Published in final edited form as:

IEEE J Biomed Health Inform. 2016 September ; 20(5): 1225–1231. doi:10.1109/JBHI.2016.2574201.

Integrative Analysis of Proteomic, Glycomic, and Metabolomic Data for Biomarker Discovery

Minkun Wang,

Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA

Guoqiang Yu, and

Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA

Habtom W. Ressom* [Senior Member, IEEE]

Department of Oncology, Georgetown University, Washington, DC 20057, USA

Abstract

Studies associating changes in the levels of multiple biomolecules including proteins, glycans, glycoproteins, and metabolites with the onset of cancer have been widely investigated to identify clinically relevant diagnostic biomarkers. Advances in liquid or gas chromatography mass spectrometry (LC-MS, GC-MS) have enabled high-throughput qualitative and quantitative analysis of these biomolecules. While results from separate analyses of different biomolecules have been reported widely, the mutual information obtained by partly or fully combining them has been relatively unexplored. In this study, we investigate integrative analysis of proteins, N-glycans, and metabolites to take advantage of complementary information to improve the ability to distinguish cancer cases from controls. Specifically, SVM-RFE algorithm is utilized to select a panel of proteins, N-glycans, and metabolites based on LC-MS and GC-MS data previously acquired by analysis of blood samples from two cohorts in a liver cancer study. Improved performances are observed by integrative analysis compared to separate proteomic, glycomic, and metabolomic studies in distinguishing liver cancer cases from patients with liver cirrhosis.

Index Terms

multi-omic data integration; machine learning; systems biology; cancer biomarker discovery

I. INTRODUCTION

Characterizing the association of biomolecules such as proteins, glycans, glycoproteins, and metabolites with cancer has proven to be a promising strategy to discover candidate biomarkers. Glycosylation is one of the most common post-translational modifications of proteins. Altered patterns of glycosylation have been associated with various diseases and many currently used cancer biomarkers. In particular protein glycosylation is relevant to liver pathology because of the major influence of this organ on the homeostasis of blood glycoproteins. Characterizing glycan modifications of proteins in complex proteomes is

*Corresponding author: Habtom W. Ressom (hwr@georgetown.edu).

challenging as glycosylation can occur on multiple sites of peptides involving the attachment of different glycans to each site. An alternative strategy to the analysis of glycoproteins is the study of proteins and protein-associated glycans [1, 2]. Metabolites are molecular fingerprints of what cells do at a particular point in time; they can reveal early signs of cancers when the chances for cure are highest. Because these biomolecules are members of strongly intertwined biological pathways and are highly interactive with each other, integrative analysis offers a great opportunity to help interpret such interactions and to identify reliable biomarkers.

We previously performed separate analyses of proteins and N-linked glycans released from proteins in blood by using liquid chromatography coupled with mass spectrometry (LC-MS) [3]. Also, we used gas chromatography coupled with mass spectrometry (GC-MS) to analyze metabolites in blood [4]. We detected proteins, N-glycans, and metabolites significantly altered in hepatocellular carcinoma (HCC) cases compared to patients with liver cirrhosis using univariate statistical methods. However, multivariate statistical or machine learning methods are desirable to improve the ability to discriminate the cases from controls by taking advantage of the mutual information within the molecules detected by a single omic study as well as the combination of molecules from multiple omic studies. The integrative analysis will allow us to investigate if the synergy of the three omic studies leads to improved performance in distinguishing cases from controls compared to the a single omic study. We recently reported improvement achieved in discriminating HCC cases from cirrhotic controls using a panel of proteins and N-glycans selected by integrating proteomic and glycomic datasets [5].

In this paper, we consider three datasets we previously generated by proteomic, glycomic, and metabolomic analysis of blood samples from HCC cases and patients with liver cirrhosis to identify proteins, N-glycans, and metabolites that are significantly altered in HCC versus cirrhosis. The goal of this research is to evaluate the improvement in disease classification achieved by integrating the data from the three studies. To select multi-omic based features that lead to highly discriminant classification, we used a model, in which feature selection and classification methods are embedded. To accomplish this, we chose support vector machine-recursive feature elimination (SVM-RFE) [6] due to its wide application and flexibility to use as an embedded method that helps recognize relevant patterns in the feature space, while reducing dimensionality to overcome the risk of overfitting. Through a 10-fold cross-validation, we evaluated the classification performances of the features selected from each omic study as well as the combined features. In addition, we split the samples into training and test sets to evaluate the performance of the selected features on an independent set. We observed that improved performances can be achieved through the integrative analysis compared to a single omic study.

The remaining part of this paper is organized as follows. Section II briefly summarizes the experimental design used for acquisition of proteomic, glycomic, and metabolomic datasets. Also, this section describes our feature selection and disease classification methods based on datasets acquired by the three omic studies. Section III presents the results we obtained in selecting optimal features from each omic study as well as the integrated multi-omic dataset. Section IV concludes the paper with summary and future goals.

II. Materials and Methods

A. Experimental Design

The proposed integrative analysis is performed on LC-MS-based proteomic and glycomic datasets and GC-MS-based metabolomic dataset we acquired by analysis of blood samples from HCC cases and patients with liver cirrhosis recruited in Egypt and the U.S. [7, 8]. The participants in Egypt and the U.S. were recruited through protocols approved by the Ethics Committee at Tanta University Hospital and the Institutional Review Board at Georgetown University, respectively. Specifically, adult patients were recruited from the outpatient clinics and inpatient wards of the Tanta University Hospital (TU cohort) in Tanta, Egypt and from the hepatology clinics at MedStar Georgetown University Hospital (GU cohort) in Washington, DC, USA. The TU cohort consists of a total of 89 subjects (40 HCC cases and 49 patients with liver cirrhosis), and the GU cohort comprises of 116 subjects (57 HCC cases and 59 patients with liver cirrhosis).

Fig. 1 depicts the overall workflow of our experimental design. Briefly, targeted quantitative analysis of selected proteins and N-glycans in blood samples was performed by multiple reaction monitoring (MRM) using a Dionex 3000 Ultimate nano-LC system (Dionex Sunnyvale, CA) interfaced to TSQ Vantage mass spectrometer (Thermo Scientific, San Jose CA). The targets were selected from our previous LC-MS-based untargeted proteomic and glycomic analyses and by text mining. Also, metabolites selected from a previous untargeted study were subjected for a targeted analysis in blood samples by selected ion monitoring (SIM) using an Agilent 7890A GC interfaced to a single quadrupole Agilent 5975C MSD (Agilent Technologies, Santa Clara, CA). The datasets from these omic studies were analyzed using Skyline [9], GPA [10], and SIMAT [11], respectively. Results from univariate statistical analysis have been previously reported in [4, 7, 8]. In the following, we introduce how we integrate the three datasets for feature selection that lead to improved performance on disease classification.

B. Feature Selection and Classification

Feature selection techniques can be generally organized into three categories: filter, wrapper, and embedded methods [12]. Filter methods are efficient and scalable to high-dimensional data analysis however they ignore feature dependencies and the interaction with the classifiers. Wrapper methods consider the model hypothesis search within the feature subset selection. A common drawback of these methods is that they have a higher risk of overfitting issue than filter methods and are very computationally intensive. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally costly than wrapper methods. Because a thorough comparison among various feature selection methods and classifiers or determination of the most suitable ones is not the primary goal of this paper, we chose an embedded method implemented in SVM-RFE due to its wide application and flexibility for high dimensional data. Linear SVMs were trained to classify samples in case and control groups using features from each of the three omic studies (proteomics, glycomics, and metabolomics) separately and by combining features from the three. Equation (1) presents the decision function in SVM model for an input sample x_i .

$$D(\mathbf{x}_t) = \mathbf{w} \cdot \mathbf{x}_t + b, \text{ where}$$

$$\mathbf{w} = \sum_k \alpha_k y_k \mathbf{x}_k \text{ and } b = \langle y_k - \mathbf{w} \cdot \mathbf{x}_k \rangle. \quad (1)$$

The feature weight vector $\mathbf{w}$ determined by support vectors is used as feature ranking criterion by the recursive feature elimination (RFE) algorithm [6]. SVM-RFE eliminates redundant features iteratively and yields better and more compact feature subsets. The major steps include 1) training the SVM classifier; 2) ranking the features according to weight vector $\mathbf{w}$ of the learned SVM; 3) eliminating features with the smallest ranking criterion; 4) retraining SVM model with the remaining features; 5) estimating the performance of the model using cross-validation to check if the optimal subset is obtained. In this paper, we applied SVM-RFE to select highly discriminative sets of proteins, N-glycans, and metabolites as well as features selected from an integrated set consisting of proteins, N-glycans, and metabolites. At each iteration, we started from the entire feature list, trained an SVM classifier with linear kernel, and estimated the average classification accuracy based on a 10-fold cross-validation. The feature with minimum weight assigned by the classifier was removed at the end of each iteration until the feature subset was empty. Additionally, we split the samples into training and test sets. The performance of the features selected using the training set were evaluated on the test set.

III. Results and Discussion

A. Integrative Analysis of Proteins and N-Glycans

We first perform the integrative analysis between proteomic and glycomic datasets with regards to their biological relations (i.e., glycosylation). Additional integration of metabolomic dataset is evaluated in the second part to further elucidate of the benefit of integrative analysis.

Datasets from targeted analyses of 101 proteins (represented by Uniprot IDs) and 82 N-glycans (characterized by the number of five monosaccharides: GlcNAc, mannose, galactose, fucose, and NeuNAc) were considered here for integrative analysis. We used SVF-RFE to select the most relevant features based on analysis of the two separate datasets obtained from targeted analysis of proteins and N-glycans and a third dataset obtained by concatenating the two datasets. Fig. 2a depicts the distributions of the LC-MS datasets from the proteomic and glycomic studies. Fig. 2b presents the log-transformed datasets that resemble normal distributions. To make the two datasets compatible for integration, we performed Z-score normalization (Fig. 2c). This step ensures features from protein and glycan lists are treated equally in the feature selection procedure by SVM-RFE.

Separate SVM-RFE models were trained for each of the three datasets. We started from the whole feature list in each dataset, and eliminated one feature in each iteration step till feature set was empty. At each step, we randomly partitioned the samples into 10 subsets. We tested

the performance of classifying one of the 10 subsets using the SVM classifier trained based on the other nine subsets. The average classification performance (i.e., accuracy, sensitivity, and specificity) was evaluated at each iteration step.

Figs. 3a and 3b depict the classification accuracy achieved at each iteration step for the top 50 features selected from the three datasets in the TU and GU cohorts, respectively. Also, the figures show the optimal number of features that leads to the best classification accuracy. We observed that, in most iteration steps, features selected from the integrated dataset yield higher accuracies compared to the same number of features selected from either the glycomic or proteomic dataset.

Receiver operating characteristics (ROC) curves were estimated by varying the SVM threshold parameter ($y_f = \hat{\mathbf{w}} \cdot \mathbf{x} - \hat{b}$). The 95% confidence intervals of area under the ROC (AUC) were calculated using bootstrap method with 1000 resampled replicates. Table I shows the disease classification performance with optimal subset of features in each dataset of the TU cohort. As shown in the table, SVM-RFE selected 29 out of 82 N-glycans and 15 out of 101 proteins as the optimal number of features. Among these, 13 glycans and 5 proteins were also selected as significantly altered in cases versus controls through univariate statistical test [7, 8]. Out of 183 integrated features, 7 proteins and 2 N-glycans in a panel were selected by SVM-RFE. The panel includes 2 that were also found significant in the univariate statistical analysis. The integrative analysis led to a significantly smaller number of features with a slight improvement on the disease classification accuracy compared to those selected by analysis of individual datasets. This phenomenon is observed consistently across the entire iteration steps, as illustrated in Fig. 3a.

Similar results are obtained in the GU cohort (Table II), in which SVM-RFE selected 18 proteins and 5 N-glycans in a panel yielded better performance than 22 proteins or the 8 glycans selected by analysis of individual datasets. Among the 23 features selected by the integrative analysis, four N-glycans and 10 proteins were also reported as significant by univariate statistical analysis. As shown in Fig. 3b, the integrative analysis yielded improved performance compared to the analysis based on the individual datasets in the majority of the iteration steps. In both cohorts, we captured features with synergic contributions to the discrimination, which provide complementary information to univariate analysis. Although we did not observe overlapping features between the optimal sets of features in the two cohorts, we were able to achieve AUCs greater than 0.73 when we trained SVMs based on the data the integrated panel learned from TU cohort and tested it on the GU cohort, and vice versa. In addition, we investigated the performance for each dataset by setting the feature size to five. We compared the performances of the best five features selected by SVM-RFE from each of the three datasets. While the integrative analysis outperformed the analysis based on individual dataset in TU cohort (Table III), both the integrated features and the protein features led to similar performances in the GU cohort (Table IV).

B. Integrative Analysis of Proteins, N-Glycans, and Metabolites

We present here the improvement in disease classification by including a dataset from a targeted analysis of 50 metabolites in blood samples. Thus, a total of 233 features (101 proteins, 82 N-glycans, and 50 metabolites) were considered for integrative analysis. The

same normalization method was applied when merging features from the new dataset. Table V presents the performance of features selected by SVM-RFE from the metabolites only and the improvement achieved by combining the metabolites with proteins and glycans in the TU cohort. From the 50 metabolites, SVM-RFE selected 14 that showed better performance than those selected from the protein and N-glycan list presented in Table I on the same TU cohort representing 89 participants. A panel consisting of 10 proteins, 5 glycans, and 6 metabolites selected from the integrated dataset outperformed all other panels selected by SVM-RFE from single omic dataset or by combining proteomic and glycomic datasets.

Fig. 4a shows the classification accuracy at each iteration step for the top 50 features from three single datasets and two integrated datasets. We observe that the two integrated datasets (colored in red and magenta) have overall higher classification accuracies than any of the single omic datasets. Although the addition of metabolites to proteins and N-glycans did not improve the classification accuracy when relatively smaller number of features are selected, a more stable and discriminative performance is achieved as the feature size increases. We also evaluated the classification performance of the list concatenated from three feature subsets selected by SVM-RFE separately (i.e., 29 N-glycans, 15 proteins, and 14 metabolites). This approach resulted in a classification accuracy of 0.79 and an AUC of 0.94 with 95% CI at (0.86, 0.97), which is worse than the performance of 21 features selected by combining the three omic datasets prior to application of SVM-RFE as presented in TABLE V.

We performed integrative analysis of proteomic, glycomic, and metabolomics datasets acquired by analysis of blood samples from 44 subjects in the GU cohort. Since the number of overlapping samples in the three omic datasets is different from the number of overlapping proteomic and glycomic datasets reported in Tables II and IV, we repeated all multivariate analyses for appropriate comparison. Table VI presents the performances of features selected from each of the three datasets as well as two integrated datasets. A panel of 10 features consisting of 4 proteins, 3 N-glycans, and 3 metabolites led to the best performance. Seven of these 10 features were also reported previously to have shown statistically significant changes in HCC vs. cirrhosis [4, 7, 8]. As illustrated in Fig. 4b, features selected from integrated datasets tend to have the best classification accuracy in most iterations. Integration of metabolites with proteins and N-glycans improves the classification accuracy as the number of features increases. Concatenating the three proteins and ten N-glycans, with the four metabolites selected independently from each omic dataset achieves a classification accuracy of 0.97 and an AUC of 0.99, which is about the same performance obtained by the ten features selected from the integrated omic dataset in the GU cohort, which resulted in accuracy of 0.98 and AUC of 0.99.

We would like to emphasize that the performance evaluations presented in Tables V and VI represent the average 10-fold cross-validation results based on all samples in each cohort. Less sensitivity and specificity are expected when the selected features are tested on an independent set due to potential overfitting issue. To address this, we evaluated the model performance by using 70% of samples (balanced in case and control groups) as a training set and the remaining 30% as a testing set. When selecting features, we used the same 10-fold cross-validation on the 70% of samples (nine subsets for training, the remaining one for

validation). Though the classification accuracies on the testing set using selected features, decrease in both cohorts, improved classification performance is observed by using the integrative analysis compared to a single omic study (Table VII).

IV. Conclusion

In this study, we investigated the benefit of an integrative analysis of proteomic, glycomic, and metabolomic datasets in improving our ability to distinguish HCC cases from patients with liver cirrhosis. Through SVM-RFE, a panel of features was selected from 101 proteins, 82 N-glycans, and 50 metabolites acquired by targeted analysis of blood samples using LC-MS and GC-MS. Complementary to univariate statistical methods, the integrative analysis utilizes mutual information among features to select a panel of features with improved ability to discriminate biologically distinct groups. In this study, we observe that features selected by merging the proteomic, glycomic, and metabolomic datasets lead to better disease classification accuracy compared to those selected from one or two of the three datasets. We would like to emphasize that the improvement achieved by the integrative analysis was observed not only in using SVM-RFE, but also through other methods such as a sequential feature selection coupled with quadratic discriminant analysis. We believe that integration of multi-omic data by multivariate statistical or machine learning methods, combined with pathway-centric and network-based approaches, will help not only in identifying a panel of biomarkers that leads to improved diagnosis but also in gaining insight into the molecular mechanisms of cancer.

Acknowledgments

Research supported by NIH Grants R01CA143420 and R01GM086746.

References

1. Fuster MM, Esko JD. The sweet and sour of cancer: Glycans as novel therapeutic targets. *Nat. Rev. Cancer*. 2005; 5(7):526–542. [PubMed: 16069816]
2. Blomme B, Van Steenkiste C, Callewaert N, Van Vlierberghe H. Alteration of protein glycosylation in liver diseases. *J. Hepatol*. 2009; 50(3):592–603. [PubMed: 19157620]
3. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat. Clin. Pract. Oncol*. 2008; 5(10):588–599. [PubMed: 18695711]
4. Nezami Ranjbar MR, Luo Y, Di Poto C, Varghese RS, Ferrarini A, Zhang C, Sarhan NI, Soliman H, Tadesse MG, Ziada DH, Roy R, Ressim HW. GC-MS based plasma metabolomics for identification of candidate biomarkers for hepatocellular carcinoma in Egyptian cohort. *PLoS One*. 10(6):e0127299.
5. Wang, M., Yu, G., Ressim, HW. Engineering in Medicine and Biology Society (EMBC), 37th Annual International Conference of the IEEE. IEEE; 2015. Integrative analysis of LC-MS based glycomic and proteomic data; p. 8185-8188.
6. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach. Learning*. 2002; 46(1–3):389–422.
7. Tsai T, Wang M, Di Poto C, Hu Y, Zhou S, Zhao Y, Varghese RS, Luo Y, Tadesse MG, Ziada DH, Ressim HW. LC-MS profiling of N-glycans derived from human serum samples for biomarker discovery in hepatocellular carcinoma. *Journal of Proteome Research*. 2014; 13(11):4859–4868. [PubMed: 25077556]
8. Tsai T, Song E, Zhu R, Di Poto C, Wang M, Luo Y, Varghese RS, Tadesse MG, Ziada DH, Desai CS, Shetty K, Mechref Y, Ressim HW. LC-MS/MS based Serum Proteomics for Identification of

Candidate Biomarkers for Hepatocellular Carcinoma. *Proteomics*. 2015; 15(13):2369–2381. [PubMed: 25778709]

9. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010; 26(7):966–968. [PubMed: 20147306]
10. Wang, M., Yu, G., Mechref, Y., Resson, HW. IEEE International Conference on Bioinformatics and Biomedicine Workshop (BIBM 2013). Shanghai, China: 2013 Dec. GPA: An algorithm for LC/MS based glycan profile annotation.
11. Nezami Ranjbar MR, Di Poto C, Wang Y, Resson HW. SIMAT: GC-SIM-MS data analysis tool. *BMC Bioinformatics*. 2015;16–259. [PubMed: 25591662]
12. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007; 23(19):2507–2517. [PubMed: 17720704]

Biographies

Minkun Wang received the B.S. degree in electrical engineering from University of Science and Technology of China, Hefei, China, in 2012. He is currently working toward the PhD degree in the Department of Electrical and Computer Engineering at Virginia Tech. He is also a research assistant at the Lombardi Comprehensive Cancer Center, Georgetown University. His research focuses on applications of statistical and machine learning methods for omic data analysis including LC-MS data preprocessing, multi-omic data integration, and deconvolution of heterogeneous data.

Guoqiang Yu received the B.S. degree in electronic engineering from Shandong University, Shandong, China in 2001, the M.S. degree in electrical engineering from Tsinghua University, Beijing, China in 2004 and the Ph.D. degree in electrical engineering from Virginia Tech in 2011. He is currently an assistant professor in Department of Electrical and Computer Engineering at Virginia Tech. His research interests include machine learning, signal and image processing, applied statistics, and their applications to developing bioinformatics and systems genetics tools for integrated modeling and analyses of various human diseases.

Habtom W. Resson received B.Sc. and M.Sc. degrees in Electrical Engineering from Addis Ababa University, Addis Ababa, Ethiopia in 1989 and 1992, respectively, and a Ph.D. degree in Electrical Engineering from University of Kaiserslautern, Kaiserslautern, Germany, in 1999. He is currently a Professor in the Department of Oncology and the Director of the Genomics and Epigenomics Shared Resource at Georgetown University Medical Center, Washington, DC, USA. His research interests focus on cancer biomarker discovery using multi-omic approaches.

Figure 1.
Workflow of integrative analysis of multi-omic data.

Figure 2.

The distributions of raw glycomic (orange) and proteomic (cyan) datasets (a); log-transformed data (b); data after log-transformation and Z-score normalization.

Figure 3. Classification accuracy at each iteration step for the top 50 features from glycomic (green), proteomic (blue), and integrated datasets (red) in the TU and GU cohorts. The optimal numbers of features (indicated by triangles) correspond to the best classification accuracy (indicated by circles).

Figure 4. Classification accuracy at each iteration step for the top 50 features from proteomic (blue), glycomic (green), metabolomic (yellow), integrated proteomic and glycomic (red), and integrated proteomic, glycomic, and metabolomic (matenga) datasets in the TU and GU cohorts.

TABLE I

Performance Comparison Based on the Optimal Number of Features Selected in the TU Cohort

TU Cohort	<i>Glycomic</i>	<i>Proteomic</i>	<i>Integrated (P & G)</i>
Accuracy	0.77	0.84	0.87
Sensitivity	0.82	0.83	0.90
Specificity	0.75	0.84	0.85
AUC (95% CI)	0.87 (0.78, 0.93)	0.93 (0.83, 0.97)	0.92 (0.81, 0.97)
Optimal Number of Features	29/82	15/101	9/183
Selected Features ^b	[25000]	[43000] ^a	P01024 ^a
	[53111] ^a	[34100] ^a	P02743
	[63402] ^a	[43202] ^{a,c}	P02750
	[53313]	[53000] ^a	P02753 ^a
	[53323]	[33101]	P02763
	[34110]	[63403] ^a	P03952
	[53311] ^c	[53311] ^c	P04004 ^a
	[43110] ^a	[53010]	P05160
	[63413] ^a	[53302] ^a	P06727
	[53411]	[34101]	P0C0L4
	[63423]	[63404] ^a	P13598 ^a
	[53312]	[29000]	P13796
	[53101] ^a	[73514]	P22891 ^a
	[53201]	[43202] ^{a,c}	P27918
	[2 10 000]		P35858

^aSignificant (p value = 0.05) in univariate statistical analysis^bN-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs^cIsomers with different retention times

The results of best performing methods are marked in bold.

TABLE II

Performance Comparison Based on the Optimal Number of Features Selected in the GU Cohort

GU Cohort	<i>Glycomic</i>	<i>Proteomic</i>	<i>Integrated (P & G)</i>
Accuracy	0.77	0.88	0.91
Sensitivity	0.79	0.86	0.89
Specificity	0.75	0.91	0.93
AUC (95% CI)	0.83 (0.71, 0.91)	0.95 (0.89, 0.98)	0.96 (0.89, 0.99)
Optimal Number of Features	8/82	22/101	23/183
Selected Features ^b		O75015 O75636 ^a	O75015 O75636^a
		P00748 ^a P01023 ^a	P01023^a P01034^a
		P01877 ^a P02741	P01877^a P02771^a
	[43100] ^a	P02766 P02771 ^a	P04278 P05155
	[53313]	P02790 P04278	P05452^a P08294
	[53000] ^a	P05155 P05452	P13796 P41222^a
	[43212]	P06727 P13796	P61626^a Q13201^a
	[53411]	P27169 ^a P41222 ^a	Q15848^a Q96KN2
	[53312]	P49747 ^a P61626 ^a	[43100]^a [53313]
	[53200]	P61769 ^a	[53000]^a [43200]^a
	[63434]	Q15848 ^a	[53411] [53200]
		Q96KN2	[53111]^a
		Q9Y6R7 ^a	

^aSignificant (p value = 0.05) in univariate statistical analysis^bN-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs

The results of best performing methods are marked in bold.

TABLE III
Performance Comparison on the Top Ranking five Featured Selected in the TU Cohort.

TU cohort	Glycomic	Proteomic	Integrated (P & G))
Accuracy	0.68	0.79	0.83
Sensitivity	0.71	0.79	0.82
Specificity	0.67	0.79	0.85
AUC (95% CI)	0.77 (0.65, 0.59)	0.88 (0.77, 0.94)	0.89 (0.80, 0.95)
Number of Selected Features	5/82	5/101	5/183

Selected
Features

- Complement C3^a
- Serum amyloid P-component
- Alpha-1-acid glycoprotein 1
- Coagulation factor XIII B chain
- Apolipoprotein A-IV

- Serum amyloid P-component
- Alpha-1-acid glycoprotein 1
- Coagulation factor XIII B chain
- Apolipoprotein A-IV

^aSignificant (p value = 0.05) proteins in univariate statistical analysis. N-glycans that found significant (p value = 0.05) in univariate statistical analysis are shown in boxes.

The results of best performing methods are marked in bold.

TABLE IV

Performance Comparison on the Top Ranking five Featured Selected in the GU Cohort.

GU cohort	Glycomic	Proteomic	Integrated (P & G)
Accuracy	0.74	0.80	0.80
Sensitivity	0.74	0.82	0.82
Specificity	0.75	0.79	0.79
AUC (95% CI)	0.82 (0.70, 0.89)	0.85 (0.74, 0.92)	0.87 (0.77, 0.93)
Number of Selected Features	5/82	5/101	5/183

Selected Features

- Ficolin-3^a
- Alpha-2-macroglobulin^a
- Plasma protease C1 inhibitor
- Tetranectin^a
- Prostaglandin-H2 D-isomerase^a

- Ficolin-3^a
- Alpha-2-macroglobulin^a
- Plasma protease C1 inhibitor
- Sex hormone binding globulin

Annotation

N-acetylglucosamine

mannose

galactose

fucose

NeuNAc

^aSignificant (p value = 0.05) proteins in univariate statistical analysis. N-glycans that found significant (p value = 0.05) in univariate statistical analysis are shown in boxes.

The results of best performing methods are marked in bold.

TABLE V

Performance Comparison Based on the Optimal Number of Features Selected in the TU Cohort

TU Cohort	<i>Metabolomics</i>		<i>Integrated (P + G + M)</i>	
Accuracy	0.86		0.90	
Sensitivity	0.91		0.91	
Specificity	0.84		0.89	
AUC (95% CI)	0.93 (0.84, 0.97)		0.99 (0.95, 0.99)	
Optimal # of Features	14/50		21/233	
Selected Features ^b	•	L-glutamic acid ^a	•	P01024 ^a
	•	L-valine ^a	•	P01591
	•	L-(+) lactic acid ^a	•	P02743 ^a
	•	N-acetyl-5-hydroxytryptamine	•	P02763
	•	L-threonine	•	P05160 ^a
	•	Diglycerol	•	P06727
	•	Urea	•	P13591
	•	Arachidic acid	•	P13598 ^a
	•	Trans-aconitic acid	•	P22891 ^a
	•	L-proline	•	P35858
	•	N, N-dimethyl-1 4-phenylenediamine	•	[43000]
	•	D-glucose	•	[53000] ^a
	•	L-serine	•	[63423]
	•	L-cystine	•	[28000]
			•	[66012] ^a
			•	L-glutamic acid ^a
			•	L-valine ^a
			•	L-(+) lactic acid ^a
			•	L-threonine
			•	Urea
			•	L-cystine

^aSignificant (p value = 0.05) in univariate statistical analysis^bN-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs

The results of best performing methods are marked in bold.

TABLE 6

Performance Comparison Based on the Optimal Number of Features Selected in the GU Cohort (44 Samples)

GU Cohort	Proteomics	Glycomics		Metabolomics	Integrated (P + G)		Integrated (P+G+M)		
Accuracy	0.89		0.91	0.84	0.98		0.98		
Sensitivity	0.94		0.87	0.85	0.95		0.96		
Specificity	0.85		0.95	0.83	0.99		0.99		
AUC (95% CI)	0.87 (0.72, 0.96)		0.97 (0.89, 0.99)	0.91 (0.77, 0.97)	0.99 (0.95, 0.99)		0.99 (0.95, 0.99)		
Optimal Number of Features	3/101	10/82		4/50	15/183		10/233		
Selected Features ^b	• O75636 ^a • P00736 • P00751 ^a	[43100] ^a		• Ethanolamine • L-(+) lactic acid ^a • Oxalic acid • Putrescine	• O75636 ^a	• [53313]	• O75636 ^a	• [53111] ^a • Malonic acid • Putrescine • Sorbose ^a	
		[53313]	[53411] ^a		• P01023 ^a	• [34110]	• P01876 ^a		
		[53000] ^a	[43201]		• P02774 ^a	• [43110] ^a	• P14151		
		[34110]	[63402]		• P04278	• [43200] ^a	• P41222 ^a		
		[53100]	[53111] ^a		• P16070	• [43201]	• [43100] ^a		
		[53302]			• P41222 ^a	• [73514]	• [43100] ^a		
					• P80108 ^a	• [53111] ^{a,c}	• [53101] ^a		

^aSignificant (p value = 0.05) in univariate statistical analysis

^bN-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs

^cIsoners with different retention times.

The results of best performing methods are marked in bold.

TABLE VII

Classification Performance on Independent Samples

Accuracy	Prote-omics	Glyc-omics	Metabol-omics	Integrated (P + G)	Integrated (P+G+M)
TU cohort	0.70	0.59	0.78	0.82	0.85
GU cohort	0.71	0.78	0.64	0.86	0.86

The results of best performing methods are marked in bold.