Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation

Kshitij Khatri¹, Joshua A. Klein^{1,2}, and Joseph Zaia^{1,2}

¹Department of Biochemistry, Center for Biomedical Mass Spectrometry, Boston University Medical Campus, 715 Albany Street, Boston, MA 02118, USA

²Bioinformatics Program, Boston University, 24 Cummington Mall, Boston, MA 02215, USA

Abstract

In order to interpret glycopeptide tandem mass spectra, it is necessary to estimate the theoretical glycan compositions and peptide sequences, known as the search space. The simplest way to do this is to build a naïve search space from sets of glycan compositions from public databases and to assume that the target glycoprotein is pure. Often, however, purified glycoproteins contain co-purified glycoprotein contaminants that have the potential to confound assignment of tandem mass spectra based on naïve assumptions. In addition, there is increasing need to characterize glycopeptides from complex biological mixtures. Fortunately, liquid chromatography-mass spectrometry (LC-MS) methods for glycomics and proteomics are now mature and accessible. We demonstrate the value of using an informed search space built from measured glycomes and proteomes to define the search space for interpretation of glycoproteomics data. We show this using α -1-acid glycoprotein (AGP) mixed into a set of increasingly complex matrices. As the mixture complexity increases, the naïve search space balloons and the ability to assign glycopeptides with acceptable confidence diminishes. In addition, it is not possible to identify glycopeptides not foreseen as part of the naïve search space. A search space built from released glycan glycomics and proteomics data is smaller than its naïve counterpart while including the full range of proteins detected in the mixture. This maximizes the ability to assign glycopeptide tandem mass spectra with confidence. As the mixture complexity increases, the number of tandem mass spectra per glycopeptide precursor ion decreases, resulting in lower overall scores and reduced depth of coverage for the target glycoprotein. We suggest use of α -1-acid glycoprotein as a standard to gauge effectiveness of analytical methods and bioinformatics search parameters for glycoproteomics studies.

Correspondence to: Joseph Zaia.

Kshitij Khatri and Joshua A. Klein contributed equally to this work.

Published in the topical collection *Glycomics, Glycoproteomics and Allied Topics* with guest editors Yehia Mechref and David Muddiman.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-016-9970-5) contains supplementary material, which is available to authorized users.

Conflict of interests The authors have no conflicts of interest.

Keywords

Glycoproteomics; Integrated-omics; Glycoinformatics; Mass spectrometry; Glycosylation; Glycomics; Alpha-1-acid glycoprotein

Introduction

While analytical methods have developed to the point that it is now feasible to acquire high quality MS and tandem MS data on glycopeptides, a consensus regarding confidence in site-specific glycosylation assignments has yet to emerge. This results in part from the fact that glycoproteomics methods are still maturing [1]; however, the lack of well-characterized standard glycoproteins also limits the ability to assess methods unambiguously. Given the central role of glycosylation in physiology, the field requires clear understanding of the appropriate uses and limitations of both mature and emerging glycoproteomics methods.

For our purposes, glycoproteomics is the determination of site-specific glycosylation of glycoproteins. This includes efforts to characterize glycopeptides in a bottom-up fashion, whereby glycoproteins are digested with proteolytic enzymes and glycopeptides with a single site of glycosylation are identified using LC-MS. The scope of glycoproteomics ranges from characterization of recombinant or purified glycoproteins to discovery of glycopeptides from complex biological matrices, including tissue, blood, and urine. Thus, as in proteomics, the complexity ranges over several orders of magnitude.

Determination of accurate masses defines the range of glycopeptide compositions with respect to peptide and glycan present; however, even for the most accurate measurements, the information is ambiguous [2]. One approach to reducing the ambiguity is to identify clusters of glycopeptide masses that differ by monosaccharide units [3, 4] and/or to digest using specific exoglycosidases [5]. The glycopeptide ions can also be selected for tandem mass spectrometry. Collisional excitation of glycopeptides, including collision-induced dissociation (CID), collisionally activated dissociation (CAD), and higher-energy collisional dissociation (HCD), preferentially dissociates the glycan portion of the glycopeptide. The extent of dissociation is higher for non-resonant dissociation such as those observed in beam type (Q-TOF, triple quadrupole and HCD) dissociation than for resonant dissociation observed in trapped ion instruments. It is necessary to observe peptide backbone fragments in glycopeptide tandem MS for confident assignment of glycopeptide identity [6–8]. Because the glycopeptide ions are protonated, monosaccharide rearrangements occur [9], making assignment of glycan topologies from collisional dissociation tandem mass spectra risky.

Investigators have developed approaches to determine the mass of the peptide portion of glycopeptides using the presence of ions corresponding to peptide+n (monosaccharide), where $n = 0-3$, in the tandem mass spectra [4, 10–13]. Provided sufficiently high collision energy is used, glycopeptide tandem mass spectra contain product ions from dissociation of the peptide backbone [14, 15] that are useful for assigning the glycopeptide sequence [12, 16]. Other investigators have used ion trap multistage tandem MS to isolate and dissociate the peptide portion of the molecule [11].

Electron activated dissociation (ExD) preferentially cleaves the peptide portion of glycopeptides to produce a sequence ladder [17]. Thus, the use of a combination of collisional dissociation and ExD produces complementary information on the glycan and peptide, respectively [18]. Electron transfer dissociation (ETD) [19–22] has proven very useful for the analysis of glycopeptides [23]. Investigators have therefore developed schemes whereby the presence of signature glycan oxonium ions in collisional tandem mass spectra are used to trigger subsequent ETD tandem mass spectra [24]. While the extent of dissociation of the peptide backbone is higher for ETD, the overall efficiency is lower than collisional dissociation, requiring more abundant precursor ions to accommodate longer reaction times.

In order for the field to reach consensus regarding analytical and bioinformatics methods for glycoproteins, it is important to examine the underlying assumptions. One assumption regards the structural complexity of glycoproteins. Thus, glycoprotein analysis can be viewed as determining which among a set of theoretical structures is the most probable based on the known or assumed sample mixture. The set of theoretical glycopeptides, known as a search space, consists of the peptide sequences, multiplied by number of possible glycan structures, multiplied by any other structural/chemical variants (such as oxidations, deamidations, etc.). Ideally we would like to use the mass spectral data to calculate confidence values for the search space so as to identify the site-specific glycosylation present. In order to do this, the search space must be defined accurately. The best way to define the proteins present in the sample is to measure the proteome. Likewise, analytical measurement of the released glycome is the best way to determine the glycans present. The purpose of this article is to examine the effectiveness of glycoproteomics under scenarios where the search space is defined based on naïve assumptions versus using glycomics and proteomics data and the effects of increasing matrix complexity on data acquisition and interpretation.

Materials and methods

Standard human serum glycoproteins were purchased from Sigma-Aldrich (St. Louis, MO). Glycoproteins were reduced with dithiothreitol and alkylated using iodoacetamide. The samples were then subjected to tryptic digestion and individual glycoprotein tryptic digests were mixed to generate samples of scaling complexity levels as below:

Alpha-1-acid glycoprotein (AGP) (complexity level 1)

Glycoprotein mix 3: AGP + transferrin + fetuin

Glycoprotein mix 4: AGP+ transferrin + fetuin + haptoglobin

Glycoprotein mix 5: AGP + transferrin + fetuin + haptoglobin + alpha-2-macroglobulin

The highest complexity level analyzed was AGP in pooled human serum (Sigma-Aldrich, St. Louis, MO). Since, serum contains very high abundances of albumin and immunoglobulin G (IgG), it was necessary to deplete these resident proteins to observe coverage on the glycoproteins of interest [25–27]. Albumin and IgG were depleted using ProteoExtract®

Albumin/IgG Removal Kit (EMD Millipore, Billerica MA). Depleted serum and glycoprotein mixes were subjected to sample preparation and analysis using a multi-pronged approach, as described previously [28].

Proteomics analyses

Tryptic samples were deglycosylated using peptide *N*-glycosidase *F* (PNGaseF, New England Biolabs, Ipswich MA) and subjected to triplicate C18 LC-MS/MS analyses using a Q-Exactive Plus mass spectrometer (Thermo Scientific, San Jose, CA) equipped with an Advion NanoMate nanoESI source, coupled to a Waters NanoAcquity nanoLC system. A Waters™ Xbridge™ reversed-phase column (150 $\mu\text{m} \times 100 \text{ mm}$) with 1.7 μm BEH C18 resin and a Waters™ trap column (180 $\mu\text{m} \times 20 \text{ mm}$) packed with 5 μm Symmetry™ C18 stationary phase were used for online desalting and separation of the proteomics samples. The mass spectrometer was programmed to acquire data-dependent tandem MS, using instrument parameters described previously [28]. Glycoproteins used in complexity mixtures were analyzed both individually and as mixtures of increasing complexity.

Glycomics analyses

N-Linked glycans were released from tryptic digests using PNGaseF. Released glycans were separated from deglycosylated peptides using C18 reversed-phase spin columns (Thermo Pierce, Rockford IL). Glycans were desalted using PD MiniTrap G-10 columns (GE Healthcare, Piscataway NJ), as per the manufacturer's instructions. Glycan samples were analyzed, in triplicate, using HILIC-MS using an Agilent™ 6520 LC-MS system with a chip-cube nanoESI source, as described previously [29].

Glycoproteomics analyses

Tryptic digests of glycoprotein mixes and depleted serum were analyzed using a previously described chip LC-MS platform [15]. Glycopeptide enrichment was performed online using a HILIC trapping column followed by C18 reversed-phase separation and data-dependent tandem MS analysis.

Proteomics data analysis

Proteomics data were analyzed using Peaks Studio 7.5 [30] (Bioinformatics Solutions Inc., Waterloo ON). Triplicate data files for individual samples were combined in Peaks Spider Searches, which combine database, error-tolerant PTM searching and de novo sequencing. Data were searched against a Uniprot protein database, using a precursor mass tolerance of 10 ppm and a product ion mass tolerance of 0.1 Da. A maximum of two missed cleavages per peptide were specified; carbamidomethylation at cysteine was used as a fixed modification and deamidation (Asn, Gln), oxidation (Met), acetylation (Lys and N-terminus), Na adduct (Asp, Glu, C-term), and pyroglutamination (Gln at N-terminus) were used as variable modifications in the database searches. The PTM search examined the data for presence of 484 variable modifications from the Unimod database [31] and possible mutations leading to amino acid substitutions. The results from combination of database, PTM, and de novo searches were accepted at a false discovery rate (FDR) threshold of

0.1 %. Peptides with spectrum matches of $-10\log p$ values 15 or higher were used for downstream processing and building a glycoproteomics database.

Glycomics data analysis

Glycomics data were deconvoluted and deisotoped using DeconTools (version 1.0.5501) [32, 33]. Scan-specific monoisotopic peaklists from LC-MS analyses were matched against a combinatorically generated glycan composition database, using the following parameters:

Monosaccharide lower and upper bounds:

Hexose_{3–10}

HexNAc_{2–8}

Fucose(deoxyhexose)_{0–5}

NeuAc_{0–4}

Rules:

Fucose(dHex) $\leq$ HexNAc – 1

NeuAc $\leq$ HexNAc – 2

Database generation and matching were performed using a 10-ppm mass error tolerance using GlycReSoft [34].

Glycoproteomics data analysis

Informed databases for integrated omics were generated as described previously [28], except that the only glycans which were included from the glycomics analysis had an aggregated abundance greater than twice the mean abundance of all glycan matches. Tandem mass spectra were identified by first recalculating the precursor ion monoisotopic mass and charge, followed by a database search procedure with a precursor mass tolerance of 10 ppm. For each glycopeptide which fell within the acceptable mass range, theoretical product ions for b, y, b plus HexNAc, y plus HexNAc, and intact peptide plus incremental losses of saccharide units (known as “stub ions”), were assigned, as shown in the Electronic Supplementary Material (ESM) Fig. S1A. The tandem spectra were deconvoluted and peaks were matched for each theoretical product ion with an error tolerance of 10 ppm, constructing a glycopeptide-spectrum match (GSM).

For each GSM, a score based upon peptide backbone coverage and presence of stub ions was computed and scaled to be between 0.0 and 1.0. For each spectrum where multiple glycopeptides could be assigned, all GSMs tied for the highest score were added to a results set.

For each experiment, a forward-sequence (target) and a reverse-sequence-with-valid-sequon (decoy) database was searched, and a q -value without PIT as described by Käll et al. [35] was computed for the paired result sets. Target spectra which had a q -value < 0.05 were selected for inclusion in the final reported results.

Results

The complexity of a sample mixture has implications for both the analytics and informatics methods used in data acquisition and data analysis. Assumptions about sample complexity or purity can have serious consequences on the quality of data and information generated. Because population heterogeneity is inherent in glycoconjugates, the number of glycosylated molecular forms propagates quickly with added heterogeneity from bottom-up sample preparation methods and presence of other PTMs. Additional proteins or glycoproteins in the sample matrix will impact the analytical and informatics performance. Therefore, we assessed the effects of sample matrix complexity on analytical methods and informatics. We also demonstrated how integration of different data domains improves estimation of complexity and keeps false discovery rates in check.

Effects of assumptions about glycoprotein search space

Typical assumptions in glycoprotein characterization are related to sample purity. Most search spaces do not account for presence of contaminating proteins/glycoproteins in the sample analyzed. For biotherapeutic formulations, these contaminants are often components of the culture system used for production of the biomolecule. For biological samples derived from animals or plants, co-purified molecules from the sample matrix are commonly present. Commercially available standard glycoproteins are commonly used for method development and validation and are usually assumed to be pure; however, while the target glycoprotein may be the most abundant molecule present in the preparations, it is not safe to assume that it is the only (glyco)protein present. We assessed purity of commercially available AGP and transferrin glycoprotein standards using proteomics. Table S1 (see ESM) shows the top scoring proteins identified in the searches. It is clear from the proteomics results that several contaminating glycoproteins are present in the samples. Furthermore, Fig. 1A shows the total number of proteins identified in a proteomics analysis of each purified glycoprotein. While all standard glycoproteins originated from serum or plasma, each has its own unique subset of contaminants, which are co-purified despite the enrichment and purification steps.

Evaluation of search space construction methods: naïve versus informed

In addition to contaminating proteins present in the sample, presence of PTMs, non-specific cleavages, or missed cleavages in bottom-up sample preparation methods can also confound the data analysis. In the absence of a proteomics experiment, it is almost impossible to gauge the number and type of protein variants in the sample. Traditional proteomics search methods require the user to specify a list of modifications and proteolytic cleavage types to be considered, which also limits the search space to assumptions [36].

Assuming 180 common glycoforms and the presence of deamidation (N/Q), oxidation (M), and water loss, a single glycopeptide from transferrin (QQHLFGSNVTDCSGSFCLFR) will have 1440 structural possibilities. The degree to which the naïve approach estimates accurately the FDR depends on the assumptions; unfortunately, in the absence of data, there is no way of evaluating the correctness of these assumptions. As a result, the naïve method may over- or underestimate search space size. Underestimation arises when a glycoprotein is

assumed incorrectly to be pure. This situation leads to false confidence in the results when FDR is calculated for glycopeptides derived from the pure glycoprotein. The presence of contaminating glycoproteins in the sample means that the search space is in reality much larger than assumed and that the FDR calculated from the assumed search space underestimates the rate of false identifications.

Overestimation of search space size is a result of including too many glycoproteins and glycoforms and may lead to inability to differentiate true from false identifications; in such cases many high quality tandem mass spectra will have spurious matches against components in the search space. In summary, the validity of an FDR calculation depends on the assumptions of search space size. To the extent that the search space accurately estimates sample complexity, the FDR correctly expresses the confidence of glycopeptide assignments.

Error-tolerant searches and de novo sequencing tools have gained popularity in proteomics [30, 37–39]. These allow the user to analyze data without bias from pre-defined search parameters to maximize coverage of the proteome with respect to the different molecular forms present. Glycosylation adds a large molecular weight and considerable macro and microheterogeneity, and occurs in addition to other non-glycan modifications present on proteins or peptides. Glycosylation also introduces glycosidic product ions in tandem MS, not found for unglycosylated peptides, complicating the identification process. As a result, proteomics search engines that perform error-tolerant searches are unable to handle complex glycosylation as a modification. For similar reasons, emerging software tools that are capable of assigning glycopeptide compositions are unable to perform large error-tolerant searches for other PTMs and therefore rely on user specified search spaces. We assert that the most efficient circumvention to this problem is to use the identifications from proteomics data to inform the search space for glycoproteomics.

To prove this, we used mzIdentML proteomics search results from a combination of error-tolerant and de novo searches and the confidently identified peptide molecular forms to build a search space for all possible glycopeptidoforms present in our samples. Similarly, to define our glycoform search space, we performed glycomics searches on our released glycan pools and iteratively combined the results with the proteomics identifications that contained glycosylation sites. Figure 1 compares the sizes for a naïve search space (B), built for glycoprotein mixture 3 based on assumptions about glycoproteins and PTMs present versus an informed search space (C), built using the proteomics search results to define the range of glycoproteins and PTMs present in the sample and a further constrained informed search space (D) that included glycomics data. Strikingly, the proteomics and glycomics informed search space in panel D had less than half the number of total theoretical glycopeptides compared to the naïve search space in panel B. While proteomics information helped identify all contaminating glycoproteins in the sample, it can lead to an explosion of the search space size, which can render the glycopeptide search process inefficient; glycomics helped restrict the number of glycoforms to those that are actually detected in the glycomics analyses. It is therefore important to use both proteomics and glycomics in glycopeptide search space construction.

The site-specific glycoproteomes of transferrin and alpha-1-acid glycoprotein

Figure 2 shows the base peak chromatogram (A) for size-enriched glycopeptides from commercially sourced human transferrin. Chromatographic peaks corresponding to glycopeptides were identified based on the extracted ion chromatogram for the HexNAc and Hex-HexNAc oxonium ions (B and C). We interpreted the glycopeptide tandem mass spectra using a search space containing putative transferrin glycopeptides; this analysis identified three of the EIC peaks as transferrin glycopeptides. Next, we acquired proteomics data on the PNGaseF deglycosylated sample and found that hemopexin, a protein with five *N*-glycosylation sequons, was in the sample, in addition to transferrin (ESM Table S1B). When we added hemopexin to the search space, we identified abundant chromatographic peaks corresponding to hemopexin glycopeptides.

α 1-Acid glycoprotein contains two protein isoforms with five sequons, each [15, 40, 41]. To construct a naïve glycopeptide search space, we allowed 0–2 missed tryptic cleavage sites, methionine oxidation, and 0–1 sites of deamidation for asparagine-containing glycopeptides. Table 1 shows the naïve search space size calculated assuming pure AGP, AGP plus haptoglobin, and AGP plus haptoglobin plus transferrin. We then compared the results from our informed versus naïve search spaces.

The search space size ballooned 10-fold for the three protein mixture relative to the hypothetically pure AGP protein. We acquired proteomics data for human AGP and found that in addition to the two expected isoforms, the sample contained haptoglobin, a protein that contains four *N*-glycosylation sequons, among the top scoring identified glycoproteins (ESM Table S1A). For both AGP and haptoglobin, there were a large number of hypothetical glycopeptides. Tryptic haptoglobin had glycopeptides containing two sequons, thus the high mass and large number of hypothetical haptoglobin glycopeptides, as seen in ESM Fig. S3.

Because of the sample complexity, we compared the ability to assign glycopeptide tandem mass spectra for a naïve search space versus a search space informed by proteomics and released glycan glycomics (Fig. 3A). The glycomics results for AGP are presented in ESM Fig. S2.

Figure 3B, compares glycoproteomics search results using naïve versus informed hypotheses. The height of the curve at the red bar tells us how many predictions were accepted at our confidence threshold. The informed hypothesis retains substantially more predictions above the confidence threshold, compared to the naïve hypothesis. Therefore, use of a search space informed by proteomics and glycomics data boosts our ability to assign glycopeptides from tandem mass spectral data with confidence. The profile of site-specific AGP glycosylation informed by proteomics and glycomics is shown in Fig. 5.

Confidence in assignment of site-specific glycosylation of an analyte of interest in an increasingly complex matrix

Most glycoproteomics experiments rely on data-dependent acquisition that is directly affected by the complexity of the sample matrix. With increase in matrix complexity, the ability of the analytical methods to interrogate the analyte of interest decreases. The mass

spectrometer may not be able to detect glycopeptides from the glycoprotein of interest depending on the relative abundance of the impurities and the dynamic range of the instrument. Moreover, even when the glycopeptides of interest are within the dynamic range, data-dependent acquisition methods spend less time on collecting tandem MS on these analytes in presence of contaminating glycopeptides/peptides.

We constructed mixtures with scaling complexity by mixing together fixed quantities of different glycoprotein samples. Figure 4 shows how data quality for the AGP glycopeptides of interest deteriorates when complexity increases. Figure 4A shows overlaid base peak chromatograms for AGP alone, glycoprotein mix 3, glycoprotein mix 5 and serum (see the Experimental section for mixture definitions). With each increase in complexity level, the chromatogram became visibly more crowded. The non-AGP glycopeptides in the sample caused saturation of the column binding capacity leading to a decrease in intensity of the glycoprotein/peptides of interest, as seen in Fig. 4B. This could be a problem when trying to get quantitative LC-MS data, since sample matrix complexity could affect the dynamic range of the assay at both the column binding capacity and mass spectrometer detection levels. Figure 4D shows how the precursors were selected for tandem MS as sample complexity grows. The colored dots trailing along the *x*-axis represent tandem MS for different precursors. The long trails of colored dots in the 10–20 and 40–60 represent tandem MS on low abundance contaminants that elute over long periods of the LC gradient. In lower complexity samples, these contaminants were repeatedly selected for tandem MS due to the absence of other precursors eluting in these retention time ranges. In the 20–40-min retention time range where most glycopeptides were found to elute in the AGP chromatogram, multiple tandem MS were acquired for each precursor. This also allows proper triggering of tandem MS for the precursor selected at the chromatographic peak apex and resulted in good quality tandem mass spectra. On the other hand, the distribution of precursors in glycoprotein mix 5 and serum was very sparse and in most cases only a single tandem mass spectrum was collected per precursor. For a better comparison of the complexity of these samples, in Fig. 4C, we compared the average number of tandem MS per precursor for each complexity level. This showed that as the total number of unique precursors increased with increasing complexity, the time-spent acquiring MS/MS on each precursor diminishes. This became a serious limiting factor for low abundance glycopeptide glycoforms, which were either not selected for tandem MS or for which poor quality tandem MS were acquired. Thus, as the complexity increased, in glycoprotein mixes and serum, the mass spectrometer diminished in ability to handle the increase in co-eluting peptides/ glycopeptides and acquired fewer tandem mass spectra per precursor, resulting in poor data quality and lower identification rates.

The effect of increasing complexity was reflected in the depth of glycoform coverage for AGP glycosylation sites as shown in Fig. 5. The most comprehensive assignment of glycopeptides came from the AGP only sample. As mixture complexity increased, the overall scores diminished, resulting in fewer assignments falling above the scoring threshold. This was due to the decreasing number of tandem mass spectra per precursor ion resulting from increased mixture complexity and the finite scanning speed of the instrument. The AGP precursor ion abundances decreased with mixture complexity. This was likely due to a combination of ionization effects and the limited capacity of the glycopeptide

enrichment step, which limited the amount of total protein/peptides loaded on to the analytical column. Since the bar plots represent the sum of fragment ion abundances for all scans matching a glycopeptide, the precursor that had the most high-scoring spectral matches also show up as the most abundant matches on the bar plots.

It was important to integrate data from proteomics and glycomics domains, for accurate estimation of the FDR and for setting confidence thresholds for accepting matched glycopeptide spectra for all complexity levels, as already shown in Fig. 3. While a tandem MS may be acquired for all eluting glycopeptides, good spectral quality is essential for confident matching above the FDR thresholds. For some of the AGP glycosylation sites shown in Fig. 5, glycoprotein mix 5 appeared to have more glycopeptide matches with glycoprotein mix 3. This was generally due to differences in the profile of co-eluting contaminating peptides that affect tandem MS acquisition.

Discussion

For most proteomics and glycoproteomics experiments, multiple sample clean-up and fractionation steps are used for reducing complexity before online LC-MS. While sample fractionation increases glycopeptide coverage, overall matrix complexity impacts the analysis of our molecule of interest. Since, heterogeneity of glycoforms is a major challenge in the analysis of glycoconjugates, the problem of defining the correct boundaries for data analysis is amplified many fold over non-glycosylated molecules. It is therefore important to account for the matrix contribution during data analysis so that assumptions about sample purity do not affect confidence in results. A better understanding of the matrix comes from integrating different data dimensions and prevents erroneous assignment of contaminant data to the analyte of interest, thereby lowering the false discovery rates. Since this step helps determine the nature of contaminants present in the sample matrix, it also guides choice of the appropriate methods required for further sample purification. This can greatly impact successful method development in an industry setting for analysis of biologicals [42]. Collectively, these results demonstrate the effect of assumptions about sample complexity on the false discovery rates and therefore confidence in analyses. Analytical method development benefits greatly from proper data analysis and results visualization to identify any issues that limit performance.

While method performance is heavily dependent on instrument hardware and software, it is essential to establish standard methods and datasets to use as benchmarks for tracking any changes in instrument and method suitability both within an experiment and over long-term use. Alpha-1-acid glycoprotein is reasonably complex, widely available, and can be handled by most LC-MS analytical platforms used for glycoproteomics. AGP therefore serves as a good standard for method development, validation, and suitability tracking.

We demonstrated that predetermination of sample complexity is of utmost importance, when dealing with complex biological matrices. In the absence of empirically derived information, the multiplicity of molecular forms in glycoproteomics can easily overwhelm analytical methods that are not setup keeping sample and matrix complexity in mind. The ramifications are even more serious on the informatics end. Contaminating proteins, sample processing

artefacts, unusual PTMs, mutations, and above all glycoforms can confound glycoproteomics data and severely inflate false discovery rates. Measurement of the proteome of glycoprotein sample constrains the search space size by defining the peptide variants actually present. This allows inclusion of contaminating glycoproteins detected in the proteomics data but were not included in the naïve search space, while preventing addition of unnecessary contaminants based on assumptions about matrix that may not be present in the sample, thus fixing the proteomics search space. In addition, profiling of the released glycans reduces the list of glycans present in the sample and efficiently constrains the search space size. Thus, proteomics and glycomics data help develop an efficient search space that prevents over- or underestimation of the glycosylated molecular forms present in a sample.

Naïve searches fail to eliminate unreasonable hypotheses because of the complexity of real life glycoprotein samples. As a result, the level of false identifications cannot be determined accurately. Informed searches allow unbiased definition of search spaces, thus helping determine confidence levels in assignments, accurately.

We conclude that sample complexity at which acceptable rates of false identifications occur need to be defined using data. We suggest use of AGP as a glycoprotein standard of complexity suitable to test the glycoproteomics workflows. This will help harmonize efforts in the glycoproteomics community and enable a better comparison of data quality and controls. Further, we suggest controls in which AGP is doped into the sample matrix to demonstrate that acceptable rate of false identification results. This will help the experimentalists identify and eliminate any flaws in their analytical strategies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding was provided from NIH grants P41GM105603 and R21CA177476. Thermo-Fisher Scientific provided access to the Q-Exactive Plus mass spectrometer used in this work.

References

1. Leymarie N, Griffin PJ, Jonscher K, Kolarich D, Orlando R, McComb M, Zaia J, Aguilan J, Alley WR, Altmann F, Ball LE, Basmallick L, Bazemore-Walker CR, Behnken H, Blank MA, Brown KJ, Bunz S-C, Cairo CW, Cipollo JF, Daneshfar R, Desaire H, Drake RR, Go EP, Goldman R, Gruber C, Halim A, Hathout Y, Hensbergen PJ, Horn DM, Hurum D, Jabs W, Larson G, Ly M, Mann BF, Marx K, Mechref Y, Meyer B, Möglinger U, Neusüss C, Nilsson J, Novotny MV, Nyalwidhe JO, Packer NH, Pompach P, Reiz B, Resemann A, Rohrer JS, Ruthenbeck A, Sanda M, Schulz JM, Schweiger-Hufnagel U, Sihlbom C, Song E, Staples GO, Suckau D, Tang H, Thaysen-Andersen M, Viner RI, An Y, Valmu L, Wada Y, Watson M, Windwarder M, Whittall R, Wührer M, Zhu Y, Zou C. Interlaboratory study on differential analysis of protein glycosylation by mass spectrometry: the ABRF glycoprotein research multi-institutional study 2012. *Mol Cell Proteomics*. 2013; doi: 10.1074/mcp.M113.03064
2. Desaire H, Hua D. When can glycopeptides be assigned based solely on high-resolution mass spectrometry data? *Int J Mass Spectrom*. 2009; 287:21–6. DOI: 10.1016/j.ijms.2008.12.001

3. Mayampurath AM, Wu Y, Segu ZM, Mechref Y, Tang H. Improving confidence in detection and characterization of protein N-glycosylation sites and microheterogeneity. *Rapid Commun Mass Spectrom.* 2011; 25:2007–19. DOI: 10.1002/rcm.5059 [PubMed: 21698683]
4. Wu Y, Mechref Y, Klouckova I, Mayampurath A, Novotny MV, Tang H. Mapping site-specific protein N-glycosylations through liquid chromatography/mass spectrometry and targeted tandem mass spectrometry. *Rapid Commun Mass Spectrom.* 2010; 24:965–72. DOI: 10.1002/rcm.447 [PubMed: 20209665]
5. Wang WT, LeDonne NC, Ackerman B, Sweeley CC. Structural characterization of oligosaccharides by high-performance liquid chromatography, fast-atom bombardment-mass spectrometry, and exoglycosidase digestion. *Anal Biochem.* 1984; 141:366–81. DOI: 10.1016/0003-2697(84)90057-5 [PubMed: 6496945]
6. Hu H, Khatri K, Klein J, Leymarie N, Zaia J. A review of methods for interpretation of glycopeptide tandem mass spectral data. *Glycoconj J.* 2015; :1–12. DOI: 10.1007/s10719-015-9633-3
7. Dallas DC, Martin WF, Hua S, German JB. Automated glycopeptide analysis—review of current state and future directions. *Brief Bioinform.* 2013; 14:361–74. DOI: 10.1093/bib/bbs045 [PubMed: 22843980]
8. Hu H, Khatri K, Zaia J. Algorithms and design strategies towards automated glycoproteomics analysis. *Mass Spectrom Rev.* 2016; :n/a–n/a. DOI: 10.1002/mas.21487
9. Wührer M, Deelder AM, van der Burgt YEM. Mass spectrometric glycan rearrangements. *Mass Spectrom Rev.* 2011; 30:664–80. DOI: 10.1002/mas.20337 [PubMed: 21560141]
10. Joenväärä S, Ritamo I, Peltoniemi H, Renkonen R. N-Glycoproteomics—an automated workflow approach. *Glycobiology.* 2008; 18:339–49. DOI: 10.1093/glycob/cwn013 [PubMed: 18272656]
11. Wu S-W, Liang S-Y, Pu T-H, Chang F-Y, Khoo K-H. Sweet-Heart—an integrated suite of enabling computational tools for automated MS2/MS3 sequencing and identification of glycopeptides. *J Proteomics.* 2013; 84:1–16. DOI: 10.1016/j.jprot.2013.03.026 [PubMed: 23568021]
12. Lynn K-S, Chen C-C, Lih TM, Cheng C-W, Su W-C, Chang C-H, Cheng C-Y, Hsu W-L, Chen Y-J, Sung TY. MAGIC: an automated N-linked glycoprotein identification tool using a Y1-ion pattern matching algorithm and in silico MS2 approach. *Anal Chem.* 2015; doi: 10.1021/ac5044829
13. Strum JS, Nwosu CC, Hua S, Kronewitter SR, Seipert RR, Bachelor RJ, et al. Automated assignments of N- and O-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures. *Anal Chem.* 2013; 85:5666–75. DOI: 10.1021/ac4006556 [PubMed: 23662732]
14. An Y, Cipollo JF. An unbiased approach for analysis of protein glycosylation and application to influenza vaccine hemagglutinin. *Anal Biochem.* 2011; 415:67–80. DOI: 10.1016/j.ab.2011.04.018 [PubMed: 21545787]
15. Khatri K, Staples GO, Leymarie N, Leon DR, Turiák L, Huang Y, et al. Confident assignment of site-specific glycosylation in complex glycoproteins in a single step. *J Proteome Res.* 2014; 13:4347–55. DOI: 10.1021/pr500506z [PubMed: 25153361]
16. He L, Xin L, Shan B, Lajoie GA, Ma B. GlycoMaster DB: software to assist the automated identification of N-linked glycopeptides by tandem mass spectrometry. *J Proteome Res.* 2014; 13:3881–95. DOI: 10.1021/pr401115y [PubMed: 25113421]
17. Håkansson K, Cooper HJ, Emmett MR, Costello CE, Marshall AG, Nilsson CL. Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptide to yield complementary sequence information. *Anal Chem.* 2001; 73:4530–6. DOI: 10.1021/ac0103470 [PubMed: 11575803]
18. Mechref Y. Use of CID/ETD mass spectrometry to analyze glycopeptides. *Curr Protoc Protein Sci* Editor Board John E Coligan. 2012; A1 0 12:Unit-12.1111. doi: 10.1002/0471140864.ps1211s68
19. Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A.* 2004; 101:9528–33. DOI: 10.1073/pnas.0402700101 [PubMed: 15210983]
20. Viner RI, Zhang T, Second T, Zabrouskov V. Quantification of post-translationally modified peptides of bovine α -crystallin using tandem mass tags and electron transfer dissociation. *J Proteomics.* 2009; 72:874–85. DOI: 10.1016/j.jprot.2009.02.005 [PubMed: 19245863]

21. Scott NE, Parker BL, Connolly AM, Paulech J, Edwards AVG, Crossett B, et al. Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of *Campylobacter jejuni*. *Mol Cell Proteomics MCP*. 2011; 10:M000031–MCP201. DOI: 10.1074/mcp.M000031-MCP201 [PubMed: 20360033]
22. Chalkley RJ, Thalhammer A, Schoepfer R, Burlingame AL. Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proc Natl Acad Sci U S A*. 2009; 106:8894–9. DOI: 10.1073/pnas.0900288106 [PubMed: 19458039]
23. Catalina MI, Koeleman CAM, Deelder AM, Wührer M. Electron transfer dissociation of N-glycopeptides: loss of the entire N-glycosylated asparagine side chain. *Rapid Commun Mass Spectrom RCM*. 2007; 21:1053–61. DOI: 10.1002/rcm.2929 [PubMed: 17311219]
24. Zhao P, Viner R, Teo CF, Boons G-J, Horn D, Wells L. Combining high-energy c-trap dissociation and electron transfer dissociation for protein O-GlcNAc modification site assignment. *J Proteome Res*. 2011; 10:4088–104. DOI: 10.1021/pr2002726 [PubMed: 21740066]
25. Anderson NL, Anderson NG. The human plasma proteome history, character, and diagnostic prospects. *Mol Cell Proteomics*. 2002; 1:845–67. DOI: 10.1074/mcp.R200007-MCP200 [PubMed: 12488461]
26. Echan LA, Tang H-Y, Ali-Khan N, Lee K, Speicher DW. Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *PROTEOMICS*. 2005; 5:3292–303. DOI: 10.1002/pmic.200401228 [PubMed: 16052620]
27. Zhang A, Sun H, Yan G, Han Y, Wang X. Serum proteomics in biomedical research: a systematic review. *Appl Biochem Biotechnol*. 2013; 170:774–86. DOI: 10.1007/s12010-013-0238-7 [PubMed: 23609910]
28. Khatri K, Klein JA, White MR, Grant OC, Leymarie N, Woods RJ, Hartshorn KL, Zaia J. Integrated omics and computational glycobiology reveal structural basis for influenza A virus glycan microheterogeneity and host interactions. *Mol Cell Proteomics*. 2016; mcp.M116.058016. doi: 10.1074/mcp.M116.058016
29. Staples GO, Naimy H, Yin H, Kileen K, Kraiczek K, Costello CE, et al. Improved hydrophilic interaction chromatography LC/MS of heparinoids using a chip with postcolumn makeup flow. *Anal Chem*. 2009; 82:516–22. DOI: 10.1021/ac901706f
30. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics MCP*. 2012; 11:M111.010587. doi: 10.1074/mcp.M111.010587
31. Creasy DM, Cottrell JS. Unimod: protein modifications for mass spectrometry. *Proteomics*. 2004; 4:1534–6. DOI: 10.1002/pmic.200300744 [PubMed: 15174123]
32. Horn DM, Zubarev RA, McLafferty FW. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom*. 2000; 11:320–32. DOI: 10.1016/S1044-0305(99)00157-9 [PubMed: 10757168]
33. Jaitly N, Mayampurath A, Littlefield K, Adkins JN, Anderson GA, Smith RD. Decon2LS: an open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics*. 2009; 10:87. doi: 10.1186/1471-2105-10-87 [PubMed: 19292916]
34. Maxwell E, Tan Y, Tan Y, Hu H, Benson G, Aizikov K, et al. GlycReSoft: a software package for automated recognition of glycans from LC/MS data. *PLoS One*. 2012; 7:e45474. doi: 10.1371/journal.pone.0045474 [PubMed: 23049804]
35. Käll L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res*. 2008; 7:29–34. DOI: 10.1021/pr700600n [PubMed: 18067246]
36. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994; 5:976–89. DOI: 10.1016/1044-0305(94)80016-2 [PubMed: 24226387]

37. Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*. 2002; 2:1426–34. DOI: 10.1002/1615-9861(200210)2:10<1426::AID-PROT1426>3.0.CO;2-5 [PubMed: 12422359]
38. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*. 1994; 66:4390–9. [PubMed: 7847635]
39. Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem*. 2003; 75:1307–15. DOI: 10.1021/ac026199a [PubMed: 12659190]
40. Treuheit MJ, Costello CE, Halsall HB. Analysis of the five glycosylation sites of human alpha 1-acid glycoprotein. *Biochem J*. 1992; 283:105–12. [PubMed: 1567356]
41. Nishi K, Ono T, Nakamura T, Fukunaga N, Izumi M, Watanabe H, et al. Structural insights into differences in drug-binding selectivity between two forms of human alpha1-acid glycoprotein genetic variants, the A and F1*S forms. *J Biol Chem*. 2011; 286:14427–34. DOI: 10.1074/jbc.M110.208926 [PubMed: 21349832]
42. Rathore AS, Winkle H. Quality by design for biopharmaceuticals. *Nat Biotechnol*. 2009; 27:26–34. DOI: 10.1038/nbt0109-26 [PubMed: 19131992]

Fig. 1.

a Venn diagram showing number and overlap of proteins identified from proteomics analyses of individual serum glycoproteins and albumin and IgG depleted serum. Only proteins identified at a 1 % FDR with two or more unique peptides were included. Comparison of a naïve search space (**b**) with proteomics informed (**c**) and proteomics + glycomics informed (**d**) search spaces for glycoprotein mixture 3. *x*-axes show the Swissprot identifiers for glycoproteins included in the hypotheses and *y*-axis shows number of

glycopeptides per glycoprotein considered. Glycoprotein standards AGP (A1AG), fetuin (FETUA), and transferrin (TRFE) are shown on the *left-most side* of the *x*-axes

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 2.
Base peak chromatogram of enriched transferrin glycopeptides stacked over extracted ion chromatograms for saccharide oxonium ions, generated by MS/MS, indicating the presence of glycopeptides in the eluting peaks

Fig. 3.

Comparison of the performance of naïve versus informed search spaces for an AGP glycoproteomics result. **a** Number of glycopeptide matches over a range of confidence thresholds (q -value) based on glycoproteomics searches followed by FDR calculation by decoy database searches. *Colored lines* represent different hypotheses as indicated in the legend. **b** A comparison of number of glycopeptides and decoys matches for naïve and informed search spaces at different MS^2 score thresholds

Fig. 4. Effects of increasing sample complexity on analytical methods. **a** Base peak chromatograms (MS only) for samples with different levels of complexity. **b** Extracted ion chromatograms showing the abundance of an AGP glycopeptide (LVPVPITNATLDR-Hex₆, HexNAc₅, NeuAc₃) in the different complexity levels. **c** The average number of tandem mass spectra acquired per precursor across samples of different complexity. **d** Distributions of precursors selected for tandem MS across a LC-tandem MS experiment for samples with scaling complexity levels

Fig. 5. Grouped bar plots showing identified site-specific AGP glycoforms in samples of different complexity levels and depleted serum. The different glycoforms for each glycosylation site are listed on the *x*-axes while the *y*-axes indicate summed abundances of all fragment ions in the matched glycopeptide spectra. *Different colored bars* are specific for pure AGP, glycoprotein mixtures, or depleted serum as indicated in the legend

Table 1

Search space sizes for AGP and AGP mixed with transferrin when calculated with assumptions about proteoforms and glycoforms (naïve) and using data from proteomics and glycomics experiments

Sample	Naïve search space size	Informed search space size
Naïve pure AGP	14,577	2000
Naïve contaminated AGP (AGP + haptoglobin)	45,371	2085
Naïve AGP + transferrin (AGP + haptoglobin + transferrin)	115,910	25,850