Electronic Supplementary Material (ESI) for Chemical Science.

Supplementary Information

Systematic and Site-specific Analysis of N-Sialoglycosylated Proteins on the Cell Surface by Integrating Click Chemistry and MS-based Proteomics

Weixuan Chen, Johanna M. Smeekens and Ronghu Wu*

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

*Correspondence to: Ronghu Wu (Ronghu.Wu@chemistry.gatech.edu), Phone: 404-385-1515, Fax: 404-894-7452.

Supplementary Tables:

Table S1: Identification of 395 unique N-sialoglycosylation sites on cell surface proteins in HEK293T cells

Table S2: Identification of 128 unique sites on N-sialoglycoproteins in the secretome of HEK293T cells

Table S3: Identification of 1,120 N-sialoglycosylation sites on membrane proteins in HEK293T cells

Table S4: Identification of 541 N-sialoglycosylation sites on cell surface proteins of breast cancer cells

Table S5: N-sialoglycoproteins belonging to type I, II single-pass, multi-pass and peripheral membrane protein (UniProt)

Table S6: Computational results of signal peptide, transmembrane domain and non-classic secretion pathway analysis

Table S7: Quantification of 406 unique N-sialoglycopeptides on the cell surface of breast cancer cells

Table S8: Up-regulated N-sialoglycoproteins related to cell adhesion

Table S9: Domain analysis of up-regulated N-sialoglycoproteins in MDA-MB-231 cells

Table S10: Up-regulated N-sialoglycoproteins based on important domains

Table S11: Quantification of 342 N-sialoglycosylation sites on the cell surface of breast cancer cells

Table S12: Identification of 325 unique N-sialoglycosylation sites on 88 CDs

Table S13: Quantification of 119 N-sialoglycoprotein sites on 57 CDs in breast cancer cells

SUPPLEMENTARY TEXT

EXPERIMENTAL PROCEDURES

Materials

Complete protease inhibitors were purchased from Roche Applied Sciences, sequencing grade trypsin was from Promega, and Lysyl endopeptidase (Lys-C) from Wako. Dulbecco's Modified Eagle Medium (DMEM), N-acetylmannosamine (ManNAc), DBCO-Fluor545 and phosphate buffered saline (PBS) were from Sigma-Aldrich. Fetal bovine serum (FBS), dialyzed FBS, Ac_4ManNAz and high capacity NeutrAvidin resin were bought from Fisher. DBCO-sulfo-biotin was from Click Chemistry Tools, and streptavidin-fluorescein isothiocyanate (streptavidin-FITC) from Genetex. 4–12% Nu-PAGE® Bis-Tris gels, and lithium dodecyl sulfate buffer were ordered from Invitrogen.

Cell lines, cell culture, metabolic labeling and click chemistry

HEK293T and MCF-7 cells were kindly provided by Dr. Gang Bao (Georgia Institute of Technology). MDA-MB-231, the triple negative breast cancer cell line was purchased from ATCC. All cell lines were cultured following the instructions provided by ATCC. HEK293T cells were incubated in DMEM until 40% confluency, then the medium was discarded and the monolayer was rinsed with PBS two times. The cells were further incubated in DMEM supplemented with 30 μM ManNAz or ManNAc for 24 hours [1]. The labeling medium was collected and the cells were washed twice with PBS and incubated in PBS containing 200 μM DBCO-sulfo-biotin or DBCO-Fluor545 for 1 hour with gentle agitation at 4 °C. Cells were collected by centrifugation (5 minutes at 300 g) and washed with PBS twice.

Breast cancer cell culture with SILAC

Two types of breast cancer cells with distinctive invasiveness were chosen: MCF-7 cells with low invasiveness, and MDA-MB-231 cells, which is one of the most invasive breast cancer cells.[2, 3] Both types of cells were cultured through stable isotope labeling by amino acids in cell culture (SILAC), in addition to the Ac_4ManNAz labeling. MCF-7 cells were grown in media containing unlabeled lysine and arginine (light), while MDA-MB-231 cells were grown in media with the addition of heavy lysine (+8 Da; lysine ($^{13}C_6$, $^{15}N_2$)) and arginine (+6 Da; arginine ($^{13}C_6$)). In order to prevent the possible conversion of arginine to proline [4], the medium was supplemented with 200 mg/L L-proline.

Cell imaging

HEK293T cells were maintained in DMEM. Wells of a 2-well Lab-Tek™ II Chambered Coverglass (Thermo Scientific, 155379) were coated with 0.2% gelatin at 37 °C for 30 minutes. Gelatin solution was removed, and ~120K ManNAz-treated HEK293T cells were plated into the wells and treated with 30 μM ManNAz. Twenty-four hours later, cells were washed twice with PBS, and incubated with solution A (PBS containing both 200 μM DBCO-sulfo-biotin and 200 μM DBCO-Fluor545) or B (PBS containing only 200 μM DBCO-Fluor545) for 1 hour at 4 °C. Cells incubated with solution B were further incubated with PBS containing 200 μM DBCO-sulfo-biotin for 1 hour at 4 °C. After incubation with the reagents containing DBCO, cells were washed with PBS twice and treated with solution C (containing 2μg/mL streptavidin-FITC) for 30 minutes at 4 °C. Cells were then washed with PBS again and the Chambered Coverglasses were mounted on a confocal microscope (Zeiss LSM 700). Fluor545 was visualized under the microscope using a TRITC filter set (555/28 excitation and 617/73 emission), and FITC using a GFP

filter set (470/40 excitation and 525/50 emission). The control experiments followed the same procedure, but used ManNAc instead of ManNAz. Labeled MDA-MB-231 and MCF-7 cells were growing in the chambered coverglasses and incubated with solution B for 1 hour at 4 °C. After washed with PBS twice, cells were visualized using the same microscope (Zeiss LSM 700) with 555 nm excitation.

Sialoglycoprotein tagging and detection in cell lysates

HEK293T cells were seeded in T175 flasks and treated with the ManNAz or ManNAc (30 μM for 1 day) in DMEM. The cells were then harvested with nonenzymatic dissociation buffer; ~2×10^7 cells were collected and washed with PBS. Harvested cells were subjected to different click chemistry procedures. For cell surface protein tagging, the cell pellets were washed twice with PBS and incubated in PBS containing 200 μM DBCO-Fluor545 for 1 hour with end-over-end rotation at 4 °C. For the membrane compartment tagging, the cell pellets were suspended in ice-cold buffer containing 100 mM HEPES (pH 7.5), 25 μg/mL digitonin, and protease inhibitor mixture (Roche Diagnostics), and then incubated with rotation at 4 °C for 10 minutes, followed by the centrifugation at 2,000 *g* to pellet the cells. The supernatant was discarded and the cell pellets were washed with ice-cold PBS twice. The purified cell pellets were incubated with 200 μM DBCO-Fluor545 in PBS for 1 hour at 4 °C. After the conjugation, cell pellets were washed with ice-cold PBS containing 10 mM DTT twice to remove and quench unreacted reagents. The pellets for surface protein detection were treated with digitonin then lysed in RIPA buffer (50 mM HEPES, pH=7.5, 150 mM NaCl, 0.1% SDS, protease inhibitors) and the pellets for membrane protein detection were lysed in RIPA directly. Protein concentrations were measured using the BCA protein assay (Pierce). To detect sialoglycoproteins in different cell extracts, ~20 μg of labeled or unlabeled protein lysate was resolved by SDS-PAGE. The Fluor545 signal and SimplyBlue staining bands were visualized and documented using a GE Typhoon Trio+ Fluorescence/Phospho-Imager system.

Cell lysis, protein extraction and digestion

The cell pellets were suspended in ice-cold buffer containing 100 mM HEPES, pH=7.5, 25 μg/ml digitonin, and protease inhibitor mixture (Roche Diagnostics); the cell suspension was rotated end-over-end at 4 °C for 10 minutes and then centrifuged at 2,000 *g* to pellet the cells. The supernatant was discarded and the cell pellets were washed in ice-cold PBS twice. The cell pellets were resuspended in RIPA buffer with protease inhibitor cocktail and Benzonase (1 U/mL) and the suspension was incubated with shaking at 4 °C for 1 hour. After complete solubilization of nuclei and digestion of genomic DNA, the lysate was centrifuged at 25,000 *g* for 20 minutes. The supernatant was collected and proteins (about 2 mg total proteins for HEK293T cells, and about 10 mg for the SILAC experiment) were reduced by 5 mM DTT (56 °C, 25 min) and alkylated with 15 mM iodoacetamide (RT, 30 minutes in the dark). Proteins were purified by the methanol/chloroform precipitation method and resolublized in 50 mM HEPES (pH=8.2), 0.1 M urea, 5% ACN, digested with Lys-C (the ratio of Lys-C and protein was about 1:50) at 31ºC for 15 hours followed by trypsin digestion at 37 ºC for 4 hours. Digestion was quenched by the addition of 10% TFA to a final concentration of 0.4%, and the resulting peptides were purified using a Sep-Pak tC18 cartridge (Waters).

Biotinylated peptide enrichment and PNGase F treatment

The peptide mixture was dissolved in 100 mM PBS (pH=7.4), and incubated with the high capacity NeutrAvidin resin for one hour with rotation. After incubation, the beads were transferred into a spin column and washed five times with PBS and one time with Milli-Q water. Enriched peptides were eluted three times by incubation with 8 M guanidine (pH=1.5) at 56 ºC for 2 minutes each. The eluted peptides

were purified using a Sep-Pak tC18 cartridge. The enriched samples were dried in a lyophilizer for at least 24 hours. The completely dried samples were dissolved in heavy-oxygen water containing 40 mM NH_4HCO_3 and treated with PNGase F (lyophilized powder from Sigma Aldrich) at 37 ºC for 3 hours. PNGase F removed glycans on Asn and as a result, Asn was converted to Asp with a tag of ^{18}O. The deglycosylated peptide samples were desalted again using a stage tip and eluted with three eluting solutions containing 20%, 50% and 80% ACN, respectively.

LC-MS/MS analysis

Purified and dried peptide samples were dissolved in a solvent with 5% ACN and 4% formic acid (FA), and 2 μL was loaded onto a microcapillary column packed with C18 beads (Magic C18AQ, 5 μm, 200 Å, 100 μm x 16 cm) using a WPS-3000TPLRS autosampler (UltiMate 3000 Thermostatted Pulled Loop Rapid Separation Wellplate Sampler, Dionex). Peptides were separated by reversed-phase chromatography using an UltiMate 3000 binary pump with a 90-minute gradient of 4-30% ACN (in 0.125% FA) and detected in a hybrid dual-cell quadrupole linear ion trap – Orbitrap mass spectrometer (LTQ Orbitrap Elite, ThermoFisher, with Xcalibur 3.0.63 software) using a data-dependent Top 20 method.[5] For each cycle, one full MS scan (resolution: 60,000) in the Orbitrap at 10^6 AGC target was followed by up to 20 MS/MS in the LTQ for the most intense ions. Selected ions were excluded from further analysis for 90 s each. Ions with a single or unassigned charge were not sequenced. Maximum ion accumulation times were 1000 ms for each full MS scan and 50 ms for each MS/MS scan.

Database searching and data filtering

The raw files recorded by MS were converted into mzXML format. Precursors for MS/MS fragmentation were checked for incorrect monoisotopic peak assignments while refining precursor ion mass measurements.[6] All MS/MS spectra were then searched using the SEQUEST algorithm (version 28).[7] Spectra were matched against a database encompassing sequences of all proteins in the UniProt Human (*Homo sapiens*) database (downloaded in February 2014) containing common contaminants such as keratins. Each protein sequence was listed in both forward and reversed orientations to estimate false discovery rate (FDR) of peptide identification. The following parameters were used for the database search: 20 ppm precursor mass tolerance; 1.0 Da product ion mass tolerance; fully tryptic digestion; up to two missed cleavages; variable modifications: oxidation of methionine (+15.9949) and O^{18} tag of Asn (+2.9883); fixed modifications: carbamidomethylation of cysteine (+57.0214).

The target-decoy method was employed to evaluate and further control FDRs of glycopeptide identification.[8, 9] Linear discriminant analysis (LDA) was utilized to distinguish correct and incorrect peptide identifications using numerous parameters such as XCorr, ΔCn, and precursor mass error.[6] This approach has been used in previous works[10-13]. After scoring, peptides less than six amino acids in length were discarded and peptide spectral matches were filtered to a less than 1% FDR based on the number of decoy sequences in the final data set. The dataset was restricted to N-sialoglycopeptides when determining FDRs.

Glycosylation site localization

To assign glycosylation site localizations and measure the assignment confidence, we applied a probabilistic algorithm[14] that considers all potential glycosylation sites on a peptide and uses the presence or absence of experimental fragment ions unique to each to calculate a ModScore. The ModScore, which is similar to AScore, indicates the likelihood that the best site match is correct when compared with the

next best match. This has also been used previously.[15, 16] We considered sites with a score ≥19 ($P \leq 0.01$) to be confidently localized.

Bioinformatic analysis

Membrane protein annotations were extracted from the UniProt database (Peripheral membrane protein [UniProt subcellular location ID: SL-9903], Multi-pass membrane protein [SL-9909], Single-pass type I membrane protein [SL-9905], Single-pass type II membrane protein [SL-9906], Single-pass type III membrane protein [SL-9907], Single-pass type IV membrane protein [SL-9908], and Transmembrane beta strand [UniProt keyword ID: KW-1134]). For the proteins whose information was unavailable on the UniProt website, further sequence analysis was performed. Briefly, the signal sequence, unconventional secretion and transmembrane segment of the undefined protein sequences were predicted by two prediction algorithms SecretomeP (http://www.cbs.dtu.dk/services/SecretomeP/) and Phobius (http://phobius.sbc.su.se/). Domain analysis was carried out by using available information from the UniProt database, the InterPro database and the online prediction software SUPERFAMILY (http://supfam.org/SUPERFAMILY/hmm.html). Protein functional annotations were obtained using DAVID (http://david.abcc.ncifcrf.gov/home.jsp) [17] and the Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System.[18, 19]

Quantification of unique N-sialoglycopeptides and N-sialoglycosylation sites

Identified N-sialoglycopeptides in breast cancer cells with different invasiveness were quantified. In order to obtain accurate ratios, signal-to-noise (S/N) ratios for both heavy and light peptides were required to be larger than 3. If the S/N ratio for one (either heavy or light) peptide was lower than 3, the other one (light or heavy, respectively) was required to be larger than 5. In addition, if a glycopeptide is exclusively

identified in one sample, the ratio reported here is not a real ratio, but the S/N ratio of the corresponding peptide. With the criteria of S/N being over 3, any quantified exclusive N-sialoglycopeptides must be up- or down- regulated.

Based on the dataset of quantified unique N-sialoglycopeptides, we further performed the quantification of N-sialoglycosylation sites. It required that the quantified peptide contain only one site, and the ModScore was >19 to make sure the site was well-localized.

Domain analysis

Another two examples are shown in Figure S6. PTPRJ (CD148) is receptor-type tyrosine-protein phosphatase eta, which dephosphorylates or contributes to the dephosphorylation of many proteins including EGFR, LYN, SRC, MAPK1, MAPK3 and TJP1. It plays a critical role in cell adhesion, migration, proliferation and differentiation, and is involved in the regulation of human T cell activation [20]. This protein contains nine fibronectin type III domains in the extracellular part and one tyrosine phosphatase domain in the intracellular part. All 13 identified N-sialoglycosylation sites were located within the fibronectin type III domain. From the N-terminus, each of the first four domains contained at least two sialoglycosylation sites, and all three quantified sites were significantly up-regulated in invasive cells.

Another example is integrin alpha-3 (ITGA3, CD49c), which is a receptor for fibronectin, laminin, collagen, epiligrin, thrombospondin, and is involved in cell-matrix adhesion [21]. There are 14 predicted N-glycosylation sites listed on the UniProt website, and ten of them were identified in this work (N86, 107, 265, 511, 573, 605, 656, 697, 926 and 969). Seven sites were quantified and they were all highly up-regulated with ratios ranging from 3.7 to 28.3 in MDA-MB-231 cells. The transmembrane region is from

amino acid residue 992-1014, and the cytoplasmic region is from 1015-1051. Sialic acid bound to the terminus of the N-glycans at these sites can dramatically increase the water solubility and cause proteins to carry more negative charges under physiological conditions, which will result in decreased cell adhesion and increased cell invasiveness and migration.

SUPPLEMENTARY REFERENCES

1 S. T. Laughlin and C. R. Bertozzi, *Nat. Protoc.*, 2007, **2**, 2930-2944.

2 M. Morini, M. Mottolese, N. Ferrari, F. Ghiorzo, S. Buglioni, R. Mortarini, D. M. Noonan, P. G. Natali and A. Albini, *Int. J. Cancer*, 2000, **87**, 336-342.

3 R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J. P. Coppe, F. Tong, T. Speed, P. T. Spellman, S. DeVries, A. Lapuk, N. J. Wang, W. L. Kuo, J. L. Stilwell, D. Pinkel, D. G. Albertson, F. M. Waldman, F. McCormick, R. B. Dickson, M. D. Johnson, M. Lippman, S. Ethier, A. Gazdar and J. W. Gray, *Cancer Cell*, 2006, **10**, 515-527.

4 S. C. Bendall, C. Hughes, M. H. Stewart, B. Doble, M. Bhatia and G. A. Lajoie, *Molecular & Cellular Proteomics*, 2008, **7**, 1587-1597.

5 W. Haas, B. K. Faherty, S. A. Gerber, J. E. Elias, S. A. Beausoleil, C. E. Bakalarski, X. Li, J. Villen and S. P. Gygi, *Molecular & Cellular Proteomics*, 2006, **5**, 1326-1337.

6 E. L. Huttlin, M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausolei, J. Villen, W. Haas, M. E. Sowa and S. P. Gygi, *Cell*, 2010, **143**, 1174-1189.

7 J. K. Eng, A. L. McCormack and J. R. Yates, *J. Am. Soc. Mass Spectrom.*, 1994, **5**, 976-989.

8 J. M. Peng, J. E. Elias, C. C. Thoreen, L. J. Licklider and S. P. Gygi, *J. Proteome Res.*, 2003, **2**, 43-50.

9 J. E. Elias and S. P. Gygi, *Nat. Methods*, 2007, **4**, 207-214.

10 X. Du, S. J. Callister, N. P. Manes, J. N. Adkins, R. A. Alexandridis, X. Zeng, J. H. Roh, W. E. Smith, T. J. Donohue, S. Kaplan, R. D. Smith and M. S. Lipton, *J. Proteome Res.*, 2008, **7**, 2595-2604.

11 L. Kall, J. D. Canterbury, J. Weston, W. S. Noble and M. J. MacCoss, *Nat. Methods*, 2007, **4**, 923-925.

12 J. Y. Zhang, J. Ma, L. Dou, S. F. Wu, X. H. Qian, H. W. Xie, Y. P. Zhu and F. C. He, *Molecular & Cellular Proteomics*, 2009, **8**, 547-557.

13 R. H. Wu, W. Haas, N. Dephoure, E. L. Huttlin, B. Zhai, M. E. Sowa and S. P. Gygi, *Nat. Methods*, 2011, **8**, 677-683.

14 S. A. Beausoleil, J. Villen, S. A. Gerber, J. Rush and S. P. Gygi, *Nat. Biotechnol.*, 2006, **24**, 1285-1292.

15 W. X. Chen, J. M. Smeekens and R. H. Wu, *J. Proteome Res.*, 2014, **13**, 1466-1473.

16 W. X. Chen, J. M. Smeekens and R. H. Wu, *Molecular & Cellular Proteomics*, 2014, **13**, 1563-1572.

17 D. W. Huang, B. T. Sherman and R. A. Lempicki, *Nat. Protoc.*, 2009, **4**, 44-57.

18 H. Y. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano and P. D. Thomas, *Nucleic Acids Res.*, 2005, **33**, D284-D288.

19 P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Y. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan and A. Narechania, *Genome Res.*, 2003, **13**, 2129-2141.
20 S. G. Tangye, J. H. Phillips, L. L. Lanier, J. E. de Vries and G. Aversa, *J. Immunol.*, 1998, **161**, 3249-3255.
21 R. O. Hynes, *Cell*, 2002, **110**, 673-687.

SUPPLEMENTARY FIGURES

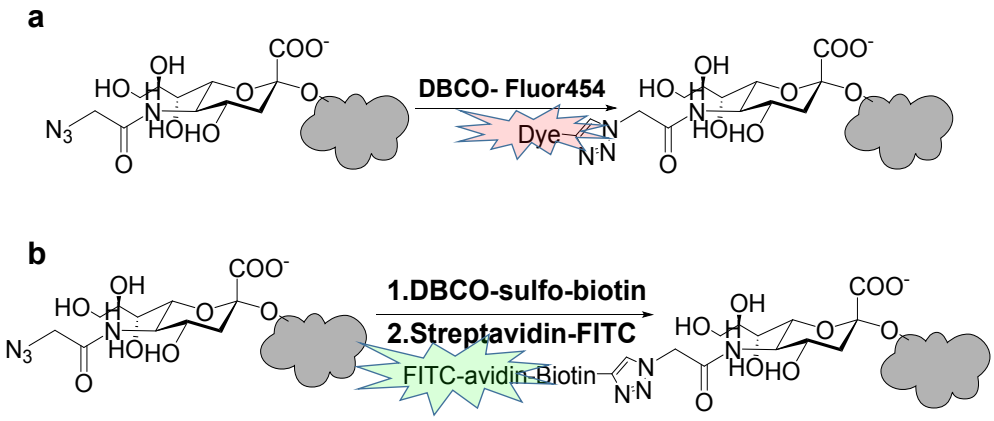

Figure S1. (a) The click chemistry reaction between DBCO-Fluor545 and a sialoglycoprotein containing the azido group labeled with Ac_4ManNAz. (b) The reaction between labeled proteins and DBCO-sulfo-biotin, followed by incubation with streptavidin-FITC.

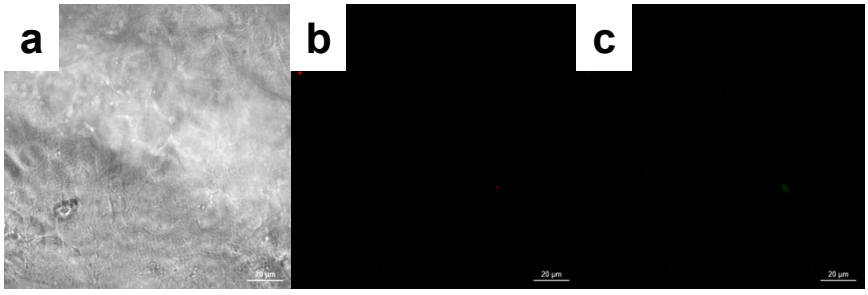

Figure S2. Microscopic results for unlabeled cells. (a) Bright field images of cells (scale bar is 20 μm); (b) Fluorescence signals of unlabeled cells reacted with DBCO-Fluor545 *via* copper-free click chemistry, and no red signal was detected; (c) Fluorescence signals of unlabeled cells reacted with DBCO-sulfo-biotin, followed by incubation with streptavidin-FITC, and no green signal was detected.

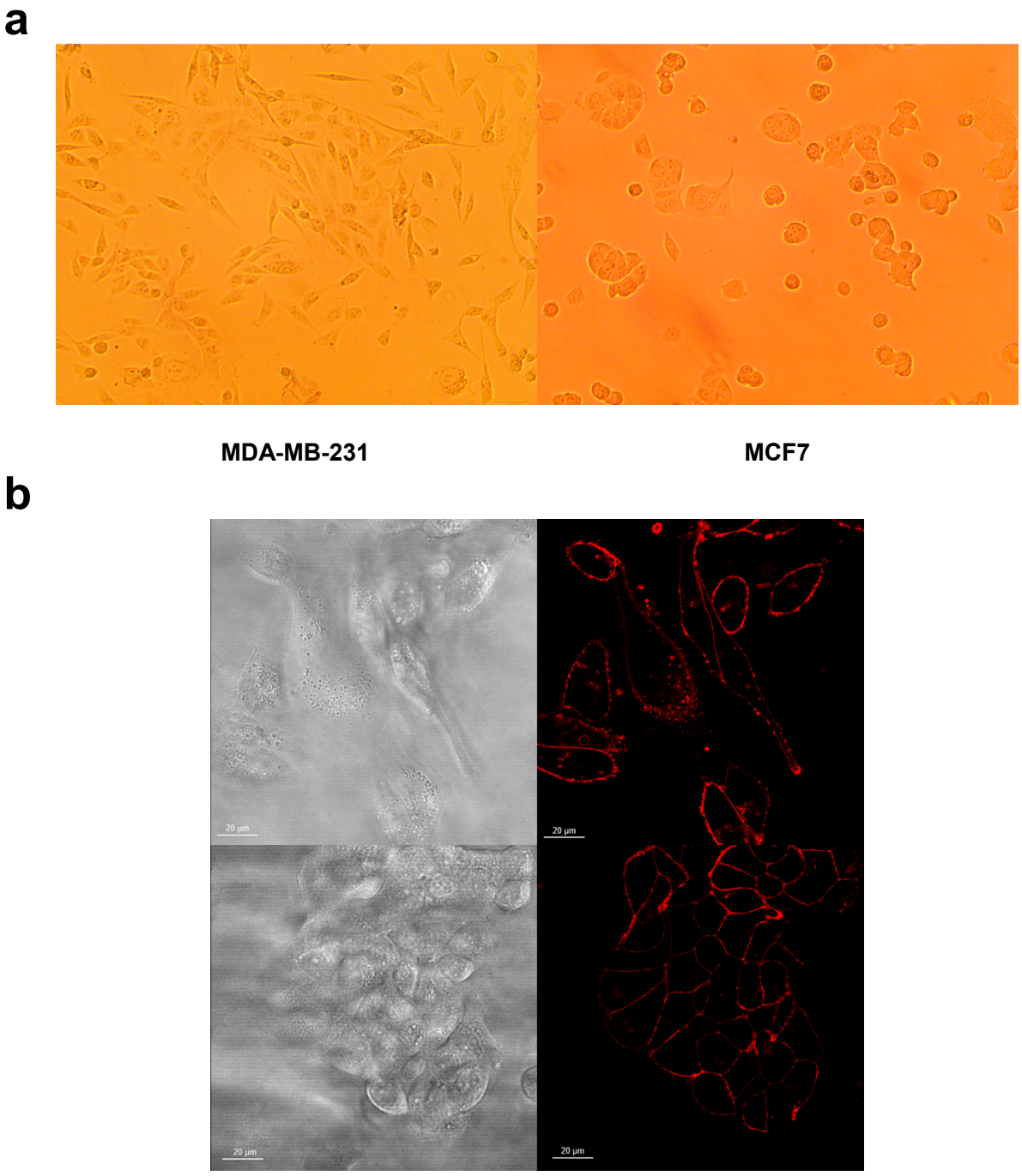

Figure S3. Cell morphology results of invasive MDA-MB-231 and non-invasive MCF-7 cells. (a) Microscopic results of the two types of cells. (b) After the two types of labeled cells were tagged by DBCO-Fluor545, red fluorescence signal was detected (scale bar is 20 µm). MDA-MB-231 cells are in the top row and MCF-7 cells in the bottom one.

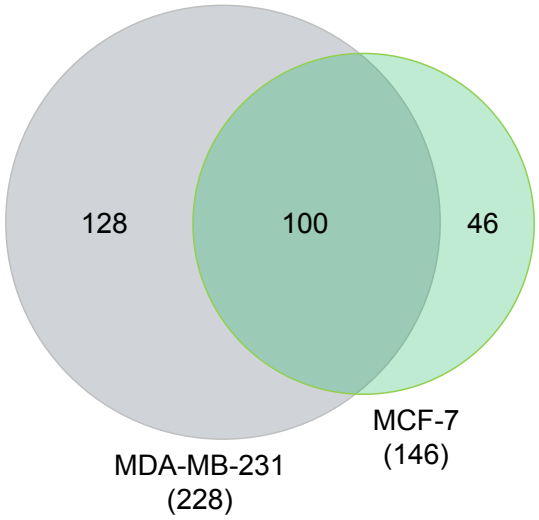

Figure S4. Overlap of cell surface N-sialoglycoproteins identified in MDA-MB-231 and MCF-7 cells.

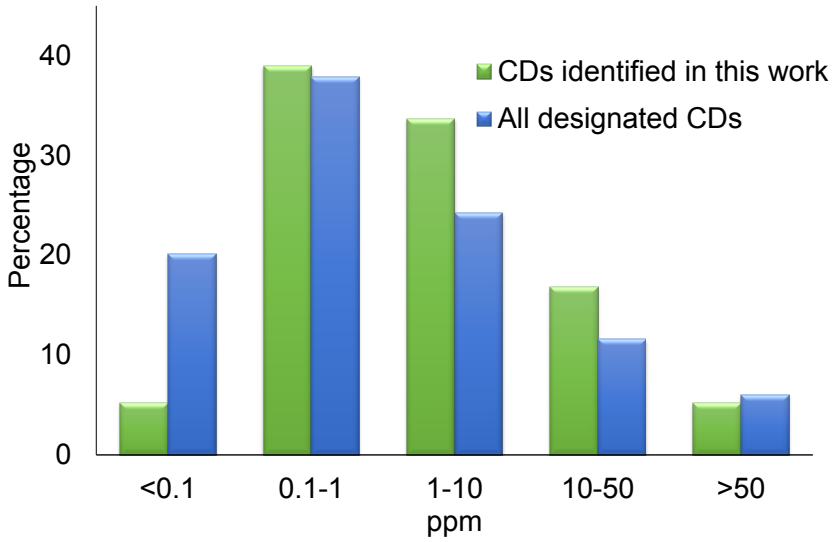

Figure S5. Abundance distribution of CDs identified from this work compared to all designated CDs (abundance values are from an on-line database (PaxDb)).

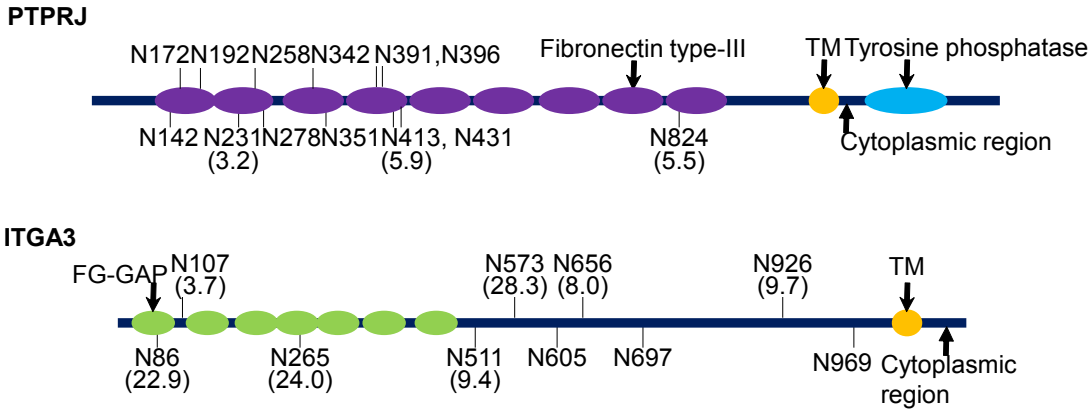

Figure S6. Identified and quantified N-sialoglycosylation sites on the cell surface proteins PTPRJ (CD148) and ITGA3 (CD49c).