Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2016 August ; 2016: 3437–3440. doi:10.1109/EMBC.2016.7591467.

Multi-omic Approaches for Characterization of Hepatocellular Carcinoma

Habtom W. Ressom[1,*], **Cristina Di Poto**[1], **Alessia Ferrarini**[1], **Yunli Hu**[2], **Mohammad R. Nezami Ranjbar**[1], **Ehwang Song**[2], **Rency S. Varghese**[1], **Minkun Wang**[1], **Shiyue Zhou**[2], **Rui Zhu**[2], **Yiming Zuo**[1], **Mahlet G. Tadesse**[1], and **Yehia Mechref**[2]

[1]Georgetown University, Washington, DC

[2]Texas Tech University, Lubbock, TX

Abstract

Multi-omic approaches offer the opportunity to characterize complex diseases such as cancer at various molecular levels. In this paper, we present transcriptomic, proteomic/glycoproteomic, glycomic, and metabolomic (TPGM) data we acquired by analysis of liver tissues from hepatocellular carcinoma (HCC) cases and patients with liver cirrhosis. We evaluated changes in the levels of transcripts, proteins, glycans, and metabolites between tumor and cirrhotic tissues by statistical methods. We demonstrated the potential of multi-omic approaches and network analysis to investigate the interactions among these biomolecules in the progression of liver cirrhosis to HCC. Also, we showed the significance of multi-omic approaches to identify pathways altered in HCC.

I. Introduction

In a typical disease characterization or biomarker discovery study using high-throughput omic technologies, statistical methods are used to identify genes or proteins that are differentially expressed between two or more biologically distinct groups. However, gene or protein signatures selected by independent but similar studies tend to have only few in common [1]. This lack of reproducibility is partly due to the fact that biomolecules at different levels (e.g., genes, proteins, glycoproteins, and metabolites) are members of strongly intertwined biological pathways and are highly interactive with each other. Multi-omic approaches and network-based analysis offer the opportunity to help interpret such interactions and to characterize complex diseases. Network-based methods are increasingly applied in omic studies to gain insight into the underlying metabolic, cell signaling, protein-protein interaction (PPI), and gene regulatory networks. Also, they are utilized for integrative analysis of multi-omic data and identification of aberrant pathways [2, 3]. It was previously reported that network-based analysis of gene expression data helped characterize liver specific networks, including biological functions, signaling pathways, and transcription factors that are potentially dysregulated in hepatocellular carcinoma (HCC) [4]. The authors were able to elucidate major signaling pathways related to tumorigenicity in HCC by

*Corresponding author.

combining computational methods and functional characterization. Also, it has been demonstrated in cancer studies that the clustering of individual biomarkers through sub-networks or network motifs can improve the diagnostic accuracy and biomarker robustness [5]. Additionally, network-based analysis of omic data has been used to identify markers to predict patient prognosis [6].

In this paper, we used transcriptomic, proteomic/glycoproteomic, glycomic, and metabolomic (TPGM) data we acquired by analysis of liver tissues to compare biomolecules and pathways altered in HCC cases vs. patients with liver cirrhosis. The interactions among these biomolecules are investigated using network analysis. The paper is organized as follows. Section II describes the samples we analyzed in this study, the analytical methods and platforms used for omic data acquisition, and software tools and algorithms applied for data analysis of multi-omic data. Section III presents the results we obtained by analysis each omic dataset and by integrating multi-omic data. Finally, Section IV summarizes our work and lists some future goals.

II. Materials and Methods

A. Samples analyzed by multi-omic approaches

Human liver tissues from 10 participants recruited at MedStar Georgetown University Hospital through a protocol approved by the Georgetown IRB were considered in this study. All subjects provided signed informed consent forms. The tissues represent 5 HCC cases and 5 patients with liver cirrhosis (CIRR). Subjects in cases and controls were matched by gender, age, ethnicity and BMI (Table I). The tissues from the HCC cases include 5 tumor (HCC) and 5 adjacent cirrhotic (ADJ-CIRR) tissues. Thus, a total of 15 liver tissues were analyzed by various platforms to acquire TPGM data.

B. Acquisition and analysis of TPGM data

Transcriptomics—RNA samples were extracted from the 15 liver tissues. Following RNA quality assessment using Agilent Bioanalyzer, 14 samples that met the quality criterion were analyzed by Illumina HiSeq 2500 using 125 bp pair-end. The RNA-seq data contained an average of 33M reads per sample. The fastq files were imported into Partek Flow for quality assessment and analysis of the RNA-seq data. Alignment was performed using the spliced transcripts alignment (STAR) algorithm, which applies sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. The aligned reads were quantified to the transcriptome through an Expectation Maximization (EM) method. Differential expression analysis was performed using the gene-specific analysis (GSA) statistical model implemented in Partek. Specifically, paired and unpaired *t*-tests were applied to determine significant changes in expression in HCC *vs.* ADJ-CIRR and HCC *vs.* CIRR, respectively. We used DAVID for pathway enrichment analysis of the significant genes against a background of all detected genes.

Proteomics/glycoproteomics—We analyzed proteins/glycoproteins in the 15 liver tissues using a Dionex 3000 Ultimate nano-LC system interfaced to an LTQ Orbitrap Velos mass spectrometer. Following protein extraction and trypsin digestion, the samples were

online-purified using Acclaim PepMap100 C18 cartridge (3 μm, 100Å, Dionex). The purified samples were then separated using Acclaim PepMap100 C18 capillary column (75 μm id × 150 mm, 2 μm, 100Å, Dionex). The separation of the digests was achieved at 350 nl/min flow rate with scan time set to 120 min. The mass spectrometer was operated with two scan events. The first scan event was a full FTMS scan of 380–2000 *m/z* with a mass resolution of 15,000 at *m/z* of 400. The second scan event was collision induced dissociation (CID) MS/MS of parent ions selected from the first scan event with an isolation width of 3 *m/z*. The CID MS/MS was performed on the five most intense ions observed from the first MS scan event. The LC-MS/MS data were analyzed using MaxQuant, where the quantitation of proteins was based on ion intensity. MS/MS spectra were searched against the UniProt human protein database using Andromeda. We considered a decoy database of modified reversed protein sequences and 257 common contaminants during the search. Only proteins with more than two identified peptides were considered. The false discovery rate (FDR) was set at 0.01 for identification of peptides and proteins. Prior to the statistical analysis, identifications from reversed sequences and contaminants were removed, ion intensities were log-transformed, and missing value imputation was applied. We used Perseus, an accompanying tool to MaxQuant, for statistical analysis.

For analysis of N-linked glycopeptides in the 15 tissues, enrichment by cotton hydrophilic interaction liquid chromatography (HILIC) microcolumns was performed following protein extraction and digestion prior to analysis by the LTQ Orbitrap Velos mass spectrometer coupled with Dionex 3000 Ultimate nano-LC system. With CID mode of fragmentation, the diagnostic sugar oxonium ions, sequential neutral losses of glycosyl residues, and Y1 ions of peptide backbones were detected and utilized to infer the complete glycopeptide information. The acquired LC-MS/MS data were analyzed based on EICs using MASIC. We determined glycan sequences (i.e., topology or cartoon-graph representation) for site-specific glycopeptides using GlycoSeq [7]. Starting with a Y1 ion, the glycan topology was grown using CID fragmentation peaks while adhering to N-linked glycan synthesis rules. Peptide backbone was determined by matching mass against glycoprotein database.

Following paired and unpaired t-tests to select significantly altered proteins and glycopeptides in HCC vs. CIRR, we used DAVID for pathway enrichment analysis.

Glycomics—We analyzed N-glycans enzymatically removed from proteins extracted from the 15 tissues. The procedure includes release, puri cation, reduction, and permethylation of glycans. Permethylated N-glycans were separated by the Dionex 3000 Ultimate nano-LC system with an Acclaim PepMap C_{18} column (75 μm × 15 cm, 2 μm, 100 Å) at 55 °C. The flow rate of nanopump was set to 350 nl/min. The nano-LC system was interfaced to an LTQ Orbitrap Velos hybrid mass spectrometer, where each MS full scan (*m/z* range 500–2000) was followed by five MS/MS scans of the most intense ions. The glycomic data were analyzed using a pipeline consisting of in-house-developed algorithms and open-source software tools. Specifically, DeconTools was used for deisotoping of the mass spectra. A Savitzky-Golay filter was applied to smooth the ion trace and a first-order derivative of a Gaussian kernel was used to detect the position of the peaks. Simultaneous multiple alignment (SIMA) was applied for matching peaks detected in multiple runs. Charge and adduct states were clustered using our in-house tool for glycan profiling analysis (GPA).

Monosaccharide composition (characterized by the number of N-acetylglucosamine, mannose, galactose, fucose, and N-acetylneuraminic acid) were assigned through accurate mass matching (<2 ppm). Paired and unpaired *t*-tests were used to determine significant N-glycans in comparing HCC vs. CIRR.

Metabolomics—We used three platforms (GC-TOF-MS, GCxGC-TOF-MS and LC-QTOF-MS) for metabolomic analysis of the 15 tissues. The tissues were homogenized and metabolite extraction was performed in one single step for both GC-MS and LC-MS analyses. Briefly, 20 mg of liver tissue was homogenized with 1 mL of pre-chilled Isopropanol:Acetonitrile:Water (3:3:2) in order to extract the metabolites and precipitate the proteins. Samples were then centrifuged and the resulting supernatant was divided into two, and concentrated to dryness in speedvac. The dried samples were kept at −20°C until subsequent steps.

Prior to GC-MS-TOF MS and GCxGC-TOF-MS analysis, one aliquot was reconstituted in 500 μL of water, of which 100 μL were dried-out, lyophilize (−50°C) and derivatized in order to protect functional groups and increase the volatility and the thermostability of the analytes. Dried samples were mixed with 20 μL of MEOX reagent (20 mg/mL in pyridine) and heated at 60°C for 1 hour. This was followed by reaction with 80 μL of MSTFA at 60°C for 1 hour. The MEOX reagent was spiked with octafluoronaphthalene (OFN) to monitor system performance and fatty acid methyl esters (C8, C9, and C10 – C28 even) for retention time reference. The reaction products were cooled down to room temperature, vortex-mixed and analysed. Samples for LC-QTOF-MS analysis were directly reconstituted in 200 μL of mobile phase with spiked-in debrisoquine and 4-nitrobenzoic acid for quality assessment of positive and negative mode analysis, respectively. Metabolites were analysed using Agilent 7890 GC with dual stage modulator, MPS2 autosampler and LECO Pegasus HT, equipped with an electron ionisation source and TOF analyzer. GC-TOF-MS data were acquired using both splitless and split 10:1 injection mode to compensate for very large or very low concentration range of tissue metabolites. Similarly, GCxGC-TOF-MS data were generated using two splits ratios (20:1 and 40:1). We used ChromaTOF with True Signal Deconvolution package for data pre-processing, including baseline calculation, peak finding, deconvolution and identification. LECO's Statistical Compare software tool was used for alignment of the GC-MS data. Spectral similarity searches against the NIST and Fiehn libraries were performed to determine the identities of the analytes.

LC-MS data were acquired by analysis of the metabolite extracts using Waters ACQUITY UPLC coupled to XEVO G2 QTOF, operating in positive and negative polarity. LC-QTOF-MS data were first converted into Network Common Data Format (NetCDF) using DataBridge Program from the MassLynx software (Waters). Peak detection, alignment, and ion annotation were performed using XCMS and CAMERA. Putative metabolite identification of selected significant ions detected by LC-MS was performed using MetaboSearch, a mass-based tool we developed previously to obtain putative IDs by combining information retrieved from Human Metabolome DataBase, Madison Metabolomics Consortium Database, Metlin, and LipidMaps.

Both GC-MS and LC-MS data were normalized using total protein amount calculated by the BCA test. Paired and unpaired *t*-test were applied to determine significant changes in metabolite levels in HCC vs. CIRR. Metabolic pathway enrichment analysis was performed using MetaboAnalyst.

C. Multi-omic data analysis

In addition to searching overlaps among lists of significantly altered biomolecules and enriched pathways derived from each of the omic study, we performed pathway enrichment analysis using DAVID based on a combined list of significant genes, proteins, and glycoproteins. MetaboAnalyst was used for metabolic pathway enrichment analysis by combining significant genes and metabolites. Ingenuity Pathway Analysis Tool (IPA) was applied to identify pathways associated to a combined list of significant genes, proteins, glycoproteins, and metabolites. For network analysis of the TPGM data, we computed the Spearman correlation coefficients between each pair of variables, and performed permutation test to select the significant pairs to build networks. The results of pairwise analyses from the multi-omic data were then merged together (Figure 1). From the constructed network, we extracted relevant subnetworks guided by the statistical significance of the changes in the levels of the corresponding biomolecules between HCC and CIRR. Corresponding z-scores were obtained to assess the significance of change in intensity level for each biomolecule in the network. A combined z-score was calculated for a specific sub-network A of k biomolecules using $z_c = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} z_i$. Then, a greedy search algorithm was applied to identify sub-networks with global maximum combined z-scores by using Cytoscape plugin jActiveModules. The significance of the identified subnetworks was assessed by comparing their combined z-scores with those from random networks.

III. Results and Discussion

Transcriptomics—As shown in Table 2, from a total of 20,510 mapped genes, we identified 499 genes with statistically significant changes in expression in both HCC vs. ADJ-CIRR (paired analysis) and HCC vs. CIRR, with an FDR < 5%. Of these, 240 had a fold change > 2. Pathway enrichment analysis of the 499 genes yielded five pathways (folate biosynthesis, systemic lupus erythematosus, arginine and proline metabolism, sphingolipid metabolism, and fatty acid metabolism pathways).

Proteomics/glycoproteomics—We detected 933 proteins, of which 146 showed significant changes in ion intensities in HCC vs. CIRR ($p < 0.05$). In comparing HCC vs. ADJ-CIRR, we found 110 significant proteins cirrhosis, with 39 overlapping between the lists from the two comparisons. Also, we detected 934 glycopeptides representing 83 glycoproteins. We found 19 out of 934 glycopeptides with significant changes in HCC vs. CIRR. In comparing HCC vs. ADJ-CIRR, we found 12 significant glycopeptides. Five glycopeptides were found significant in both comparisons. Pathway analysis of the 217 proteins against a background of 933 proteins detected by LC-MS yielded six pathways (spliceosome, ECM-receptor interaction, glycine serine and threonine metabolism, focal adhesion, ribosome, and glycolysis/gluconeogenesis). From the identified 933 proteins, we

extracted 296 potential glycoproteins by searching the N-glycosylation sites (indicated by asparagine-X-serine/threonine sequons, i.e., N-X-S/T, X is any amino acid except proline). Pathway analysis of 296 glycoproteins using DAVID yielded five pathways: valine, leucine and isoleucine degration, fatty acid metabolism, glycolysis/gluconeogenesis, butanoate metabolism, and pyruvate metabolism.

Glycomics—From 43 N-glycans we detected, 11 had significant changes in ion intensities in HCC vs. CIRR, and 16 in HCC vs. ADJ-HCC, with 2 N-glycans overlapping between the two comparisons.

Metabolomics—A total of 720 and 427 ions were detected from splitless and split 10:1 GC-MS data and 6,119 and 672 from LC-MS data in the positive and negative ion mode, respectively. Table 2 presents the number of metabolites with putative identifications that showed significant changes in their levels in HCC vs. CIRR and HCC vs. ADJ-CIRR by analysis of the GC-MS and LC-MS data in different modes. These metabolites belong to various biochemical categories including acids, di-acids, amino acids, bases, sugars, phosphorylated sugars, sugar alcohols, fatty acids, nucleosides, nucleotides, mono- di- & tri-acylglycerides, lysophosphatidylglycerol lipids, glycerolipids, sphingolipids, and carnitines. Among them, amino acids and phospholipids are up-regulated in HCC vs. CIRR, whereas carboxylic acids, bile acids and long chain carnitines are down-regulated. From the combined list of significant metabolites from GC-MS and LC-MS analysis, we were able to determine the identities of 48 metabolites, which we used for metabolic pathway enrichment analysis. Pathways such as aminoacyl-tRNA biosynthesis, valine, leucine and isoleucine biosynthesis, nitrogen metabolism, valine, leucine and isoleucine degradation, and glycine, serine and threonine metabolism were identified as the top five pathways by MetaboAnalyst.

Multi-omic data analysis—Comparing the transcriptomic and proteomic data, we found 13 genes with significant changes at both transcript and protein levels. Also, we found four pathways (spliceosome, ECM-receptor interaction, glycine serine and threonine metabolism, focal adhesion) identified relevant to HCC by both significant transcripts and proteins found in comparing HCC vs. CIRR. An integrated metabolic pathway analysis was performed using MetaboAnalyst on significant genes and metabolites combined to evaluate whether the observed genes and metabolites in a particular pathway are significantly enriched. Aminoacyl-tRNA biosynthesis, valine, leucine and isoleucine biosynthesis, arginine and proline metabolism, folate biosynthesis, and alanine, aspartate and glutamate metabolism pathways were selected as the top five pathways.

A combined analysis of the significant genes, proteins, glycoproteins, and metabolites found in each omic study was performed using IPA. Specifically, 499 genes, 217 proteins, 296 glycoproteins, and 48 metabolites were mapped into IPA, which identified remodeling of epithelial adherens junctions, tRNA charging, ILK signaling, EIF2 signaling, and glycolysis pathways as the top five pathways.

Finally, we considered 90 glycosylation genes, 933 proteins, 41 glycans, and 83 glycoproteins for network analysis. We constructed an integrated network by pairwise correlation analyses (i.e., genes vs. glycans, proteins vs. glycans, proteins vs. glycoproteins,

and glycans vs. glycoproteins). Figure 2 depicts the top two significant subnetworks selected from the integrated network using a greedy search algorithm. As shown in the figure, biomolecules with smaller *p*-values are connected together due to a significant correlation between them. We believe this types of interaction are likely to help researchers generate testable hypothesis for subsequent experimental verification.

IV. Conclusion and Future Work

In this study, we investigated multi-omic approaches for characterization of HCC. Although the study was performed with small sample size, we were able to show the potential of multi-omic approaches as well as network analysis to investigate the interactions among these biomolecules in the progression of liver cirrhosis to HCC. Also, we identified pathways altered in HCC in both proteomics/glycoproteomics and metabolomics including valine, leucine and isoleucine degradation and glycine serine and threonine metabolism pathway. Future work will focus on performing multi-omic analysis on tissues and blood samples from larger cohorts and improving our approach for integrative analysis including the incorporation of prior knowledge into our data-driven network inference method, utilization of sparse generalized canonical correlation analysis (sGCCA) for a comprehensive integration of TPGM data. We anticipate these methods will enable us to verify the findings from this this pilot study and to unravel novel associations of various biomolecules with HCC.

Acknowledgments

This work is in part supported by the National Institutes of Health Grants R01CA143420 and R01GM086746 awarded to HWR.

References

1. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A. 2006; 103(15):5923–5928. [PubMed: 16585533]
2. Gibbs DL, Gralinski L, Baric RS, McWeeney SK. Multi-omic network signatures of disease. Frontiers in Genetics. 2013; 4:309.
3. Gevaert O, Villalobos V, Sikic BI, Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus. 2013; 3(4):20130013. [PubMed: 24511378]
4. Liu CH, Chen TC, Chau GY, Jan YH, Chen CH, Hsu CN, Lin KT, Juang YL, Lu PJ, Cheng HC, Chen MH, Chang CF, Ting YS, Kao CY, Hsiao M, Huang CY. Analysis of protein-protein interactions in cross-talk pathways reveals CRKL protein as a novel prognostic marker in hepatocellular carcinoma. Mol Cell Proteomics. 2013; 12(5):1335–1349. [PubMed: 23397142]
5. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007; 3:140. [PubMed: 17940530]
6. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009; 27(2):199–204. [PubMed: 19182785]
7. Song E, Mayampurath A, Yu C, Tang H, Mechref Y. Glycoproteomics: Identifying the glycosylation of prostate specific antigen at normal and high isoelectric points by LC–MS/MS. J Proteome Res. 2014; 13(12):5570–5580. [PubMed: 25327667]

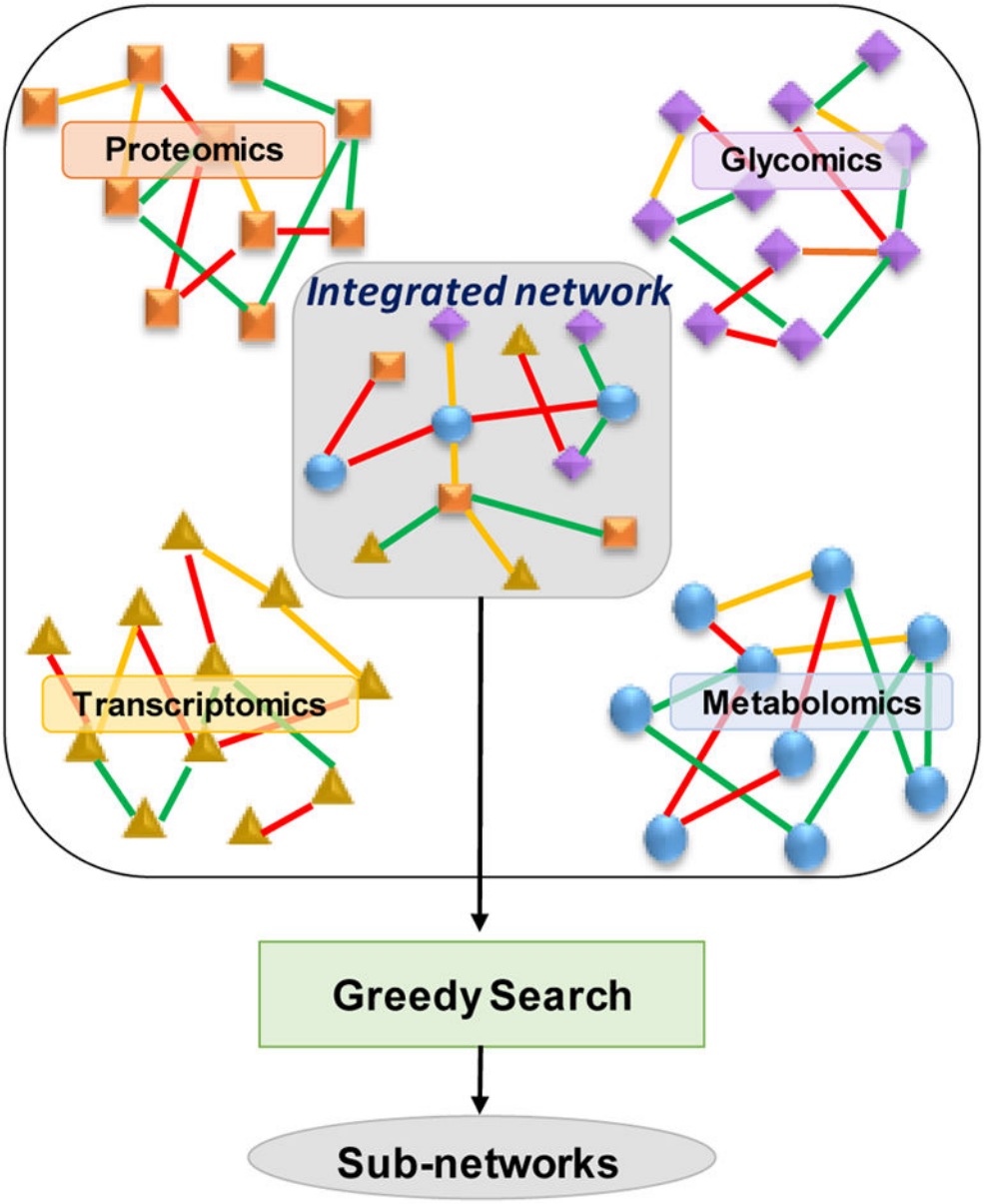

Figure 1.
Network-based integration of TPGM data.

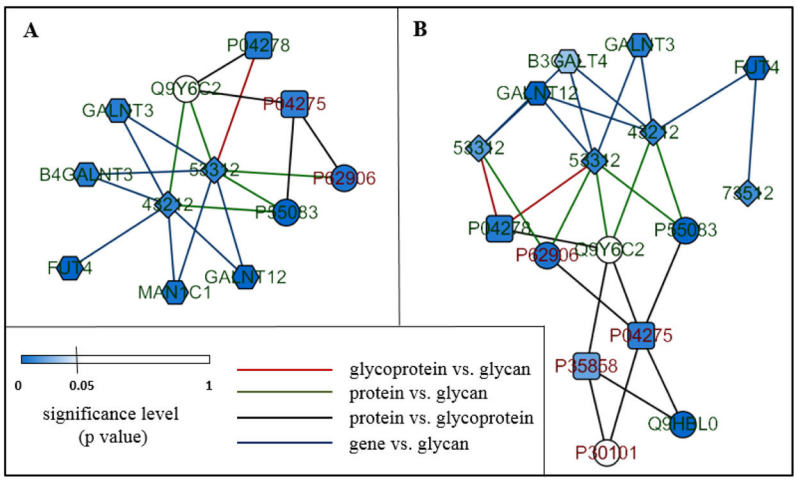

Figure 2.

Subnetworks from integrative analysis of genes (hexagon), glycans (diamond), proteins (circle), and glycoproteins (rectangle). Red-labeled biomolecules are up-regulated in HCC group, green are down-regulated.

Table I

Samples used for multi-omic analysis

		HCC (N=5)	CIRR (N=5)
Age	*Mean (SD)*	65 (7.3)	62 (2.3)
Gender	*Male*	3	3
BMI	*Mean (SD)*	31.3	30.3
Ethnicity	*Black*	1	1
	White	4	4
HCV Serology	*HCV Ab+*	2	2
HBV Serology	*HBs Ab+*	2	3
	HBs Ag−	5	5
Smoking	*Yes*	1	2
Alcohol	*Yes*	2	2
HCC Stage	*Stage I*	3	
	Stage II	1	
	Stage III	1	

Table II

Significant biomolecules identified by analysis of multi-omic data.

Omic data	Platform	Mode	# of features	HCC vs. CIRR (p<0.05)	HCC vs. ADJ-CIRR (p<0.05)	# of overlaps
Transcriptomics	RNA-Seq		20,510	2667 *	2625 *	499
Proteomics	LC-MS		933	146	110	39
Glycoproteomics	LC-MS		934	19	12	5
Glycomics	LC-MS		144	11	16	4
Metabolomics	GC-MS	Splitless	720	23	16	2
		Split 10:1	427	9	20	0
	GCxGC-MS	Split 20:1	1,408	28	18	2
		Split 40:1	1,222	27	16	3
	LC-MS	Postive	6,119	173	67	9
		Negative	672	35	20	8

* The significant values for RNA-Seq is based on FDR <0.05

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2018 April 24.