RESEARCH ARTICLE

The glycolyzer: Automated glycan annotation software for high performance mass spectrometry and its application to ovarian cancer glycan biomarker discovery

Scott R. Kronewitter[1], Maria Lorna A. De Leoz[1], John S. Strum[1], Hyun Joo An[1], Lauren M. Dimapasoc[1], Andrés Guerrero[1], Suzanne Miyamoto[2], Carlito B. Lebrilla[1] and Gary S. Leiserowitz[3]

[1] Department of Chemistry, University of California, Davis, CA, USA
[2] Division of Hematology/Oncology, UC Davis Cancer Center, Sacramento, CA, USA
[3] Division of Gynecologic Oncology, UC Davis Medical Center, Sacramento, CA, USA

Human serum glycomics is a promising method for finding cancer biomarkers but often lacks the tools for streamlined data analysis. The Glycolyzer software incorporates a suite of analytic tools capable of identifying informative glycan peaks out of raw mass spectrometry data. As a demonstration of its utility, the program was used to identify putative biomarkers for epithelial ovarian cancer from a human serum sample set. A randomized, blocked, and blinded experimental design was used on a discovery set consisting of 46 cases and 48 controls. Retrosynthetic glycan libraries were used for data analysis and several significant candidate glycan biomarkers were discovered via hypothesis testing. The significant glycans were attributed to a glycan family based on glycan composition relationships and incorporated into a linear classifier motif test. The motif test was then applied to the discovery set to evaluate the disease state discrimination performance. The test provided strongly predictive results based on receiver operator characteristic curve analysis. The area under the receiver operator characteristic curve was 0.93. Using the Glycolyzer software, we were able to identify a set of glycan biomarkers that highly discriminate between cases and controls, and are ready to be formally validated in subsequent studies.

Received: May 22, 2011
Revised: May 18, 2012
Accepted: May 22, 2012

Keywords:
Biomarkers / Clinical glycomics / Data processing / Glycoproteomics / Human serum / Ovarian cancer

1 Introduction

Glycans are a common post-translational modification of proteins that consist of complex arrangements of monosaccharides that vary in size, linkage, and composition. They are instrumental to the vitality of higher organisms and are currently of considerable interest as a source for serum-based biomarkers [1–8]. Glycan cancer biomarkers are of particular importance because changes in glycosylation have been observed in globally released glycans from the serum of cancer patients [4–11] and on glycans released from targeted glycoproteins [12, 13]. Mass spectrometry is widely used for studying glycans because most compounds in a complex mixture can be simultaneously detected and identified. The masses and ionization characteristics of glycans are suitable for most modern mass spectrometers. However, the vast amount of glycan data makes it difficult to extract and organize information from mass spectra.

There have been several methods for annotating glycans incorporating combinatorial approaches (GlycoMod [14]),

Correspondence: Dr. Gary S. Leiserowitz, Division of Gynecologic Oncology, UC Davis Medical Center, 4860 Y Street, Suite 2500, Sacramento, CA 95817, USA
E-mail: gsleiserowitz@ucdavis.edu
Fax: +1-916-734-6034

Abbreviations: FFTs, fast Fourier transforms; **GCC-SPE**, graphitized carbon cartridge solid-phase extraction; **RMS**, root mean squared; **ROC**, receiver operating characteristic curves; **TPI**, total peak intensity

Colour Online: See the article online to view Figs. 4–5 in colour.

empirical databases (GlycoSuiteDB [15, 16], SWEET-DB [17], BOLD [18], KEGG [19], and EUROCarbDB [20]), glycobiology-oriented glycan library models (Cartoonist [21], Retrosynthetic Glycan Network Libraries [22]), and tandem mass spectrometry processing algorithms (StrOligo [23], GlySpy and OSCAR [24, 25], GlycoPeakFinder [26], Glyco-Fragment [27, 28], GlycoWorkbench [29]). However, there has been little attention paid to raw glycan spectra processing. Vakhrushev and co-workers developed the SysBioWare software for processing and annotating raw glycan mass spectra [30]. This program includes several basic features including background subtraction, peak detection, noise thresholding, and data processing tools including preprocessing, smoothing, peak selection, and isotope grouping. Additional tools in the software used for glycan processing include a difference calculator that can use monosaccharide masses and a rudimentary biological filter that uses logical monosaccharide ratio statements entered by the user.

We have developed an integrated software annotation program for glycan biomarker discovery that is referred to as The Glycolyzer. The Glycolyzer contains a full data analysis pipeline in one software package to allow for minimal user intervention. The software was written in IgorPro (WaveMetrics, Portland, OR, USA) language and the source code is available from our group website (chemgroups.ucdavis.edu/~lebrilla/Glycolyzer.zip) or by request from the authors. Although IgorPro is required to run the software, the algorithms can be viewed with text editors. Future versions will be written in a more common programming language. The mass spectrum analysis software is a graphical user interface-based program designed for processing and analyzing carbohydrate mass spectra with a focus on clinical glycan biomarker discovery. The Glycolyzer has similarities with the SysBioWare software but goes further by incorporating a full analysis pipeline including additional algorithms for calibration, theoretical retrosynthetic library based glycan annotation, and statistical hypothesis testing. The overall workflow, including Fourier transform ion cyclotron resonance (FT-ICR) and general mass spectra data processing, is shown in Fig. 1. Calibrated deconvoluted data from LC-MS experiments can be used as well by bypassing the internal preprocessing algorithms and proceeding directly to the annotation and statistics part of the pipeline.

We used this software to discover serum-based glycan biomarkers for epithelial ovarian cancer. Epithelial ovarian cancer is the most dangerous of the gynecologic malignancies due to its propensity for late detection when most patients present advanced stages of the disease. It currently lacks diagnostic tests that are effective for screening and early detection. There are a limited number of FDA approved blood tests available to assist in the diagnosis and monitoring of ovarian cancer, including CA 125 and HE4 [31], but the value of these tests is largely limited to monitoring disease status after treatment, or assessing the risk of malignancy when an ovarian mass has already been detected. CA 125 is elevated in only 50% of Stage I cancers, so it is not a sensitive test for

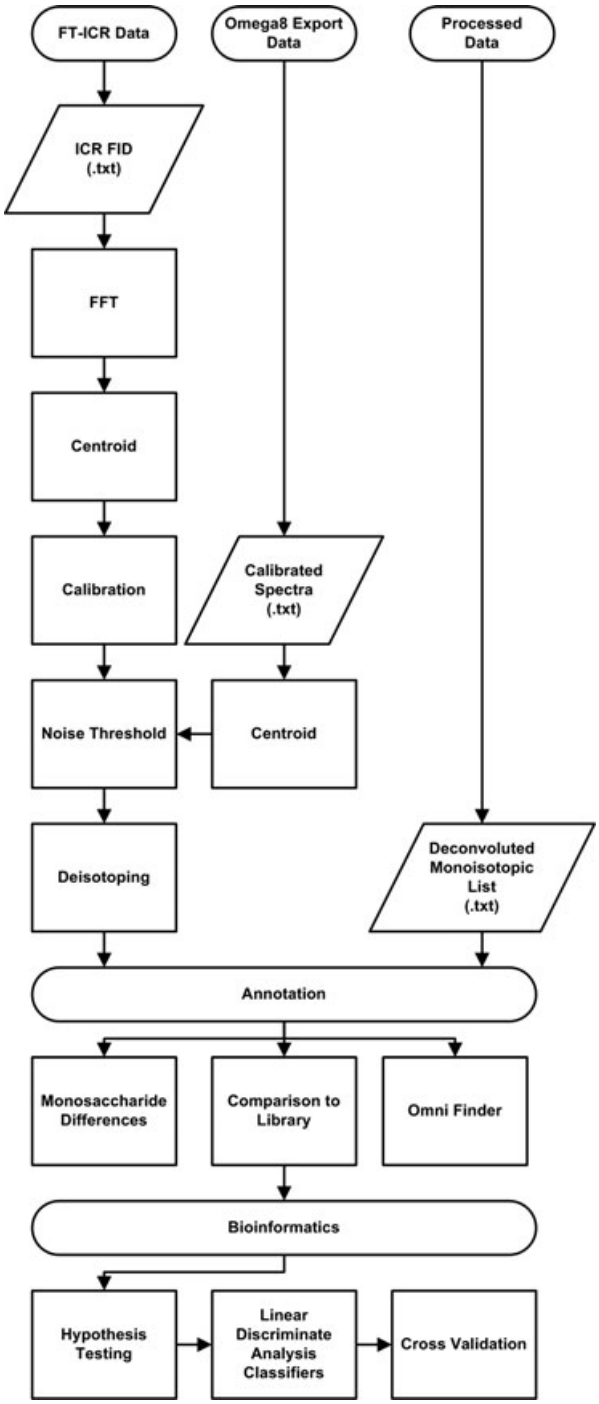

Figure 1. The overall data workflow for the Glycolyzer.

early detection. It is also rather nonspecific, especially in premenopausal women, leading to many false-positive results that require diagnostic intervention [31]. Thus, new novel serum-based biomarkers with improved sensitivity and specificity would be highly desirable.

We have pursued glycomics analysis using mass spectrometry to detect glycans that are altered either by

monosaccharide composition and/or have increased/decreased amounts when comparing the serum from patients with ovarian cancer cases and healthy controls [7, 32]. Although a number of informative glycans were found that distinguished between cases and controls, detection of the glycan mass peaks has been hampered by the lack of useful bioinformatics analytic techniques. Use of the Glycolyzer software provides a platform that can substantially automate the analysis of complex mass spectrometry data, allowing for detection and annotation of informative glycans. Glycan annotation employs a novel theoretical glycan library that has recently been published [22].

As a demonstration of the utility of this software, we have used it to distinguish a unique set of glycans in a carefully selected group of matched cases and controls. Briefly, human serum N-linked glycans (N-glycans) were profiled with the theoretical retrosynthetic N-glycan library and experimental profiles were developed based on 46 control samples. The work presented here demonstrates the high throughput capabilities of the current methodology on a matched set of cases and controls. The methodology includes isolation of N-linked glycans from human serum, mass spectrometry using MALDI-FT-ICR, and then bioinformatic evaluation with the Glycolyzer software. Based on these results, the discovery set is appropriate for use in a clinical validation study to evaluate the robustness of the candidate markers presented here.

2 Materials and methods

2.1 Human serum samples

Approval for this research protocol using clinical data and human serum samples was obtained from the Institutional Review Board of the University of California, Davis Medical Center. Human serum samples were obtained through a formal data use agreement with the Gynecologic Oncology Group (GOG). The subjects either had epithelial ovarian cancer (cancer cases) or were healthy volunteers (healthy controls). All serum samples arrived frozen and were transferred to a −75°C freezer prior to processing.

The discovery set included healthy controls ($n = 48$) and ovarian cancer cases ($n = 46$). The discovery set samples were aged matched by 5-year intervals to avoid confounding effects (40–45, 46–50, 51–55, 56–60, and 61–65 years). Disease status (case versus control) and age block were blinded outside of our laboratory prior to chemical analysis. The samples were blocked into eight sets of 12 samples (each block contained six controls and six cancer cases) with relatively even balancing of subject ages. Following mass spectrometry data collection and annotation using the Glycolyzer software, the samples were unblinded for statistical analysis.

N-glycan release and extraction from human serum for the discovery set was carried out by the optimized methods described by Kronewitter et al. [33]. Briefly, 100 μL of serum was mixed with 100 μL digestion buffer (pH 7.5, 100 mM ammonium bicarbonate, 10 mM dithiothreitol) and heated in boiling water for 2 min to denature the proteins. After cooling to room temperature, 2.0 μL Peptide N-glycosidase F (PNGase F, 500 000 units/mL, glycerol free, New England BioLabs, Ipswich, MA, USA) were added and the mixture was incubated in a microwave reactor for 20 min at a constant power of 20 W. An 800 μL aliquot of chilled ethanol was then added to precipitate peptides and proteins. The solution was frozen in a −75°C freezer for 60 min and then centrifuged at 13 300 revolutions per minute for 20 min (5415 D, Eppendorf AG, Hamburg, Germany). After centrifuging, 700 μL of supernatant was removed from the precipitate and dried in a Savant AES 2010 centrifugal evaporator (Thermo Fischer Scientific, Waltham, MA, USA). PNGase F-released glycans were then purified by graphitized carbon cartridge solid-phase extraction (GCC-SPE) with an automated Gilson GX-274 ASPEC liquid handler. GCC-SPE cartridges (150 mg bed weight, 4 mL cartridge volume) were acquired from Alltech (Deerfield, IL, USA). Three fractions of glycans were collected using increasing amounts of acetonitrile (ACN): 4 mL each of 10% ACN/H_2O (v/v), 20% ACN/H_2O (v/v), and 40% ACN/H_2O (v/v) with 0.05% trifluoroacetic acid. Each fraction was collected and dried in a centrifugal evaporator apparatus. Fractions were reconstituted in nanopure water prior to mass spectrometry. Mass spectra were recorded on an external source MALDI-FT-ICR instrument (HiResMALDI, IonSpec Corporation, Irvine, CA, USA) equipped with a 7.0 T superconducting magnet and a pulsed 355 nm Nd:YAG laser. Five spectra were collected for each sample: 10% ACN and 20% ACN fractions in the positive mode and the 40% ACN fraction in the negative mode. A total of 1410 FT-ICR spectra were collected the 94 samples. The spectra were collected in blocks (blocked by SPE fraction). The samples from the blinded, randomized, sample set were analyzed sequentially on the same instrument over 2–3 days to maintain constant sample detection conditions. The mass spectra collection conditions were optimized for reproducibility by controlling several instrumental parameters during operation. The ultra-high vacuum base pressure was maintained lower than 1×10^{-10} Torr (measured with an ion gauge). Cooling gas was used to kinetically cool the ions during ion accumulation in a hexapole prior to transfer to the ICR cell. The cooling gas pump down rate was controlled via the initial system pressure. The initial system pressure chosen was between 1×10^{-10} and 5×10^{-10} Torr prior to ionization and subsequent accumulation and detection. Fixing the initial pressure allowed for replicate pressure conditions in the ICR cell during detection. Under these conditions, the average coefficient of variation of glycan intensities from technical replicates from the same MALDI spot ranges from 12% to 17% [33].

3 Data analysis algorithms

The Glycolyzer is a software package consisting of a graphical user interface and several modular data processing

algorithms that can be linked to each other in a user defined order. All the algorithms are integrated into the platform's user interface and can be run in series as a data analysis pipeline. Different degrees of processed data can be loaded into the software. For example, the analytical signal from the FT-ICR (ICR transient or free induction decay) can be loaded directly into the start of the pipeline and processed, or the analytical signal can be processed externally to the Glycolyzer software via instrument software (e.g. Omega8, IonSpec) and loaded in at a later point in the data analysis pipeline. This allows data from other types of mass spectrometers to be used as long as the data are already calibrated. If external software deconvolution is preferred rather than the Glycolyzer's built-in deconvolution algorithm, exogenous deconvoluted monoisotopic masses can be loaded directly and the rest of the Glycolyzer's analysis pipeline can still be applied.

3.1 Automatic spectra processing

Data analysis for clinical glycan sample sets requires many automated steps to assure rapid and consistent data handling. The Glycolyzer automates the full data analysis pipeline starting with the analytical signal from the instrument and concluding with biomarker elucidation. The general modules included are: data importing and exporting, FT-ICR signal preprocessing, internal calibration, noise threshold calculation, peak picking, isotope grouping and filtering, glycan annotation, intensity normalization, missing value filling, multiple spectra averaging, hypothesis testing, and multiple testing corrections. The glycans that pass the rigorous multiple testing corrected hypothesis tests are considered to be candidate biomarkers and can be incorporated into data classifiers and their diagnostic performance evaluated.

3.2 Data importing/exporting

Importing data from text files is facilitated by the Glycolyzer's graphical user interface. Raw ICR transients, mass spectra, or deconvoluted monoisotopic mass lists can be loaded in as single files or as a batch. The modular pipeline of the Glycolyzer allows the user to select appropriate analysis algorithms for the data type loaded. Different levels of data preprocessing previously applied to file are taken into account by allowing data to start at different parts in the analysis pipeline.

3.3 FT-ICR preprocessing

Fast Fourier transforms (FFTs) were performed on raw data transients obtained from Omega8 (IonSpec) data acquisition software. The analog-to-digital conversion rate, magnet strength, number of zero fills, and apodization window are specified by the user. In this study, one zero fill was used during the Fourier transform along with a Blackman apodization window. A one second transient was used. In addition, the user is able to truncate the length of the transients prior to applying the FFT to improve quantification by reducing the dampening effects inherent to ICR transients. The FFT converts the transients from the time domain to the frequency domain. Many apodization windows for smoothing out the peak shapes, such as the commonly used Blackman and Hamming windows, are included in the user interface.

3.4 High mass accuracy spectra calibration

High mass accuracy calibration was used for the clinical samples. The error was generally less than 5 parts per million (ppm) root-mean-squared (RMS) mass difference of calibrant ions from calculated values across a data set. Smaller errors, e.g. 1–2 ppm, have been obtained for glycan standards (data not shown) but is challenging for large data sets. Accurate calibrations allow for accurate mass determination of unknowns. For FT-ICR instruments, the free induction decay transients need to be converted into mass spectra via the FFT and calibration equations. The Glycolyzer's internal calibration algorithm performs a six-point calibration using six common glycan ions in each spectrum. A serum N-glycan mass profile, derived from 46 healthy controls, was used to identify the six best ions for calibrating human serum N-glycan spectra [22]. The six calibrant ions were selected from the set of 28 glycans detected in 100% of the samples. The calibrant masses were converted to the frequency domain via the following standard calibration equation [34–36]:

$$m/z(f) = \frac{A}{f - B}.$$

Each calibrant ion mass was aligned to its respective monoisotopic peak in each spectrum. To identify the monoisotopic peak for alignment, the first step is to isotope-filter the frequency data and highlight monoisotopic peaks. Monoisotopic peak selection in the frequency domain is different from the mass domain because the isotopologue distributions are reversed and the neutron mass differences between isotopologue ions are nonlinear in the frequency domain. For this reason, a novel deisotoping algorithm was developed specifically for the frequency domain and presented here. Finally, graphs containing the monoisotopic-highlighted experimental data surrounding each of the six calibrant ions are presented to the user for a final visual inspection. If the wrong peak is selected by the computer, the user can manually reselect the correct peak with arrow buttons then continue to the calibration algorithm and subsequent samples. The manual inspection step ensures proper calibration of densely packed spectra that are hard to decipher with computer algorithms alone. Final calibration is performed by fitting the calibration equation to the calibration ions to find the equation coefficients. The optimized calibration is facilitated by a CurveFit function built into IgorPro that is based on the Levenberg–Marquardt

algorithm. The Omega8's (IonSpec) and the Glycolyzer's internal calibration methods are compared in Supporting Information Fig. S1, where 12 spectra were calibrated and their RMS mass deviations from known values recorded.

3.5 Noise threshold

Separating the signal from the noise is important for peak annotation and reliable quantification. To threshold a spectrum, a limit of detection (LOD) line is calculated. All peaks above the line are considered signal and all peaks below the line are classified as noise. One option for dynamically assigning a LOD is to manually set the threshold to a relative percentage of the base peak. A user-selected threshold is problematic because the cutoff is arbitrary and independent of the noise and background. In contrast, we apply different threshold settings based on the standard deviation and mean intensity of the noise. The mean intensity of the noise is calculated by the average mean of all the peak intensities in the spectra since the number of noise data points greatly outweighs the signal. Commonly, the lower LOD is set at three-sigma above the mean noise level, but we used six-sigma above the mean noise level to further reduce the number of falsely annotated noise peaks.

The standard deviation of the noise is calculated from a histogram of all intensities in the spectrum. This histogram is presented in Supporting Information Fig. S2. The most common intensity in the histogram is the noise level used as the standard deviation. Noise removal by threshold cutoffs drastically improves processing time since the subsequent algorithms are only applied to the signal. Alternately, the standard deviation of the noise is calculated from the full-width-at-half maximum of the distribution. However, the standard deviation from this method is smaller and produces a lower threshold line. Although lower threshold cutoffs allow for higher sensitivity, they also result in less specificity as noise peaks can be detected above the threshold. This algorithm works well for data collected in this study because there are significantly more noise peaks than signal peaks detected in a spectrum.

3.6 Peak picking

The Glycolyzer program requires that each peak has a maximum and contain at least three data points. The centroid mass of each peak is derived by fitting a parabola to the top three points in each peak via parabolic regression. The fit parabola provides a centroid mass and a corrected intensity. Apex-based intensities are used for ICR spectra because peak line shapes and corresponding areas are affected by many variables not directly related to the number of ions in the ICR cell [37]. In contrast, intensities calculated by the area under the curve (AUC) work well for TOF since the TOF detectors are based on counting ions.

3.7 Isotope grouping

Current mass spectrometers commonly resolve glycans into their isotopologues. High resolving power presents the opportunity to identify the monoisotopic peak for further annotation and analysis. Several research groups have developed isotope grouping algorithms [38–42]. The Glycolyzer's general isotope grouping workflow is based on the Thorough High Resolution Analysis of Spectra by Horn (THRASH) algorithm [43] with several modifications pertinent to MALDI ionization and glycans.

One significant improvement is the Glycolyzer's ability to separate overlapping clusters of isotopologues. Rather than using subtractive methods for deconvoluting overlapping distributions, the theoretical overlapped models are reconstructed to reduce the propagation of fitting errors in the residual spectra. The reconstructive approach is similar to the LASSO method applied by Du and co-workers [44]; however, our model generation is permutated rather than regressed with automatic variable selection. A simplified workflow is presented in Supporting Information Fig. S3.

The first step for deconvolving the spectra is to identify an isotopic cluster. A cluster is a set of ions spaced apart by an isotope mass unit equal to 1.00235 Da [43], or a fraction depending on the charge state. The fraction is equal to the isotope unit divided by the charge state. A cluster can contain more than one isotopic distribution if multiple distributions overlap. Overlapped isotopic distributions are common in glycan spectra because chromatographic separation prior to mass spectrometry is typically not performed. MALDI mass spectrometry has the favorable characteristic of only producing ions with a single charge. This eliminates the need for charge deconvolution because the spacing between isotopologues is always a full isotope mass unit rather than a fractional mass related to higher charge states.

Isotope clusters are found in the spectra by a neighbor peak-finding algorithm. The algorithm looks for neighboring peaks around a principal ion that are one isotope mass unit away in both directions. A mass-error tolerance is applied to this calculation to provide a window for locating a neighboring peak apex. This mass error window allows for proper detection of neighboring peaks despite imperfect peak shapes and centroid errors. If a neighboring peak apex is within the error window, it is added to the cluster and the algorithm continues searching for additional ions to add to the cluster. Additional ions are found by making the newly added ion the principle ion and repeating the neighboring peak selection process. This peak finding process continues until there are no neighboring ions to add. If a large mass-error tolerance is selected by the user, the clustering algorithm may falsely include a second cluster if the spectrum is densely populated. However, this type of error will be corrected later in the algorithm when the cluster is deconvoluted (see below). However, if the error is too small, the tail end of a cluster may be broken off and form a second cluster. This condition results in assignment of extra false-positive monoisotopic peaks.

The second step for deconvolution is to create synthetic isotopic distributions. Depending on the type of molecules detected in the spectra, the isotope distributions will change. Peptide mass spectra are often simulated with the use of an averagine. An averagine unit represents, by mass and elemental composition, the average mass of an amino acid that occurs in human proteins. Unknown peptide masses can be converted to elemental compositions by dividing the unknown mass by the averagine mass (111.1254 Da) to find the number of averagine units and then multiplying the number of units by the averagine elemental composition ($C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$) [45]. However, N-glycans have compositions that differ from peptides. Because of the need for sugar-based isotope distribution models, an averagose model was established by An et al. [46]. Subsequently, a direct glycan analogue to the averagine was presented by Vakhrushev et al. [30], which included an average monosaccharide unit based on an equal weighting of hexose, N-acetylhexosamine, fucose, and neuraminic acid monosaccharides. We now propose a similar averagose for modeling N-glycans based on the theoretical libraries and experimentally derived glycan profiles. Experimental serum profiles generated from applying theoretical libraries to experimental spectra provide a more accurate estimate of a human averagose. The proposed serum averagose is $C_{6.0000}H_{9.8124}N_{0.3733}O_{4.3470}S_{0.0}$ with an average mass of 156.64662 Da (sulfur was included as a place holder since it is not typically seen in our spectra). This new more specific averagose is compared against theoretical isotope distributions modeled from the estimates of elemental compositions with Poisson distributions [46]. Supporting Information Fig. S4 demonstrates that both Vakhrushev and Glycolyzer methods produce characteristics similar to the exact elemental composition model. In addition, a peptide averagine was used for glycans and a relatively poor fit was obtained compared to averagose methods. For deisotoping purposes, reducing the length of the ICR transients from 1.0 to 0.5 s (1 048 576–524 288 data points) improved the chi-squared fit of the model to the experimental data.

Next, we improved processing performance by filtering the clusters based on how many isotopic distributions are present in a cluster. Extensive deconvolution is not needed on single ion clusters and is reserved for larger clusters containing several monoisotopic ions and respective distributions. If there are multiple maxima within a cluster, multiple ions are expected and complete deconvolution is performed.

Theoretical isotopologue intensity distributions are then calculated based on an averagose model. If there is only one expected ion in the cluster, a simple theoretical model with one ion is created. However, if multiple ions need to be deconvoluted, combinations of multiple distributions are needed. Overlapping isotope group data reconstruction is accomplished by applying a nonlinear set of 16 ratios between the intensity of multiple clusters to build the model. Sixteen ratios of ion intensities are used to span two orders of magnitude with a small number of steps. This decreases the computer processing overhead while maintaining the desired deconvolution sensitivity. A nonlinear set of ratios are chosen to have greater detail for the ratios close to unity while the more apparent larger ratios are still included. The synthetic models are created with varying amounts of mass unit offsets between the theoretical ions. The number of unit offsets is limited by the number of ions in the cluster to further speed up the processing.

Finally, the complete models are multiplied by an alignment matrix and the individual fits are evaluated with a chi-squared test. The best chi-squared fit alignment is decomposed to identify the monoisotopic peaks and the results are recorded. The monoisotopic and isotopologue peaks are assigned and the theoretical values are subtracted from the spectrum. The algorithm then repeats clustering on the nonannotated portion of the spectra. This process repeats until all the ions above the noise threshold are assigned.

3.8 Glycan annotation

The Glycolyzer provides two methods for annotating peaks using accurate mass: development mode and high-throughput mode. Tools available for use in the development mode include a broad combinatorial method for making theoretical glycans and calculating monosaccharide differences from the spectra. The brute-force combinatorial method can be adapted with biological rules input by the user. Similar "biological filters" have been described in the literature to reduce the quantity of nonsensical glycan compositions [14, 30, 47]. OmniFinder, a dynamic algorithm similar to GlycoMod [14], creates a list of all the mathematically possible glycans or glycopeptides within specified monosaccharide and/or amino acid compositions and searches for them in the spectra. The list is comprehensive but includes a high degree of false-positive hits. The nonsensical glycan false hits are largely eliminated with an array of glycan filters based on known biology.

Another useful tool in the development mode is a glycan peak relationship finder. Mass differences consistent with monosaccharide masses are indicative of an ion being a glycan or a glycoconjugate. This can be helpful with variable or unknown head groups. This information is also helpful for determining families of glycans that differ by one monosaccharide. Finding these differences require processed spectra that only contain monoisotopic masses because many extraneous differences will be found that include associated isotopologues. A stem-and-leaf algorithm is employed to find differences because error bars can be applied to each side of the difference. The stem-and-leaf algorithm starts by looking for imprecise monosaccharide differences and iteratively focuses in on the differences with the least RMS mass error. The adaptive algorithm allows the difference finder to work on poorly calibrated spectra. Calibrated spectra often yield RMS mass errors in the several hundreds of parts-per-billion range for monosaccharide differences. The high accuracy of

correctly matched pairs allows for easy differentiation of true assignments from false ones.

In our glycomics studies, high-throughput annotation was achieved by bounding the glycan composition possibilities to a targeted list of N-glycans. A recently published theoretical glycan library or experimentally derived glycan profile was used as a basis for annotation [22]. In short, the N-glycan library was generated by degrading fully glycosylated complex, hybrid, and high mannose type glycans all the way to the N-linked core. The glycome is bounded by the extent of glycosylation of the starting point glycans. The retrosynthetic degradation provides a well-defined comprehensive list. Subset profiles were rapidly established by scanning the N-glycan library across a set of samples and matching the masses to well calibrated, highly resolved, peaks with masses within a 15 RMS ppm mass error cutoff. Mass profile establishment is critical for advancement from the development stage to the high-throughput biomarker analysis.

Implementation of glycan libraries improves the biomarker detection sensitivity because it focuses the hypothesis testing to only glycan masses. Reducing the number of tests allows for respective performance gains from the multiple testing corrections. The Bonferroni multiple-testing corrections help avoid inflated Type-1 error rates. The size of the glycan profiles is large enough to test all the glycans of interest but still small enough for significant changes to be detected.

The combinatorial glycan method (generating a library by iterating over all possible monosaccharide combinations) was compared with the theoretical glycan library method by examining the fraction of compositions consistent with the library to those that are not. The number of inconsistent combinatorial compositions increases with increasing tolerances for mass assignments. This trend is shown in Supporting Information Fig. S5. The drawback of using an unfiltered combinatorial library is that it generates between 40% and 60% false compositions depending on whether protonated masses or sodiated masses are used; assuming a 15 RMS ppm mass error cutoff. There are more false compositions in the sodiated mass list because of the allowed proton-sodium exchange common to the carboxylic acid group of sialic acid. The sodium substituted cation takes on a multiple sodiated form $[M + (1 + x)Na - (x)H]^+$, where x can be equal to or less than the number of exchangeable acid groups. An N-glycan biological filtered method is not included for comparison because the N-glycan filter is inherent with the theoretical retrosynthetic theoretical N-glycan library [22]. All of the rules are included in the glycan networks and initial starting point ions. Additionally, multiple mass error windows are included for comparison. Supporting Information Fig. S5 depicts the importance of high mass accuracy measurements and shows that as the mass error tolerance increases, the number of false assignments increases.

Since many glycans are present in families that are related by monosaccharides, identifying these differences in spectra helps confirm compositions without the need for tandem mass spectrometry or glycosidase digestion. It is critical that each spectrum is reduced to only monoisotopic peaks prior to searching for monosaccharide differences.

4 Statistics

4.1 Normalization

Normalizing spectra intensities is one of the most important operations in mass spectrometry analysis. It affects intensity values more than any other data operation. The Glycolyzer includes several normalization options: base peak intensity, total ion intensity, total peak intensity (TPI), total library intensity, and select library intensity. Base peak intensity normalization converts peak intensities to a percentage relative to the most intense peak in the spectrum. However, changes in the base peak's intensity cannot be observed and subsequent perturbations to it are propagated to other ions in the spectra. Total ion intensity is based on a sum of all data present in the unprocessed spectrum. Dividing ion intensities by the mean of all ion intensities will normalize the spectrum primarily to the noise level because of the relative sparseness of the ions as compared to the noise. TPI involves normalizing the spectra to the average peak intensity based on only peaks above the noise threshold. This is similar to the method used by Barkauskas et al. on a prostate cancer study [48] and focuses the normalization to intense peaks. The total library intensity option is similar to the TPI except that only annotated peak intensities contribute to the mean total intensity divisor. This allows normalizing by only the ions of interest (N-glycans in this case). The select library intensity normalization further focuses the normalization divisor by including only a select subset of the annotated ions. Prior information on the frequency of detection of library ions in a data set (the percentage of samples containing the ion) can be used to rank the ions so only glycan ions with high detection rate are used for normalization calculations.

Although the different normalization methods tested on this data set produced slightly different sets of significant ions, there was a high degree of similarity between results because the methods all used a constant divisor and only varied by the different sets of ions used to calculate the divisor. The strongest biomarkers were found significant regardless of normalization method. The results from this study are based on the TPI method.

4.2 Spectra averaging

Collecting multiple spectra of the same sample greatly improves the precision of the measurement. As the number of spectra, N, increases, the standard deviation decreases inversely proportional to the square root of N [33]. Replicate spectra can be processed with the Glycolyzer providing the user with two options to incorporate them. The most common

method averages specific ions intensities from each technical replicate together prior to statistical analysis. This works best when target ions are detected in all spectra. An alternate method is to take the highest value of each ion from the set of replicates to use as the value. This situation represents the best-case scenario of data from the sample. This can help overcome some of the variability from the MALDI ionization process, where cold spots on the matrix produce only the most intense ions. Each ion needs only be detected above the threshold in one sample of a given set of replicates to be included. Standard spectra averaging of specific ion intensities were used for the five technical replicates acquired in the discovery set.

4.3 Missing values

When extracting glycan library masses from the data, some of the ions in the profile are not detected in the data above the noise threshold or are missed by deisotoping errors. The absence of a peak is useful when monitoring the frequency of detection of an ion (presence or absence) across a sample set but often causes problems with downstream statistics calculations. The solution to the missing data implemented here is to look below the threshold and find the largest peak within a prescribed mass error window. Filling in noise values for missing peaks should result in higher quality biomarkers because the former zero values will skew distributions of low intensity ions that are near the noise threshold cutoff. However, very low intensity peaks can be over represented if the number of zeroes is greater than the number of detected peaks across a data set. Although a potential problem, this scenario typically does not lead to an increased number of false-positive biomarkers because the glycans with large amounts of missing values will not pass the strict hypothesis tests due to high variance caused by the randomness of the low intensity peaks used for data filling.

4.4 Multiple statistical hypothesis testing

Each glycan annotated by the theoretical profile is subjected to hypothesis testing to determine if any changes are significant. Five technical replicate FT-ICR spectra from each sample are averaged prior to hypothesis testing. The natural logarithm of the intensities is used for testing to prevent the most intense ions in the spectra from overwhelming the less intense species. Furthermore, taking logarithms of the intensities improves the assumption of constant error variance and makes the data better suited for standard statistical testing [4]. Two-tailed *t*-tests were used for hypothesis testing. Due to the large amount of independent glycans tested in this manner, multiple testing corrections should be employed. Bonferroni corrections are implemented to add rigor to the testing by maintaining the family-wise error rate. Glycans with significant changes in intensity are found when they pass the hypothesis testing ($p < 0.05$) and the Bonferroni multiple testing correction ($n = 101$ for the number of glycan masses in the library).

4.5 Linear classifier motif tests

The significant markers that passed the *t*-test were combined into a motif test that leverages deviations in case intensities from control mean intensities. Combining multiple markers into a diagnostic panel has been shown previously to improve discrimination [49, 50]. To obtain a score for each sample, each glycan in the motif test is weighted by the difference between the mean control ion intensity and the mean case ion intensity. The larger the difference between the mean is, the larger the weighting factor. The scoring scheme was set up by adding the absolute value of the marker ion deviations from the control mean. This allows the summation of positive and negative deviations found in the biomarker motif test. The net score is used to classify unknown samples; whereas the samples consistent with the motifs, and thus larger deviations, score higher. A separate motif test was developed for each ACN fraction. The results are summarized with receiver operating characteristic curves (ROC) and evaluated by the AUC. The AUC is calculated by geometric integration. Applying motif tests to the discovery set provided high AUC results for the three fractions: 10% (0.89), 20% (0.87), and 40% (0.88). When weighted evenly across the 10%, 20%, and 40% fractions, a linear combination of the motif test scores can be linearly combined into an overall test metric. The overall test improves sensitivity and specificity and increases ROC AUC to 0.93. The ROC curve results are included in Fig. 2.

4.6 Data modeling

The data analysis pipeline was evaluated by modeling the data with perturbation analysis. Synthetic case and control mass spectra were created with perturbed intensities. A representative sample spectrum was selected and used to seed new spectra. Each intensity value was modified with a multiplicative factor generated randomly from a normal distribution using a Box–Muller simulation [51]. Several data sets were generated to include distributions in intensity values that produced coefficients of variation of 10%, 20%, 30%, 40%, and 60%. The randomization was evaluated by comparing two sets of unperturbed control spectra. After data processing, no significant biomarkers were detected ($p = 0.05$) indicating the data is sufficiently randomized in the model. An example plot of 48 simulated control spectra with a coefficient of variation of 60% is included in Supporting Information Fig. S6. At each coefficient of variation perturbation, two sets of 48 spectra (one for case and the other for control) where the case set contained one glycan ion with its mean intensity value increased by 5%, 10%, 25%, 50%, 100%, or 150% relative to the control. This change in abundance simulates the effect of a

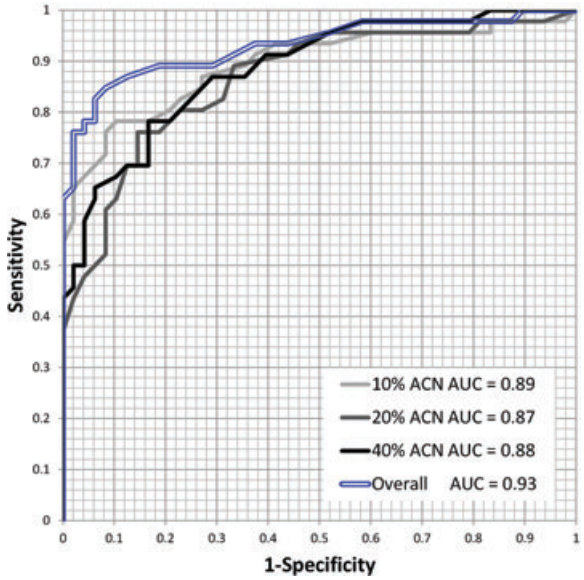

Figure 2. Receiver operating characteristic curve results from applying the glycan motif test to the discovery sample set. The area under the curves represents a high degree of specificity and sensitivity for all three fractions independently. The "overall" trace is an evenly weighted linear combination of the motif scores from the 10%, 20%, and 40% fractions.

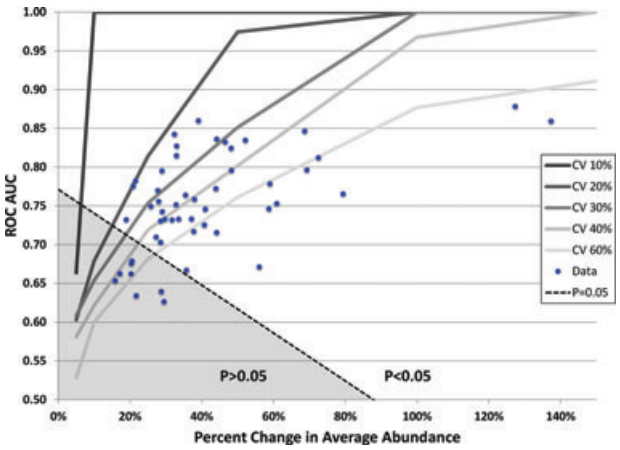

Figure 3. Evaluation of the performance of the Glycolyzer classifier using controlled modeled data. The trend lines demonstrate how the ROC analysis responds with variation within a data set and separation in average intensities between the data sets. The dots correspond to the glycans significantly changing between disease states. The dashed line corresponds to an approximated $p = 0.05$ cut-off value determined from the modeled data. Experimental biomarkers below the simulated $p = 0.05$ line are marked with * in Table 1.

case biomarker ion deviating in intensity from the control. The resulting case sets contained one theoretical biomarker with increasing perturbations that could be used to test the Glycolyzer's ability to detect biomarker changes.

The synthetic sets of raw spectra were processed with the Glycolyzer's preprocessing and statistical algorithms and the *p*-values and ROC AUC values were recorded. The hypothesis testing analysis was based off of the single biomarker programmed into the model and the same multiple testing corrections were applied ($N = 101$). Trend lines depicting the relationship between percent change in abundance and ROC AUC are plotted in Fig. 3. Approximating the ROC AUC values at $p = 0.05$ using linear interpolation of the data allows for the calculation of a $p = 0.05$ cut-off line. Plotting the interpolated ROC AUC values versus interpolated percent abundance change is shown in Supporting Information Fig. S7. The $p = 0.05$ cut-off line is presented as a dashed line in Fig. 3. Modeled values higher than this dashed line in Fig. 3 would pass the hypothesis tests. The glycan biomarkers detected from the experimental data were overlaid to demonstrate how well the experimental data followed the trends and pass the modeled *p*-value cut-off line. A total of 85% of the experimentally determined biomarkers were above the modeled cut-off line.

5 Discussion

The Glycolyzer successfully calibrated and processed 1410 transients from the ovarian cancer discovery set and identified several candidate glycan biomarkers. The unified software approach streamlined the data analysis and allowed for results to be obtained on the same day the last spectrum was collected. The empirically derived serum glycan profile [22] (containing 101 glycan masses) developed in-house was used to filter the data. After statistical analysis, 51 glycan candidate biomarkers (39 glycan masses due to detection in multiple fractions) were identified using a $p < 0.05$ cutoff for Bonferroni corrected *p*-values ($n = 101$). The candidate markers with their monosaccharide compositions and *p*-values are summarized in Table 1 . The full list of 101 glycans monitored is included in Supporting Information Fig. S8 and the compositions are included in Supporting Information Fig. S9. The glycan log mean intensities and standard deviation are also presented.

The data quality and processing improvements can be observed by selecting biomarker m/z 1809.63 from Table 1 as a case study. Plotting all of the mass spectra from the controls and juxtaposing it to all of the spectra from the cases shows that even without data processing, the abundance has decreased on average. This is shown with the mass spectra zoom profiles in Fig. 4. The data processing improvements to the data can be exemplified using box plots in which the 0, 25, 50, 75, and 100 percentiles are shown for the normal and cases in Fig. 5. The logarithm and normalization procedures applied tightened up the data distributions, produced more symmetric distributions and biomarker discernment fidelity.

Twelve of the glycans identified as significant were detected in more than one elution fraction. Interestingly, all 12 were detected in multiple fractions and had consistent trends of increasing or decreasing intensities. Although it is possible that glycan isomers were crudely separated along the SPE

Table 1. Candidate glycan biomarkers found with Bonferroni corrected *p*-values less than 0.05

ID	ACN elution fraction	Exact *m/z*	RMS ppm error	Intensity relative percent change	Intensity Log_{10} control mean	Intensity Log_{10} control CV	Intensity Log_{10} cases mean	Intensity Log_{10} cases CV	Bonferroni corrected *p*-value	Hex	HexNAc	Fucose	NeuAc	Na/H	ROC	Motif test ROC AUC
1	10%	1257.423	7.5	−43.9%	3.99	6.5%	3.74	7.6%	1.1×10^{-3}	5	2	0	0	0	0.77	89.4
2	10%	1298.449	10.8	−33.0%	3.24	4.7%	3.06	5.6%	6.4×10^{-5}	4	3	0	0	0	0.81	
3	10%	1419.475	1.8	−27.3%	5.19	3.0%	5.05	4.2%	3.0×10^{-2}	6	2	0	0	0	0.71	
4	10%	1444.507	10.5	−35.5%	3.50	5.8%	3.31	6.2%	9.4×10^{-4}	4	3	1	0	0	0.76	
5	10%	1460.502	9.6	−48.2%	4.02	6.1%	3.73	8.6%	3.1×10^{-4}	5	3	0	0	0	0.80	
6	10%	1485.534	5.1	60.9%	5.09	3.4%	5.30	5.6%	5.0×10^{-3}	3	4	1	0	0	0.75	
7	10%	1501.529	6.6	−38.0%	4.38	4.7%	4.17	6.1%	1.9×10^{-3}	4	4	0	0	0	0.76	
8	10%	1542.555	5.3	79.4%	4.62	5.8%	4.87	7.1%	8.3×10^{-3}	3	5	0	0	0	0.77	
9	10%	1647.586	2.7	−27.8%	5.09	2.6%	4.95	3.4%	9.2×10^{-4}	4	4	1	0	0	0.77	
10	10%	1663.581	9.0	−48.2%	4.30	4.8%	4.01	6.7%	5.9×10^{-6}	5	4	0	0	0	0.82	
11	10%	1793.644	10.3	−39.1%	2.96	4.5%	2.75	6.2%	6.4×10^{-8}	4	4	2	0	0	0.86	
12	10%	1809.639	7.6	−52.2%	4.30	5.7%	3.98	6.2%	5.2×10^{-7}	5	4	1	0	0	0.83	
13	10%	1850.666	3.6	−21.5%	5.36	2.2%	5.26	2.3%	2.7×10^{-3}	4	5	1	0	0	0.78	
14	10%	1996.724	8.9	−33.0%	3.52	3.4%	3.35	4.8%	3.8×10^{-6}	4	5	2	0	0	0.83	
15	10%	2012.719	3.0	−32.4%	5.02	2.5%	4.85	3.1%	2.0×10^{-6}	5	5	1	0	0	0.84	
16	10%	2028.714	10.3	−28.9%	2.96	4.2%	2.81	5.2%	4.7×10^{-5}	6	5	0	0	0	0.80	
17	10%	2158.777	10.5	−46.5%	3.44	6.1%	3.16	7.0%	1.4×10^{-6}	5	5	2	0	0	0.83	
18	20%	1257.423	7.3	−35.7%	4.69	4.8%	4.50	10.1%	3.7×10^{-6}	5	2	0	0	0	0.67	86.5
19	20%	1339.476	9.9	58.7%	3.15	6.8%	3.35	9.1%	1.0×10^{-10}	3	4	0	0	0	0.75	
20	20%	1444.507	10.9	−40.7%	3.97	4.8%	3.74	8.9%	3.5×10^{-11}	4	3	1	0	0	0.73	
21	20%	1485.534	4.2	69.3%	5.31	4.2%	5.54	5.0%	5.6×10^{-19}	3	4	1	0	0	0.80	
22	20%	1501.529	10.0	−28.5%	4.11	3.0%	3.96	6.4%	7.5×10^{-11}	4	4	0	0	0	0.70	
23	20%	1542.555	11.5	59.0%	2.67	5.8%	2.87	8.9%	1.5×10^{-14}	3	5	0	0	0	0.78	
24	20%	1606.560	11.1	−21.7%	3.60	8.0%	3.49	11.0%	4.8×10^{-2}	5	3	1	0	0	0.63*	
25	20%	1647.586	7.5	−18.9%	5.73	2.0%	5.63	2.5%	3.5×10^{-11}	4	4	1	0	0	0.73	
26	20%	1663.581	7.5	−37.2%	4.44	4.5%	4.24	7.0%	2.8×10^{-14}	5	4	0	0	0	0.73	
27	20%	1688.613	8.8	44.2%	3.85	6.4%	4.01	6.2%	1.0×10^{-9}	3	5	1	0	0	0.72	
28	20%	1704.608	10.2	20.6%	3.16	4.9%	3.25	6.5%	1.2×10^{-3}	4	5	0	0	0	0.68*	

Table 1. Continued

ID	ACN elution fraction	Exact *m/z*	RMS ppm error	Intensity relative percent change	Intensity Log_{10} control mean	Intensity Log_{10} control CV	Intensity Log_{10} cases mean	Intensity Log_{10} cases CV	Bonferroni corrected *p*-value	Hex	HexNAc	Fucose	NeuAc	Na/H	ROC	Motif test ROC AUC
29	20%	1793.644	6.6	−28.6%	3.60	5.0%	3.45	9.6%	3.2×10^{-6}	4	4	2	0	0	0.73	
30	20%	1809.639	6.4	−44.1%	5.50	3.1%	5.24	4.3%	8.8×10^{-33}	5	4	1	0	0	0.84	
31	20%	1825.634	5.4	−37.8%	3.35	8.1%	3.15	10.3%	6.0×10^{-10}	6	4	0	0	0	0.72	
32	20%	1955.697	9.5	−41.0%	3.49	7.1%	3.27	9.5%	1.3×10^{-14}	5	4	2	0	0	0.75	
33	20%	1976.659	10.1	56.1%	2.82	8.3%	3.01	15.4%	1.2×10^{-5}	5	4	0	1	1	0.67	
34	20%	1996.724	9.8	−15.8%	2.98	4.2%	2.90	6.1%	7.7×10^{-3}	4	5	2	0	0	0.65*	
35	20%	2012.719	4.6	−20.4%	4.48	4.3%	4.38	5.6%	2.0×10^{-4}	5	5	1	0	0	0.68*	
36	20%	2122.717	10.0	28.7%	2.54	6.9%	2.65	12.1%	1.2×10^{-2}	5	4	1	1	1	0.64*	
37	20%	2158.777	7.7	−32.8%	3.48	5.9%	3.30	6.8%	2.1×10^{-13}	5	5	2	0	0	0.75	
38	20%	2174.772	7.0	−31.7%	3.61	5.7%	3.44	6.2%	1.8×10^{-13}	6	5	1	0	0	0.73	
39	20%	2325.796	9.5	29.4%	2.49	7.4%	2.60	12.6%	8.1×10^{-3}	5	5	1	1	1	0.63*	
40	20%	2377.851	9.8	−17.2%	2.72	5.5%	2.64	6.1%	3.4×10^{-3}	6	6	1	0	0	0.66*	
41	20%	3271.090	7.8	−20.3%	2.55	9.3%	2.45	6.7%	1.0×10^{-2}	8	5	2	2	2	0.66*	
42	40%	1274.453	18.2	−20.9%	3.08	4.0%	2.97	4.3%	8.1×10^{-3}	4	3	0	0	0	0.78	88.4
43	40%	1727.601	12.5	29.0%	3.59	4.3%	3.70	3.4%	1.3×10^{-2}	5	3	0	1	0	0.74	
44	40%	2159.812	15.9	28.0%	3.21	3.7%	3.32	4.4%	1.0×10^{-2}	3	6	3	0	0	0.76	
45	40%	2441.870	5.8	137.4%	3.97	5.1%	4.35	7.2%	8.5×10^{-8}	6	5	1	1	0	0.86	
46	40%	2513.854	16.8	72.5%	3.25	5.4%	3.49	6.0%	3.0×10^{-6}	12	2	1	0	0	0.81	
47	40%	2515.907	16.6	29.7%	3.01	3.7%	3.12	5.0%	6.4×10^{-3}	7	6	1	0	0	0.73	
48	40%	2732.966	14.3	33.6%	3.20	4.4%	3.33	4.9%	6.2×10^{-3}	6	5	1	2	0	0.73	
49	40%	2754.948	15.5	68.7%	3.07	4.0%	3.30	6.6%	1.6×10^{-6}	6	5	1	2	1	0.85	
50	40%	2807.003	14.3	127.4%	3.16	5.5%	3.52	7.8%	4.7×10^{-9}	7	6	1	1	0	0.88	
51	40%	2816.039	16.8	25.8%	2.96	4.1%	3.06	4.2%	1.1×10^{-2}	4	7	3	1	0	0.75	

The "ACN Elution Fraction" shows which acetonitrile fraction the glycan was detected and found to be significantly changing. Several glycans eluted in multiple fractions and were found significant in both fractions. "Exact *m/z*" is the calculated mass to charge ratio of the sodiated ion in the 10% and 20% fraction or the deprotonated ion in the 40% fraction. "Intensity Relative Percent Change" depicts the magnitude that the normalized glycan intensity is increasing or decreasing in the cancer state relative to control. The "Intensity Log_{10} Mean" corresponds to the logarithm of the mean intensity of the glycan from either the controls or cancer cases and the "Intensity Log_{10} CV" is the coefficient of variation of that mean. "Bonferroni Corrected p-value" is the *t*-test statistic corrected for 101 independent tests. The monosaccharide composition symbols are abbreviated: Hex: hexose; HexNAc: N-acetylglucosamine; Fucose: deoxyhexoe; NeuAc: neuraminic acid; and Na/H: sodium cation substitution for proton. "ROC AUC" stands for the area under the curve of a ROC plot for a specific ion. An asterisks indicates a borderline biomarker because the area under the curve was not greater than the modeled $p = 0.05$ significance line. The "Motif Test ROC AUC" shows the area under the curve of a ROC plot for all the ions combined into a motif test for each fraction.

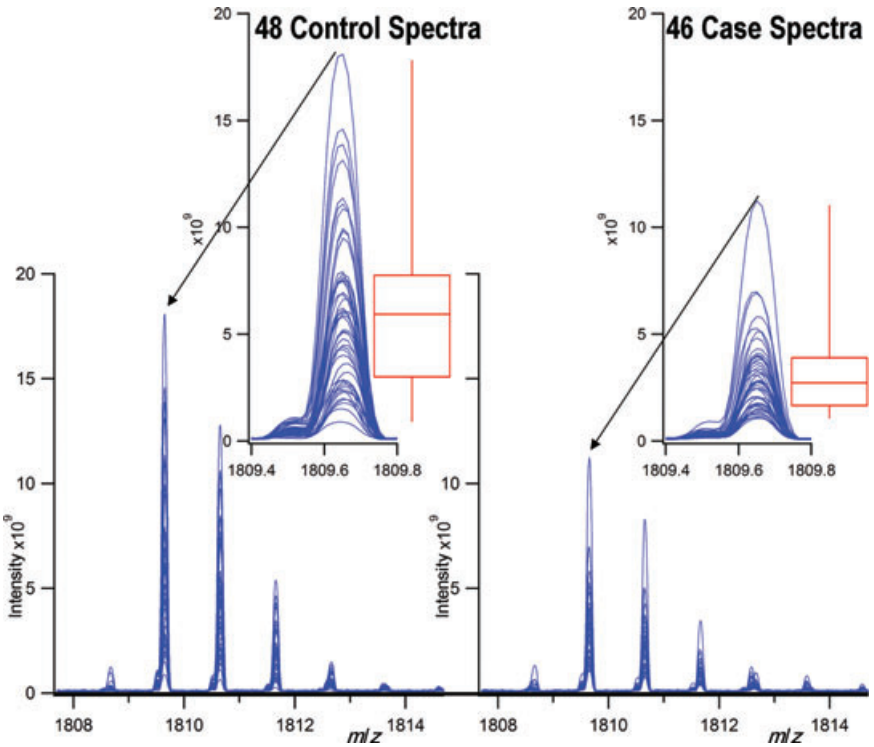

Figure 4. Comparison of mass spectra centered on the isotope envelope of *m/z* 1809.639 and its monoisotopic mass. The left plots correspond to the unprocessed, averaged data from the controls while the right plots correspond to the unprocessed, average data from the cases. Since the plots overlap, box plots were included to show the 0, 25, 50, 75, and 100 percentiles.

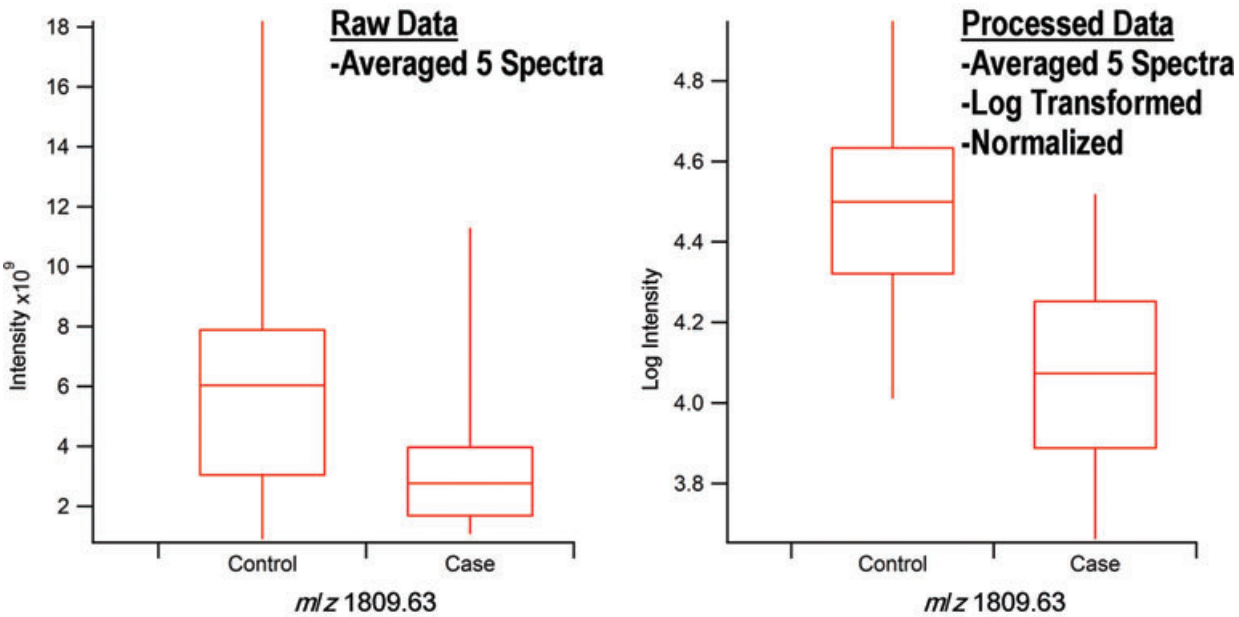

Figure 5. Improvements from data processing. The left box plots correspond to the unprocessed, averaged data while the right box plots correspond to the same data after it is log transformed and normalized.

fractional lines, the constant trends of the glycans across fractions suggest a split fractionation of single glycan structure. Several glycan compositions were detected with and without fucose. When the fucosylated/nonfucosylated pairs were detected in more than one fraction, the fucosylated form was more intense in the later fraction. This is consistent with the elution order observed with graphitized columns and LC/MS.

N-glycans are synthesized enzymatically with glycosyltransferases and glycosidases and are built up one monosaccharide unit at a time. This process results in families of glycans that differ from each other by only one monosaccha-

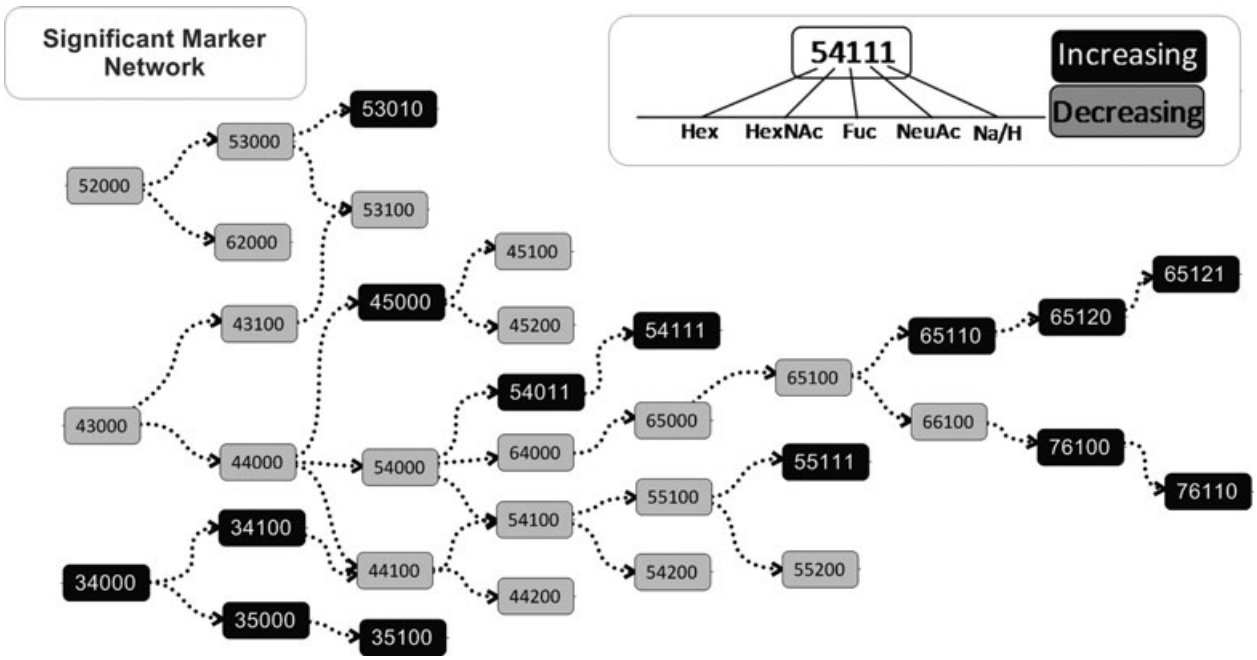

Figure 6. A network of the statistically significant glycan biomarkers. The compositions correspond to the number of units: Hexose-N-Acetylhexosamine-Fucose-Neuraminic Acid-Sodium Substituted Proton.

ride. The statistically significant biomarkers detected in this study primarily come from a single family. A glycan network in Fig. 6 shows 34 of the 39 glycan nodes can be linked either directly or indirectly to all other glycans in the family by single monosaccharide links. The glycan network is also coded with the change in intensity trends between the controls and the cancer cases. Detection of glycans in families increases the confidence of the glycan annotations because additional orthogonal information is used beyond exact mass.

Most of the neutral glycans were decreasing in intensity while most of the glycans containing sialic acid were increasing in intensity. However, one set of neutral glycans, consisting of a subfamily of glycans located in the bottom left of Fig. 6 ($Hex_3HexNAc_4$, $Hex_3HexNAc_4$ Fuc_1, $Hex_3HexNAc_5$, Hex_3Hex NAc_5Fuc_1) was increasing. Specifically, this increase in the FA2 glycan ($Hex_3HexNAc_4Fuc_1$) is consistent with reported increases detected in ovarian cancer patients from serum IgG and whole serum [13]. The FA2 nomenclature is described by Gornik et al. [52]. Kim et al. also reported the same increase in FA2 levels in serum [53]. Eight significant glycan biomarkers containing sialic acid were detected in the 20% and 40% fractions. Increasing changes in sialylated glycan intensities are consistent with other reports in literature where sialylated glycans have been implicated in cancer detection and metastases [54–56]. Many sialic acid containing and sialic acid free glycan pairs, such as $Hex_5HexNAc_4$ and $Hex_5HexNAc_4$ $NeuAc_1$, showed a trend of increasing sialic acid containing and decreasing sialic acid free intensities. This conflicting trend is consistent with the sialic acid free glycans being used as substrates for upregulated sialyltransferases, which produce sialylated glycans. The sialic acid pairs and mean intensities are listed in Table 2. Although we cannot confirm that the sialylated/nonsialylated pairs have the same core structure, the trends seem intriguing and require further analysis.

Six of the eight glycan biomarkers contain fucose and which may indicate the presence of sialyl Lewis X motifs. Sialyl Lewis X has been documented as a marker for inflammation [57, 58] and its aberrant expression has been implicated in tumor formation and metastasis [59]. Additional structural studies would demonstrate whether the fucose is located on the core or antennae.

6 Concluding remarks

The Glycolyzer software removes the data analysis bottleneck and drastically decreases the time to results for a clinical glycomics study. Each piece of the biomarker discovery pipeline can now be perturbed and evaluated now that the full pipeline is in place and metrics for evaluating the system have been established. Prior to the Glycolyzer, manual calibration and data analysis from a 94-sample set would take several weeks to months. The Glycolyzer can accomplish the same task in a matter of hours. The magnitude of samples processed in this study demonstrates the potential for high-throughput analysis for discovery and validation studies in the future, both for ovarian cancer and other malignancies. In addition, the Glycolyzer allowed us to identify a panel of glycan biomarkers with high sensitivity and specificity that are appropriate for

Table 2. Sialic acid containing or sialic acid free pairs of glycans. The sialic acid free glycans decreased significantly while the sialic acid containing glycans increased significantly

ID	ACN elution fraction	Exact *m/z*	Intensity relative percent change	Intensity Log_{10} control mean	Intensity Log_{10} control CV	Intensity Log_{10} cases mean	Intensity Log_{10} cases CV	Hex	HexNAc	Fucose	NeuAc	Na/H
16	10%	2028.714	−28.9%	2.96	4.2%	2.81	5.2%	6	5	0	0	0
45	40%	2441.870	137.4%	3.97	5.1%	4.35	7.2%	6	5	1	1	0
48	40%	2732.966	33.6%	3.20	4.4%	3.33	4.9%	6	5	1	2	0
49	40%	2754.948	68.7%	3.07	4.0%	3.30	6.6%	6	5	1	2	1
35	20%	2012.719	−20.4%	4.48	4.3%	4.38	5.6%	5	5	1	0	0
39	20%	2325.796	29.4%	2.49	7.4%	2.60	12.6%	5	5	1	1	1
26	20%	1663.581	−37.2%	4.44	4.5%	4.24	7.0%	5	4	0	0	0
10	10%	1663.581	−48.2%	4.30	4.8%	4.01	6.7%	5	4	0	0	0
33	20%	1976.659	56.1%	2.82	8.3%	3.01	15.4%	5	4	0	1	1
36	20%	2122.717	28.7%	2.54	6.9%	2.65	12.1%	5	4	1	1	1
5	10%	1460.502	−48.2%	4.02	6.1%	3.73	8.6%	5	3	0	0	0
43	40%	1727.601	29.0%	3.59	4.3%	3.70	3.4%	5	3	0	1	0

formal validation testing. Although the case subjects in this study were diagnosed with ovarian cancer, it is possible that the biomarkers are noncancer specific and could represent an inflammatory response. This would need to be investigated in subsequent studies.

GlycanFinder algorithms were influenced in part by IgorPro code used for combinatorial glycans model building developed by Brian H. Clowers. Eric D. Dodds helped develop the Fast Fourier Transform algorithm use to transform the raw transient data. Sample selection, age matching, and blinding were performed by Donald A. Barkauskas and David M. Rocke. In addition, insight into the statistical treatment of data was provided by them as well. Anding Fan helped develop an application that produced the raw transient text data files from instrument specific data files for use in the Glycolyzer. We gratefully acknowledge the financial support provided by the National Institute of Health RO1 GM049077. Support was also provided by a gift from the National Ovarian Cancer Coalition (NOCC), Sacramento Chapter (to G.S.L.); a UC Davis Health Systems Research Award (to K.S.L), and an Ovarian Cancer Research Fund (OCRF) Award (to G.S.L). We also acknowledge the Gynecologic Oncology Group Tissue Bank for providing the serum sample sets used in this study.

The authors have declared no conflict of interest.

7 References

[1] Lebrilla, C. B., An, H. J., The prospects of glycan biomarkers for the diagnosis of diseases. *Mol. Biosyst.* 2009, *5*, 17–20.

[2] Packer, N. H., von der Lieth, C. W., Aoki-Kinoshita, K. F., Lebrilla, C. B. et al., Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11–13, 2006). *Proteomics* 2008, *8*, 8–20.

[3] Turnbull, J. E., Field, R. A., Emerging glycomics technologies. *Nat. Chem. Biol.* 2007, *3*, 74–77.

[4] Barkauskas, D. A., An, H. J., Kronewitter, S. R., de Leoz, M. L. et al., Detecting glycan cancer biomarkers in serum samples using MALDI FT-ICR mass spectrometry data. *Bioinformatics* 2009, *25*, 251–257.

[5] de Leoz, M. L. A., An, H. J., Kronewitter, S., Kim, J. et al., Glycomic approach for potential biomarkers on prostate cancer: profiling of N-linked glycans in human sera and pRNS cell lines. *Disease Markers* 2008, *25*, 243–258.

[6] Alley, W. R., Jr., Madera, M., Mechref, Y., Novotny, M. V., Chip-based reversed-phase liquid chromatography-mass spectrometry of permethylated N-linked glycans: a potential methodology for cancer-biomarker discovery. *Anal. Chem.* 2010, *82*, 5095–5106.

[7] An, H. J., Miyamoto, S., Lancaster, K. S., Kirmiz, C. et al., Profiling of glycans in serum for the discovery of potential biomarkers for ovarian cancer. *J. Proteome Res.* 2006, *5*, 1626–1635.

[8] Bones, J., Mittermayr, S., O'Donoghue, N., Guttman, A. et al., Ultra performance liquid chromatographic profiling of serum N-glycans for fast and efficient identification of cancer associated alterations in glycosylation. *Anal. Chem.* 2010, *82*, 10208–10215.

[9] An, H. J., Kronewitter, S. R., de Leoz, M. L. A., Lebrilla, C. B., Glycomics and disease markers. *Curr. Opin. Chem. Biol.* 2009, *13*, 601–607.

[10] Lebrilla, C. B., An, H. J., The prospects of glycan biomarkers for the diagnosis of diseases. *Mol. Biosyst.* 2009, *5*, 17–20.

[11] Li, B., An, H. J., Kirmiz, C., Lebrilla, C. B. et al., Glycoproteomic analyses of ovarian cancer cell lines and sera from ovarian cancer patients show distinct glycosylation changes in individual proteins. *J. Proteome Res.* 2008, *7*, 3776–3788.

[12] Meany, D. L., Zhang, Z., Sokoll, L. J., Zhang, H. et al., Glycoproteomics for prostate cancer detection: changes in serum PSA glycosylation patterns. *J. Proteome Res.* 2008, *8*, 613–619.

[13] Saldova, R., Royle, L., Radcliffe, C. M., Abd Hamid, U. M. et al., Ovarian cancer is associated with changes in glycosylation in both acute-phase proteins and IgG. *Glycobiology* 2007, *17*, 1344–1356.

[14] Cooper, C. A., Gasteiger, E., Packer, N. H., GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* 2001, *1*, 340–349.

[15] Cooper, C. A., Harrison, M. J., Wilkins, M. R., Packer, N. H., GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.* 2001, *29*, 332–335.

[16] Cooper, C. A., Joshi, H. J., Harrison, M. J., Wilkins, M. R. et al., GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res.* 2003, *31*, 511–513.

[17] Lo, A., Bunsmann, P., Bohne, A., Lo, A. et al., SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res.* 2002, *30*, 405–408.

[18] Cooper, C. A., Wilkins, M. R., Williams, K. L., Packer, N. H., BOLD—a biological O-linked glycan database. *Electrophoresis* 1999, *20*, 3589–3598.

[19] Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K. et al., KEGG as a glycome informatics resource. *Glycobiology* 2006, *16*, 63R–70R.

[20] von der Lieth, C. W., Freire, A. A., Blank, D., Campbell, M. P. et al., EUROCarbDB: an open-access platform for glycoinformatics. *Glycobiology* 2011, *21*, 493–502.

[21] Goldberg, D., Sutton-Smith, M., Paulson, J., Dell, A., Automatic annotation of matrix-assisted laser desorption/ionization N-glycan spectra. *Proteomics* 2005, *5*, 865–875.

[22] Kronewitter, S. R., An, H. J., de Leoz, M. L., Lebrilla, C. B. et al., The development of retrosynthetic glycan libraries to profile and classify the human serum N-linked glycome. *Proteomics* 2009, *9*, 2986–2994.

[23] Ethier, M., Figeys, D., Perreault, H., N-glycosylation analysis using the StrOligo algorithm. *Methods Mol. Biol.* 2006, *328*, 187–197.

[24] Lapadula, A. J., Hatcher, P. J., Hanneman, A. J., Ashline, D. J. et al., Congruent strategies for carbohydrate sequencing. 3. OSCAR: an algorithm for assigning oligosaccharide topology from MSn data. *Anal. Chem.* 2005, *77*, 6271–6279.

[25] Zhang, H., Singh, S., Reinhold, V. N., Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. *Anal. Chem.* 2005, *77*, 6263–6270.

[26] Maass, K., Ranzinger, R., Geyer, H., Lieth, C.-W. v. d. et al., "Glyco-Peakfinder"—*denovo* composition analysis of glycoconjugates. *Proteomics* 2007, *7*, 4435–4444.

[27] Lohmann, K. K., Lieth, C.-W. v. d., GLYCO-FRAGMENT: a web tool to support the interpretation of mass spectra of complex carbohydrates. *Proteomics* 2003, *3*, 2028–2035.

[28] Lohmann, K. K., von der Lieth, C.-W., GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res.* 2004, *32*, W261–W266.

[29] Ceroni, A., Maass, K., Geyer, H., Geyer, R. et al., GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.* 2008, *7*, 1650–1659.

[30] Vakhrushev, S. Y., Dadimov, D., Peter-Katalinic, J., Software platform for high-throughput glycomics. *Anal. Chem.* 2009, *81*, 3252–3260.

[31] Moore, R. G., Jabre-Raughley, M., Brown, A. K., Robison, K. M. et al., Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass. *Am. J. Obstet. Gynecol.* 2010, *203*, 228.e1–228.e6.

[32] Leiserowitz, G. S., Lebrilla, C., Miyamoto, S., An, H. J. et al., Glycomics analysis of serum: a potential new biomarker for ovarian cancer? *Int. J. Gynecol. Cancer* 2007, *18*, 470–475.

[33] Kronewitter, S. R., de Leoz, M. L., Peacock, K. S., McBride, K. R. et al., Human serum processing and analysis methods for rapid and reproducible N-glycan mass profiling. *J. Proteome Res.* 2010, *9*, 4952–4959.

[34] Shi, S. D. H., Drader, J. J., Freitas, M. A., Hendrickson, C. L. et al., Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry. *Int. J. Mass Spectrom.* 2000, *195–196*, 591–598.

[35] Zhang, L. K., Rempel, D., Pramanik, B. N., Gross, M. L., Accurate mass measurements by Fourier transform mass spectrometry. *Mass Spectrom. Rev.* 2005, *24*, 286–309.

[36] Francl, T. J., Sherman, M. G., Hunter, R. L., Locke, M. J. et al., Experimental determination of the effects of space charge on ion cyclotron resonance frequencies. *Int. J. Mass Spectrom.* 1983, *54*, 189–199.

[37] Marshall, A. G., Comisarow, M. B., Parisod, G., Theory of Fourier-transform ion-cyclotron resonance mass spectroscopy-lii. 1. Relaxation and spectral-line shape in Fourier-transform ion resonance spectroscopy. *J. Chem. Phys.* 1979, *71*, 4434–4444.

[38] Wehofsky, M., Hoffmann, R., Hubert, M., Spengler, B., Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substance-class specific analysis of complex samples. *Eur. J. Mass Spectrom.* 2001, *7*, 39–46.

[39] Maleknia, S. D., Downard, K. M., Charge ratio analysis method to interpret high resolution electrospray Fourier transform—ion cyclotron resonance mass spectra. *Int. J. Mass Spectrom.* 2005, *246*, 1–9.

[40] Zhang, X. A., Asara, J. M., Adamec, J., Ouzzani, M. et al., Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics* 2005, *21*, 4054–4059.

[41] Kaur, P., O'Connor, P. B., Algorithms for automatic interpretation of high resolution mass spectra. *J. Am. Soc. Mass Spectrom.* 2006, *17*, 459–468.

[42] Tabb, D. L., Shah, M. B., Strader, M. B., Connelly, H. M. et al., Determination of peptide and protein ion charge states by Fourier transformation of isotope-resolved mass spectra. *J. Am. Soc. Mass Spectrom.* 2006, *17*, 903–915.

[43] Horn, D. M., Zubarev, R. A., McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 2000, *11*, 320–332.

[44] Du, P. C., Angeletti, R. H., Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Anal. Chem.* 2006, *78*, 3385–3392.

[45] Senko, M. W., Beu, S. C., McLafferty, F. W., Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 1995, *6*, 229–233.

[46] An, H. J., Tillinghast, J. S., Woodruff, D. L., Rocke, D. M. et al., A new computer program (GlycoX) to determine simultaneously the glycosylation sites and oligosaccharide heterogeneity of glycoproteins. *J. Proteome Res.* 2006, *5*, 2800–2808.

[47] Chu, C. S., Ninonuevo, M. R., Clowers, B. H., Perkins, P. D. et al., Profile of native N-linked glycan structures from human serum using high performance liquid chromatography on a microfluidic chip and time-of-flight mass spectrometry. *Proteomics* 2009, *9*, 1939–1951.

[48] Barkauskas, D. A., An, H. J., Kronewitter, S. R., de Leoz, M. L. et al., Detecting glycan cancer biomarkers in serum samples using MALDI FT-ICR mass spectrometry data. *Bioinformatics* 2009, *25*, 251–257.

[49] Zhu, C. S., Pinsky, P. F., Cramer, D. W., Ransohoff, D. F. et al., A framework for evaluating biomarkers for early detection: validation of biomarker panels for ovarian cancer. *Cancer Prev. Res. (Phila.)* 2011, *4*, 375–383.

[50] Cramer, D. W., Bast, R. C., Jr., Berg, C. D., Diamandis, E. P. et al., Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. *Cancer Prev. Res. (Phila.)* 2011, *4*, 365–374.

[51] Box, G. E. P., Muller, M. E., A note on the generation of random normal deviates. *Ann. Math. Statist.* 1958, *29*, 610–611.

[52] Gornik, O., Royle, L., Harvey, D. J., Radcliffe, C. M. et al., Changes of serum glycans during sepsis and acute pancreatitis. *Glycobiology* 2007, *17*, 1321–1332.

[53] Kim, Y. G., Jeong, H. J., Jang, K. S., Yang, Y. H. et al., Rapid and high-throughput analysis of N-glycans from ovarian cancer serum using a 96-well plate platform. *Anal. Biochem.* 2009, *391*, 151–153.

[54] Dwek, M. V., Ross, H. A., Leathem, A. J., Proteome and glycosylation mapping identifies post-translational modifications associated with aggressive breast cancer. *Proteomics* 2001, *1*, 756–762.

[55] Kyselova, Z., Mechref, Y., Al Bataineh, M. M., Dobrolecki, L. E. et al., Alterations in the serum glycome due to metastatic prostate cancer. *J. Proteome Res.* 2007, *6*, 1822–1832.

[56] Alley, W. R., Jr., Novotny, M. V., Glycomic analysis of sialic acid linkages in glycans derived from blood serum glycoproteins. *J. Proteome Res.* 2010, *9*, 3062–3072.

[57] De Graaf, T. W., Van der Stelt, M. E., Anbergen, M. G., van Dijk, W., Inflammation-induced expression of sialyl Lewis X-containing glycan structures on alpha 1-acid glycoprotein (orosomucoid) in human sera. *J. Exp. Med.* 1993, *177*, 657–666.

[58] Brinkman-van der Linden, E. C. M., de Haan, P. F., Havenaar, E. C., van Dijk, W., Inflammation-induced expression of sialyl Lewisx is not restricted to α1-acid glycoprotein but also occurs to a lesser extent on α1-antichymotrypsin and haptoglobin. *Glycoconjugate J.* 1998, *15*, 177–182.

[59] Ohyama, C., Tsuboi, S., Fukuda, M., Dual roles of sialyl Lewis X oligosaccharides in tumor metastasis and rejection by natural killer cells. *EMBO J.* 1999, *18*, 1516–1525.

