analytical chemistry

Article

pubs.acs.org/ac

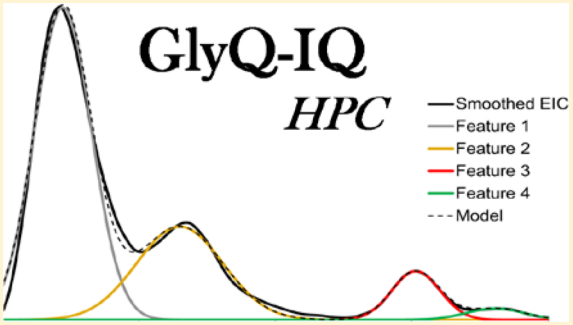
GlyQ-IQ: Glycomics Quintavariate-Informed Quantification with High-Performance Computing and GlycoGrid 4D Visualization

Scott R. Kronewitter, Gordon W. Slysz, Ioan Marginean, Clay D. Hagler, Brian L. LaMarche, Rui Zhao, Myanna Y. Harris, Matthew E. Monroe, Christina A. Polyukh, Kevin L. Crowell, Thomas L. Fillmore, Timothy S. Carlson, David G. Camp II, Ronald J. Moore, Samuel H. Payne, Gordon A. Anderson, and Richard D. Smith*

Biological Sciences Division, Pacific Northwest National Laboratory, P.O. Box 999, Richland, Washington 99352, United States

Supporting Information

ABSTRACT: Glycomics quintavariate-informed quantification (GlyQ-IQ) is a biologically guided glycomics analysis tool for identifying N-glycans in liquid chromatography−mass spectrometry (LC−MS) data. Glycomics LC−MS data sets have convoluted extracted ion chromatograms that are challenging to deconvolve with existing software tools. LC deconvolution into constituent pieces is critical in glycomics data sets because chromatographic peaks correspond to different intact glycan structural isomers. The biological targeted analysis approach offers several key advantages to traditional LC−MS data processing. *A priori* glycan information about the individual target's elemental composition allows for improved sensitivity by utilizing the exact isotope profile information to focus chromatogram generation and LC peak fitting on the isotopic species having the highest intensity. Glycan target annotation utilizes glycan family relationships and in source fragmentation in addition to high specificity feature LC−MS detection to improve the specificity of the analysis. The GlyQ-IQ software was developed in this work and evaluated in the context of profiling the N-glycan compositions from human serum LC−MS data sets. A case study is presented to demonstrate how GlyQ-IQ identifies and removes confounding chromatographic peaks from high mannose glycan isomers from human blood serum. In addition, GlyQ-IQ was used to generate a broad human serum N-glycan profile from a high resolution nanoelectrospray-liquid chromatography−tandem mass spectrometry (nESI-LC−MS/MS) data set. A total of 156 glycan compositions and 640 glycan isomers were detected from a single sample. Over 99% of the GlyQ-IQ glycan-feature assignments passed manual validation and are backed with high-resolution mass spectra.



Glycosylation strongly affects how proteins are folded and maintain the proper structure required for interacting with other proteins and their environment. The fusion of proteomics, glycomics, glycoproteomics, and mass spectrometry has the capability of completely characterizing glycoproteins and determine which glycans are attached to which glycosylation site on the protein backbone.[1] Including glycan profiling with bottom up glycoproteomics analyses technologies has been shown to decrease the false positive identification rate by bounding the relatively large amount of possible glycans to experimental evidence.[2]

N-Glycosylation is a widespread posttranslational modification of proteins commonly found covalently attached to asparagine. N-Glycans are complex branched biopolymers of monosaccharides constructed by glycosidase and glycosyltransferase enzymatic reactions in the Golgi and endoplasmic reticulum (ER) cellular organelles. The nontemplate driven process produces families of glycans that relate to each other by one enzymatic step and resulting in glycans related to each other by a difference of one monosaccharide.[3,4] As a result, end-product glycan mixtures contain multiple isomer forms based on different monosaccharide connectivity. Determining the monosaccharide composition of glycans and how many isomers are present is an important step toward in-depth structural characterization analysis. Characterizing glycan structures provides a basis for, among other things, insights on structure−function relationships present in biological systems.

Glycan isomer profiling can be accomplished by coupling liquid chromatography with mass spectrometry (LC−MS). Glycomics annotation of LC−MS data involves two critical aspects of analysis: feature detection and glycan assignment. LC−MS features are commonly defined by m/z peak intensities (e.g., for their isotope profile) in the mass spectra dimension and chromatographic elution profile from their extracted ion chromatogram (EIC). For simplicity, the isotopic envelope for a species can be collapsed or "deisotoped" to a single value. In the case of glycan annotation, exact monoisotopic mass from

Received: January 13, 2014
Accepted: May 31, 2014
Published: May 31, 2014

each feature can then be matched to glycan masses calculated theoretically or via experimentally generated libraries to develop a composition profile that attempts to describe all of the monosaccharides present in the glycan mixture. Since many glycans have isomer structures, the structures can be often separated chromatographically and retention times for each isomer assigned.

Several informatics tools have been developed that facilitate one or multiple aspects of the glycan annotation process and vary in the analytical features included, type of mass spectrometry data required, and overall sensitivity and specificity of the results.[5−8] When chromatographic information is available as in LC−MS experiments, LC−MS features are commonly detected by assembling the monoisotopic masses derived from each mass spectrum and then annotated with the glycomics tool.[6,7,9−11] In general, glycomics data is challenged with partially resolved chromatographic peaks (common for glycan isomers) and convoluted isotopic profiles. Although efforts to deconvoluted overlapping isotope distributions have been previously explored (NITPICK,[12] Glycolyzer,[6] MultiGlycan[9]), they generally work solely in the mass spectra isotope profile space and are limited to averagine[13]/averagose[14] based isotope model approximations. Several advances have been made in chromatographic processing for LC−MS/MS proteomics,[15] yet the algorithms have not been applied to glycomic challenges where closely related isomers are common. In addition to the advanced chromatographic deconvolution presented here, this algorithm leverages exact chemical formula based isotopic profiles and insource fragmentation information to support feature detection and annotation.

We have developed a new targeted software application, Glycomic Quintavariate Informed Quantification (GlyQ-IQ) to analyze and annotate enzymatically released N-glycan LC−MS data sets with high-performance computing. GlyQ-IQ includes all of the signal processing and spectra averaging, peak detection, targeted deisotoping, feature finding and glycan composition annotation required to convert raw data files into annotated results. The Informed Quantification (IQ) framework includes robust liquid chromatography processing that includes LC peak modeling and robust correlations between extracted ion chromatograms to increase annotation confidence. In addition, the robust correlations are used to identify insource fragmentation and increase the specificity of the analysis. Insource fragmentation is identified when glycans and glycan fragments chromatographic elution profiles correlate.

This work describes the methodology behind the GlyQ-IQ algorithms and the use for N-glycan profiling of N-glycans enzymatically cleaved from glycoproteins. A case study is also included involving characterization of high mannose glycans in human serum and how insource fragmentation can be used to differentiate intact glycans from insource fragments and confirm glycan composition assignments. Problems related to interfering peaks in EICs, correlation coefficients involving nonideal data, and imparting glycan biology into the annotation process are discussed. This work is broken into two parts focusing on data processing in the mass spectra and chromatography dimensions. The software is available online at http://omics.pnl.gov/software.

METHODS

Materials. All materials were of high purity, and the same pooled human blood serum (male, blood type AB, not heat inactivated) was used for all analyses (Sigma-Aldrich, St. Louis, MO). All other chemicals and procedures used were consistent with those previously described for 100 μL of serum to produce reduced, nonderivatized glycans.[16] Briefly, N-glycans were enzymatically cleaved from serum glycoproteins with PNGase F in acidified reaction conditions (pH 5.5). The enzymatic cleavage was catalyzed with a microwave reactor (CEM Discover, Matthews, NC). The released glycans were purified by 80% ethanol precipitation/centrifugation and reduced with sodium borohydride before further purification and desalting. A dual cartridge (C8, GCC) automated solid phase extraction and desalting was performed using a Gilson GX-274 ASPEC liquid handler (Middleton, WI) with customized algorithms optimized for this process. The aqueous glycan samples were first passed through a C8 cartridge to remove any residual peptides or lipids and the flow-through was directly loaded onto a graphite cartridge for salt and small molecule removal.

Data Acquisition. Data was acquired using a constant flow high-performance liquid chromatography (HPLC) system (Agilent Technologies, Santa Clara, CA) coupled to a Velos Pro Orbitrap Mass spectrometer with 100K mass resolution. The LC column was packed with 3 μm diameter Hypercarb porous graphitized carbon particles (Thermo Scientific) and had a length of 70 cm long × 360 μm o.d. × 75 μm i.d. (Polymicro Technologies Inc., Phoenix, AZ) and a 1 cm sol−gel frit for media retention.[17] Human blood serum N-glycans previously identified with GlyQ-IQ software were used to populate a mass list for targeted fragmentation. High-resolution (60k mass resolution) collision induced dissociation (CID, energy 30 for 10 ms) and higher energy collision dissociation (HCD, energy 20 for 0.1 ms) were used for fragmentation.

The atomicity of the glycan targets makes the IQ based algorithms attractive for parallel computing since they do not depend on each other for the bulk of the analysis. Each target requires a large amount of processing because of the multiplicative effect of target charge states, isomers, possible insource fragments, and insource fragment charge states. For example, if 4 charge states and 4 fragment ions are considered for a single target, 512 additional subtargets need to be discovered and processed. This multiplicative effect on targets is illustrated in Figure 1. For each target, EICs need to be determined, theoretical isotope profiles generated and compared to mass spectra, and LC curves modeled and fit.

When thousands of targets are searched, millions of subtargets need to be correlated and analyzed. The GlyQ-IQ software was deployed on a Windows 2012 R2 based HPC cluster with a head node and 1 504 compute cores. The HPC acceleration reduced the runtime by 99.92% (550× faster) when compared to executing the same job on a single core processor. Additional HPC Compute Cluster and Microsoft Windows Azure Cloud deployment information was included in the Supplementary Text S1 and Supplementary Figure 1 in the Supporting Information.

Mass Spectrometric Data Processing. *Targeted Deisotoping.* Isotopic profiles were evaluated by comparing a theoretically generated isotopic profile for exact targeted elemental compositions with the experimental data. The experimental isotopic envelope (or profile) abundances are populated by extracting all the masses contained in the theoretical isotopic envelope from the spectra. A user determined number of spectra (e.g., 9 spectra in this study, see below) surrounding the EIC peak apex are averaged together to increase the ion statistics of the isotopic envelope

dx.doi.org/10.1021/ac501492f | Anal. Chem. 2014, 86, 6268−6276

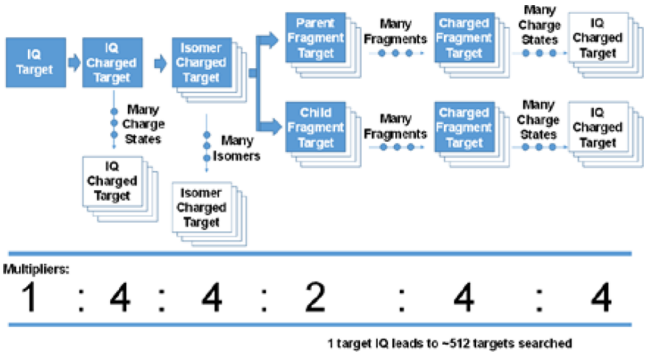


Figure 1. Glycan compositions detected in LC−MS experiments can be in multiple forms that need to be identified in the data sets (isomers, charge states). The multiple forms need to be cross correlated and searched for consistency. The multipliers used in this example represent 1 glycan target containing 4 charge states, 4 isomers, searching for fragments in both directions (2, greater or smaller by one monosaccharide), 4 monosaccharide differences to search for, and 4 charge states possible for the fragment ions.

and signal-to-noise ratio (e.g., noise decrease by a factor of ~3, square root of $N = 9$).

A detectable area filter is applied to the experimental isotopic envelope so that experimental profiles which are missing critical abundant ions are removed. A key feature of this filter is it improves detection of high quality features independent of mass range as shown in Supplementary Figure 2 in the Supporting Information which contains examples ranging from 1000 to 5500 Da. Removing mass dependency bias from the algorithm improves the quality of the result and allows for an even response when an isotope fit score cutoff is applied. The total isotopic profile area is calculated from the theoretical isotopic profile through integration by summing the abundances of the individual isotopic peaks. The critical required ions are determined by including the most abundant ions contributing to 75% of the total isotope area. Experimental profiles missing the critical ions are rejected. Additional ions beyond the critical ions are included if present in the experimental data.

Isotope Profile Fit Scoring. All charge states that produce ions within the m/z range covered by the spectra are considered independently. For the ions of one charge state, the intensities are extracted and the fit scores are based on a modified chi squared test. Modifications to the fit score calculations used in Decon2LS[18] and THRASH[19] were incorporated here to account for the number of ions detected in the isotope profile. The fit score used in GlyQ-IQ is divided by the number of ions detected in the profile to decrease the score biases against high masses with several isotopes. The equation is included as eq 1 below where E is the experimental data, T is the theoretical data, and n and N are the number of observed isotopes.

$$\text{fit score} = \frac{\sum_{i=0}^{n} 100(E_i - T_i)^2}{N \sum_{i=0}^{n} T_i^2} \quad (1)$$

Additional details explaining the effect of the modified equation are included in the Supplementary Text S2 and Supplementary Figures 3 and 4 in the Supporting Information.

A strict 0.10 fit score cutoff was chosen to increase the specificity of feature detection and resulted in fewer features that need to be removed by manual inspection; 99% of the glycans assigned with GlyQ-IQ passed the manual inspection process and include ions with a mass that fall in the range from 912 up to 4500 Da. A total of 689 glycan isomer peaks from this data set were detected by GlyQ-IQ (including 45 insource fragmentation fragments see below) and 685 passed manual validation. The sensitivity was also increased as well since large masses with formerly relatively high (i.e., poor) fit scores were rescored to lower (i.e., good) values that fall in the acceptable range (see Supplementary Figure 4 in the Supporting Information). The additional larger masses detected pass manual inspection and are included in the 99% success rate calculation. However, masses greater than 4500 Da are still detectable but the sucesss rate is decreased because the monoisotopic mass is often no longer present and the 1 Da penalty peaks have less discriminiation (as noted below).

Increased Mass Spectra Data Quality. Mass spectra averaging provided by the C# MSFileReader DLL (Thermo Scientific, San Jose, CA) was used to increase sensitivity and improve the ion statistics in the glycan isotopic profiles. Provided the noise level is relatively constant and randomly distributed within a few mass spectra scans during a LC−MS run (corresponding to a fraction of a second up to a few seconds of time depending on detector acquisition rate and amount of MS/MS scans collected), signal averaging can decrease the effective noise level of the measurement by the square of the number of spectra and improve isotope fit scores. In addition, spectral averaging incorporates isotopic profile information from several spectra into the observed profile which improves the accuracy of intensity distribution in the profile. This is especially important for ions with low abundance that have insufficient ion statistics to produce a well-defined profile.

Implementing raw spectra averaging shows a striking advantage in sensitivity (at constant specificity) when applied to our human serum data as shown in Supplementary Figure 5 in the Supporting Information. For this data set, the optimal number of spectra averaged is 9 and was selected by the number that produced the lowest average fit score and highest number of manually verified glycan annotations. Implementing optimal signal averagine, the number of glycan features detected increased by 56%, the average fit scores decreased by 15%, and the GlyQ-IQ analysis runtime increased by 38%.

1-Dalton Errors. One common problem with deisotoping mass spectra is false annotations caused by 1 Da (Da) mass errors in monoisotopic mass assignments. The framework implemented here attempts to mitigate this effect by searching for an ion 1 Da lower than the monoisotopic mass, populating its abundance, and use it as a penalty value in the fit score calculation. Since the fit score equation used here is based on the sum of the squared errors, any ion present will contribute a significant difference (as compared to zero abundance in the model) and increase the fit score. If the intensity of the precluding peak is large, the increase in fit score caused by the penalty ion error is typically sufficient to increase the score beyond the 0.10 cutoff used here. An example is shown in Figure 2A,B where two ions partially coelute (compound A, monoisotopic m/z 881.996, fifth isotope at 883.334 and Compound B (target), monoisotopic m/z 883.663) and produce a multiplexed isotope pattern. Ions that have higher than the fit score cutoff due to the penalty ion contribution are triaged for advanced chromatographic analysis. This is particularly beneficial for the relatively low mass range (500−4500 Da) and less helpful as the mass increases due to the

dx.doi.org/10.1021/ac501492f | Anal. Chem. 2014, 86, 6268−6276

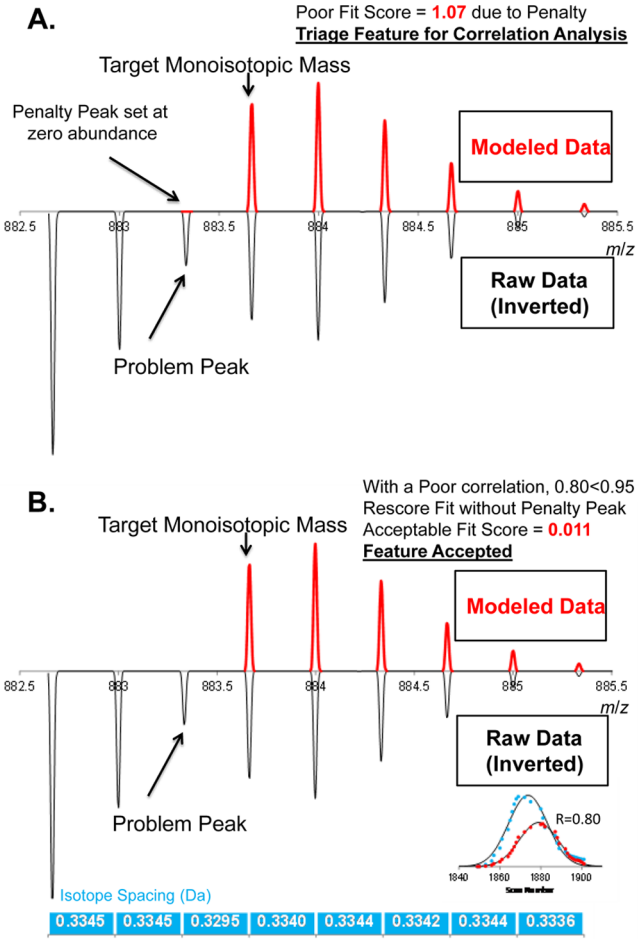


Figure 2. Example depicting the penalization of the isotope fit score calculations when an unrelated peak lower in mass (1 Da, 0.33 m/z at 3+ charge) is observed in the mass spectra. (A) Fit score calculated when modeled data is fit to the observed data with 1 Da penalty assigned. The high fit score greater than 0.1 cutoff triages further chromatographic analysis. (B) EICs from the most abundant isotope m/z and the penalty peak m/z are modeled, fit, and correlated. Failed EIC correlations trigger a rescoring of the isotope profile while excluding the penalty peak because it was deemed not part of the targeted distribution. The large score decrease from 1.07 to 0.011 leads to a correct assignment of the target. The mass differences between the 3+ charge state isotopes are shown in blue at the bottom.

decreasing contribution of the monoisotopic mass intensity to the overall distribution.

Triaged isotopic profiles provide at least two possible solutions to dealing with this problem and finding the correct answer for the experimental data: (1) the simplest answer (with the least amount of assumptions) is the raw data fits a single compound (compound A, 1 or more Da lower in mass) with a different elemental composition, or (2) the distribution is a convolution of two near perfectly overlapping isotope profiles, but offset in mass, containing 2 or more multiplexed ions of target compound B and lower mass compound A. Implementing chromatographic information can unambiguously separate the problem cases by using EIC profiles and correlations. If the extracted ion chromatogram of compounds A and B have peaks within the point range of the most abundant isotope and correlate greater than our 0.95 cutoff (see below), case 1 is deemed correct and the desired target (compound B) is considered not present. If the EIC do not coelute ($R < 0.95$) as in case 2, the target ion is discrete and rescored and tested for acceptance.

Chromatogram Processing. *EIC Generation.* GlyQ-IQ is a chromatography centric data analysis tool which involves significant EIC processing. Since each target has an empirical formula and subsequent theoretical isotopic distribution, the most abundant isotope can be chosen as a target for chromatogram generation. EIC widths are calculated based on the data using the following method. The largest peak (consistent with a target isotopic profile) in the default EIC window is chosen and its full-width at half-maximum (fwhm) is calculated in the mass domain. This width is then divided in half (or a user defined divisor) and the data driven ppm mass tolerance is used to center the EIC extraction mass window for subsequent target processing.

Smoothing. EICs often contain significant amounts of variation, particularly for lower intensity species due to the stochastic limitations, microinstabilities in the electrospray process, and other sources of measurement noise. To eliminate artifact peaks and focus the algorithms on chromatographic peaks of interest, a digital Savitzky-Golay (SG) smoothing filter (degree 2) is applied. Windowed extracted ion peak chromatograms are typically used in these algorithms and are bounded by a scan range of interest. However, since the SG filter is a moving average based filter, edge effects are accounted for by buffering all EICs with extra points (2 times the number of smoothing window points) from the full range peak chromatogram. A 9 point SG smoothing filter window is applied to Man_8 from human blood serum and plotted in Supplementary Figure 6 in the Supporting Information that shows that the smoothed data sufficiently represents the experimental chromatograms and does not introduce new artifact peaks. SG smoothing was chosen to better distinguish the peak centroid, peak width, and peak concavity required for robust correlations, and such buffered EICs preserve the data quality and have faster execution time than full range EICs.

Peak Detection. Chromatographic peaks are detected and filtered from the smoothed extracted peak chromatograms to reduce chromatographic noise. Nevertheless noise and artifact peaks still need to be identified and filtered out of the smoothed data, including peaks and shoulders with too few data points.

A set of candidate peaks is established via three point differential peak detection and then the centroids were determined to find the apex. The minima of each peak is also calculated and used to determine how many points constitute half the peak shape (center point−minima point, etc.). The half peak shape is used instead of a full peak shape because it allows the better detection of partially resolved species. If the peak has less than a minimum amount of points (2 in this case) on either side, it is removed.

Chromatographic Artifact Peak Removal. A layer of complexity is observed on postsmoothed data because single point peaks (a peak consisting of 3 points, e.g., 2 baseline points surrounding 1 nonzero point) can resemble broader peaks once smoothed. Five points are often considered the minimum needed for defining a peak apex and width reliably.[20] However, this criterion does not directly map to SG smoothed data because the number of points defining a peak increases as the smoothing window point number increases. Examples of the smoothing effect broadening simple peaks are presented in Supplementary Figure 7 in the Supporting Information where peaks defined with one or five points are smoothed with

dx.doi.org/10.1021/ac501492f | Anal. Chem. 2014, 86, 6268−6276

different SG windows (the five points peak series is inverted). The relationship between an acceptable number of points and the SG windows can be calculated by plotting the number of points across a raw peak versus the number of points across a smoothed peak (at various smoothing levels). This is shown to be linear in Supplementary Figure 8 in the Supporting Information. A minimum of 4 points per side (>8 points for full peak) is set for a 9 point SG smooth. This is sufficient to remove three point peak equivalent peaks from the raw data and improve confidence in peak assignments by removing artifacts caused by single point peaks. Using this equation presented in Supplementary Figure 8 in the Supporting Information, the peak detection parameters are automatically set based on the smoothing parameter supplied by the user.

Partially Resolved Chromatographic Peaks. An asymmetric peak shifting parameter is included for better shoulder detection of partially overlapping peaks. This allows up to two points to be shifted from one side to the other for the case where shoulder peaks are well-defined on the nonconvolved side (min number of points + 2) and have less than the number of points on the convolved side (min number of points − 2). Two points were chosen to maintain the 8 points across the peaks at the smoothed level (9 point SG window).

EIC Peak Shape Fitting. After peaks with insufficient data points are removed, each peak is modeled with a Gaussian function and the models were compared to the data. This helps determine the characteristic curve associated with the peak and allows integration to estimate relative abundance. A Gaussian peak shape is fit to each peak using a Levenberg−Marquardt algorithm (ALGLIB[21]) using a 0.85 R^2 coefficient of determination cutoff. As shown in Supplementary Figure 9 in the Supporting Information, the coefficient of determination cutoff value was chosen by fitting a Pareto distribution to the distribution of coefficients of determination and setting the cutoff to the 99th percentile (0.85).

For peaks with noisy or tailing bottoms that do not fit Gaussian line shapes, points are removed from the bottom until the optimization algorithm converges. This helps remove low signal-to-noise peaks and peak tailing from the fit while maintaining the peak centroid which is the most determinate factor in the correlation value calculation and ultimately improving correlation score robustness. Although other peak shapes have been used in the past to model chromatographic peak shapes,[22] defining the peak tops and apex is the most important parameter for correlations and insource fragmentation. If chromatographic peak tailing is present in the elution profiles, other functions could be implemented such as expanded Gaussian models.[23] For this study, each peak in the EIC is deconvoluted into its constituents by Gaussian model fitting. A sample deconvolution of the Man_8 EIC from human blood serum is shown in Figure 3.

When a sufficient fit is achieved for a single peak, the coefficients for the corresponding Gaussian model are obtained and used for integrating the area under the curve. The quantification area is calculated numerically using the trapezoidal rule over 100 points. The fit coefficients are also used for calculating a new interpolated peak shape model optimal for peak interchromatogram correlations.

Chromatogram Correlations. *Pierson Product-Moment Correlation (R).* The robustness of correlation analysis was increased by mapping correlation values to the same scale so that a constant cutoff could be used regardless of peak height, resolution, and offsets. Addressing for different sampling point

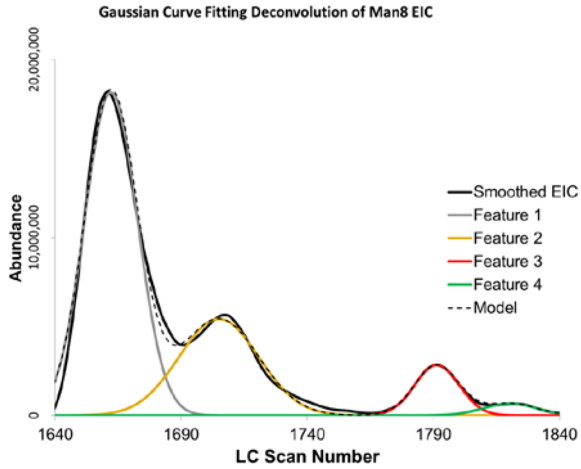


Figure 3. Deconvolution of Man_8 EIC into Gaussian peak shapes. Summing the individual deconvoluted features provides the model dotted line that is consistent with the smoothed EIC. Observing multiple isobaric chromatographically separated compositions with different elution times correspond to different chemical isomeric structures.

densities across the LC peaks was addressed by modeling the peak shapes and interpolating a constant number of points. Additional information is included in the Supplementary Text S3 and Supplementary Figure 11 in the Supporting Information.

Insource Fragmentation Determination and Future Target Detection. Insource fragmentation occurs due to labile covalent glycosidic bonds between monosaccharides that are broken during ion formation or transport through the interface prior to the m/z analysis. Generally optimizing electrospray ionization source conditions for decreased fragmentation is achieved by decreasing the energy imparted to the ions during ionization and transport into lower pressure regions of the instrument.[24] However, decreasing the energy imparted to the ion−solvent clusters decreases desolvation and results in lower signal intensity. Although nonderivatized glycans are detected in the native form (with the exclusions of the reducing-end modifications) they are generally more susceptible to insource fragmentation than their derivatized counterparts (e.g., permethylated).[25]

One type of insource fragments considered here consists of the loss of one monosaccharide unit from the parent. The loss of a monosaccharide indicates the peak target compound is a glycan. This is detectable by correlating EICs for the parent and the fragment since both will exhibit the same chromatographic elution profile. Comparisons are made between the parent and fragment across all allowable detectable charge states for all allowable monosaccharides. For example, this allows for a 3^+ charged parent to be correlated with a 2^+ charged fragment. The correlation coefficients were calculated for the human blood serum glycan targets (including isomers) and their insource fragments (including larger compositions) and the counts are presented as a histogram in Supplementary Figure 12 in the Supporting Information. Correlation coefficients with values greater than 0.95 are used to distinguish insource fragmentation. The 0.95 cutoff was determined based on the fitting a Pareto distribution to the correlation coefficient trace in Supplementary Figure 12 in the Supporting Information and establishing the cutoff at the 99th percentile.

dx.doi.org/10.1021/ac501492f | *Anal. Chem.* 2014, 86, 6268−6276

In addition to the parent-fragment relationships, parent-to-larger-composition relationships are also considered which indicate that the target compound is a glycan fragment. In this case, glycans with an extra monosaccharide beyond the target's composition are searched for and correlated. Correlating hits become future targets in subsequent analysis and are added to the glycan library. Discrete glycan isomers are identified in the case when no other larger compositions are detected.

When insource fragmentation is detected, it confirms that the compositional assignment assigned by accurate mass is correct and the peak is glycan related. All 86 glycans features detected with insource fragmentation were consistent with its assigned monosaccharide composition. Insource fragmentation detection helps confirm feature detection and largely eliminates the need for manual inspection of the data. The quintavariate glycans (insource fragmentation detection + the other 4 pieces of information) provided a 100% acceptance rate for manual feature inspection. In addition, the insource fragmentation removes the identified glycan fragments from the accepted list thus decreasing the false positive glycan hits reported.

Charge State Correlation. Chromatographic centric algorithms process EIC for each charge state separately and common chromatographic peaks apexes can shift slightly. Modeled chromatographic peaks from neighboring charge states are correlated to determine consistency. A histogram of R correlation values from charge state correlations (a monoisotopic mass at different charge states) is presented in Supplementary Figure 13 in the Supporting Information. A cutoff value ($R > 0.95$) was determined based on Pareto curve fitting and a 99th percentile cutoff. Once correlated, coeluting charge states are combined by averaging the monoisotopic masses detected and summing the abundances. Summing the abundance values from each charge state provides an aggregative abundance value for each glycan species.

Incorporating Glycan Characteristics. *Glycan Family Relationships.* Glycans are created enzymatically in the Golgi and endoplasmic reticulum by a complex process involving the addition and subtraction of monosaccharides. Consequently glycans are often detected in families that differ by one or more monosaccharides and can serve as a glycan signature. Thus, we set a key requirement that for all glycans reported, at least one other glycan was also detected that differed by one monosaccharide and that used a single linkage clustering algorithm to identify glycans families.

Tandem Mass Spectrometry. Targeted high resolution tandem mass spectrometry was performed to add an additional level of verification beyond the GlyQ-IQ paradigm. GlyQ-IQ results from an initial precursor MS only run were used to populate a targeted fragmentation list for the Velos Pro Orbitrap mass spectrometer. The 60k resolution MS/MS scans were used to provide accurate diagnostic masses (preferential in HCD) and monosaccharide differences (preferential in CID). Ions detected from the list were fragmented with CID and HCD and annotated with glycan diagnostic ions and monosaccharide differences. The characteristic diagnostic oxonium ions and monosaccharide differences used to confirm glycan compositions are presented in Supplementary Table 1 in the Supporting Information.[26] At least one diagnostic ion or monosaccharide mass difference is required for acceptance at this additional validation level.

Data Visualization. *GlyQ-IQ Viewer.* The GlyQ-IQ feature viewer is based on the SIPPER viewer engine[27] facilitates the rapid visualization and review of LC−MS features. The GlyQ-IQ viewer uses the IQ base results files and works well with GlyQ-IQ output files. The GlyQ-IQ viewer was used to drill down into the raw data and display a smoothed EIC and averaged mass spectrum corresponding to the chromatographic peak and isotopic envelope of the result. Each result was viewed in one screen and accepted or denied by a user before being exported to from the final result list. Glycan family relationships are calculated after false hits have been removed. A screen shot of the GlyQ-IQ viewer is included in as Supplemental Figure 14 in the Supporting Information.

GlycoGrid 4D Visualization. A GlycoGrid is a fast and efficient method for viewing and comparing N-glycan composition profiles on a single plot.[16] The GlycoGrid 4D visualization software presented here plots the four compositions (hexose, HexNAc, fucose, and sialic acid) in a four-dimensional grid and denotes detection of a composition by coloring a grid square and populating the grid square with a number denoting the quantity of isomers detected (font too small to be displayed here). Distinct peaks in the EIC that are not related to other compositions via insource fragmentation are considered glycan isomers. Because of the increased size of the retrosynthetic glycan library from our previous publications,[4,16] the GlycoGrid dimensions have increased in size, respectively. Zoom functionality has been incorporated to focus on individual glycan compositions or families and compare across multiple data sets. When zoomed in, the number of isomers (chromatographic separated features) is displayed. The software is a Windows Model View ViewModel application written in C# and is available online at http://omics.pnl.gov/software along with a description of the features incorporated.

RESULTS AND DISCUSSION

Results. *Human Serum High Mannose Glycans.* The high mannose glycans in human serum provide a good case study because there are a relatively fewer compositions relative to complex or hybrid glycans. In this work, separable chromatographic peaks of isobaric masses consistent with glycan compositions are considered isomers. Further structural analysis of isomeric peaks is still required for detailed linkage analysis. Chromatographic features were either annotated as glycan isomers with all 5 GlyQ-IQ variables including supporting insource fragmentation (5-Var), confirmed glycan fragments by insource fragmentation (Fragment) or detected with the other 4 GlyQ-IQ variables (exact mass, Isotope fit score, LC profile coefficient of determination, and glycan family relationships). Glycans with discrete chromatographic peaks without coeluting larger glycan masses are considered glycan isomers. A table of the high mannose glycan isomers is presented in Supplementary Table 3 in the Supporting Information. An example of the insource fragmentation detection for convoluted isomers of Man_6 has been selected and presented in Supplementary Figure 15 in the Supporting Information. EIC from Man_5 and Man_7 are depicted to show inter EIC relationships of family glycans. Both 1+ and 2+ charge starts are shown and the 2+ is inverted for clarity. Tie lines show how many glycans have mapped elution peaks indicating insource fragmentation and do not represent specific isomers present in the solution.

Human Blood Serum N-Glycan Profile. A total of 156 N-glycan compositions and 685 isomers were detected with the GlyQ-IQ software within -0.6 ± 2 ppm mass error. In total, 46 of those compositions had confirmation with insource fragmentation, and targeted MS/MS confirmed an additional

dx.doi.org/10.1021/ac501492f | Anal. Chem. 2014, 86, 6268−6276

60 (32 were confirmed with both). The glycan profile results are presented in a GlycoGrid 4D image (Figure 4) and a

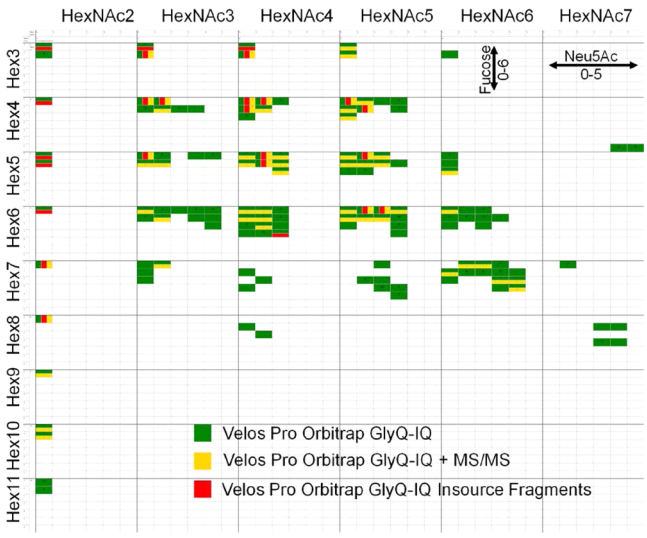


Figure 4. GlycoGrid 4D of 156 human blood serum glycan compositions annotated with GlyQ-IQ (green) and confirmed with CID or HCD (yellow). Glycans identified as insource fragments are depicted in red. The major *Y* axis corresponds to the number of hexose and the minor *Y* corresponds to the number of fucose. The major *X* corresponds to the number of *N*-acetylhexosamine and the minor *X* corresponds to the number of sialic acid.

detailed GlyQ-IQ output table (Supplementary Table 4 in the Supporting Information). The GlycoGrid clearly depicts which glycans were detected (green) and which were confirmed with tandem MS/MS (yellow). Glycans detected and identified as insource fragments are depicted in red. Although the entire prospective ion list was used for targeted fragmentation, 96 compositions were not selected for fragmentation during the run due to low abundance or limited instrument duty cycle.

Insufficient fragmentation information derived from low abundance glycans restricts the overall glycan profile achievable with MS/MS confirmation and also highlights the added information obtained with an MS-based approach. Of the 156 compositions, 29% were confirmed at the 5-variate level via insource fragmentation and 38% with tandem MS/MS.

For comparison, 96 glycan compositions were detected in our previous work from a different aliquot of the same human blood serum sample and comparable sample processing and HPLC (45 cm vs the present 70 cm long).[16] Note that different column lengths will have a larger effect on the total number of discrete isomers separated and detected but the compositions are roughly comparable because we are not limited to only tandem mass spectrometry for annotations and subsequently instrument duty cycle limited. Previously, the data was acquired on an Agilent 6538 Q-TOF MS in contrast to the Velos Pro Orbitrap MS used here.[16] Although the number of glycan compositions is consistent, GlyQ-IQ was able to confidently determine how many isomer peaks are present for each composition and remove false positives assignments caused by insource fragmentation.

DISCUSSION

Targeted Approach. The Glycomic Quintavariate Informed Quantification (GlyQ-IQ) software has been developed to more effectively reveal N-glycan compositions in high resolution LC−MS data sets. Informed quantification (IQ) based algorithms are targeted chromatograph centric algorithms and subscribe to the paradigm that compounds are known or predicted before the analysis takes place; i.e., an array of assumed possible target compounds inform the analysis. This helps restrict the search space and improves success because the compound can be accurately modeled *a priori* before searching for its fingerprint in the data and quantifying its amount. GlyQ-IQ is an extension of this design applied to N-glycan profiling.

Since glycan synthesis is not a template driven process (as compared to peptides and proteins which can leverage the genome), target libraries need to be experimentally determined or predicted *in silico*. The retrosynthetic glycan library approach[28] has been expanded into an informed targeted approach where each composition in the glycan libraries is considered a distinct target. Briefly, the retrosynthetic glycan approach is to constrain the library to the largest glycan structure from each glycan class postulated by the glycosyltransferases present in the system (or experimentally detected). Then it is assumed that if there is sufficient glycan machinery to create the largest structure, any glycan containing fewer monosaccharides may be present. This helps bound the size of the glycome and imparts biological rules for glycan compositions.[28] The glycan library approach allows for searching for adducted glycan by incorporating adducts into the target list and researching the data set. The high speed of the HPC deployment of GlyQ-IQ makes multiple searches with large numbers of targets a pedestrian task. The library implemented here is bounded by the glycan rules presented previously[28] and modified to allow for increased sialylation. The 2 195 target compositions searched are bounded in monosaccharide count as follows (hexose 3−12, *N*-acetylhexosamine 2−8, fucose 0−7, Neu5Ac 0−9).

Knowing the monoisotopic mass of the glycan and elemental composition upfront, we are able to search the LC−MS data space for its mass and exact isotopic envelope. Better isotopic envelope models improve the fit scores, increasing discrimination between correct and false fits. The targets are readily parallelized because each target is processed atomically and does not rely on other targets until the finalization step. The first step is to create a set of charged targets that correspond to each charge state of interest (determined from the data) and process them independently. The general flow of a charged target is to discover associated candidate LC peaks in an EIC (at that charge state) in the time domain and validate each LC peak's isotope profile in the mass domain to remove false hits such as within-tolerance isobaric isotopes and noise peaks. For each acceptable target LC peak, candidate insource fragments are searched for and their EIC generated. The chromatographic peaks of the target and the fragment are subsequently modeled and the modeled peaks are robustly correlated. During the finalization steps, a triage of results analyses is performed to differentiate which LC peaks are intact glycans or glycan fragments.

Confidence via Quintavariate Metrics. Orthogonal pieces of information can be orchestrated at the feature detection level and glycan assignment level to annotate glycans with a confidence greater than each level independently. The mass dimension is characterized by modeling isotopic distributions, fitting it to the data, and scoring how well the experimental data fits the model (Fit Score). Observed isotope profiles that fit well provide the monoisotopic mass useful for exact mass

dx.doi.org/10.1021/ac501492f | *Anal. Chem.* 2014, 86, 6268−6276

Table 1. Table of Different Degrees of Confidence in Annotations[a]

		basic	relationship	insource fragmentation	fragmentation with relationship
		trivariate	quadravariate		quintavariate
variate	category	3-Var	4-Var		5-Var
1	mass accuracy (ppm)	7	7	7	7
2	isotope fit score (0−1)	0.1	0.1	0.1	0.1
3	EIC LC peak size (scans)	9	9	9	9
3	LC Peak model fit (0−1)	0.85	0.85	0.85	0.85
4	monosaccharide relationships (yes,no)		yes		yes
5	insource fragmentation correlation (0−1)			0.95	0.95

[a]Variables 2 and 3 correspond to the two-dimensional quality of the LC−MS feature detection, and variables 1, 4, and 5 correspond to the glycan annotation of the feature.

measurements and determining the mass measurement accuracy (typically measured in parts-per-million, PPM). The LC dimension is often divided up into EICs in which a single mass (±a few ppm mass tolerance) is traced out over time. Elution profiles can be characterized by smoothing the data to remove noise, fitting a model peak shape and then scoring how well the model fits the data. Since many glycans have sequential, positional, and linkage isomers (multiple structures with the same mass and elemental composition), the EIC can have multiple peaks, one corresponding to each glycan isomer structure. In the absence of structural information, isobaric glycan features with multiple LC elution peaks are considered glycan isomers.

Glycan discovery is based on multivariate detection, and each additional orthogonal variable improves confidence of assignment. Up to five orthogonal measurements are implemented: mass measurement error, isotopic envelope fit score, LC peak shape, monosaccharide family relationships, and monosaccharide insource fragmentation. A table consisting of the different confidence levels of annotation is included in Table 1. Acceptable metrics in all five measurements are required for glycan compositional confirmation and the four variables (excluding insource fragmentation) required for annotation (see the Methods section). Since most glycans did not have detectable insource fragmentation, much of the annotation was determined by the remaining four variables. In combination with glycan family relationships, this indicates all annotated glycan reported have at least 8 variables within tolerance supporting the assignment.

CONCLUSION

GlyQ-IQ is a software application for glycan MS based upon an algorithm centric nontargeted analyses approach. Combining spectral averaging, mass-independent fit score calculations, tight fit score tolerances, and a 1 Da interfering peak deconvolution result in over 99% of all glycans identified passed manual inspection and juxtaposition to high-resolution mass spectra. In addition to high specificity LC−MS feature detection, key integration of glycan relationships and insource fragmentation detection further increased the annotation confidence and decreased the false positive rate. Hig- performance computation allowed for 550-fold speed up which translates hours of runtime into a few minutes for large glycome searches (2 156 candidate glycan targets). A total of 156 N-glycan compositions were detected and with 640 intact glycan isomer peaks provide an improved perspective on the glycans present on human blood serum glycoproteins.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rds@pnnl.gov.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Portions of this work were supported by the U.S. DOE office of Biological and Environmental Research Pan-omics project of the Genome Sciences Program, as well as by the NIH GMS Grant P41 GM103493-11. Work was performed in the EMSL, a DOE-BER national scientific user facility PNNL. High-performance computing research was performed using PNNL Institutional Computing at Pacific Northwest National Laboratory. The Microsoft Azure Research was made possible by a Windows Azure Research Pass Grant. We also acknowledge Daniel Fay and Wen-ming Ye from Microsoft Research (http://azure4research.com, Redmond, WA), Magnus Mårtensson from Mårtensson Consulting (Malmö, Sweden), and Alan Smith from Active Solutions (Stockholm, Sweden) for their expertise and support with the Microsoft Windows Azure Cloud Deployment. PNNL is a multiprogram national laboratory operated by Battelle Memorial Institute for the DOE under Contract DE-AC05-76RLO 1830.

REFERENCES

(1) An, H. J.; Froehlich, J. W.; Lebrilla, C. B. *Curr. Opin. Chem. Biol.* **2009**, *13*, 421−426.
(2) Strum, J. S.; Nwosu, C. C.; Hua, S.; Kronewitter, S. R.; Seipert, R. R.; Bachelor, R. J.; An, H. J.; Lebrilla, C. B. *Anal. Chem.* **2013**, *85*, 5666−5675.
(3) Goldberg, D.; Bern, M.; North, S. J.; Haslam, S. M.; Dell, A. *Bioinformatics* **2009**, *25*, 365−371.
(4) Kronewitter, S. R.; An, H. J.; de Leoz, M. L.; Lebrilla, C. B.; Miyamoto, S.; Leiserowitz, G. S. *Proteomics* **2009**, *9*, 2986−2994.
(5) Goldberg, D.; Sutton-Smith, M.; Paulson, J.; Dell, A. *Proteomics* **2005**, *5*, 865−875.
(6) Kronewitter, S. R.; De Leoz, M. L.; Strum, J. S.; An, H. J.; Dimapasoc, L. M.; Guerrero, A.; Miyamoto, S.; Lebrilla, C. B.; Leiserowitz, G. S. *Proteomics* **2012**, *12*, 2523−2538.
(7) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008**, *7*, 1650−1659.
(8) Maass, K.; Ranzinger, R.; Geyer, H.; von der Lieth, C. W.; Geyer, R. *Proteomics* **2007**, *7*, 4435−4444.

(9) Yu, C. Y.; Mayampurath, A.; Hu, Y.; Zhou, S.; Mechref, Y.; Tang, H. *Bioinformatics* **2013**, *29*, 1706−1707.
(10) Vakhrushev, S. Y.; Dadimov, D.; Peter-Katalinic, J. *Anal. Chem.* **2009**, *81*, 3252−3260.
(11) Maxwell, E.; Tan, Y.; Tan, Y.; Hu, H.; Benson, G.; Aizikov, K.; Conley, S.; Staples, G. O.; Slysz, G. W.; Smith, R. D.; Zaia, J. *PLoS One* **2012**, *7*, e45474.
(12) Renard, B. Y.; Kirchner, M.; Steen, H.; Steen, J. A. J.; Hamprecht, F. A. *BMC Bioinf.* **2008**, *9*, 355.
(13) Senko, M. W.; Beu, S. C.; McLaffertycor, F. W. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.
(14) An, H. J.; Tillinghast, J. S.; Woodruff, D. L.; Rocke, D. M.; Lebrilla, C. B. *J. Proteome Res.* **2006**, *5*, 2800−2808.
(15) Rost, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinovic, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmstrom, J.; Malmstrom, L.; Aebersold, R. *Nat. Biotechnol.* **2014**, *32*, 219−223.
(16) Marginean, I.; Kronewitter, S. R.; Moore, R. J.; Slysz, G. W.; Monroe, M. E.; Anderson, G.; Tang, K.; Smith, R. D. *Anal. Chem.* **2012**, *84*, 9208−9213.
(17) Maiolica, A.; Borsotti, D.; Rappsilber, J. *Proteomics* **2005**, *5*, 3847−3850.
(18) Jaitly, N.; Mayampurath, A.; Littlefield, K.; Adkins, J. N.; Anderson, G. A.; Smith, R. D. *BMC Bioinf.* **2009**, *10*, 87.
(19) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320−332.
(20) Barkauskas, D. A.; An, H. J.; Kronewitter, S. R.; de Leoz, M. L.; Chew, H. K.; de Vere White, R. W.; Leiserowitz, G. S.; Miyamoto, S.; Lebrilla, C. B.; Rocke, D. M. *Bioinformatics* **2009**, *25*, 251−257.
(21) Bochkanov, S.; Bystritsky, V. *ALGLIB*; 2013 (www.alglib.net).
(22) Goodman, K. J.; Brenna, J. T. *Anal. Chem.* **1994**, *66*, 1294−1301.
(23) Schulz-Trieglaff, O.; Pfeifer, N.; Gropl, C.; Kohlbacher, O.; Reinert, K. *BMC Bioinf.* **2008**, *9*, 423.
(24) Gabelica, V.; De Pauw, E. *Mass Spectrom. Rev.* **2005**, *24*, 566−587.
(25) Zaia, J. *OMICS: J. Integr. Biol.* **2010**, *14*, 401−418.
(26) Huddleston, M. J.; Bean, M. F.; Carr, S. A. *Anal. Chem.* **1993**, *65*, 877−884.
(27) Slysz, G. W.; Steinke, L.; Ward, D. M.; Klatt, C. G.; Clauss, T. R. W.; Purvine, S. O.; Payne, S. H.; Anderson, G. A.; Smith, R. D.; Lipton, M. S. *J. Proteome Res.* **2014**, *13*, 1200−1210.
(28) Kronewitter, S. R.; An, H. J.; de Leoz, M. L.; Lebrilla, C. B.; Miyamoto, S.; Leiserowitz, G. S. *Proteomics* **2009**, *9*, 2986−2994.