O-Pair Search with MetaMorpheus for O-glycopeptide characterization

Lei Lu1,2,3, Nicholas M. Riley2,4, Michael R. Shortreed5, Carolyn R. Bertozzi2,3 and Lloyd M. Smith1,6

We report O-Pair Search, an approach to identify O-glycopeptides and localize O-glycosites. Using paired collision- and electron-based dissociation spectra, O-Pair Search identifies O-glycopeptides via an ion-indexed open modification search and localizes O-glycosites using graph theory and probability-based localization. O-Pair Search reduces search times more than 2,000-fold compared to current O-glycopeptide processing software, while defining O-glycosite localization confidence levels and generating more O-glycopeptide identifications. Beyond the mucin-type O-glycopeptides discussed here, O-Pair Search also accepts user-defined glycan databases, making it compatible with many types of O-glycosylation. O-Pair Search is freely available within the open-source MetaMorpheus platform at https://github.com/smith-chem-wisc/MetaMorpheus.

Mass spectrometry (MS) is the gold standard for interrogating the glycoproteome, enabling the localization of glycans to specific glycosites1–3. Recent applications of electron-driven dissociation methods have shown promise in localizing modified O-glycosites even in multiply glycosylated peptides4. Yet, standard approaches for interpreting MS/MS spectra are ill-suited to the heterogeneity of O-glycopeptides. Perhaps the most challenging problem for O-glycopeptide analysis is mucin-type O-glycosylation, which is abundant on many extracellular and secreted proteins and is a crucial mediator of immune function, microbiome interaction and biophysical forces imposed on cells, among others4. Mucin-type O-glycans are linked to serine and threonine residues through an initiating N-acetylgalactosamine (GalNAc) sugar, which can be further elaborated into four main backbone fragments that do not retain glycan masses (but with a pre-defined index). These O-glycosites occur most frequently in long serine/threonine rich sequences (Supplementary Fig. 1), such as PTS mucin tandem repeat domains, which exist with macroheterogeneity defined by the occurrence of O-glycosylation and with microheterogeneity defined by a large number of potential O-glycans4–11. The number of serine and threonine residues present in glycopeptides derived from mucin-type O-glycoproteins, combined with the consideration of dozens of potential O-glycans at each site, leads to a combinatorial explosion when generating databases of theoretical O-glycopeptides to consider for each tandem MS (MS/MS) spectrum (Supplementary Note 1).

Current O-glycoproteomic analysis pipelines are unable to search for multiply O-glycosylated peptides within reasonable time frames even for simple mixtures of O-glycoproteins, much less for proteome-scale experiments. Recent efforts to combat this search time issue have forgone site localization for the more expedient option of identifying only the total glycan mass on a peptide backbone4. While effective at lowering time costs, this sacrifices valuable information about site-specific modifications, which is often the goal of intact glycopeptide analysis in the first place. Such an approach also fails to report the number and composition of individual glycans for multiply glycosylated O-glycopeptides, where multiple smaller oligosaccharides may represent the same mass as a larger single glycan or a combination of different oligosaccharides.

Open modification searches and combinations of peptide database searching with de novo glycan sequencing have also recently been reported, but neither address the time issues that challenge analysis of highly modified O-glycopeptides4–11. Moreover, electron-driven dissociation methods are required to localize O-glycosites4,11–13, yet current software tools fail to capitalize on combinations of collision- and electron-based fragmentation spectra that are acquired for the same precursor ion. This is coupled with a general lack of ability to confidently localize glycans within multiply glycosylated O-glycopeptides.

Here, we describe the O-Pair Search strategy implemented in the MetaMorpheus platform to provide a pipeline for rapid identification of O-glycopeptides and subsequent localization of O-glycosites using paired collision- and electron-based dissociation spectra collected for the same precursor ion.

Results

Developing the O-Pair Search approach. O-Pair Search first uses an ion-indexed open search14 of higher energy collisional dissociation (HCD) spectra to rapidly identify combinations of peptide sequences and total O-glycan masses, which are generated through combinations of entries in an O-glycan database. Graph-theoretical localization15–21 then defines site-specific O-glycan localizations using ions present in EThC spectra (electron-transfer dissociation with HCD supplemental activation) (Fig. 1b). The HCD–EThC spectral pair in Fig. 1c exemplifies this synergistic approach: the HCD spectrum enables confident sequence assignment via peptide backbone fragments that do not retain glycan masses (but with a precursor mass showing a defined mass shift matching a combination of O-glycans), while the subsequent EThC spectrum enables localization of all four O-glycosites (gold) even with the presence of four other unmodified potential sites. Localization is followed by localization probability calculations using an extension of the phosphoRS12 algorithm used for phosphosite localization (maximum score of 1), in addition to scoring of fine scoring (which includes calculation of Y-type ions) and false discovery rate (FDR) calculations performed separately for O-glycosylated and nonmodified peptides.

We also introduce here the concept of localization levels, which is the culmination of the O-Pair Search (Fig. 1d). Inspired by early development of de novo glycan sequencing, we define three localization levels: base (the raw mass), site (the specific position on the peptide), and glycan (the specific sugar structure).
adoption of class levels for phosphopeptide localizations and more recently for proteoforms, we developed this classification system to more accurately describe the quality and confidence of glycopeptide and glycosite identifications. Level 1 glycopeptide identifications indicate that all glycans identified in the total glycan mass modification are localized to specific serine and threonine residues with a localization probability >0.75. Glycopeptides with glycosite assignments with localization probabilities <0.75 are assigned as Level 1b, even though they are still identified as localized by the graph theory approach. Level 1b assignments also occur when a glycosite is assigned without the presence of sufficient spectral evidence (for example, fragments cannot explain a glycosite, but the sequence contains only one serine or threonine). We currently borrow the 0.75 cutoff from phosphopeptide precedents, empirical determination of localization cutoffs will likely need to be explored in future work using libraries of synthetic glycopeptide standards, as has been done with phosphopeptides. That said, such libraries are currently difficult to generate. Level 2 assignments occur when at least one glycosite is assigned a glycan based on spectral evidence, but not all glycans can be assigned unambiguously. Level 3 identifications represent a confident match of glycopeptide and total glycan mass, but no glycosites can be assigned unambiguously. Indeed,
Level 3 glycopeptides (such as those reported in HCD-only methods by default) are still useful to note the presence of glycosylated residues somewhere in a given sequence. Overall, our classification system provides a straightforward approach to qualify glycoproteomic datasets without having to exclude confident identifications that have no site-specific information.

In addition to localization-level assignments, O-Pair Search also reports matched peptide and glycan fragment ion series and their intensities for each of the paired spectra, the presence of N-glycosylation sequons to identify potentially confounding assignments, and localization probabilities for all sites, both localized and not. As with all glycopeptide-centric workflows, O-Pair Search reports comparisons with limited ability to comment on topology of glycans. That said, oxonium ions can help distinguish between certain glycans, such as the ratio of 138.055 and 144.0655 m/z to differentiate HexNAc residues as GlcNAc or GalNAc. O-Pair Search reports this ratio by default to aid with interpretation. For example, 95.7% of identifications with localized H1N1 glycans have a 138/144 ratio less than two, suggesting GalNAc-Gal rather than Man-GlcNAc for H1N1 identity. Furthermore, during the review process for this manuscript, another approach using fragment ion-indexed open searching of glycopeptide spectra (but without localization as discussed here) was reported from Nesvizhskii and coworkers, the group who developed the highly efficient Byonic search.

O-Pair Search dramatically decreased search times, with approximately 45- and 2,100-fold faster searches than Byonic when considering two or four glycans per peptide, respectively (Fig. 2c). O-Pair Search required approximately 30 s to complete a search considering four glycans per peptide, while the Byonic search was terminated after the search failed to complete in over 33,000 min (roughly 3.5 weeks, no reported search progress for over 1 week). Improvements in search speed are accompanied by roughly 2–3-fold increases in the number of localized glycosites identified.

In addition to more than doubling the number of total identified spectra, O-Pair Search identified most spectra that Byonic returned as GlycoPSMs for both HCD (Fig. 2d) and EThcD (Fig. 2e) scans, and most (around 95%) of the shared identified scans mapped to the same glycopeptide (Fig. 2f). These searches were completed using a FASTA file containing sequences only for the four mucin standards, which highlights the impracticality of O-glycopeptide searches in Byonic for complex mixtures. Moreover, O-Pair Search performed localization calculations and reported localization levels within the reported search time while Byonic spectra had to be further processed after the search to obtain localization information (see Methods). Supplementary Fig. 4 compares microheterogeneity seen at localized glycosites between O-Pair Search and Byonic searches.

The ability to rapidly search O-glycopeptide data allowed us to vary the number of O-glycans to consider per peptide for easy evaluation of optimal search conditions. Figure 2g shows that search times remain shorter than 1 min when considering four glycans per peptide, while up to eight glycans can be considered per peptide in searches requiring less than 20 min. Allowing for more glycans per peptide does not change the spectral assignments to various glycopeptides (Supplementary Fig. 5), indicating the robustness of O-Pair Search identifications. The number of nonmodified identifications remained similarly constant (Supplementary Fig. 6). Similarly, different glycan databases can be searched within reasonable time frames (Supplementary Fig. 7).

Evaluating O-Pair Search performance. We first compared O-Pair Search to Byonic, the most commonly used O-glycopeptide identification software. Byonic, which uses a lookup peaks approach to speed up search times relative to traditional database searching, can also search HCD and EThcD spectra, although it is agnostic of paired spectra originating from the same precursor. To benchmark performance, we used a recently published dataset of O-glycopeptides from mucin glycoproteins using a combination of trypsin and the mucin-specific protease SteC, which cleaves only in glycosylated mucin domains. This data originates from sequential digestion of four recombinant mucin standards (CD43, MUC16, PSGL-1 and Gp1ba), using SteC to cleave mucin domains followed by N-glycan removal with PNGaseF and trypptic digestion.

We initially searched a file with HCD and EThcD paired spectra from this dataset, using a protein sequence file of the four mucins (to minimize Byonic search times) and a glycan database of 12 common O-glycans presented in Fig. 1a (ref. 20). The number of compositions considered in each search is not simply 12, however, but instead is a combination of possible compositions determined by the number of glycans allowed per glycopeptide. When four glycans are allowed per peptide, this actually represents 439 different mass offset values; that is, the number of unique masses present in 1,819 different glycan combinations (Supplementary Note 1). O-Pair Search identified more localized (Fig. 2a) and total (Fig. 2b) O-glycopeptide spectral matches (glycoPSMs) than Byonic when allowing either two or three glycans per peptide (Supplementary Fig. 2 and Fig. 2, respectively). This holds true even when relaxing the scoring thresholds used to obtain confident Byonic identifications (Supplementary Fig. 3). Note, all O-Pair Search identifications represent two spectra from an HCD–EThcD spectral pair. Conversely, Byonic is agnostic to paired scans, meaning identifications can come from HCD and EThcD spectra that were collected for the same precursor (pair) or from spectra identified separately from their paired counterpart. Byonic identifications are grouped into HCD–EThcD pairs (where paired scans identified the same O-glycopeptide), HCD alone and EThcD alone. The last two cases are where an identification came only from an HCD scan or EThcD scan, but the other spectrum in the pair did not return a hit.
The search times required for byonic and O-Pair Search when considering two, three and four glycans per peptide (DNF, did not finish). The number of identifications are considered for the localized O-Pair Search data, and three glycans per peptide were allowed for both searches.

Comparing the number of localized (1b) and total glycoPSMs for HCD-pd–eThcD data collected from Stce digestions of four recombinant mucin standards. Note, only Level 1 and 1b identifications were considered for the localized O-Pair Search data, and three glycans per peptide were allowed for both searches.

O-Pair Search was used to process files from a published urinary O-glycopeptide study that previously reported Protein Prospector (Prot. Pros.) and byonic results. Here, T otal indicates all identifications; that is, the sum of all localization-level identifications.

O-Pair Search performance with several protein database backgrounds. Here, T otal indicates all identifications; that is, the sum of all localization-level identifications.

O-Pair Search performance when considering a range of glycans allowed per peptide.

O-Pair Search performance when considering a range of glycans allowed per peptide.

Fig. 2 | Performance of O-Pair Search for O-glycopeptide characterization. a, b, Comparing the number of localized (a) and total glycoPSMs (b) returned from Byonic and from O-Pair Search for HCD-pd–EThcD data collected from Stce digestions of four recombinant mucin standards. Note, only Level 1 and 1b identifications are considered for the localized O-Pair Search data, and three glycans per peptide were allowed for both searches. c, The table compares the search times required for Byonic and O-Pair Search when considering two, three and four glycans per peptide (DNF, did not finish). The number of localized glycosites identified by the searches is also provided, which correspond to the GlycoPSMs in a and Supplementary Fig. 3. NA, not applicable. d-f, Overlap in identified spectra between O-Pair Search and Byonic is compared for both HCD (d) and EThcD scans (e), and f provides the proportion of shared identified scans that map to the same glycopeptide. g, O-Pair Search performance when considering a range of glycans allowed per peptide. h, O-Pair Search performance with several protein database backgrounds. Here, Total indicates all identifications; that is, the sum of all localization-level identifications. i, Estimated FDRs resulting from various entrapment database searches (Std, standard; Muc, mucin). j, O-Pair Search was used to process files from a published urinary O-glycopeptide study that previously reported Protein Prospector (Prot. Pros.) and Byonic results. k, Protein Prospector reports localized glycosites, which we converted into our localization-level system and compared with O-Pair results. l, Results from several O-Pair searches of Fraction 1 (three files), Fraction 2 (two files) and all ten files available from the urinary O-glycopeptide study. Supplementary Note 4 details the files used for each panel.
Large-scale O-glycoproteomics with O-Pair Search. Finally, we applied O-Pair Search to a large dataset of urinary O-glycopeptides, which has been analyzed in a number of studies\(^{19-22}\). The raw data for this dataset represent glycopeptides purified from urine from four donors using affinity chromatography with wheat germ agglutinin and is available through the MassIVE repository (MSV000083070).\(^{8}\) Pap et al. provide identifications from Protein Prospector and Byonic for EThcD scans from Fraction 1 (the ‘shoulder fraction,’ three raw data files available) and Fraction 2 (the ‘GlcNAc fraction,’ two raw data files available) and providing site-specific localization for the open-search step. Others have shown that starting method development may be needed to optimize data acquisition.

We expanded our analysis of this dataset to explore the use of a larger glycan database (32 glycans versus 12) and the effect of searching with more glycans allowed per peptide (five versus two). Figure 2I compares results from these different search parameters for Fraction 1, Fraction 2 and all ten files available for download from the urinary O-glycopeptide dataset. In Fraction 1, the larger O-glycan database boosted identifications for Fraction 1, and increased in Fraction 2 and the entire dataset as a whole. This indicates that Fraction 1 likely harbored glycopeptides with more diverse glycans while most of the dataset did not. Also, according to the original study, Fraction 2 contained more multiply modified O-glycopeptides, which may produce less efficient peptide backbone fragments sufficient for reliable identification. Conversely, considering more glycans per peptide provided slight benefits in all cases. By requiring only a few hours to perform a whole proteome-search with a variety of glycopeptide possibilities, O-Pair Search provides a flexible platform to explore O-glycoproteomics data. When considering only Level 1 and 1b GlycoPSMs, our results represent 447 unique O-glycopeptides with localized O-glycosylations, and O-Pair Search identified 354 localized O-glycosylations in total when allowing five glycans per peptide from the 12-glycan database.

Discussion

In all, we show that O-Pair Search can reduce O-glycopeptide search times by \(>2,000\) times over the most widely used commercial glycopeptide search tool, Byonic. Additionally, O-Pair Search identifies more O-glycopeptides than Byonic and provides O-glycosite localizations using graph theory and localization probabilities. O-Pair Search also introduces a new classification scheme to unify data reporting across the glycoproteomic community. These localization levels are automatically calculated by O-Pair Search to indicate if all (Level 1), at least one (Level 2) or none (Level 3) of the O-glycosylates are confidently localized. We further demonstrate the use of O-Pair Search by searching a large published dataset of urinary O-glycopeptides, increasing the number of glycopeptides identified and providing site-specific localization for \(>350\) O-glycosylates. We also note that O-Pair Search allows user-specified glycan databases to enable unbiased searches of a variety of glycosylation classes, including O-fucose, O-mannose and O-glucose. That said, these O-glycans often lack monosaccharides that generate the most commonly used oxonium ions for product-dependent methods, so more method development may be needed to optimize data acquisition. A report published while this work was under review also described the classification of mucin-type O-glycans using B and Y type; we will seek to incorporate this into our workflow\(^{26}\). Current limitations include the reliance on HCD to generate good peptide fragmentation for the open-search step. Others have shown that starting with EThcD data may be a viable option\(^{27}\), although this also brings several inherent challenges. This may be alleviated in our approach by indexing both HCD and EThcD spectra for open searching. Even so, true peptides may still rank low among all the peptide candidates. Scoring refinement could improve all open-search approaches (including ours), especially in complex datasets where many precursor candidates must be considered. Also, it remains difficult for any glycoproteomics software to identify glycan isomers.
(that is, same composition, different connectivity), a challenge not addressed here. Perhaps better automation of this could be achieved as more studies are published with O-glycopeptides modified with defined glycan structures. Regardless, O-Pair Search can process both product-dependent and standard acquisition methods with a variety of O-glycan databases, making it a flexible tool for a variety of O-glycoproteomics applications.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-00985-5.

Received: 18 May 2020; Accepted: 21 September 2020; Published online: 26 October 2020

References
1. Ouyang, X., Qin, H. & Ye, M. Recent advances in methods for the analysis of glycoproteins and glycoconjugates. Curr. Opin. Struct. Biol. 62, 56–69 (2020).
2. You, X., Qin, H. & Ye, M. Recent advances in methods for the analysis of protein O-glycosylation at proteome level. J. Sep. Sci. 41, 248–261 (2018).
3. Suttapitsakul, S., Sun, F. & Wu, K. Recent advances in glycoproteomics analysis by mass spectrometry. Anal. Chem. 92, 267–291 (2020).
4. Riley, N. & Coom, J. I. The role of electron transfer dissociation in modern proteomics. Anal. Chem. 90, 40–64 (2018).
5. Reily, C., Stewart, T. J., Renfrow, M. B. & Novak, J. Glycosylation in health and disease. Nature 15, 346–366 (2019).
6. Brockhausen, I. & Stanley, P. in Essentials in Glycobiology (eds Varki, A. & Stanley, P.). (Oxford University Press, 2010).
7. 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. Cellular Proteomics 17, 2–18 (2018).
8. Pap, A., Klement, E., Hunyadi-Gulyas, E., Darula, Z. & Medzihradszky, K. F. Status report on the high-throughput characterization of O-glycopeptide mixtures. J. Am. Soc. Mass Spectrom. 29, 1210–1220 (2018).
9. 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).
10. Papan, A., Tisnadee, E., Medizhradszky, K. F. & Darula, Z. Novel O-linked sialoglycans structures in human urinary glycoproteins. Mol. Omics. 16, 156–164 (2020).
11. Khoo, K. H. Advances toward mapping the full extent of protein site-specific O-GaINAc glycosylation that better reflects underlying glycogenic complexity. Curr. Opin. Struct. Biol. 56, 146–154 (2019).
12. Mao, J. et al. A new searching strategy for the identification of O-linked glycopeptides. Anal. Chem. 91, 3852–3859 (2019).
13. Izhazam, A. R. A. & Scott, N. E. Open database searching enables the identification and comparison of bacterial glycoproteomes without defining glycan compositions prior to searching. Mol. Cell. Proteomics https://doi.org/10.1074/mcp.T112.02100 (2020).
14. Huang, J. et al. Development of a computational tool for automated interpretation of intact O-glycopeptide tandem mass spectra from single proteins. Anal. Chem. 92, 6777–6784 (2020).
15. Riley, N. M., Malaker, S. A., Driessen, M. & Bertozzi, C. R. Optimal dissociation methods differ for N- and O-glycopeptides. J. Proteome Res. 19, 3286–3301 (2020).
16. Solntsev, S. K., Shortreed, M. R., Frey, B. L. & Smith, L. M. Enhanced global post-translational modification discovery with MetaMorphus. J. Proteome Res. 17, 1844–1851 (2018).
17. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat. Methods 14, 513–520 (2017).
18. Liu, X. et al. Identification of ultra-modified proteins using top-down tandem mass spectra. J. Proteome Res. 12, 5830–5838 (2013).
19. Frank, A. M., Pesavento, J. J., Muzzen, C. A., Kelleher, N. L. & Pezner, P. A. Interpreting top-down mass spectra using spectral alignment. Anal. Chem. 80, 2499–2505 (2008).
20. Pezner, P. A., Dancik, V. & Tang, C. L. Mutation-tolerant protein identification by mass spectrometry. Comput. Biol. 7, 777–787 (2001).
21. Park, J. et al. Informed Proteomics: open-source software package for top-down proteomics. Nat. Methods 14, 909–914 (2017).
22. Taus, T. et al. Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. 10, 5354–5362 (2011).
23. Olsen, J. V. et al. Global, In Vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127, 635–648 (2006).
24. Smith, L. M. et al. A five-level classification system for peptide identification. Nat. Methods 16, 939–940 (2019).
25. Marx, H. et al. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nat. Biotechnol. 31, 557–564 (2013).
26. Halimi, A. et al. Assignment of saccharide identities through analysis of oxonium ion fragmentation profiles in LC–MS/MS of glycopolypeptides. J. Proteome Res. 13, 6024–6032 (2014).
27. Polasky, D. A., Yu, F., Teo, G. C. & Nesvizhskii, A. I. Fast and comprehensive O- and N-glycoproteomics analysis with MSFragger-Glyco. Nat. Methods https://doi.org/10.1038/s41592-020-00967-9 (2020).
28. Bern, M., Kil, Y. J. & Becker, C. Byonic: advanced peptide and protein identification software. Curr. Proteom. Bioinform. 40, 13.20.1–13.20.14 (2012).
29. Bern, M., Cai, Y. & Goldberg, D. Lookspk peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal. Chem. 79, 1393–1400 (2007).
30. Malaker, S. A. et al. The mucin-selective protease StcE enables molecular and functional analysis of human cancer-associated mucins. Proc. Natl. Acad. Sci. USA 116, 7278–7287 (2019).
31. Choo, M. S., Pan, 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).
32. Klein, J. & Zaia, J. Relative retention time estimation improves N-glycopeptide identifications by LC–MS/MS. J. Proteome Res. 19, 2113–2121 (2020).
33. Khatri, K., Klein, J. A. & Zaia, J. Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. Anal. Bioanal. Chem. 409, 607–618 (2017).
34. Liu, M. Q. et al. Pol Glyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. Nat. Commun. 8, 438 (2017).
35. Lee, L. Y. et al. Toward automated N-glycopeptide identification in glycoproteomics. J. Proteome Res. 15, 3904–3915 (2016).
36. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. J. Am. Soc. Mass Spectrom. 27, 1719–1727 (2016).
37. Chalkley, R. J., Medizhradszky, K. F., Darula, Z., Pap, A. & Baker, P. R. The effectiveness of filtering glycopeptide peak list files for Y ions. Mol. Omics. 16, 147–155 (2020).
38. Baker, P. R., Trinidad, J. C. & Chalkley, R. J. Modification site localization scoring integrated into a search engine. Proteomics https://doi.org/10.1002/ proteomics.2011.M11.00878 (2011).
39. Park, G. W. et al. Classification of mucin-type O-glycopeptides using higher-energy collisional dissociation in mass spectrometry. Anal. Chem. 92, 9772–9781 (2020).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. © The Author(s), under exclusive licence to Springer Nature America, Inc. 2020
We provide the hypothetical example illustrated in Fig. 1b to aid understanding of the graph-theoretical model. The example O-glycopeptide contains eight O-glycosites. The glycan group consists of two glycans A and B. Either of the two glycans can occupy any one of the eight positions subject to the following requirements: a maximum of two glycans can be on the peptide, only one glycan is allowed per position; and each glycan can appear only once on a given peptide. For this example, there are 56 total (Supplementary Note 1) different modified forms in the graph. The weight of nodes vertically aligned is determined by the number of associated theoretical fragment ions. In the example, the nodes associated with amino acid S9 can be matched to theoretical fragments c9, c10, c11, 29, 210 and z11. The path highlighted in orange represents that the peptide is modified on S9 with glycan A and S12 with glycan B.

Site-specific localization probability. We use an iterative method to track the localization scores from all the potential paths of the graph to calculate site-specific localization probability of a glycoPSM. These scores are integrated with a random event-based localization method similar to a method described previously in PhosphoRS2. The integer part of the localization is the MetaMorpheus score, k, which is the total number of matched peaks. This is applied to a cumulative binomial distribution for calculating probability P as follows:

\[ P = \sum_{i=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \]

In the formula, n is the number of theoretical fragment ions and p is the probability of randomly matching a single theoretical fragment ion given spectral tolerances.

One difference from PhosphoRS is that the extracted peak depth is not optimized to achieve maximal differentiation. Finally, localization level is assigned by considering the ambiguity of paths, the matched fragment ions corresponding to each localized O-glycosite and the site-specific probabilities.

Data analysis. All searches were performed on a PC running Windows 10 Education (x1909), with two 2.20 GHz Intel Xeon Silver 4114 CPU processors with 64Gb of installed RAM. Up to 40 virtual processors were available to use for searching. Generally, 16 cores were used per search, but variations were used as described in the text. An O-glycan database of 12 common O-glycans was used for all searches, except for the 32-glycan database used for the urinary O-glycopeptide dataset as described in Fig. 2, which was compiled using literature sources10. Both glycan databases are provided as Supplementary Data 3. Data from these analyses are available in the Supplementary Information. A FASTA database of the four standard mucins used in the literature data (CD43, MUC16, PSGL-1 and Opa1b) were used for all searches unless otherwise noted, and known signaling peptide sequences were removed from the FASTA entries.

Byonic searching. All searches were performed with MetaMorpheus, an open-source search software useful for a variety of different applications including bottom-up, top-down, PTM discovery, crosslink analysis and label-free quantification. O-Pair is optimally designed for identifying O-glycopeptides from tethered collision- and electron-based dissociation spectra collected from the same precursor ion. However, it is also capable of identifying O-glycopeptides from spectra obtained using other fragmentation schemes and modalities. Before the beginning the open search, MetaMorpheus tracks precursors available from the data and also calculates precursors for potential co-isolated peptides16. O-Pair Search occurs in three stages (Fig. 1a): (1) identification of peptide candidates using an ion-indexed open search; (2) localization of O-glycosites with a graph-based localization algorithm and (3) calculation of site-specific localization probabilities. On completion of these stages, the O-glycopeptide localization levels (Fig. 1d) are determined and reported along with the FDRs, which are currently estimated using the target-decoy strategy.

Ion-indexed open search. O-Pair Search uses ion-indexed open search12 to quickly identify peptide candidates for each spectrum. O-glycosylation is a labile modification and O-glycopeptides under collision-based dissociation in MS generate peptide backbone fragment ions rarely retaining the glycans. Thus, even though an O-glycopeptide can be modified with multiple O-glycans, an O-glycopeptide HCD spectrum could be searched to determine the amino acid backbone without considering the O-glycans.

In an ion-indexed open search, a lookup table is created that includes a complete set of theoretical target and decoy fragment masses from the entire protein database, each labeled with the peptide from which it is derived. A collection of all peptides with fragments matching any peak in a given MS2 spectrum is assembled. The peptide candidates are then chosen from these peptides with the most matching fragments. The usage of an ion-indexed algorithm avoids unnecessary comparisons between experimental and theoretical spectra and makes it unnecessary to consider the variety of posttranslational modifications that might be present. The peptide candidates with the highest scores are retained for glycan identification and localization.

For each peptide candidate retained from the open search, the mass difference between the unmodified peptide backbone and the experimental precursor mass is computed. The mass difference is hypothesized to be the sum of all glycans masses on the peptide. We refer to the collection of glycans on a given peptides as the glycan group: mass of glycan group = precursor mass – peptide mass. All glycan groups whose mass equals the mass difference within the specified mass tolerance are considered as glycan group candidates for glycosite localization.

Graph-based localization. The graph algorithm is specially optimized for O-glycosite localization. A directed acyclic graph is constructed to represent all possible O-glycan modified forms of a peptide candidate and each of its corresponding glycan group candidates. If a peptide candidate corresponds to several different glycan group candidates within the mass tolerance limitation, several graphs are constructed.

The graph is constructed from left to right, beginning with a ‘Start’ node at the N-terminal side of the peptide and ending with an ‘End’ node at the C-terminal side. Nodes, vertically aligned, are added to the graph for each corresponding serine or threonine because these amino acids are the only two allowed for O-glycosylation. One vertical node represents an unoccupied and is labeled with ‘N’. Vertical nodes are then added, one for each potential glycan at the current position. These are labeled ‘A’, ‘B’ and so on. Additional vertical nodes are added representing combinations of glycans that may have occurred for the portion of the peptide represented by that vertical column of the graph. Combination nodes are labeled, for example, A + B. These nodes are added at each serine and threonine. Next, adjacent nodes are connected by edges representing the accumulation of glycans across the peptide backbone. Nodes that are not possible given the constraints of the total peptide mass, which stipulate the number and kinds of glycans on the peptide remain disconnected. This process culminates in a graph representing all possible glycopeptides, where each individual continuous path from Start to End represents one unique glycopeptide.

Next, we associate theoretical fragment ions with each node. Here we need to make clear which amino acids and glycans from the peptide are included. Beginning at the N terminus, the node represents the peptide up to and including the amino acid listed for the node. Beginning at the C terminus, the node represents the peptide up to, but not including, the amino acid listed for the node. The two portions of a peptide associated with a node are complementary to each other and do not cross over. Each node has associated with it all possible theoretical peptide fragment masses whose accumulated mass can be uniquely attributed to the glycopeptide segment containing the amino acids up to that point. The MetaMorpheus score for the entire peptide Tempo is calculated from all nodes in the path plus the fraction of spectrum intensity attributable to the matched fragments. The glycopeptide with the highest MetaMorpheus score can be extracted with dynamic programming and is designated as the match and reported in the results.
occur on a glycopeptide candidate and the number of times each O-glycan could occur per peptide. For most searches following the results obtained in Fig. 2g, the maximum Oglycan allowed was set to five unless otherwise noted. Under ‘search parameters,’ both ‘use provided precursor’ and ‘deconvolute precursors’ were checked. Peak trimming was not enabled and top N peaks and minimum ratio were set to 1,000 and 0.01, respectively. In silico digestion parameters were set to generate decoy proteins using reversed sequences, and the initiator methionine feature was set to ‘variable.’ The maximum modification isoforms allowed was 1,024, and the minimum and maximum peptide length values were set to 5 and 60, respectively. The protease was set to semitrypsin with two missed cleavages allowed, unless otherwise noted (Supplementary Fig. 4). The number of database partitions was set to one unless noted below. Precursor and product mass tolerances were 10 and 20 ppm, respectively, and the minimum score allowed was three. The maximum number of threads, that is, cores, was varied as described in the text, with 16 cores being the default used in this study unless otherwise noted. Modifications were set as carbamidomethyl on C as fixed, and oxidation on M and deamidation on N as variable.

O-Pair Search produces two separate PSM files, one for nonglycopeptides and one for glycopeptides. The numbers of nonglycopeptide identifications were calculated using filtering the single_peptide file to include only target PSMs with q values less than 0.01. The same target and q value filterings were used for O-glycopeptide identifications in the glyco_peptide file. Localization-level assignments were calculated using the provided outputs following target and q value filtering, and all were confirmed manually for data represented in Fig. 2a–l. The UpSet plot in Supplementary Fig. 5 was made using https://asntech.shinyapps.io/intervene/ (ref. 42).

Entrainment databases used for Fig. 2h were compiled from several different sources. The canonical mucin database (20 entries) was compiled using annotated mucins available at http://www.medkem.gu.se/mucinbiology/databases, ref. 43. The FBS database (86 entries) was generated from data provided in Shin et al. 44. The database of CD markers, that is, cluster of differentiation markers known to be cell surface molecules, was downloaded from the Human Protein Atlas (https://www.proteinatlas.org/). The E. coli, yeast and mouse proteome databases were retrieved from the Uniprot Consortium. 45 The CHO secretome was downloaded from Park et al. 46 Sequences for the four mucin standards in the mixture that was analyzed were appended to each entry. See Supplementary Note 3 for more discussion about the databases used. For searches performed with each of these databases, the number of database partitions was set to 16, and 16 cores were also used for each search.

The FDR was calculated after filtering target and q values < 0.01 in the glyco_peptide file, by taking the ratio of the total number of GlycoPSMs that did not originate from the four mucin standard proteins (false positives) to the total number of GlycoPSMs in the supplemental material in ref. 31, raw files 170919_11.raw, 170921_06.raw and 170922_04.raw correspond to Frac 1. Raw files 170919_08.raw and 170921_03.raw are the only two files available for download from MassIVE that are from Fraction 2. We processed those sets of three and two files as Fraction 1 and Fraction 2, respectively, and then processed all ten files available for download from MassIVE, as indicated in Fig. 2l. Identifications from Protein Prospector and Byonic provided in the supplemental material from ref. 47 were used from all three search conditions provided (described in detail in ref. 47), with duplicate identifications between the searches removed. To convert Protein Prospector identifications to our localization levels scheme, all identifications containing ‘@’ but not ‘|’ were classified as Level 1 or 1b, because @ indicates a modification assigned at a specific residue while | indicates an ambiguous assignment. Level 2 identifications were then added by including GlycoPSMs that included an @, whether or not other characters indicating ambiguity were present, because @ meant at least one modification was localized.

Statistical analysis. The q values for PSMs and peptide IDs are estimated using the target-decoy strategy from a search with concatenated targets and decoys. The q value calculation is the number of cumulative decoys (false positives) divided by the number of cumulative targets (true positives). The count of identified PSMs/peptides at a specified FDR is the count of true positive PSMs/peptides with q less than the specified FDR. Graphs were produced using Origin 2020. Graphs produced from these data replicates contained error bars of one standard deviation and have dot plots overlaid on top bar graphs to show individual replicate values.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The data used in this manuscript are available through the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD017646 (ref. 47) and via MassIVE with identifier MSV00083070 (ref. 48). Processed data using Byonic and Protein Prospector for the urinary O-glycopeptide dataset were downloaded from ref. 47.

Code availability
O-Pair Search is available in MetaMorphes (<v0.0.307 for HCD–ETD data and v0.0.308 for HCD–HCD and HCD–secHCD data), and is open source and freely available at https://github.com/smith-chem-wisc/MetaMorphes under a permissive license. All source code was written in Microsoft C# with .NET CORE 3.1 using Visual Studio.

References
40. Xu, G., Goonatilleke, E., Wongkham, S. & Lebrilla, C. B. Deep structural analysis and quantitation of O-linked glycans on cell membrane reveal high abundances and distinct glycomic profiles associated with cell type and stages of differentiation. Anal. Chem. 92, 3758–3768 (2020).
41. Wenger, C. D. & Coon, J. J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J. Proteome Res. 12, 1377–1386 (2013).
42. Khan, A. & Mathelier, A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. BMC Bioinformatics 18, 287, https://doi.org/10.1186/s12859-017-1708-7 (2017).
43. Lang, T. et al. Searching the evolutionary origin of epithelial mucus protein components—mucins and FCGRP. Mol. Biol. Evol. 33, 1921–1936 (2016).
44. Shin, J. et al. Use of composite protein database including search result sequences for mass spectrometric analysis of cell secretome. PLoS ONE 10, e0121692 (2015).
45. Uhlen, M. et al. Tissue-based map of the human proteome. Science 347, 1260419–1260419 (2015).
46. Bateman, A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D556–D515 (2019).
47. Park, J. H. et al. Proteomic analysis of host cell protein dynamics in the culture supernatants of antibody-producing CHO cells. Sci. Rep. 7, 44246 (2017).
48. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 47, D442–D450 (2019).

Acknowledgements
We appreciate discussions with Z. Rolfs, R.J. Millikin and other Smith group members to enhance software analysis speed and address challenges in implementing ideas. This work was supported by National Institute of Health (NIH) grant no. R35 GM126914 awarded to L.M.S. and grant no. R01 CA200423 awarded to C.R.B., as well as with support from the Howard Hughes Medical Institute. N.M.R. was funded through an NIH Predoctoral to Postdoctoral Transition Award (grant no. K00 CA212454-03).

Author contributions
L.L. and N.M.R. contributed equally to this work. L.L. conceived the project and software design, wrote software, analyzed data and wrote the paper. N.M.R. conceived the project and software design, advised on software development, analyzed most of the data and wrote the paper. M.R.S. designed software and supervised the project. C.R.B. and L.M.S. supervised the project. All authors discussed results and edited the paper.

Competing interests
C.R.B. is a cofounder and Scientific Advisory Board member of Lycia Therapeutics, Palleo Pharmaceuticals, Enable Bioscience, Redwood Biosciences (a subsidiary of Catalent) and InterVenn Biosciences, and a member of the Board of Directors of Eli Lilly & Company.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-09885-5.
Correspondence and requests for materials should be addressed to L.M.S.
Reprints and permissions information is available at www.nature.com/reprints.
Editor recognition statement Arunima Singh was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a  Confirmed

☐  The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐  A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐  The statistical test(s) used AND whether they are one- or two-sided

Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐  A description of all covariates tested

☐  A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐  A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐  For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

Give P values as exact values whenever suitable.

☐  For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐  For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐  Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about: availability of computer code

Data collection  O-Pair search tool were benchmarked using publicly available data.

Data analysis  MS data processing was performed using O-Pair (MetaMorpher 0.0.307 and 0.0.308, https://github.com/nera-chicago/MetaMorpher), Byonic (Version 3.7.4). Several plots were prepared using Origin 2020. MetaMorpher is developed with C# 8.0 with .NET CORE 3.1 using Visual Studio 2019.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about: availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data that support the findings of this study are public available.

The raw data are available from the ProteomeXchange (http://www.proteomexchange.org/) Consortium via the PRIDE partner repository with the dataset identifier PXD017646 and from MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp) with the project identifier MSV000083070.

The processed data using Byonic and Protein Prospector for the urinary O-glycoside data set was downloaded from https://doi.org/10.1021/jasms.8b05844.

The canonical mucin database (20 entries) was compiled using annotated mucins available at http://www.mediprima.com/mucinbiology/databases. The FBS database (86 entries) was generated from data provided from (https://doi.org/10.1371/journal.pone.0121692). The database of CD markers, i.e., cluster of...
differentiation markers known to be cell surface molecules, was downloaded from the Human Protein Atlas [https://www.proteinatlas.org/]. The E. coli, yeast, and mouse proteome databases were retrieved from the Uniprot Consortium. The CHO secretome was downloaded from (doi:10.1038/srep44246) Sequences for the four mucin standards in the mixture that was analyzed were appended to each. The database files are supplied as supplementary data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No statistical methods were used to predetermine sample sizes. The software benchmarks assessed the numbers of glycopeptides identified from different mass spectrometry datasets, including datasets originated from 4 standard mucin proteins and human urinary samples. Protein databases with different sizes from focused database to large entrap database are used. In all cases, the benchmarks follow standard practice in the field and were considered sufficient.

Data exclusions

No data were excluded from the analysis.

Replication

Three technical data replicates of electron-driven dissociation and higher collision dissociation data are used in two parts of analysis. All attempts at replication were successful.

Randomization

Randomization is not relevant for this study. The study presents computational benchmarks, wherein the same data sets was analyzed using different software tools.

Blinding

Blinding is not relevant for this study. The study does not involve any experiments. The data processing was performed in an unbiased automatic way.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a

Involved in the study

- Antibodies
- Eukaryotic cell lines
- Palaeontology and archaeology
- Animals and other organisms
- Human research participants
- Clinical data
- Dual use research of concern

Methods

n/a

Involved in the study

- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging