



Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2019 July ; 2019: 243–247. doi:10.1109/EMBC.2019.8857049.

MOTA: Multi-omic integrative analysis for biomarker discovery

Ziling Fan

Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, Washington, DC

Yuan Zhou, Habtom W. Ressom

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC

Abstract

Recent advancement of omic technologies provides researchers with opportunities to search for disease biomarkers at the systems level. However, selection of biomarker candidates from a large number of molecules involved at various layers of the biological system is challenging. In this paper, we propose multi-omic integrative analysis (MOTA), a network-based method that uses information from multi-omic data to identify candidate disease biomarkers. We evaluated the performance of MOTA in selecting disease-associated molecules from four sets of multi-omic data representing three cohorts of hepatocellular carcinoma (HCC) cases and patients with liver cirrhosis. The results demonstrate that MOTA leads to selection of more biomarker candidates that shared by two different cohorts compared to traditional statistical methods. Also, the networks constructed by MOTA allow users to investigate biological significance of the selected biomarker candidates.

Keywords

multi-omic; network analysis; feature selection; liver cancer; liver cirrhosis; biomarkers

I. INTRODUCTION

Recent development in high-throughput omic technologies enables biomarker discovery from different layers of central dogma. Tremendous amount of effort has been devoted to select disease-associated biomolecules by analyzing data obtained via different ‘omic’ experiments (genomics, transcriptomics, metabolomics). However, integrative analysis of multi-omic data is a challenging task due to the complexity of the biological system.

It has been recognized that different feature selection approaches may generate different sets of biomarkers [1]. A traditional approach to biomarker selection includes statistical methods such as Student’s t-test and ANOVA, by which biomolecules with significant change in their expression level between distinct biological groups (normal vs. disease; untreated vs.

Corresponding author: hwr@georgetown.edu.

* The work presented in this paper is supported by NIH grants U01CA185188 and R01GM123766 awarded to H.W.R

treated) are selected. One obvious drawback of this category of methods is that it neglects that biomolecules in a biological system are highly intertwined and interact with each other. To this end, network-based methods have become an intuitive way to reconstruct biological networks and investigate disease-associated changes in the interaction of biomolecules at the systems level. For example, relevance network is a widely used method to model biological system due to its simplicity [2]. It measures ‘relevance’ by correlation or mutual information between two biomolecules and set a threshold to determine whether they are relevant or not. However, this method fails to distinguish direct and indirect association especially when dealing with high-dimensional omic datasets. Gaussian graphical models (GGM) are increasingly used to overcome this drawback. The advantage of these methods is they estimate the conditional dependency between two features in a dataset by removing the effect brought by others using partial correlation. In the case of large p and small n problem, which is typical omic studies, one can use graphical LASSO to directly estimate a sparse precision matrix of the sample with which partial correlation between each feature pair can be calculated. We will pursue this idea to investigate interaction of biomolecules from one omic dataset.

However, interactions between biomolecules within single layer (interactions between proteins, mRNAs, metabolites, etc.) are insufficient to depict a holistic picture of a biological system. Different biological layers are under a tightly coordinated or regulatory relationship. For example, transcription of mRNA from its DNA template is under the control of its transcription factors which generally are protein molecules. Similarly, metabolic reactions are carried out by enzymes which are also protein molecules or complexes. Therefore, an integrative framework taking into account interactions over different biological layers helps to gain a more comprehensive understanding of biological systems. Similar with idea of constructing intra-omic connections, correlation-based methods can also be applied to investigate inter-omic connections. It has been reported that multivariate methods outperform univariate correlation analysis for calculating correlation between different datasets [3]. Canonical correlation analysis (CCA) and partial least squares regression (PLS) are two commonly used multivariate approaches to explore associations of features between two omic studies [4,5]. Whereas CCA aims to find weighted linear combinations of variables maximizing the correlation between two datasets, PLS focuses on covariance rather than correlation.

In this paper, we propose a novel approach, MOTA (multi-omic integrative analysis), which prioritizes biomarker candidates from one omic dataset by integrating it with other omic datasets. The method starts by building a differential network based on changes in partial correlation (intra-omic) and canonical correlation (inter-omic) between distinct biological groups. Specifically, we use regularized version of CCA (rCCA) that allows us to deal with the large p small n problem for correlation between features in two omic datasets. A MOTA score is calculated considering both the connectivity of nodes (features) in the network and the significance level based on differential expression analysis by statistical methods. We tested MOTA using four sets of multi-omic data obtained by analysis of sera and liver tissues from three cohorts of hepatocellular carcinoma (HCC) cases and patients with liver cirrhosis (CIRR). The results show that MOTA tends to select more consistent biomarker candidates when applied to omic datasets from different cohorts.

The rest of the paper is organized as follows. Section II Introduces MOTA and multi-omic datasets used for evaluation. Section III presents the results obtained by using MOTA on different multi-omic datasets. Finally, Section IV concludes our work and proposes future directions.

II. MULTI-OMIC INTEGRATIVE ANALYSIS (MOTA)

Figure 1 shows the MOTA's framework consisting of three steps:

Step 1: calculate partial correlation (pc) using graphical LASSO for each biomolecule pair in each biological group based on metabolomic dataset; calculate differential partial correlation to determine intra-omic connections.

Step 2: calculate canonical correlation (cc) for each biomolecule pair based on metabolomic dataset and another omic dataset using rCCA; compute differential correlation to determine inter-omic connections.

Step 3: calculate MOTA score for each node in metabolomic dataset for prioritization of biomarker candidates.

A. Partial correlation calculation using graphical LASSO

Graphical LASSO is used to build sparse graphs that mimic the properties of biological networks by adding a LASSO penalty when estimating the inverse covariance matrix (i.e. precision matrix) [6]. The advantage of partial correlation, which is calculated using precision matrix, is that it removes indirect association caused by other features in the dataset. Graphical LASSO maximizes the following penalized log-likelihood

$$\log \det \Theta - \text{tr}(S\Theta) - \rho \|\Theta\|_1 \quad (\text{Eq. 1})$$

where Θ is the precision matrix; S is the sample covariance matrix, tr denotes trace, $\|\Theta\|_1$ represents the ℓ_1 norm of Θ , which is the sum of the absolute values of all elements in Θ . Precision matrix for both biological groups are calculated using graphical LASSO and partial correlation for each biomolecular pair in each biological group using Eq. 2.

$$pc_{ij} = - \frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \quad (\text{Eq. 2})$$

The change in partial correlation for each biomolecular pair between two biological groups is calculated using Eq. 3. A permutation test is used to determine the statistical significance of pc . An edge connecting two nodes will be built if pc falls into the 2.5% tails on either end of the empirical distribution curve for $\Delta \widetilde{pc}$

$$\Delta pc_{ij} = pc_{ij}^{(1)} - pc_{ij}^{(2)} \quad (\text{Eq. 3})$$

B. Canonical correlation calculation using regularized canonical correlation analysis (rCCA)

Belonging to multivariate statistical method, rCCA is widely used to associate high-dimensional omic measurements obtained from different platforms (e.g. metabolomics, transcriptomics, proteomics, etc.) [7]. Let $X = \{x_1, x_2, \dots, x_p\}$ and $Y = \{y_1, y_2, \dots, y_q\}$ denote $n \times p$ and $n \times q$ matrices, respectively; n is the number of samples; p and q are the number of variables. Furthermore, let the i th column of matrix X is denoted by X_i and the j th column of matrix Y by Y_j . We assume that the columns of X and Y are standardized (i.e. a mean of 0 and variance of 1) and $p \geq q$. rCCA computes two vectors $a^1 = \{a_1^1, a_2^1, \dots, a_p^1\}^T$ and $b^1 = \{b_1^1, b_2^1, \dots, b_q^1\}$ to maximize $\text{cor}(Xa^1, Yb^1)$:

$$\rho_1 = \text{cor}(U^1, V^1) = \max_{a, b} \text{cor}(Xa^1, Yb^1)$$

$$s.t. \text{var}(Xa^1) = \text{var}(Yb^1) = 1 \quad (\text{Eq. 4})$$

Where $U^1 = Xa^1$, $V^1 = Yb^1$ are the first canonical variates and ρ_1 is the first canonical correlation. Higher order canonical variates and canonical correlations can be computed in a similar manner. For $s = 2, 3, \dots, p$:

$$\rho_s = \max_{a, b} \text{cor}(Xa^s, Yb^s)$$

$$s.t. \text{var}(Xa^s) = \text{var}(Yb^s) = 1 \quad (\text{Eq. 5})$$

under additional constraint that $\text{cor}(Xa^s, Xa^t) = \text{cor}(Yb^s, Yb^t) = 0, 1 \leq t < s \leq p$.

In case of $n < \max(p, q)$, the sample covariance matrices are singular so their inverses are undefined. To resolve this, rCCA includes a regularization term to the diagonal elements of the covariance matrix [8]:

$$\Sigma_{XX}(\beta_1) = S_{XX} + \beta_1 I_p, \quad \Sigma_{YY}(\beta_2) = S_{YY} + \beta_2 I_q \quad (\text{Eq. 6})$$

where S_{XX} and S_{YY} are the sample covariance matrices; the regularization parameters β_1 and β_2 are determined through a standard cross validation procedure.

The relationship between X_j and Y_j is measured as described previously [9]. In MOTA, we first calculate group specific canonical correlation for each feature pair between two omic datasets for both biological groups ($cc_{ij}^{(1)}, cc_{ij}^{(2)}$). Next, the change in canonical correlation of each biomolecular pair between two biological groups are calculated using Eq. 7. We draw an edge in the resulting graph if $|\Delta cc_{ij}|$ is above a certain threshold (0.5 was used in this work).

$$\Delta cc_{ij} = cc_{ij}^{(1)} - cc_{ij}^{(2)} \quad (\text{Eq. 7})$$

C. MOTA score calculation

The network obtained by MOTA consists of intra-omic connections calculated using graph LASSO and inter-omic connections with other omic datasets calculated using rCCA. The final step in MOTA is to calculate a MOTA score for each feature (node) in the intra-omic network. Briefly, p-value (p_k) for each biomolecule is converted to z-score as shown in Eq. 8. A MOTA score for each node is defined by Eq. 9.

$$|z_k| = \Phi^{-1} \left(1 - \frac{p_k}{2} \right) \quad (\text{Eq. 8})$$

where Φ^{-1} is the inverse cumulative distribution function of the standard Gaussian distribution.

$$M_k = \sum_{d \in conn} |z_d| + |z_k| \quad k \in \text{centric dataset} \quad (\text{Eq. 9})$$

where Z_d indicates the combined z-score of all nodes which are connected to the target node k from one omic dataset calculated using Stouffer's Z-score method shown as Eq. 10 and $conn$ indicates each omic dataset to which there is at least one node connected to the target node k .

$$Z \sim \frac{\sum_{i=1}^m z_i}{\sqrt{m}} \quad (\text{Eq. 10})$$

D. Multi-omic datasets

We tested our algorithm using four sets of multi-omic data. Briefly, blood samples from 89 patients (40 HCC cases and 49 cirrhotic controls) recruited at Tanta University (TU) Hospital [10] and 84 patients (40 HCC cases and 44 cirrhotic controls) recruited at Georgetown University (GU) Hospital [11] were analyzed by targeted metabolomics, glycomics, and proteomics. These are designated as TU and GU1 cohorts, respectively, in Table 1. Additionally, blood and liver tissue samples from 65 patients (40 HCC cases and 25 cirrhotic controls) recruited at Georgetown University Hospital [12,13] were analyzed. This cohort is designated as GU2 in Table 1. Metabolomics and glycoproteomics datasets (GU2a datasets) were acquired by analysis of blood sample from each patient. Metabolomics, mRNA expression profiling, and miRNA expression profiling datasets (GU2a datasets) were acquired by analysis of liver tissue samples.

III. RESULTS AND DISCUSSION

A. Integrative analysis of TU datasets

We applied student t-test on TU metabolomics dataset to select metabolites with significant changes in their levels between HCC cases and cirrhotic controls. Also, we used MOTA to integrate the TU metabolomic dataset with proteomic and glycomic datasets. Table 2 shows metabolites ranked by t-test and MOTA. As shown in the table, ethanolamine ranked fifth by t-test but its ranking was elevated to first by MOTA. As the integrative network in Fig. 2 illustrates, the node representing ethanolamine (bottom) has one of the largest size (large MOTA score) due to high node degree (the number of connected nodes) and its connection with several low p-value metabolites such as L-glutamic acid, L-Valine2 and GDCA (intra-omic) and proteins, O75636 (Ficolin3) and P06276 (Cholinesterase) (inter-omic). In contrast to ethanolamine, lactic acid (upper left) ranks second by t-test and 12th by MOTA. From Fig. 2, we observe that lactic acid has fewer intra-omic connections compared to ethanolamine (9 vs. 16) and the connected nodes have lighter color (higher p-value). Furthermore, it does not have any inter-omic connection with proteins or glycans. As a result, its ranking decreased dramatically by MOTA compared to t-test.

Each edge (both intra-omic and inter-omic) connecting two nodes (features) in the network calculated by MOTA indicates change in correlation of the two features within two biological groups. For example, an edge may indicate that two biomolecules positively correlated in control group may have no or negative correlation in the case group. We investigated the biological significance of the edges reconstructed by MOTA as illustrated in Table 3. For example, TCDCA and TCA, which are involved in bile acids metabolism in liver, are connected because of the significant change of correlation between the HCC and cirrhotic groups. Similarly, significant change in correlation resulted in the connectivity between Ficolin3 and glycocholic acid (GCA); physical interaction between these two molecules has been reported [14]. We calculated MOTA scores under the assumption that strong candidates tend to be differentially expressed and be surrounded by differentially expressed neighbors. Taking into account the interactions of connected features, the MOTA score carries not only the statistical significance of its own and connected nodes, but also their biological significance in terms of regulatory relationship.

B. Ranking biomarker candidates via TU & GU1 datasets

One primary application of parsimonious feature selection is to achieve accurate classification of disease with limited number of features [15]. In order to make it widely applicable, an ideal feature selection algorithm is expected to generate consistent results when applied on data acquired by analysis of samples from different cohorts but the same disease condition.

Table 4 presents a comparison of feature ranking results obtained by analysis of the TU to GU1 datasets using t-test, INDEED [16] and MOTA. While INDEED uses a network-based method for ranking biomolecules from an omic study based on differential partial correlation, MOTA builds on INDEED to utilize information from multi-omic data for improved ranking. As shown in the table, ethanolamine's ranking is elevated from 5th by t-

test to 1st by MOTA compared to student t-test in TU cohort. In the GU1 cohort, ethanolamine is ranked 1st by both t-test and MOTA. The ranking of lactic acid in the TU cohort decreased from second by t-test to 12th by MOTA; lactic acid is ranked 25th in the GU1 dataset (not shown in the table). These examples demonstrate that MOTA tends to help converge the ranking results from two cohorts. As shown in Table 4, while only two common metabolites between the TU and GU1 cohorts are selected by t-test and INDEED, four common metabolites are selected by MOTA. We conclude from these results that multi-omic approach is likely to lead to more consistent ranking.

C. GU2 datasets

We tested our algorithm on GU2a and GU2b multi-omic datasets which are independent to the GU1 datasets. GU2a denotes metabolomic and glycoproteomic datasets acquired by analysis of serum samples, whereas GU2b denotes metabolomic and transcriptomic (mRNA and miRNA expression profiling) datasets acquired by analysis of liver tissues. Since a large number of statistically significant metabolites were found by t-test (Table 5), we used MOTA to rank the metabolites based on biological and statistical significance. In addition, we took advantage of metabolomic data acquired by analysis of sera and liver tissues from the same cohort to investigate potential overlaps between metabolite biomarker candidates in serum and tissue. We believe that such overlapping candidates are likely to be biologically relevant. Table 6 shows overlapping serum and tissue metabolites selected by t-test, INDEED, and MOTA. We treated as the same metabolites if the mass-to-charge ratio (m/z) difference between the analytes is within 10 ppm. While MOTA and t-test each identified three analytes that overlap between serum and tissue (two by INDEED) in the top 50-ranked list, MOTA and INDEED gave rise to two analytes overlapping between serum and tissue (one by t-test) in the top 30-ranked list. We obtained multiple putative IDs for four of the five analytes in Table 6. For example, putative IDs for $m/z=498.2896$ (negative mode) include taurodeoxycholic acid, taurodeoxycholate; and for $m/z=195.093193$ (positive mode) include methiuron, aminoacridine and benzene. By analyzing these two datasets, we demonstrated that MOTA can be used as a prioritization method when a large number of significantly changed molecules are found by statistical methods.

IV. CONCLUSION

In this paper, we introduce multi-omic integrative analysis (MOTA), a network-based method for disease biomarker discovery. Using four sets of multi-omic data representing three cohorts, we demonstrated that MOTA allows the selection of biomarker candidates shared by two cohorts compared to t-test. Also, the networks constructed by MOTA allow evaluation biological significance of biomarker candidates. Future work will focus on pathway analysis to further interpret the networks created by MOTA.

Acknowledgment

The authors wish to thank the shared resources at Georgetown University, especially the Genomics & Epigenomics Shared Resource (GESR), the Proteomics & Metabolomics Shared Resource (PMSR), and the Histopathology & Tissue Shared Resource (HTSR) for their help and support.

REFERENCES

- [1]. Armitage EG, Barbas C (2014). Metabolomics in cancer biomarker discovery: current trends and future perspectives. *J Pharm Biomed Anal*, 87:1–11. [PubMed: 24091079]
- [2]. Butte AJ, Kohane IS (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, 418–29. [PubMed: 10902190]
- [3]. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, Laurila PP, Kangas AJ, Soininen P, Savolainen MJ, Viikari J, Kähönen M, Perola M, Salomaa V, Raitakari O, Lehtimäki T, Taskinen MR, Järvelin MR, Ala-Korpela M, Palotie A, ... de Bakker PI (2012). Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS genetics*, 8(8), e1002907. [PubMed: 22916037]
- [4]. Lê cao KA, González I, Déjean S (2009). integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855–6. [PubMed: 19706745]
- [5]. Lê Cao KA, Martin PG, Robert-Granié C, Besse P (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10:34. [PubMed: 19171069]
- [6]. Friedman J, Hastie T, Tibshirani R (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–41. [PubMed: 18079126]
- [7]. Krumsiek J, Bartel J, Theis FJ (2016). Computational approaches for systems metabolomics. *Curr Opin Biotechnol*, 39:198–206. [PubMed: 27135552]
- [8]. Leurgans SE, Moyeed RA and Silverman BW (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.725–740.
- [9]. Zuo Y, Yu G, Zhang C, Ressom HW A new approach for multi-omic data integration. In: *Bioinformatics and Biomedicine (BIBM)*, 2014 IEEE International Conference On: 2014 p. 214–7. IEEE.
- [10]. Nezami Ranjbar MR, Luo Y, Di Poto C, Varghese RS, Ferrarini A, Zhang C, Sarhan NI, Soliman H, Tadesse MG, Ziada DH, Roy R, Ressom HW (2015). GC-MS based plasma metabolomics for identification of candidate biomarkers for hepatocellular carcinoma in Egyptian cohort. *PLoS One* 10(6):e0127299 [PubMed: 26030804]
- [11]. Di Poto C, He S, Varghese RS, Zhao Y, Ferrarini A, Su S, Karabala A, Redi M, Mamo H, Rangnekar AS, Fishbein TM, Kroemer AH, Tadesse MG, Roy R, Sherif ZA, Kumar D, Ressom HW (2018). Identification of race-associated metabolite biomarkers for hepatocellular carcinoma in patients with liver cirrhosis and hepatitis C virus infection. *PLoS One* 13(3):e0192748 PMID: 29538406 [PubMed: 29538406]
- [12]. Di Poto C, Wang M, Su S, Ma J, Ressom HW “Identification of glycoprotein biomarkers for hepatocellular carcinoma.” Abstract for a poster presentation at ASMS 2017 Meeting, June 4–8, 2017, Indianapolis, IN.
- [13]. Ferrarini A, Di Poto C, Varghese RS, Ressom HW “Tracking aberrant pathways in hepatocellular carcinoma using metabolomics: from tissue alterations to blood biomarkers.” Abstract for a poster presentation at Metabolomics 2015, June 29–July 2, 2015, Burlingame, CA.
- [14]. Hatakeyama T, Murakami K, Miyamoto Y, and Yamasaki N (1996) An assay for lectin activity using microtiter plate with chemically immobilized carbohydrates. *Anal. Biochem* 237, 188–192. [PubMed: 8660564]
- [15]. Buchholz M, Kestler HA, Bauer A, et al. (2005). Specialized DNA arrays for the differentiation of pancreatic tumors. *Clin Canc Res* 15:11(22):8048–54.
- [16]. Zuo Y, Cui Y, Di Poto C, Varghese RS, Yu G, Li R, & Ressom HW (2016). INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery. *Methods* 111:12–20. 10.1016/j.ymeth.2016.08.015 [PubMed: 27592383]

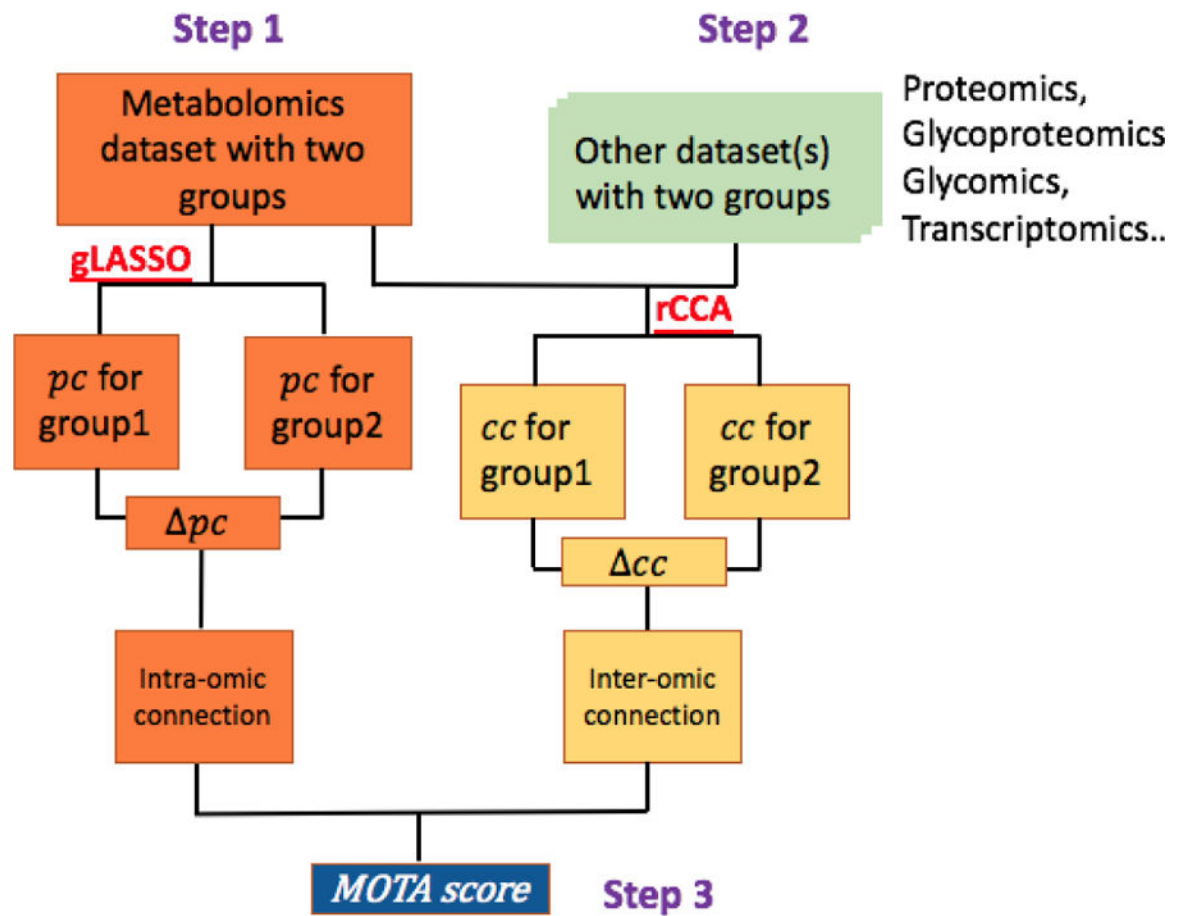


Figure 1.

Framework of MOTA, demonstrating how metabolomic data are integrated with other omic datasets to select disease-associated metabolite. pc: partial correlation, cc: canonical correlation.

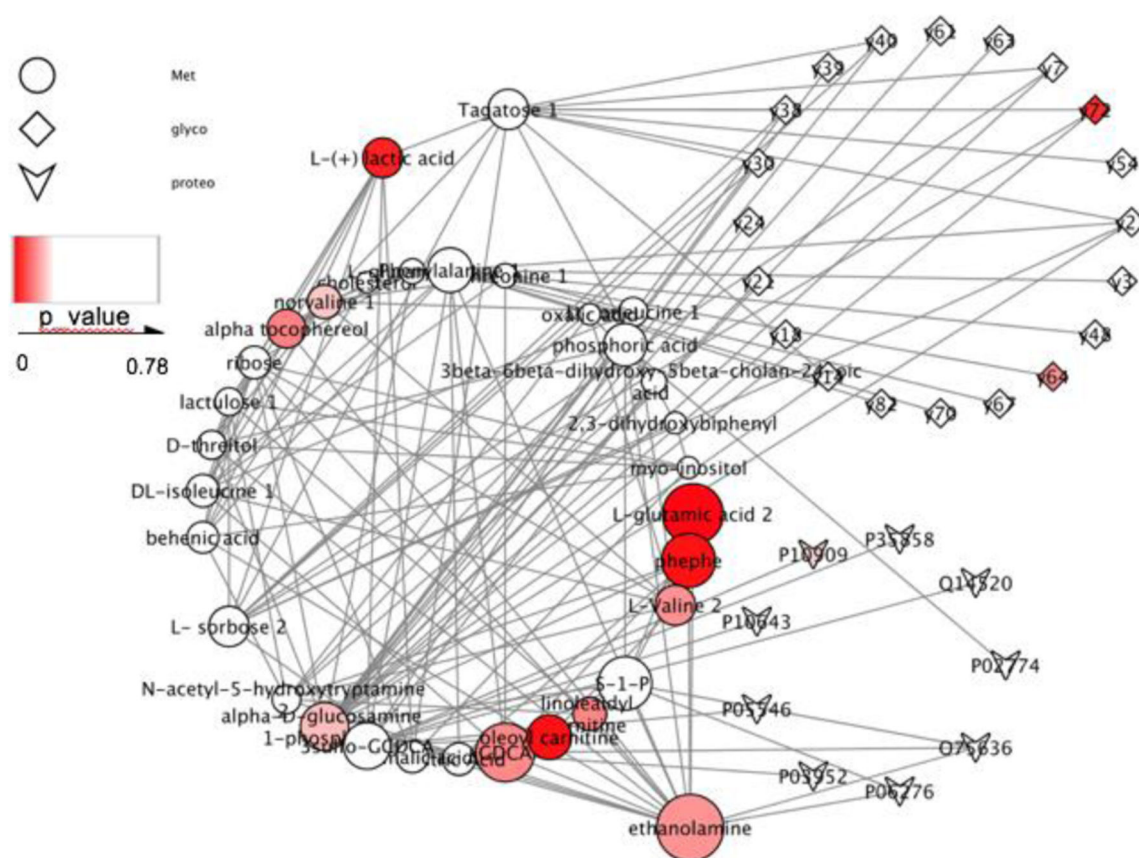


Figure. 2.
Network constructed by MOTA using TU datasets

Table 1.

MULTI-OMIC DATA ACQUIRED FROM THREE COHORTS. NUMBER OF FEATURES IN EACH OMIC DATASET AND NUMBER OF SERUM & TISSUE SAMPLES WITH MULTI-OMIC DATASETS ARE SHOWN IN PARENTHESES.

Cohorts	Datasets	Omic Studies (number of features)	Serum		Tissue	
			HCC	CIRR	HCC	CIRR
TU	CTU datasets	Metabolomics (66) Glycomics (82) Proteomics (100)	40 39	49 (48)		
GU1	GU1 datasets	Metabolomics (53) Glycomics (82) Proteomics (101)	40 (40)	44 (44)		
GU2	GU2a datasets	Metabolomics (3,150) Glycoproteomics (8,540)	40 (37)	25 (24)		
	GU2b datasets	Metabolomics (3,672) mRNA profiling (27,523) miRNA profiling (2,543)			40 (37)	25 (24)

Table 2.

RANKING OF METABOLITES IN TU DATASETS.

ID	P-Value	Rank	MOTA Score	Rank
L-glutamic acid 2	1.32E-07	1	6.7	2
L-(+) lactic acid	0.0022	2	4.3	12
alpha tocophereol	0.0086	3	4.2	13
L-Valine2	0.0094	4	4.4	11
ethanolamine	0.0097	5	7.4	1
glucosamine 1-phosphate	0.0123	6	5.3	3
norvaline 1	0.0130	7	3.6	16
Citric Acid	0.0227	8	3.5	20
L-norleucine 1	0.0292	9	3.1	26
L- sorbose 2	0.0384	10	4.4	10
Tagatose 1	0.0389	11	4.4	9
behenic acid	0.0444	12	3.4	22
D-malicacid	0.0897	13	3.4	21
DL-isoleucine 1	0.0913	14	3.5	18
cholesterol	0.1031	15	2.2	43
2,3-dihydroxybiphenyl	0.1073	16	2.3	36
Phenylalanine 1	0.1224	17	4.9	5
D-threitol	0.1479	18	3.1	27
lactulose 1	0.1591	19	3.0	29
phosphoric acid	0.1685	20	4.7	6
hydroxytryptamine 2	0.1821	21	3.0	28
L-glutamine 2	0.1878	22	2.6	32
ribose	0.1908	23	3.5	18
myo-inosrtol	0.1939	24	2.3	37
L-threonine 1	0.2014	25	2.5	33
oxalic acid	0.2041	26	2.2	40
diglycerol 2	0.2049	27	2.4	34
L-threonine 2	0.2774	28	3.7	15
L-mimosine 2	0.2825	29	5.3	4
urea	0.2971	30	3.1	25
6-hydroxy caproic acid	0.3156	31	2.1	44

Table 3.

EXAMPLES OF BIOLOGICAL SIGNIFICANCE OF EDGES CALCULATED BY MOTA USING THE TU DATASETS.

Category	Compound	Correlation in CIRR	Correlation in HCC	Differential Correlation	Significant	Note
Inter-omic Network (metabolomics)						
1	tyramine vs. L-Valine2	0.11	0.11	0	X	-
2	Lactulose vs. hydroxytryptamine	-0.047	-0.05	0.01	X	-
3	TCDCA vs. TCA	0.09	0.39	-0.29	✓	Bile acids metabolism in liver
4	L-threonine 1 vs. Glycine	0.20	0	0.20	✓	threonine is a precursor of glycine
Inter-omic Network (metabolomics & other omics datasets)						
4	Ficlin3 vs. GCA	0.50	0.02	0.58	✓	Ficlin3 has Lectin activity; binds with sugar part of other molecules
4	Cholinesterases vs. ethanolamine	0.54	0.03	0.61	✓	Cholinesterases bind with glycopospholipids through ethanolamine

Table 4.

THE NUMBER OF BIOMAKER CANDIDATES OVERLAPPING BETWEEN THE TU AND GU1 DATASETS RANKED USING T-TEST, INDEED AND MOTA.

Ranking Using student t-test (p-value)		
	GU1 Cohort	TU Cohort
1	ethanolamine	L-glutamic acid 2
2	phenylalanine	L-(+) lactic acid
3	sorbose	alpha tocophereol
4	L-pyrogutamic acid	L-Valine 2
5	glycine	ethanolamine
6	linoleic acid	alpha-D-glucosamine 1-phosphate
7	Creatinine	norvaline 1
8	lauric acid	Citric Acid
9	ribitol /arabitol	L-norleucine 1
10	D-threitol	sorbose
Ranking Using INDEED		
	GU1 Cohort	TU Cohort
1	phenylalanine	L-glutamic acid 2
2	L-pyrogutamic acid	ethanolamine
3	ethanolamine	L-Valine 2
4	Creatinine	L-(+) lactic acid
5	D-threitol	alpha tocophereol
6	glycine	alpha-D-glucosamine 1-phosphate
7	diglycerol	norvaline 1
8	lauric acid	DL-isoleucine 1
9	trans-aconitic acid	ribose
10	ribose	Citric Acid
Ranking Using MOTA		
	GU1 Cohort	TU Cohort
1	ethanolamine	ethanolamine
2	phenylalanine	L-glutamic acid 2
3	L-phrogutamic acid	alpha-D-glucosamine 1-phoshate
4	glycine	L-mimosine 2
5	Creatinine	phenylalanine
6	D-threitol	phosphoric acid
7	DL-isoleucine	Creatinine
8	L-homoserine	L-Proline
9	2,3-dihydroxybiphenyl	Tagatose 1
10	L-proline	L-sorbose 2

Table 5.NUMBER OF STATISTICALLY SIGNIFICANT FEATURES OBTAINED BY ANALYSIS OF GU_{2A} & GU_B DATASETS.

		Metabolomics	Glycoproteomics	
Serum	Total No. of detected features	3,150	8,540	
	Statistically significant features (p<0.05)	1,336	51	
Tissue		Metabolomics	Transcriptomics	
			mRNA	miRNA
	Total No. of detected features	3,672	27,523	2,543
	Statistically significant features (p<0.05)	786	549	125

Table 6.

OVERLAPPING M/Z VALUES OF SERUM AND TISSUE METABOLITES IN THE TOP 30-RANKED AND TOP 50-RANKED CANDIDATES SELECTED BY T-TEST AND MOTA FROM THE GU2A AND GU2B DATASETS (N: NEGATIVE MODE; P: POSITIVE MODE).

	Student t-test		INDEED		MOTA	
	Serum	Tissue	Serum	Tissue	Serum	Tissue
Top 30	464,2832 (p)	484,2826 (p)	498.2896 (N)	498.2886 (N)	299.1391 (P)	299.1396 (P)
			781.5725 (P)	718.5726 (P)	781.5725 (P)	781.5726 (P)
	Serum	Tissue	Serum	Tissue	Serum	Tissue
Top 50	464.2832 (P)	464.2829 (P)	464.2829 (N)	464.2886 (N)	299.1391 (P)	299.1397 (P)
	498.2896 (N)	498.2886 (N)	718.5725 (P)	718.5726 (P)	718.5725 (P)	718.5726 (P)
	299.1391 (P)	299.1397 (P)			195.0932 (P)	195.0896 (P)