Supporting Information

Surface Glycoproteomic Analysis Reveals that both Unique and Differential Expression of Surface Glycoproteins Determine the Cell Type

Suttipong Suttapitugsakul, Lindsey D. Ulmer, Chendi Jiang, Fangxu Sun, and Ronghu Wu*

School of Chemistry and Biochemistry and the Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

*Corresponding author: Phone: 404-385-1515; Fax: 404-894-7452.

E-mail: ronghu.wu@chemistry.gatech.edu

ABSTRACT:

Proteins on the cell surface are frequently glycosylated and they are essential for cells. Surface glycoproteins regulate nearly every extracellular event, but compared to proteins, comprehensive and site-specific analysis of surface glycoproteins is much more challenging and dramatically understudied. Here, combining metabolic labeling, click-chemistry and enzymatic reactions, and mass spectrometry-based proteomics, we characterized surface glycoproteins from eight popular types of human cells. This integrative and effective method allowed for the identification of 2,172 N-glycosylation sites and 1,047 surface glycoproteins. The distribution and occurrence of N-glycosylation sites were systematically investigated, and protein secondary structures were found to have dramatic influence on glycosylation sites. As expected, most sites are located on disordered regions. For the sites with the motif N-!P-C, about one third of them are located on helix structures while those with the motif N-!P-S/T prefer strand structures. There is almost no correlation between the number of glycosylation sites and protein length, but the number of sites corresponds well with the frequencies of the motif. Quantification results reveal that besides cell-specific glycoproteins, the uniqueness of each cell type further arises from differential expression of surface glycoproteins. The current research indicates that multiple surface glycoproteins including their abundances need to be considered for cell classification rather than a single cluster of differentiation (CD) protein normally used in conventional methods. These results provide valuable information to the glycoscience and biomedical communities and aid in the discovery of surface glycoproteins as disease biomarkers and drug targets.

Supplementary Tables in Excel Files

Table S1. Glycosylation sites from eight types of cells identified from two biological replicate experiments

Table S2. Prediction of transmembrane region and signal peptide of protein identified with no membrane information from UniProt

Table S3. Prediction of the structure at each asparagine residue and the solvent accessibility

Table S4. Calculated entropy values and LFQ intensities of proteins from MaxQuant

Table S5. Estimation of absolute protein abundance with iBAQ

Cell Culture

Human cell lines, including HEK293T, HeLa, Jurkat, K562, MCF7, and U266, were from the American Type Culture Collection (ATCC). HeLa and PANC1 cell lines were a generous gift from Professor M.G. Finn's lab. HEK293T cell line was from Dr. Gang Bao's lab. Cell lines were not authenticated. HEK293T, HeLa, HepG2, MCF7, and PANC1 cells were maintained in high-glucose Dulbecco's Modified Eagle's Medium (DMEM, Sigma-Aldrich) containing 10% fetal bovine serum (FBS, Corning). Suspension cell lines, including Jurkat, K562, and U266, were maintained in RPMI-1640 medium (Sigma-Aldrich) containing 10% FBS. All cells were grown in a humidified incubator at 37 ^{o}C with 5% carbon dioxide (CO_2).

Metabolic Labeling and Click-Chemistry Reaction

Adherent cells were cultured in high-glucose DMEM medium until the cells reached ~50% confluency. The cells were labeled with 100 µM N-azidoacetylgalactosamine-tetraacylated ($Ac_4GalNAz$, Click Chemistry Tools) in low-glucose DMEM medium (Sigma-Aldrich) with 10% FBS. Suspension cells were cultured in RPMI-1640 medium until the cell density was ~$7x10^5$ cells/mL as determined by hemocytometry and trypan blue staining and labeled similarly to the adherent cells. After 24 hours of metabolic labeling, the adherent cells were washed twice with Dulbecco's Phosphate Buffered Saline (DPBS, Sigma-Aldrich) while the suspension cells were centrifuged at 300 g for 5 minutes to remove the medium and washed twice with DPBS similarly. Adherent cells were tagged with 100 µM water-soluble dibenzocyclooctyne (DBCO)-biotin (Click Chemistry Tools) in Cellstripper solution (Corning) for 1 hour in the humidified incubator. Suspension cells were labeled similarly except that DPBS was used instead of the Cellstripper solution. The reaction was quenched with 10 mM dithiothreitol (DTT, Sigma-Aldrich). The cell pellets were washed twice with ice-cold DPBS and kept on ice until the next steps.

Protein Extraction and Peptide Purification

The cell pellets were incubated with a buffer containing 25 µg/mL digitonin (Sigma-Aldrich), 150 mM sodium chloride (NaCl, Sigma-Aldrich), 50 mM N-(2-hydroxyethyl)piperazine-N'-2-ethanesulfonic acid (HEPES, Sigma-Aldrich, pH=8.2), and 1 tablet/10 mL cOmplete ULTRA Tablets protease inhibitor cocktail (Roche) at 4 ^{o}C for 10 minutes on an end-over-end rotator. The suspension was centrifuged at 2,000 g for 10 minutes, and the supernatant was removed. An ice-cold lysis buffer containing 0.5% sodium deoxycholate (SDC, Sigma-Aldrich), 50 mM HEPES (pH=8.2), 150 mM NaCl, 20 units/mL universal nuclease for cell lysis (Pierce), and 1 tablet/10 mL cOmplete ULTRA Tablets protease inhibitor cocktail was added to the cell pellets. After the lysis at 4 ^{o}C for 45 minutes on an end-over-end rotator, the suspension was centrifuged at 25,830 g for 10 minutes. The supernatant was collected and reduced with 5 mM DTT at 56 ^{o}C for 25 minutes and subsequently alkylated with 14 mM iodoacetamide (Sigma-Aldrich) for 30 minutes in the dark. The alkylation reaction was quenched by incubating with DTT to the final concentration of 5 mM in the dark for another 15 minutes.[1] Proteins were purified and pelleted with methanol/chloroform precipitation and digested with sequencing grade modified trypsin (Promega) at 37 ^{o}C for 16 hours (enzyme:substrate ratio of ~1:100) in a buffer containing 5% acetonitrile (ACN, Sigma-Aldrich), 1.6 M urea (Sigma-Aldrich), and 50 mM HEPES (pH=8.2). The digestion was quenched by adding trifluoroacetic acid (TFA, Sigma-Aldrich) to the final concentration of 0.4%, and the pH was checked to be lower than ~2. The peptides were desalted using a Sep-Pak Vac tC18 cartridge (Waters) and dried in a vacuum concentrator.

LC-MS/MS Analysis and Database Searching

The peptides were dissolved in a solution containing 5% ACN and 4% FA and were separated by a Dionex UltiMate 3000 UHPLC system (Thermo Fisher Scientific) with a microcapillary column containing C18 beads (Magic C18AQ, 3 µm, 200 A^{o}, 75 µm*16 cm) packed in-house. A total of ~1 µg of peptides was loaded into the column by a Dionex WPS-3000TPL RS autosampler (Thermostatted Pulled Loop Rapid Separation Nano/Capillary Autosampler). Peptides were separated by reversed-phase liquid chromatography (LC) using an UltiMate 3000 binary pump with 80-minute gradients of 4-25%, 10-38%, and 15-50% ACN containing 0.125% FA for the three fractions, respectively. The LC is coupled to an LTQ Orbitrap Elite Hybrid Mass Spectrometer (Thermo Scientific) with Xcalibur software (version 3.0.63). MS/MS analysis was performed with a data-dependent Top20 method.[2-3] For each cycle, a full MS scan (resolution: 60,000) in the Orbitrap with 1 million automatic gain control (AGC) target was followed by up to 20 MS/MS in the LTQ for the most intense ions. Selected ions were excluded from further sequencing for 90 seconds. Ions with singly or unassigned charge were not sequenced. Maximum ion accumulation times were 1,000 ms for each full MS scan and 50 ms for MS/MS scans.

Raw MS files were analyzed by MaxQuant (version 1.6.2.3).[4] MS spectra were searched against the human proteome database downloaded from UniProt containing common contaminants using the integrated Andromeda search engine.[5] Glycopeptides were searched separately for the identification experiments. All default parameters were left unchanged, except adding variable modification for asparagine deamidation (+2.9883 Da) for glycosylation site determination and 3 maximum missed cleavages. In the quantification experiments, all raw files were searched together with the three files from the same experiment grouped together. Label-free quantification was also enabled with the LFQ min ratio count of 1, the match-between-runs option was enabled, asparagine deamidation modification was used in protein quantification, and the iBAQ option was enabled. The false discovery rates (FDR) were kept at 0.01 at the peptide spectrum match, protein, and site decoy fraction levels.

Bioinformatic Analysis

Data analyses were performed with Perseus[6] and Excel. Glycopeptides were filtered to only contain the canonical sequences (N-X-S/T) and non-canonical sequence (N-X-C), where X is any amino acid except proline, for N-linked glycosylation. Human membrane protein information was extracted from UniProt database: SL-9905 (single-pass type I membrane proteins), SL-9906 (single-pass type II membrane proteins), SL-9907 (single-pass type III membrane protein), SL-9908 (single-pass type IV membrane proteins), SL-9909 (multi-pass membrane proteins), and SL-9903 (peripheral membrane proteins). For those whose membrane information is not available, further sequence analyses were performed using Phobius (phobius.sbc.su.se), which predicts the transmembrane and signal peptide regions of proteins.[7] SecretomeP (cbs.dtu.dk/services/SecretomeP) was used to further predict protein secretion through non-classical secretory pathways with the cutoff score of 0.6.[8]

Gene ontology-based enrichment analysis was performed on Gene Ontology Consortium website (http://www.geneontology.org). Fisher's exact test was used to calculate the P values and only those with P<0.05 were included. Residue solvent accessibility and structure were predicted using NetsurfP (version 1.1).[9] The structure (helix, strand, or coil) with the highest probability among the others was assigned a structure for the residue.

For the quantification experiments, the glycopeptide LFQ intensity was extracted from the peptides.txt table and limited to only glycopeptides (with the deamidation sites). The glycoprotein intensity was calculated by summing the peptide LFQ intensities together. The final LFQ intensity for each cell line was an average of the intensities between two biological duplicate experiments. iBAQ was used to estimate the absolute protein abundance ranking.[10] iBAQ intensity was calculated manually by dividing the summed glycopeptide intensity by the number of theoretical tryptic peptides, which was extracted from the proteinGroups.txt table. Shannon's entropy was calculated the same way as the previous report.[11] The entropy was calculated using the formula H(S)=$-\sum_t \; p(S_t)\ln p(S_t)$, where t is the protein index, $p(S_t)$ is the ratio of the protein LFQ intensity to the summed LFQ intensity of the protein. Because of the missing values, 1/8 was added to the raw LFQ intensity so that the natural log can be calculated.

Unsupervised hierarchical clustering was performed with Perseus. Euclidean distance was used to calculate the distance. Protein intensity was converted to a $\log_2$ scale before further analysis. For the heat map generation, missing values were imputed with a normal distribution (width=0.3, shift=1.8) before Z-score transformation.[6] ANOVA was performed with Perseus (S_0=0.5, Benjamini-Hochberg FDR=0.05). The proteins were filtered so that they contain at least 8 out of 16 valid values in order to reduce the effect of quantifying low-abundance surface glycoproteins.[12]

Protein interaction network was processed using Cytoscape.[13] Interaction information was extracted from STRING database with the high confidence cutoff (score=0.7).[14] Pathway analysis was performed with the Cluego plugin of Cytoscape.[15] All default parameters were used.

Cell-Surface Glycoprotein Interactions and Pathway Analyses

To explore the connection network among the quantified proteins, surface glycoprotein interactions with high confidence score were extracted from STRING database.[14] The network is complexed with several modes of interactions between these proteins including binding, catalysis, reaction, inhibition, and post-translation modification (Figure S4A). Proteins in the integrin family have the highest degree of interactions with other proteins, with ITGB1 interacting with the greatest number of proteins, including enzymes such as receptor-type tyrosine-protein phosphatase C, adhesion molecules such as neural cell adhesion molecule 11, and transporters such as basigin. Integrin beta-1 also interacts with other proteins in the integrin family, including ITGAV, ITGA1, ITGA2, ITGA3, and ITGA5, that are globally expressed. Integrin proteins are well-known surface proteins that play important roles at the cell interface, including acting as adhesion molecules and receptors. In fact, we identified most members of proteins in the integrin alpha family, and all proteins from the integrin beta family were identified, from ITGB1 to ITGB 8. Epidermal growth factor receptor (EGFR), a receptor tyrosine kinase with several roles in protein signaling, is another protein interacting with some integrin proteins and several other groups of proteins, such as those in the ephrin family, and transferrin receptor protein 1 (Figure S4A).

Different interaction networks also arose for each cell line, with proteins in the integrin family at the center of the network (Figure S5A for HEK293T cells, an enlarged figure is in Figure S6). About 40% of the quantified proteins do not have any interactions with the others. It is highly probable that these proteins interact with others outside of the current list or their interactions have not been reported yet. However, even with 100 additional proteins extracted from outside of our dataset, some proteins, such as choline transporter-like protein 1, do not show any interactions despite that they are globally expressed in all these cell lines.

Some cell-surface glycoproteins, especially receptors, regulate signal transductions and are involved in many pathways. We assigned these quantified surface glycoproteins into 213 KEGG pathway annotations (Figure S4B). As expected, most of these proteins are annotated as cell adhesion molecules (CAMs). With the specific surface glycoprotein expression patterns, several pathways arise depending on the function of that cell type. For example, those from the axon guidance process were identified from HEK293T cells, a neuronal-originated cell line, while proteins involved in the leukocyte trans-endothelial migration pathway were found in Jurkat cells (Figure S5B). Several signaling pathways were also identified, with 51 surface glycoproteins involved in the PI3K-Akt signaling pathway and 23 proteins in the RAP1 signaling pathways, for example. Other pathways only contain a few surface glycoproteins. These numbers are small compared to the phosphoproteome since the majority of signal transduction pathway is conducted through protein phosphorylation, and the numbers of phosphoproteins, kinases, and phosphatases far exceed the number of cell-surface glycoproteins.[11]

Table S6. Examples of cell-specific surface glycoproteins. The LFQ intensity is shown for each protein in each cell line. See the method section for the calculation of Shannon's entropy.

Major cell type	UniProt ID	Name	Annotation	Site	Shannon's entropy	HEK293T	HeLa	HepG2	Jurkat	K562	MCF7	PANC1	U266B1
HEK293T	P21802	FGFR2	Fibroblast growth factor receptor 2	N318	4.80E-06	3.3E+6	0	0	0	0	0	0	0
	P23471	PTPRZ1	Receptor-type tyrosine-protein phosphatase zeta	N105	2.64E-05	5.4E+5	0	0	0	0	0	0	0
HeLa	P42262	GRIA2	Glutamate receptor 2	N256	1.85E-06	0	9.0E+6	0	0	0	0	0	0
	P55285	CDH6	Cadherin-6	N399, N437	2.74E-05	0	5.2E+5	0	0	0	0	0	0
HepG2	P05534	HLA-A	HLA class I histocompatibility antigen, A-24 alpha chain	N110	6.43E-06	0	0	2.4E+6	0	0	0	0	0
	P07307	ASGR2	Asialoglycoprotein receptor 2	N102, N170	3.16E-07	0	0	5.8E+7	0	0	0	0	0
Jurkat	Q6GTX8	LAIR1	Leukocyte-associated immunoglobulin-like receptor 1	N69	4.18E-06	0	0	0	3.8E+6	0	0	0	0
	P06127	CD5	T-cell surface glycoprotein CD5	N116, N241	5.02E-07	0	0	0	3.6E+7	0	0	0	0
K562	P04629	NTRK1	High affinity nerve growth factor receptor	N67, N188, N262	1.59E-07	0	0	0	0	1.2E+8	0	0	0
	Q6UWB1	IL27RA	Interleukin-27 receptor subunit alpha	N76	1.26E-05	0	0	0	0	1.2E+6	0	0	0
MCF7	Q8IZF3	ADGRF4	Adhesion G protein-coupled receptor F4	N61, N250, N686	3.52E-05	0	0	0	0	0	4.0E+5	0	0
	Q96NY8	NECTIN4	Nectin-4	N281	4.66E-06	0	0	0	0	0	3.4E+6	0	0
PANC1	Q02297	NRG1	Pro-neuregulin-1, membrane-bound isoform	N164	7.92E-06	0	0	0	0	0	0	1.9E+6	0
	Q04900	CD164	Sialomucin core protein 24	N146	6.31E-05	0	0	0	0	0	0	2.1E+5	0
U266B1	Q02223	TNFRSF17	Tumor necrosis factor receptor superfamily member 17	N42	4.17E-06	0	0	0	0	0	0	0	3.8E+6
	Q96FE7	PIK3IP1	Phosphoinositide-3-kinase—interacting protein 1	N66	2.32E-06	0	0	0	0	0	0	0	7.1E+6

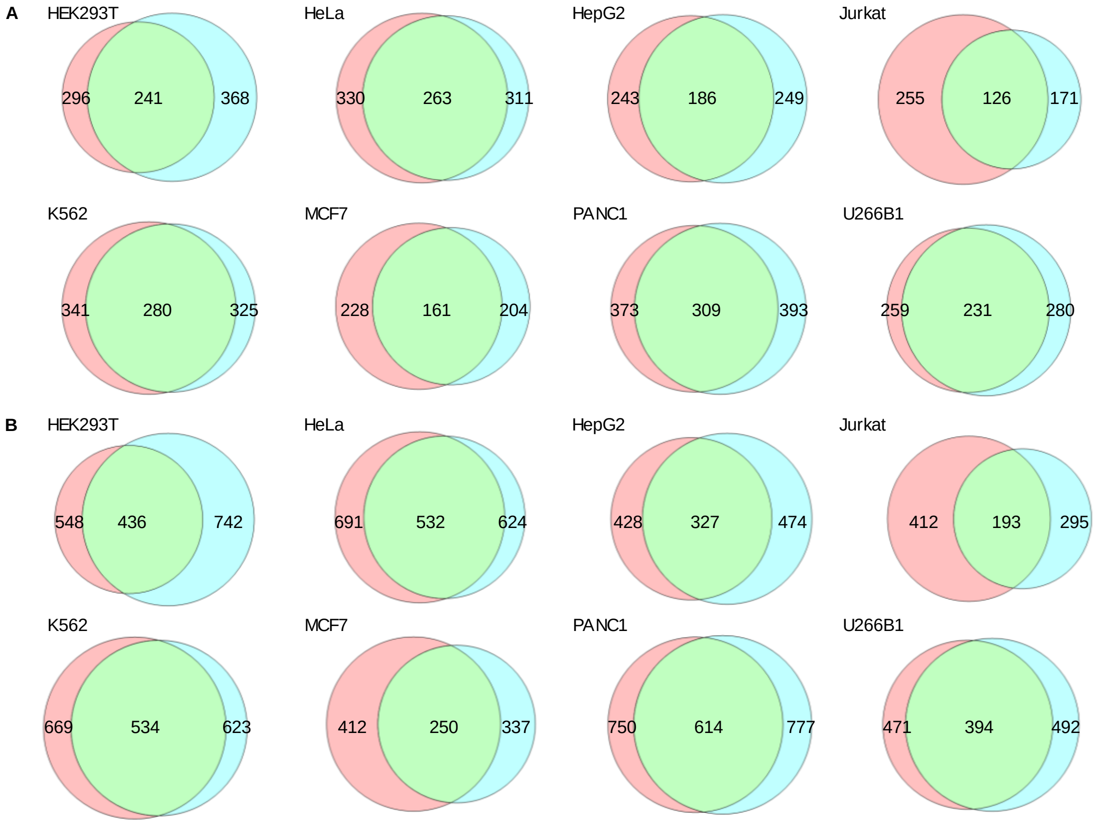

Figure S1. Reproducibility of the identification of cell-surface glycoproteins from biological duplicate experiments. (**A**) Overlap of cell-surface glycoproteins identified from each cell line. (**B**) Overlap of cell-surface glycosylation sites identified from each cell line.

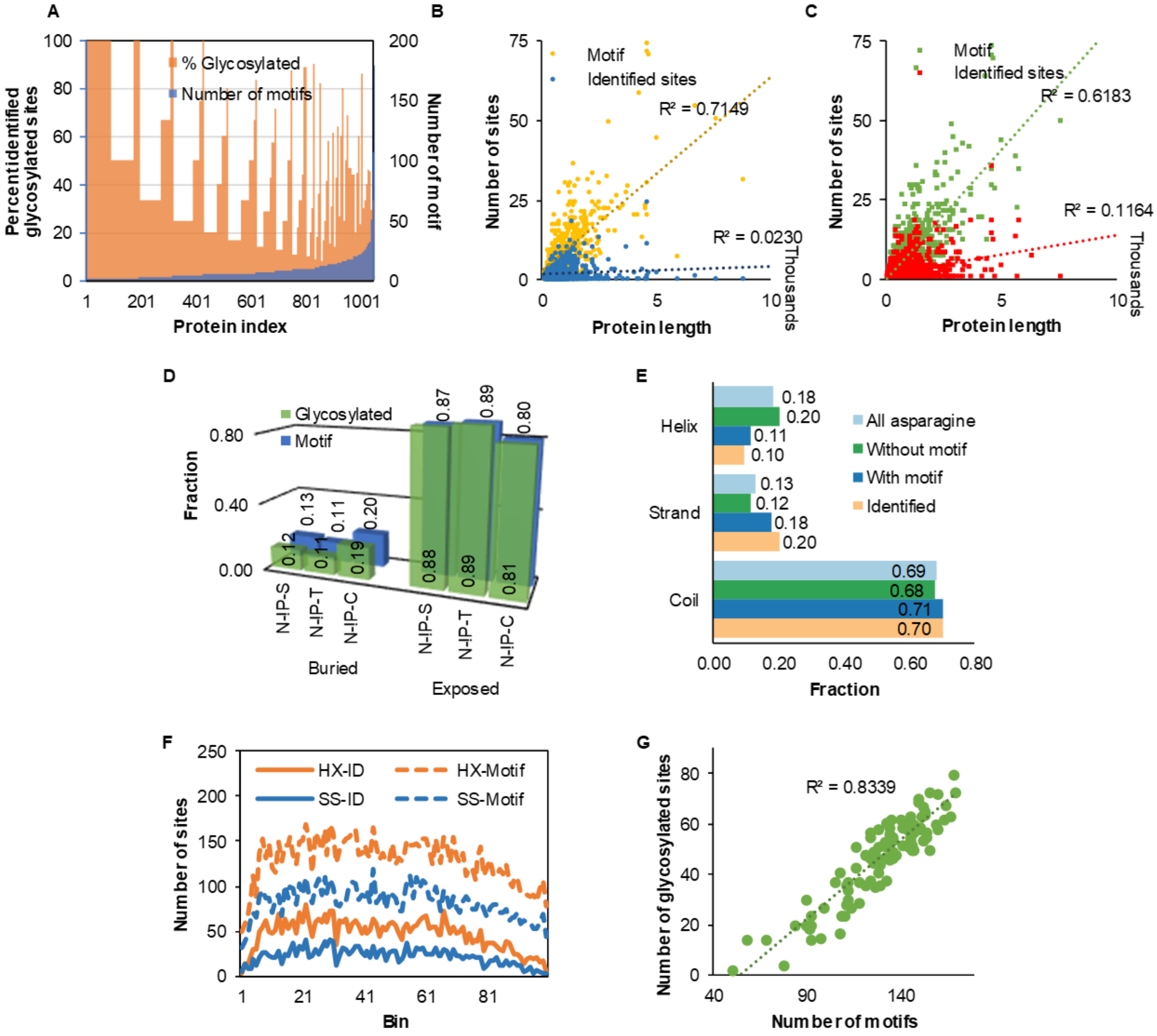

Figure S2. Site-specific analysis of surface protein glycosylation. (**A**) The number of glycosylation motifs, including all N-!P-S/T/C motifs, of all identified proteins and their percentages of the identified glycosylation sites. (**B**) The correlation of the protein length and the numbers of glycosylation motifs and identified sites from this work. (**C**) The correlation of the protein length and the numbers of glycosylation motifs and identified sites from Xiao et al. (Ref. 40). (**D**) Distribution of the solvent accessibility at each N-linked glycosylation motif. (**E**) Structure prediction of all asparagine residues from the identified surface glycoproteins. (**F**) Glycosylation motifs and site distribution when each protein length is divided into 100 bins from this work (SS) compared to those identified from Xiao et al., 2018 (**G**) The correlation between the number of glycosylated sites and the number of motifs in each bin for the data set from Xiao et al., 2018.

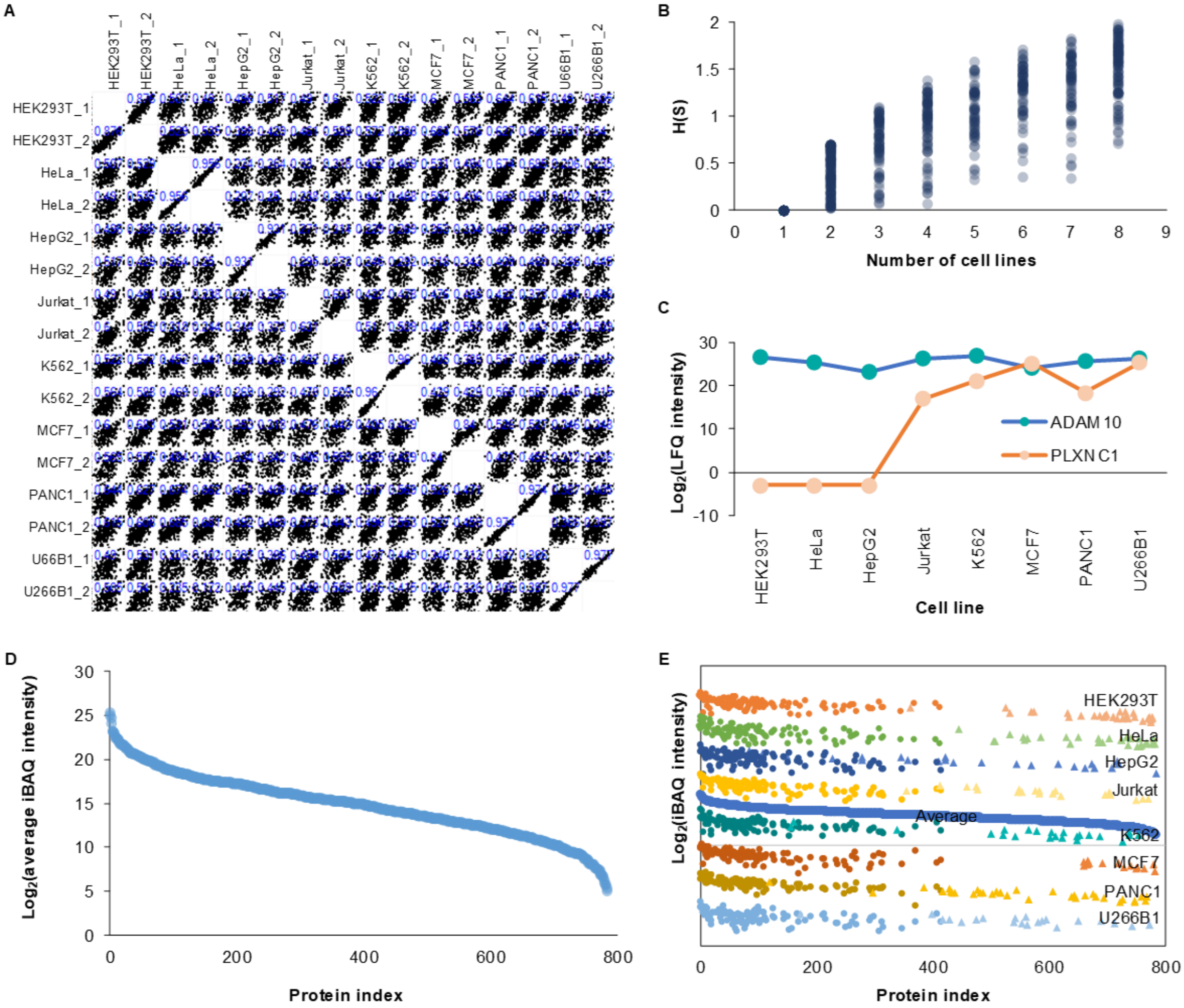

Figure S3. Label-free quantification of cell-surface glycoproteins. (**A**) The correlation of $\log_2$-transformed LFQ intensity from biological duplicate experiments or between cell lines. Pearson correlation is shown on the plot. Zeros values were omitted. (**B**) Distribution of Shannon's entropy (H(S)) of proteins found in different number of cell lines. (**C**) Example of two proteins with different Shannon's entropies and their LFQ intensities across the cell lines after missing value imputation. ADAM10 has the entropy of 1.89 while PLXNC1 has the entropy of 0.83. (**D**) Average iBAQ intensity of proteins from all cell lines. (**E**) Abundance ranking of cell-specific and globally expressed proteins from all cell lines, similar to Figure 3D. The $\log_2$-tranformed intensities for each cell line were offset by different values so the plot can visibly be seen. The data points from the same cell line are shown in the same color. Those that are globally expressed are shown in circles while those specific to the cell line are shown in triangles.

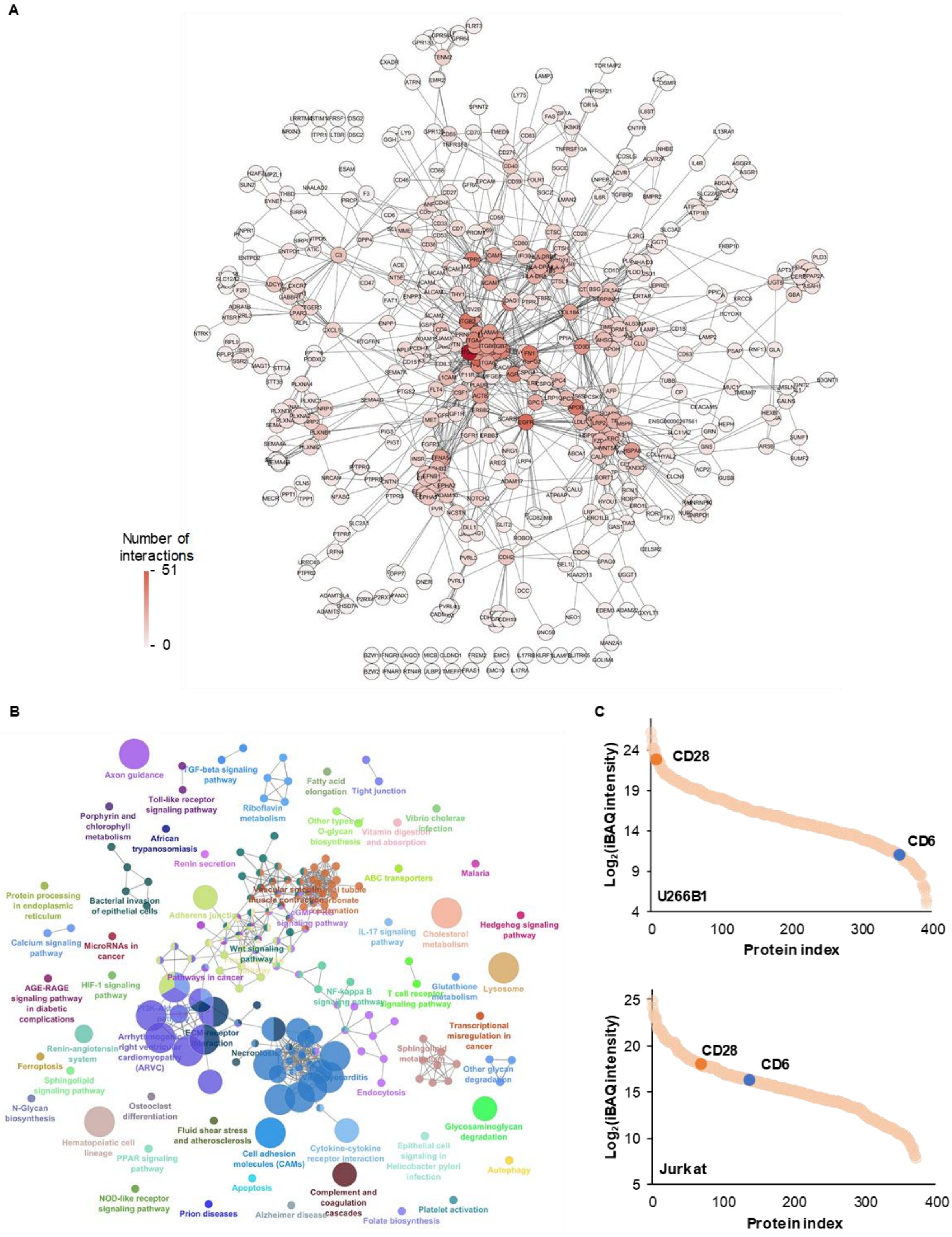

Figure S4. Cell-surface glycoprotein interaction and CD proteins. (**A**) Cell-surface glycoprotein interaction of all quantified surface glycoproteins with the data extracted from STRING. Those

without any interactions were omitted. Data was processed by Cytoscape. (**B**) KEGG pathway analysis of all quantified surface glycoproteins. Data were processed with ClueGO plugin on Cytoscape. Default parameters were used. The size of each node corresponds to the significance of the term (a larger node means that the corrected P value is lower). The term with the highest significance is labeled on the plot. (**C**) Estimated abundance ranking of CD28 and CD6 from U266B1 and Jurkat cells.

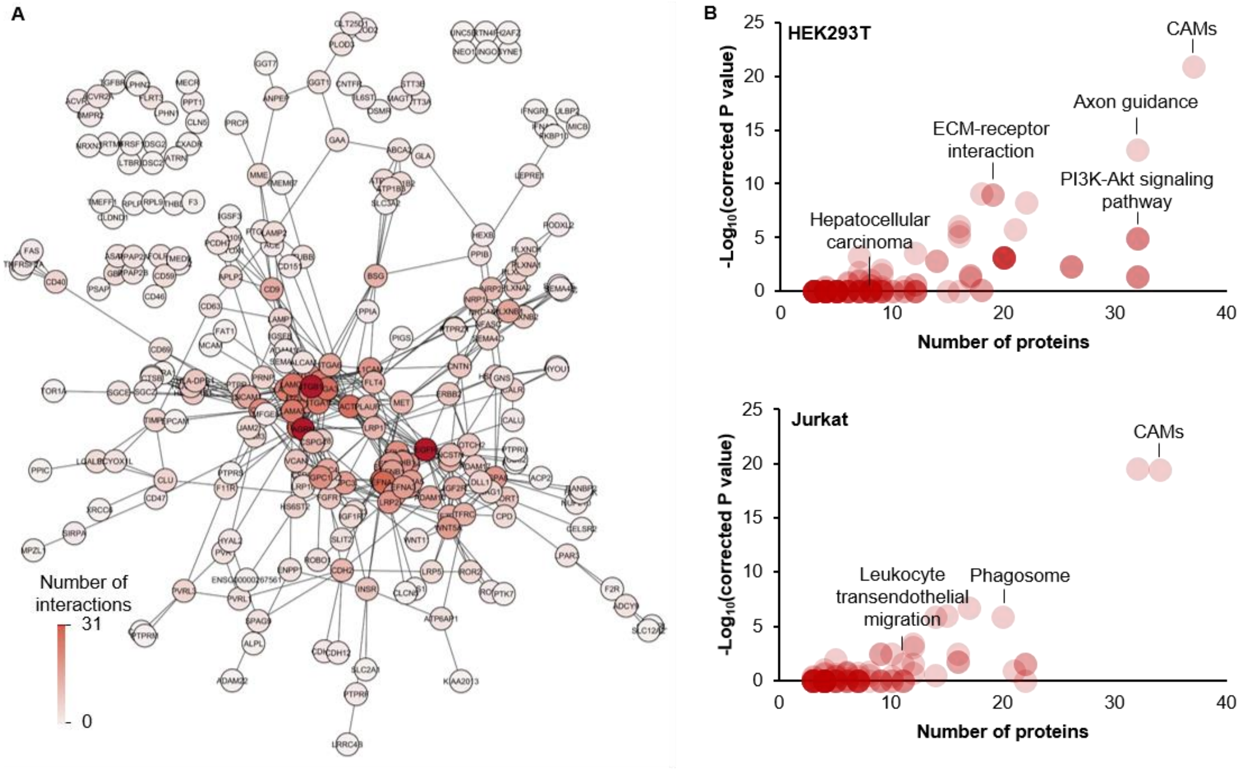

Figure S5. Cell-surface glycoprotein interactions and their roles in biological pathways. (**A**) Interaction network of surface glycoproteins from HEK293T cells extracted from STRING database. Those without any interactions are not shown. See Figure S6 for a larger view of this Figure. (**B**) KEGG pathway analysis shows enriched pathways in HEK293T (top) and Jurkat (bottom) cells, respectively.

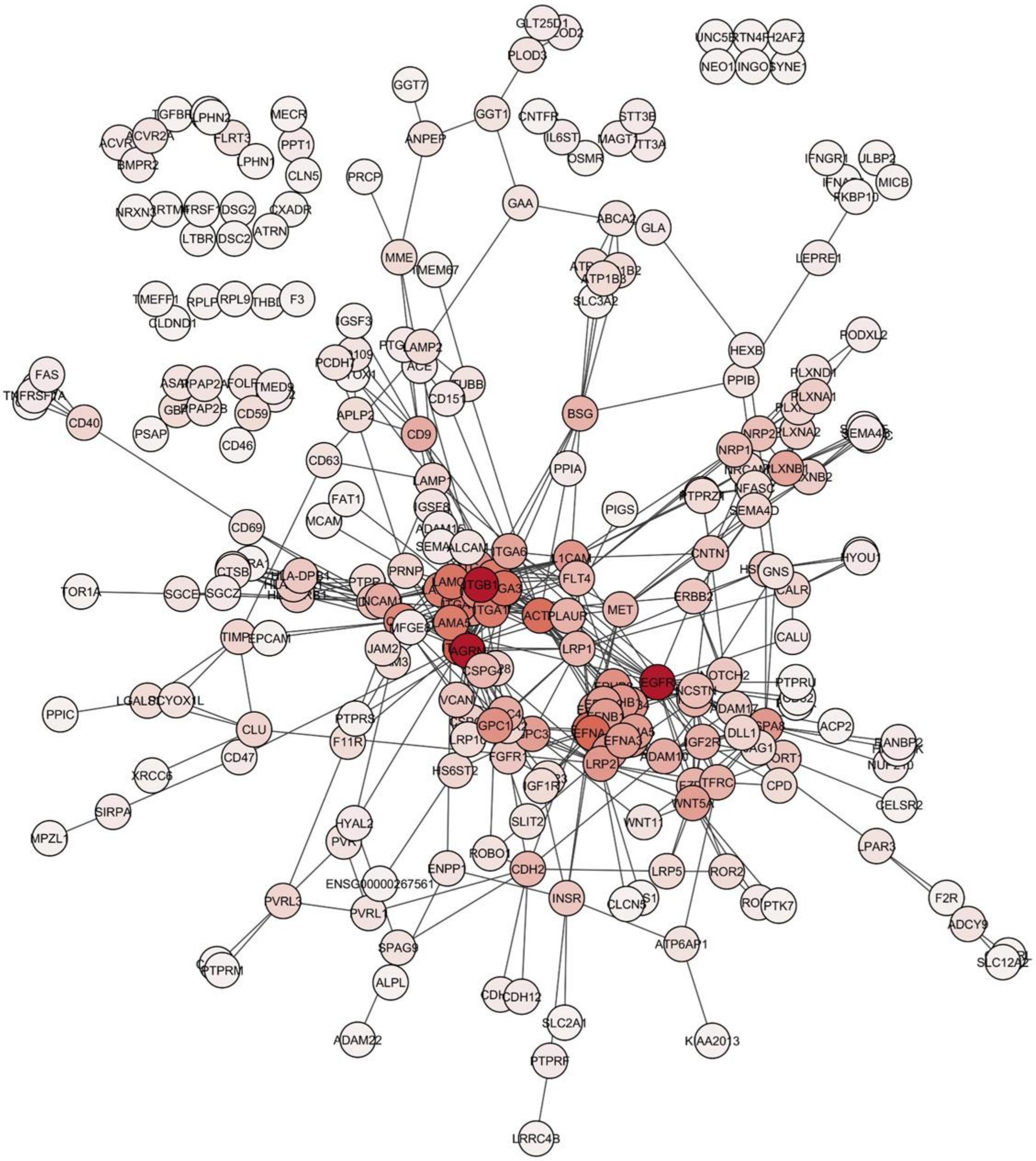

Figure S6. An enlarged view of Figure 5A.

References:

(1) Suttapitugsakul, S.; Xiao, H.; Smeekens, J.; Wu, R., Evaluation and optimization of reduction and alkylation methods to maximize peptide identification with MS-based proteomics. *Mol. Biosyst.* **2017,** *13* (12), 2574-2582.

(2) Xiao, H.; Wu, R., Global and site-specific analysis revealing unexpected and extensive protein S-GlcNAcylation in human cells. *Anal. Chem.* **2017,** *89* (6), 3656-3663.

(3) Smeekens, J. M.; Xiao, H.; Wu, R., Global analysis of secreted proteins and glycoproteins in saccharomyces cerevisiae. *J. Proteome Res.* **2017,** *16* (2), 1039-1049.

(4) Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008,** *26* (12), 1367-1372.

(5) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011,** *10* (4), 1794-1805.

(6) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **2016,** *13* (9), 731-740.

(7) Kall, L.; Krogh, A.; Sonnhammer, E. L., A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004,** *338* (5), 1027-1036.

(8) Bendtsen, J. D.; Jensen, L. J.; Blom, N.; Von Heijne, G.; Brunak, S., Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des. Sel.* **2004,** *17* (4), 349-356.

(9) Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C., A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **2009,** *9*, 51.

(10) Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M., Global quantification of mammalian gene expression control. *Nature* **2011,** *473* (7347), 337-342.

(11) Huttlin, E. L.; Jedrychowski, M. P.; Elias, J. E.; Goswami, T.; Rad, R.; Beausoleil, S. A.; Villen, J.; Haas, W.; Sowa, M. E.; Gygi, S. P., A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **2010,** *143* (7), 1174-1189.

(12) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M., Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **2012,** *11* (3), M111 014050.

(13) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003,** *13* (11), 2498-2504.

(14) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C., STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015,** *43* (Database issue), D447-452.

(15) Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W. H.; Pages, F.; Trajanoski, Z.; Galon, J., ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **2009,** *25* (8), 1091-1093.