





OPEN

Identification and characterization of intact glycopeptides in human urine

Fernando Garcia-Marques¹, Keely Fuller¹, Abel Bermudez¹, Nikhiya Shamsher¹, Hongjuan Zhao², James D. Brooks^{1,2}, Mark R. Flory³ & Sharon J. Pitteri¹  

Glycoproteins in urine have the potential to provide a rich class of informative molecules for studying human health and disease. Despite this promise, the urine glycoproteome has been largely uncharacterized. Here, we present the analysis of glycoproteins in human urine using LC–MS/MS-based intact glycopeptide analysis, providing both the identification of protein glycosites and characterization of the glycan composition at specific glycosites. Gene enrichment analysis reveals differences in biological processes, cellular components, and molecular functions in the urine glycoproteome versus the urine proteome, as well as differences based on the major glycan class observed on proteins. Meta-heterogeneity of glycosylation is examined on proteins to determine the variation in glycosylation across multiple sites of a given protein with specific examples of individual sites differing from the glycosylation trends in the overall protein. Taken together, this dataset represents a potentially valuable resource as a baseline characterization of glycoproteins in human urine for future urine glycoproteomics studies.

Glycosylation of proteins encompasses an extraordinarily diverse post-translational modification class wherein complex carbohydrate structures are linked, principally via hydroxyl (O-linkage) and asparagine (N-linkage) attachments, to the side chains of polypeptide backbones. The glycan additions, in myriad branched-chain structures, are formed from building blocks of mannose, high mannose, fucose, and sialic acid and are assembled, modified, and removed by a highly regulated set of enzymatic activities in cells¹. The diversity of glycosylation modifications on proteins contributes significantly to the estimated millions of proteoforms, or protein variant isoforms, estimated to comprise the full cellular proteome². Since proteins are the direct effectors of most biological processes, a more thorough characterization of this modification class is critical to both a better understanding of cellular mechanisms and for discovery of new disease biomarkers.

At a functional level, emerging data indicates that protein glycosylation is integrated with fundamental biological processes including cell–cell recognition, signal transduction, and protein trafficking³, and protein glycosylation figures prominently in health and disease¹. Multiple congenital disorders have long been associated with inborn glycosylation defects that often present with phenotypes including neurologic abnormalities and intellectual disabilities⁴. Furthermore, a growing body of evidence demonstrates critical linkages between protein glycosylation and cancer progression. For example, in serous ovarian carcinoma, glycoproteomic signatures can be used to classify disease subtypes that have distinct clinical outcomes⁵. Discrete classes of protein glycosylation have been observed in tissue specimens collected from patients with benign prostatic hyperplasia that can be distinguished from prostate cancer⁶. Separate studies in prostate cancer have identified specific glycoproteomic profiles that correlate with tumor aggressiveness and include markers potentially specific to metastatic disease^{7,8}. The wealth of glycoproteomic data demonstrate that glycoproteins specifically mark cancers and suggest that these glycosylation changes may play critical roles in promoting important biological state changes that occur during tumor development, invasion, and metastasis, underscoring the importance of gaining a better understanding of their biological functions.

Glycoproteins, often secreted from cells and tissues and detectable in body fluids, also provide a rich source of noninvasive biomarkers for cancer detection. For example, non-invasive detection and treatment monitoring in pancreatic cancer diagnosis involves assaying for glycosylated proteins in the serum that are detected by

¹Canary Center at Stanford for Cancer Early Detection, Department of Radiology, Stanford University School of Medicine, 3155 Porter Drive MC5483, Palo Alto, CA 94304, USA. ²Department of Urology, Stanford University School of Medicine, Stanford, CA 94305, USA. ³Cancer Early Detection Advanced Research (CEDAR) Center, Knight Cancer Institute, Oregon Health & Science University, Portland, OR 97239-3098, USA. ✉email: spitteri@stanford.edu

the monoclonal antibody CA19-9 which is known to bind a specific sialyl glycan class⁹. Measurement of serum glycoprotein prostate serum antigen (PSA) is widely used to screen for and monitor prostate cancer. PSA demonstrates a number of glycosylation modifications, resulting in PSA “glycoforms”, that can differ between benign and malignant prostate tissues and have been proposed as candidate biomarkers that improve PSA performance as a screening biomarker^{10,11}. While substantial biomarker discovery work has focused on plasma and serum from blood, much less effort has been directed at measuring biomarkers in more accessible fluid types such as urine. Urine has been shown to harbor a rich variety of modified proteins with over 2600 glycoproteins identified to date¹². While urinary glycoproteomics is a logical frontier for discovery of biomarkers for urologic indications such as prostate, kidney and bladder cancers, glycoprotein alterations in urine have also been described for organs not in direct continuity with the urinary tract, including liver, lung and stomach cancers¹².

Given the potential of urine glycoproteins as disease biomarkers, there is a need to develop improved urinary glycoproteomic workflows aimed toward biomarker discovery in this highly accessible biofluid. One roadblock of comprehensive glycoproteomic profiling in urine and other body fluids has been the inability of mass spectrometry (MS) and downstream analytic tools to effectively identify the vast diversity of attached glycan structures. To circumvent this challenge, MS-based glycoproteomic profiling efforts have often employed enzymatic removal of glycans prior to downstream MS, typically involving shotgun mode data acquisition on derivative tryptic peptide digests. In this approach, glycosylated amino acids retain a small chemical adduct after enzymatic deglycosylation that is detectable by MS. While these studies have provided a critical foundation by identifying glycosylated amino acid residues, including those in urine, the ability to understand more granular information encoded in the complex glycan structures themselves is forfeited because of the upfront deglycosylation step¹³. More recently, methods for characterization intact glycopeptides and determining structures for attached glycans using MS with companion bioinformatic tools like Byonic have enabled mass-based identification of a growing list of peptide-attached glycans¹⁴. In addition, the efficiency of glycoproteomic workflows continue to improve with the advent of chromatographic strategies including hydrophilic-lipophilic balance (HLB) and C18-facilitated reversed-phase chromatography modes and combined approaches^{13,15,16}, as well as hydrophilic interaction (HILIC) chromatography¹⁷, that have improved sample desalting and glycoprotein enrichment for urinary samples. Finally, choice of mass spectrometry configuration also has been shown to markedly impact the efficiency of glycopeptide detection. High-energy C trap-based dissociation (HCD) has emerged as an effective choice for peptide fragmentation improving the ability to detect and analyze structures of intact glycopeptides¹⁸. Here, we present an optimized workflow that combines elements of the sample processing steps and employs a new database of intact glycopeptides, resulting in an eightfold improvement in urinary glycopeptide detection over prior reports.

Methods

Glycoproteomics workflow

Pooled urine from healthy individuals was purchased from Innovative Research and 5 and 10 mL aliquots of pooled urine were each concentrated to 200 μ L using 4 mL Amicon filters (Sigma-Aldrich) via multiple rounds of centrifugation (45 min, 3500 \times g, 4 $^{\circ}$ C). Filters were washed with an additional 3 mL of 50 mM ammonium bicarbonate (Sigma-Aldrich) and the concentrated solution was split into 3 aliquots of equal volumes. Each aliquot was adjusted to 120 μ L with 50 mM ammonium bicarbonate solution followed by the addition of 12 μ L or 14 μ L of 10% sodium dodecyl sulfate (SDS, Invitrogen) for the initial 5 mL and 10 mL urine aliquots respectively. The disulfide bonds on cysteine residues on concentrated proteins were reduced with 5 μ L of 200 mM Tris(2-carboxyethyl) phosphine (TCEP) (Sigma-Aldrich) at 70 $^{\circ}$ C for 1 h. The free thiol groups were alkylated with 7.5 μ L of 200 mM iodoacetamide (Acros Organics) followed by an incubation of 45 min at room temperature in the dark. Proteins were precipitated with 1 mL of cold acetone and stored overnight at -20 $^{\circ}$ C. Samples were centrifuged at 14,000 \times g for 10 min at 4 $^{\circ}$ C. Acetone was removed and the protein pellets were allowed to dry for 5 min. Pellets were reconstituted with 80 μ L of 50 mM ammonium bicarbonate and vortexed. Urinary proteins were digested with 2 μ g of sequencing grade modified trypsin enzyme (Thermo Fisher Scientific) for 18 h at 37 $^{\circ}$ C. The three separate 120 μ L aliquots from each urine sample were combined, vortexed, and glycopeptides were enriched using strong anion exchange and electrostatic repulsion hydrophilic interaction chromatography (SAX-ERIC) as described previously¹⁹. Briefly, the SOLA SAX solid phase extraction column was equilibrated with 3 mL of acetonitrile (Fisher Scientific), activated with 3 mL of 100 mM triethylammonium acetate (Fluka, Honeywell), and followed by adding 3 mL of 1% trifluoroacetic acid (TFA, Sigma-Aldrich) in water (Fisher Scientific). 3 mL of equilibration solution consisting of 95% acetonitrile with 1% TFA in water were passed through the SOLA SAX column for equilibration. The combined tryptic peptides were diluted with 3 mL of equilibration solution, loaded, and passed through the column at a rate of 1 mL/min. Non-binding peptides were washed off with 6 mL of equilibration solution. Then, glycopeptides were eluted from the column by adding two 850 μ L aliquots of 50% acetonitrile with 0.1% TFA in water followed by another two 850 μ L aliquots of 5% acetonitrile with 0.1% TFA in water. The resulting glycopeptides were dried down using a speed vacuum (LabConco) and further fractionated using a high pH reversed-phase fractionation kit (Thermo Fisher Scientific) following manufacturer’s recommended protocol. The fractionated glycopeptides were dried down using a speed vacuum and reconstituted with 12 μ L of 0.1% formic acid (Fisher Scientific) in HPLC MS grade water (Fisher Scientific) for LC/MS–MS analysis.

A Dionex Ultimate Rapid Separation Liquid Chromatography system (Thermo Fisher Scientific) was used to load 10 μ L of the reconstituted glycopeptides onto a PEPMAP 100 C18 5 μ m trap column (Thermo Fisher Scientific) with a flow rate set at 5 μ L/min for 10 min. Glycopeptides were separated by reversed-phase chromatography on a 25 cm long C18 analytical column (New Objective) packed in-house with BEH C18, 130 Å , 1.7 μ m particle size (Waters). An external column heater (MSWIL) was used to heat the analytical column to

60 °C. Glycopeptides were eluted by changing the mixture of mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile). The gradient program consisted of holding mobile phase B at 2% for the first 10 min, slowly ramped up to 35% over the next 85 min, followed by an increase to 85% over 5 min with a 5 min hold. The analytical column was re-equilibrated for 15 min prior to the next sample injection. The flow rate throughout the gradient was set to 0.3 µL/min. Eluted glycopeptides were analyzed using an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher Scientific). The cycle time was set at Top-speed for 3 s with an MS1 mass scan range of 375–2000 m/z and Orbitrap resolution of 120,000. The normalized AGC target was set to 250 percent and the maximum injection time to auto. The most abundant precursor ions were fragmented with higher energy collisional dissociation (HCD) and with a collision energy set to 38%. Dynamic exclusion was enabled for 15 s with the mass tolerance to 10 ppm. The normalized AGC target for the MS2 was set to 200%. MS2 fragments were detected in the Orbitrap with a mass resolution of 30,000 with injection time set to auto.

Byonic software 4.0.12 (Protein Metrics) was used to search raw files against a focused-human-urine protein database (2021; 2421) generated from shotgun proteomics of urine and the 309 mammalian N-glycan library provided in the Byonic software for glycopeptide identification. Parameters included trypsin digestion with a maximum of two missed cleavages and precursor mass tolerance of 10 ppm. Fixed cysteine carbamidomethylation and variable methionine oxidation, asparagine deamination, and N-glycan modification on asparagine contained within a N-X-S/T (where X can be any amino acid except proline) N-glycosylation amino acid consensus sequence were also specified.

Data processing and analysis

Peptide identifications were filtered for Byonic Score greater than 150, and log probability greater than 1.5. The mass difference between two fucoses and one sialic acid is 1 Da and can lead to misidentifications of glycopeptides when the incorrect monoisotopic peak is identified. We corrected for this problem by: (1) selecting glycopeptide identifications containing two or more fucoses where the mass accuracy was determined to be greater than – 1 Da, (2) determining the maximum number of sialic acids possible (= the number of hexoses minus 3), and (3) if the maximum number of sialic acids in the glycopeptide was less than the maximum number of sialic acids possible, two fucoses were replaced by one sialic acid in the glycopeptide identification.

Glycopeptides were classified according to the numbers of each combined sugar into seven glycan types (high mannose, hybrid, complex undecorated, complex sialylated, complex fucosylated, complex fucosylated plus sialylated, and other) using the decision tree in the Supplementary Fig. 1. All glycan structures containing two or less HexNAc, and three or less Hex were classified as “other”. If the number of HexNAc equaled two and the number of Hex was greater than three, the glycans were classified as “high mannose”. If the number of HexNAc was greater than or equal to 3, the number of Hex was greater than or equal to 3, and the number of Hex in the glycan main core was lower than HexNAc, the glycan was classified as “hybrid glycan”. For complex glycans: (1) those that did not contain Fuc or NeuAc, were classified as “complex undecorated”, (2) those containing NeuAc but no Fuc were classified as “complex sialylated”, (3) those containing Fuc but no NeuAc were classified as “complex fucosylated”, and (4) those containing both Fuc and NeuAc, were classified as “complex fucosylated plus sialylated”.

Each identified glycoprotein was quantified using each assigned identification and distributed according to each glycan type after correction (if needed), dividing each spectral count per protein and glycan type by the total number of spectral counts per glycoprotein.

To better understand the relationship between the total urine proteome and the urine glycoproteome, we compared our glycoproteome data and a reference urine proteome²⁰ by applying an overrepresentation analysis using the protein annotations according to GO biological process, GO cellular component, GO molecular function, biological pathway, protein domain, and site of expression, using the total human proteome as background. The analysis included only protein categories with greater than four proteins, and adjusted p-value of enrichment lower than 0.01, in at least one the datasets. Using these same criteria, we analyzed the protein sets determined by significant correlation ($P < 0.01$) against each of the seven glycan types considered in the analysis.

Results

Characterization of urine glycoproteome

Tandem mass spectrometry-based glycoproteomic analysis of pooled human urine samples collected from healthy individuals resulted in 45,303 total high quality intact N-linked glycopeptides (i.e. glycan attached to peptide backbone, GSMs). These glycopeptides corresponded to 8135 unique combinations of peptide sequences and glycan structures that mapped to 751 glycosites on 347 unique glycoproteins (Supplementary Table 1). Approximately 50% of the total identified glycopeptides corresponded to five abundant proteins (Fig. 1A). Uromodulin (UMOD), the most abundant protein in urine, accounted for ~28% of the total identified glycopeptides. Approximately 70% of the glycoproteins were identified by more than one unique glycopeptide (Fig. 1B). Furthermore, 44% of the glycoproteins had two or more unique glycosites for which we were able to characterize the glycan composition (Fig. 1B). The glycan compositions on the glycopeptides included more than 269 unique structures which were classified by glycan group (Fig. 1C, Supplementary Table 2). The complex decorated glycans (complex fucosylated, complex sialylated, and complex fucosylated and sialylated) were the most abundant glycan structures, with complex fucosylated and sialylated being the most abundant types. Complex undecorated, high mannose, hybrid, and other glycan structures were less abundant, comprising 21.3% of the total glycans.

To further study the biological properties of glycoproteins in urine, we performed gene enrichment analysis using Gene Ontology (GO) databases and biological pathway, protein domain, and site of expression databases. For each of these classes, enrichment analyses of the glycoproteins we identified and a large urine proteome dataset were performed²⁰ (Fig. 1D, Supplementary Table 3). GO biological process including immune response

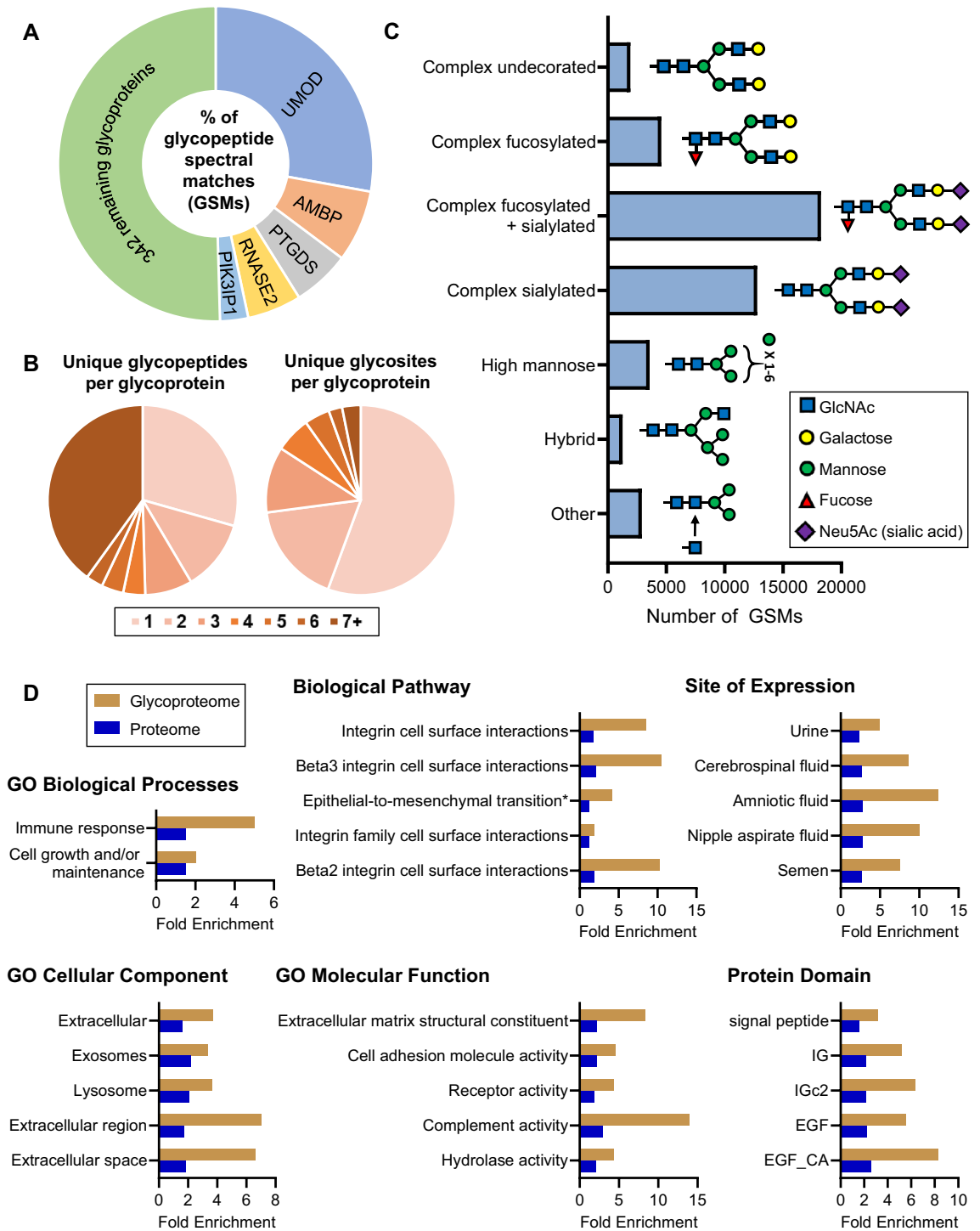


Figure 1. (A) Proportion of total identified glycopeptides (GSMs) corresponding to respective glycoproteins. (B) The number of unique glycopeptides identified per glycoprotein and the number of unique glycosites characterized per glycoprotein. (C) The number of glycopeptide spectral matches (GSMs) corresponding to different classes of glycans. An example putative glycan structure from each class is shown. (D) Gene enrichment analysis of identified urine glycoproteins compared to the urine proteome. All gene sets are significant (Bonferroni-corrected $p < 0.01$ in the urine glycoproteome and urine proteome unless otherwise noted). The top most significant gene sets in the urine glycoproteome are shown. *Not significant in the urine proteome.

and cell growth and/or maintenance were the most significantly enriched categories in the urine glycoproteome and showed higher fold enrichment compared to the urine proteome. Integrin-related cell surface interactions

and epithelial-to-mesenchymal transition were the most highly enriched biological pathways in the urine glycoproteome. Not surprisingly in the urine glycoproteome, “urine” was the most significantly (lowest adjusted p-value) enriched site of expression, with cerebrospinal fluid, amniotic fluid, nipple aspirate fluid, and semen also showing significant enrichment.

The GO cellular components showed significant enrichment in urine glycoproteins for extracellular, exosomes, lysosome, and extracellular region/space, an expected finding since secreted proteins are commonly glycosylated. Molecular functions showed a significant enrichment in the urine glycoproteome dataset including proteins associated with the extracellular matrix structural constituents, cell adhesion molecule activity, receptor activity, complement activity, and hydrolase activity. Protein domains related to signal peptide, immune response (e.g. IG and IGc2) and EGF (a domain present on cell surface proteins) were found to be significantly enriched in the urine glycoproteome.

Classification of glycoproteins by dominant glycan type

Glycoproteins can be classified by the predominant putative type of glycan (e.g. complex undecorated, complex fucosylated, complex sialylated, complex fucosylated + sialylated, high mannose, hybrid, or other). For each glycan type, the spectral counts were normalized to the total number of identified glycopeptides per protein and the glycoproteins were then clustered using a Pearson correlation analysis to display proteins with a significant correlation ($P < 0.01$) based on the glycan types (Fig. 2). For most of the glycoproteins (76%), complex structures were the predominant glycan type, and of the complex glycans, complex fucosylated and sialylated were most commonly observed (Figs. 1C and 2). Notably, four immunoglobulin proteins were found in the complex fucosylated glycoprotein cluster.

The only protein with a complex undecorated dominant glycan type was thrombospondin-1 (*THBS1*), an adhesive glycoprotein that mediates cell-to-cell and cell-to-matrix interactions²¹. As shown in Fig. 3, we detected 19 total glycopeptides, 84% of which are complex undecorated glycans, that map to a single glycosite (ASN1067) in the C-terminal region of thrombospondin-1, with the remaining glycans including complex fucosylated (11%) and hybrid (5%) types. According to Interpro and Gene Ontology cross-reference databases, this glycoprotein domain is present in proteins involved in calcium binding and cell adhesion.

Since the complex fucosylated and sialylated glycoproteins were the most common glycan type in the urine glycoproteome, we performed a gene enrichment analysis to compare these glycoproteins to the human proteome (Fig. 4A). We observed significant enrichment of proteins related to immune response. The complex fucosylated and sialylated glycoproteins were enriched for extracellular, exosomes, plasma membrane, and lysosome proteins, as well as molecular functions related to receptor activity.

Using the same approach, we compared complex sialylated glycoproteins, the second most abundant glycan type, to all human proteins (Fig. 4B). Once again, there was enrichment for glycoproteins involved in the immune response. Interestingly, complex sialylated proteins were enriched in similar cellular compartments compared to the complex fucosylated and sialylated glycoproteins, showing enrichment in extracellular, exosomes, and lysosome-related proteins. Complex sialylated proteins also showed significant enrichment in defense/immunity protein activity and protease inhibitor activity, as well as signal peptide and SERPIN (a specific type of protease inhibitor) domain.

When we compared predominantly high mannose glycoproteins to all human proteins (Fig. 4C), we did not observe significant enrichment in any GO biological processes, GO molecular functions, or protein domains. However, we did observe enrichment in subcellular locations of extracellular, exosomes, and lysosome.

Identification and characterization of protein glycosites

The information provided by intact glycoproteomics analysis allows the identification of the specific amino acid that is glycosylated in a protein (i.e. glycosite) and characterization of the glycan composition at the glycosite. In this study, we evaluated the meta-heterogeneity²² (i.e. the variation in glycosylation across multiple sites of a given protein), to determine whether dominant glycan types differed between individual glycosites within a single protein. We compared the observed glycan species for the overall protein to the glycans observed at each individual glycosite. Figure 5 shows proteins (with three or more characterized glycosites) that have significant meta-heterogeneity, as defined by the dominant glycan type on one or more glycosite(s) differing from the glycosylation information from combining information across glycosites for a given protein. For example, cubilin (CUBN), an endocytic receptor important in metabolism by facilitating the uptake of lipoproteins, vitamins and iron^{23–27}, shows predominantly complex fucosylated glycans when the protein is viewed as a whole (Fig. 6A). However, when examining the individual glycosites, CUBN showed a high degree of meta-heterogeneity with the dominant glycans for N₇₈₁/N₁₈₀₂, N₂₄₀₀, N₂₉₂₃, and N₃₄₅₇ dominated by complex fucosylated, complex sialylated, high mannose, and complex undecorated glycans respectively.

An additional example of a protein exhibiting meta-heterogeneity in glycosylation is uromodulin (UMOD). UMOD is the most abundant glycoprotein in urine and contributes to colloid osmotic pressure, retards passage of positively charged electrolytes, prevents urinary tract infections, and inhibits formation of liquid containing supersaturated salts and subsequent formation of salt crystals²⁸. We identified 12,614 total glycopeptides corresponding to 1485 unique glycopeptides, containing 269 unique glycan structures that mapped to five glycosites on UMOD (Fig. 6B). The overall protein was found to have the highest percentage of sialylated and complex fucosylated glycans, consistent with a recent study specifically characterizing glycans from UMOD in urine²⁹ and which was consistent with four of the five glycosites. However, N₂₇₅ exhibited a strikingly different pattern with 71% of all glycopeptide identifications contain high mannose structures. Interestingly, glycan structures at this site differed from a previous study³⁰ which identified only high mannose structures at N₂₇₅. To determine if the difference in glycosylation on this specific site may be due to amino acid's location on the three dimensional

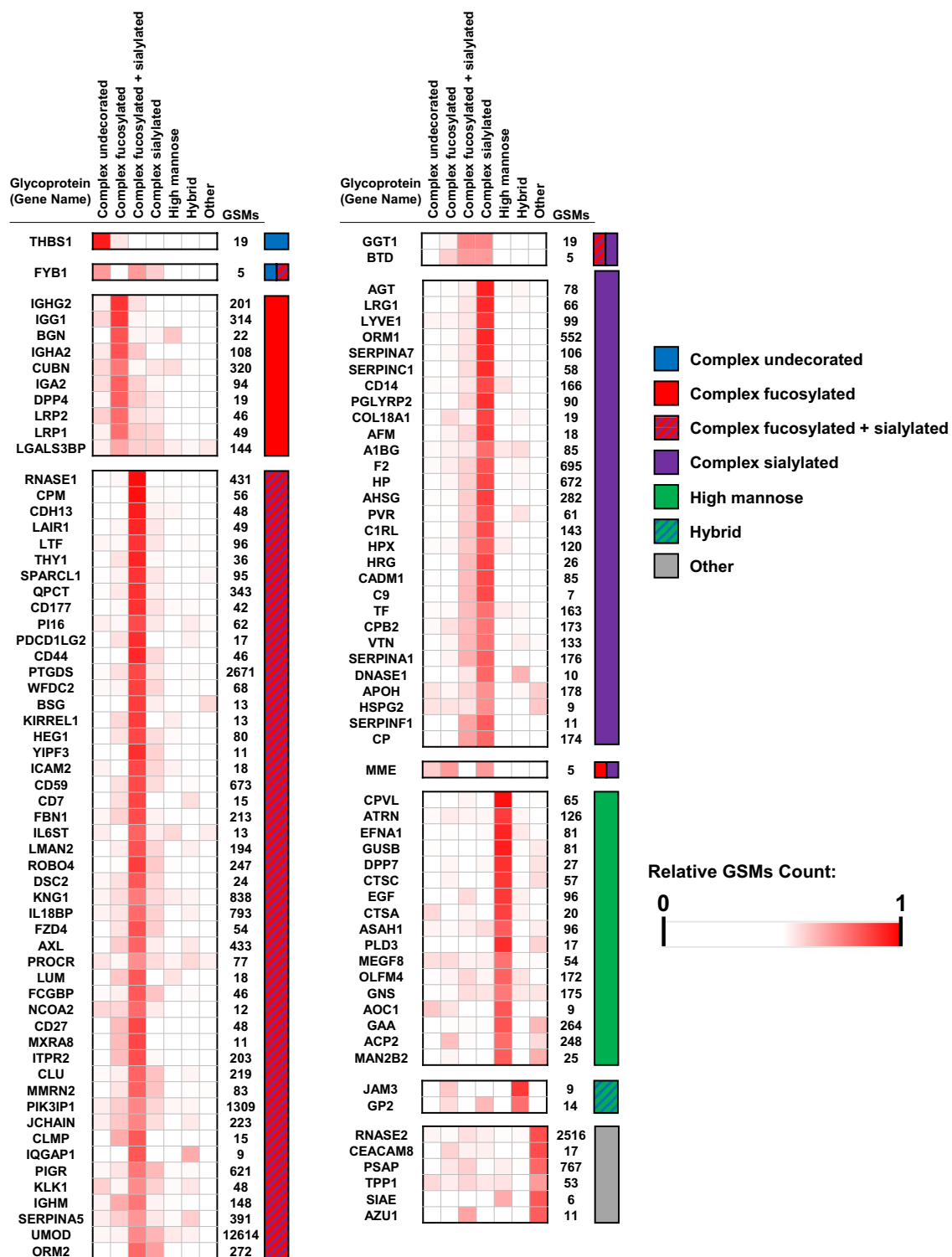


Figure 2. Heatmaps showing the distributions of glycan types identified on proteins. The relative intensity of each glycan class is calculated based on the number of GSMs for the corresponding glycopeptides. Proteins are clustered by dominant glycan class.

structure of the protein, we calculated the residue depth³¹ of each glycosite as shown in Fig. 6B. Interestingly, N₂₇₅ had a larger residue depth than the other four glycosites which were mapped to more solvent-accessible areas on the protein. These results are consistent with the previous observation that protein structure dictates formation of N-glycan type³². Therefore, it is possible that the dominant high mannose glycans on N₂₇₅ may be, at least partially explained by the limited accessibility of the glycosite therefore restricting access to the glycans at that site by glycosyltransferases.

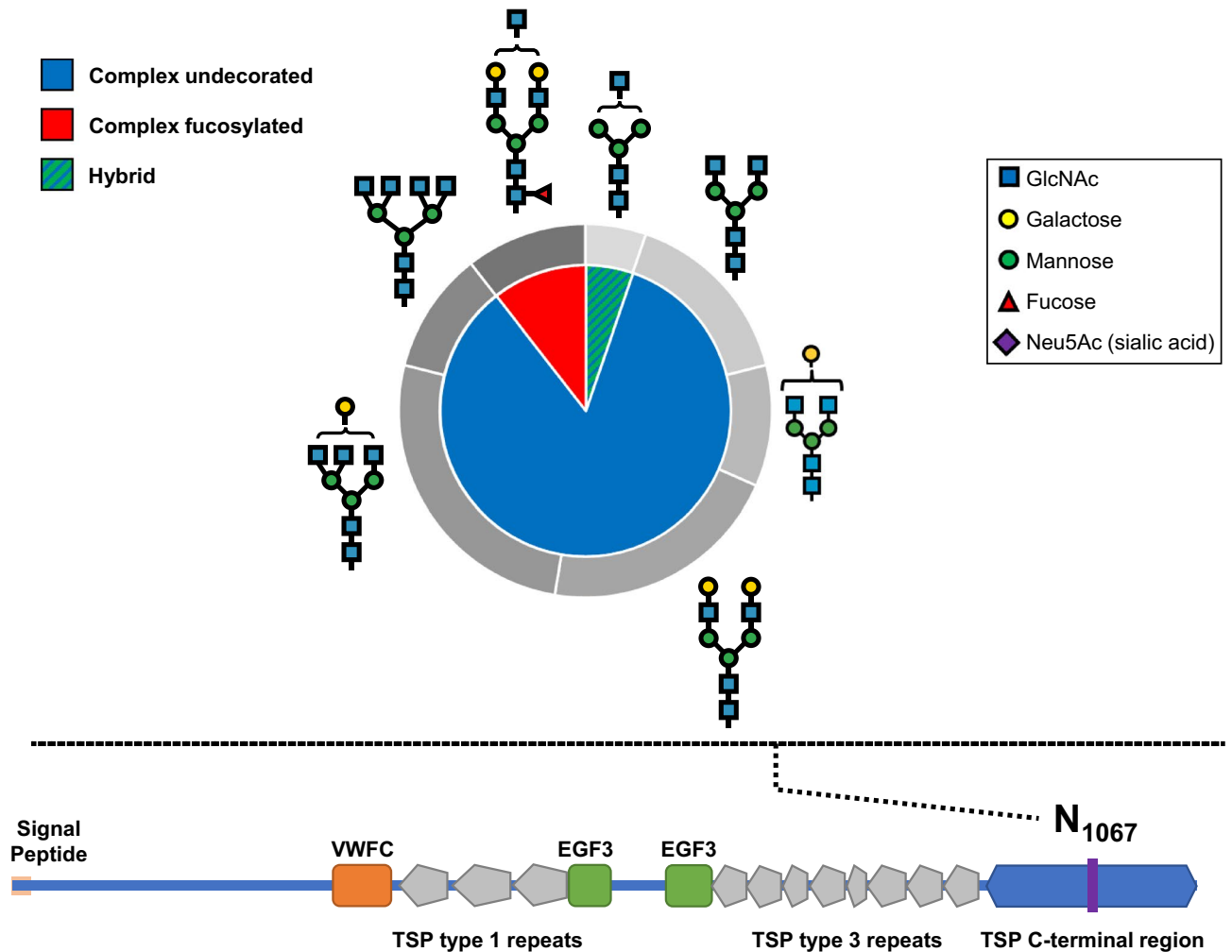


Figure 3. Glycans identified on asparagine-1067 (N_{1067}) of thrombospondin (THBS1). Classes of glycans are shown in the inner pie chart and the breakdown of glycans corresponding to the respective classes are shown in the outer doughnut chart. A schematic of the THBS1 protein sequence is shown at the bottom with the different protein domains and the purple line indicating the location of N_{1067} .

Alterations in glycosylation of proteins such as cubilin and uromodulin can have profound implications for protein structural conformation, molecular interactions, and biological functions due to the pivotal role of glycosylation in protein folding, stability, trafficking, and receptor-ligand interactions³. In the case of cubilin, a protein expressed in renal and intestinal cells, and a receptor involved in renal reabsorption and cellular transport, changes in glycosylation patterns can impact its ligand-binding affinity and transport efficiency³³. Modulations in the glycosylation profile of cubilin may disrupt the proper recognition and binding of specific ligands, potentially compromising its role in renal transport processes and overall renal function³³. Similarly, uromodulin, is a glycoprotein exclusively produced in the kidney, where it participates in urine concentration regulation and kidney defense mechanisms³⁴. Alterations in uromodulin glycosylation can affect its structural stability, intracellular trafficking, and interactions with other urinary components and proteins. These modifications may influence uromodulin's polymerization, ion transport regulation, and involvement in immune responses within the kidney.

The specific consequences of glycosylation changes in cubilin and uromodulin are likely dependent on the nature and site of the altered glycans. These changes can result in functional modifications, modulated protein-protein interactions, modified receptor binding affinities, and potential effects on intracellular signaling pathways. Elucidating the precise effects of these glycosylation alterations is crucial for comprehending the underlying mechanisms and their implications for physiological processes and diseases.

In summary, urine is a highly attractive sample type for developing clinical assays, and this study describes deep analysis of the urine glycoproteome spanning more than four orders of magnitude of dynamic range of protein abundance. We describe the protein composition of the urine glycoproteome and how that differs from the overall urine proteome. This study also provides detailed characterization of glycans on specific glycosites of identified proteins and examples of meta-heterogeneity in glycosylation are shown with possible explanation by residue depth. This dataset may also find utility as a resource for future studies as a baseline characterization of pooled normal human urine.

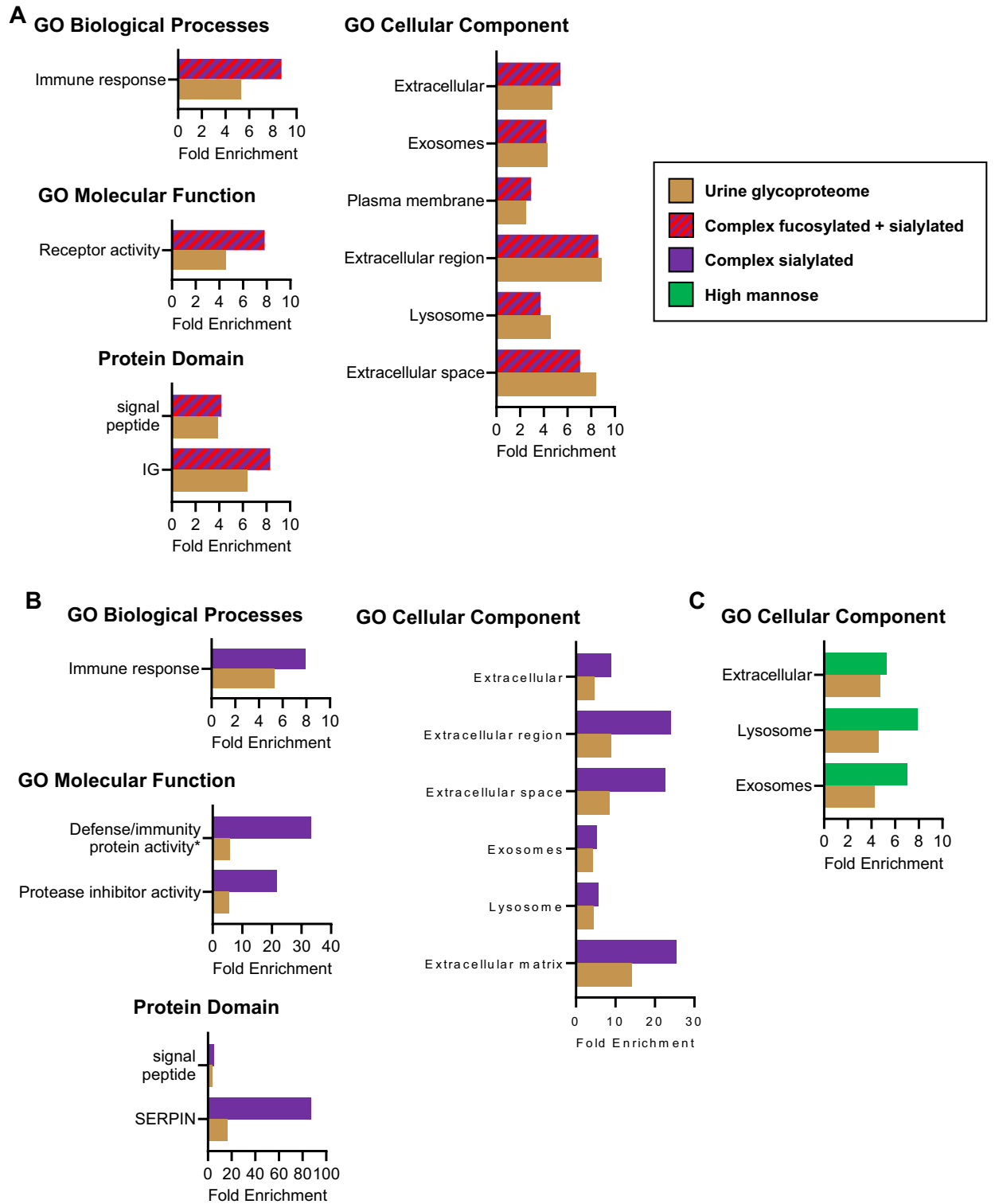


Figure 4. Gene enrichment analysis of glycoproteins with dominant (A) complex fucosylated and sialylated glycans, (B) complex sialylated glycans, and (C) high mannose glycans, compared to the overall urine glycoproteome. All gene sets are significant (Bonferroni-corrected $p < 0.01$ in the urine glycoproteome and urine sub-glycoproteome) unless otherwise noted. The top most significant gene sets in each urine subglycoproteome are shown. *Not significant in the urine glycoproteome.

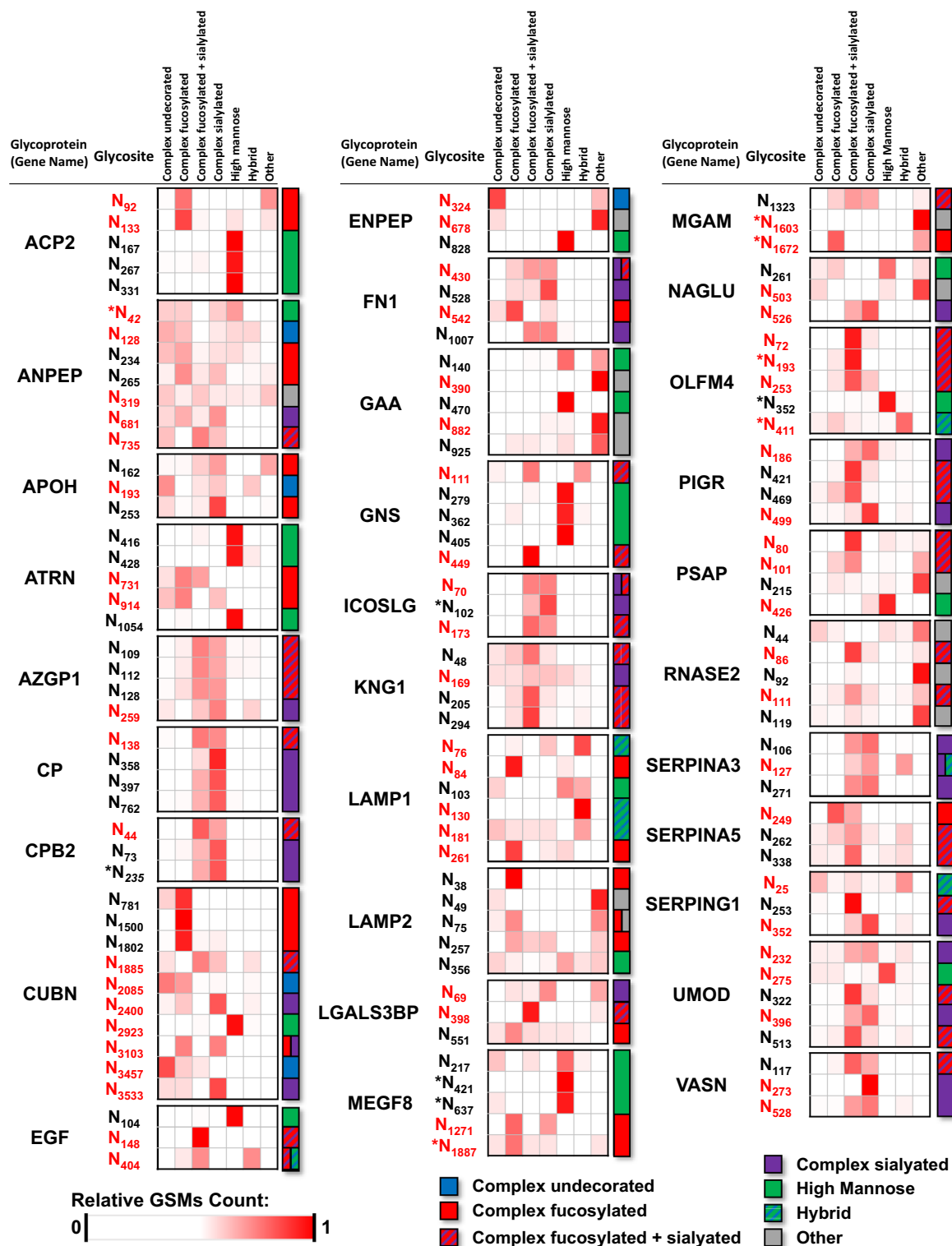


Figure 5. Heatmaps showing the distributions of glycan class by specific protein glycosite. The relative intensity of each glycan class is calculated based on the number of GSMs for the corresponding glycopeptide. Sites in red indicate that the dominant glycan class at the site is different than the dominant glycan class for the overall protein. *Indicates that the glycosite is not annotated in UniProt.

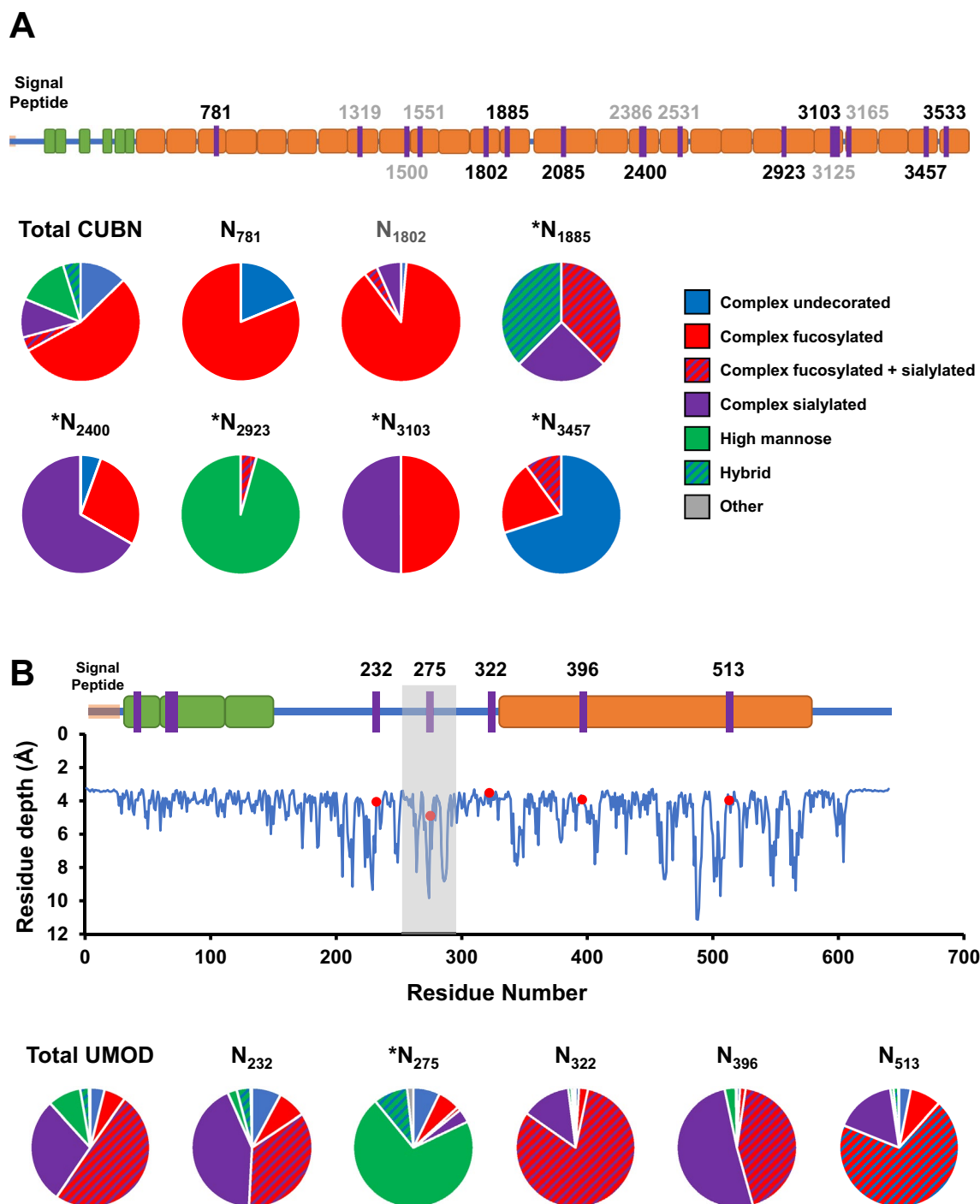


Figure 6. (A) Top: schematic of protein sequence for cubilin (CUBN). Glycosites are indicated by purple lines and numbers with gray numbers referring to glycosites that were not identified with more than two total glycopeptides. Green and orange indicate EGF-like and CUB domains respectively. Bottom: Pie charts indicate the composition of glycan types identified for the overall protein (Total) and at the individual sites. *Indicates that the dominant glycan type at the site differs from the overall dominant glycan type on the protein. (B) Top: schematic of protein sequence for uromodulin (UMOD). Red indicates ZP domain in UMOD. Middle: calculated residue depth for amino acids in UMOD. Red dots indicate characterized glycosites. Unless otherwise specified, information in (A) applies to (B).

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD038923.

Received: 31 May 2023; Accepted: 30 January 2024

Published online: 14 February 2024

References

1. Reily, C., Stewart, T. J., Renfrow, M. B. & Novak, J. Glycosylation in health and disease. *Nat. Rev. Nephrol.* **15**, 346–366. <https://doi.org/10.1038/s41581-019-0129-4> (2019).
2. Aebersold, R. *et al.* How many human proteoforms are there?. *Nat. Chem. Biol.* **14**, 206–214. <https://doi.org/10.1038/nchembio.2576> (2018).
3. Schjoldager, K. T., Narimatsu, Y., Joshi, H. J. & Clausen, H. Global view of human protein glycosylation pathways and functions. *Nat. Rev. Mol. Cell Biol.* **21**, 729–749. <https://doi.org/10.1038/s41580-020-00294-x> (2020).
4. Wenzel, D. M. & Olivier-Van Stichelen, S. The O-GlcNAc cycling in neurodevelopment and associated diseases. *Biochem. Soc. Trans.* <https://doi.org/10.1042/BST20220539> (2022).
5. Pan, J. *et al.* Glycoproteomics-based signatures for tumor subtyping and clinical outcome prediction of high-grade serous ovarian cancer. *Nat. Commun.* **11**, 6139. <https://doi.org/10.1038/s41467-020-19976-3> (2020).
6. Kawahara, R. *et al.* The complexity and dynamics of the tissue glycoproteome associated with prostate cancer progression. *Mol. Cell Proteom.* **20**, 100026. <https://doi.org/10.1074/mcp.RA120.002320> (2021).
7. Chen, J., Xi, J., Tian, Y., Bova, G. S. & Zhang, H. Identification, prioritization, and evaluation of glycoproteins for aggressive prostate cancer using quantitative glycoproteomics and antibody-based assays on tissue specimens. *Proteomics* **13**, 2268–2277. <https://doi.org/10.1002/pmic.201200541> (2013).
8. Liu, Y. *et al.* Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovers N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness. *Mol. Cell Proteomics* **13**, 1753–1768. <https://doi.org/10.1074/mcp.M114.038273> (2014).
9. Luo, G. *et al.* Roles of CA19-9 in pancreatic cancer: Biomarker, predictor and promoter. *Biochim. Biophys. Acta* **1875**, 188409. <https://doi.org/10.1016/j.bbcan.2020.188409> (2021).
10. Gratacos-Mulleras, A. *et al.* Characterisation of the main PSA glycoforms in aggressive prostate cancer. *Sci. Rep.* **10**, 18974. <https://doi.org/10.1038/s41598-020-75526-3> (2020).
11. Llop, E. *et al.* Improvement of prostate cancer diagnosis by detecting PSA glycosylation-specific changes. *Theranostics* **6**, 1190–1204. <https://doi.org/10.7150/thno.15226> (2016).
12. Sun, S. *et al.* N-GlycositeAtlas: A database resource for mass spectrometry-based human N-linked glycoprotein and glycosylation site mapping. *Clin. Proteomics* **16**, 35. <https://doi.org/10.1186/s12014-019-9254-0> (2019).
13. Kawahara, R., Saad, J., Angeli, C. B. & Palmisano, G. Site-specific characterization of N-linked glycosylation in human urinary glycoproteins and endogenous glycopeptides. *Glycoconj. J.* **33**, 937–951. <https://doi.org/10.1007/s10719-016-9677-z> (2016).
14. Belczacka, I. *et al.* Urinary glycopeptide analysis for the investigation of novel biomarkers. *Proteom. Clin. Appl.* **13**, e1800111. <https://doi.org/10.1002/prca.201800111> (2019).
15. Shen, Y., Xiao, K. & Tian, Z. Site- and structure-specific characterization of the human urinary N-glycoproteome with site-determining and structure-diagnostic product ions. *Rapid Commun. Mass Spectrom.* **35**, e8952. <https://doi.org/10.1002/rcm.8952> (2021).
16. Chen, S. Y. *et al.* Glycans, glycosite, and intact glycopeptide analysis of N-linked glycoproteins using liquid handling systems. *Anal. Chem.* **92**, 1680–1686. <https://doi.org/10.1021/acs.analchem.9b03761> (2020).
17. Nakatani, A. I., Mohler, C. E. & Hughes, S. Chain conformation of polymers adsorbed to clay particles: Effects of charge and concentration. *Soft Matter* **17**, 6848–6862. <https://doi.org/10.1039/d1sm00674f> (2021).
18. Liu, L., Qin, H. & Ye, M. Recent advances in glycopeptide enrichment and mass spectrometry data interpretation approaches for glycoproteomics analyses. *Se Pu* **39**, 1045–1054. <https://doi.org/10.3724/SPJ.1123.2021.06011> (2021).
19. Bermudez, A. & Pitteri, S. J. Enrichment of intact glycopeptides using strong anion exchange and electrostatic repulsion hydrophilic interaction chromatography. *Methods Mol. Biol.* **2271**, 107–120. https://doi.org/10.1007/978-1-0716-1241-5_8 (2021).
20. Zhao, M. *et al.* A comprehensive analysis and annotation of human normal urinary proteome. *Sci. Rep.* **7**, 3024. <https://doi.org/10.1038/s41598-017-03226-6> (2017).
21. Simantov, R. *et al.* Histidine-rich glycoprotein inhibits the antiangiogenic effect of thrombospondin-1. *J. Clin. Invest.* **107**, 45–52. <https://doi.org/10.1172/JCI9061> (2001).
22. Caval, T., Heck, A. J. R. & Reiding, K. R. Meta-heterogeneity: Evaluating and describing the diversity in glycosylation between sites on the same glycoprotein. *Mol. Cell Proteom.* **20**, 100010. <https://doi.org/10.1074/mcp.R120.002093> (2021).
23. Nykjaer, A. *et al.* Cubilin dysfunction causes abnormal metabolism of the steroid hormone 25(OH) vitamin D(3). *Proc. Natl. Acad. Sci. U S A* **98**, 13895–13900. <https://doi.org/10.1073/pnas.241516998> (2001).
24. Kozyraki, R. *et al.* The human intrinsic factor-vitamin B12 receptor, cubilin: Molecular characterization and chromosomal mapping of the gene to 10p within the autosomal recessive megaloblastic anemia (MGA1) region. *Blood* **91**, 3593–3600 (1998).
25. Kozyraki, R. *et al.* Megalin-dependent cubilin-mediated endocytosis is a major pathway for the apical uptake of transferrin in polarized epithelia. *Proc. Natl. Acad. Sci. U S A* **98**, 12491–12496. <https://doi.org/10.1073/pnas.211291398> (2001).
26. Kozyraki, R. *et al.* The intrinsic factor-vitamin B12 receptor, cubilin, is a high-affinity apolipoprotein A-I receptor facilitating endocytosis of high-density lipoprotein. *Nat. Med.* **5**, 656–661. <https://doi.org/10.1038/9504> (1999).
27. Fyfe, J. C. *et al.* The functional cobalamin (vitamin B12)-intrinsic factor receptor is a novel complex of cubilin and amnionless. *Blood* **103**, 1573–1579. <https://doi.org/10.1182/blood-2003-08-2852> (2004).
28. Bates, J. M. *et al.* Tamm-Horsfall protein knockout mice are more prone to urinary tract infection: Rapid communication. *Kidney Int.* **65**, 791–797. <https://doi.org/10.1111/j.1523-1755.2004.00452.x> (2004).
29. Li, H. *et al.* Uromodulin isolation and its N-glycosylation analysis by nanoLC-MS/MS. *J. Proteome Res.* **20**, 2662–2672. <https://doi.org/10.1021/acs.jproteome.0c01053> (2021).
30. Weiss, G. L. *et al.* Architecture and function of human uromodulin filaments in urinary tract infections. *Science* **369**, 1005–1010. <https://doi.org/10.1126/science.aaz9866> (2020).
31. Xu, D., Li, H. & Zhang, Y. Protein depth calculation and the use for improving accuracy of protein fold recognition. *J. Comput. Biol.* **20**, 805–816. <https://doi.org/10.1089/cmb.2013.0071> (2013).
32. Thaysen-Andersen, M. & Packer, N. H. Site-specific glycoproteomics confirms that protein structure dictates formation of N-glycan type, core fucosylation and branching. *Glycobiology* **22**, 1440–1452. <https://doi.org/10.1093/glycob/cws110> (2012).
33. Udagawa, T. *et al.* Amnionless-mediated glycosylation is crucial for cell surface targeting of cubilin in renal and intestinal cells. *Sci. Rep.* **8**, 2351. <https://doi.org/10.1038/s41598-018-20731-4> (2018).
34. Patabandige, M. W., Go, E. P. & Desaire, H. Clinically viable assay for monitoring uromodulin glycosylation. *J. Am. Soc. Mass Spectrom.* **32**, 436–443. <https://doi.org/10.1021/jasms.0c00317> (2021).

Acknowledgements

This work is supported by National Institutes of Health U54 DK130065 to JDB and U01 CA226051 to SJP and JDB; and by the International Alliance for Cancer Early Detection (SJP, MRE, and JDB).

Author contributions

Conception and design (A.B., K.F., S.J.P.), acquisition of data (K.F., A.B., N.S.), analysis and interpretation of data (F.G.M., K.F., A.B., N.S.), writing, reviewing, and revision of manuscript (F.G.M., K.F., A.B., N.S., H.Z., J.D.B., M.R.F., S.J.P.).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-53299-3>.

Correspondence and requests for materials should be addressed to S.J.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024