Journal of proteome research

Cite This: J. Proteome Res. 2019, 18, 3985−3998

Article

pubs.acs.org/jpr

Diagnostic and Prognostic Performance of Blood Plasma Glycan Features in the Women Epidemiology Lung Cancer (WELCA) Study

Yueming Hu,[†] Shadi Ferdosi,[†,∥] Erandi P. Kapuruge,[†] Jesús Aguilar Diaz de Leon,[†] Isabelle Stücker,[‡] Loredana Radoï,[‡,§] Pascal Guénel,[‡] and Chad R. Borges*[,†]

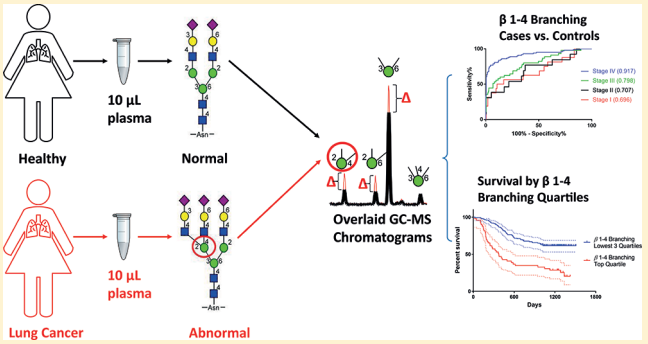
[†]School of Molecular Sciences and The Biodesign Institute, Arizona State University, Tempe, Arizona 85287, United States

[‡]CESP (Center for Research in Epidemiology and Population Health), Cancer and Environment Team, INSERM UMS1018, University Paris-Sud, University Paris-Saclay, 94800 Villejuif, France

[§]Faculty of Dental Surgery, University Paris Descartes, 75006 Paris, France

Supporting Information

ABSTRACT: Lung cancer is the leading cause of cancer death in women living in the United States, which accounts for approximately the same percentage of cancer deaths in women as breast, ovary, and uterine cancers combined. Targeted blood plasma glycomics represents a promising source of noninvasive diagnostic and prognostic biomarkers for lung cancer. Here, 208 samples from lung cancer patients and 207 age-matched controls enrolled in the Women Epidemiology Lung Cancer (WELCA) study were analyzed by a bottom-up glycan "node" analysis approach. Glycan features, quantified as single analytical signals, including 2-linked mannose, α2−6 sialylation, β1−4 branching, β1−6 branching, 4-linked GlcNAc, and antennary fucosylation, exhibited abilities to distinguish cases from controls (ROC AUCs: 0.68−0.92) and predict survival in patients (hazard ratios: 1.99−2.75) at all stages. Notable alterations of glycan features were observed in stages I−II. Diagnostic and prognostic glycan features were mostly independent of smoking status, age, gender, and histological subtypes of lung cancer.



KEYWORDS: *glycans, lung cancer, women, diagnostic, prognostic, survival, plasma, sialylation, fucosylation, branching*

INTRODUCTION

Lung cancer accounts for approximately 25% of all U.S. cancer deaths, making it the leading cause of U.S. cancer deaths.[1] More than half of lung cancer patients are diagnosed at an advanced stage: about 33% and 40% of lung cancer patients are diagnosed at stage IIIB and IV, respectively,[2] primarily due to a lack of early stage symptoms. The five-year survival rate of stage IV patients is only ∼5%.[1] Conversely, if lung cancer can be detected before it escapes the lungs, five-year survival rates usually exceed 50%.[1] Therefore, to improve the outcomes of lung cancer patients, a major clinical priority is to detect lung cancer early. Recently, the National Lung Screening Trial (NLST) applied low dose chest computed tomography (LDCT) in older, high-risk individuals and achieved 20% reduction in lung cancer mortality. Yet the positive screening rate in this study was 24.2%, of which 96.4% were false-positive results.[3] The high false-positive rate may lead to additional clinical tests, emotional distress, and unnecessary treatments, as well as unnecessary time and costs spent. Thus, a reliable and highly specific noninvasive blood test could help to reduce the false-positive and overdiagnosis rate of CT scans.

Biomarkers from easily accessible biofluids, such as blood plasma or serum (P/S), could potentially be used as a noninvasive and cost-effective way to improve lung cancer diagnosis and screening. Numerous P/S biomarkers for lung cancer have been extensively studied, including proteins (such as cytokeratin 19 fragments[4,5] and carcinoembryonic antigen[6,7]), miRNAs (such as miR-34[8] and miR-182[9,10]), methyl-DNA (such as P16[11] and BRMS1[12]), and circulating tumor cells.[13] However, biomarkers with improved sensitivity and specificity are still needed.

Aberrant glycosylation is a well-established hallmark of cancer and seems to facilitate the metastasis of various tumor cells.[14] Thus, blood P/S glycomics represents a promising source for a new generation of cancer biomarkers. At present, almost all P/S glycomics studies focus on the analysis of intact glycans—primarily N-linked glycans, with O-linked and lipid-linked glycans usually excluded. Generally, a great many intact glycan structures need to be investigated in order to fully capture and quantify the cancer-specific behavior of one unique glycan feature, such as core fucosylation, α2−6 sialylation, or β1−4 branching.[15] Glycan node analysis is a molecularly bottom-up approach to P/S glycomics developed by Borges et al. in 2013 that focuses on monosaccharides and linkage specific glycan

Received: July 8, 2019
Published: September 30, 2019

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

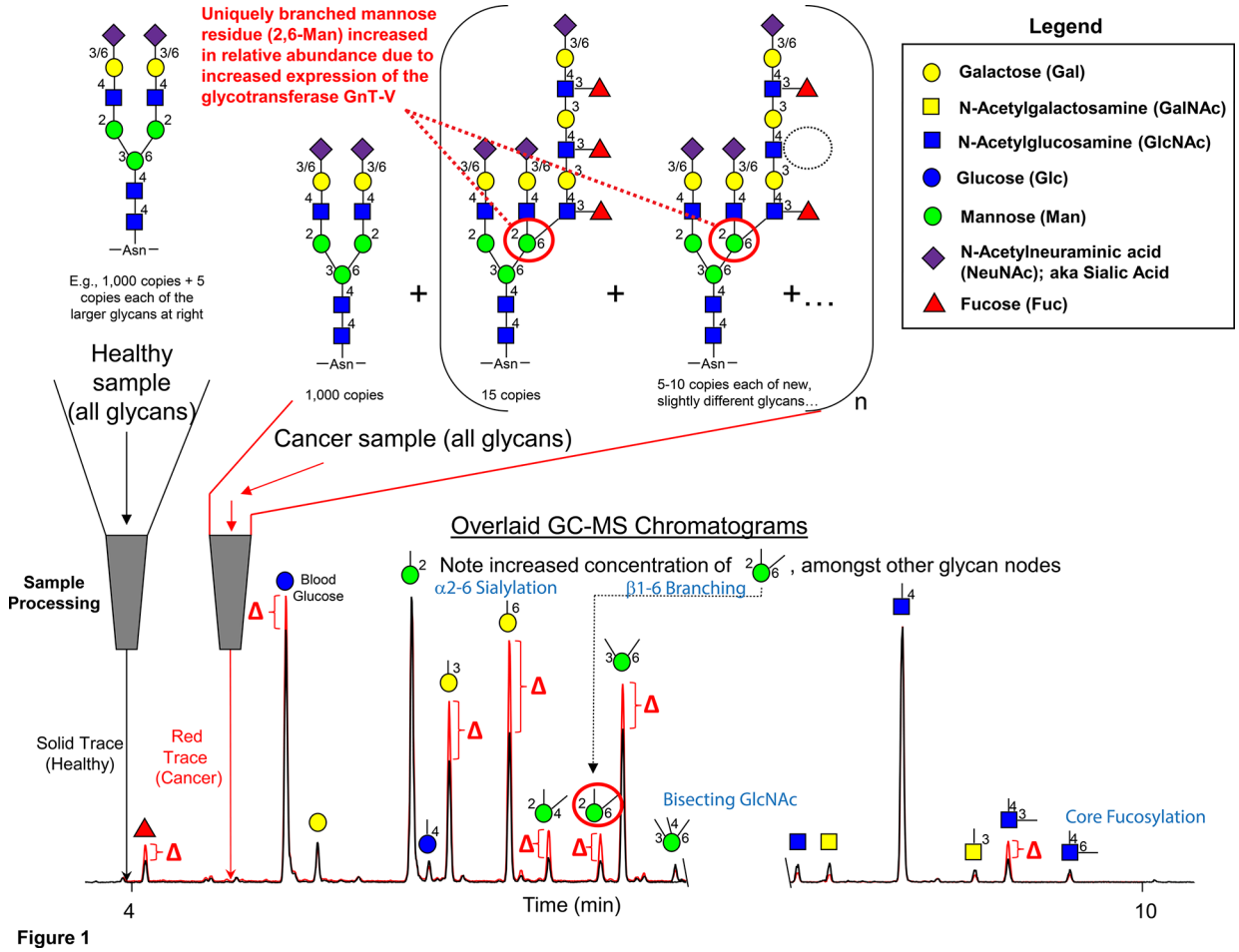


Figure 1. Conceptual overview of the glycan "node" analysis concept, which essentially consists of applying glycan linkage (methylation) analysis to whole biofluids. Intact normal and abnormal glycans including O-glycans, N-glycans, and glycolipids are processed and transformed into partially methylated alditol acetates (PMAAs, Figure 2), each of which corresponds to a particular monosaccharide-and-linkage-specific glycan "node" in the original polymer. As illustrated, analytically pooling together the glycan nodes from among all the aberrant intact glycan structures provides a more direct surrogate measurement of abnormal glycosyltransferase activity than any individual intact glycan while simultaneously converting unique glycan features such as "core fucosylation", "α2−6 sialylation", "bisecting GlcNAc", and "β1−6 branching" into single analytical signals. Actual extracted ion chromatograms from 9-μL blood plasma samples are shown. Numbers adjacent to monosaccharide residues in glycan structures indicate the position at which the higher residue is linked to the lower residue. This figure was adapted with permission from ref 16. Copyright 2013 American Chemical Society.

"nodes" rather than the intact glycan structures.[16−20] This approach captures all P/S glycans including N-, O-, and lipid-linked glycans and breaks them down into monosaccharides that maintain their original linkage information. In short, the method involves the application of glycan linkage (methylation) analysis to whole biofluids (Figures 1 and 2). Uniquely in this approach, linkage-related glycan features are captured and quantified as single analytical signals, rather than being spread across numerous intact glycans that bear the specific feature. For example, 6-linked galactose and 2,6-linked mannose, corresponding to α2−6 sialylation and β1−6 branching, respectively, are both captured as single chromatographic peak areas (Figure 1). In addition, numerous glycan nodes serve as direct surrogates for the activities of specific glycosyltransferases (GTs)—enzymes that facilitate the construction of glycans.

Recently, we have applied glycan node analysis to several cancer case-control studies, including pancreatic,[19] ovarian,[19] prostate,[19] bladder,[20] breast,[18] and lung[16,19] cancer cohorts. The purpose of this study was to further validate glycan node analysis as a means of detecting and predicting patient outcomes in lung cancer. In addition, though glycan node analysis has been analytically validated in the past,[16,17] we felt it was important to conduct a more comprehensive stability study than that which we have previously reported. The cohort of specimens to which we had access that most readily lent itself to addressing both of these goals was from a study of lung cancer in women. Interestingly, there are several important gender differences in lung cancer, including the facts that (1) after adjusting for the number of cigarettes smoked, women have a 3-fold greater risk of lung cancer than men,[21−24] (2) never-smoker women are at significantly greater risk for lung cancer than men,[25] and (3) women tend to have better survival rates than men.[26,27] As such, we felt that for any differences observed in this study relative to our previously reported results in lung cancer,[19] it would also be important to look for any existing gender-based differences in glycan nodes as they may occur in the context of lung cancer.

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

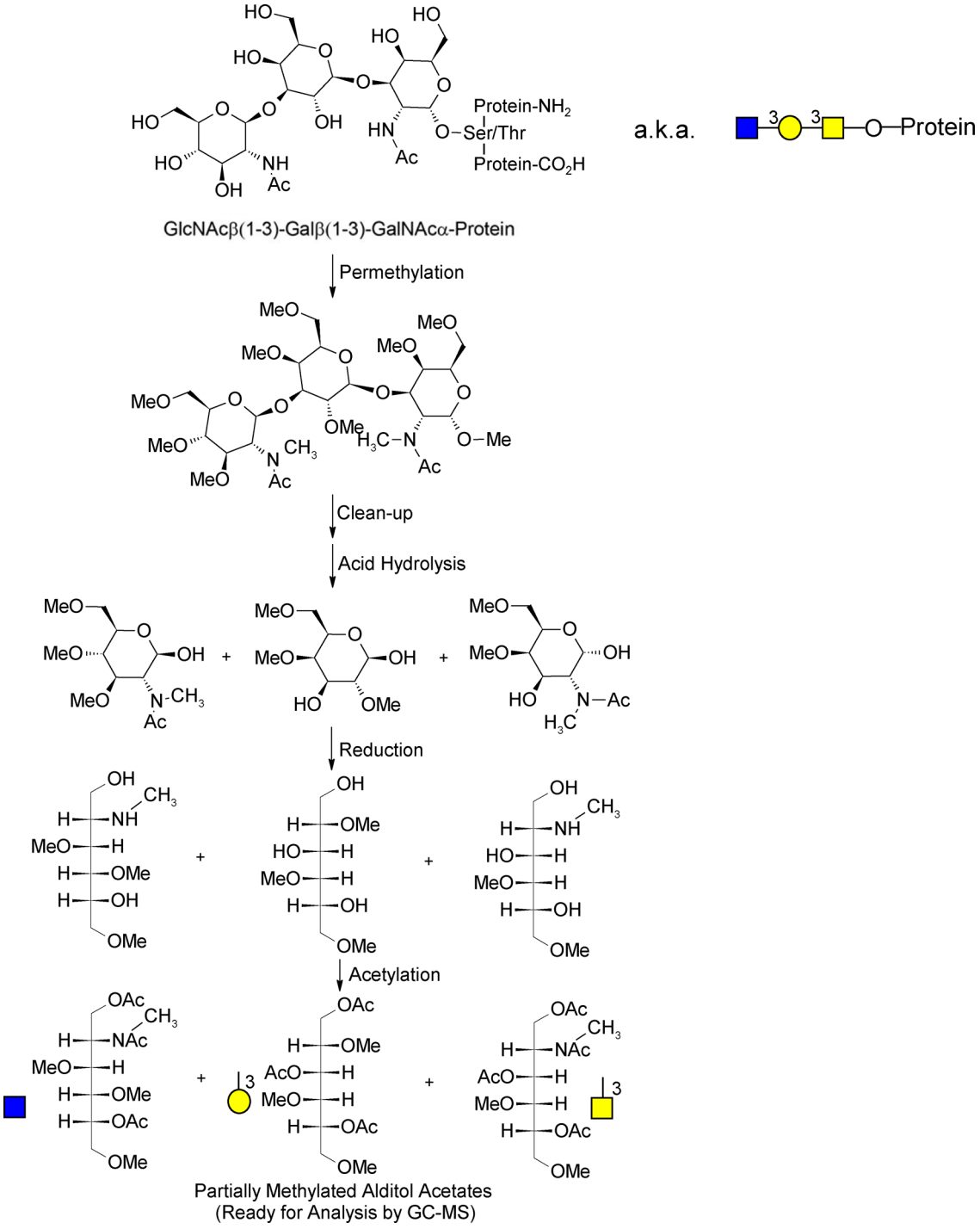


Figure 2. Molecular overview of the glycan "node" analysis procedure. For glycans from blood plasma and other biofluids, O-linked glycans are released during permethylation, while N-linked glycans and glycolipids are released during acid hydrolysis. The unique pattern of methylation and acetylation in the final partially methylated alditol acetates (PMAAs) corresponds to the unique glycan "node" in the original glycan polymer and provides the molecular basis for separation and quantification by GC−MS. Figure adapted with permission from ref 16. Copyright 2013 American Chemical Society.

MATERIALS AND METHODS

Materials and Samples

Materials. Heavy, stable-isotope-labeled D-glucose (U−$^{13}C_6$, 99%; 1,2,3,4,5,6,6-D7, 97−98%) was obtained from Cambridge Isotope Laboratories (Tewksbury, MA). Acetone was acquired from Avantor Performance Materials (Center Valley, PA). Methanol was purchased from Honeywell Burdick & Jackson (Muskegon, MI). Acetonitrile and methylene chloride were obtained from Fisher Scientific (Fair Lawn, NJ). Dimethyl sulfoxide (DMSO), iodomethane (99%, Cat. No. I8507), chloroform, trifluoroacetic acid (TFA), ammonium hydroxide, sodium borohydride, acetic anhydride, sodium acetate, and sodium hydroxide beads (20−40 mesh, Cat. No. 367176) were acquired from Sigma-Aldrich. Pierce spin columns (900 μL volume) were purchased from ThermoFisher Scientific

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

(Waltham, MA, Cat. No. 69705). GC−MS autosampler vials and Teflon-lined pierceable caps were obtained from ThermoFisher Scientific. GC consumables were acquired from Agilent (Santa Clara, CA); MS consumables were obtained from Waters (Milford, MA).

Plasma and Serum Samples. All specimens were collected in compliance with the Declaration of Helsinki principles. Once collected, they were coded and deidentified to protect patient identities.

Women Epidemiology Lung Cancer (WELCA) Set. EDTA plasma samples from stage I−IV lung cancer patients and age-matched controls were collected at 12 different collection centers in France.[26] This study was approved by the Institutional Review Board of the French National Institute of Health and Medical Research and by the French Data Protection Authority (IRB-Inserm, no. 3888 and CNIL no. C13-52). As part of the WELCA Study, all-female lung cancer patients were recruited between September 2014 and December 2017, and age-matched all-female controls were recruited between June 2015 and December 2017. All women living in Paris and the lle de France area, newly diagnosed with lung cancer, were considered as eligible cases. Age-matched controls were randomly sampled from women living in the same area without a history of lung cancer. All peripheral blood samples were drawn and processed following a written standardized protocol.[26] Briefly, after transport to the laboratory at 4 °C, blood samples collected in tubes containing EDTA additive were spun for 15 min at 3000 rpm and 4 °C in a standard centrifuge. Then the collected plasma samples were aliquoted and periodically transported on dry ice to the central repository for final storage at −80 °C. No freeze−thaw cycles occurred prior to shipment to Arizona State University (Borges lab) for analysis. A detailed profile of the clinical characteristics of the patients in this WELCA study is given in Table S1.

Dual Gender Lung Cancer Set. Sodium heparin plasma samples from a lung cancer study consisting of patients and controls in both genders were collected by Dr. Xifeng Wu at the University of Texas MD Anderson Cancer Center. Even though it is a glycosaminoglycan itself, heparin possesses monomer units that are predominately carboxylated, sulfated or both, and thus cannot be directly detected by the analytical methodology used in this study. As we have reported previously, there are only negligible differences between glycan nodes measure in heparin plasma vs EDTA plasma or serum,[19] and thus direct comparisons were made for these three types of biospecimens. Venous blood samples were collected from newly diagnosed and histologically confirmed lung cancer patients prior to therapy at the MD Anderson Cancer Center hospital. Blood samples of age-, gender-, smoking-, and ethnicity-matched controls were collected at the Kelsey-Seybold Clinic. All blood samples were collected since 1995 and processed following the same SOP. These specimens has previously been described.[19]

Stage I-Only Lung Cancer Set (Also Dual Gender). Serum samples for dual gender stage I lung adenocarcinoma patients were collected together with age-, gender-, and smoking-status-matched controls, under NYU IRB approval at the NYU Langone Medical Center by Dr. Harvey Pass. Arterial blood samples were drawn from fasting patients undergoing surgery between September 2006 to August 2013 to remove one or more lung nodules that were detected during a CT scan. A pathological exam of the excised nodules was performed to determine whether nodules were benign or malignant. Serum was collected under a standardized procedure. These specimens have previously been described.[19]

Plasma Samples for the Stability Study. The samples employed for the ex vivo thawed-state stability study included EDTA plasma samples from three healthy male and two healthy female donors. These samples were aliquoted and stored at different temperatures over the course of a year, with their matched control aliquots stored continuously at −80 °C. The mistreatment conditions included 10 days at −20 °C, 90 days at −20 °C, 360 days at −20 °C, 2 days at 4 °C, 90 days at 4 °C, and 1 day at 25 °C. At the end of the 360-day time point, glycan node analysis was performed on all the mistreated sample aliquots and their matched control aliquots.

Additional Biospecimen Details. A summary of the case-control sample sets discussed in this study is provided in Table S2. A 300 mL plasma sample from an individual donor was obtained from BioIVT, which served as a quality control sample to ensure batch-to-batch quantitative reproducibility. All specimens were stored at −80 °C prior to analysis.

Experimental Procedures

The glycan node analysis procedure was adapted from Borges et al.[16,17]

Permethylation, Nonreductive Release, and Purification of Glycans. Nine microliters (9 μL) of blood plasma and 1 μL of a 5 mM solution of heavy-labeled D-glucose (U-$^{13}C_6$, 99%; 1,2,3,4,5,6,6-D7, 97−98%) and *N*-acetyl-D-[UL-$^{13}C_6$]-glucosamine were mixed in a 1.5 mL Eppendorf tube, followed by the addition of 270 μL of DMSO. About 0.7 g sodium hydroxide beads were collected in a Pierce spin column (900 μL volume) and washed once with 350 μL of acetonitrile (ACN) followed by two rinses with 350 μL of DMSO. The plasma sample was mixed in with 270 μL of DMSO and 105 μL of iodomethane followed by immediate mixing. The whole mixture was then added to the preconditioned NaOH beads in the plugged microfuge spin column. After occasional gentle stirring the sample solution in NaOH column for 11 min, the microfuge spin column was unplugged and spun for 30 s at 5000 rpm (1000*g* in a fixed-angle rotor). The collected sample solution was quickly transferred into 3.5 mL of 0.5 M NaCl solution in 0.2 M sodium phosphate buffer (pH 7) within a silanized 13 × 100 mm glass test tube. To maximize glycan recovery, the NaOH beads were then washed twice by 300 μL of ACN, with all spin-throughs immediately transferred into the same silanized glass test tube. To perform liquid/liquid (L/L) extraction, 1.2 mL of chloroform was added to each test tube, which was then capped and shaken well. After brief centrifugation to separate the layers, the aqueous layer (top) was discarded and then replaced by a fresh aliquot of 3.5 mL of 0.5 M NaCl solution in 0.2 M sodium phosphate buffer (pH 7). After three L/L extraction rounds, the chloroform layer was finally recovered and dried under a gentle stream of nitrogen in a heater block set to 74 °C.

Hydrolysis, Reduction, and Acetylation. To perform TFA hydrolysis, each sample was mixed with 2 M TFA (325 μL) and incubated at 121 °C for 2 h, which was then dried under a gentle stream of nitrogen in a heater block set to 74 °C. To reduce the sugar aldehydes, each sample was incubated at room temperature for 1 h after dissolution in 475 μL of freshly made 10 mg/mL sodium borohydride in 1 M ammonium hydroxide. To remove excess borate, 63 μL of methanol (MeOH) was added and dried under nitrogen, followed by adding 125 μL of 9:1 (v/v) MeOH:acetic acid. Samples were then dried under nitrogen and then fully dried in a vacuum desiccator for 20 min.

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

The last step is acetylation of nascent hydroxyl groups, in which 18 μL of deionized water was added to each test tube to dissolve any precipitates. After adding 250 μL of acetic anhydride and sonicating in a water bath for 2 min, each sample was incubated for 10 min at 60 °C, followed by mixing with 230 μL of concentrated TFA and incubated again at 60 °C for 10 min. To clean up the sample mixture, L/L extraction was performed twice after adding 1.8 mL of dichloromethane and 2 mL of deionized water to each test tube. With the aqueous layer (top layer) discarded for each round, the organic layer of each sample was then transferred to a silanized autosampler vial, dried under nitrogen and reconstituted in 120 μL of acetone, which was then capped in preparation for injection onto the GC−MS.

Gas Chromatography−Mass Spectrometry. An Agilent Model A7890 gas chromatograph (equipped with a CTC PAL autosampler) coupled to a Waters GCT (time-of-flight) mass spectrometer was employed to analyze the prepared samples. For each sample, 1 μL of the 120 μL total volume was injected onto a hot (280 °C), silanized glass liner (Agilent Cat. No. 5183−4647) containing a small plug of silanized glass wool at a split ratio of 20:1. A 30-m DB-5 ms GC column was used to separate different sample components, facilitated by the carrier gas (helium) with a 0.8 mL/min flow rate. The GC oven temperature was initially kept at 165 °C for 0.5 min, then increased to 265 °C at a rate of 10 °C/min, followed by immediate ramping to 325 °C at a rate of 30 °C/min, and finally held at 325 °C for 3 min. Sample components eluted from GC column were subjected to electron ionization (70 eV, 250 °C). Positive-ion mode mass spectra from individual TOF pulses over a m/z range of 40−800 were summed every 0.1s. Daily tuning and calibration of the mass spectrometer was performed with perfluorotributylamine to ensure reproducible relative abundances of EI ions and mass accuracy within 10 ppm.

Data Analysis

Data Processing. Quanlynx 4.1 software was employed to integrate the summed extracted-ion chromatogram (XIC) peak areas for all glycan nodes. The peak areas were automatically integrated and manually verified, then exported to a spreadsheet for further analysis.

Two possible normalization approaches were considered: (1) individual hexoses were normalized to heavy glucose, and individual *N*-acetylhexosamines (HexNAcs) were normalized to heavy *N*-acetyl glucosamine (GlcNAc); (2) individual hexoses were normalized to the sum of all endogenous hexoses, and individual HexNAcs were normalized to the sum of all endogenous HexNAcs. The second normalization approach tends to provide better interbatch reproducibility (<9% average CV for the six most elevated glycan nodes), but the first approach performs better in identifying the potential increases of all glycan nodes in the patient groups relative to the control group while maintaining a reasonable interbatch % CV (i.e., <21%). Thus, results reported below are based on normalization with heavy glucose and heavy GlcNAc, unless otherwise stated. The raw data of all XIC peak areas for all samples, together with the normalized data by the two normalization approaches and % CV values for batch-to-batch quality control (QC) samples are provided in a spreadsheet available as Supporting Information.

Statistical Analysis

For the glycan node data of each cohort, outliers were removed by log-transformation and the ROUT method at Q = 1% using GraphPad Prism 7. Outlier-removed data were then reverse transformed by taking the antilog of each value. To identify differences between cohorts, the Kruskal−Wallis test followed by the Benjamini−Hochberg false discovery correction procedure was performed at a 5% false discovery rate using GraphPad Prism 7. RStudio Version 1.0.143 was used to compare different receiver operating characteristic (ROC) curves by Delong's test or Bootstrap test. The ROC curves shown in figures were plotted by GraphPad Prism 7. Correlation of glycan nodes with age or smoking pack-years were assessed via Spearman's rank correlation in GraphPad Prism 7. Stage-by-stage multivariate modeling was performed using multivariate logistic regression in RStudio Version 1.0.143, with assessment carried out by leave-one-out-validation, and model selection done using a best subsets procedure. The ability of specific glycan nodes to predict lung cancer survival was evaluated with Cox proportional hazards regression model in SAS 9.4. And GraphPad Prism 7 was applied to generate survival curves and perform associated log-rank Mantel-Cox tests.

RESULTS

Study Highlights

- Striking increases in glycan nodes that serve as direct indicators of α2−6 sialylation, β1−4 branching, β1−6 branching, and antennary fucosylation in stage III−IV lung cancer. Similar increases also observed for 2-linked mannose and 4-linked *N*-acetylglucosamine, both of which are associated with total glycosylation levels, especially N-glycans (Figure 3).
- Significant increases in these glycan nodes in stage I−II lung cancer, with a general trend for increasing prevalence from stage I−IV (Figure 3).
- Minimal dependence of glycan nodes on smoking status, age, and histological type of lung cancer.
- The top quartiles of all six glycan nodes listed in bullet point 1 above predict all-cause mortality across all stages of lung cancer combined (Figure 6).
- Glycan nodes corresponding to α2−6 sialylation and β1−4 branching are particularly good at predicting survival in stage IV patients (Figure S6).

Glycan Node Stability in Plasma

Cancer patient enrollment for the WELCA study took place at 12 different sites. In some cases, samples were permitted to sit overnight at 4 °C prior to final processing and storage at −80 °C. In other cases, sample aliquots were temporarily stored at −20 °C prior to shipment a few weeks later to the central repository where they were kept long-term at −80 °C. As such, assessment of the stability of glycan nodes in EDTA plasma kept at room temperature, 4 °C, and −20 °C for varying lengths of time was assessed.

Five EDTA plasma samples from separate healthy donors (three male and two female), were aliquoted and temporarily kept at −20 °C for 10, 90, and 360 days, 4 °C for 2 or 90 days, room temperature for 1 day, or kept continuously at −80 °C. Samples kept temporarily at temperatures warmer than −80 °C were compared with their respective control aliquots kept continuously at −80 °C. The glycan nodes that are typically present at >1% relative abundance within their respective hexose or HexNAc class were measured and normalized to heavy, stable isotope-labeled glucose and GlcNAc internal standards or, alternatively, normalized to the sum of endogenous hexoses or HexNAcs. No significant differences were observed in the data sets normalized to the sum of endogenous hexoses/HexNAcs.

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

Table 1. Statistically Significant Differences between Cohorts within the WELCA Study[a]

Glycan Node[b]	Control vs Stage I	Control vs Stage II	Control vs Stage III	Control vs Stage IV	Stage I vs Stage II	Stage I vs Stage III	Stage I vs Stage IV	Stage II vs Stage III	Stage II vs Stage IV	Stage III vs Stage IV
t-Fucose	ns	ns	iiii	iiii	ns	ns	ns	ns	ns	ns
t-Gal	ii	ns	iiii	iiii	ns	ns	ns	ns	i	ns
2-Man	iii	i	iiii	iiii	ns	ns	ns	ns	i	ns
4-Glc	iii	i	iiii	iiii	ns	ns	ns	ns	ns	ns
3-Gal	ns	ns	ns	iiii	ns	ns	ns	ns	iii	ii
6-Gal	ii	ns	iiii	iiii	ns	ns	i	ns	ii	i
3,4-Gal	ii	ii	iiii	iiii	ns	ns	ns	ns	ns	ns
2,4-Man	i	i	iiii	iiii	ns	ns	iii	ns	ii	ii
2,6-Man	iii	ii	iiii	iiii	ns	ns	ns	ns	ns	ns
3,6-Man	ns	ns	ii	iiii	ns	ns	ns	i	iii	i
3,6-Gal	ns	ns	i	iiii	ns	ns	ns	ns	iii	ii
3,4,6-Man	ii	ns	ii	ii	i	ns	ns	ns	ns	ns
t-GlcNAc	ns	ns	iiii	ii	ns	ns	ns	ii	ns	ns
4-GlcNAc	ii	ns	iiii	iiii	ns	ns	i	i	ii	ns
3-GlcNAc	ns	ns	iiii	iiii	ns	ns	ns	ns	ns	ns
3-GalNAc	ns	ns	ii	iiii	ns	ns	ns	ns	ns	ns
3,4-GlcNAc	ns	ii	iiii	iiii	ns	i	ii	ns	ns	ns
4,6-GlcNAc	iii	ns	iiii	iiii	ii	ns	ns	ii	ii	ns
3,6-GalNAc	ns	ns	ii	iiii	ns	ns	ns	ns	ns	ns

[a]Heavy, stable isotope labeled glucose (Glc) and GlcNAc were utilized to normalize Hexose and HexNAc data, correspondingly. [b]Kruskal–Wallis test followed by Benjamini–Hochberg false discovery correction procedure at 95% confidence level is given. "ns" stands for "not significant". "i" indicates $p < 0.05$. "ii" indicates $p < 0.01$. "iii" indicates $p < 0.001$, and "iiii" indicates $p < 0.0001$.

When normalized to heavy, stable isotope-labeled glucose and GlcNAc internal standards, the only significant difference observed was an increase in 6-linked galactose for samples stored at room temperature for 1 day ($p = 0.033$; Table S3). Thus, under the mildly adverse conditions to which some of the specimens in this study may have been exposed (less than a day at 4 °C or up to a few weeks at −20 °C), glycan nodes were found to be stable.

Notably, a study of the impact of plasma vs serum matrices on glycan nodes was previously reported in this journal.[19] Differences observed were modest and did not impact the biological results of either the previous study or this one.

Altered Glycan Features in Stage I–IV Patients

Basic clinical characteristics and *n*-values of the WELCA sample set were described in the Materials and Methods section and Table S1. All 207 control and 208 stage I–IV patient samples were randomized and analyzed in 27 batches. Within each control and case sample, a total of 19 glycan "nodes" were measured. The relative abundances of each of these nodes contributed at least 1% of the total hexose or total *N*-acetylhexosamine (HexNAc) signal. Data from each of the 19 glycan nodes were normalized to heavy, isotope-labeled glucose and GlcNAc internal standards. Statistically significant differences were detected in each cancer stage relative to the control cohort: 10, 6, 18, and 19 out of 19 glycan nodes were increased in stage I, II, III, and IV, respectively (Table 1). Data for each glycan node normalized to the sum of endogenous hexoses or HexNAcs were analyzed analogously (Table S4). This revealed shifts in glycan compositions in stage I–IV patients vs controls. However, because quantitative changes in glycans tended to outpace glycan compositional changes (as we have previously observed[19]) this normalization procedure was not as sensitive in distinguishing age-matched controls from lung cancer patients at each stage.

Six glycan nodes were found to be significantly elevated at nearly every stage in lung cancer patients relative to the age-matched controls, and these included: 2-linked mannose (2-Man) and 4-linked *N*-acetylglucosamine (4-GlcNAc), both of which are associated with total glycosylation levels especially for N-glycans;[14] 6-linked galactose, corresponding to α2–6 sialylation;[16] 2,4-linked mannose, corresponding to β1–4 branching;[16] 2,6-linked mannose, corresponding to β1–6 branching;[16] and 3,4-linked GlcNAc, which primarily corre-

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985–3998

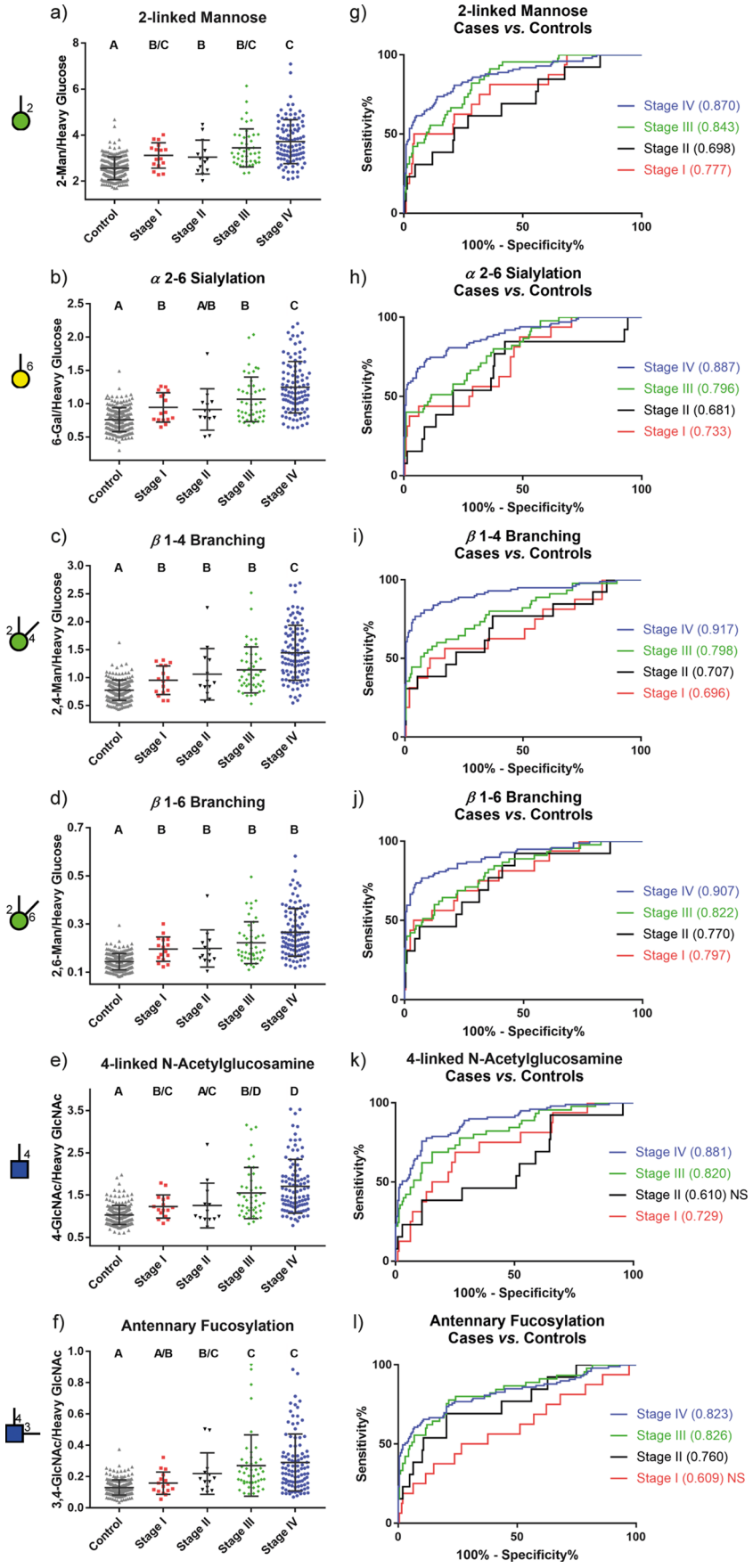


Figure 3. Univariate distribution (a–f) and ROC curves (g–l) for the six top performing glycan nodes in the WELCA study. The Kruskal–Wallis test was performed followed by the Benjamini–Hochberg false discovery correction procedure. Different letters at the top of data points in panels a–f demonstrate statistically significant differences between groups; any overlap between groups in any of the letter(s) assigned to the groups indicates a

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985–3998

Figure 3. continued

lack of significant differences between groups. ROC curves for stage I−IV lung cancer cases vs controls are provided in panels g−l. Areas under the ROC curves are provided in parentheses next to the specified stages. "NS" next to the AUC values indicates that the ROC curve is not statistically significant. All raw glycan node data in this figure were normalized to heavy glucose or heavy GlcNAc prior to data analysis.

sponds to antennary fucosylation[16] (Figure 3). The latter four nodes were among the top five most elevated nodes in our previously reported lung cancer study.[19] The receiver operating characteristic (ROC) curve c-statistics (areas under the curve, AUCs) for these six glycan nodes in stage I−IV patients vs controls ranged (with two exceptions) from 0.68 to 0.92 (Figure 3).

For most of these six glycan nodes there were significant differences between stages (Figure 3), but the most robust differences tended to be between stage IV and stage I−II patients. 2,4-Man, the glycan node indicative of β1−4 branching, was the best at differentiating stage IV vs all other stages of lung cancer. ROC curves showing the ability of β1−4 branching to distinguish between stage IV and all other stages are provided in Figure S1.

Prominent Early-Stage Alteration

Relative to the age-matched controls, five of the six top performing glycan node markers in stage I patients, and four in stage II patients, were significantly increased (Figure 3a−f). In addition, the ROC c-statistics (AUC) of these glycan nodes were mostly statistically significant and ranged from 0.68 to 0.80 (with one exception). The notable alterations of glycan nodes in early stages were not previously observed for other lung cancer sets, such as the dual gender lung cancer set and stage I-only lung cancer set (which was also dual gender) reported in our previous work[19] (*n*-values for these studies are provided in Table S2). Stage-specific ROC curves from the WELCA study and these other two studies were statistically compared (Table S5) and are shown side-by-side in Figure 4. Significant differences were observed for β1−6 branching when the ROC curve of the stage I cohort of the WELCA set was compared to that of the stage I-only lung cancer set and that of the stage I cohort of the dual gender lung cancer set. When comparing ROC curves for the stage IV cohorts of the WELCA set and the dual gender lung cancer set, significant differences were found for three glycan features including α2−6 sialylation, β1−4 and β1−6 branching. Since all the lung cancer patients and age-matched controls involved in the WELCA set were female, the gender dependence of these glycan node markers in early stages was evaluated in the other two lung cancer sets, which included patients and controls from both sexes. When sample set and stage were held constant, the ROC curves of the two sexes were compared for each glycan node using Delong's test or the Bootstrap test (Table S6). No significant differences were observed, however, indicating the early stage clinical performance characteristics of the six glycan node markers were independent of gender.

Negligible Dependence on Smoking-Status, Age, and Histological Type

No significant alteration of five out of the six top performing glycan node markers was observed when each individual glycan node was separately analyzed for differences among never-smokers, previous smokers and current smokers within the WELCA study control cohort. The only exception was 3,4-linked GlcNAc (corresponding to antennary fucosylation), which was slightly elevated in current smokers relative to previous smokers (Figure S2). Spearman's rank correlation analysis demonstrated no statistically significant correlation with smoking pack-years in the control cohort, both for all control patients and control patients with smoking history (smoking pack-year >0). Together, these data revealed that the top performing glycan node markers within the control cohort had negligible dependence on smoking status. (A parallel analysis within the cancer patient cohort was not conducted due to the confounding correlation between smoking and lung cancer.)

The average ages of the control and case cohorts were nearly identical (61.2 and 61.6, respectively; Table S1). After pooling all data from the cases (all stages) and controls, 3,4-linked GlcNAc, corresponding to antennary fucosylation, was found to be weakly correlated with age (correlation coefficient $r = 0.159$, $p = 0.0016$; Figure 5a). When the control and case cohorts were analyzed separately, a significant correlation with age for 3,4-linked GlcNAc was only observed in the control cohort (correlation coefficient $r = 0.205$, $p = 0.0031$). No statistically significant correlations with age were observed for the other five top performing glycan nodes (Table S7). When the population was divided into smaller age groups, only 3,4-linked GlcNAc showed significant differences between pairs of decades; if the control and case cohorts were investigated in isolation, 3,4-linked GlcNAc within the controls in particular indicated a distinct upward pattern in more advanced age groups (Figure 5b). The same phenomenon was observed in the male-only controls of the dual gender lung cancer set. However, with the exception of 3,4-GlcNAc, these analyses indicated a lack of dependence of glycan node markers on age.

The effect of lung cancer histological subtypes on the six glycan nodes was evaluated in the stage IV non-small-cell lung cancer (NSCLC) subcohort (i.e., the largest single-stage subcohort available; Table S1). For each glycan node marker, ROC curves of the three histological subtypes of NSCLC—adenocarcinoma, squamous cell carcinoma, and large cell carcinoma—were compared pairwise by Delong's test or Bootstrap test (Figure S3, Table S8). No statistically significant differences between histological subtypes of NSCLC were discovered for any glycan node marker.

These findings on glycan node independence from smoking status, age and histological type are consistent with our previously reported findings from other lung cancer case/control studies.[19]

Role of Gender in Defining Plasma Glycans

The WELCA studied consisted entirely of women. Thus, to evaluate the role of gender in plasma glycan nodes, we turned to control patient data from the "large lung cancer" cohort of our previous study.[19] This set of cancer-free patients consisted of plasma samples from 123 males and 76 females. Since we had not previously done so, we looked for gender differences in all 19 glycan nodes evaluated in the WELCA study and found significant decreases in 3,4-linked GlcNAc (the node that corresponds to antennary fucosylation) as well as total fucose in females relative to males—regardless of whether data were normalized to heavy Glc/GlcNAc or to the sum of endogenous hexoses/HexNAcs ($p < 0.05$ or lower after applying the Benjamini−Hochberg false discovery correction procedure).

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

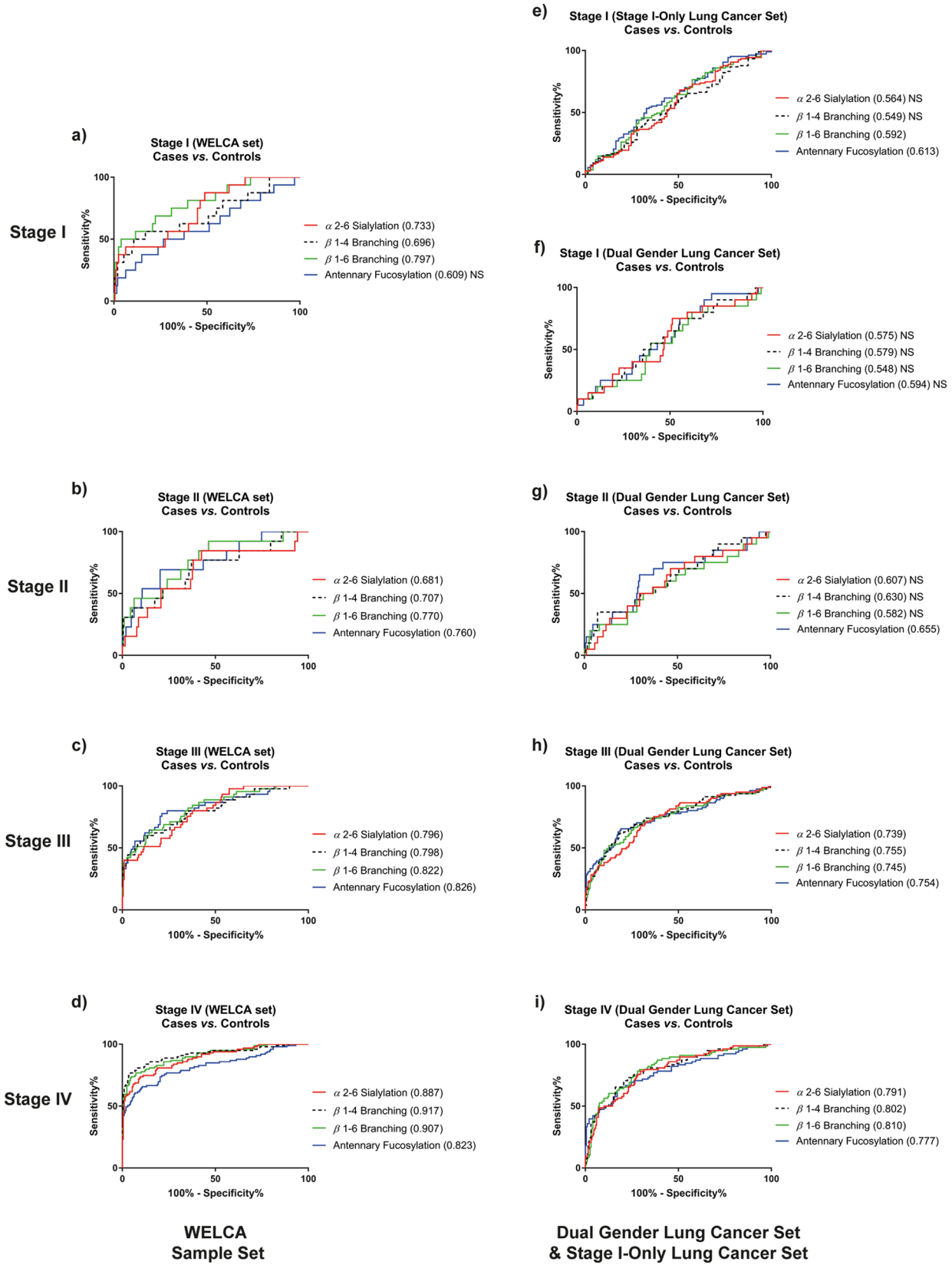


Figure 4. ROC curves for four top-performing glycan nodes in stage I–IV within different lung cancer case-control sets. Four glycan nodes with highly ranked performance in all three sample sets are shown. The ROC curves from WELCA sample set are illustrated in panels a–d. In panels e–i are ROC curves from the other two lung cancer sets, aligned with the WELCA data by stage: Dual Gender Lung Cancer set (f–i) and Stage I Only Lung Cancer set (also dual gender; e). *n*-values of each group are provided in Table S2. Results from stage-specific statistical comparisons between studies are provided in Table S5 and summarized in the main text.

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985–3998

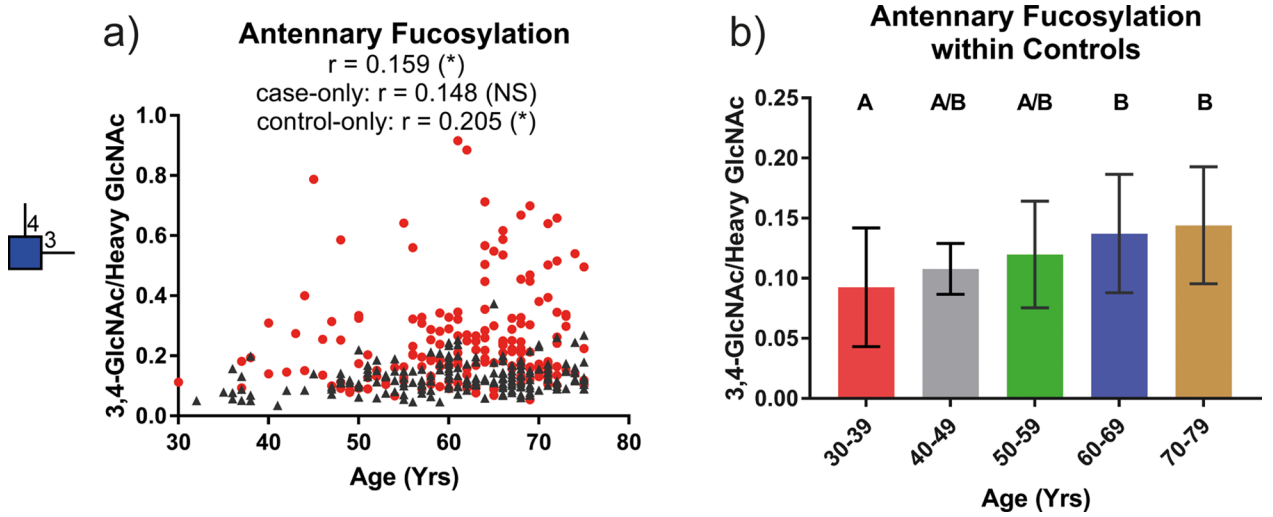


Figure 5. Correlation between age and a top performing glycan node, 3,4-linked GlcNAc, in the WELCA study. (a) Spearman's rank correlation was performed with cases and controls together, with coefficients provided above the data points. "*" next to the coefficient indicates that the Spearman's rank correlation was statistically significant with $p < 0.0083$ (Bonferroni-corrected cutoff). Possible correlations between age and 3,4-linked GlcNAc for the case cohort and the control cohort, separately, were also evaluated, with the corresponding coefficients provided. "NS" indicates no significant correlation. Controls are indicated by black triangles and cases by red dots. (b) The Kruskal–Wallis test was performed followed by the Benjamini–Hochberg false discovery correction procedure to identify differences between age groups in the control cohort. Different letters at the top of bars demonstrate statistically significant differences between groups; any overlap between groups in any of the letter(s) assigned to the groups indicates a lack of significant differences between groups. For the other five top performing glycan nodes not shown in this figure, no statistically significant associations with age were found.

These observed increases in antennary fucosylation agree with previously published findings on studies of women of approximately the same age.[28,29] Moreover, in the WELCA study we found that the observed increase in antennary fucosylation with age (Figure 5) agreed with that previously observed by Reiding et al.[29]

Notably, we also observed increases in 2,4-linked mannose, corresponding to β1−4 branching, in women compared to men ($p < 0.05$ for both heavy Glc/GlcNAc and endogenous normalizations). These findings align with those from Kneževi ć et al.[28] and Reiding et al.[29] in which they found modest increases in triantennary and tetrantennary glycans in women relative to men—though for this glycan feature only the study of Kneževi ć et al. revealed a statistically significant difference.[28]

Total Glycosylation and Multivariate Model of Glycan Features

The clinical performance characteristics of total glycosylation (i.e., total hexoses, total HexNAcs, and the sum of total hexoses and total HexNAcs) were evaluated and compared to individual glycan node markers on a stage-by-stage basis (Table S9). Results of ROC curve comparisons by paired Delong's tests demonstrated that total glycosylation cannot distinguish stage I–IV cases from controls better than individual glycan node markers.

Additionally, multivariate logistic regression models were built and compared with the clinical performance characteristics of individual glycan nodes at each stage (Figure S4). Fully cross-validated multivariate logistic regression models were no better at detecting lung cancer than the top-performing individual glycan node at each respective stage. Again, these results were consistent with our previous observations in lung cancer.[19]

Prediction of All-Cause Mortality

To evaluate the ability of the six glycan nodes to predict all-cause mortality, glycan node data were broken into quartiles and analyzed by Cox proportional hazards regression, with adjustment for age, smoking status, and cancer stage (Table S10). First and foremost, for patients in all four stages, the top quartiles of all six glycan node markers predicted all-cause mortality with hazard ratios in the range of 2−3 and $p < 0.01$, relative to all other quartiles combined. The different rates of death for the top quartile versus all other quartiles for each glycan node marker are illustrated by survival curves (Figure 6).

When focusing on stage III and IV patients, the top quartiles of all six glycan node markers predicted all-cause mortality with hazard ratios in the range of 2−3 and $p < 0.05$ (Table S10) relative to all other quartiles combined (survival curves shown in Figure S5). Similar results were observed for stage IV patients only (Table S10). However, when stage III patients were analyzed alone, the hazard ratios of all six glycan nodes were not significantly different from 1 ($p > 0.05$), indicating the relative risk of death was not detectably different between patients in the top quartile vs all other quartiles of each glycan node. 6-linked galactose (corresponding to α2−6 sialylation) and 2,4-linked mannose (corresponding to β1−4 branching) were significantly different between stages III and IV (Figure 3b,c). Stage-specific survival curves for these glycan nodes are provided in Figure S6.

Overall, these results for glycan node-based prediction of mortality vary slightly, but are largely consistent with our previously reported results on the ability of α2−6 sialylation and branched mannose residues to predict all-cause mortality in lung cancer.

DISCUSSION

Consistency of Specific Glycan Feature Changes in Lung Cancer

Six out of 19 quantified glycan nodes, corresponding to total glycosylation levels (especially for N-glycans), α2−6 sialylation, β1−4 branching, β1−6 branching, and antennary fucosylation, were significantly elevated in the WELCA lung cancer patients relative to age-matched controls. These findings in the WELCA

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

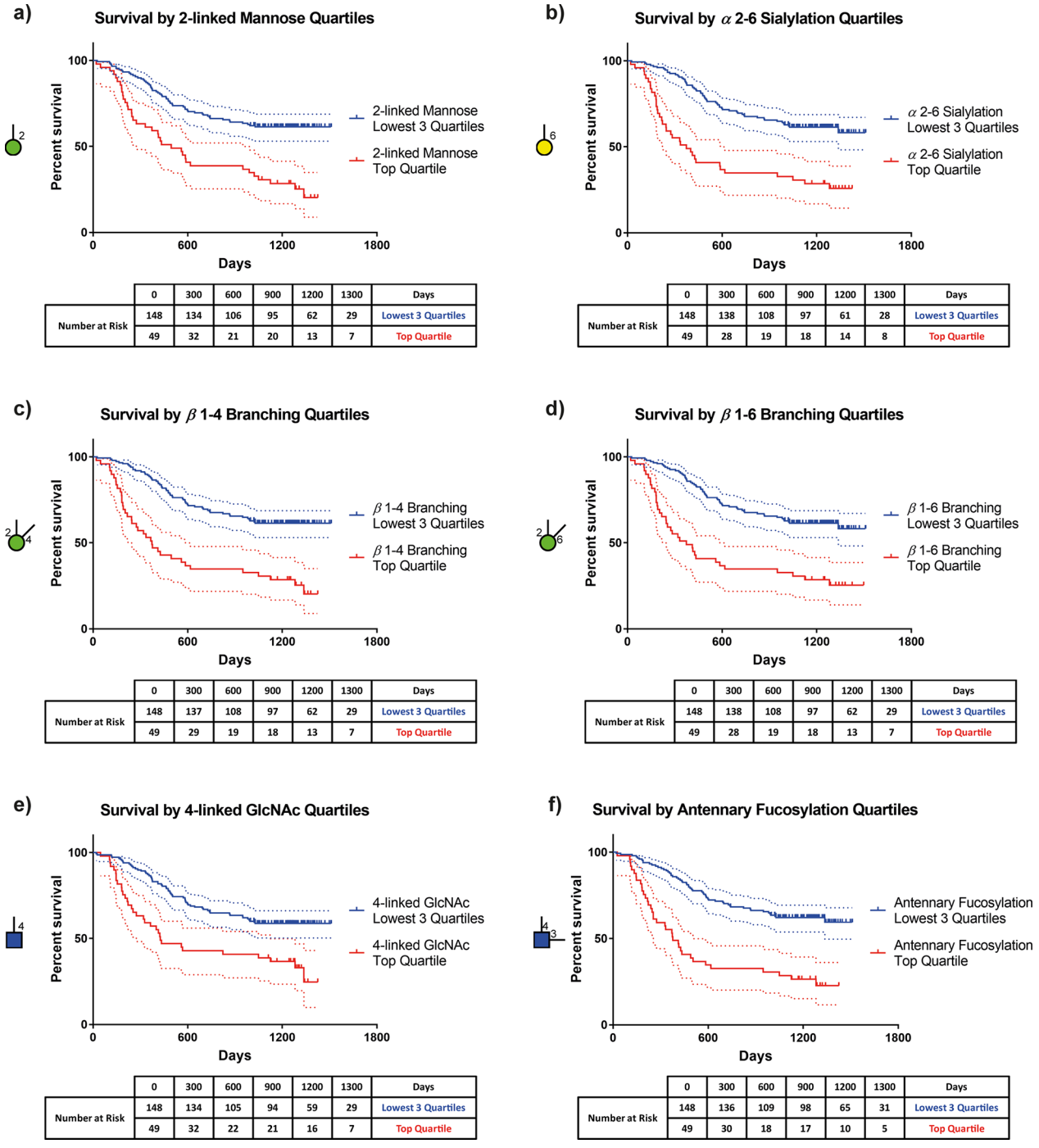


Figure 6. Survival curves for the six top performing glycan nodes for all stages combined. In each panel, the top quartile of specified glycan node is compared to all other quartiles combined. According to results of log-rank Mantel−Cox test, the survival curves within each panel are significantly different ($p < 0.0001$). Dotted lines represent 95% confident intervals. The median duration of follow-up for deceased patients (until death) was 406 days; for those that remained alive, it was 1253 days. The median follow-up time for all patients was 1057 days.

set are highly consistent with our previously reported lung cancer study on a dual gender lung cancer set,[19] which also demonstrated the distinct increase of the latter four glycan features within stage III−IV cases compared to their respective control cohorts.

Our observations of the glycan node-based feature changes in lung cancer patients are closely aligned with the intact glycan changes reported in lung cancer by Vasseur and colleagues.[30] Their intact glycan analysis results primarily revealed significant increases in antennary fucosylation, as well as fucosylated tri- and tetra-antennary N-glycans—findings that are in line with increases observed here in β1−4 branching and β1−6 branching.

Consistency in Prediction of Survival

The six top performing glycan nodes-based features in this study were not only able to distinguish lung cancer patients from age-matched controls, but were also able to predict all-cause mortality in the WELCA set—a finding that agrees well with the survival-predicting nodes in our previously reported study on the dual gender lung cancer set.[19] Similar discoveries regarding

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

the prognostic capacity of P/S glycans have also been reported by other groups. Hashimoto and colleagues[31] suggested that specific glycoforms of serum α_1-acid glycoprotein (AGP) seemed to predict progression and mortality of several carcinomas, including lung cancer. According to their follow-up studies, patients who had the AGP glycoforms that contained highly fucosylated and branched sugar chains tended to have a poor prognosis. Besides the glycan features discussed above, another good prognostic predictor of lung cancer is the sialyl Lewis X epitope (SLe^x),[32] which consists of α2−3 sialylation instead of α2−6 sialylation. The progression and survival in non-small-cell[33−35] and small-cell lung cancer[36] can both be predicted by SLe^x.

Most clinical trials require that enrolled patient life expectancy exceed three months such that a benefit from treatment can be observed—yet formal guidelines are generally not provided to facilitate this prediction.[37] Glycan nodes representing α2−6 sialylation and β1−4 branching both performed well as prognostic indicators of survival within stage IV patients (Figure S6), and as such they may be able to provide some clinical utility toward this end. A prospective study would be required to validate them for this purpose.

Early Stage Changes in Glycan Nodes

Unlike the other two lung cancer sets that we have reported on previously,[19] some glycan node-based features were substantially altered in the WELCA lung cancer patients at stages I−II (Figure 4). Even though a relatively low number of early stage samples were measured (n = 16 and 13 for stage I and II, respectively), statistically significant elevations were detected in most of the six glycan node markers, alongside comparatively high ROC c-statistics. Outside of a statistical anomaly, there are two possible noncancer related causes for this phenomenon. First, since the lung cancer patients and controls enrolled in the WELCA study are all female, a distinct gender dependence of glycan features may exist, especially in early stages. However, this possibility was not evidenced by the observation that no significant difference was detected between men and women in stage I and II of the dual gender lung cancer set, as well as in the stage I-only lung cancer set, which was also dual gender (Table S6). The second possible explanation is that the nonsmoking-matched controls of the WELCA set may have lower relative abundances of all the glycan nodes of interest relative to the smoking-matched controls for other lung cancer sets. In the WELCA set most controls were never-smokers, but the cancer patients were mainly current-smokers (Table S1), suggesting that smoking history might possibly contribute to increases in some glycan nodes. Taken together with the observation that the top performing glycan node markers within the control cohort had near-negligible dependence on smoking status (Figure S2), smoking appears to contribute to slight, but mostly statistically insignificant elevation of glycan nodes. Smoking is undoubtedly bad for the liver,[38] which secretes approximately half of all circulating glycoproteins.[39,40] Nevertheless, these results that indicate only a mild contribution of smoking to alterations in circulating glycan nodes is in full agreement with results from our previous study of glycan nodes in lung cancer patients in which controls were smoking status-matched to the lung cancer patients, and in which only minor impacts of smoking on glycan nodes within the control population were observed.[19]

Role of Gender

Many studies have reported important gender differences in lung cancer between men and women, in terms of histological type, tobacco exposure, and survival and treatment response.[41,42] Here, by comparison with our previously conducted studies,[19] no obvious gender differences were detected with regard to P/S glycan features. Smoking is the primary risk factor for lung cancer. However, a large percentage of women with lung adenocarcinoma—between 20% and 30% in Western countries and nearly 80% in Asian countries—are nonsmokers.[26] Hence, some female-specific risk factors for lung cancer must exist and may play vital roles in lung cancer development, progression and survival; these may include hormonal factors and occupational risk factors in female occupations—as suggested by Stücker et al.[26] Therefore, studies focused on lung cancer in women, especially on the gender specific risk factors, should garner further attention as they promise to disentangle the etiology of lung cancer in women.

CONCLUSIONS

As represented by glycan nodes, blood plasma glycans were found to be stable under a variety of less-than-ideal sample storage conditions. The diagnostic and prognostic capacity of plasma glycan features in stage I−IV lung cancer—as represented by monosaccharide and linkage-specific glycan nodes—were validated in the WELCA case-control study. Significant elevation of α2−6 sialylation, β1−4 branching, β1−6 branching, antennary fucosylation, and total N-glycosylation level was observed in almost every stage of lung cancer relative to age-matched control groups. Early stage detection was stronger than we have previously observed,[19] but this observation may have been related to the lack of smoking status-matching between cases and controls in the WELCA study. Nevertheless, alteration of glycan features in lung cancer was found to be almost completely independent of smoking status, age, and histological subtypes of lung cancer. The six most-elevated glycan features predicted all-cause mortality in lung cancer patients after adjusting for age, smoking status, and cancer stage. No gender-based differences were discovered in glycan features associated with lung cancer.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00457.

Table S1: Basic Clinical Characteristics and n-values of the WELCA Sample Set; Table S2: Stage and Gender Composition of Three Lung Cancer Sample Sets and Their Subcohorts; Table S3: Comparison of Glycan Node Stability at Different Conditions Relative to Control Aliquots Stored at −80 °C; Table S4: Statistically Significant Differences between Cohorts within the WELCA Study; Table S5: Stage-by-Stage ROC Comparison of the Top Performing Glycan Nodes; Table S6: Comparison of Top Performing Glycan Nodes in Male vs Female Patients with Early Stage Lung Cancer; Table S7: Correlation Between Age and the Top Performing Glycan Nodes in the WELCA Cases (all stages) and, Separately, Controls; Table S8: Comparison of the Top Performing Glycan Nodes in Different Histological Types; Table S9: Stage-by-Stage Comparison of Total Glycosylation with Individual Glycan Feature; Table S10: Survival Prediction by the Top Performing Glycan Nodes in All Stages, Stage III and IV Combined, Stage III Only, and Stage IV Only;

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

Figure S1: ROC curves for β1−4 branching for stage IV vs each other stage of non-small-cell lung cancer (NSCLC); Figure S2: Connection between antennary fucosylation and smoking status within the WELCA control group; Figure S3: ROC curves for the six top performing glycan nodes within different histological subtypes of non-small-cell lung cancer (NSCLC); Figure S4: Multivariate logistic regression models for stage I−IV patients from the WELCA data set; Figure S5: Survival curves of the six top performing glycan nodes for stage III and IV combined; Figure S6: Survival curves for the two top performing glycan nodes that were significantly different between stages III and IV: Stage III patients alone and stage IV patients alone (PDF)

Raw and normalized chromatographic peak areas for all WELCA samples analyzed in this study (XLSX)

AUTHOR INFORMATION

Corresponding Author

*Tel: 480 727 9928. E-mail: chad.borges@asu.edu.

ORCID

Chad R. Borges: 0000-0002-8122-3438

Present Address

‖Seer INC, South San Francisco, California 94080, United States.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are grateful to Drs. Harvey Pass of NYU and Xifeng Wu of the University of Texas MD Anderson Cancer Center for providing samples from which previously collected data[19] were employed for comparative purposes. The research reported here was supported in part by the National Cancer Institute of the National Institutes of Health under award no. R33 CA191110 (to C.B.). The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institutes of Health. The WELCA study was supported by the French "Institut National du Cancer" (grant# 2013-132), the "Fondation de France" (grant #2015-60747) and the "Ligue Nationale Contre le Cancer" (grant # PRE2015.LNCC).

ABBREVIATIONS

WELCA, women epidemiology lung cancer; ROC, receiver operating characteristic; AUC, area under the curve; P/S, plasma or serum; GTs, glycosyltransferases; HexNAcs, *N*-acetylhexosamines; GlcNAc, *N*-acetyl glucosamine; QC, quality control; NSCLC, non-small-cell lung cancer (NSCLC).

REFERENCES

(1) Siegel, R. L.; Miller, K. D.; Jemal, A. Cancer statistics, 2019. *Ca-Cancer J. Clin.* **2019**, *69*, 7−34.

(2) Provencio, M.; Isla, D.; Sánchez, A.; Cantos, B. Inoperable stage III non-small cell lung cancer: Current treatment and role of vinorelbine. *J. Thorac. Dis.* **2011**, *3*, 197−204.

(3) National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **2011**, *365*, 395−409.

(4) Okamura, K.; Takayama, K.; Izumi, M.; Harada, T.; Furuyama, K.; Nakanishi, Y. Diagnostic value of CEA and CYFRA 21−1 tumor markers in primary lung cancer. *Lung cancer* **2013**, *80*, 45−49.

(5) Xu, Y.; Xu, L.; Qiu, M.; Wang, J.; Zhou, Q.; Xu, L.; Wang, J.; Yin, R. Prognostic value of serum cytokeratin 19 fragments (Cyfra 21−1) in patients with non-small cell lung cancer. *Sci. Rep.* **2015**, *5*, 9444.

(6) Arrieta, O.; Villarreal-Garza, C.; Martínez-Barrera, L.; Morales, M.; Dorantes-Gallareta, Y.; Peña-Curiel, O.; Contreras-Reyes, S.; Macedo-Pérez, E. O.; Alatorre-Alexander, J. Usefulness of serum carcinoembryonic antigen (CEA) in evaluating response to chemotherapy in patients with advanced non small-cell lung cancer: a prospective cohort study. *BMC Cancer* **2013**, *13*, 254.

(7) Wang, X.-B.; Li, J.; Han, Y. Prognostic significance of preoperative serum carcinoembryonic antigen in non-small cell lung cancer: a meta-analysis. *Tumor Biol.* **2014**, *35*, 10105−10110.

(8) Bianchi, F.; Nicassio, F.; Marzi, M.; Belloni, E.; Dall'Olio, V.; Bernard, L.; Pelosi, G.; Maisonneuve, P.; Veronesi, G.; Di Fiore, P. P. A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer. *EMBO Mol. Med.* **2011**, *3*, 495−503.

(9) Zheng, D.; Haddadin, S.; Wang, Y.; Gu, L.-Q.; Perry, M. C.; Freter, C. E.; Wang, M. X. Plasma microRNAs as novel biomarkers for early detection of lung cancer. *Int. J. Clin. Exp. Pathol.* **2011**, *4*, 575−586.

(10) Shen, J.; Todd, N. W.; Zhang, H.; Yu, L.; Lingxiao, X.; Mei, Y.; Guarnera, M.; Liao, J.; Chou, A.; Lu, C. L.; et al. Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab. Invest.* **2011**, *91*, 579−587.

(11) Belinsky, S. A.; Klinge, D. M.; Dekker, J. D.; Smith, M. W.; Bocklage, T. J.; Gilliland, F. D.; Crowell, R. E.; Karp, D. D.; Stidley, C. A.; Picchi, M. A. Gene promoter methylation in plasma and sputum increases with lung cancer risk. *Clin. Cancer Res.* **2005**, *11*, 6505−6511.

(12) Balgkouranidou, I.; Chimonidou, M.; Milaki, G.; Tsarouxa, E.; Kakolyris, S.; Welch, D.; Georgoulias, V.; Lianidou, E. Breast cancer metastasis suppressor-1 promoter methylation in cell-free DNA provides prognostic information in non-small cell lung cancer. *Br. J. Cancer* **2014**, *110*, 2054−2062.

(13) Hou, J.-M.; Krebs, M.; Ward, T.; Sloane, R.; Priest, L.; Hughes, A.; Clack, G.; Ranson, M.; Blackhall, F.; Dive, C. Circulating tumor cells as a window on metastasis biology in lung cancer. *Am. J. Pathol.* **2011**, *178*, 989−996.

(14) Varki, A.; Kannagi, R.; Toole, B. P. Glycosylation Changes in Cancer. In *Essentials of Glycobiology*, 2nd ed.; Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2009; pp 617−632.

(15) Ruhaak, L. R.; Miyamoto, S.; Lebrilla, C. B. Developments in the identification of glycan biomarkers for the detection of cancer. *Mol. Cell. Proteomics* **2013**, *12*, 846−855.

(16) Borges, C. R.; Rehder, D. S.; Boffetta, P. Multiplexed surrogate analysis of glycotransferase activity in whole biospecimens. *Anal. Chem.* **2013**, *85*, 2927−2936.

(17) Zaare, S.; Aguilar, J. S.; Hu, Y. M.; Ferdosi, S.; Borges, C. R. Glycan Node Analysis: A Bottom-up Approach to Glycomics. *J. Visualized Exp.* **2016**, *111*, No. e53961.

(18) Hu, Y.; Borges, C. R. A spin column-free approach to sodium hydroxide-based glycan permethylation. *Analyst* **2017**, *142*, 2748−2759.

(19) Ferdosi, S.; Rehder, D. S.; Maranian, P.; Castle, E. P.; Ho, T. H.; Pass, H. I.; Cramer, D. W.; Anderson, K. S.; Fu, L.; Cole, D. E. C.; Le, T.; Wu, X.; Borges, C. R. Stage Dependence, Cell-Origin Independence, and Prognostic Capacity of Serum Glycan Fucosylation, β1−4 Branching, β1−6 Branching, and α2−6 Sialylation in Cancer. *J. Proteome Res.* **2018**, *17*, 543−558.

(20) Ferdosi, S.; Ho, T. H.; Castle, E. P.; Stanton, M. L.; Borges, C. R. Behavior of blood plasma glycan features in bladder cancer. *PLoS One* **2018**, *13*, No. e0201208.

(21) Gasperino, J. Gender is a risk factor for lung cancer. *Med. Hypotheses* **2011**, *76*, 328−331.

(22) Osann, K. E.; Anton-Culver, H.; Kurosaki, T.; Taylor, T. Sex differences in lung-cancer risk associated with cigarette smoking. *Int. J. Cancer* **1993**, *54*, 44−48.

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

(23) Risch, H. A.; Howe, G. R.; Jain, M.; Burch, J. D.; Holowaty, E. J.; Miller, A. B. Are female smokers at higher risk for lung cancer than male smokers? A case-control analysis by histologic type. *Am. J. Epidemiol.* **1993**, *138*, 281−293.
(24) Pope, M.; Ashley, M.; Ferrence, R. The carcinogenic and toxic effects of tobacco smoke: are women particularly susceptible? *J. Gender-Specific Med.* **1999**, *2*, 45−51.
(25) Wakelee, H. A.; Chang, E. T.; Gomez, S. L.; Keegan, T. H.; Feskanich, D.; Clarke, C. A.; Holmberg, L.; Yong, L. C.; Kolonel, L. N.; Gould, M. K.; et al. Lung cancer incidence in never-smokers. *J. Clin. Oncol.* **2007**, *25*, 472−478.
(26) Stücker, I.; Martin, D.; Neri, M.; Laurent-Puig, P.; Blons, H.; Antoine, M.; Guiochon-Mantel, A.; Brailly-Tabard, S.; Canonico, M.; Wislez, M.; et al. Women Epidemiology Lung Cancer (WELCA) study: reproductive, hormonal, occupational risk factors and biobank. *BMC Public Health* **2017**, *17*, 324.
(27) Wang, B.-Y.; Huang, J.-Y.; Cheng, C.-Y.; Lin, C.-H.; Ko, J.-L.; Liaw, Y.-P. Lung cancer and prognosis in Taiwan: a population-based cancer registry. *J. Thorac. Oncol.* **2013**, *8*, 1128−1135.
(28) Knežević, A.; Gornik, O.; Polašek, O.; Pučić, M.; Redžić, I.; Novokmet, M.; Rudd, P. M.; Wright, A. F.; Campbell, H.; Rudan, I.; Lauc, G. Effects of aging, body mass index, plasma lipid profiles, and smoking on human plasma N-glycans. *Glycobiology* **2010**, *20*, 959−969.
(29) Reiding, K. R.; Ruhaak, L. R.; Uh, H.-W.; El Bouhaddani, S.; van den Akker, E. B.; Plomp, R.; McDonnell, L. A.; Houwing-Duistermaat, J. J.; Slagboom, P. E.; Beekman, M.; Wuhrer, M. Human plasma N-glycosylation as analyzed by matrix-assisted laser desorption/ionization-Fourier transform ion cyclotron resonance-MS associates with markers of inflammation and metabolic health. *Mol. Cell. Proteomics* **2017**, *16*, 228−242.
(30) Vasseur, J. A.; Goetz, J. A.; Alley, W. R., Jr; Novotny, M. V. Smoking and lung cancer-induced changes in N-glycosylation of blood serum proteins. *Glycobiology* **2012**, *22*, 1684−1708.
(31) Hashimoto, S.; Asao, T.; Takahashi, J.; Yagihashi, Y.; Nishimura, T.; Saniabadi, A. R.; Poland, D. C.; van Dijk, W.; Kuwano, H.; Kochibe, N.; et al. α1-Acid glycoprotein fucosylation as a marker of carcinoma progression and prognosis. *Cancer* **2004**, *101*, 2825−2836.
(32) Arnold, J. N.; Saldova, R.; Hamid, U. M. A.; Rudd, P. M. Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation. *Proteomics* **2008**, *8*, 3284−3293.
(33) Mizuguchi, S.; Inoue, K.; Iwata, T.; Nishida, T.; Izumi, N.; Tsukioka, T.; Nishiyama, N.; Uenishi, T.; Suehiro, S. High serum concentrations of Sialyl Lewisx predict multilevel N2 disease in non−small-cell lung cancer. *Annals of surgical oncology* **2006**, *13*, 1010−1018.
(34) Mizuguchi, S.; Nishiyama, N.; Iwata, T.; Nishida, T.; Izumi, N.; Tsukioka, T.; Inoue, K.; Kameyama, M.; Suehiro, S. Clinical value of serum cytokeratin 19 fragment and sialyl-Lewis x in non-small cell lung cancer. *Annals of thoracic surgery* **2007**, *83*, 216−221.
(35) Mizuguchi, S.; Nishiyama, N.; Iwata, T.; Nishida, T.; Izumi, N.; Tsukioka, T.; Inoue, K.; Uenishi, T.; Wakasa, K.; Suehiro, S. Serum Sialyl Lewisx and cytokeratin 19 fragment as predictive factors for recurrence in patients with stage I non-small cell lung cancer. *Lung cancer* **2007**, *58*, 369−375.
(36) Iwata, T.; Nishiyama, N.; Nagano, K.; Izumi, N.; Tsukioka, T.; Chung, K.; Hanada, S.; Inoue, K.; Kaji, M.; Suehiro, S. Preoperative serum value of sialyl Lewis X predicts pathological nodal extension and survival in patients with surgically treated small cell lung cancer. *J. Surg. Oncol.* **2012**, *105*, 818−824.
(37) Clément-Duchêne, C.; Carnin, C.; Guillemin, F.; Martinet, Y. How accurate are physicians in the prediction of patient survival in advanced lung cancer? *Oncologist* **2010**, *15*, 782−789.
(38) El-Zayadi, A.-R. Heavy smoking and liver. *World J. Gastroenterol* **2006**, *12*, 6098−6101.
(39) Anderson, N. L.; Anderson, N. G. The human plasma proteome history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **2002**, *1*, 845−867.
(40) Baker, E. S.; Liu, T.; Petyuk, V. A.; Burnum-Johnson, K. E.; Ibrahim, Y. M.; Anderson, G. A.; Smith, R. D. Mass spectrometry for translational proteomics: progress and clinical implications. *Genome Med.* **2012**, *4*, 63.
(41) Locher, C.; Debieuvre, D.; Coëtmeur, D.; Goupil, F.; Molinier, O.; Collon, T.; Dayen, C.; Le Treut, J.; Asselain, B.; Martin, F.; et al. Major changes in lung cancer over the last ten years in France: the KBP-CPHG studies. *Lung Cancer* **2013**, *81*, 32−38.
(42) Debieuvre, D.; Oster, J.-P.; Riou, R.; Berruchon, J.; Levy, A.; Mathieu, J.-P.; Dumont, P.; Leroy-Terquem, E.; Tizon-Couetil, V.; Martin, F.; et al. The new face of non-small-cell lung cancer in men: Results of two French prospective epidemiological studies conducted 10 years apart. *Lung Cancer* **2016**, *91*, 1−6.

DOI: 10.1021/acs.jproteome.9b00457
J. Proteome Res. 2019, 18, 3985−3998

Diagnostic and Prognostic Performance of Blood Plasma Glycan Features in the Women Epidemiology Lung Cancer (WELCA) study

Yueming Hu[a], Shadi Ferdosi[a, †], Erandi P. Kapuruge[a], Jesús Aguilar Diaz de Leon[a], Isabelle Stücker[b], Loredana Radob[b,c], Pascal Guénel[b], and Chad R. Borges[a,]*

[a] School of Molecular Sciences and The Biodesign Institute at Arizona State University, Tempe, AZ, 85281

[b] CESP (Center for Research in Epidemiology and Population Health), Cancer and Environment team, INSERM UMS1018, University Paris-Sud, University Paris-Saclay, Villejuif, France

[c] University Paris Descartes, Faculty of Dental Surgery, Paris, France

* Author to whom correspondence should be addressed: Chad R. Borges, The Biodesign Institute at Arizona State University, P.O. Box 876401, Tempe, AZ 85287. Tel 480-727-9928; email: chad.borges@asu.edu

†Current address: Seer INC, South San Francisco, CA, 94080

Table of Contents

Table S1. Basic Clinical Characteristics and n-values of the WELCA Sample Set.

		WELCA set	
		Controls	**Cases**
Age [a]		61.2 ± 9.73	61.6 ± 9.04
Smoking Status	Never-Smoker	90	36
	Previous-Smoker	72	52
	Current-Smoker	45	98
	Unknown	0	22
Staging	Stage I	N/A	16
	Stage II	N/A	13
	Stage III	N/A	45
	Stage IV	N/A	99
	Unknown Stage	N/A	35
Tumor Histological Types	SCLC [b] – Located	N/A	7
	SCLC - Disseminated	N/A	12
	NSCLC [c] - Adenocarcinoma	N/A	131
	NSCLC - Squamous cell carcinoma	N/A	25
	NSCLC - Large cell carcinoma	N/A	13
	NSCLC - Adenosquamous	N/A	1
	NSCLC - Sarcoma	N/A	3
	Unknown	N/A	16

[a] Age in years ± SD.

[b] Small cell lung cancer

[c] non-small cell lung cancer

Table S2. Stage and Gender Composition of Three Lung Cancer Sample Sets and Their Sub-Cohorts.

Name of Sample Set	Plasma or Serum	Controls (M/F)	Stage I (M/F)	Stage II (M/F)	Stage III (M/F)	Stage IV (M/F)
WELCA set	plasma	0/207	0/16	0/13	0/45	0/99
Dual Gender Lung Cancer set	plasma	123/76	14/6	12/8	50/31	47/31
Stage I Only Lung Cancer set	serum	28/45	33/74	–	–	–

Table S3. Comparison of Glycan Node Stability at Different Conditions Relative to Control Aliquots Stored at -80 °C.[a]

Glycan Node[b]	10 days at -20 °C	90 days at -20 °C	360 days at -20 °C	2 days at 4 °C	90 days at 4 °C	1 day at 25 °C
t-Fucose	ns	ns	ns	ns	ns	ns
t-Gal	ns	ns	ns	ns	ns	ns
2-Man	ns	ns	ns	ns	ns	ns
4-Glc	ns	ns	ns	ns	ns	ns
3-Gal	ns	ns	ns	ns	ns	ns
6-Gal	ns	ns	ns	ns	ns	*
2,4-Man	ns	ns	ns	ns	ns	ns
2,6-Man	ns	ns	ns	ns	ns	ns
3,6-Man	ns	ns	ns	ns	ns	ns
3,4,6-Man	ns	ns	ns	ns	ns	ns
t-GlcNAc	ns	ns	ns	ns	ns	ns
4-GlcNAc	ns	ns	ns	ns	ns	ns
3-GalNAc	ns	ns	ns	ns	ns	ns
3,4-GlcNAc	ns	ns	ns	ns	ns	ns
4,6-GlcNAc	ns	ns	ns	ns	ns	ns

[a]Heavy, stable isotope labeled glucose (Glc) and GlcNAc were utilized to normalize Hexose and HexNAc data, correspondingly.

[b]Results of Friedman test followed by Dunn's post hoc test at 95% confidence level are given. "ns" stands for "not significant". "*" indicates $p < 0.05$.

Table S4. Statistically Significant Differences between Cohorts within the WELCA Study[a].

Glycan Node[b]	Control vs Stage I	Control vs Stage II	Control vs Stage III	Control vs Stage IV	Stage I vs Stage II	Stage I vs Stage III	Stage I vs Stage IV	Stage II vs Stage III	Stage II vs Stage IV	Stage III vs Stage IV
t-Fucose	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
t-Gal	ns	ns	dddd	dddd	ns	dd	ns	ns	ns	ns
2-Man	ns	ns	ns	dd	ns	ns	ns	ns	d	d
4-Glc	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
3-Gal	ns	dd	dd	ns	ns	ns	ns	ns	ns	ns
6-Gal	ns	ns	i	iiii	ns	ns	ii	ns	ns	ns
3,4-Gal	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
2,4-Man	ns	ns	ii	iiii	ns	ns	iiii	ns	ns	iiii
2,6-Man	ns	ii	iii	iiii	ns	ns	ns	ns	ns	ns
3,6-Man	ns	dd	ns	ns	dd	ns	ns	ns	ii	ns
3,6-Gal	ns	dd	ns	ns	ns	ns	ns	ns	ii	ns
3,4,6-Man	ns	dd	ddd	dddd	d	ns	ddd	ns	ns	dd
t-GlcNAc	ns	d	dd	dddd	ns	ns	ddd	ns	ns	ddd
4-GlcNAc	ns	ns	ns	iiii	ns	ns	ns	ns	ns	ns
3-GlcNAc	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
3-GalNAc	ns	ns	dddd	dddd	ns	ns	dd	ns	ns	ns
3,4-GlcNAc	ns	ii	iiii	iiii	i	i	ii	ns	ns	ns
4,6-GlcNAc	ns	d	dd	dddd	dd	dd	dddd	ns	ns	d
3,6-GalNAc	ns	ns	dd	dddd	ns	ns	ns	ns	ns	ns

[a]Hexose data were normalized to the sum of endogenous hexoses, and HexNAc data were normalized to the sum of endogenous HexNAcs.

[b]Kruskal-Wallis test followed by Benjamini–Hochberg false discovery correction procedure at 95% confidence level is given. “ns” stands for “not significant”. “i” and “d” stands for “increased” and “decreased”. i/d indicates $p < 0.05$. ii/dd indicates $p < 0.01$. iii/ddd indicates $p < 0.001$, and iiii/dddd indicates $p < 0.0001$.

Table S5. Stage-by-Stage ROC Comparison of the Top Performing Glycan Nodes.

Stages[a]	Glycan Feature	A: ROC AUC of set A B: ROC AUC of set B	*p*-value of Delong's test for two ROC curves[b]
Stage I **Set A: WELCA Set** **Set B: Stage I-Only Lung Cancer Set**	α 2 - 6 Sialylation	A: 0.733 B: 0.564	0.031 (NS)
	β 1 - 4 Branching	A: 0.696 B: 0.549	0.112 (NS)
	β 1 - 6 Branching	A: 0.797 B: 0.592	0.008
	Antennary Fucosylation	A: 0.609 B: 0.613	0.965 (NS)
Stage I **Set A: WELCA Set** **Set B: Dual Gender Lung Cancer Set**	α 2 - 6 Sialylation	A: 0.733 B: 0.575	0.092 (NS)
	β 1 - 4 Branching	A: 0.696 B: 0.579	0.264 (NS)
	β 1 - 6 Branching	A: 0.797 B: 0.547	0.008
	Antennary Fucosylation	A: 0.609 B: 0.594	0.885 (NS)
Stage II **Set A: WELCA Set** **Set B: Dual Gender Lung Cancer Set**	α 2 - 6 Sialylation	A: 0.681 B: 0.607	0.509 (NS)
	β 1 - 4 Branching	A: 0.707 B: 0.630	0.489 (NS)
	β 1 - 6 Branching	A: 0.770	0.071 (NS)

		B: 0.582	
	Antennary Fucosylation	A: 0.760	0.302 (NS)
		B: 0.655	
Stage III **Set A: WELCA Set** **Set B: Dual Gender Lung Cancer Set**	α 2 - 6 Sialylation	A: 0.796	0.241 (NS)
		B: 0.739	
	β 1 - 4 Branching	A: 0.798	0.407 (NS)
		B: 0.755	
	β 1 - 6 Branching	A: 0.822	0.119 (NS)
		B: 0.745	
	Antennary Fucosylation	A: 0.826	0.161 (NS)
		B: 0.754	
Stage IV **Set A: WELCA Set** **Set B: Dual Gender Lung Cancer Set**	α 2 - 6 Sialylation	A: 0.887	0.009
		B: 0.791	
	β 1 - 4 Branching	A: 0.917	0.002
		B: 0.802	
	β 1 - 6 Branching	A: 0.907	0.008
		B: 0.810	
	Antennary Fucosylation	A: 0.822	0.307 (NS)
		B: 0.777	

[a]The WELCA set was compared to the Dual Gender Lung Cancer Set and Stage I-Only Lung Cancer Set (also dual gender). *N*-values of each group are shown in Table S2. Actual ROC curves are shown in Fig. 4.

[b]"NS" indicates no significant difference between the two compared ROC curves. The significant levels of *p* values are adjusted by Bonferroni multiple comparison correction: $p > 0.013$ (NS), $p < 0.013$ (*), $p < 0.003$ (**), $p < 0.0003$ (***).

Table S6. Comparison of Top Performing Glycan Nodes in Male vs. Female Patients with Early Stage Lung Cancer.

Lung Cancer Sets[a]	Glycan Feature	A: ROC AUC of set A B: ROC AUC of set B	*p*-value of Delong's test for two ROC curves[b]
Dual Gender Lung Cancer Set **Stage I** **Male vs Female** **Set A:** **Male Patients (n = 14)** **vs Male Controls (n = 123)** **Set B:** **Female Patients (n = 6)** **vs Female Controls (n = 76)**	2-linked Mannose	A: 0.618 B: 0.544	0.537 (NS)
	α 2 - 6 Sialylation	A: 0.646 B: 0.553	0.393 (NS) [c]
	β 1 - 4 Branching	A: 0.629 B: 0.539	0.464 (NS) [c]
	β 1 - 6 Branching	A: 0.598 B: 0.579	0.870 (NS) [c]
	4-linked GlcNAc	A: 0.592 B: 0.537	0.686 (NS)
	Antennary Fucosylation	A: 0.552 B: 0.702	0.206 (NS)
Dual Gender Lung Cancer Set **Stage II** **Male vs Female** **Set A:** **Male Patients (n = 12)** **vs Male Controls (n = 123)** **Set B:**	2-linked Mannose	A: 0.628 B: 0.579	0.719 (NS)
	α 2 - 6 Sialylation	A: 0.626 B: 0.602	0.873 (NS)
	β 1 - 4 Branching	A: 0.633 B: 0.618	0.918 (NS)
	β 1 - 6 Branching	A: 0.594 B: 0.559	0.826 (NS)

Female Patients (n = 8) vs Female Controls (n = 76)	4-linked GlcNAc	A: 0.648	0.533 (NS)
		B: 0.564	
	Antennary Fucosylation	A: 0.737	0.382 (NS)
		B: 0.615	
Stage I-Only Lung Cancer Set **Male vs Female** **Set A: Male Patients (n = 33) vs Male Controls (n = 28)** **Set B: Female Patients (n = 74) vs Female Controls (n = 45)**	2-linked Mannose	A: 0.655	0.247 (NS)
		B: 0.547	
	α 2 - 6 Sialylation	A: 0.585	0.766 (NS)
		B: 0.557	
	β 1 - 4 Branching	A: 0.448	0.219 (NS) [c]
		B: 0.563	
	β 1 - 6 Branching	A: 0.450	0.0825 (NS) [c]
		B: 0.616	
	4-linked GlcNAc	A: 0.650	0.508 (NS)
		B: 0.589	
	Antennary Fucosylation	A: 0.632	0.811 (NS)
		B: 0.610	

[a]Comparisons are made for stage I and II of the Dual Gender Lung Cancer Set, and the Stage I Only Lung Cancer Set. Unpaired Delong's test or Bootstrap test are applied to compare two ROC curves.

[b]"NS" indicates no significant difference between the two compared ROC curves. The significant levels of p values are adjusted by Bonferroni multiple comparison correction: $p > 0.0083$ (NS), $p < 0.0083$ (*), $p < 0.0017$ (**), $p < 0.00017$ (***).

[c]p-value is from Bootstrap test instead of Delong's test, because Delong's test should not be applied to ROC curves with different directions and the stratification of Bootstrap is especially useful if groups are not balanced.

Table S7. Correlation Between Age and the Top Performing Glycan Nodes in the WELCA Cases (all stages) and, Separately, Controls.

Case/Control[a]	Glycan Feature	Correlation coefficient (r)	*p*-value of Spearman's rank correlation[b]
Case	2-linked Mannose	0.102	0.168 (NS)
	α 2 - 6 Sialylation	0.073	0.324 (NS)
n = 208	β 1 - 4 Branching	0.072	0.329(NS)
	β 1 - 6 Branching	0.091	0.217 (NS)
	4-linked GlcNAc	0.030	0.681 (NS)
	Antennary Fucosylation	0.148	0.044 (NS)
Control	2-linked Mannose	0.039	0.577 (NS)
	α 2 - 6 Sialylation	0.075	0.283 (NS)
n = 207	β 1 - 4 Branching	0.047	0.501 (NS)
	β 1 - 6 Branching	0.103	0.142(NS)
	4-linked GlcNAc	-0.101	0.148 (NS)
	Antennary Fucosylation	0.205	0.0031

[a]Spearman's rank correlation coefficients and *p* values are provided for the six top performing glycan features in all cases (n = 208) and controls (n = 207).

[b]"NS" indicates no significant correlation between age and the corresponding glycan feature. The significant levels of *p* values are adjusted by Bonferroni multiple comparison correction: $p > 0.0083$ (NS), $p < 0.0083$ (*), $p < 0.0017$(**).

Table S8. Comparison of the Top Performing Glycan Nodes in Different Histological Types.

Histological Types[a]	Glycan Feature	A: ROC AUC of set A B: ROC AUC of set B	*p*-value of Delong's test for two ROC curves[b]
Adenocarcinoma vs Squamous cell carcinoma **Set A: Adenocarcinoma vs Controls** **Set B: Squamous cell carcinoma vs Controls**	2-linked Mannose	A: 0.854 B: 0.926	0.071 (NS) [c]
	α 2 - 6 Sialylation	A: 0.878 B: 0.939	0.130 (NS) [c]
	β 1 - 4 Branching	A: 0.908 B: 0.960	0.114 (NS) [c]
	β 1 - 6 Branching	A: 0.906 B: 0.939	0.539 (NS) [c]
	4-linked GlcNAc	A: 0.877 B: 0.899	0.702 (NS) [c]
	Antennary Fucosylation	A: 0.815 B: 0.861	0.608 (NS) [c]
Adenocarcinoma vs Large cell carcinoma **Set A: Adenocarcinoma vs Controls** **Set B: Large cell carcinoma vs**	2-linked Mannose	A: 0.854 B: 0.860	0.957 (NS) [c]
	α 2 - 6 Sialylation	A: 0.878 B: 0.817	0.647 (NS) [c]
	β 1 - 4 Branching	A: 0.908 B: 0.828	0.586 (NS) [c]
	β 1 - 6 Branching	A: 0.906 B: 0.808	0.402 (NS) [c]

Controls	4-linked GlcNAc	A: 0.877	0.934 (NS) [c]
		B: 0.869	
	Antennary Fucosylation	A: 0.815	0.706 (NS) [c]
		B: 0.757	
Squamous cell carcinoma **vs** **Large cell carcinoma** **Set A: Squamous cell carcinoma vs Controls** **Set B: Large cell carcinoma vs Controls**	2-linked Mannose	A: 0.926	0.556 (NS)
		B: 0.860	
	α 2 - 6 Sialylation	A: 0.939	0.406 (NS)
		B: 0.817	
	β 1 - 4 Branching	A: 0.960	0.426 (NS)
		B: 0.828	
	β 1 - 6 Branching	A: 0.939	0.332 (NS)
		B: 0.808	
	4-linked GlcNAc	A: 0.899	0.804 (NS)
		B: 0.869	
	Antennary Fucosylation	A: 0.861	0.578 (NS)
		B: 0.757	

[a]Comparisons are made for stage IV patients with various histological types of non-small cell lung cancer (NSCLC) vs. all controls. The n-values for the different histological sets are as following. Adenocarcinoma set: n = 70; Squamous cell carcinoma set: n = 8; Large cell carcinoma set: n = 5; Controls: n = 207. Unpaired Delong's test or Bootstrap test are used to compare two ROC curves.

[b]"NS" indicates no significant difference between the two compared ROC curves. The significant levels of *p* values are adjusted by Bonferroni multiple comparison correction: $p > 0.0083$ (NS).

[c]*p*-value is from Bootstrap test instead of Delong's test, because the stratification of Bootstrap is especially useful if groups are not balanced.

Table S9. Stage-by-Stage Comparison of Total Glycosylation with Individual Glycan Feature.

Stages[a]	Glycan Feature		ROC AUC	*p*-value of Delong's test for two ROC curves[b]
Stage I	A	β 1 - 6 Branching	0.797	A vs B: 0.280 (NS) A vs C: 0.196 (NS) A vs D: 0.289 (NS)
	B	Total Hexoses	0.750	
	C	Total HexNAcs	0.730	
	D	Total Hexoses and HexNAcs	0.755	
Stage II	A	β 1 - 6 Branching	0.770	A vs B: 0.024 (NS) A vs C: 0.091 (NS) A vs D: 0.017
	B	Total Hexoses	0.674	
	C	Total HexNAcs	0.627	
	D	Total Hexoses and HexNAcs	0.679	
Stage III	A	2-linked Mannose	0.843	A vs B: 0.938 (NS) A vs C: 0.676 (NS) A vs D: 0.196 (NS)
	B	Total Hexoses	0.844	
	C	Total HexNAcs	0.830	
	D	Total Hexoses and HexNAcs	0.860	
Stage IV	A	β 1 - 4 Branching	0.917	A vs B: 0.021 (NS) A vs C: 0.159 (NS)
	B	Total Hexoses	0.892	
	C	Total HexNAcs	0.891	

	D	Total Hexoses and HexNAcs	0.907	A vs D: 0.280 (NS)

[a]For each stage, the individual top performing glycan node with the largest area under curve (AUC) value was selected to compare to total hexoses (sum of all hexose glycan nodes), total HexNAcs (sum of all HexNAc glycan nodes) and total Hexoses and HexNAcs (sum of all glycan nodes). A paired Delong's test was utilized to compare two ROC curves.

[b]"NS" indicates no significant difference between the two compared ROC curves. The significant levels of p values are adjusted by Bonferroni multiple comparison correction: $p > 0.017$ (NS), $p < 0.017$ (*), $p < 0.0033$ (**).

Table S10. Survival Prediction by the Top Performing Glycan Nodes in All Stages, Stage III and IV Combined, Stage III Only and Stage IV Only.

Stage Involved	Glycan Feature	Cox proportional hazards regression model[a]			
		p-value[b]	Hazard Ratio	Lower bound at 95% CL	Upper bound at 95% CL
All stages	2-linked Mannose	0.0003	2.39	1.49	3.83
	α 2 - 6 Sialylation	0.0002	2.48	1.53	4.03
n = 197	β 1 - 4 Branching	< 0.0001	2.70	1.66	4.41
	β 1 - 6 Branching	0.0002	2.54	1.57	4.12
	4-linked GlcNAc	0.0066	1.99	1.21	3.26
	Antennary Fucosylation	< 0.0001	2.75	1.70	4.42
Stage III & IV	2-linked Mannose	0.0059	2.09	1.24	3.52
	α 2 - 6 Sialylation	0.0029	2.16	1.30	3.58
n = 138	β 1 - 4 Branching	0.0014	2.29	1.38	3.82
	β 1 - 6 Branching	0.0014	2.29	1.38	3.82
	4-linked GlcNAc	0.0148 (NS)	1.98	1.14	3.42
	Antennary Fucosylation	0.0011	2.45	1.43	4.18
Stage III	2-linked Mannose	0.3291 (NS)	1.69	0.59	4.81
	α 2 - 6 Sialylation	0.1551 (NS)	2.12	0.75	5.99
n = 44	β 1 - 4 Branching	0.5653 (NS)	1.35	0.49	3.75
	β 1 - 6 Branching	0.1685 (NS)	2.04	0.74	5.65
	4-linked GlcNAc	0.2910 (NS)	1.68	0.64	4.42
	Antennary Fucosylation	0.0769 (NS)	2.61	0.90	7.58

Stage IV **n = 94**	2-linked Mannose	0.0131 (NS)	2.19	1.18	4.05
	α 2 - 6 Sialylation	0.0051	2.42	1.30	4.50
	β 1 - 4 Branching	0.0011	2.82	1.51	5.26
	β 1 - 6 Branching	0.0032	2.53	1.36	4.67
	4-linked GlcNAc	0.0220 (NS)	2.19	1.12	4.31
	Antennary Fucosylation	0.0081	2.31	1.24	4.31

[a]Cox proportional hazards regression model *p* values and hazard ratios for the top quartile for each glycan node vs. all other quartiles combined, and lower and upper bound at 95% confident limits of hazard ratios are provided.

[b]"NS" indicates no statistically significance between hazard ratio and 1, representing no difference in the relative risk of death, comparing patients in the top quartile vs. all other quartiles of the respective glycan node. The significant levels of *p* values are adjusted by Bonferroni multiple comparison correction: $p > 0.0083$ (NS), $p < 0.0083$ (*), $p < 0.0017$ (**), $p < 0.00017$ (***).

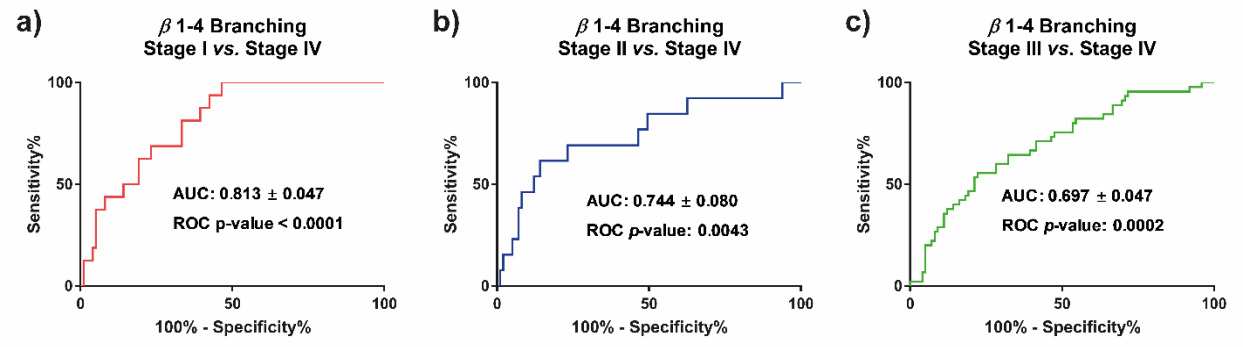


Figure S1. ROC curves for β1-4 branching for stage IV vs. each other stage of non-small cell lung cancer (NSCLC). *N*-values of each group are provided in **Table S1**. ROC curves for stage I-III lung cancer cases vs. controls are provided in panels a-c. Areas under the ROC curves (AUC) values and *p* values are provided under each ROC curve

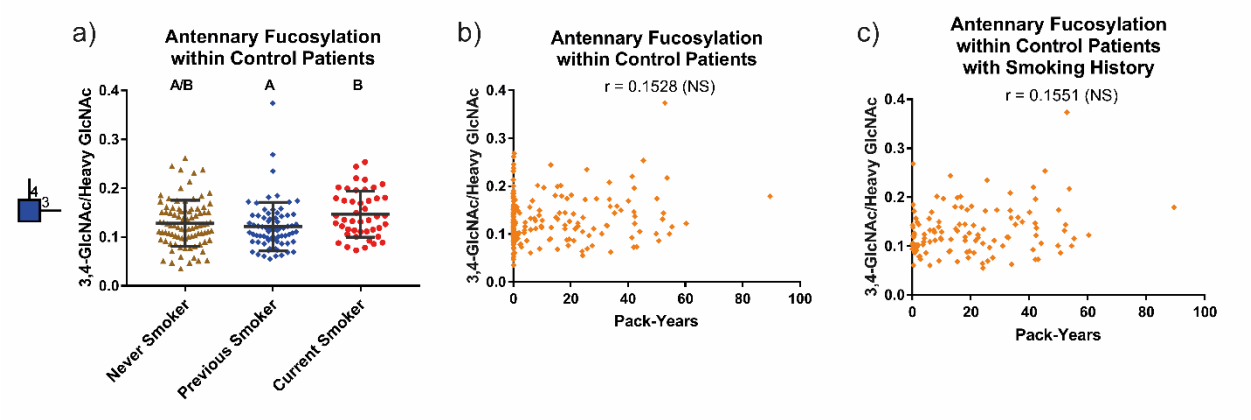


Figure S2. Connection between antennary fucosylation and smoking status within the WELCA control group. (a) The univariate distributions of antennary fucosylation within the control group are shown, subdivided by smoking status. Different letters above the data points indicate statistically significant differences between groups as detected by the Kruskal-Wallis test followed by the Benjamini-Hochberg FDR correction procedure. Spearman's rank correlation between antennary fucosylation and smoking pack-years for (b) all control patients and (c) control patients with smoking history (smoking pack-year > 0). Correlation coefficients are provided above the data points. "NS" next to the correlation coefficient demonstrates a lack of statistical significance. For the other five top performing glycan nodes not shown in this figure, no statistically significant associations with smoking were found.

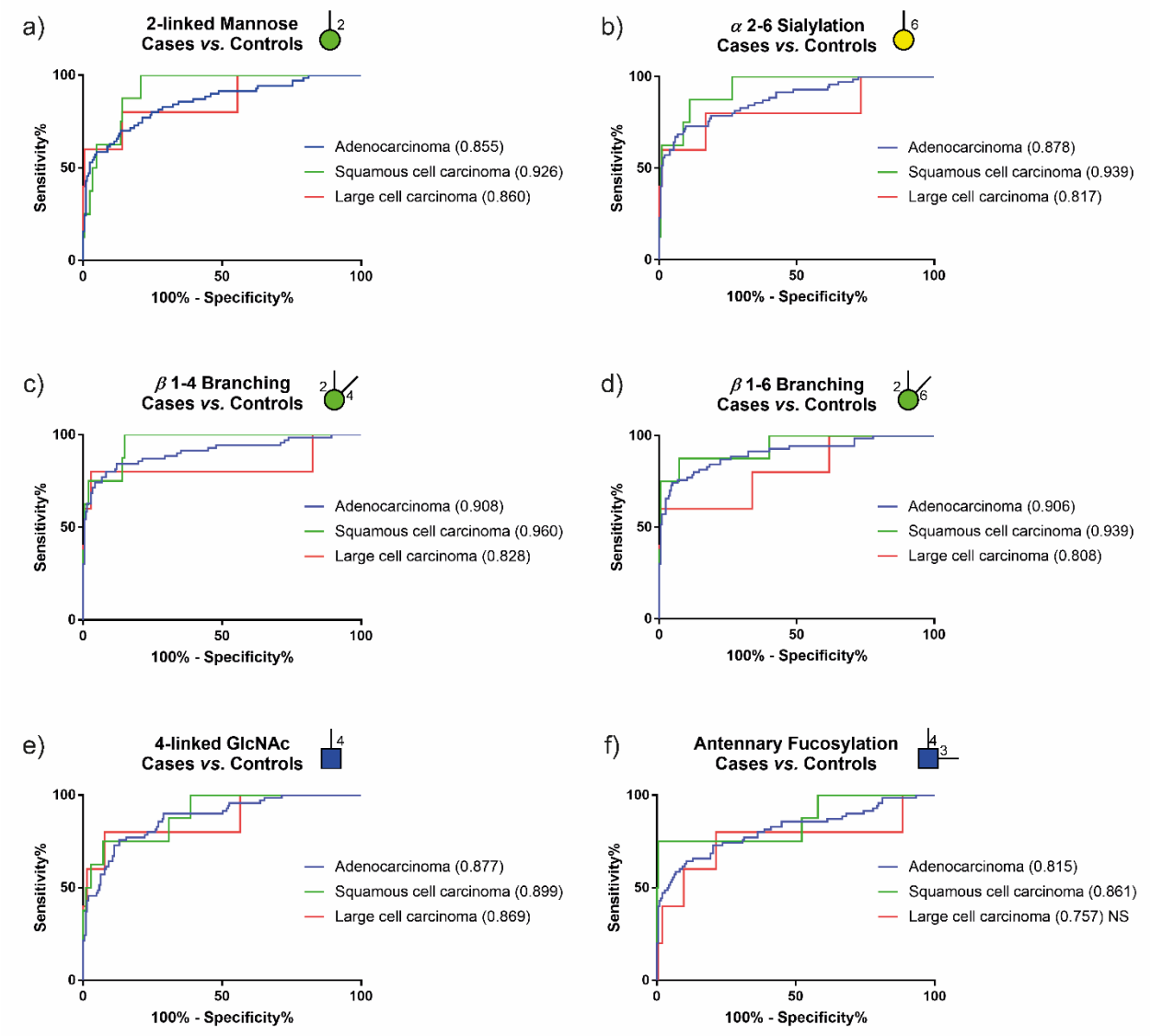


Figure S3. ROC curves for the six top performing glycan nodes within different histological subtypes of non-small cell lung cancer (NSCLC). *N*-values of each group are provided in **Table S8**. Results of unpaired Delong's test and Bootstrap test indicated no significant differences between ROC curves of different histology subtypes of NSCLC (see **Table S8**). ROC AUC values are provided in parenthesis next to the specified histological subtypes. "NS" next to the AUC values indicates no significant difference was found between cases and controls.

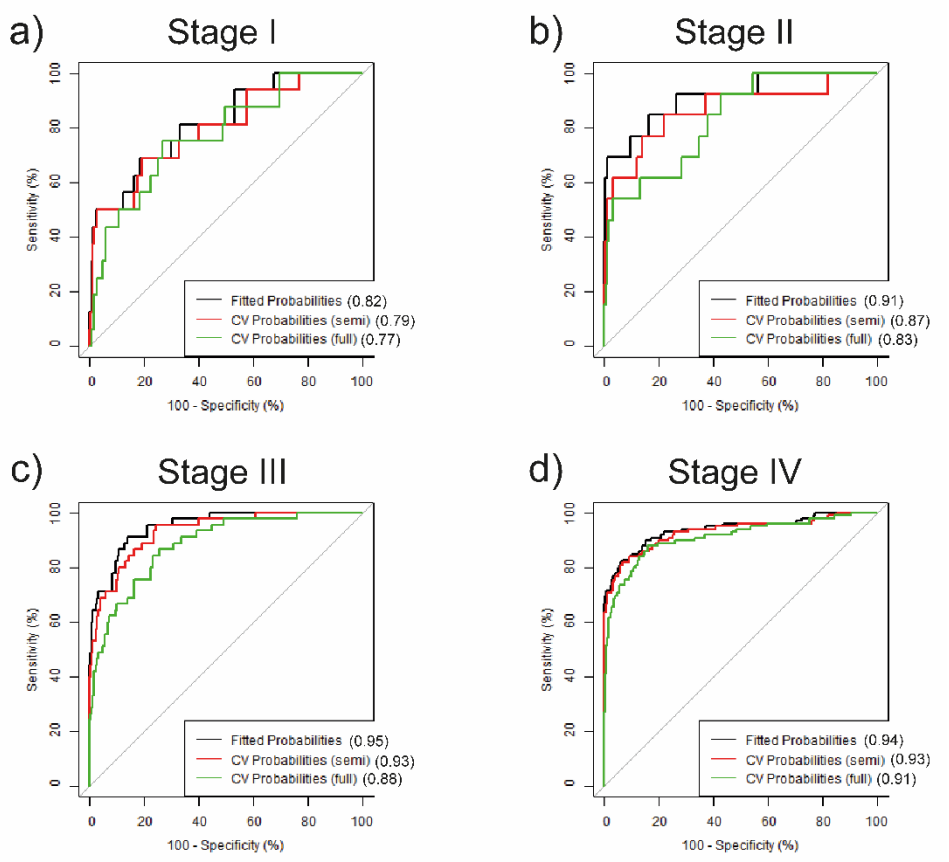


Figure S4. Multivariate logistic regression models for stage I–IV patients from the WELCA data set. Three multivariate logistic regression models were built and corresponding ROC curves were plotted for each stage with different fitting procedures: (1) fitted once on the complete data set and acquired probability (referred to as "Fitted Probabilities") with no use of cross-validation; (2) fitted once on the complete data set, cross-validated with fixed predictors but mobile parameter estimates at each iteration (predicted probability referred to as "CV Probabilities (semi)"); and (3) refitted at each iteration of cross-validation (corresponding probability demonstrated as "CV Probabilities (full)"). ROC AUC values are provided in parenthesis next to the specified models. For each stage, the ROC curve of the best performing individual glycan node was selected and compared to the fully validated multivariate model. No significant differences were detected (Delong's test).

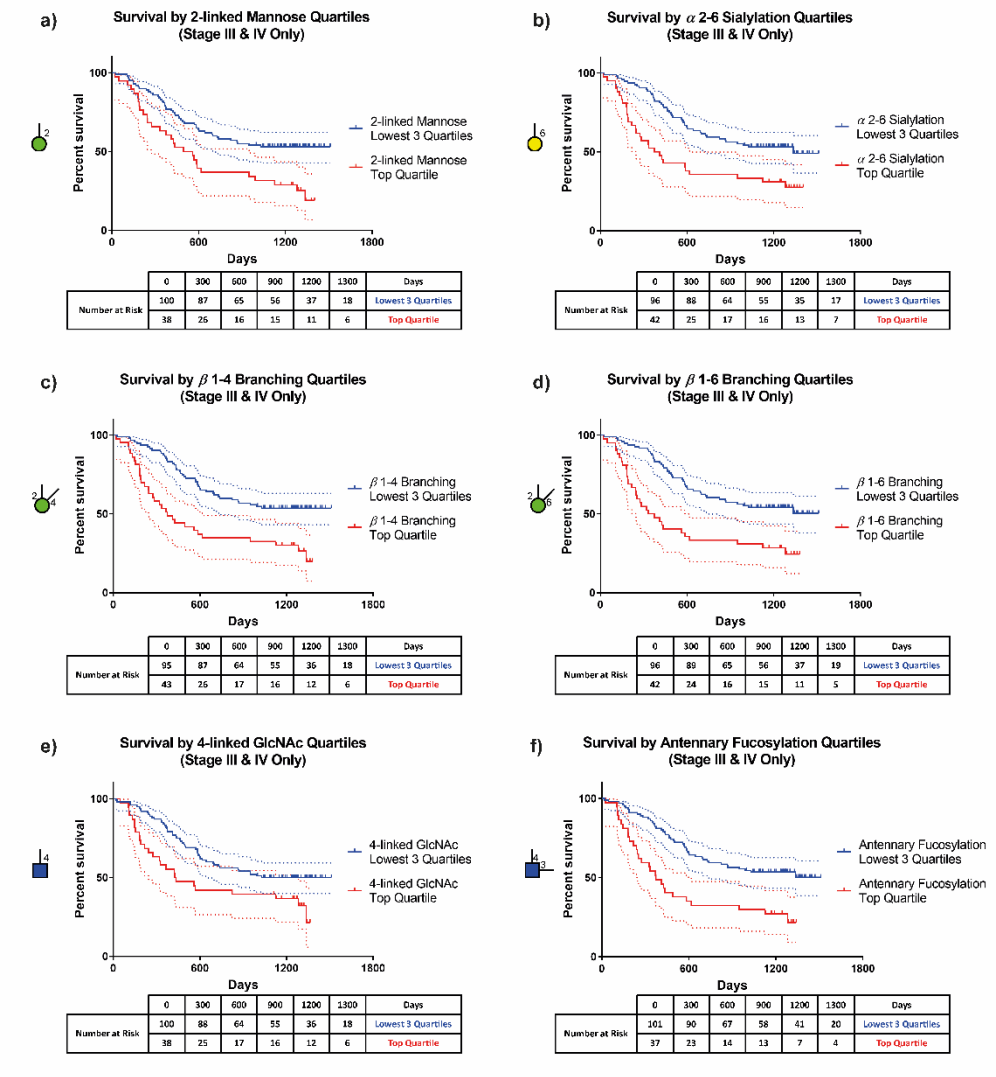


Figure S5. Survival curves of the six top performing glycan nodes for stage III and IV combined. In each panel, the top quartile of specified glycan node is compared to all other quartiles combined. According to the results of a log-rank Mantel-Cox test, the survival curves within each panel are significantly different ($p < 0.05$). Dotted lines represent 95% confident intervals. The median duration of follow-up for patients that died, until death, was 393 days; for survivors this value was 1264 days. The median total follow-up time for all patients was 908 days.

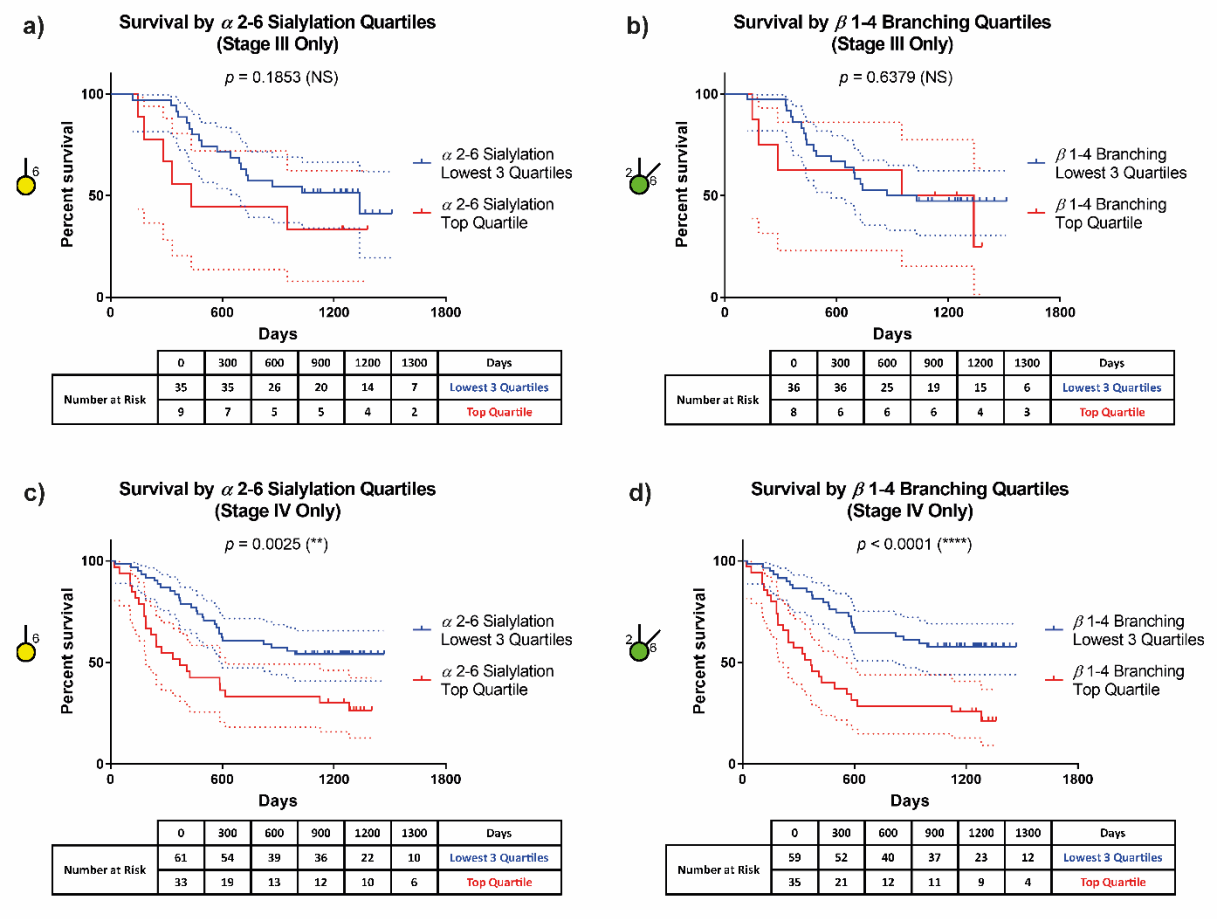


Figure S6. Survival curves for the two top performing glycan nodes that were significantly different between stages III and IV: Stage III patients alone and stage IV patients alone. The top α2-6 Sialylation quartile is compared to all other quartiles combined for stage III patients (panel a) and stage IV patients (panel c). Similarly, the top β1-4 Branching quartile is compared to all other quartiles combined for stage III patients (b) and stage IV patients (d). In each plot, the *p* value of the log-rank Mantel-Cox test is provided, indicating whether significant differences were determined for the two survival curves compared in each plot ("NS" indicates no significant difference, "**" and "****" demonstrate significant difference with $p < 0.01$ and $p < 0.0001$). Dotted lines represent 95% confident intervals. In stage III samples, the median duration of follow-up for patients that died, until death was 458 days; for survivors this value

was 1247 days. The median total follow-up time for all stage III patients was 989 days. In stage IV samples, the median duration of follow-up for patients that died, until death was 357 days; for survivors this value was 1273 days. The median total follow-up time for all stage IV patients was 844 days.