification processes can be found in the Supplementary Materials S2. The exported results file consisting of m/z, retention time and relative peak intensity was submitted for further statistical analysis. We established various statistical techniques, such as missing value estimation, log10 transformation and Z score scaling; thus, the features could be more comparable in MetaAnalyst 5.0. The data handling process is depicted in Supplementary Figure S2B, similar to the proteomic process. Any differential variables that fulfilled all the limitations were considered significant: 1) p-value <0.05; and 2) Fold change ≥1.5.

Functional enrichment analysis

The R package "ClusterProfiler" was applied to conduct Gene Oncology (GO) enrichment analysis (Yu et al., 2012). The interaction network between the proteomics and metabolomics and the functional enrichment of the differential metabolites used MetaboAnalyst 5.0 (http://www.metabo analyst.ca). The "WGCNA" package was applied to find characteristic markers of every group (Langfelder and Horvath, 2008; Langfelder and Horvath, 2012). GSEA application (version 4.1.0) was applied to perform GSEA hallmark analysis. The "ClueGo" module of Cytoscape (version 3.9.0, United States) was launched to conduct and display the functional enrichment results (Shannon et al., 2003).

Statistical analysis

Unless specially mentioned above, R (version 4.1.1) was used to perform all the analyzes and construct all the figures. All above tests were two-sided and p-value ≤ 0.05 was regarded as statistically significant. The R package "pwr" (version 1.3–0) has already been applied to calculate the minimum samples required for the analysis.

Results

Human samples and clinical data

A total of 85 patients were enrolled in our final analysis, including 29 TSC-RAML, 29 S-AML and 27 renal cyst (CY) patients. Among the 29 TSC-RAML patients, 25 had double samples, namely, the pre-treatment (pre_TSC) and post-treatment plasma (post_TSC) samples. The basic clinical information of all enrolled patients is shown in Table 1, and the workflow of this study is depicted in Figure 1.

In terms of mutations in the 29 TSC patients, 9 had nonsense mutations, 6 had shift frame mutations, 6 did not have any mutations detected, 4 had missense mutations and 4 had other mutations (2 with base deletions, 1 with an insertion and 1 splicing variation), which can be seen in Supplementary Table S1.

Regarding the treatment effect of everolimus, the results showed that after 3–6 months of treatment, 92% (23/25) of patients experienced tumor reduction, and more than half (56%, 14/25) of patients reached the endpoint of 50% tumor reduction (as depicted in Figure 1B).

The proteome of TSC-RAML, S-AML and renal cyst

Quantitative proteomic data of one hundred and ten plasma samples based on the DIA mode were created. After processing the raw data, a total of 997 proteins remained for further analysis (the process can be seen in Supplementary Figure S1A).

TABLE 1 The baseline information of all enrolled patients.

Items	TSC-RAML		S-AML	Renal cyst
	Pre_treatment	Post_treatment		
Cases (n)	29	25	29	27
Age (years)	29	29.5	39	47
—	(14, 42)	(18, 42)	(15, 54)	(13,78)
Gender (M/F)	11/18	9/16	5/24	13/14

First, t-distributed stochastic neighbor embedding (t-SNE) was applied, and distinctions within the subgroups could be observed, although there was some overlap within the S-AML vs. the CY and post_TSC vs. pre_TSC (Figure 2A). Then, we performed gene co-expression clustering, pathway analysis and functional module classification by means of weighted gene correlation network analysis (WGCNA) and "ClueGO". All 997 proteins were classified into eight whole proteome coexpression clusters (WP-CC), among which the "WP-CC 1" module was positively and significantly associated with TSC-RAML but negatively associated with renal cysts and S-AML (Figure 2B). In addition, the cluster of "WP-CC 2" demonstrated the same tendency. The proteins within the two clusters were then enrolled into the functional analysis and displayed by the "ClueGO". Interestingly, the proteins in the two rewired clusters were mainly enriched in the glycosaminoglycan catabolic process, regulation of phosphatidylinositol 3-kinase signaling and cell-matrix adhesion (WP-CC 1, Figure 2C) and glycosaminoglycan catabolic process, regulation of smooth muscle cell migration and proliferation, extracellular matrix disassembly, and regulation of phospholipase activity pathways (WP-CC 2, Figure 2D).

Comparison of the proteomes of TSC-RAML, S-AML and renal cysts

According to the threshold (FC $\geq$ 2, $p \leq 0.05$), there were 198 differentially expressed (DE) proteins in the pre-treatment TSC-RAML group compared with the renal cyst group, including 73 upregulated and 125 downregulated molecules (Figure 3A, above). Gene oncology (GO) functional enrichment revealed that there were several dysregulated pathways, including platelet degranulation, blood coagulation, hemostasis, cell-matrix adhesion and humoral immune response within the two groups (Figure 3B, above). Since the GO functional enrichment of DE proteins may neglect pivotal information regarding the interactive mechanism, we additionally applied gene set enrichment analysis (GSEA) regarding hallmarks with the molecular signature database (MSigDB v7.4). The results of GSEA hallmark analysis demonstrated that compared with the renal cyst group, the pre-treatment TSC-RAML group possessed two significantly upregulated and seven significantly downregulated pathways (Figures 3C, D). As expected, the angiogenesis pathway was significantly upregulated in the plasma of TSC-RAML patients, which was in accordance with the pathological process of angiomyolipoma biosynthesis (Xian et al., 2011). In addition, KRAS signaling up was upregulation.

Due to the characteristic symptoms and specific mutations of the TSC1 or TSC2 genes, TSC-RAML is quite different from S-AML in many aspects, including multifocal, a larger tumor volume and a higher incidence of tumor rupture, which is the main cause of death among adult TSC-RAML patients (Amin et al., 2017; Lam et al., 2018). Therefore, we also analyzed the plasma proteins in TSC-RAML and S-AML to illustrate their differences at the proteome level.

According to the differential analysis, we identified 174 DE proteins, namely, 77 upregulated and 97 downregulated proteins (Figure 3A, middle). Similarly, the GO enrichment analysis of all DE proteins suggested dysregulated blood coagulation, hemostasis, etc., (Figure 3B, middle). Furthermore, the hallmark GSEA showed that compared with S-AML, TSC-RAML had high targets of angiogenesis and the K-RAS signaling up pathway, which was quite similar to the results of TSC vs. renal cysts (Figures 3C, D).

Differential analysis was also carried out within the post_TSC versus pre_TSC groups to assess the effect of everolimus on plasma proteomics. With the corresponding cutoff value, 40 upregulated and 28 downregulated molecules were observed (presented in Figure 3A, below), and the GO analysis revealed altered nuclear-transcribed mRNA catabolic process, mRNA catabolic process, and protein targeting to ER pathways after everolimus treatment (Figure 3B, below). The GSEA pathway analysis revealed upregulated MYC targets V1, estrogen response late, interferon gamma response and the mTORC1 signaling pathway. Interestingly, treatment with everolimus reversed almost all of the altered pathways caused by the TSC gene mutations (Figure 4).

Diagnostic and prognostic role of serum proteomics in TSC-RAML

To find potential biomarkers that could not only distinguish TSC-RAML from renal cysts and S-AML but also predict the response to everolimus, the DE proteins within the different subgroups were analyzed.

Finally, 34 intersecting molecules were observed (Supplementary Figure S2A). The top 11 upregulated and downregulated proteins are shown in Table 2. From the expression pattern, we can clearly see that most upregulated proteins returned to normal levels after everolimus treatment and *vice versa*.

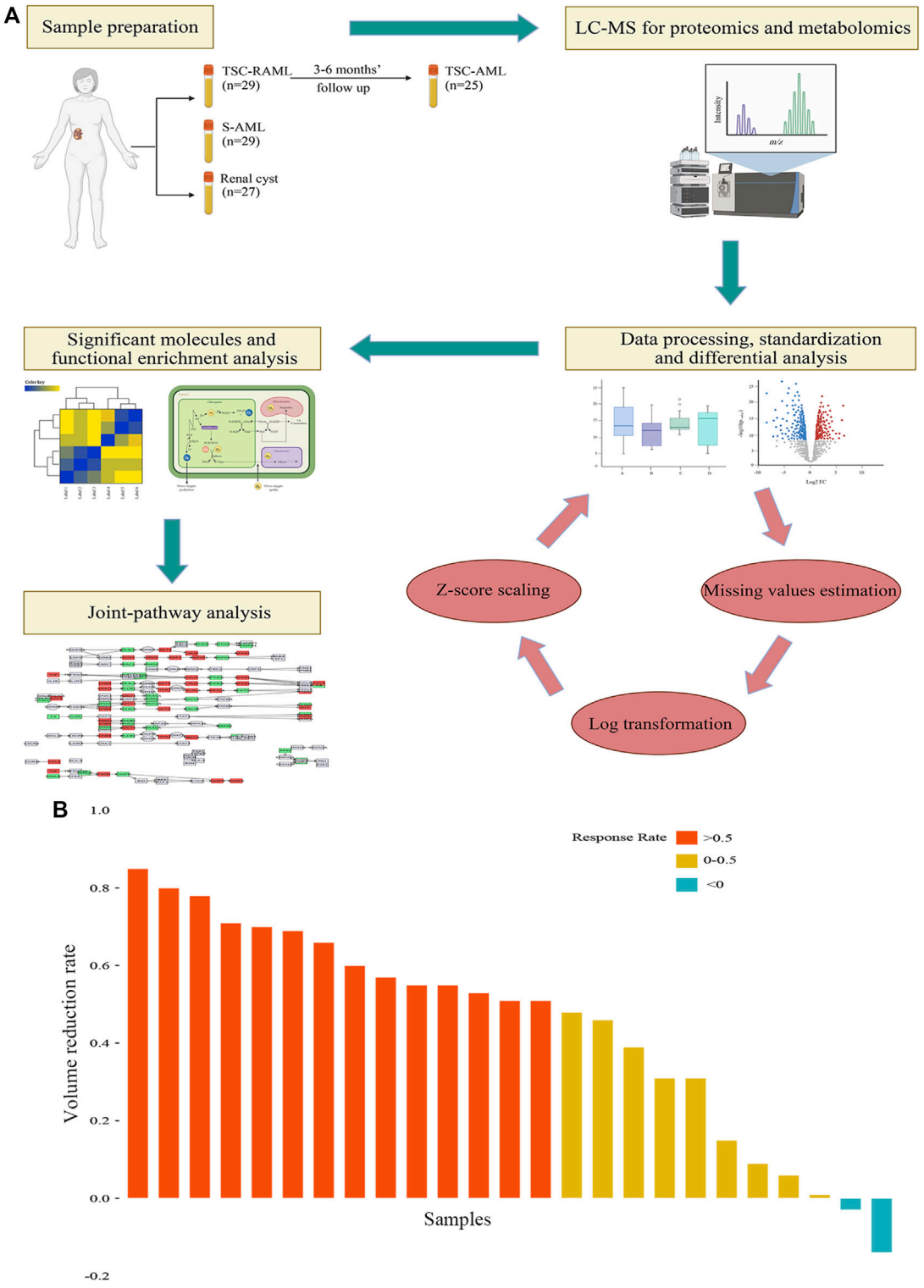


FIGURE 1
(A) The workflow of our study. **(B)** The tumor volume reduction rate of TSC-RAML after everolimus treatment.

From the AUC value, we found that these proteins could perfectly distinguish TSC-RAML from renal cysts and S-AML and within the treatment groups. Furthermore, we compared the correlation between the protein level and maximum tumor volume burden. After applying Pearson analysis, we identified five proteins (out of the 34 intersected DE proteins) positively correlated with the maximum renal angiomyolipoma ($p < 0.05$), namely, PCSK1N, PMEL, HK1, GOT2 and SPTBN2 (as presented in

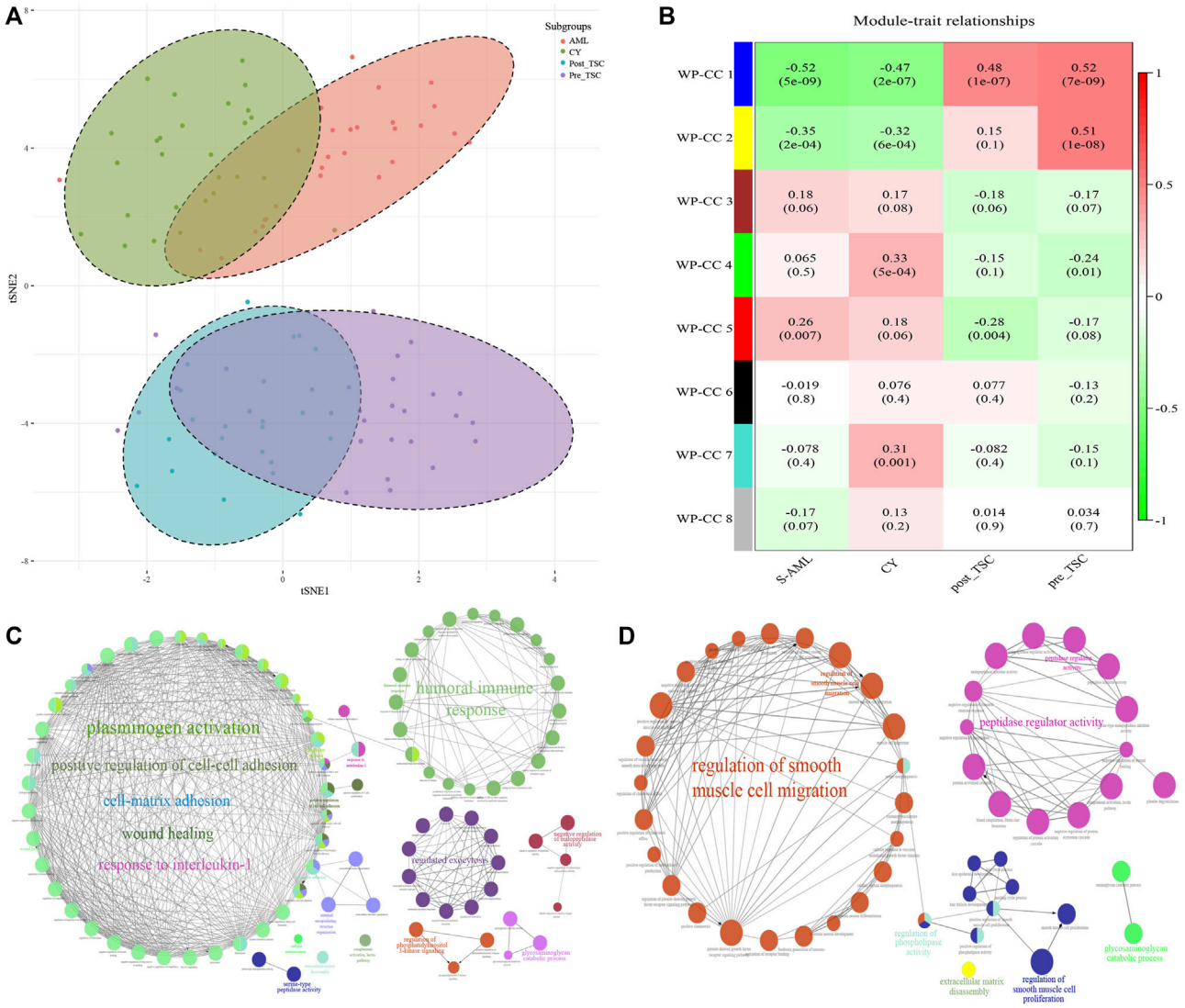


FIGURE 2
Proteomics profile of all enrolled patients. **(A)** t-SNE analysis revealed the unique proteome of TSC-RAML. **(B)** WGCNA identified 8 whole proteome co-expression clusters (WP-CC). **(C)** Functional analysis of the "WP-CC 1" cluster showed dysregulated cell-matrix adhesion. **(D)** Functional analysis of the "WP-CC 2" cluster showed altered regulation of smooth muscle cell migration.

Supplementary Figures 2B–F). Since VEGF-D has been previously proven to be a gold standard biomarker of TSC, we compared the expression level of the intersected proteins with VEGF-D, and many of the proteins demonstrated better discrimination (Figure 5).

The metabolomics of TSC-RAML, S-AML and renal cysts

To describe the metabolomic profiling of TSC-RAML, S-AML and renal cysts, UPLC-MS was applied to measure the concentrations of small metabolites.

Using the same samples and methods for grouping, we measured the plasma metabolites of 110 samples. After pre-analytical data processing (including quality control, missing value estimation, log transformation and Z score scaling), we identified a total of 517 metabolites for further analysis (Supplementary Figure S1B).

First, an unsupervised t-SNE analysis (Figure 6A) was launched, and from the results we can clearly see that there was a distinguished altered metabolomic component within the 4 subgroups, especially with the TSC (including pre-treatment and post-treatment TSC-RAML) vs. renal cyst and S-AML, illustrating the specific metabolomic profiling of TSC-RAML.

Similarly, to find the characteristic metabolomic clusters of TSC-RAML, WGCNA was applied and six whole metabolome coexpression clusters (WM-CC) were constructed, within which "WM-CC 1", "WM-CC 4" and "WM-CC 5" were significantly correlated with TSC-RAML (Figure 6B). The metabolite expression levels of the different modules were obviously different within subgroups (Figure 6C). Furthermore, the pathway enrichment of the three distinguished modules

10.3389/fmolb.2023.1000248

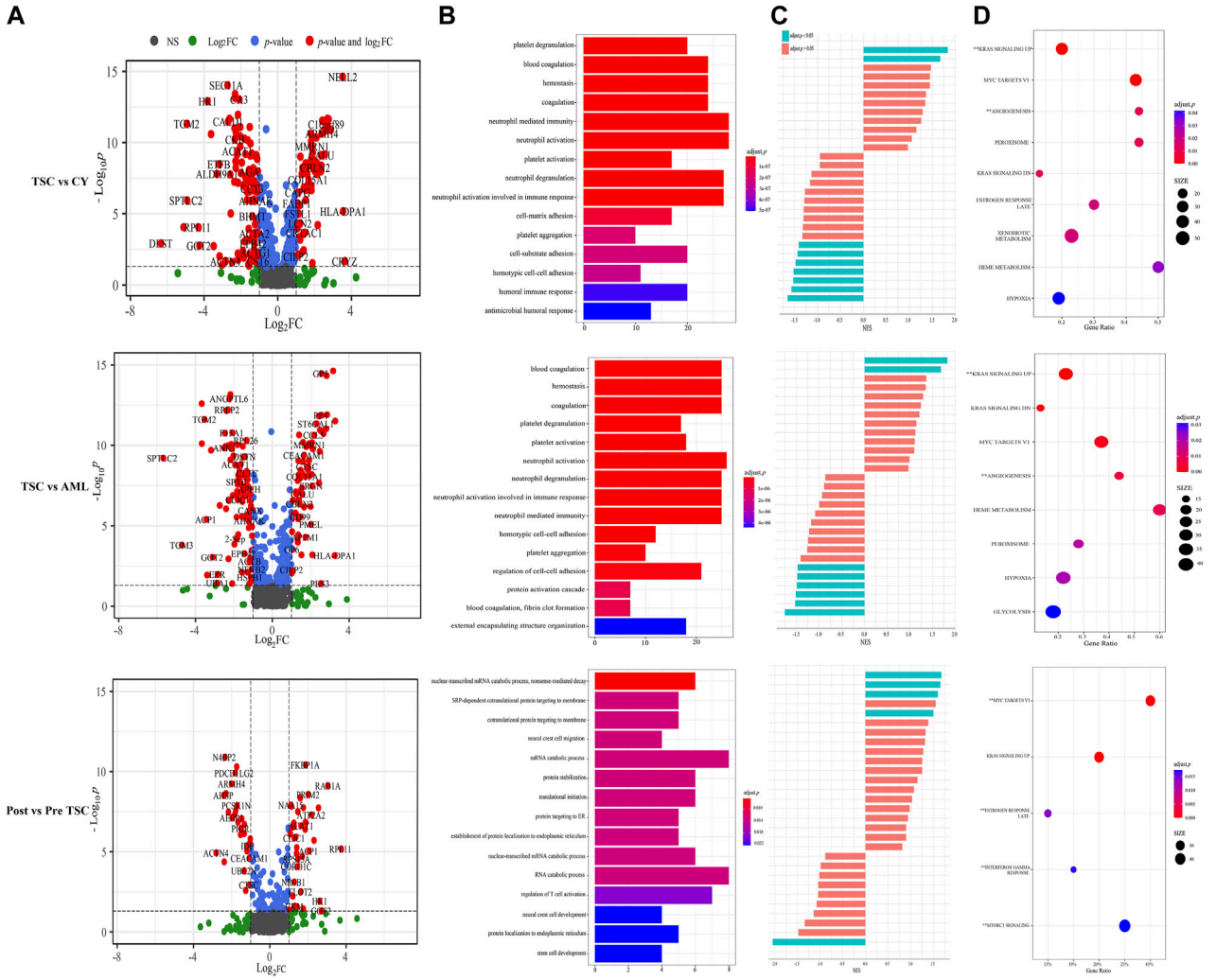


FIGURE 3
Comparison within different groups and functional analysis of DE proteins. **(A)** Volcano plot based on the threshold. **(B)** GO enrichment of DE proteins. **(C)** GSEA items of the proteomics. **(D)** GSEA enrichment of all of the statistically significant items (**significantly upregulated; without annotation: significantly downregulated).

illustrated their altered metabolomic patterns, including upregulated arginine biosynthesis, cysteine and methionine metabolism as well as downregulated amino sugar and nucleotide sugar metabolism and tryptophan metabolism of TSC-RAML (Figures 6D–F).

The comparison metabolomics of TSC-RAML versus renal cysts and sporadic AML and altered metabolomic profiles after everolimus treatment

Similar to proteomics, the 110 samples were first divided into 4 subgroups (pre-treatment TSC-RAML, post-treatment TSC-RAML, renal cysts and S-AML). When comparing pre-treatment TSC-RAML vs. renal cysts, there were 272 differentially expressed metabolites, namely, 116 upregulated and 156 downregulated metabolomic molecules (depicted as a volcano plot in Figure 7A, above). The pathway analysis revealed dysregulated tryptophan metabolism, arginine biosynthesis and glycerophospholipid metabolism (Figure 7B, above). In addition, the joint pathway that integrates DE proteins and metabolites revealed a critically dysregulated metabolism, including the citrate cycle, tryptophan metabolism and pyruvate metabolomic disturbance (Figure 7C, above).

For the TSC-RAML vs. S-AML group, a total of 283 DE metabolites were confirmed, which included 106 upregulated and 177 downregulated metabolites (as depicted in Figure 7A, middle). The pathway analysis revealed altered D-glutamine and D-glutamate metabolism, nitrogen metabolism and porphyrin and chlorophyll metabolism (Figure 7B, middle). The joint pathway analysis stressed the dysregulated glucose metabolism and nitrogen metabolism (Figure 7C, middle).

Regarding the metabolomic effect of everolimus treatment, 22 DE metabolites were identified for the post-treatment vs. pre-treatment TSC-RAML, including 9 upregulated and 13 downregulated

10.3389/fmolb.2023.1000248

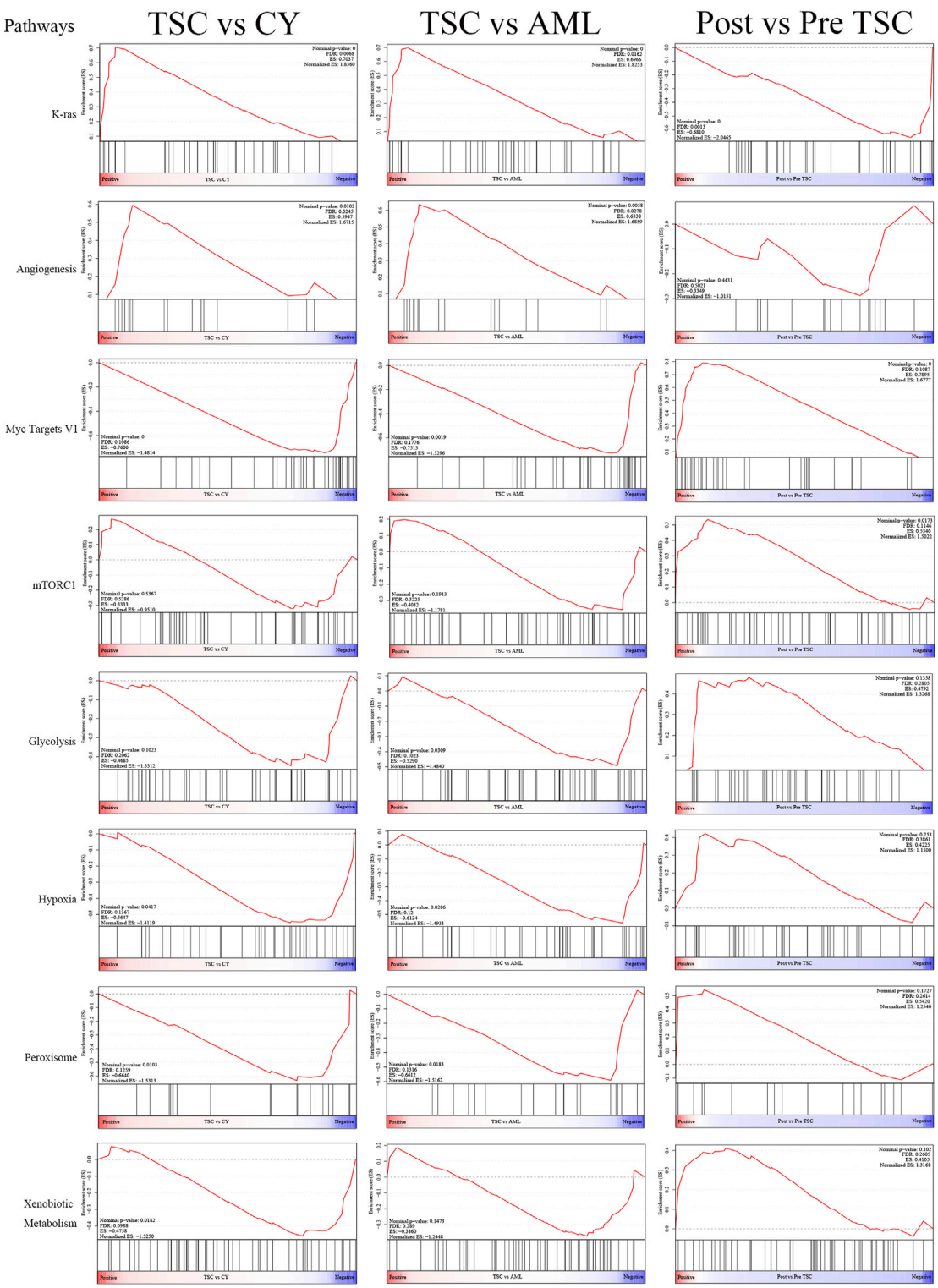


FIGURE 4
The display of significant GSEA enrichment results within the different subgroups.

metabolites (Figure 7A, below). The pathway analysis showed that everolimus treatment changed many pathways, including pyrimidine metabolism and tryptophan metabolism (Figure 7B, below). The joint pathway analysis showed many altered amino acid and nucleotide metabolism pathways (Figure 7C, below).

Potential diagnostic and prognostic metabolite biomarkers of TSC-RAML

To discover both prognostic and diagnostic metabolites, we chose the intersected DE metabolites within different groups. As

TABLE 2 Potential diagnostic and prognostic proteins of TSC-RAML.

Proteins	TSC vs. CY			TSC vs. S-AML			Post vs. Pre TSC		
	FC	*p*-value	AUC	FC	*p*-value	AUC	FC	*p*-value	AUC
PMEL	29.470	$7.738*10^{-16}$	0.98	4.013	$8.235*10^{-6}$	0.80	0.417	0.003	0.72
N4BP2	23.410	$8.112*10^{-23}$	0.96	7.094	$4.641*10^{-15}$	0.93	0.197	$1.245*10^{-11}$	0.88
PCSK1N	19.641	$1.319*10^{-22}$	0.97	8.969	$2.364*10^{-15}$	0.94	0.306	$1.287*10^{-8}$	0.85
AEBP1	18.236	$4.138*10^{-20}$	0.96	19.609	$3.524*10^{-21}$	0.99	0.254	$7.560*10^{-8}$	0.87
TGFBR3	6.789	$2.314*10^{-12}$	0.94	4.141	$6.437*10^{-9}$	0.90	0.395	$7.270*10^{-7}$	0.82
SDHA	6.475	$2.174*10^{-12}$	0.90	2.772	$3.488*10^{-9}$	0.85	0.300	$4.956*10^{-11}$	0.90
CEACAM1	4.274	$4.455*10^{-11}$	0.85	3.148	$4.153*10^{-10}$	0.84	0.479	$2.463*10^{-5}$	0.74
PIGR	3.507	$8.261*10^{-11}$	0.85	2.587	$7.037*10^{-9}$	0.82	0.351	$3.544*10^{-7}$	0.86
COL15A1	3.203	$4.183*10^{-8}$	0.81	3.276	$8.893*10^{-9}$	0.84	0.437	$9.088*10^{-6}$	0.80
PDCD1LG2	3.115	$9.367*10^{-8}$	0.80	2.088	$2.152*10^{-6}$	0.78	0.283	$1.221*10^{-10}$	0.85
SFTPD	2.865	$1.106*10^{-6}$	0.77	3.761	$8.652*10^{-10}$	0.85	0.204	$2.319*10^{-9}$	0.87
GOT2	0.052	$1.775*10^{-3}$	77.1	0.116	$9.052*10^{-3}$	0.76	6.584	0.049	0.65
RPS3	0.054	$6.529*10^{-18}$	0.97	0.079	$7.896*10^{-11}$	0.90	2.662	$5.470*10^{-6}$	0.81
ACP1	0.061	$7.662*10^{-27}$	0.98	0.093	$4.018*10^{-6}$	0.95	4.153	$8.919*10^{-6}$	0.96
HK1	0.071	$1.245*10^{-13}$	0.96	0.473	$3.383*10^{-4}$	0.75	6.133	0.012	0.62
UBA1	0.090	$1.889*10^{-3}$	0.72	0.137	0.036	0.63	6.512	0.049	0.67
NAA15	0.112	$9.777*10^{-3}$	0.76	0.095	0.012	0.73	2.142	$1.210*10^{-8}$	0.86
CALD1	0.154	$3.338*10^{-12}$	0.94	0.109	$2.028*10^{-10}$	0.91	5.864	$1.831*10^{-8}$	0.85
FLOT2	0.171	$9.640*10^{-6}$	0.84	0.0890	$1.248*10^{-22}$	0.96	3.083	0.003	0.78
RPS9	0.204	$2.093*10^{-9}$	0.88	0.434	0.041	0.50	3.710	$1.493*10^{-7}$	0.82
YWHAH	0.216	$7.075*10^{-10}$	0.88	0.277	$1.788*10^{-9}$	0.87	3.033	$4.405*10^{-9}$	0.84
ACAT1	0.217	$3.956*10^{-10}$	0.92	0.262	$1.596*10^{-9}$	0.89	3.101	$2.377*10^{-7}$	0.86

a result, 13 DE metabolites were selected (Supplementary Figure S3A), and the corresponding data are presented in Table 3. After assessing the 13 metabolite levels with the maximum tumor volume with Pearson correlation analysis, we did not find any metabolites associated with the maximum tumor volume burden (Supplementary Figures 3B–F). The relative expression levels of some critical metabolites are depicted in Figure 8, from which we can clearly see that treatment with everolimus could reverse the altered metabolite levels caused by the TSC mutations.

Discussion

In summary, our proteomics analysis found an upregulated angiogenesis pathway, while metabolomics showed the multiple altered amino acid pathways, such as the arginine biosynthesis, tryptophan metabolism and glutamate metabolism. In addition, plasma proteins such as PMEL and metabolites such as S-adenosylmethionine showed potential diagnostic and prognostic functions, demonstrating a significant role in translational medicine, which fills a knowledge gap in this field.

Functional analysis of proteomics

From the GSEA functional enrichment of TSC-RAML, we found that the angiogenesis pathway was significantly upregulated compared with both renal cysts and sporadic AML patients. The WGCNA cluster and ClueGO enrichment also identified characteristic upregulation of smooth muscle cell migration and proliferation in TSC-RAML patients. As the name suggests, angiomyolipoma is comprised of different proportions of proliferative blood vessels, smooth muscle and adipose tissues (Lam et al., 2018). Arbiser, J. L. et al. proved that TSC-associated benign neoplasms, including renal angiomyolipoma, are highly vascular and possess the ability to synthesize and secrete VEGF

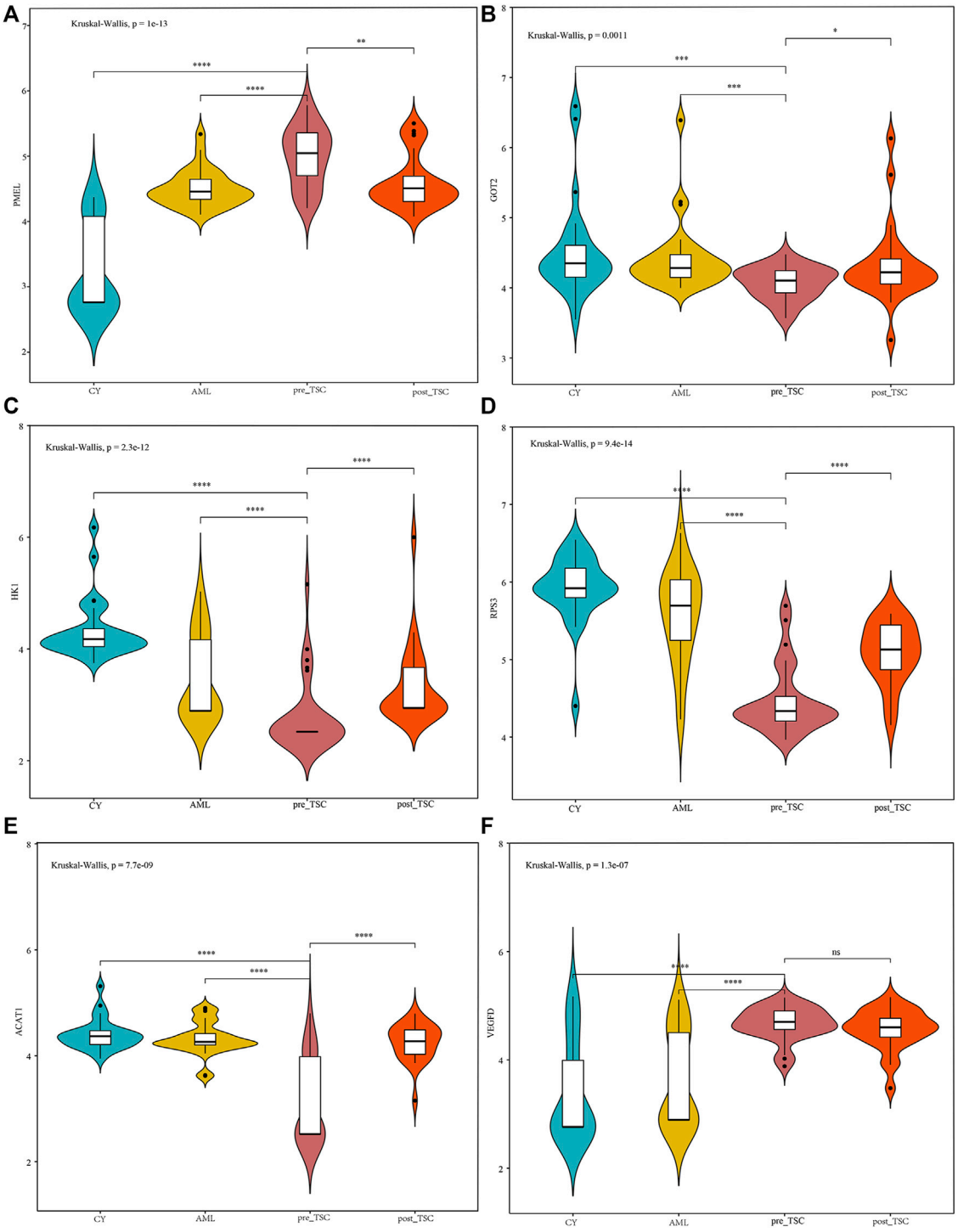


FIGURE 5
The relative expression level of PMEL **(A)**, GOT2 **(B)**, HK1 **(C)**, RPS3 **(D)**, ACAT1 **(E)** and VEGFD **(F)** within different groups. The above sentence should be added after "Relative proteins levels of some important molecules based on the UPLC-MS results.

in vitro (Arbiser et al., 2002). Later, researchers found that mTOR1 plays a central role in the process of angiogenesis through multifactorial ways, including promoting VEGF-A expression by HIF-1α dependent and HIF-1α independent mechanism (Dodd et al., 2015). Based on this hypothesis, additional experiments have suggested that a combination of rapalogs (Rapamycin and its analogs) and angiogenesis inhibitors, such as everolimus plus sorafenib, may

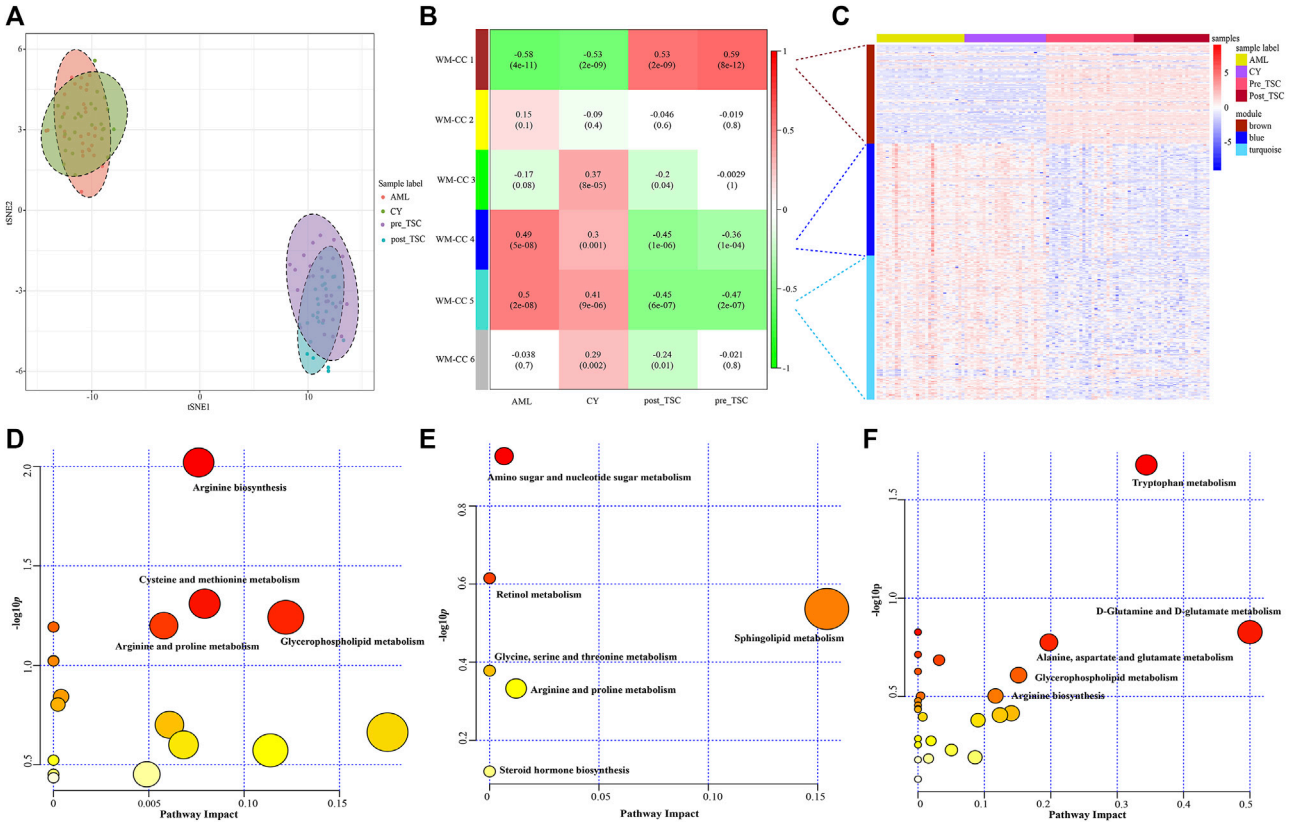


FIGURE 6
Metabolomics profile of all enrolled patients. **(A)** t-SNE analysis revealed the distinguished metabolomic pattern of TSC-RAML. **(B)** WGCNA identified six whole metabolites co-expression clusters (WM-CCs). **(C)** The relative expression levels of metabolites within three differential modules. **(D)** Pathway analysis based on the "WM-CC 1" module revealed altered arginine biosynthesis. **(E)** Pathway analysis based on the "WM-CC 4" module showed altered amino sugar and nucleotide sugar metabolism. **(F)** Pathway analysis based on the "WM-CC 5" module demonstrated altered tryptophan metabolism.

significantly decrease the tumor size and improve the therapeutic efficacy by inhibiting mTORC1 and the mitogen-activated protein kinase (MAPK) pathway (Yang et al., 2017), which is superior to the treatment with single rapalogs alone. Another study also found that angiogenesis inhibitors (sunitinib and bevacizumab) have therapeutic effects on TSC-related tumors, although they are not as effective as rapamycin (Woodrum et al., 2010).

Another significantly upregulated pathway relative to renal cysts and sporadic AML was K-RAS pathway activation, which has been proven to play a critical role in the tumorigenesis of various cancers and therefore has been implicated as a cancer target during the past few years, such as in pancreatic ductal adenocarcinoma (Mehra et al., 2021), lung cancer (Chu, 2020), and breast cancer (Gupta et al., 2020). As an important downstream target of the K-RAS signaling pathway, the role of PI3K-Akt-mTOR axis in tumor occurrence and development has been validated by a variety of researchers (Hillmann and Fabbro, 2019). Although many drugs targeting the K-RAS pathway have been explored to induce tumor regression in other diseases (Kinross et al., 2011; Hillmann and Fabbro, 2019), the evidence for their use in TSC-RAML is limited. Therefore, our results may provide new ideas for the treatment of rapamycin-resistant TSC-RAML.

Protein biomarkers for differential diagnosis and everolimus effect prognostication

In our analysis, we discovered that the plasma pre-melanosome protein PMEL, antigen for HMB-45, demonstrates good differential (AUC of TSC vs. CY: 0.98; AUC of TSC vs. AML: 0.80) and prognostic ability (AUC of Post vs. Pre TSC: 0.72), as depicted in Table 2. In addition, the PMEL level was also associated with the tumor burden ($r = 0.55$, $p < 0.001$, as depicted in Figure Supplementary Figure S2E). To the best of our knowledge, our study is the first to discover the latent role of plasma PMEL in diagnosing and predicting the outcome of TSC-RAML. Pigment cell-specific PMEL is an extraordinarily well-conserved type I transmembrane glycoprotein mainly engaged in the formation of fibrillar sheets within melanosomes (Watt et al., 2013), and it is associated with melanocyte-related diseases and pathological neurodegeneration, such as Alzheimer's Disease (AD) and Parkinson's disease (PD) (Watt et al., 2013).

In 2001, Stone, C. H. assessed the relationship between the immunophenotypic and ultrastructural profile of renal angiomyolipoma and found that all 27 renal angiomyolipomas stained positive for HMB-45, regardless of their identification as epithelioid, spindle, or adipocytic cells, suggesting all components were coming from a common cell ancestor and providing a unitarian

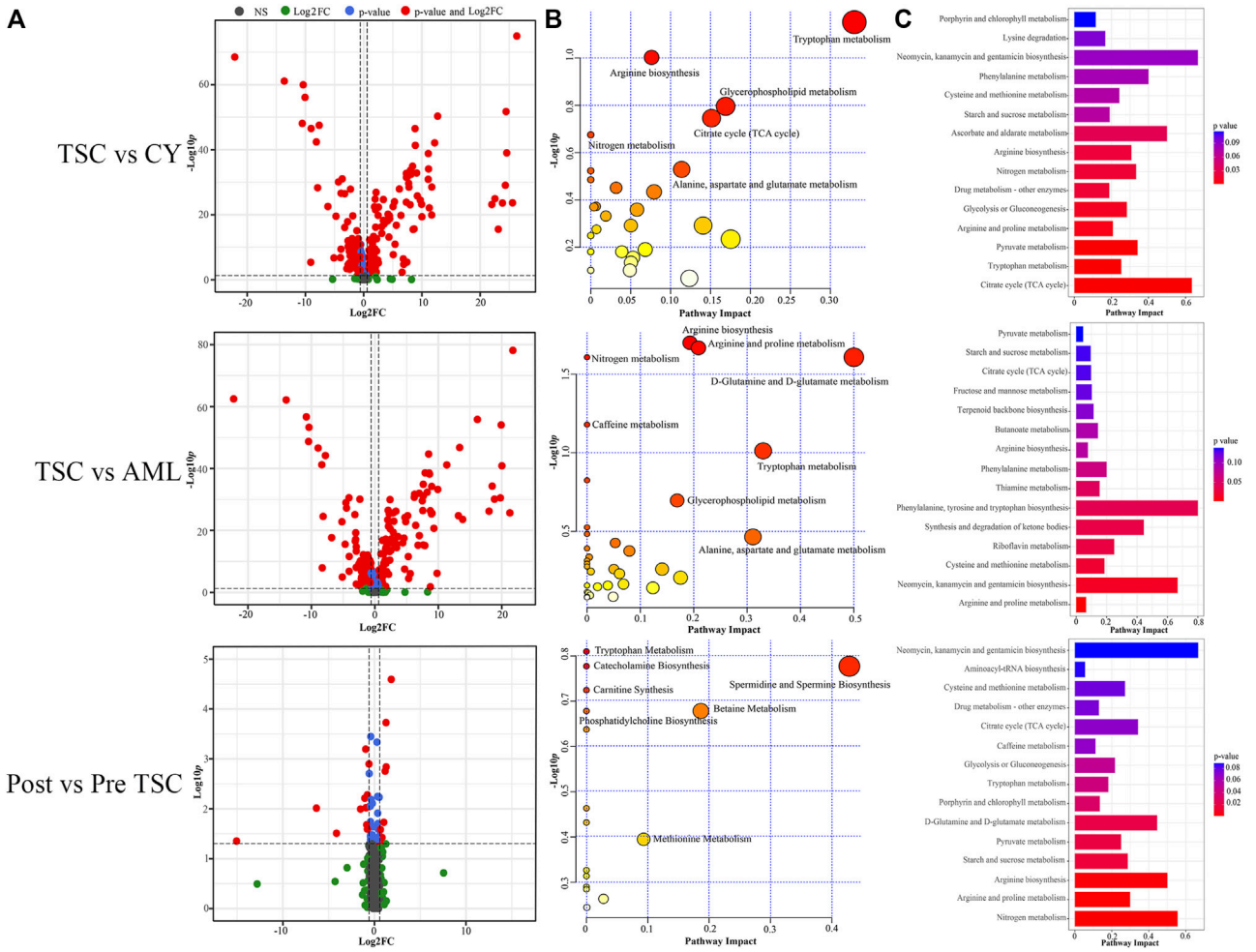


FIGURE 7
Comparison within different groups and functional analysis of the DE metabolomics. **(A)** Volcano plot based on the threshold. **(B)** Pathway enrichment of DE metabolites showed the many altered metabolomic pattern within different groups, including tryptophan metabolism. **(C)** Joint pathway analysis integrating the DE proteins and metabolites.

concept for renal angiomyolipoma (Stone et al., 2001). In addition to angiomyolipoma, pulmonary LAM cells are also positive for HMB-45 (Venyo, 2016; Guo et al., 2020), indicating that neural crest cells, a kind of migratory, multipotent embryonic cell, maybe the cell origin for LAM and other TSC-related tumors (Delaney et al., 2014). Two recent published articles have found the relative reduction of T lymphocytes within the tumor microenvironment for TSC related LAM (Guo et al., 2020) and AML (Wang et al., 2022b), suggesting adoptive transferred PMEL-specific $CD8^+$ T cells may be effective because this cytotoxic T cells can specifically attack PMEL + tumor cells (Hanada et al., 2019; Han et al., 2020).

Another protein, PCSK1N, also called proSAAS, an inhibitor of prohormone convertase 1 (PC1) activity produced by neuroendocrine cells, has been proven to be a biomarker for many neurological disorders, including Alzheimer's disease (AD), Pick's disease, and the Parkinsonism-dementia complex (Shakya et al., 2020; van Steenoven et al., 2020). Encoded by the *PSCK1N* gene, ProSAAS was initially identified as a neuroendocrine-specific proprotein convertase binding protein and was classified into the granin family of proteins (Shakya et al., 2020). In addition, proSAAS can be proteolytically processed into a large number of active neuropeptides, including SAAS, PEN and LEN, all of which have been regarded as neurotransmitters (Khoonsari et al., 2019). Several proteomic and transcript studies have found elevated proSAAS protein levels in cerebrospinal fluid and upregulated proSAAS expression in the brain during Alzheimer's progression (McDermott et al., 2019).

More than 90% of TSC patients have central nervous system abnormalities, including cortical or subcortical tubers, subependymal nodules, giant cell astrocytoma, and white matter migration lines (Curatolo et al., 2015). These pathological lesions can lead to many neurological symptoms, such as epilepsy and tuberous sclerosis-associated neuropsychiatric disorders (TANDs). In our analysis, we found that plasma PCSK1N was significantly elevated compared with renal cyst (FC = 19.6, $p = 1.3^{*}10^{-22}$) and S-AML (FC = 8.97, $p = 2.36^{*}10^{-5}$) but was reduced

TABLE 3 Potential diagnostic and prognostic metabolites of TSC-RAML.

Metabolites	TSC vs. CY			TSC vs. S-AML			Post vs. Pre TSC		
	FC	*p*-value	AUC	FC	*p*-value	AUC	FC	*p*-value	*AUC*
Lucyoside K	6,657.429	$4.883*10^{-51}$	1.0	73,025.948	$1.443*10^{-56}$	1.0	0.662	$1.26*10^{-3}$	0.76
His Trp	2,171.442	$1.311*10^{-31}$	1.0	36.0719	$1.762*10^{-25}$	0.98	1.507	0.0483	0.68
Pro Pro Glu Phe	404.041	$2.352*10^{-16}$	1.0	111.556	$5.911*10^{-16}$	0.99	1.783	0.0377	0.67
Inosine	16.936	$8.377*10^{-19}$	0.98	5.760	$2.746*10^{-14}$	0.95	0.635	0.0240	0.67
Dipropyl sulfide	3.947	$5.690*10^{-4}$	0.79	4.141	$9.994*10^{-3}$	0.73	0.3487	0.0101	0.72
S-Adenosylmethionine	3.037	$2.174*10^{-12}$	0.81	2.236	$3.155*10^{-4}$	0.76	0.581	0.0256	0.71
Gly Trp Glu Ser	0.0671	$3.920*10^{-10}$	0.94	0.0608	$2.444*10^{-12}$	0.95	3.569	$2.549*10^{-5}$	0.87
Adenosine 3′-monophosphate	0.105	$8.144*10^{-17}$	0.96	0.0389	$2.938*10^{-16}$	0.95	0.0565	0.0310	0.57
3,4-Methylenedioxymethamphetamine (MDMA)	0.1478	$1.383*10^{-18}$	1.0	0.123	$5.726*10^{-20}$	1.0	0.481	0.00613	0.71
Gly Asp Ala Ala	0.156	$6.092*10^{-11}$	0.96	0.131	$1.910*10^{-15}$	0.97	3.569	$2.549*10^{-5}$	0.76
Aspartyl-Tryptophan	0.198	$8.160*10^{-10}$	0.94	0.131	$4.811*10^{-14}$	0.99	2.429	0.00145	0.79
Ketotifen-N-glucuronide	0.265	$8.674*10^{-6}$	0.83	0.363	$6.097*10^{-4}$	0.76	0.5145	$6.686*10^{-4}$	0.75
alpha-Terpineol formate	0.547	$1.776*10^{-8}$	0.91	0.524	$6.846*10^{-7}$	0.87	0.6539	0.00913	0.70

dramatically after everolimus treatment (FC of post vs. pre = 0.3, $p = 1.29*10^{-8}$), which indicated that plasma PCSK1N may be a useful marker for TSC.

Furthermore, some other biomarkers, including SDHA, GOT2, HK1 and ACAT1, are involved in metabolic processes, including glucose metabolism, and amino acid and fatty acid metabolism, indicating the wide reprogramming of vital metabolic and biochemical processes caused by *TSC* gene mutations (Lam et al., 2018).

Functional analysis of metabolomics

In our comparative metabolomic analysis, we also found characteristic plasma metabolomic patterns of TSC-RAML, including arginine biosynthesis, glutamine and glutamate metabolism, tryptophan metabolism, and glycerophospholipid metabolism, which was consistent with the joint pathway analysis integrating proteomics and metabolomics (as depicted in Figures 6, 7). The overactivated mTOR pathway caused by dysfunctional hamartin or tuberin could lead to a subsequent metabolic alteration to sustain necessary proliferation and survival, including aberrant metabolism of amino acids, glucose, nucleotides, fatty acids and lipids. On the other hand, the altered metabolites, particularly amino acids such as arginine and glutamine (Wolfson and Sabatini, 2017), could reversely stimulate mTOR *via* RAS related GTP binding proteins (Mossmann et al., 2018), resulting in positive feedback. As one of the several amino acids that can directly activate the mTOR pathway, arginine can modulate cellular signaling pathways through many mechanisms, such as been transformed into the cytoplasm by solute carriers (SLCs) or by binding to L-amino acid receptor, G-protein coupled receptor GPRCA6 (Chen et al., 2021). In contrast, deprivation of arginine could convert Rag GTPases into an inactive state and lead to the immediate deactivation of mTORC1 (Darnell et al., 2018), thus suppressing the growth and inducing cell death of various cancer types, and corresponding clinical trials are being conducted (Chen et al., 2021). Our metabolomic analysis showed that arginine biosynthesis was significantly upregulated and that the fold change in L-arginine could even reach 2.183 and 1.89 compared with renal cysts and AML, respectively ($p < 0.01$), which suggested that arginine-targeted drugs or an arginine-light diet may be a promising choice for TSC-RAML patients.

In contrast to arginine directly activating the mTOR pathway, glutamine can activate mTOR through a Rag GTPase-independent pathway and it requires the participation of ADP-ribosylation factor 1 (Arf1) (Yan et al., 2020). In addition, TSC-deficient cells have also demonstrated increasing consumption of glutamine to engage in an overactive tricarboxylic acid cycle (which has already been depicted in Figure 7) and create the antioxidant agent glutathione (Lam et al., 2018). Another important pathway, glycerophospholipid metabolism, which has been reported by Bottolo, L. et al. in their research regarding TSC-related LAM, was associated with the severity of lung disease and total body burden of LAM (Bottolo et al., 2020). In our study, however, glycerophospholipids showed an upregulated tendency but it did not reach statistical significance ($p > 0.05$). We think the difference may be due to the inner heterogeneity with TSC-LAM and TSC-RAML and the limited samples within our two studies. Therefore, larger sample size and more centers should be engaged to validate these results.

10.3389/fmolb.2023.1000248

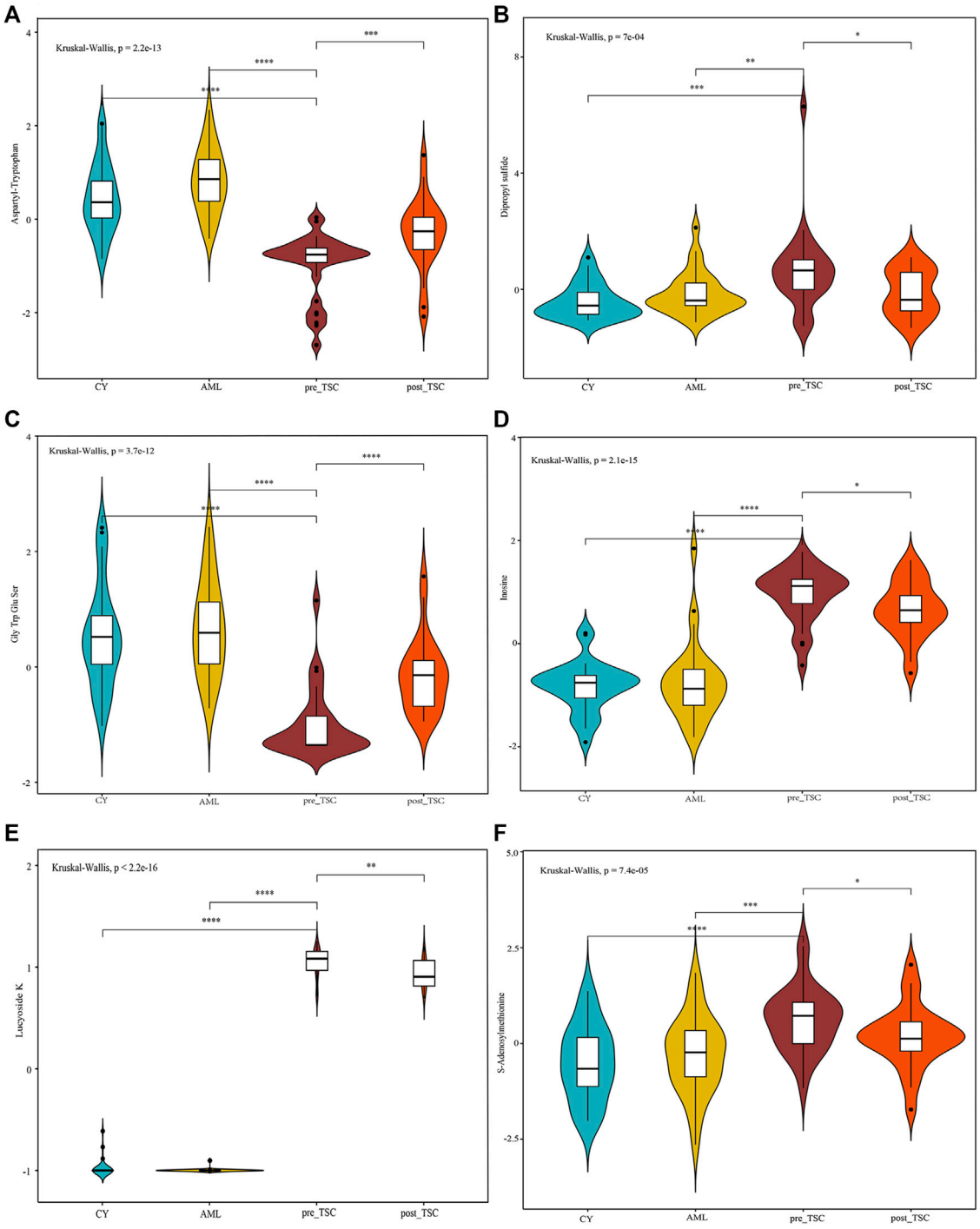


FIGURE 8
Relative metabolite levels of some important molecules based on the UPLC-MS results. The relative expression level of Aspartyl-Tryptophan **(A)**, Dipropyl sulfide **(B)**, Gly Trp Glu Ser **(C)**, Inosine **(D)**, Lucyoside K **(E)** and S-Adenosylmethionine **(F)** within different groups.

Metabolite biomarkers for differential diagnosis and prognosis prediction

Regarding prognostic and diagnostic biomarkers, several metabolites attracted our attention, including S-adenosylmethionine, inosine, and adenosine 3′-monophosphate. As one of the most important methyl donors, S-adenosylmethionine (SAM) plays a critical role in the methylation of multiple biological processes, including DNA, RNA and histone methylation as well as the synthesis of creatine and phosphatidylcholine (Elango, 2020; Menezo et al., 2020), which may

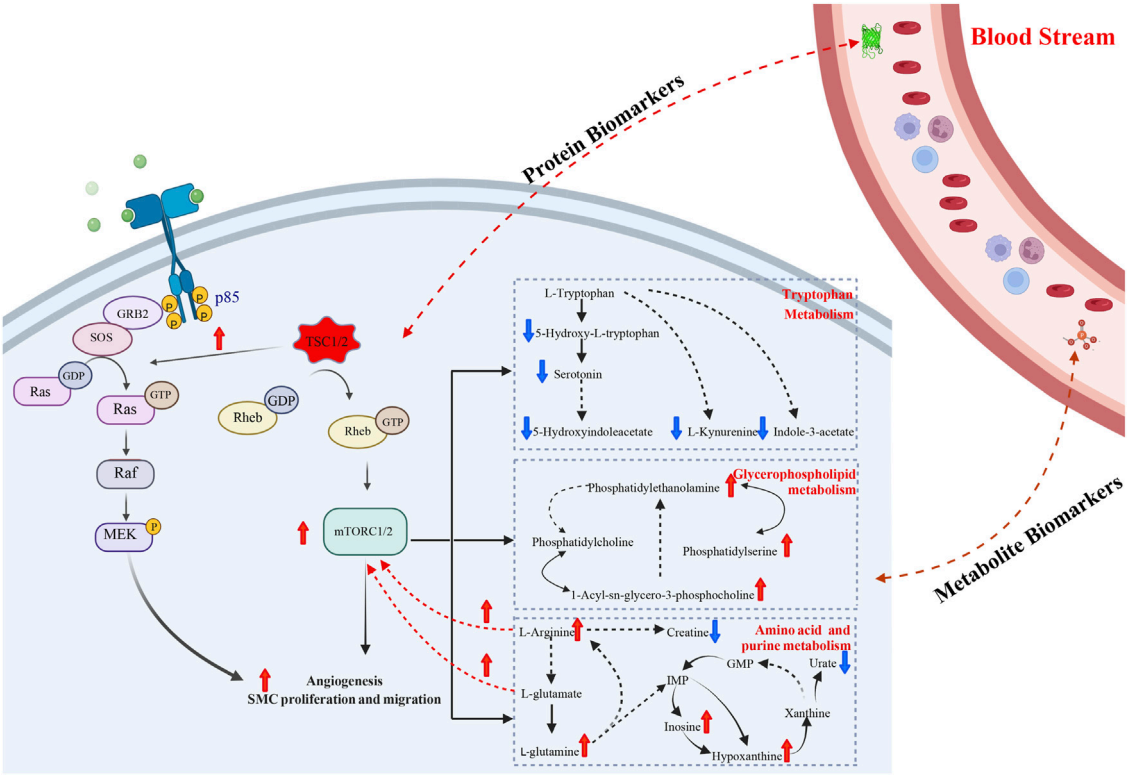


FIGURE 9
The main summary of discovery in our multi-omics analysis.

be the reason why the level of inosine showed the same tendency as SAM (as depicted in Figure 8). Researchers have found that intracellular SAM can be detected by SAMTOR, a sensor for SAM binding with KICTOR, thus leading to mTORC1 activation and autophagy suppression (Kitada et al., 2021). In addition, SAM is also the sole donor of aminopropyl groups, which have been proven to be overexpressed in various cancers and are vital for cell proliferation (Kaiser, 2020). SAM mainly originates from methionine and ATP under the catalysis of methionine adenosyltransferase (MAT). MAT contains three isozymes in mammals. MAT1 and MAT3 are limited in hepatocytes, while MAT2 are widely expressed in almost all tissues (Alam et al., 2022). Accumulating evidence suggested that SAM and its enzyme MAT2A are closed related with tumorigenesis of various cancers, like colon and breast cancers (Alam et al., 2022). Targeting SAM or MAT2A has proven beneficial among several type of cancers, especially in *MTAP*-deleted cancers (Bruce et al., 2021; Kalev et al., 2021). We suggest that a high level of plasma SAM could satisfy a higher demand for nutritional supplies and altered methylation pattern to sustain tumor progression, indicating that SAM could be a potential pharmacological target, and further research is required.

As retrospective research, our study has some potential disadvantages. First, the small sample size due to the essence of rare disease and lack of external validation may limit the wide application of biomarkers in clinical. To overcome this drawback, we are carrying out multi-center cooperation and the result will be published once finished. In addition, although we have discovered many biomarkers for TSC-RAML, more vitro and vivo experiments targeting the molecules should be carried out to explore the inner mechanism.

In conclusion, we integrated the plasma proteomics and metabolomics data of TSC-RAML and discovered altered unique pathways as well as potential prognostic and diagnostic biomarkers (as summarized in Figure 9). Our results provide new thoughts regarding the underlying mechanism of TSC-RAML and potential drug targets for future research.

Data availability statement

The data presented in the study are deposited in the Genome Sequence Archive (https://ngdc.cncb.ac.cn/bioproject/) under the accession number 'PRJCA014958'.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of Peking Union Medical College Hospital and the Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences. The patients/participants provided their written informed consent to participate in this study.

Author contributions

YuZ, WS, and LJ designed the study, analyzed and interpreted the data. ZW, XL, and WW conducted the experiments, analyzed data and wrote the manuscript under the instruction of YuZ, WS, and LJ. XW, YaZ, and ZL collected the samples and clinical data, and revised the manuscript. JX, HS, JW, and YY assisted in the process of conducting experiments, paper writing and revising. All authors reviewed and approved the manuscript for publication.

Funding

This research was partially supported by the National High Level Hospital Clinical Research funding from Peking Union Medical College Hospital (2022-PUMCH-A-151).

Acknowledgments

All the enrolled patients should be given our sincere appreciation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1000248/full#supplementary-material

References

Alam, M., Shima, H., Matsuo, Y., Long, N. C., Matsumoto, M., Ishii, Y., et al. (2022). mTORC1-independent translation control in mammalian cells by methionine adenosyltransferase 2A and S-adenosylmethionine. *J. Biol. Chem.* 298 (7), 102084. doi:10.1016/j.jbc.2022.102084

Amaral, A. F., de Oliveira, M. R., Dias, O. M., Arimura, F. E., Freitas, C. S. G., Acencio, M. M. P., et al. (2019). Concentration of serum vascular endothelial growth factor (VEGF-D) and its correlation with functional and clinical parameters in patients with lymphangioleiomyomatosis from a Brazilian reference center. *Lung* 197 (2), 139–146. doi:10.1007/s00408-018-00191-3

Amin, S., Lux, A., Calder, N., Laugharne, M., Osborne, J., and O'Callaghan, F. (2017). Causes of mortality in individuals with tuberous sclerosis complex. *Dev. Med. Child Neurol.* 59 (6), 612–617. doi:10.1111/dmcn.13352

Arbiser, J. L., Brat, D., Hunter, S., D'Armiento, J., Henske, E. P., Arbiser, Z. K., et al. (2002). Tuberous sclerosis-associated lesions of the kidney, brain, and skin are angiogenic neoplasms. *J. Am. Acad. Dermatol.* 46 (3), 376–380. doi:10.1067/mjd.2002.120530

Behsaz, B., Bode, E., Gurevich, A., Shi, Y. N., Grundmann, F., Acharya, D., et al. (2021). Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. *Nat. Commun.* 12 (1), 3225. doi:10.1038/s41467-021-23502-4

Bissler, J. J., Kingswood, J. C., Radzikowska, E., Zonnenberg, B. A., Frost, M., Belousova, E., et al. (2013). Everolimus for angiomyolipoma associated with tuberous sclerosis complex or sporadic lymphangioleiomyomatosis (EXIST-2): A multicentre, randomised, double-blind, placebo-controlled trial. *Lancet (London, Engl.)* 381 (9869), 817–824. doi:10.1016/S0140-6736(12)61767-X

Blomme, A., Ford, C. A., Mui, E., Patel, R., Ntala, C., Jamieson, L. E., et al. (2020). 2,4-dienoyl-CoA reductase regulates lipid homeostasis in treatment-resistant prostate cancer. *Nat. Commun.* 11 (1), 2508. doi:10.1038/s41467-020-16126-7

Bottolo, L., Miller, S., and Johnson, S. R. (2020). Sphingolipid, fatty acid and phospholipid metabolites are associated with disease severity and mTOR inhibition in lymphangioleiomyomatosis. *Thorax* 75 (8), 679–688. doi:10.1136/thoraxjnl-2019-214241

Bruce, J. P., To, K. F., Lui, V. W. Y., Chung, G. T. Y., Chan, Y. Y., Tsang, C. M., et al. (2021). Whole-genome profiling of nasopharyngeal carcinoma reveals viral-host co-operation in inflammatory NF-κB activation and immune escape. *Nat. Commun.* 12 (1), 4193. doi:10.1038/s41467-021-24348-6

Cai, Y., Guo, H., Wang, W., Li, H., Sun, H., Shi, B., et al. (2018). Assessing the outcomes of everolimus on renal angiomyolipoma associated with tuberous sclerosis complex in China: A two years trial. *Orphanet J. rare Dis.* 13 (1), 43. doi:10.1186/s13023-018-0781-y

Chen, C. L., Hsu, S. C., Ann, D. K., Yen, Y., and Kung, H. J. (2021). Arginine signaling and cancer metabolism. *Cancers* 13 (14), 3541. doi:10.3390/cancers13143541

Chu, Q. S. (2020). Targeting non-small cell lung cancer: Driver mutation beyond epidermal growth factor mutation and anaplastic lymphoma kinase fusion. *Ther. Adv. Med. Oncol.* 12, 1758835919895756. doi:10.1177/1758835919895756

Curatolo, P., Moavero, R., and de Vries, P. J. (2015). Neurological and neuropsychiatric aspects of tuberous sclerosis complex. *Lancet Neurol.* 14 (7), 733–745. doi:10.1016/S1474-4422(15)00069-1

Dabora, S. L., Franz, D. N., Ashwal, S., Sagalowsky, A., DiMario, F. J., Jr., Miles, D., et al. (2011). Multicenter phase 2 trial of sirolimus for tuberous sclerosis: Kidney angiomyolipomas and other tumors regress and VEGF- D levels decrease. *PloS one* 6 (9), e23379. doi:10.1371/journal.pone.0023379

Darnell, A. M., Subramaniam, A. R., and O'Shea, E. K. (2018). Translational control through differential ribosome pausing during amino acid limitation in mammalian cells. *Mol. Cell* 71 (2), 229–243.e11. doi:10.1016/j.molcel.2018.06.041

de la Calle Arregui, C., Plata-Gómez, A. B., Deleyto-Seldas, N., García, F., Ortega-Molina, A., Abril-Garrido, J., et al. (2021). Limited survival and impaired hepatic fasting metabolism in mice with constitutive Rag GTPase signaling. *Nat. Commun.* 12 (1), 3660. doi:10.1038/s41467-021-23857-8

Delaney, S. P., Julian, L. M., and Stanford, W. L. (2014). The neural crest lineage as a driver of disease heterogeneity in Tuberous Sclerosis Complex and Lymphangioleiomyomatosis. *Front. Cell Dev. Biol.* 2, 69. doi:10.3389/fcell.2014.00069

Dodd, K. M., Yang, J., Shen, M. H., Sampson, J. R., and Tee, A. R. (2015). mTORC1 drives HIF-1α and VEGF-A signalling via multiple mechanisms involving 4E-BP1, S6K1 and STAT3. *Oncogene* 34 (17), 2239–2250. doi:10.1038/onc.2014.164

Elango, R. (2020). Methionine nutrition and metabolism: Insights from animal studies to inform human nutrition. *J. Nutr.* 150, 2518s–2523s. doi:10.1093/jn/nxaa155

Franz, D. N., Belousova, E., Sparagana, S., Bebin, E. M., Frost, M., Kuperman, R., et al. (2013). Efficacy and safety of everolimus for subependymal giant cell astrocytomas associated with tuberous sclerosis complex (EXIST-1): A multicentre, randomised, placebo-controlled phase 3 trial. *Lancet (London, Engl.)* 381 (9861), 125–132. doi:10.1016/S0140-6736(12)61134-9

Guo, M., Yu, J. J., Perl, A. K., Wikenheiser-Brokamp, K. A., Riccetti, M., Zhang, E. Y., et al. (2020). Single-cell transcriptomic analysis identifies a unique pulmonary lymphangioleiomyomatosis cell. *Am. J. Respir. Crit. Care Med.* 202 (10), 1373–1387. doi:10.1164/rccm.201912-2445OC

Gupta, G. K., Collier, A. L., Lee, D., Hoefer, R. A., Zheleva, V., Siewertsz van Reesema, L. L., et al. (2020). Perspectives on triple-negative breast cancer: Current treatment strategies, unmet needs, and potential targets for future therapies. *Cancers* 12 (9), 2392. doi:10.3390/cancers12092392

Han, F., Dellacecca, E. R., Barse, L. W., Cosgrove, C., Henning, S. W., Ankney, C. M., et al. (2020). Adoptive T-cell transfer to treat lymphangioleiomyomatosis. *Am. J. Respir. Cell Mol. Biol.* 62 (6), 793–804. doi:10.1165/rcmb.2019-0117OC

Hanada, K. I., Yu, Z., Chappell, G. R., Park, A. S., and Restifo, N. P. (2019). An effective mouse model for adoptive cancer immunotherapy targeting neoantigens. *JCI insight* 4 (10), e124405. doi:10.1172/jci.insight.124405

Henske, E. P., Jóźwiak, S., Kingswood, J. C., Sampson, J. R., and Thiele, E. A. (2016). Tuberous sclerosis complex. *Nat. Rev. Dis. Prim.* 2, 16035. doi:10.1038/nrdp.2016.35

Hillmann, P., and Fabbro, D. (2019). PI3K/mTOR pathway inhibition: Opportunities in oncology and rare genetic diseases. *Int. J. Mol. Sci.* 20 (22), 5792. doi:10.3390/ijms20225792

Kaiser, P. (2020). Methionine dependence of cancer. *Biomolecules* 10 (4), 568. doi:10.3390/biom10040568

Kalev, P., Hyer, M. L., Gross, S., Konteatis, Z., Chen, C. C., Fletcher, M., et al. (2021). MAT2A inhibition blocks the growth of MTAP-deleted cancer cells by reducing PRMT5-dependent mRNA splicing and inducing DNA damage. *Cancer Cell* 39 (2), 209–224.e11. doi:10.1016/j.ccell.2020.12.010

Khoonsari, P. E., Musunri, S., Herman, S., Svensson, C. I., Tanum, L., Gordh, T., et al. (2019). Systematic analysis of the cerebrospinal fluid proteome of fibromyalgia patients. *J. Proteomics* 190, 35–43. doi:10.1016/j.jprot.2018.04.014

Kinross, K. M., Brown, D. V., Kleinschmidt, M., Jackson, S., Christensen, J., Cullinane, C., et al. (2011). *In vivo* activity of combined PI3K/mTOR and MEK inhibition in a Kras(G12D);Pten deletion mouse model of ovarian cancer. *Mol. Cancer Ther.* 10 (8), 1440–1449. doi:10.1158/1535-7163.MCT-11-0240

Kitada, M., Ogura, Y., Monno, I., Xu, J., and Koya, D. (2021). Effect of methionine restriction on aging: Its relationship to oxidative stress. *Biomedicines* 9 (2), 130. doi:10.3390/biomedicines9020130

Krueger, D. A., Care, M. M., Holland, K., Agricola, K., Tudor, C., Mangeshkar, P., et al. (2010). Everolimus for subependymal giant-cell astrocytomas in tuberous sclerosis. *N. Engl. J. Med.* 363 (19), 1801–1811. doi:10.1056/NEJMoa1001671

Lam, H. C., Siroky, B. J., and Henske, E. P. (2018). Renal disease in tuberous sclerosis complex: Pathogenesis and therapy. *Nat. Rev. Nephrol.* 14 (11), 704–716. doi:10.1038/s41581-018-0059-6

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Langfelder, P., and Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* 46 (11), i11. doi:10.18637/jss.v046.i11

Malinowska, I. A., Lee, N., Kumar, V., Thiele, E. A., Franz, D. N., Ashwal, S., et al. (2013). Similar trends in serum VEGF-D levels and kidney angiomyolipoma responses with longer duration sirolimus treatment in adults with tuberous sclerosis. *PloS one* 8 (2), e56199. doi:10.1371/journal.pone.0056199

Martin, T. D., Chen, X. W., Kaplan, R. E., Saltiel, A. R., Walker, C. L., Reiner, D. J., et al. (2014). Ral and Rheb GTPase activating proteins integrate mTOR and GTPase signaling in aging, autophagy, and tumor cell invasion. *Mol. Cell* 53 (2), 209–220. doi:10.1016/j.molcel.2013.12.004

McCormack, F. X., Inoue, Y., Moss, J., Singer, L. G., Strange, C., Nakata, K., et al. (2011). Efficacy and safety of sirolimus in lymphangioleiomyomatosis. *N. Engl. J. Med.* 364 (17), 1595–1606. doi:10.1056/NEJMoa1100391

McDermott, M. V., Afrose, L., Gomes, I., Devi, L. A., and Bobeck, E. N. (2019). Opioid-Induced signaling and antinociception are modulated by the recently deorphanized receptor, GPR171. *J. Pharmacol. Exp. Ther.* 371 (1), 56–62. doi:10.1124/jpet.119.259242

Mehra, S., Deshpande, N., and Nagathihalli, N. (2021). Targeting PI3K pathway in pancreatic ductal adenocarcinoma: Rationale and progress. *Cancers* 13 (17), 4434. doi:10.3390/cancers13174434

Menezo, Y., Clement, P., Clement, A., and Elder, K. (2020). Methylation: An ineluctable biochemical and physiological process essential to the transmission of life. *Int. J. Mol. Sci.* 21 (23), 9311. doi:10.3390/ijms21239311

Morita, M., Gravel, S. P., Hulea, L., Larsson, O., Pollak, M., St-Pierre, J., et al. (2015). mTOR coordinates protein synthesis, mitochondrial activity and proliferation. *Cell Cycle (Georget. Tex)* 14 (4), 473–480. doi:10.4161/15384101.2014.991572

Mossmann, D., Park, S., and Hall, M. N. (2018). mTOR signalling and cellular metabolism are mutual determinants in cancer. *Nat. Rev. Cancer* 18 (12), 744–757. doi:10.1038/s41568-018-0074-8

Ranek, M. J., Kokkonen-Simon, K. M., Chen, A., Dunkerly-Eyring, B. L., Vera, M. P., Oeing, C. U., et al. (2019). PKG1-modified TSC2 regulates mTORC1 activity to counter adverse cardiac stress. *Nature* 566 (7743), 264–269. doi:10.1038/s41586-019-0895-y

Shakya, M., Yildirim, T., and Lindberg, I. (2020). Increased expression and retention of the secretory chaperone proSAAS following cell stress. *Cell Stress Chaperones* 25 (6), 929–941. doi:10.1007/s12192-020-01128-7

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Shepherd, C. W., Gomez, M. R., Lie, J. T., and Crowson, C. S. (1991). Causes of death in patients with tuberous sclerosis. *Mayo Clin. Proc.* 66 (8), 792–796. doi:10.1016/s0025-6196(12)61196-3

Sovio, U., Goulding, N., McBride, N., Cook, E., Gaccioli, F., Charnock-Jones, D. S., et al. (2020). A maternal serum metabolite ratio predicts fetal growth restriction at term. *Nat. Med.* 26 (3), 348–353. doi:10.1038/s41591-020-0804-9

Stone, C. H., Lee, M. W., Amin, M. B., Yaziji, H., Gown, A. M., Ro, J. Y., et al. (2001). Renal angiomyolipoma: Further immunophenotypic characterization of an expanding morphologic spectrum. *Archives pathology laboratory Med.* 125 (6), 751–758. doi:10.5858/2001-125-0751-RA

van Steenoven, I., Koel-Simmelink, M. J. A., Vergouw, L. J. M., Tijms, B. M., Piersma, S. R., Pham, T. V., et al. (2020). Identification of novel cerebrospinal fluid biomarker candidates for dementia with lewy bodies: A proteomic approach. *Mol. Neurodegener.* 15 (1), 36. doi:10.1186/s13024-020-00388-2

Venyo, A. K. (2016). A Review of the literature on extrarenal retroperitoneal angiomyolipoma. *Int. J. Surg. Oncol.* 2016, 6347136. doi:10.1155/2016/6347136

Wang, Z., Liu, X., Liu, X., Sun, H., Guo, Z., Zheng, G., et al. (2019). UPLC-MS based urine untargeted metabolomic analyses to differentiate bladder cancer from renal cell carcinoma. *BMC cancer* 19 (1), 1195. doi:10.1186/s12885-019-6354-1

Wang, C. Y., Lempp, M., Farke, N., Donati, S., Glatter, T., and Link, H. (2021). Metabolome and proteome analyses reveal transcriptional misregulation in glycolysis of engineered *E. coli*. *Nat. Commun.* 12 (1), 4929. doi:10.1038/s41467-021-25142-0

Wang, Z., Guo, X., Wang, W., Gao, L., Bao, X., Feng, M., et al. (2021). UPLC-MS/MS-based lipidomic profiles revealed aberrant lipids associated with invasiveness of silent corticotroph adenoma. *J. Clin. Endocrinol. metabolism* 106 (1), e273–e287. doi:10.1210/clinem/dgaa708

Wang, W., Zhao, Y., Wang, X., Wang, Z., Cai, Y., Li, H., et al. (2022a). Analysis of renal lesions in Chinese tuberous sclerosis complex patients with different types of TSC gene mutations. *Genet. Mol. Biol.* 45 (2), e20200387. doi:10.1590/1678-4685-GMB-2020-0387

Wang, Z., Liu, X., Wang, W., Wei, J., Seery, S., Xu, J., et al. (2022b). A multi-omics study of diagnostic markers and the unique inflammatory tumor micro-environment involved in tuberous sclerosis complex-related renal angiomyolipoma. *Int. J. Oncol.* 61 (5), 132. doi:10.3892/ijo.2022.5422

Watt, B., van Niel, G., Raposo, G., and Marks, M. S. (2013). PMEL: A pigment cell-specific model for functional amyloid formation. *Pigment Cell Melanoma Res.* 26 (3), 300–315. doi:10.1111/pcmr.12067

Wolfson, R. L., and Sabatini, D. M. (2017). The dawn of the age of amino acid sensors for the mTORC1 pathway. *Cell Metab.* 26 (2), 301–309. doi:10.1016/j.cmet.2017.07.001

Woodrum, C., Nobil, A., and Dabora, S. L. (2010). Comparison of three rapamycin dosing schedules in A/J Tsc2+/- mice and improved survival with angiogenesis inhibitor or asparaginase treatment in mice with subcutaneous tuberous sclerosis related tumors. *J. Transl. Med.* 8, 14. doi:10.1186/1479-5876-8-14

Xian, Z. H., Cong, W. M., Lu, X. Y., Yu, H., and Wu, M. C. (2011). Angiogenesis and lymphangiogenesis in sporadic hepatic angiomyolipoma. *Pathol. Res. Pract.* 207 (7), 403–409. doi:10.1016/j.prp.2011.04.008

Xu, K. F., Zhang, P., Tian, X., Ma, A., Li, X., Zhou, J., et al. (2013). The role of vascular endothelial growth factor-D in diagnosis of lymphangioleiomyomatosis (LAM). *Respir. Med.* 107 (2), 263–268. doi:10.1016/j.rmed.2012.10.006

Yan, S., Hui, Y., Li, J., Xu, X., Li, Q., and Wei, H. (2020). Glutamine relieves oxidative stress through PI3K/Akt signaling pathway in DSS-induced ulcerative colitis mice. *Iran. J. basic Med. Sci.* 23 (9), 1124–1129. doi:10.22038/ijbms.2020.39815.9436

Yang, H., Rudge, D. G., Koos, J. D., Vaidialingam, B., Yang, H. J., and Pavletich, N. P. (2013). mTOR kinase structure, mechanism and regulation. *Nature* 497 (7448), 217–223. doi:10.1038/nature12122

Yang, J., Samsel, P. A., Narov, K., Jones, A., Gallacher, D., Gallacher, J., et al. (2017). Combination of everolimus with sorafenib for solid renal tumors in Tsc2(+/-) mice is superior to everolimus alone. *Neoplasia (New York, NY)* 19 (2), 112–120. doi:10.1016/j.neo.2016.12.008

Young, L. R., Vandyke, R., Gulleman, P. M., Inoue, Y., Brown, K. K., Schmidt, L. S., et al. (2010). Serum vascular endothelial growth factor-D prospectively distinguishes lymphangioleiomyomatosis from other diseases. *Chest* 138 (3), 674–681. doi:10.1378/chest.10-0573

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics a J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

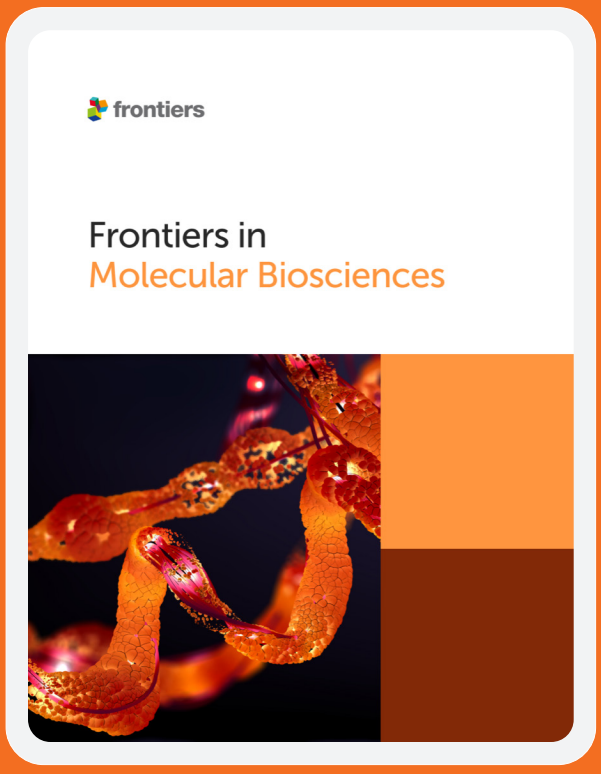
Frontiers in
Molecular Biosciences

Explores biological processes in living organisms on a molecular scale

Focuses on the molecular mechanisms underpinning and regulating biological processes in organisms across all branches of life.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact