Community Evaluation of Glycoproteomics Informatics Solutions Reveals High-Performance Search Strategies of Glycopeptide Data

Rebeca Kawahara[1], Kathirvel Alagesan[2], Marshall Bern[3], Weiqian Cao[4], Robert J Chalkley[5], Kai Cheng[6], Matthew S. Choo[7], Nathan Edwards[8,9], Radoslav Goldman[8,9,10], Marcus Hoffmann[11], Yingwei Hu[12], Yifan Huang[13], Jin Young Kim[14], Doron Kletter[3], Benoit Liquet-Weiland[15,16], Mingqi Liu[4], Yehia Mechref[13], Bo Meng[17], Sriram Neelamegham[6], Terry Nguyen-Khuong[7], Jonas Nilsson[18], Adam Pap[19,20], Gun Wook Park[14], Benjamin L. Parker[21], Cassandra L. Pegg[22], Josef M. Penninger[23,24], Toan K. Phung[22], Markus Pioch[11], Erdmann Rapp[11,25], Enes Sakalli[23], Miloslav Sanda[8,10], Benjamin L. Schulz[22], Nichollas E. Scott[26], Georgy Sofronov[15], Johannes Stadlmann[23], Sergey Y. Vakhrushev[27], Christina M. Woo[28], Hung-Yi Wu[28], Pengyuan Yang[4], Wantao Ying[17], Hui Zhang[12], Yong Zhang[17], Jingfu Zhao[13], Joseph Zaia[29], Stuart M. Haslam[30], Giuseppe Palmisano[31], Jong Shin Yoo[14,32], Göran Larson[33], Kai-Hooi Khoo[34], Katalin F. Medzihradszky[5,19], Daniel Kolarich[2], Nicolle H. Packer[1,2,35], and Morten Thaysen-Andersen[1,35]*

[1]Department of Molecular Sciences, Macquarie University, Sydney, NSW, Australia

[2]Institute for Glycomics, Griffith University Gold Coast Campus, QLD, Australia

[3]Protein Metrics Inc., Cupertino, CA, USA

[4]Institutes of Biomedical Sciences, and the NHC Key Laboratory of Glycoconjugates Research, Fudan University, Shanghai, China

[5]UCSF, School of Pharmacy, Department of Pharmaceutical Chemistry, San Francisco, CA, United States of America

[6]State University of New York, Buffalo, NY, United States of America

[7]Analytics Group, Bioprocessing Technology Institute, Agency for Science, Technology and Research, Singapore

[8]Clinical and Translational Glycoscience Research Center (CTGRC), Georgetown University, Washington, DC, United States of America

[9]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, Washington, DC, United States of America

[10]Department of Oncology, Georgetown University, Washington, DC, United States of America

[11]Max Planck Institute for Dynamics of Complex Technical Systems, Bioprocess Engineering, Magdeburg, Germany

[12]Department of Pathology, The Johns Hopkins University, Baltimore, MD, United States of America

[13]Department of Chemistry and Biochemistry, Texas Tech University, TX, United States of America

[14]Research Center of Bioconvergence Analysis, Korea Basic Science Institute, Republic of Korea

[15]Department of Mathematics and Statistics, Macquarie University, Sydney, NSW, Australia

[16]CNRS, Laboratoire de Mathématiques et de leurs Applications de PAU, E2S-UPPA, Pau, France

[17]State Key Laboratory of Proteomics, Beijing Institute of Lifeomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing, China

[18]Proteomics Core Facility, Sahlgrenska academy, University of Gothenburg, Gothenburg, Sweden

[19]BRC, Laboratory of Proteomics Research, Szeged, Hungary

[20]Doctoral School in Biology, Faculty of Science and Informatics, University of Szeged, Szeged, Hungary

[21]Department of Anatomy and Physiology, University of Melbourne, Melbourne, VIC, Australia

[22]School of Chemistry and Molecular Biosciences, University of Queensland, QLD, Australia

[23]IMBA, Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna, Austria

[24]Department of Medical Genetics, Life Sciences Institute, University of British Columbia, Vancouver, BC, Canada

[25]glyXera GmbH, Magdeburg, Germany

[26]Deparment of Microbiology and Immunology, University of Melbourne, Melbourne, VIC, Australia

[27]Copenhagen Center for Glycomics, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark

[28]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, United States of America

[29]Department of Biochemistry, Boston University Medical Campus, Boston, MA, United States of America

[30]Department of Life Sciences, Imperial College London, London, UK

[31]Instituto de Ciências Biomédicas, Departamento de Parasitologia, Universidade de São Paulo, São Paulo, SP, Brazil

[32]Graduate School of Analytical Science and Technology, Chungnam National University, Daejeon, Republic of Korea

[33]Department of Laboratory Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[34]Institute of Biological Chemistry, Academia Sinica, Taipei, Taiwan

[35]Biomolecular Discovery Research Centre, Macquarie University, Sydney, NSW, Australia

Running title: High-performance search strategies for glycoproteomics data analysis

Keywords: Glycoproteomics, glycopeptide, informatics, software, mass spectrometry

***Corresponding author**:

Dr Morten Thaysen-Andersen, PhD

Department of Molecular Sciences

Biomolecular Discovery Research Centre

Faculty of Science & Engineering

Macquarie University - Sydney

NSW-2109, Australia

Abstract

Glycoproteome profiling (glycoproteomics) remains a considerable analytical challenge that hinders rapid progress in glycobiology. The complex tandem mass spectra generated from glycopeptide mixtures require sophisticated analysis pipelines for structural determination. Diverse informatics solutions aiding the process have appeared, but their relative strengths and weaknesses remain untested. Conducted through the Human Proteome Project – Human Glycoproteomics Initiative, this community study comprising both developers and expert users of glycoproteomics software is the first to evaluate the relative performance of current informatics solutions for comprehensive glycopeptide analysis. High-quality LC-MS/MS-based glycoproteomics datasets of *N*- and *O*-glycopeptides from serum proteins were shared with all teams. The relative team performance for efficient glycopeptide data analysis was systematically established through multiple orthogonal performance tests. Excitingly, several high-performance glycoproteomics informatics solutions and tools displaying a considerable performance potential were identified. While the study illustrated that significant informatics challenges remain in the analysis of glycopeptide data as indicated by a high discrepancy between the reported glycopeptides, a substantial list of commonly reported high-confidence glycopeptides could be extracted from the team reports. Further, the team performance profiles were correlated to the many study variables, which revealed important performance-associated search settings and search output variables, some intuitive others unexpected. This study concludes that diverse informatics solutions for comprehensive glycopeptide data analysis exist within the community, points to several high-performance search strategies, and specifies key variables that may guide future software developments and assist the experimental decision-making of practitioners in glycoproteomics.

Introduction

Protein glycosylation, the attachment of complex carbohydrates (glycans) to discrete sites on proteins, plays diverse roles in biology[1]. Glycoproteomics, the system-wide analysis of intact glycopeptides, has evolved from proteomics and glycomics by the recognised benefits of concertedly studying the glycan structures, their modification sites and their protein carriers at scale within a single experiment[2, 3].

Facilitated by recent advances in separation science, mass spectrometry (MS) and informatics, glycoproteomics has matured over the past decade and is now ready to tackle biological questions and generate new insights into the complex glycoproteome in biological systems from a new angle[4-7]. While liquid chromatography tandem mass spectrometry (LC-MS/MS)-based glycoproteomics studies routinely report thousands of *N*- and *O*-linked glycopeptides[8-13], accurate and confident identification of glycopeptides from large volumes of mass spectral data remain a significant bottleneck in the field. The annotation process of glycopeptide tandem mass spectra is highly error-prone due to the challenging task of correctly assigning both the glycan composition, modification site and peptide carrier[14-16]. As a result, glycopeptides reported in glycoproteomics papers are frequently incorrectly identified or suffer from ambiguous annotation even in studies attempting to control the false discovery rate (FDR) of such assignments. Isomeric monosaccharide compositions displaying identical elemental compositions e.g. hexose (Hex, $C_6H_{10}O_5$) + *N*-acetylneuraminic acid (NeuAc, $C_{11}H_{17}NO_8$) *versus* deoxyhexose (dHex, $C_6H_{10}O_4$) and *N*-glycolylneuraminic acid (NeuGc, $C_{11}H_{17}NO_9$) (Δm = 0 Da), (un)expected peptide side chain modifications e.g. Met oxidation (16 Da) and Met alkylation (57 Da) matching monosaccharide mass differences e.g. Hex *vs* dHex (Δm = 16 Da), and dHex *vs* *N*-acetylhexosamine (HexNAc, Δm = 57 Da), and near-isobaric modifications e.g. $dHex_2$ and NeuAc (Δm = 1 Da) that are frequently misassigned as precursor

isotope differences, are examples of the informatics challenges associated with glycopeptide mass spectral assignment.

Diverse fragmentation modes including resonance-activation collision-induced dissociation (CID) performed on ion trap instruments and beam-type CID (higher-energy collisional dissociation, HCD) and electron-transfer dissociation (ETD) typically achieved with quadrupole time-of-flight (Q-TOF) and Orbitrap mass spectrometers have proven beneficial for large-scale glycopeptide analysis[17-20]. When applied in concert, these fragmentation strategies provide complementary structural information of glycopeptide analytes. In short, HCD-MS/MS informs on the peptide carrier and produces useful diagnostic glycan fragment ions often sufficient to crudely classify fragment spectra as glycopeptide tandem mass spectra and deduce generic glycan compositions, ETD-MS/MS may in favourable cases reveal the modification site and peptide identity, while ion trap CID-MS/MS provides details of the glycan sequence/composition and topology[21, 22]. Hybrid-type fragmentation strategies including ETciD and EThcD available on the Orbitrap Tribrid systems are becoming popular given their ability to generate information-rich chimeric glycopeptide spectra containing multiple types of fragments[23]. Accurate mass measurements (low mass error, often <5 ppm) at high resolution of precursor and product ions available on most contemporary instruments are essential for glycopeptide analysis at scale. Despite these exciting advances, unambiguous glycopeptide identification remains challenging and it is recognised that new innovative software solutions for precise glycopeptide identification from MS/MS data are required to ensure accurate glycoproteome profiling and reporting to further the glycoproteomics field[24].

Glycoproteomics has recently experienced the development of diverse informatics solutions, by both commercial and academic developers, showing promise for precise annotation and identification of intact glycopeptides from mass spectral data[25, 26]. While some of these computational tools are already relatively well established and have been applied in many

glycoproteomics studies[27], the relative performance, strengths and limitations of the software solutions available to the community remain untested leaving a critical knowledge gap that at present hinders rapid progress in the field.

Facilitated by the Human Proteome Project – Human Glycoproteomics Initiative, we here perform a comprehensive and level community-based evaluation of existing informatics solutions for large-scale glycopeptide analysis. While informatics challenges undoubtedly still exist in glycoproteomics, our study highlights that a diversity of software solutions, some already demonstrating high performance, others a considerable potential, are now available to the community. Importantly, our study has identified several key search engines, search settings and search strategies that may aid software developers and practitioners alike to improve the challenging glycoproteomics data analysis in the immediate future.

Results

Study design and overview

Two high-resolution LC-MS/MS data files (hereafter File A and B) of tryptic glycopeptides from human serum were acquired on an Orbitrap Fusion Lumos Tribrid mass spectrometer (**Figure 1a**). A synthetic *N*-glycopeptide (EVFVHPNYSK, $Hex_5HexNAc_4NeuAc_2$) from human vitamin K-dependent protein C was included as a positive control. File A and B were generated using HCD-MS/MS fragment ion-triggered-ETciD-CID-MS/MS and -EThcD-CID-MS/MS, respectively, in attempts to generate data accommodating most informatics solutions in the field. Serum is a well-characterised biospecimen displaying high molecular complexity and an abundance of both *N*- and *O*-glycoproteins[28-30]. Thus, File A and B displayed characteristics (file size, complexity and type) similar to data files typically encountered in glycoproteomics[31-34].

File A and B were shared with all 22 participating teams who classified themselves as either developers (9 teams) or expert users (13 teams) of glycoproteomics software (**Figure 1b** and **Supplementary Figure S1a-b**). All teams were asked to identify intact *N*- and *O*-glycopeptides from the shared data files and then report the identifications and their approaches in a common reporting template. Most developers (five teams) and expert users (eight teams) had significant experience in glycoproteomics (>10 years). The participants were from North America, Europe, Asia, Oceania, and South America (**Supplementary Figure S1c-d**).

Only File B was processed by all teams (**Supplementary Figure S1e**). While most participants reported on spectra acquired with multiple fragmentation methods, a few teams used only HCD- or EThcD-MS/MS for the glycopeptide identification (**Supplementary Figure S1f-g**).

A diversity of search engines was used in the study (**Supplementary Figure S1h**). Some search engines were used as stand-alone tools or in combinations with other search engines while others were applied with pre- or post-processing tools to aid the glycopeptide identification process (**Supplementary Figure S1i**).

File A and B contained a total of 8,737 and 9,776 HCD-MS/MS scans, respectively, of which 5,485 (62.8%) and 6,148 (62.9%) fragment spectra contained highly specific glycopeptide oxonium ions (*m/z* 138, 204 and 366) used to trigger the subsequent ETciD-, EThcD- and CID-MS/MS events (**Supplementary Figure S2a-b**). Within the entire set of potential glycopeptide MS/MS spectra (File A/B: 16,445/18,444, considering all fragmentation modes), only a total of 3,402 (20.7%) and 4,982 (27.0%) non-redundant (unique) glycopeptide-to-spectrum matches (glyco-PSMs) were collectively reported for File A and B by the participants. The teams primarily reported on spectra obtained using HCD-MS/MS (File A/B: 47.8%/44.5%) and EThcD-MS/MS (File B: 34.1%), while only relatively few teams used high- or low-resolution CID-MS/MS (File A/B: 9.5%/2.5%, respectively), or ETciD-MS/MS data (File A: 4.8%). Similar charge distribution was observed for the glycopeptides reported from the

different fragmentation modes (**Supplementary Figure S2c-d**). Glycopeptides carrying four charges were most frequently reported.

In total, 11 search engines were used for the core glycopeptide identification process (**Figure 1b**). The developers used 9 different glycopeptide-centric search engines including Team 1: IQ-GPA v2.5[35], Team 2: Protein Prospector v5.20.23[36], Team 3: glyXtool^{MS} v0.1.4[37], Team 4: Byonic v2.16.16[38], Team 5: Sugar Qb[39], Team 6: Glycopeptide Search v2.0alpha[40], Team 7: GlycopeptideGraphMS v1.0[41]/Byonic[38], Team 8: GlycoPAT v2.0[42], Team 9: GPQuest v2.0[43]. Amongst the 13 expert users, 10 teams (~75%) used Byonic (Team 10, 11, 13, 15-18, 20-22), while only a single team used Protein Prospector (Team 12), SugarQB/Sequest HT (Team 14), and Mascot (Team 19) (**Figure 1b** and **Supplementary Table S1**).

More than 100 different types of data were collected via the comprehensive reporting template. The reports, which were carefully checked for compliance to the study guidelines, covered details of the team, identification strategy, and identified glycopeptides (**Figure 1c**). Amongst many other details, information of the search settings was captured including the protease specificity, permitted peptide modifications, mass tolerance, post-search filtering criteria (**Supplementary Table S1**), and the applied glycan search space (**Supplementary Table S2**). Most of the 13 key search settings (SS1-SS13, **Table 1**) varied considerably across teams.

In addition, 37 types of output data arising from the glycopeptide identification process spanning both numerical and categorical variables were captured via the reporting template or were manually extracted from the reports (**Supplementary Table S3**). Data for the numerical output variables were compiled and compared (**Supplementary Table S4**). Analysis of 9 key search output variables (SO1-SO9) e.g. observed glycopeptide LC retention time, *m/z*, actual mass error, charge state, glycan mass, and peptide length (**Table 1**) from File B (processed by all teams) revealed that the reported *N*- and *O*-glycopeptides, as expected, showed different properties including different LC retention times (higher for *N*-glycopeptides), charge states

(higher for *O*-glycopeptides), actual mass error (lower for *N*-glycopeptides), peptide length (longer *O*-glycopeptides) and glycan mass (higher for *N*-glycopeptides) while other properties including the observed precursor *m/z* amongst other traits did not differ between the reported *N*- and *O*-glycopeptides (**Supplementary Figure S3**). Analysis of the SO1-SO9 data also demonstrated that some teams employed search strategies that led to highly discrepant output. For example, and without being able to link these observations directly to performance, the developers of Glycopeptide Search (Team 6) and GlycopeptideGraph (Team 7) reported glycopeptides with unusually low (average z = ~3+) and high (~5.5+) charge states relative to all other teams (~4.5+), **Supplementary Figure S3c**. These comparisons of output data may be valuable for the developer teams to better understand and eventually improve their software.

The relative team performance was assessed using 11 independent performance tests that served to comprehensively evaluate the identification accuracy (specificity) of the reported glycopeptides (glycan composition and source glycoprotein) and the glycoproteome coverage (sensitivity), two key performance characteristics in glycoproteomics (**Figure 1d**). Specifically, six (N1-N6) and five (O1-O5) independent performance tests were carefully designed and applied to assess the relative performance for *N*- and *O*-glycopeptide data analysis for all teams (**Table 2** and **Supplementary Table S5-S15**). Firstly, the ability to detect the synthetic *N*-glycopeptide spiked into the samples was assessed across teams (N1). Further, the glycan compositions (N2, N6, O1, O5) and source glycoproteins (N3, O2) of the reported glycopeptides were compared to the previously established serum glycome and against known serum glycoproteins[4, 28, 44-47]. In another test, the glycoproteome coverage was assessed by comparing the number of non-redundant glyco-PSMs (unique peptide sequence and glycan composition) reported by each team (N4, O3). Finally, the ability of each team to identify glycopeptides commonly reported by most participants ("consensus glycopeptides") was assessed (N5, O4). The performance scores were separately compared amongst the developer

and user teams (not between the two groups since they were given slightly different tasks) and correlated to the search settings (SS1-SS13) and search outputs (SO1-SO9) using seven advanced statistical methods to uncover software solutions and search variables for high-performance glycopeptide data analysis (**Figure 1e** and **Supplementary Table S16-S17**).

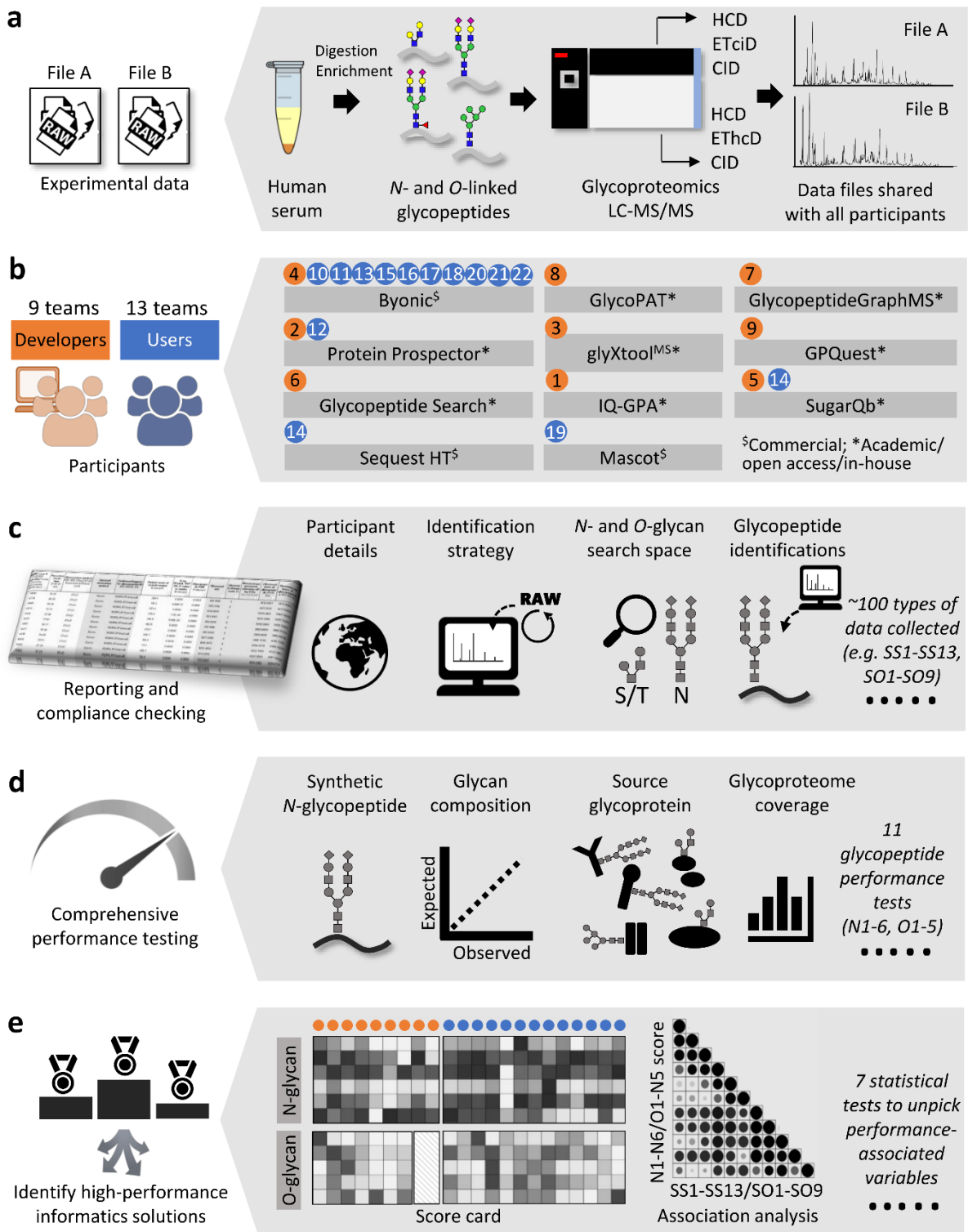

Figure 1. Study overview. **a**. Two glycoproteomics datasets of human serum (File A and B) were shared with all participants. **b**. The 22 participating teams comprised 9 developers (orange) and 13 expert users (blue, team identifiers indicated). Diverse search engines were used to complete the study. **c**. Following the glycopeptide identification process, teams completed a common reporting template capturing details of the team and their identification strategy including the search settings (SS1-SS13) and search output variables (SO1-SO9), **Table 1**. Reports were carefully checked for compliance with the study guidelines. **d**. Complementary performance tests were used to evaluate the ability of each team to identify *N*-glycopeptides (N1-N6) and *O*-glycopeptides (O1-O5), see **Table 2** for performance tests. **e**. The performance profiles were compared across the developers and expert users (not between the two groups) and used to rank teams. The performance scores were also tested for correlation to the search settings and search output to identify performance-associated variables.

Table 1. Overview of important study variables including key search settings (SS1-SS13) and search output variables (SO1-SO9). *Num, numerical variable; Cat, categorical variable. NG, *N*-glycosylation; OG, *O*-glycosylation. #Average of output data reported by each team. AA, amino acid residues. See **Supplementary Table S1-S4** for details. The reported glycopeptides waere not included as a search output; this search output was instead used to score the glycoproteome coverage (see performance test N4 and O3).

Search setting variables		**Type ***	**Range or definition of category** (counts of teams)
SS1	*N*-glycan search space	Num	23 - 381 unique glycan compositions
SS2	*O*-glycan search space	Num	3 - 223 unique glycan compositions
SS3	Search engine(s) applied	Cat	Byonic (11), Protein Prospector (2), GlycoPAT (1), GlycopeptideGraph (1), glyXtool^{MS} (1), GPQuest (1), Other (5)
SS4	Type of search engine	Cat	Academic/open access/in-house = 0 (7), Commercial = 1 (15)
SS5	Spectral calibration post-acquisition?	Cat	No = 0 (18), Yes = 1 (4)
SS6	Protease specificity	Cat	Non-tryptic (*N*- or *O*-ragged or non-specific) = 0 (8), Tryptic = -1 (14)
SS7	Missed cleavages permitted	Num	0 - 2 missed cleavages
SS8	Variable peptide modification(s) (non-glycan)	Num	0 - 14 non-glycan modification types
SS9	Maximum glycans per peptide	Num	1 - 5 glycans/peptide
SS10	Maximum other variable modifications	Num	0 - 5 variable modifications/peptide
SS11	Precursor ion mass error permitted	Cat	Low (<5 ppm) = 0 (6), Medium (5-10 ppm) = 1 (14), High (>10 ppm) = 2 (2)
SS12	Product ion mass error permitted	Cat	Low (<5 ppm) = 0 (0), Medium (5-10 ppm) = 1 (9), High (>10 ppm) = 2 (13)
SS13	Peptide/protein FDR (use of decoys/contaminant database?)	Cat	No decoy/contaminant = 0 (3), Only decoy or contaminant = 1 (9), Both decoy and contaminant = 2 (10)

Search output variables		Type *	Range (*N*- or *O*-glycosylation)
SO1	Glycopeptide LC retention time#	Num	41.7 - 57.2 min (NG) 26.2 - 55.8 min (OG)
SO2	Glycopeptide *m/z* (observed)#	Num	*m/z* 712.9 - 1199.4 (NG) *m/z* 619.6 – 1229.0 (OG)
SO3	Glycopeptide charge state (observed)#	Num	z = 3.9 - 4.3 (NG) z = 3.1 - 5.4 (OG)
SO4	Glycopeptide monoisotopic precursor selection off by X (positive values)#	Num	0 - 2.1 Da (NG) 0 - 2.0 Da (OG)
SO5	Glycopeptide mass ([M+H], observed)#	Num	3144.6 - 4913.0 Da (NG) 1892.9 - 6057.1 Da (OG)
SO6	Glycopeptide actual mass error (observed, positive values)#	Num	0.5 - 2.8 ppm (NG) 1.1 - 5.9 ppm (OG)
SO7	Glycopeptide length#	Num	16.9 - 26.8 AA (NG) 14.5 - 38.9 AA (OG)
SO8	Glycan mass (calculated from reported glycopeptides)#	Num	1880.8 - 2410.1 Da (NG) 195.8 - 2216.8 Da (OG)
SO9	Total number of reported glycopeptides	Num	49 - 2122 glyco-PSMs (NG) 5 - 578 glyco-PSMs (OG)

Table 2. Overview of the independent performance tests applied to establish the relative team performance for *N*- and *O*-glycopeptide analysis. Each performance test was used to score each team using normalised quantitative values (range 0-1). *The scoring of some tests relied on robust literature on human serum glycosylation i.e. Clerc et al., 2016[28], Sun et al., 2018[44], Yabu et al., 2014[45], Darula et al., 2016[46], Yang et al., 2018[47] and Ye et al., 2019[4].

Performance tests		Description of scoring
Tests relevant for the N-glycopeptide analysis		
N1	Synthetic *N*-glycopeptide	Average of accuracy (specificity, %) and coverage (sensitivity, %) of the identification of the synthetic glycopeptide (**Supplementary Table S5**)
N2	*N*-glycan composition*	Pearson correlation (R^2) between expected (based on Clerc et al., 2016[28]) and observed glycan distribution in human serum (**Supplementary Table S6)**
N3	Source *N*-glycoprotein*	Specificity and sensitivity of reported source glycoproteins relative to expected serum glycoproteins (based on Clerc et al., 2016[28], Sun et al., 2018[44]) (**Supplementary Table S7**)
N4	*N*-glycoproteome coverage	Number of unique glycopeptides (unique peptide sequence + glycan composition) (**Supplementary Table S8**)
N5	Commonly reported 'consensus' *N*-glycopeptides	Proportion of reported glycopeptides out of the consensus glycopeptides commonly reported by most teams (>50% of the 22 teams) (**Supplementary Table S9**)
N6	NeuGc and multi-Fuc *N*-glycopeptides	Average of reported non-NeuGc and non-Fuc ≥2 containing glyco-PSMs out of total PSMs. Only applicable for teams using NeuGc and Fuc ≥2 glycans in glycan search space (**Supplementary Table S10**)

Tests relevant for the O-glycopeptide analysis		
O1	*O*-glycan composition*	Pearson correlation (R^2) between expected (based on Yabu et al., 2014[45]) and observed glycan distribution in human serum (**Supplementary Table S11**)
O2	Source *O*-glycoprotein*	Specificity and sensitivity of reported source glycoproteins relative to expected serum glycoproteins (based on Darula et al. 2016[46], Yang et al., 2018[47] and Ye et al., 2019[4] (**Supplementary Table S12**)
O3	*O*-glycoproteome coverage	Number of unique glycopeptides (unique peptide sequence + glycan composition) (**Supplementary Table S13**)
O4	Commonly reported 'consensus' *O*-glycopeptides	Proportion of reported glycopeptides out of the consensus glycopeptides commonly reported by most teams (>30% of the 20 teams) (**Supplementary Table S14**)
O5	NeuGc and multi-Fuc *O*-glycopeptides	Average of reported non-NeuGc and non-Fuc ≥2 containing glyco-PSMs out of total. Only applicable if NeuGc and Fuc ≥2 containing glycans were included in glycan search space (**Supplementary Table S15**)

Overview of the reported glycopeptides

The below analyses were carried out using data reported from File B processed by all teams. The total *N*-glyco-PSMs reported by the 22 teams ranged from 49 to 2,122 covering between 9 and 168 source glycoproteins (**Figure 2a** and **Supplementary Table S3**). In line with literature of human serum *N*-glycosylation[28, 29], the reported *N*-glycopeptides carried mainly complex-type *N*-glycans (92.6%, average of all teams) while only relatively few were reported to carry oligomannosidic (6.4%) and truncated ($Hex_{<4}HexNAc_{<4}$) (1.0%) *N*-glycans. The applied *N*-glycan search space (the *N*-glycan database that was searched against) spanned an equally wide range from 23 to 381 glycan compositions comprising mostly complex-type *N*-glycans (89.1%, average of all teams) and the less heterogeneous oligomannosidic (5.9%) and truncated (5.0%) *N*-glycans. Despite being perceived as an important search variable in glycoproteomics, only weak associations were found between the size of the *N*-glycan search space and the reported *N*-glyco-PSMs (Pearson R = 0.340). Unexpected glycan compositions including NeuGc- and multi-Fuc-containing complex-type *N*-glycans, which are absent or negligible features of human serum glycoproteins[28,48-50], were not only included in the glycan

search space (up to 26.5% and 28.9%, respectively), but also reported (up to 20.6% and 5.0%, respectively) by some teams. The absence of NeuGc glycopeptides in the shared MS/MS data was supported by a lack of diagnostic fragment ions for NeuGc (*m/z* 290/308) (data not shown). Thus, glycopeptides reported with NeuGc- and multi-Fuc-containing glycans were considered as incorrect identifications in this study.

Collectively, 2,556 unique (non-redundant) *N*-glycopeptides (defined herein as unique peptide sequences and glycan compositions), covering 320 different source *N*-glycoproteins and spanning 424 different *N*-glycan compositions were reported across all teams (**Figure 2b**, **Supplementary Table S6-S7** and **Supplementary Table S9**). Of these, only 43 *N*-glycopeptides (1.7%), 26 source *N*-glycoproteins (8.1%) and 28 *N*-glycan compositions (6.6%) were commonly reported by at least 75% of the teams. Most identifications were reported by less than 25% of the teams. Slightly less discordance was observed when the identifications were compared amongst the teams within the developer or expert user groups. Within the 10 teams using the Byonic search engine, 157 of 1,867 *N*-glycopeptides (8.4%) were commonly reported by at least 75% of the teams, but most identifications were still reported by less than 25% of the teams.

Significantly fewer *O*-glycopeptides were identified by the participants ranging from 5 to 578 *O*-glyco-PSMs (**Figure 2c** and **Supplementary Figure S3j**). Most *O*-glycopeptides were, as expected, reported to carry $Hex_1HexNAc_1NeuAc_1$ and $Hex_1HexNAc_1NeuAc_2$[45, 51]. The *O*-glycan search space varied from 3 to 223 glycan compositions. Similar to the *N*-glycopeptide analysis, only weak associations were observed between the size of the *O*-glycan search space and the reported *O*-glyco-PSMs (Pearson R = 0.344). Instead, many other types of associations were identified (discussed below). While seven teams included NeuGc in the applied *O*-glycan search space (up to 9.0%), only four teams reported NeuGc *O*-glycopeptides (up to 7.3%). In addition, 12 teams included multi-fucosylated glycans in the *O*-glycan search space (up to

43.5%); 11 of those teams reported multi-fucosylated *O*-glycopeptides (average of 28.6%, up to 61.8%). Both NeuGc and multi-fucosylated *O*-glycans are absent or negligible features of human serum *O*-glycoproteins as supported by literature[45, 51] and by the absence of NeuGc diagnostic fragment ions in the analysed MS/MS data as mentioned earlier. In principal, the reported multi-fucosylated *O*-glycan compositions could in some cases arise from discrete *O*-glycan sites on the same peptide; however, since *O*-glycosylation sites were inconsistently and/or ambiguously reported between teams (see below) we were not able to assess this aspect further.

Collectively, 1,192 unique *O*-glycopeptides covering 231 unique source *O*-glycoproteins and 288 unique *O*-glycan compositions were identified, but only relatively few *O*-glycopeptides were commonly reported by the teams. Only 3 *O*-glycopeptides (0.3%), 6 source *O*-glycoproteins (2.6%) and 7 *O*-glycan compositions (2.4%) were commonly reported by at least 50% of the teams (**Figure 2c**). The identification rate of these commonly reported 'consensus' glycopeptides was marginally higher when assessed within specific subsets of teams. For example, 20 of 674 *O*-glycopeptides (3.0%) were reported by at least 50% of the 10 Byonic users; however, most identifications were still reported by less than 25% of those teams.

Despite the discrepant reporting by teams, a high-confidence list spanning 163 *N*- and 23 *O*-glycopeptides commonly reported by more than 50% and 30% of teams, respectively, could be extracted from the team reports. Importantly, these commonly assigned consensus glycopeptides mapped to expected serum glycoproteins e.g. α-2-macroglobulin (UniProtKB, P01023) and immunoglobulin heavy constant α1 (UniProtKB, P01876) and carried expected serum *N*-glycans e.g. $Hex_5HexNAc_4NeuAc_2$ (GlyTouCan ID, G09675DY) and $Hex_5HexNAc_4Fuc_1NeuAc_2$ (GlyTouCan ID, G22754FQ) and *O*-glycans e.g. $Hex_1HexNAc_1NeuAc_1$ (GlyTouCan ID, G65285QO) and $Hex_1HexNAc_1NeuAc_2$ (GlyTouCan ID, G84906ML). Further, the *N*- and *O*-glycans carried by these consensus glycopeptides were

biosynthetically related (**Supplementary Figure S4**) and were devoid of NeuGc and multi-Fuc features further supporting their correct identification. These high-confidence glycopeptides form an important reference to future studies of the human serum glycoproteome and have therefore been made publicly available (GlyConnect Reference ID 2943).

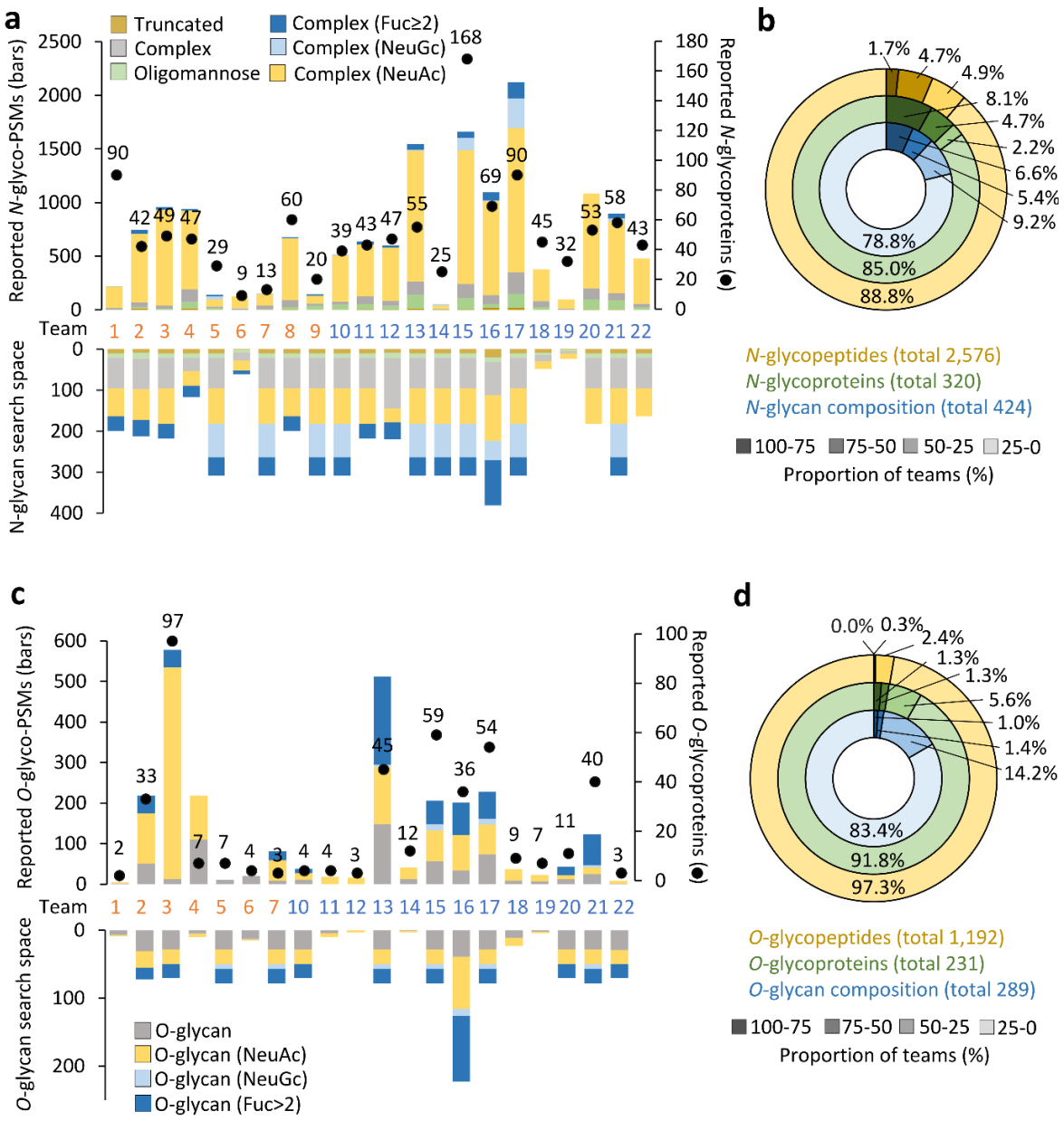

Figure 2. Overview of the *N*- and *O*-glycopeptide identifications reported by the 22 teams completing the study. Team 8 and 9 did not partake in the *O*-glycopeptide challenge. **a.** Distribution of the reported *N*-glyco-PSMs (bars), unique source *N*-glycoproteins (dots) and the applied *N*-glycan search space (mirror, see also **Supplementary Table S2-S3**) across all teams. **b.** The proportion of common *N*-glycopeptides, source *N*-glycoproteins and *N*-glycan

compositions reported across teams. **c.** Distribution of the reported *O*-glyco-PSMs (bars), unique source *O*-glycoproteins (dots) and the applied *O*-glycan search space (mirror, see also **Supplementary Table S2-S3**) across teams. **d.** The proportion of common *O*-glycopeptides, source *O*-glycoproteins and *O*-glycan compositions reported across teams. The high-confidence 'consensus' glycopeptides reported by more than 50% and 30% of all participants i.e. 163 *N*- and 23 *O*-glycopeptides, respectively, have been made publicly available (GlyConnect Reference ID 2943).

Identification of high-performance informatics solutions for N-glycoproteomics

The relative team performance for accurate and comprehensive *N*-glycopeptide identification was established using six independent performance tests (N1-N6) (**Table 2** and **Supplementary Table S5-S10**). Amongst these performance tests, N1 scored the ability of teams to accurately identify the synthetic *N*-glycopeptide in the sample. Similar to the other performance tests, the synthetic *N*-glycopeptide test was used to establish the relative team performance (**Supplementary Figure S5**). Data from N1 also confirmed that glycopeptides were preferentially identified in charge 4+ using HCD- and EThcD-MS/MS (even when high-quality MS/MS data from other charge states and fragmentation modes were available) as supported by observations made across the entire dataset (**Supplementary Figure S2**).

In line with the literature[2, 13], most teams employed HCD- and/or EThcD-MS/MS for the glycopeptide identification. While these two fragmentation modes surprisingly displayed similar performance in the *N*- and *O*-glycan composition tests (N2, O1) and similar *N*- and *O*-glycoproteome coverage (N4, O3), higher scores were achieved for identifications based on EThcD- relative to HCD-MS/MS data in the tests addressing the source *N*-glycoproteins (N3) and *O*-glycoproteins (O2, albeit without reaching statistical significance) (**Supplementary Figure 6**). Importantly, glycosylation site localisation, not tested with this study (discussed below), is widely recognised to be more accurately achieved with EThcD-MS/MS data[5, 17].

An elaborate scorecard used to rank the developer and expert user teams was established based on the performance tests (**Figure 3a**). At a glance, the scorecard pointed to considerable team-

to-team variations in the performance profiles suggesting that the applied software and search strategies exhibit markedly different strengths and weaknesses for *N*-glycopeptide data analysis. As an example, the developers of IQ-GPA (Team 1) and GlycoPAT (Team 8) performed well (relative to other developer teams) in the *N*-glycan composition test (N2), while Protein Prospector (Team 2) and Byonic (Team 4) performed well in tests scoring the ability to identify the source *N*-glycoproteins (N3) and the *N*-glycoproteome coverage (N4).

Overall, Protein Prospector (Team 2, score 0.682), Byonic (Team 4, score 0.676) and GlycoPAT (Team 8, score 0.647) were found to be high-performance software solutions for *N*-glycopeptide data analysis herein defined as top third of performers. Notably, the scoring did not separate these three developer teams by any significant margin, but their overall scores were slightly higher than the performance of IQ-GPA (Team 1, score 0.621) and glyXtool^{MS} (Team 3, score 0.580) and significantly higher than the other software (score range 0.296 - 0.366). In support of this ranking, the four best performing expert user teams were Team 15, 13 and 11 (Byonic users, score range 0.687 - 0.777) and Team 12 (Protein Prospector user, score 0.709). Their overall performance scores were marginally higher than the performance of seven other Byonic user teams (Team 10, 16, 17, 18, 20, 21, 22, score range 0.593 - 0.679), but significantly higher than teams using SugarQb (Team 14, score 0.172) and Mascot (Team 19, score 0.413). Despite the similar overall performance amongst most user teams, not least the 10 Byonic users, their performance profiles differed markedly across the performance tests.

We then explored the scores for performance-associated variables including the search settings (SS1-SS13) and search output variables (SO1-9) using seven different statistical methods (**Figure 3b**). Many statistically strong relationships were found revealing key performance-associated variables, some intuitive others rather unexpected, that either positively or negatively correlated with the glycopeptide identification efficiency. As an example, the use of a decoy and/or contaminant database (SS13) was found to be associated with both a high *N*-

glycoproteome coverage (N4) and high performance in the synthetic glycopeptide test (N1). Search strategies that allowed for a relatively high diversity and number of non-glycan variable peptide modifications (SS8 and SS10) and a low number of glycans per peptide (SS9) were also associated with a high *N*-glycoproteome coverage (N4). As expected, allowing a relatively high number of missed peptide cleavages (SS7) and variable non-glycan modifications (SS10) in the search strategy correlated with higher glycopeptide FDRs as indicated by a high rate of NeuGc and multi-Fuc identifications (low N6 scores) (**Supplementary Table S16-17**).

The association analyses also identified many interesting relationships between the search output variables and performance (**Figure 3b**). Intuitively, teams that reported a high number of *N*-glyco-PSMs (SO9) performed well in the synthetic *N*-glycopeptide test (N1), had a higher *N*-glycoproteome coverage (N4) and identified more consensus *N*-glycopeptides (N5). Further, teams that reported glycopeptides featuring a relatively high glycan mass (SO8) were found to more often identify the correct glycan composition (N2) while teams that reported glycopeptides exhibiting relatively high molecular masses (SO5) more often identified the correct source *N*-glycoproteins (N3). Glycopeptide analytes with relatively high molecular masses (arising from large glycans and/or peptide moieties) are less likely to result in false positive identifications due to fewer theoretical glycopeptide candidates (fewer potential false positives) in the higher mass range. In addition, early LC retention time (SO1), high charge state (SO3), high glycopeptide mass (SO5), high actual mass error (SO6) and high glycan mass (SO8) were search output variables closely linked to high *N*-glycoproteome coverage (N4). Teams reporting glycopeptides with high glycopeptide masses (SO5) were more likely to identify consensus glycopeptides (N5). Finally, low actual mass error (high mass accuracy, SO6) was, as expected, associated with a low identification rate of incorrect NeuGc and multi-Fuc glycopeptides. In summary, we identified many performance-associated variables useful to guide the *N*-glycopeptide data analysis process in the future (**Table 3**).

Identification of high-performance informatics solutions for O-glycoproteomics

Protein Prospector (Team 2) displayed the highest performance in tests scoring the ability to accurately identify the source *O*-glycoproteins (O2) and to identify consensus *O*-glycopeptides (O4). On the other hand, IQ-GPA (Team 1) and glyXtool^{MS} (Team 3) were the best performing software in tests scoring the ability to accurately identify the *O*-glycan compositions (O1) and the *O*-glycoproteome coverage (O3), respectively. Overall, Protein Prospector (Team 2, score 0.613) and glyXtool^{MS} (Team 3, score 0.522) were evaluated to be high-performance software for *O*-glycoproteomics (top third performance). Amongst the expert users, Team 13, 15, 17 and 18 (all Byonic users, score range 0.473-0.701) were found to be in the high-performance band (**Figure 3c** and **Supplementary Table S16-17**).

Association analyses showed that accurate identification of the *O*-glycan compositions (O1) was linked to the use of a focused (narrow) *O*-glycan search space (SS2) and to permitting only few missed peptide cleavages (SS7) (**Figure 3d**). In addition, search strategies permitting incorrect precursor selection (SO4) and filtering of *O*-glycopeptides identified with a low actual mass error (SO6) were common approaches used by teams performing well in the *O*-glycan composition test (O1). Interestingly, the use of a broad *O*-glycan search space (SS2) was associated with accurate identification of source *O*-glycoproteins (O2), high *O*-glycoproteome coverage (O3), and a high identification rate of consensus *O*-glycopeptides (O4). Further, teams reporting identifications with low actual mass error (SO6) scored well in the *O*-glycan composition test (O1), but, notably, at the cost of a lower *O*-glycoproteome coverage (O3) and fewer consensus *O*-glycopeptides (O4). In summary, this study has identified several performance-associated variables, including many previously unknown relationships, that may enhance the *O*-glycopeptide analysis process in the future (**Table 3**).

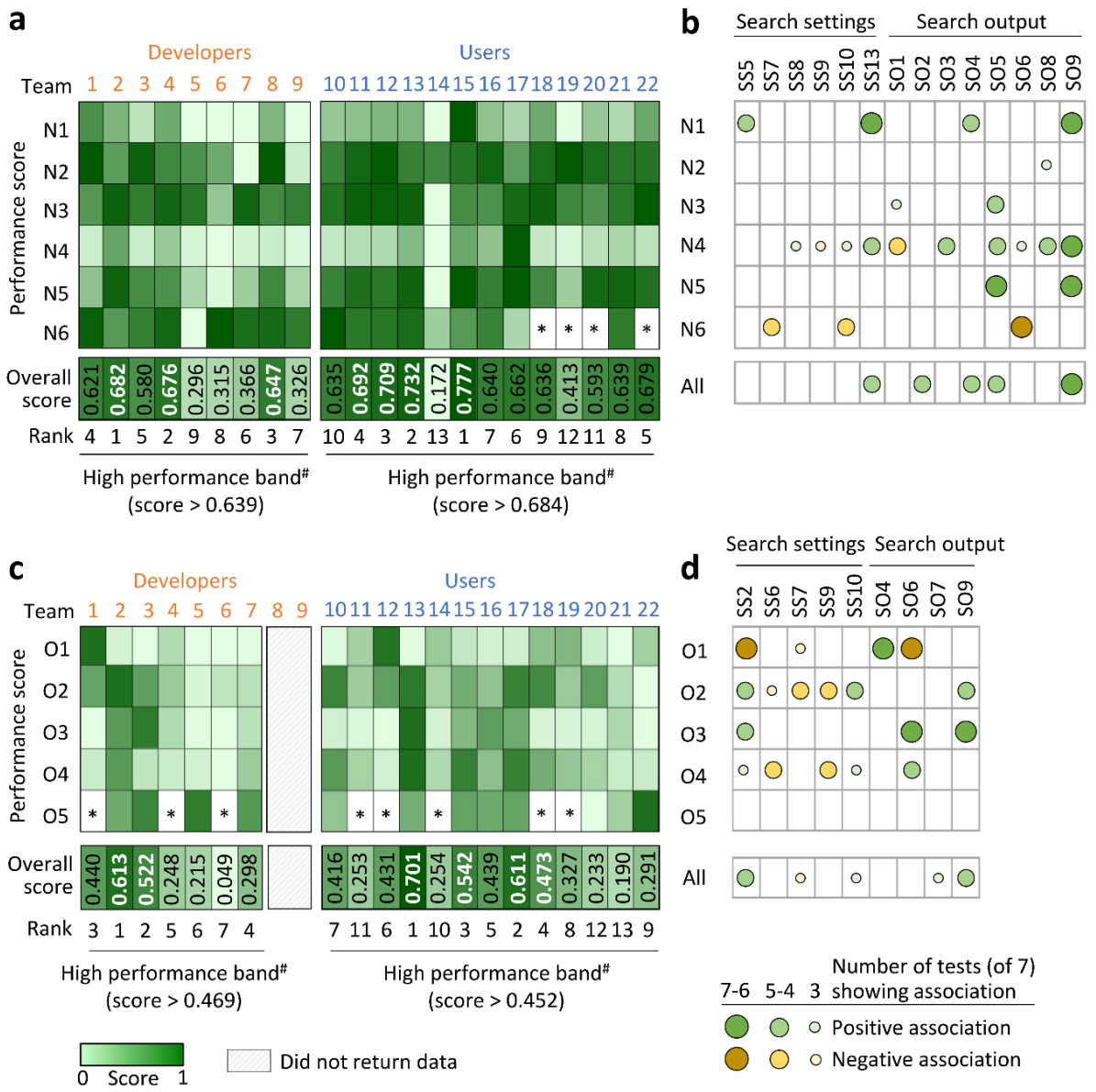

Figure 3. Team performance and identification of performance-associated variables (i.e. search settings and search outputs) as evaluated using File B. **a.** Heatmap representation of the normalised scores (range 0-1) arising from the six performance tests for *N*-glycopeptide identification (N1-N6) (see **Table 3**). **b.** Overview of search settings and search outputs that showed significant association (negative or positive) with the performance scores for *N*-glycopeptide identification as supported by multiple statistical analyses. **c.** Heatmap representation of the normalised scores (range 0-1) arising from the five performance tests for *O*-glycopeptide identification (O1-O5) (Table 3). Team 8 and 9 did not partake in the *O*-glycopeptide challenge. **d.** Overview of search settings and search outputs that showed significant associations with the performance scores for *O*-glycopeptide identification. See **Table 1** for search settings and search outputs. See **Supplementary Table S17** for statistical data. See **Supplementary Table S5-S15** for performance data. *Performance score could not be determined for these specific teams. #The top third performance scores (white bold) were grouped into a high-performance band.

Table 3. Search settings and expected search output features associated with high-performance glycoproteomics data analysis as supported by multiple association analyses (see **Figure 3b, 3d, Table 1-2** and **Supplementary Table S17** for details of study variables and reported associations). ‡The search settings outlined here may guide improved search strategies in various performance areas of a glycoproteomics experiment. ^These search output features may enhance the quality of the glycoproteomics data analysis by enabling informed post-search filtering of glycopeptide data. Only high-performance search settings and search outputs closely associated with each performance test (correlation observed for at least three of seven statistical tests) have been included in this table. The possible compromise of selected search strategies on the overall glycoproteomics performance is indicated. NA, not applicable.

Performance area	**Related test**	**High-performance search settings‡**	**High-performance search output (expected)^**	**Strategy may compromise**
Efficient *N*-glycopeptide analysis (all-round performance)	Overall score (N1-N6)	• Use decoy and/or contaminant protein database to establish peptide/protein FDR (SS13)	• High *m/z* (SO2) • High monoisotopic precursor selection off by X (SO4) • High glycopeptide mass (SO5) • High number of glyco-PSMs (SO9)	NA
Accurate *N*-glycan identification	N2	NA	• High glycan mass (Da) (SO8)	*N*-glyco-proteome coverage
Accurate identification of source *N*-glycoprotein	N3	NA	• Late LC retention time (SO1) • High glycopeptide mass (SO5)	*N*-glyco-proteome coverage
High *N*-glycoproteome coverage	N4	• Allow diversity of variable non-glycan peptide modifications (SS8) • Few glycans per peptide allowed (SS9) • Allow a higher number of variable non-glycan modifications per peptide (SS10) • Use decoy and/or contaminant protein database to establish peptide/protein FDR (SS13)	• Late LC retention time (SO1) • High charge stage (SO3) • High glycopeptide mass (SO5) • High actual mass error (SO6) • High glycan mass (SO8) • High number of glyco-PSMs (SO9)	*N*-glycopeptide identification accuracy, search time
Minimise the NeuGc and multi-Fuc FDR	N6	• Allow fewer missed cleavages of peptides (SS7)	• Low actual mass error (SO6)	*N*-glyco-proteome coverage

		• Allow fewer variable non-glycan modifications per peptide (SS10)		
Efficient *O*-glycopeptide analysis (all-round performance)	Overall score (O1-O5)	• Broad glycan database (SS2) • Few missed cleavages allowed (SS7) • Allow a higher number of variable non-glycan modifications per peptide (SS10)	• High peptide length (SO7) • High number of glyco-PSMs (SO9)	Search time
Accurate *O*-glycan identification	O1	• Use a focused glycan database (SS2) • Allow fewer missed cleavages per peptide (SS7)	• High monoisotopic precursor selection off by X (SO4) • Low actual mass error (SO6)	*O*-glyco-proteome coverage
Accuracy identification of source *O*-glycoprotein	O2	• Broad glycan database (SS2) • Trypsin fully specific (SS6) • Few missed cleavages allowed (SS7) • Few glycans per peptide allowed (SS9) • Allow a higher number of variable non-glycan modifications per peptide (SS10)	• High number of glyco-PSMs (SO9)	*O*-glyco-proteome coverage, search time
High *O*-glycoproteome coverage	O3	• Broad glycan database (SS2)	• High actual mass error (SO6) • High number of glyco-PSMs (SO9)	*O*-glycopeptide identification accuracy, search time

Discussion

This is the first comparative study to objectively discern the strength and weaknesses of current informatics solutions for large-scale glycopeptide data analysis. Excitingly, several high-performance glycoproteomics software and search strategies were identified. Amongst the 9 developer teams completing this study, Protein Prospector (Team 2) was identified as the top performing informatics solution for both *N*- and *O*-glycopeptide data analysis. Byonic (Team 4) also displayed high-performance for *N*-glycopeptide data analysis, and while this developer team only demonstrated moderate performance for *O*-glycoproteomics, four Byonic user teams (Team 13, 15, 17, 18) displayed the highest performance for *O*-glycopeptide data analysis. Protein Prospector[52] and Byonic[38], developed 10-20 years ago, have pioneered the glycopeptide informatics field and are search engines already commonly used in glycoproteomics and in studies of other post-translational modifications[13, 36, 38]. Protein Prospector is an academic/open-source tool recognised for its ability to identify modified peptides and their modification site(s) from LC-MS/MS data using a probability-difference based scoring system[36]. In fact, Protein Prospector is often a preferred software for the challenging site annotation of intact *O*-glycopeptides in particular when ET(hc)D-MS/MS data are available[5]. However, Protein Prospector does not estimate the FDR of the glycan component of glycopeptides which is regarded by the software as a non-descript PTM with an exact mass, and the software may appear less user-friendly than competing tools. Facilitated by a user-friendly and intuitive interface, precise spectral annotation and useful output reports of identified glycoproteins and glycopeptides, the commercial Byonic search engine has rapidly gained popularity (as illustrated by this study) as it enables relatively straightforward identification of peptides with known and unknown modifications including glycosylation from different types of MS/MS data. Byonic features useful fine control options that enable tailored glycopeptide searches and post-search filtering of search output based on prior

knowledge. Byonic scores and annotates multiple types of glycopeptide fragments in attempts to deduce the peptide carrier, glycan and modification site, but the FDR is treated identically for non-glycosylated peptides and glycopeptides and primarily addresses the correctness of the peptide rather than the glycan and site localisation[38].

Notably, GlycoPAT[42] (Team 8) and glyXtool^{MS} [37] (Team 3) were in our study also identified as high-performance software solutions for *N*- and *O*-glycopeptide data analysis, respectively. Furthermore, IQ-GPA[35] (Team 1) demonstrated merit for both *N*- and *O*-glycoproteomics data analysis. While all three software handle high-resolution HCD-, ETciD-, and EThcD-MS/MS data, IQ-GPA and GlycoPAT also identify glycopeptides based on high-resolution CID-MS/MS data and apply a post-search filtering step based on advanced peptide and glycan decoy methods to estimate both the peptide and glycan FDR of glycopeptide candidates. The software glyXtool^{MS} instead uses oxonium ions, the Y_1-ion (peptide-HexNAc ion), other glycopeptide-specific fragments, and peptide-specific b- and y-ions, to reduce the false-positive glycopeptide identification rate. Excitingly, these three open-source/academic tools were all recently developed (< 5 years ago), and thus, hold a considerable potential for future impact in the field.

This study also revealed several performance-associated search variables important for comprehensive and accurate glycopeptide identification. Amongst the observed associations related to the search settings, we found that a large glycan and/or protein search space (e.g. broad glycan database or permitting multiple missed cleavages and variable non-glycan modifications of peptides) as expected is associated with a high glycoproteome coverage, but at the cost of high glycopeptide FDRs (high identification rate of NeuGc and multi-Fuc glycans) and less accurate identification of the source glycoproteins. Although not tested directly with this study, such less restricted search settings also increase the search time. As an example of the many performance-associated search outputs that were observed, we found that applying a stringent mass tolerance (low mass error) and accepting incorrect precursor peak

picking were important variables to achieve confident glycan identification for both *N*- and *O*-glycopeptides. Knowledge of this relationship is particularly useful to reduce the FDR of glycopeptides carrying "difficult-to-identify" glycans such as those with near-isobaric masses (e.g. NeuAc-R and Fuc_2-R, Δm = 1.0204 Da). Several less intuitive performance-associated search settings and search outputs were also observed, relationships that will be important to guide and improve the pre- and post-search data analysis process in glycoproteomics.

This community study also illustrated the significant informatics challenges still associated with large-scale glycopeptide data analysis as indicated by the dramatic variance of identified glycopeptides reported across teams. Notably, high discordance of reported glycopeptides was even found between participants using the same software, illustrating that search variables other than the search engine also substantially impact the glycopeptide data analysis. While the 10 Byonic user teams reported a marginally higher rate of consensus glycopeptides, their spread in terms of search output data, reported glyco-PSMs and overall performance scores was of similar magnitude as the variance observed amongst other expert user teams. The variability in the applied search settings and, importantly, at the post-search filtering stage were found to be key factors contributing to the discrepant reports. Concertedly, these observations point to the importance of using both powerful search engines and tailored search settings and post-search filtering criteria to achieve efficient and accurate glycoproteomics data analysis.

Despite the considerable team-to-team variation, this study produced a consensus list of 163 *N*-glycopeptides and 23 *O*-glycopeptides from serum glycoproteins commonly reported by most teams. Importantly, these high-confidence glycopeptides carried biosynthetically-related glycans devoid of NeuGc- and multi-Fuc features in line with literature[28, 45] and mapped to known high-abundance serum proteins[4, 28, 44, 46, 47]. The consensus list has been made publicly available (GlyConnect Reference ID 2943) as it forms an important reference for future studies addressing the complexity of the human serum glycoproteome.

The design of this study including the sample type and preparation as well as the data collection method was chosen to mimic conditions typically encountered in glycoproteomics experiments while also aiming to accommodate for most informatic solutions and appeal to practitioners in the field. Multiple orthogonal performance tests were applied in attempts to ensure a fair and holistic scoring of the search engines and search strategies. Despite these efforts, it cannot be ruled out that some informatics solutions and expert users may have been excluded and teams disadvantaged by the chosen experimental design and scoring system. The resulting scorecard and ranking of software and teams should be viewed in light of these constraints and limitations common to most community-based comparison studies founded on communal data files.

In addition to reporting on the peptide and glycan component of the identified glycopeptides, teams were requested to report on site(s) of modification where possible. Since most tryptic *N*-glycopeptides only comprise a single sequon, site localisation is primarily a challenge related to *O*-glycopeptide analysis[5, 21]. Most teams indeed returned data of the *O*-glycosylation site(s), but due to highly discrepant and often inconclusive reporting of such sites and given the paucity of literature on serum *O*-glycosylation sites, we did not score the ability to localise glycosylation sites in this study.

Most software solutions currently available for glycoproteomics data analysis participated in this study. However, several glycopeptide search engines e.g. pGlyco[53], MSFragger-Glyco[54], and O-Pair Search[55] were unfortunately not represented due to LC-MS/MS data incompatibility or due to their development and/or release after the software evaluation period. Highlighting the rapid progress in the glycoproteomics informatics field, most of the software solutions that participated in this study have been significantly improved and new versions released after the evaluation period. For example, the GPQuest v2.0, GlycoPAT v1.0 and Protein Prospector v5.20.23 search engines tested herein have been superseded by more recent versions i.e. GPQuest v2.1, GlycoPAT v2.0 and Protein Prospector v.6.2.2. Follow-up studies comparing

the performance of these very latest glycoproteomics software upgrades and informatics solutions not included in this study are therefore warranted. Beyond testing the ability of developers and expert users to identify the peptide and glycan component of intact glycopeptides from LC-MS/MS data, such future comparative studies should ideally also test for their ability to accurately report on the modification site as well as the (relative) quantity of the identified glycopeptides.

In summary, this community study has documented that the field has several high-performance informatics tools available for glycoproteomics data analysis and has elucidated many performance-associated search strategies and key variables that may guide developers and users of glycoproteomics software to improved strategies for large-scale glycopeptide data analysis.

Online Methods

Study design and participants

Calls to join this study as a developer (academic/commercial) or expert user of glycoproteomics software were made across the community. In total, 22 teams comprising 9 developer and 13 expert user teams completed the study. All teams were provided with two communal high-resolution glycoproteomics LC-MS/MS data files (File A and B) and were asked to identify intact glycopeptides from these shared files, and then report back their findings using a common reporting template (see PXD024101 via the PRIDE repository). All reports were carefully inspected to ensure strict compliance with the study guidelines and to enable a fair comparison between teams. While the expert user teams were guaranteed anonymity, the developer teams were informed that their software (hence, potentially their identity) would be disclosed upon publication. The expert user teams were free to use any search engine(s) at their disposal including manual annotation and filtering of search output. Developers were asked to return the list of glycopeptide identifications directly from their own software without manual post-search filtering. The 9 developer teams employed the following glycopeptide-centric search engines: Team 1: IQ-GPA v2.5[35], Team 2: Protein Prospector v5.20.23[36], Team 3: glyXtool^{MS} v0.1.4[37], Team 4: Byonic v2.16.16[38], Team 5: Sugar Qb[39], Team 6: Glycopeptide Search v2.0alpha [40], Team 7: GlycopeptideGraphMS v1.0/Byonic[41], Team 8: GlycoPAT v2.0[42] and Team 9: GPQuest v2.0[43] (see **Supplementary Table S1** for details). The relative team performance was compared within (not between) the developer and expert user groups.

Study sample

In short, a synthetic *N*-glycopeptide from human vitamin K-dependent protein C (52 fmol, EVFVHP<u>N</u>YSK, $Hex_5HexNAc_4NeuAc_2$) was spiked into human serum (5 µg, product number

#31876, Thermo Fisher Scientific), the mixture digested using trypsin, and the resulting glycopeptides enriched (see **Extended methods** for details of the study sample and handling).

Mass spectrometry

The enriched glycopeptide mixture (~1 µg/injection) was analysed twice using two similar LC-MS/MS acquisition methods to generate File A and B. For both files, peptides were separated using C_{18} reversed phase nanoLC and detected on a Thermo Scientific™ Orbitrap Fusion™ Lumos™ Tribrid™ mass spectrometer using high-resolution MS1 and data-dependent HCD-MS/MS data acquisition. Product-dependent (oxonium ion) triggered ETciD-/CID-MS/MS (Orbitrap) and EThcD-/CID-MS/MS (ion trap) events of glycopeptide precursors were scheduled for File A and B, respectively. File A and B were provided to all participants as .raw and .mgf files (see **Extended methods** for LC-MS/MS details).

Search instructions and reporting template

All teams were requested to use a fixed protein search space (human proteome, 20,231 UniProtKB reviewed sequences), but could freely choose the *N*- and *O*-glycan search space. Teams were asked not to include xylose and any glycan substitutions in the glycan search space. The participants reported their team details, identification strategy and identified glycopeptides in a common reporting template. All reports were carefully checked for compliance to the study guideline (see **Extended methods** for search and reporting details).

Compilation and comparison of team data

Data from the returned reports were compiled (**Supplementary Table S1-S2**) including the reported *N*- and *O*-glycopeptides (**Supplementary Table S3**). Unique identifiers (IDs) for the reported glycopeptides and their glycan compositions and source glycoproteins were generated to enable comparison across teams. The glycan composition ID was written as Hex*HexNAc*Fuc*NeuAc*, where * represents the number of monosaccharide residues.

Glycopeptides adducted with Na^+ and K^+ reported by some teams were grouped with the corresponding non-adducted glycan compositions. UniProtKB identifiers were used as the source glycoprotein IDs. The glycopeptide IDs were written as the peptide sequence followed by the generic glycan composition. Comparisons between the glycan compositions, source glycoproteins and glycopeptide IDs reported by the 22 teams were performed using the "pivot table" in Excel. Variables from each identifier type were compared as summed counts across teams (see **Extended methods** for details of the data comparison between reports).

Performance testing

The relative team performance was assessed using a scoring system composed of multiple independent tests designed to holistically score the accuracy (specificity) and coverage (sensitivity) of the reported *N*- and *O*-glycopeptides. The raw scores from the individual tests (N1-N6 and O1-O5) were normalised within the range 0-1. The normalised scores were collectively used to establish an overall score (range 0-1) measuring the ability to perform efficient *N*- and *O*-glycopeptide data analysis. The overall score was utilised to rank the developer and expert user teams within each group. The performance tests included the synthetic *N*-glycopeptide test (N1), the glycan composition test (N2 and O1), the source glycoprotein test (N3 and O2), the glycoproteome coverage test (N4 and O3), the commonly reported (consensus) glycopeptide test (N5 and O4), and the NeuGc and multi-Fuc glycopeptide test (N6 and O6). The overall performance scores for *N*- and *O*-glycopeptide analysis were established by averaging the normalised scores of the individual performance tests (see **Extended methods** for details of the performance tests and the scoring system).

Statistical analysis

The normalised scores from each of the performance tasks (N1-N6 and O1-O5) and the overall performance scores across all teams were tested for associations with the search settings (SS1-

SS13) and search output variables (SO1-SO9) using seven statistical methods[56]. Only strong associations observed across a minimum of three different statistical methods were considered in this study (see **Extended methods** for details of the statistical methods).

Supplementary Information

This study contains supplementary information. Two supporting files have been provided: 1) Supplementary information containing **Extended methods** and **Supplementary Figure S1–S6** (PDF) and 2) **Supplementary Table S1–S17** (Microsoft Excel). Further, the LC-MS/MS raw data files (File A and B), the common reporting template and the deidentified but otherwise unredacted participant reports are available via ProteomeXchange with identifier PXD024101.

Acknowledgements

Drs Rosa Viner and Sergei Snovida are thanked for providing high-quality LC-MS/MS data to the study. Dr Krishnatej Nishtala is thanked for aiding the data analysis. Drs Catherine Hayes, Julien Mariethoz and Frederique Lisacek are thanked for informatics assistance. RK was supported by an Early Career Fellowship from the Cancer Institute NSW. DK is the recipient of an Australian Research Council Future Fellowship (project number FT160100344) funded by the Australian Government. GP was funded by FAPESP (n° 2018/15549-1). MTA was supported by a Macquarie University Safety Net Grant. This work was supported by the Australian Research Council Centre of Excellence in Nanoscale Biophotonics (CE140100003).

Author Contributions

Conception: NHP and MTA. Design: RK, DK, KK, GL, KM, GP, JZ, JY, SH, NHP, and MTA. Data acquisition: All study participants (Team 1-22). Data analysis: RK, MTA. Data interpretation: RK, MTA. Creation of new software used in the work: All developer teams (Team 1-9). Manuscript writing and editing: RK, DK, KK, GL, KM, GP, JZ, JY, SH, NHP, and MTA. All authors have seen and approved the manuscript.

Competing Interests Statement

All authors responsible for the conception, design, analysis and interpretation of data, as well as the manuscript writing and editing declare no conflict of interest. The study participants (Team 1-22) declare a perceived or real financial or academic conflict of interest in the study outcome, which was mitigated by excluding all study participants from the analysis and interpretation of data returned by the teams and from manuscript editing.

References

1. Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3-49 (2017).
2. Thaysen-Andersen, M., Packer, N.H. & Schulz, B.L. Maturing Glycoproteomics Technologies Provide Unique Structural Insights into the N-glycoproteome and Its Regulation in Health and Disease. *Mol Cell Proteomics* **15**, 1773-1790 (2016).
3. Chandler, K.B. & Costello, C.E. Glycomics and glycoproteomics of membrane proteins and cell-surface receptors: Present trends and future opportunities. *Electrophoresis* **37**, 1407-1419 (2016).
4. Ye, Z., Mao, Y., Clausen, H. & Vakhrushev, S.Y. Glyco-DIA: a method for quantitative O-glycoproteomics with in silico-boosted glycopeptide libraries. *Nature Methods* **16**, 902-910 (2019).
5. Pap, A., Klement, E., Hunyadi-Gulyas, E., Darula, Z. & Medzihradszky, K.F. Status Report on the High-Throughput Characterization of Complex Intact O-Glycopeptide Mixtures. *Journal of The American Society for Mass Spectrometry* **29**, 1210-1220 (2018).
6. Blazev, R. et al. Integrated glycoproteomics identifies a role of N-glycosylation and galectin-1 on myogenesis and muscle development. *Mol Cell Proteomics* (2020).
7. Kawahara, R. et al. The complexity and dynamics of the tissue glycoproteome associated with prostate cancer progression. *Mol Cell Proteomics* (2020).
8. Riley, N.M., Hebert, A.S., Westphall, M.S. & Coon, J.J. Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nat Commun* **10**, 1311 (2019).
9. Leung, K.K. et al. Broad and thematic remodeling of the surfaceome and glycoproteome on isogenic cells transformed with driving proliferative oncogenes. *Proc Natl Acad Sci U S A* **117**, 7764-7775 (2020).

10. Hu, Y. et al. Integrated Proteomic and Glycoproteomic Characterization of Human High-Grade Serous Ovarian Carcinoma. *Cell Rep* **33**, 108276 (2020).

11. Zhao, X. et al. An Integrated Mass Spectroscopy Data Processing Strategy for Fast Identification, In-Depth, and Reproducible Quantification of Protein O-Glycosylation in a Large Cohort of Human Urine Samples. *Anal Chem* **92**, 690-698 (2020).

12. Jiang, B. et al. A multi-parallel N-glycopeptide enrichment strategy for high-throughput and in-depth mapping of the N-glycoproteome in metastatic human hepatocellular carcinoma cell lines. *Talanta* **199**, 254-261 (2019).

13. Chernykh, A., Kawahara, R. & Thaysen-Andersen, M. Towards structure-focused glycoproteomics. *Biochemical Society Transactions*, 1-25 (2020).

14. Lee, L.Y. et al. Toward Automated N-Glycopeptide Identification in Glycoproteomics. *J Proteome Res* **15**, 3904-3915 (2016).

15. Darula, Z. & Medzihradszky, K.F. Carbamidomethylation Side Reactions May Lead to Glycan Misassignments in Glycopeptide Analysis. *Anal Chem* **87**, 6297-6302 (2015).

16. Riley, N.M., Malaker, S.A. & Bertozzi, C.R. Electron-Based Dissociation Is Needed for O-Glycopeptides Derived from OpeRATOR Proteolysis. *Anal Chem* **92**, 14878-14884 (2020).

17. Riley, N.M., Malaker, S.A., Driessen, M.D. & Bertozzi, C.R. Optimal Dissociation Methods Differ for N- and O-Glycopeptides. *J Proteome Res* **19**, 3286-3301 (2020).

18. Woo, C.M. et al. Development of IsoTaG, a Chemical Glycoproteomics Technique for Profiling Intact N- and O-Glycopeptides from Whole Cell Proteomes. *J Proteome Res* **16**, 1706-1718 (2017).

19. Fang, P. et al. Multilayered N-Glycoproteome Profiling Reveals Highly Heterogeneous and Dysregulated Protein N-Glycosylation Related to Alzheimer's Disease. *Anal Chem* **92**, 867-874 (2020).

20. Woo, C.M. et al. Mapping and Quantification of Over 2000 O-linked Glycopeptides in Activated Human T Cells with Isotope-Targeted Glycoproteomics (Isotag). *Mol Cell Proteomics* **17**, 764-775 (2018).

21. Darula, Z. & Medzihradszky, K.F. Analysis of Mammalian O-Glycopeptides-We Have Made a Good Start, but There is a Long Way to Go. *Mol Cell Proteomics* **17**, 2-17 (2018).

22. Wu, S.W., Pu, T.H., Viner, R. & Khoo, K.H. Novel LC-MS(2) product dependent parallel data acquisition function and data analysis workflow for sequencing and identification of intact glycopeptides. *Anal Chem* **86**, 5478-5486 (2014).

23. Reiding, K.R., Bondt, A., Franc, V. & Heck, A.J.R. The benefits of hybrid fragmentation methods for glycoproteomics. *TrAC Trends in Analytical Chemistry* **108**, 260-268 (2018).

24. Thaysen-Andersen, M., Kolarich, D. & Packer, N.H. Glycomics & Glycoproteomics: From Analytics to Function. *Mol Omics* (2020).

25. Hu, H., Khatri, K. & Zaia, J. Algorithms and design strategies towards automated glycoproteomics analysis. *Mass Spectrom Rev* **36**, 475-498 (2017).

26. Abrahams, J.L. et al. Recent advances in glycoinformatic platforms for glycomics and glycoproteomics. *Curr Opin Struct Biol* **62**, 56-69 (2020).

27. Cao, W. et al. Recent advances in software tools for more generic and precise intact glycopeptide analysis. *Mol Cell Proteomics* (2020).

28. Clerc, F. et al. Human plasma protein N-glycosylation. *Glycoconj J* **33**, 309-343 (2016).

29. Dotz, V. & Wuhrer, M. N-glycome signatures in human plasma: associations with physiology and major diseases. *FEBS Lett* **593**, 2966-2976 (2019).

30. Hoffmann, M., Marx, K., Reichl, U., Wuhrer, M. & Rapp, E. Site-specific O-Glycosylation Analysis of Human Blood Plasma Proteins. *Mol Cell Proteomics* **15**, 624-641 (2016).

31. Parker, B.L. et al. Terminal Galactosylation and Sialylation Switching on Membrane Glycoproteins upon TNF-Alpha-Induced Insulin Resistance in Adipocytes. *Mol Cell Proteomics* **15**, 141-153 (2016).

32. Zhang, Y. et al. Systems analysis of singly and multiply O-glycosylated peptides in the human serum glycoproteome via EThcD and HCD mass spectrometry. *J Proteomics* **170**, 14-27 (2018).

33. Yu, Q. et al. Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characterization. *J Am Soc Mass Spectrom* **28**, 1751-1764 (2017).

34. Darula, Z., Pap, A. & Medzihradszky, K.F. Extended Sialylated O-Glycan Repertoire of Human Urinary Glycoproteins Discovered and Characterized Using Electron-Transfer/Higher-Energy Collision Dissociation. *J Proteome Res* **18**, 280-291 (2019).

35. Park, G.W. et al. Integrated GlycoProteome Analyzer (I-GPA) for Automated Identification and Quantitation of Site-Specific N-Glycosylation. *Sci Rep* **6**, 21175 (2016).

36. Baker, P.R., Trinidad, J.C. & Chalkley, R.J. Modification site localization scoring integrated into a search engine. *Mol Cell Proteomics* **10**, M111 008078 (2011).

37. Pioch, M., Hoffmann, M., Pralow, A., Reichl, U. & Rapp, E. glyXtool(MS): An Open-Source Pipeline for Semiautomated Analysis of Glycopeptide Mass Spectrometry Data. *Anal Chem* **90**, 11908-11916 (2018).

38. Bern, M., Kil, Y.J. & Becker, C. Byonic: advanced peptide and protein identification software. *Curr Protoc Bioinformatics* **Chapter 13**, Unit13 20 (2012).

39. Stadlmann, J., Hoi, D.M., Taubenschmid, J., Mechtler, K. & Penninger, J.M. Analysis of PNGase F-Resistant N-Glycopeptides Using SugarQb for Proteome Discoverer 2.1 Reveals Cryptic Substrate Specificities. *Proteomics* **18**, e1700436 (2018).

40. Pompach, P., Chandler, K.B., Lan, R., Edwards, N. & Goldman, R. Semi-automated identification of N-Glycopeptides by hydrophilic interaction chromatography, nano-reverse-phase LC-MS/MS, and glycan database search. *J Proteome Res* **11**, 1728-1740 (2012).

41. Choo, M.S., Wan, C., Rudd, P.M. & Nguyen-Khuong, T. GlycopeptideGraphMS: Improved Glycopeptide Detection and Identification by Exploiting Graph Theoretical Patterns in Mass and Retention Time. *Anal Chem* **91**, 7236-7244 (2019).

42. Liu, G. et al. A Comprehensive, Open-source Platform for Mass Spectrometry-based Glycoproteomics Data Analysis. *Mol Cell Proteomics* **16**, 2032-2047 (2017).

43. Toghi Eshghi, S., Shah, P., Yang, W., Li, X. & Zhang, H. GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides. *Anal Chem* **87**, 5181-5188 (2015).

44. Sun, S. et al. Site-Specific Profiling of Serum Glycoproteins Using N-Linked Glycan and Glycosite Analysis Revealing Atypical N-Glycosylation Sites on Albumin and alpha-1B-Glycoprotein. *Anal Chem* **90**, 6292-6299 (2018).

45. Yabu, M., Korekane, H. & Miyamoto, Y. Precise structural analysis of O-linked oligosaccharides in human serum. *Glycobiology* **24**, 542-553 (2014).

46. Darula, Z., Sarnyai, F. & Medzihradszky, K.F. O-glycosylation sites identified from mucin core-1 type glycopeptides from human serum. *Glycoconj J* **33**, 435-445 (2016).

47. Yang, W., Ao, M., Hu, Y., Li, Q.K. & Zhang, H. Mapping the O-glycoproteome using site-specific extraction of O-linked glycopeptides (EXoO). *Mol Syst Biol* **14**, e8486 (2018).

48. Pavic, T. et al. N-glycosylation patterns of plasma proteins and immunoglobulin G in chronic obstructive pulmonary disease. *J Transl Med* **16**, 323 (2018).

49. Zaytseva, O.O. et al. Heritability of Human Plasma N-Glycome. *J Proteome Res* **19**, 85-91 (2020).

50. Gudelj, I. et al. Changes in total plasma and serum N-glycome composition and patient-controlled analgesia after major abdominal surgery. *Sci Rep* **6**, 31234 (2016).

51. Gizaw, S.T., Gaunitz, S. & Novotny, M.V. Highly Sensitive O-Glycan Profiling for Human Serum Proteins Reveals Gender-Dependent Changes in Colorectal Cancer Patients. *Anal Chem* **91**, 6180-6189 (2019).

52. Chalkley, R.J. et al. Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol Cell Proteomics* **4**, 1194-1204 (2005).

53. Liu, M.Q. et al. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nat Commun* **8**, 438 (2017).

54. Polasky, D.A., Yu, F., Teo, G.C. & Nesvizhskii, A.I. Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat Methods* **17**, 1125-1132 (2020).

55. Lu, L., Riley, N.M., Shortreed, M.R., Bertozzi, C.R. & Smith, L.M. O-Pair Search with MetaMorpheus for O-glycopeptide characterization. *Nat Methods* **17**, 1133-1138 (2020).

56. Efron, B. & Hastie, T. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. (Cambridge University Press, 2016).